

# PoseAnchor: Robust Root Position Estimation for 3D Human Pose Estimation

Jun-Hee Kim\*

Jumin Han

Seong-Whan Lee<sup>†</sup>

Korea University

{jh\_kim, juminhan, sw.lee}@korea.ac.kr

## Abstract

Standard 3D human pose estimation (HPE) benchmarks employ root-centering, which normalizes poses relative to the pelvis but discards absolute root position information. While effective for evaluation, this approach limits real-world applications such as motion tracking, AR/VR, and human-computer interaction, where absolute root position is essential. Moreover, incorporating root position into these models often leads to performance degradation. To address these limitations, we introduce PoseAnchor, a unified framework that seamlessly integrates root position estimation while improving overall pose accuracy. PoseAnchor leverages *Iterative Hard Thresholding Robust Least Squares Regression (ITRR)*, a novel robust regression approach introduced to 3D HPE for the first time. ITRR effectively mitigates the impact of noisy 2D detections, enabling more accurate root position estimation. With ITRR, PoseAnchor enables zero-shot root localization, allowing existing models to estimate absolute root positions without retraining or architectural modifications. ITRR identifies a support set of reliable joints based on their spatial relationships to achieve robust root estimation, effectively filtering out unreliable joints. Beyond zero-shot localization, PoseAnchor incorporates ITRR into a Data-Driven Training framework that selectively utilizes the support set to optimize pose learning. By dynamically filtering high-confidence joint data, PoseAnchor mitigates noise while improving robustness.

## 1. Introduction

3D Human Pose Estimation (HPE) has made significant progress; however, benchmark-centric evaluation methods often fall short in real-world applicability. The widely adopted ‘root-centering’ technique [8, 12, 23, 33, 40], while effective for benchmark performance, inherently discards root position information, which is crucial for practical deployment. While root-centered representation improves benchmark accuracy, it is like driving an autonomous vehicle

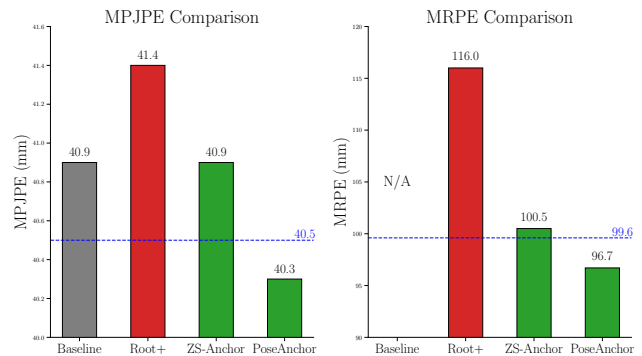


Figure 1. Mean Per Joint Position Error (MPJPE) and Mean Root Position Error (MRPE) comparison with the baseline [40]. The blue dashed line indicates the state-of-the-art (SOTA) performance threshold, based on results reported in previous works [33, 39].

without GPS—highlighting the need for absolute position information in real-world applications such as motion tracking and AR/VR.

As shown in Figure 1, the conventional ‘Baseline’ method, which relies on root-centering, is incapable of predicting root position, leading to ‘Not Applicable (N/A)’ MRPE values. This highlights that root-centering leads to a loss of root position information, which is crucial for practical deployment. Furthermore, incorporating root position prediction into root-centered models often results in performance degradation. As shown in the MPJPE comparison in Figure 1, the ‘Root+’ method, which extends the Baseline to predict root position, shows increase in MPJPE. This underscores the inherent challenge of adapting benchmark-driven methodologies to real-world deployment—balancing root position prediction with pose accuracy remains a significant hurdle. This performance degradation highlights the inherent conflict between learning relative joint relationships in root-centered models and predicting absolute root positions, disrupting the optimized representation for relative poses.

To overcome these challenges and enhance the real-world applicability of 3D HPE, we introduce PoseAnchor, a unified framework consisting of Zero-Shot Root Localization (ZS-anchor) and Data-Driven Training. PoseAnchor is a model-agnostic framework, readily integrable into various 3D HPE models without being confined to specific architec-

\*Code: [github.com/junheeeeeee/PoseAnchor](https://github.com/junheeeeeee/PoseAnchor)

<sup>†</sup>Corresponding author

tures. Unlike previous studies, which are often tailored to specific architectures, PoseAnchor provides a flexible and generalizable solution for practical 3D HPE systems. ZS-anchor estimates the root position from the predicted 3D poses without the need for additional network components or complex architectures. The root position estimation is formulated as a robust optimization problem and is solved using Iterative Hard Thresholding Robust Least Squares Regression (ITRR), which is resilient to noise and outliers. ITRR’s strength lies in its iterative refinement process, where it adaptively identifies and leverages a support set of reliable joints based on spatial consistency, effectively mitigating the influence of noisy or inaccurate joint detections commonly found in 3D HPE. Here, PoseAnchor goes a step further by utilizing the support set to filter out joint data with low reliability, integrating ITRR into data-driven training. This approach improves 3D pose accuracy and enhances root position estimation. This Data-Driven Training framework, leveraging ITRR-based support set filtering, focuses the learning process on high-confidence joint data, leading to more stable and generalizable models compared to traditional methods that treat all joint data equally.

Empirical evaluations confirm the effectiveness of PoseAnchor. Figure 1 shows that both ZS-anchor and PoseAnchor consistently outperform root-centering methods in terms of accuracy and robustness. Unlike methods that degrade when incorporating root position, PoseAnchor enhances pose estimation accuracy. These results suggest that PoseAnchor offers a practical approach to bridging the gap between benchmark research and real-world applications, increasing the applicability of 3D HPE.

Key contributions are summarized as follows:

1. **Zero-Shot Root Position Estimation:** We introduce ZS-Anchor, a zero-shot root position estimation method that enables existing 3D pose estimation models to estimate absolute root positions without retraining. This eliminates the need for supervised learning of root positions, allowing models to infer absolute root positions directly from root-centered poses.
2. **Support Set Guided Training:** We propose a data-driven training framework that leverages ITRR-based support set filtering. This mechanism effectively removes noisy joint data, stabilizes training, and enhances both pose estimation accuracy and robustness.
3. **Model-Agnostic Framework:** PoseAnchor is a model-agnostic framework that seamlessly integrates into various 3D HPE architectures without requiring modifications. This ensures broad compatibility and enhances real-world applicability.

## 2. Related Work

### 2.1. 3D Human Pose Estimation

Early 3D pose estimation methods regressed 3D coordinates from images [3, 34], but advances in 2D pose estimation [6, 32, 38] have made lifting 2D poses to 3D the dominant approach [16, 23] for its effectiveness and scalability. This decomposition improved accuracy, especially when using temporal modeling [23]. A majority of 3D pose estimation methods adopt *root-centering*, a technique that positions the pelvis joint at the origin, thereby eliminating global translation [18, 33, 40, 44]. Although effective for benchmarks, root-centering discards absolute position information, restricting its applicability to real-world tasks such as motion tracking and AR/VR. To address this, some works [20, 24, 42] directly estimate the global root position. RootNet [20] predicts the root depth separately, while LCR-Net [24] and SMAP [42] estimate absolute poses for multi-person scenarios. More recently, Ray3D [39] formulated root position estimation as a ray-based optimization problem, predicting depth along 3D rays derived from 2D keypoints and camera intrinsics. This approach improves generalization across different viewpoints, highlighting the importance of integrating camera-aware root estimation. Despite their ability to estimate absolute root position, these methods often necessitate additional supervision or introduce increased model complexity. Our approach seamlessly integrates root position estimation without requiring additional supervision, all while maintaining robustness.

### 2.2. Robust Least Squares Regression

Robust Least Squares Regression (RLSR) mitigates the sensitivity of traditional regression to noise and outliers, which are common in 2D-to-3D pose lifting. M-estimators [9], RANSAC [7], and LMS regression [25] enhance robustness by reducing the influence of outliers. Black and Rangarajan [2] further introduced a non-convex loss function widely used in vision tasks. More recently, robust regression has been applied in diverse fields such as signal processing [31, 45], economics [26], bioinformatics [15], and image processing [21, 36]. However, classical methods [10, 29, 30] often lack theoretical recovery guarantees. Trimmed inner product [5] and sub-sampling [17] approaches have been explored, but their performance is limited in large datasets. To achieve exact recovery, L1 penalty-based convex formulations [22, 35] impose strict assumptions on data distributions. Hard-thresholding methods [1] offer guarantees but depend on a user-specified corruption ratio, which is difficult to estimate and affects performance when mis-specified. Our approach refines these principles by leveraging thresholding-based robust regression to enhance root position estimation while selectively utilizing high-confidence joint data.

### 3. Methods

We propose a novel method for accurately and robustly estimating the absolute root (pelvis center) position in global 3D space, utilizing root-centered 3D poses predicted by a 3D Human Pose Estimation (HPE) model. Our approach is built upon Robust Least Squares Regression (RLSR), a technique well-known for its resilience to noise and outliers. Specifically, we employ an Iterative Hard Thresholding (ITRR) algorithm to effectively mitigate errors inherent in 3D pose estimation and 2D joint detection. To improve robustness against noise and outliers, we introduce a support set  $S$  that selectively utilizes high-confidence joint data. This ensures a more accurate estimation of the global root position. Additionally, we incorporate this root position estimation method directly into the training pipeline of the 3D HPE model, enhancing the model's capability to predict global positions while improving pose accuracy.

#### 3.1. Problem Formulation

The 3D root-centered coordinates of each joint, denoted as  $\{(X_j, Y_j, Z_j)\}_{j=1}^J$ , can be mathematically related to the root position  $(X_r, Y_r, Z_r)$  and the 2D projection coordinates  $(x_j, y_j)$  via the camera's intrinsic projection model. The projection relationship for each joint  $j$  is defined as:

$$\begin{aligned} x_j &= \frac{f_x(X_j + X_r)}{Z_j + Z_r}, \\ y_j &= \frac{f_y(Y_j + Y_r)}{Z_j + Z_r}, \end{aligned} \quad (1)$$

where  $f_x$  and  $f_y$  represent the camera's intrinsic focal lengths. By rearranging these equations to isolate  $(X_r, Y_r, Z_r)$ , we obtain:

$$\begin{aligned} x_j Z_r - f_x X_r &= f_x X_j - x_j Z_j, \\ y_j Z_r - f_y Y_r &= f_y Y_j - y_j Z_j. \end{aligned} \quad (2)$$

By extending this formulation to all  $J$  joints, we represent the problem in matrix form as:

$$A r = b, \quad (3)$$

where the root position  $r = [X_r, Y_r, Z_r]^T \in \mathbb{R}^3$  represents the 3D root position. The coefficient matrix  $A \in \mathbb{R}^{2J \times 3}$  and the target vector  $b \in \mathbb{R}^{2J}$  are given by:

$$A = \begin{bmatrix} -f_x & 0 & x_1 \\ 0 & -f_y & y_1 \\ -f_x & 0 & x_2 \\ 0 & -f_y & y_2 \\ \vdots & \vdots & \vdots \\ -f_x & 0 & x_J \\ 0 & -f_y & y_J \end{bmatrix}, \quad b = \begin{bmatrix} f_x X_1 - x_1 Z_1 \\ f_y Y_1 - y_1 Z_1 \\ f_x X_2 - x_2 Z_2 \\ f_y Y_2 - y_2 Z_2 \\ \vdots \\ f_x X_J - x_J Z_J \\ f_y Y_J - y_J Z_J \end{bmatrix}. \quad (4)$$

This compact formulation encapsulates the projection constraints of all  $J$  joints by linking the intrinsic camera parameters  $(f_x, f_y)$ , the 2D joint coordinates  $(x_j, y_j)$ , and the corresponding components of the 3D joint positions  $(X_j, Y_j, Z_j)$  into a single linear system.

In an ideal, noise-free setting, solving the least-squares problem

$$r^* = \operatorname{argmin}_{r \in \mathbb{R}^3} \|b^* - A^* r\|_2^2 \quad (5)$$

where  $r^*$  denotes the true root position,  $A^*$  is the ideal coefficient matrix, and  $b^*$  is the ideal target vector, yields the correct root position. However, in practical applications, errors in both 2D joint detection and predicted 3D coordinates inevitably arise, significantly impacting global position accuracy. Considering these joint-specific errors, we model the actual scenario more realistically as:

$$b + \Delta b = (A + \Delta A)r^*, \quad (6)$$

where  $\Delta b$  represents the error in  $b$ , and  $\Delta A$  represents the error in the matrix  $A$ . From Eq.(6), we can define the adapted true target vector for  $A$  as  $\eta^* \triangleq b + \Delta b - \Delta A r^*$ , which represents the true projection relationship given the noisy matrix  $A$  and the true root position  $r^*$ . We assume the target vector  $b$  is generated by:

$$b = \eta^* + \epsilon + c, \quad \text{where } \eta^* = A r^* \quad (7)$$

This model accounts for two types of perturbations in  $\eta_j^*$ : (1) dense but bounded noise  $\epsilon_j$  stemming from 2D joint detection and 3D pose prediction errors, and (2) potentially unbounded corruptions  $c_j$  caused by misdetections or severe joint estimation failures. Dense noise accounts for minor inaccuracies in 2D joint detections and predicted 3D joints, while sparse corruptions represent gross outliers or misdetections. We model the dense noise as  $\epsilon \sim \mathcal{N}(0, \sigma^2 I_{2J \times 2J})$ , where  $\sigma^2$  represents the variance of the dense noise. The corruption vector  $c$  is assumed to be sparse, satisfying  $\|c\|_0 \leq \alpha \cdot 2J$ , for some corruption index  $0 < \alpha < 1$ . These corruptions can be caused by outliers in the 2D joint detections or errors in the 3D pose estimation pipeline. For example, a misdetection of a limb due to occlusion can lead to a significant corruption in the corresponding projection. In such cases, the corresponding element in the corruption vector  $c$  will have a large value, indicating a significant deviation from the expected projection.

Therefore, the key challenge is to accurately estimate the true root position  $r^*$  while minimizing the adverse effects of noise and outliers present in individual joint data on global position estimation. To address this, we apply Robust Least Squares Regression (RLSR), a technique specifically designed to ensure robustness of individual data points (projection constraints of each joint). RLSR employs a robust loss function that is less sensitive to outliers or uses techniques to identify and exclude corrupted data points, allowing for more accurate regression in the presence of noise and

corruption. We formulate the root position estimation as the following optimization problem:

$$(\hat{r}, \hat{S}) = \underset{\substack{r \in \mathbb{R}^3 \\ S \subseteq \{1, 2, \dots, 2J\}: |S| \geq (1-\beta) \cdot 2J}}{\operatorname{argmin}} \sum_{i \in S} (b_i - A_{i,:} r)^2, \quad (8)$$

where  $\hat{S}$  denotes the set of corruption-free points,  $b_i$  is the  $i$ -th element of the aligned target vector  $\eta$ , and  $A_{i,:}$  is the  $i$ -th row of the coefficient matrix  $A$ . Here, the core element is the support set  $\hat{S}$ .  $\hat{S}$  represents the set of inlier data points among all joint data, i.e., data points with high confidence.

### 3.2. Robust Least Squares Regression for 3D HPE

To estimate the root position  $r^*$  in the presence of dense noise and sparse corruptions, we employ a thresholding operator-based robust regression method. This approach is inspired by the robust regression framework introduced by [1], which provides theoretical guarantees for convergence even in adversarial corruption settings. Unlike standard least squares regression, which is highly sensitive to outliers, our method utilizes a hard thresholding operator to identify and minimize the influence of corrupted observations in individual joint data. By selectively leveraging only high-confidence joint data that contributes to global position estimation, we maximize robustness and stability.

#### 3.2.1. Iterative Hard Thresholding Robust Regression

Given the observed alignment vector  $\eta$ , our goal is to solve the regression problem in Eq. (5) while discarding severely corrupted elements. To achieve this, we define the hard thresholding operator, which selects a subset of clean data points by filtering out observations with high residual errors. For a vector  $v \in \mathbb{R}^{2J}$ , the hard thresholding operator is defined as:

$$\operatorname{HT}(v, k) = \{i \in \{1, \dots, 2J\} : \sigma_v^{-1}(i) \leq k\}, \quad (9)$$

where  $\sigma_v$  is the permutation that orders elements of  $v$  in ascending order of their absolute magnitude. In essence,  $\operatorname{HT}(v, k)$  returns the indices of the top  $k$  elements with the smallest absolute values in vector  $v$ . In our research, we apply the hard thresholding operator based on residuals of individual joint data, selecting the top  $k$  joint data points with smaller residuals (i.e., those that align well with model predictions).

To solve the robust regression problem in Eq. (8), we employ an iterative hard thresholding approach, which progressively refines the estimation by eliminating unreliable data points. This method achieves robust global position estimation through a process of repeatedly selecting high-confidence data (support set  $S$ ) based on residuals of individual joint data, and iteratively updating the root position using only the selected data. Algorithm 2 details this process,

---

#### Algorithm 1 Least Squares Regression (LSR)

---

**Require:** Target vector  $b$ , coefficient matrix  $A$

**Ensure:** Estimated root position  $\hat{r}$ , residuals  $e$

1: Compute least-squares solution:

$$\hat{r} = \underset{r \in \mathbb{R}^3}{\operatorname{argmin}} \|b - Ar\|_2^2$$

2: Compute residuals:  $e = b - A\hat{r}$

3: **Return:**  $\hat{r}, e$

---

#### Algorithm 2 Iterative hard Thresholding Robust Regression (ITRR)

---

**Require:** Target vector  $b$ , coefficient matrix  $A$ , threshold parameter  $\tau$ , convergence tolerance  $\epsilon$

**Ensure:** Estimated root position  $\hat{r}$ , selected support set  $\hat{S}$

1: Initialize using Least Squares (Algorithm 1):

$$r_0, e_0 = \operatorname{LSR}(A, b)$$

2: Remove large residuals:  $S_0 = \{i : |e_0[i]| < \tau\}$

3: Compute removal ratio:  $\beta = 1 - \frac{|S_0|}{|S|}$

4: **while**  $\|e_t[S_t]\|_2 > \epsilon$  **do**

5:   Compute least-squares update:

$$r_{t+1} = \underset{r \in \mathbb{R}^3}{\operatorname{argmin}} \sum_{i \in S_t} (b_i - A_{i,:} r)^2$$

6:   Compute residuals:  $e_{t+1} = b - Ar_{t+1}$

7:   Update support set:

$$S_{t+1} = \operatorname{HT}(e_{t+1}, (1 - \beta)2J)$$

8:    $t \leftarrow t + 1$

9: **end while**

10: **Return:**  $\hat{r} = r_t, \hat{S} = S_t$

---

efficiently removing outliers and refining the root position estimate in each iteration.

Algorithm 2 iteratively evaluates the reliability of individual joint data and updates the root position focusing on high-confidence joint data, enabling highly robust global position estimation against noise and outliers, especially errors that may occur in individual joints.

#### 3.2.2. Theoretical Robustness Guarantee

Building on the theoretical framework in [1], we assume that the noise  $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$  is Gaussian, and the corruption vector  $c$  is sparse, satisfying  $\|c\|_0 \leq \alpha 2J$ .

To guarantee robust recovery of  $r^*$  under noise and corruption, we define the constants characterizing the Subset Strong Convexity (SSC) and Subset Strong Smoothness (SSS) properties of  $A$ :

$$\lambda_\gamma \leq \min_{S \in S_\gamma} \lambda_{\min}(A_S A_S^T) \leq \max_{S \in S_\gamma} \lambda_{\max}(A_S A_S^T) \leq \Lambda_\gamma \quad (10)$$

where  $S_\gamma$  represents all subsets of  $[n]$  of size  $\gamma n$ , and  $A_S$  is the corresponding submatrix of  $A$ . The terms  $\lambda_{\min}(\cdot)$  and  $\lambda_{\max}(\cdot)$  denote the smallest and largest eigenvalues, respectively. These constants help subsets of size  $\gamma \times n$  maintain a well-conditioned state, influencing numerical stability and robustness against adversarial corruption. Under these conditions, our method guarantees geometric convergence to  $r^*$  if:

$$\frac{4\sqrt{\Lambda_\beta}}{\sqrt{\lambda_{1-\beta}}} < 1. \quad (11)$$

This ensures reliable recovery even in the presence of dense noise and sparse corruption. For a more detailed proof of these theoretical guarantees, please refer to the Supplementary Material.

### 3.3. PoseAnchor Framework

Our framework consists of two key components: ZS-Anchor and PoseAnchor. Both ZS-Anchor and PoseAnchor perform root position estimation. ZS-Anchor enables root localization without retraining, while PoseAnchor further integrates root position learning into the training process. Instead of treating root position estimation as a separate post-processing step, our approach optimizes both root-centered 3D poses and absolute root positions within a unified framework.

#### 3.3.1. Estimating Root Position Using a Root-Centered 3D HPE Model

To obtain the predicted root-centered 3D pose  $P$ , we leverage a 3D Human Pose Estimation (HPE) model. The model takes as input a sequence of 2D joint detections  $\{(x_j, y_j)\}_{j=1}^J$  and extracts spatiotemporal features to infer the 3D joint coordinates:

$$P = f(\{(x_j, y_j)\}_{j=1}^J) \quad (12)$$

where  $f$  represents the 3D pose estimation model. To ensure the network estimates both relative and absolute positioning, we incorporate root position estimation into the framework by leveraging projection constraints of individual joint data:

$$\hat{r}, \hat{S} = \text{ITRR}(A, b, \tau, \epsilon) \quad (13)$$

where  $\text{ITRR}(\cdot)$  denotes the Iterative hard Thresholding Robust Regression method (Algorithm 2). The parameters  $\tau$  and  $\epsilon$  control the thresholding level and convergence tolerance respectively. Detailed justification and sensitivity

analysis for these parameters are provided in the Supplementary Material. Given the predicted root-centered 3D pose  $P$  and corresponding 2D joint detections, the root position  $\hat{r}$  is computed by solving the robust least-squares regression problem using Algorithm 2. In this process, we evaluate the reliability of individual joint data and minimize the impact of low-reliability data, ensuring robustness in global position estimation.

#### 3.3.2. Global position Reconstruction

During inference, the estimated root position is added back to the root-centered 3D coordinates to reconstruct the absolute 3D pose:

$$P_j^{\text{abs}} = P_j + \hat{r}, \quad \forall j \in \{1, \dots, J\}. \quad (14)$$

This integration ensures that the model not only predicts accurate 3D poses but also recovers global positioning information, making it suitable for real-world applications like motion tracking and human-object interaction. This capability, achieved without requiring retraining, defines ZS-Anchor. Building upon this, PoseAnchor further enhances robustness by incorporating selective training via support set.

#### 3.3.3. Selective Training via Support Set

For training, our approach leverages the original pose loss formulation, maintaining the established objectives of 3D human pose estimation. However, instead of applying this loss to all data points indiscriminately, we introduce a selective sampling mechanism based on the support set  $\hat{S}$ . Specifically, the loss is computed exclusively over the joints within the support set  $\hat{S}$ , which comprises high-confidence joint data points identified by our ITRR algorithm. This means we are sampling from the full dataset, focusing our learning signal on the most reliable instances. The loss function thus becomes:

$$\mathcal{L} = \sum_{j \in \hat{S}} \mathcal{L}_{\text{pose}}(P_j^{\text{abs}}, P_{j,\text{gt}}^{\text{abs}}). \quad (15)$$

$P_{\text{gt},j}^{\text{abs}}$  denotes the ground-truth absolute 3D pose for joint  $j$ .  $\mathcal{L}_{\text{pose}}$  represents the original pose loss formulation used in the baseline 3D HPE models. The core principle is to retain the original loss function  $\mathcal{L}_{\text{pose}}$  while strategically sampling data points for its application. By focusing training on the high-confidence subset  $\hat{S}$ , we ensure that learning is driven by the reliable signals, effectively filtering out noisy or corrupted contributions and enhancing the robustness of the trained model without requiring any additional ground truth beyond the standard pose annotations. By incorporating the support set  $\hat{S}$  into the training process, we focus learning on high-confidence joint data, thereby enhancing robustness against noisy or corrupted inputs.

## 4. Experiments

Unless explicitly stated otherwise, results presented are obtained using the MixSTE [40] model architecture as our primary baseline. We chose a strong but not state-of-the-art baseline to better demonstrate the general applicability and extensibility of our PoseAnchor framework.

### 4.1. Datasets

We validated our approach on the following widely recognized 3D human pose estimation datasets: Human3.6M (H36M) [11], a large-scale dataset comprising 3D human poses captured indoors under controlled conditions using a motion capture system, and MPI-INF-3DHP (3DHP) [19], a dataset featuring 3D human poses in both indoor and outdoor settings, providing more diverse and challenging real-world scenarios. For each dataset, we followed the standard train/test splits as in prior research. All experiments use CPN detector [6] for 2D keypoints, except Table 4, which uses ground-truth.

### 4.2. Evaluation Metrics

We evaluate our method using standard metrics for 3D human pose estimation: Mean Per Joint Position Error (MPJPE), the average Euclidean distance between predicted and ground-truth 3D joint locations after root alignment; Procrustes-aligned MPJPE (P-MPJPE), computed after optimal scaling, rotation, and translation; Absolute MPJPE (Abs-MPJPE), calculated without root alignment; Mean Root Position Error (MRPE), the average error of predicted root positions; Percentage of Correct Keypoints (PCK, %), the proportion of keypoints within 150 mm of ground-truth locations; and Area Under the Curve (AUC), the integral of the PCK curve over various thresholds. These metrics provide a comprehensive assessment of pose estimation accuracy and robustness.

### 4.3. Results

#### 4.3.1. Results on H36M

Table 1 provides a comprehensive quantitative comparison on the Human3.6M dataset. The top section reports MPJPE, where our PoseAnchor achieves superior joint position accuracy in root-centered 3D space without rigid alignment. The middle section presents P-MPJPE, showing that PoseAnchor maintains high shape accuracy, comparable to top-performing methods. The bottom section summarizes MRPE and Abs-MPJPE, where PoseAnchor achieves the lowest root position error and leading absolute 3D pose accuracy. Overall, these results demonstrate that PoseAnchor consistently outperforms or matches state-of-the-art methods across all key metrics (MPJPE, P-MPJPE, MRPE, and Abs-MPJPE) on H36M, highlighting its effectiveness in leveraging root position information for robust 3D pose estimation.

#### 4.3.2. Results on 3DHP

Table 2 presents quantitative results on the challenging MPI-INF-3DHP dataset, highlighting the robustness of PoseAnchor. While MotionAGFormer [18] achieves a marginally lower MPJPE, PoseAnchor attains state-of-the-art PCK (99.3%) and AUC (88.1%), both the highest among compared methods. Unlike MPJPE, which is sensitive to outliers, PCK and AUC provide a more robust evaluation by reducing the impact of occasional joint errors. The superior PCK and AUC scores of PoseAnchor indicate more precise and consistent joint localization within standard thresholds, even under pose variations and noise. Although PoseAnchor’s MPJPE of 17.2 mm is not the absolute lowest, it remains highly competitive given its exceptional robustness. Overall, these results underscore the reliability of PoseAnchor for 3D human pose estimation in demanding scenarios where robustness and resilience to outliers are essential.

### 4.4. Ablation Studies

#### 4.4.1. Model Agnostic Applicability

To demonstrate the model-agnostic applicability of PoseAnchor, we evaluate its performance when integrated with different baseline 3D HPE models. Specifically, we compare across four diverse architectures: Martinez et al.[16] (MLP-based), VPose[23] (CNN-based), GLA-GCN [37] (GCN-based), and MixSTE [40] (Transformer-based). For each baseline, we consider the following four approaches:

- **Baseline (Root-Centered Model):** A standard root-centered approach [40] that removes absolute root position information before pose estimation, following common 3D pose benchmarks.
- **Root+ (Root-Aware Learning):** The baseline model with its output dimension augmented to jointly regress the 3D root position and the relative 3D pose from 2D keypoints, retrained end-to-end.
- **ZS-anchor:** A zero-shot root position estimation method that uses ITRR to infer the root position through the baseline model, without additional training.
- **PoseAnchor:** Our proposed approach, which integrates root position estimation into a data-driven framework.

Table 3 compares the performance of four approaches—Baseline, Root+, ZS-anchor, and PoseAnchor—across various 3D human pose estimation architectures. The evaluation considers two categories of metrics: root-relative (MPJPE, P-MPJPE) and absolute (Abs-MPJPE, MRPE). The results demonstrate that PoseAnchor consistently achieves the best performance, significantly improving absolute pose estimation while maintaining or enhancing root-relative accuracy.

The Baseline method, following a root-centered approach, inherently sacrifices absolute pose information for benchmark performance. Root+, while enabling absolute pose

Table 1. Detailed quantitative performance comparison on the H36M dataset.  $T$  denotes the number of input frames. ( $\dagger$ ) indicates methods originally using GT 2.5D factor (removed for fair comparison). (\*) indicates the baseline method. The best and second-best results are highlighted in bold and underlined formats respectively.

MPJPE	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Vpose [23]( $T=243$ )	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
Liu <i>et al.</i> [14]( $T=243$ )	41.8	44.8	41.1	44.9	47.4	54.1	43.4	42.2	56.2	63.6	45.3	43.5	45.3	31.3	32.2	45.1
Chen <i>et al.</i> [4]( $T=243$ )	41.4	43.5	40.1	42.9	46.6	51.9	41.7	42.3	53.9	60.2	45.4	41.7	46.0	31.5	32.7	44.1
PoseFormer [43]( $T=81$ )	41.5	44.8	39.8	42.5	46.5	51.6	42.1	42.0	53.3	60.7	45.5	43.3	46.1	31.8	32.2	44.3
MixSTE [40]( $T=243$ )(*)	<b>37.6</b>	<u>40.9</u>	37.3	39.7	<b>42.3</b>	49.9	40.1	39.8	<u>51.7</u>	<b>55.0</b>	42.1	39.8	41.0	27.9	27.9	40.9
P-STMO [28]( $T=243$ )	38.9	42.7	40.4	41.1	45.6	<u>49.7</u>	40.9	39.9	55.5	59.4	44.9	42.2	42.7	29.4	29.4	42.8
PoseFormerV2 [41]( $T=243$ )	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	45.2
STCFormer-L [33]( $T=243$ )	38.4	41.2	<u>36.8</u>	<b>38.0</b>	42.7	50.5	<b>38.7</b>	<u>38.2</u>	52.5	56.8	<b>41.8</b>	<b>38.4</b>	<b>40.2</b>	<b>26.2</b>	<b>27.7</b>	<u>40.5</u>
GLA-GCN [37]( $T=243$ )	41.3	44.3	40.8	41.8	45.9	54.1	42.1	41.5	57.8	62.9	45.0	42.8	45.9	29.4	29.9	44.4
MotionBERT [44]( $T=243$ )( $\dagger$ )	39.2	41.0	37.9	39.6	43.3	49.9	40.3	40.0	52.4	58.0	43.6	41.5	41.7	28.4	28.6	41.7
MotionAGFormer-L [18]( $T=243$ )( $\dagger$ )	38.0	41.3	37.8	39.6	43.6	49.7	42.0	40.3	51.1	57.9	42.9	39.8	41.5	29.3	29.0	41.6
PoseAnchor ( $T=243$ )	<u>37.8</u>	<b>39.8</b>	<b>36.6</b>	<u>38.8</u>	<u>42.6</u>	<b>48.6</b>	<u>39.4</u>	<b>38.1</b>	<b>50.5</b>	<u>55.1</u>	<u>42.0</u>	<u>39.1</u>	<u>40.4</u>	<u>27.8</u>	<u>27.8</u>	<b>40.3</b>
P-MPJPE	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Vpose [23]( $T=243$ )	34.2	36.8	33.9	37.5	37.1	43.2	34.4	33.5	45.3	52.7	37.7	34.1	38.0	25.8	27.7	36.8
Liu <i>et al.</i> [14]( $T=243$ )	32.3	35.2	33.3	35.8	35.9	41.5	33.2	32.7	44.6	50.9	37.0	32.4	37.0	25.2	27.2	35.6
Chen <i>et al.</i> [4]( $T=243$ )	32.6	35.1	32.8	35.4	36.3	40.4	32.4	32.3	42.7	49.0	36.8	32.4	36.0	24.9	26.5	35.0
PoseFormer [43]( $T=81$ )	34.1	36.1	34.4	37.2	36.4	42.2	34.4	33.6	45.0	52.5	37.4	33.8	37.8	25.6	27.3	36.5
MixSTE [40]( $T=243$ )(*)	30.8	33.1	30.3	31.8	<u>33.1</u>	39.1	31.1	30.5	42.5	<u>44.5</u>	<u>34.0</u>	30.8	32.7	22.1	22.9	32.6
P-STMO [28]( $T=243$ )	31.3	35.2	32.9	33.9	35.4	39.3	32.5	31.5	44.6	48.2	36.3	32.9	34.4	23.8	23.9	34.4
PoseFormerV2 [41]( $T=243$ )	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	35.6
STCFormer-L [33]( $T=243$ )	<b>29.3</b>	33.0	30.7	<b>30.6</b>	<b>32.7</b>	<b>38.2</b>	<b>29.7</b>	<b>28.8</b>	42.2	45.0	<b>33.3</b>	<b>29.4</b>	<b>31.5</b>	<b>20.9</b>	<b>22.3</b>	<b>31.8</b>
GLA-GCN [37]( $T=243$ )	32.4	35.3	32.6	34.2	35.0	42.1	32.1	31.9	45.5	49.5	36.1	32.4	35.6	23.5	24.7	34.8
MotionBERT [44]( $T=243$ )( $\dagger$ )	31.2	<u>32.6</u>	30.8	31.6	33.7	<u>38.6</u>	<u>30.4</u>	29.7	42.6	46.1	34.9	30.7	32.8	22.4	23.7	32.8
MotionAGFormer-L [18]( $T=243$ )( $\dagger$ )	30.7	33.0	<u>30.1</u>	31.9	33.6	38.9	<u>33.1</u>	30.4	<u>41.1</u>	45.9	34.0	29.9	32.3	22.3	23.1	32.7
PoseAnchor ( $T=243$ )	<u>30.6</u>	<b>32.1</b>	<b>29.9</b>	<u>31.5</u>	33.5	<b>38.2</b>	30.6	<u>29.2</u>	<b>41.3</b>	<b>44.0</b>	34.2	<u>30.0</u>	<u>32.1</u>	<u>21.8</u>	<u>22.8</u>	<u>32.1</u>
Abs-MPJPE	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Vpose [23]	128.9	125.4	124.4	138.2	<u>108.2</u>	155.5	116.6	101.1	135.8	287.6	128.6	130.9	122.1	101.6	110.7	134.4
PoseFormer [43]	112.6	137.1	<u>117.6</u>	145.8	113.0	166.0	125.5	113.8	<u>128.8</u>	245.7	122.7	144.8	125.0	118.9	129.3	136.5
RIE [27]	143.2	133.2	143.9	142.7	110.9	151.4	125.9	98.4	136.4	273.4	127.5	138.9	126.8	107.3	116.0	138.4
Ray3D [39]	<b>92.9</b>	<u>97.4</u>	139.8	<b>118.6</b>	113.8	<b>105.9</b>	<b>84.5</b>	<b>74.9</b>	148.6	<b>165.7</b>	<u>116.6</u>	<u>113.9</u>	<u>98.2</u>	<u>83.6</u>	<u>87.9</u>	<u>109.5</u>
PoseAnchor	<u>94.7</u>	<b>84.5</b>	<u>102.7</u>	<u>120.5</u>	<u>93.7</u>	<u>120.7</u>	<u>93.3</u>	<u>78.2</u>	<b>102.5</b>	<u>201.7</u>	<b>92.2</b>	<b>109.7</b>	<b>83.8</b>	<b>81.3</b>	<u>89.9</u>	<b>103.3</b>
MRPE	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Vpose [23]	124.2	115.9	111.0	127.3	<u>97.6</u>	141.9	105.7	96.4	122.0	276.5	119.6	123.3	111.3	94.0	101.6	124.6
PoseFormer [43]	104.7	134.7	<u>103.9</u>	137.4	99.6	154.6	119.8	108.9	<u>108.2</u>	233.7	111.1	141.1	116.2	117.9	123.8	127.7
RIE [27]	139.1	124.5	129.9	133.1	99.2	141.4	116.3	93.5	124.0	265.9	118.4	131.3	117.1	100.4	109.2	129.6
Ray3D [39]	<b>83.7</b>	<u>86.8</u>	128.9	<b>104.8</b>	109.3	<b>91.6</b>	<b>75.0</b>	<b>65.2</b>	143.9	<b>150.5</b>	<u>108.6</u>	<b>105.7</b>	<u>88.4</u>	<b>73.9</b>	<b>77.8</b>	<u>99.6</u>
PoseAnchor	<u>91.5</u>	<b>79.7</b>	<b>94.4</b>	<u>113.6</u>	<b>84.4</b>	<u>108.5</u>	<u>90.4</u>	<u>73.2</u>	<b>90.4</b>	<u>199.1</u>	<b>81.9</b>	<u>106.4</u>	<b>74.2</b>	<u>76.2</u>	<u>85.8</u>	<b>96.7</b>

Table 2. Quantitative comparisons on MPI-INF-3DHP.  $T$ : Number of input frames. The best and second-best scores are in bold and underlined, respectively.

Method	$T$	PCK $\uparrow$	AUC $\uparrow$	MPJPE $\downarrow$
MHFormer [13]	9	93.8	63.3	58.0
MixSTE [40]	27	94.4	66.5	54.9
P-STMO [28]	81	97.9	75.8	32.2
STCFormer [33]	81	<u>98.7</u>	83.9	23.1
PoseFormerV2 [41]	81	97.9	78.8	27.8
GLA-GCN [37]	81	98.5	79.1	27.7
MotionAGFormer [18]	81	98.2	<u>85.3</u>	<b>16.2</b>
PoseAnchor	81	<b>99.3</b>	<b>88.1</b>	<u>17.2</u>

estimation through direct root regression, often suffers from degraded root-relative pose accuracy and limited absolute accuracy. ZS-anchor, leveraging zero-shot ITRR for root localization, offers a compelling balance: it recovers absolute pose effectively without requiring retraining and without compromising the root-relative pose accuracy of the baseline model. This highlights ZS-anchor’s practical utility for readily enhancing existing root-centered 3D HPE models for real-world applications requiring absolute pose information. ZS-anchor significantly reduces absolute errors compared

to Root+, demonstrating the effectiveness of ITRR-based zero-shot root localization. PoseAnchor consistently outperforms all competing approaches, achieving the lowest absolute errors across all tested models. PoseAnchor distinguishes itself as a truly model-agnostic and high-performing solution: it not only significantly enhances absolute pose estimation accuracy, crucial for real-world deployment, but also preserves or even improves the root-relative joint localization precision of diverse 3D HPE architectures. This underscores PoseAnchor’s robustness, generalizability, and superior overall pose estimation quality, making it a versatile and effective approach for advancing 3D human pose estimation in practical scenarios. By integrating root position estimation within a data-driven training framework, PoseAnchor effectively bridges the gap between benchmark-centric research and real-world applicability.

#### 4.4.2. Robustness of Our Approach Against Gaussian Noise Injection in 2D Poses

Table 4 demonstrates that while both methods exhibit comparable performance under noiseless conditions ( $\sigma = 0$ ), the baseline method (Root+) degrades markedly as the noise level increases in the 2D pose inputs. Conversely, PoseAn-

Table 3. Comparison of Root Position Estimation Methods. Baseline, Root+, ZS-anchor, and PoseAnchor.

Model	Approach	MPJPE↓	P-MPJPE↓	Abs-MPJPE↓	MRPE↓
Martinez et al. [16]	Baseline	62.9	47.7	N/A	N/A
	Root+	63.4	47.9	150.2	146.7
	ZS-anchor	62.9	47.7	147.7	140.5
	PoseAnchor	<b>61.5</b>	<b>46.9</b>	<b>143.3</b>	<b>133.7</b>
VPose [23]	Baseline	46.8	36.8	N/A	N/A
	Root+	47.0	36.9	133.1	124.9
	ZS-anchor	46.8	36.8	128.7	121.4
	PoseAnchor	<b>46.2</b>	<b>36.4</b>	<b>120.4</b>	<b>112.0</b>
GLA-GCN [37]	Baseline	44.4	34.8	N/A	N/A
	Root+	44.9	34.9	128.5	119.2
	ZS-anchor	44.4	34.8	125.8	116.6
	PoseAnchor	<b>43.5</b>	<b>33.3</b>	<b>113.8</b>	<b>106.4</b>
MixSTE [40]	Baseline	40.9	32.7	N/A	N/A
	Root+	41.4	32.9	125.2	116.7
	ZS-anchor	40.9	32.7	107.7	100.5
	PoseAnchor	<b>40.3</b>	<b>32.1</b>	<b>103.3</b>	<b>96.7</b>

chor sustains considerably more stable performance across elevated noise levels ( $\sigma = 1, 2, 5,$  and  $10$ ) applied to the 2D pose inputs. The evaluation metric used is the Abs-MPJPE. This substantiates that our approach is robust; its performance remains resilient even when the 2D pose inputs are perturbed with increasing Gaussian noise. Such robustness is indispensable for real-world applications where input data, particularly 2D joint detections, inherently contains noise.

Table 4. Performance comparison under noise (Abs-MPJPE ↓).

Method	GT + $\mathcal{N}(0, \sigma)$				
	$\sigma = 0$	1	2	5	10
Root+	85.6	92.7	117.4	259.5	524.7
PoseAnchor	82.2	82.5	82.5	156.7	366.2

#### 4.4.3. Support Set Effectiveness Analysis

We evaluated our support set’s ability to filter out unreliable joints for ZS-anchor using the H36M dataset and 2D detection error analysis. As shown in Table 5, joints filtered out by the support set had a much higher average error (79.9 pixels) than the overall average (13.6 pixels), while inlier joints had the lowest error (13.1 pixels). This demonstrates that our support set, constructed solely from spatial and projection consistency (see Section 3.1), effectively identifies and removes noisy 2D keypoints without access to ground-truth 2D labels. These findings are consistent with our robustness improvements in noise injection experiments (Table 4) and are visually supported by the qualitative examples in Figure 2.

Table 5. Average 2D Detection Error Comparison on ZS-anchor

Joint Group	Average 2D Error (Pixels)
Overall	13.6
Filtered Out	79.9
Support Set	13.1

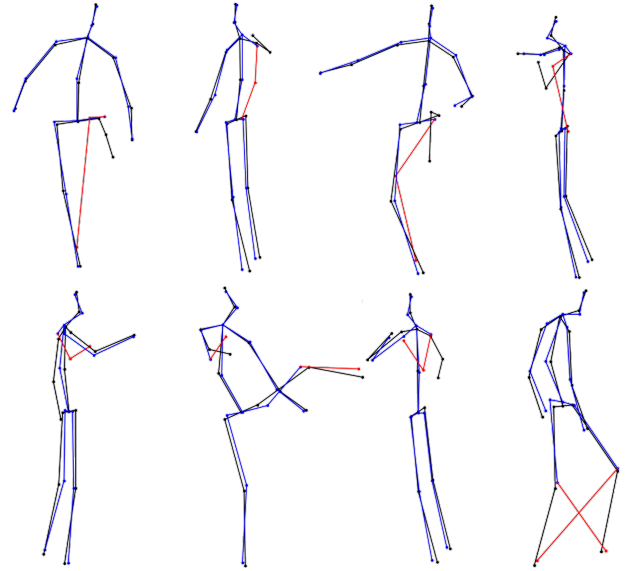


Figure 2. Qualitative Support Set Visualization. 2D keypoints are color-coded: **Black** - Ground Truth (GT), **Red** - Filtered Outliers, **Blue** - Support Set Inliers.

#### 4.4.4. Comparison with RLSR Methods

Table 6 presents the quantitative comparison of robust least squares regression (RLSR) methods for root position estimation on the H36M dataset. Compared to other RLSR methods, including TORRENT [1], our approach (ZS-anchor) demonstrates lower Abs-MPJPE and MRPE, indicating improved robustness and accuracy in root localization.

Table 6. Comparison of different RLSR methods on root position estimation using [40] in a Zero-Shot Setting.

Method	Abs-MPJPE↓ (mm)	MRPE↓ (mm)
M-estimation [9]	110.4	103.1
LMS [25]	111.0	103.7
RANSAC [7]	123.8	117.1
TORRENT [1]	109.4	102.2
ZS-anchor	<b>107.7</b>	<b>100.5</b>

## 5. Conclusion

PoseAnchor, a unified framework, bridges the gap between 3D human pose estimation benchmarks and real-world application by integrating robust root position estimation without compromising pose accuracy. ZS-anchor enables zero-shot root localization for existing models, while Data-Driven Training with support set filtering enhances robustness and accuracy. Experiments demonstrate state-of-the-art performance, surpassing root-centered and Root+ methods. Ablation studies confirm model-agnostic applicability, robustness against noise, and superiority over Root+. PoseAnchor consistently improves or maintains pose accuracy while enhancing root estimation.

## Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University), No. RS-2024-00457882, AI Research Hub Project, and Artificial Intelligence Star Fellowship Support Program to Nurture the Best Talents (IITP-2025-RS-2025-02304828)).

## References

- [1] Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. *Adv. Neural Inform. Process. Syst.*, 28, 2015. 2, 4, 8
- [2] Michael J. Black and Anand Rangarajan. On the unification of line processes, outlier rejection, and robust statistics. *International Journal of Computer Vision (IJCV)*, 19(1):57–91, 1996. 2
- [3] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7035–7043, 2017. 2
- [4] Tianlang Chen, Chen Fang, Xiaohui Shen, Yiheng Zhu, Zhili Chen, and Jiebo Luo. Anatomy-aware 3d human pose estimation with bone-based pose decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):198–209, 2021. 7
- [5] Yudong Chen, Constantine Caramanis, and Shie Mannor. Robust sparse regression under adversarial corruption. 28:774–782, 2013. 2
- [6] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7103–7112, 2018. 2, 6
- [7] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2, 8
- [8] Jumin Han, Jun-Hee Kim, and Seong-Whan Lee. Propose: Probabilistic 3d human pose estimation with instance-level distribution and normalizing flow. In *AAAI*, pages 3338–3346, 2025. 1
- [9] Peter J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1):73–101, 1964. 2, 8
- [10] Peter J Huber and Elvezio M Ronchetti. *Robust Statistics*. John Wiley & Sons, 2009. 2
- [11] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, 2013. 6
- [12] Jun-Hee Kim and Seong-Whan Lee. Toward approaches to scalability in 3d human pose estimation. In *Adv. Neural Inform. Process. Syst.* 1
- [13] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13147–13156, 2022. 7
- [14] Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen-ching Cheung, and Vijayan Asari. Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5064–5073, 2020. 7
- [15] Vera M Lourenço, Ana M Pires, and Matias Kirst. Robust linear regression methods in association studies. *Bioinformatics*, 27(6):815–821, 2011. 2
- [16] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2640–2649, 2017. 2, 6, 8
- [17] Brian McWilliams, Gabriel Krummenacher, Mario Lucic, and Joachim M Buhmann. Fast and robust least squares estimation in corrupted linear models. In *Advances in Neural Information Processing Systems*, pages 415–423, 2014. 2
- [18] Soroush Mehraban, Vida Adeli, and Babak Taati. Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 6920–6930, 2024. 2, 6, 7
- [19] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *Proc. Int. Conf. 3D Vis.*, pages 506–516, 2017. 6
- [20] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *Int. Conf. Comput. Vis.*, pages 10133–10142, 2019. 2
- [21] Imran Naseem, Roberto Togneri, and Mohammed Benamoun. Robust regression for face recognition. *Pattern Recognition*, 45(1):104–118, 2012. 2
- [22] Nam H Nguyen and Trac D Tran. Exact recoverability from dense corrupted observations via l1-minimization. *IEEE Transactions on Information Theory*, 59(4):2017–2035, 2013. 2
- [23] Dario Pavlo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7753–7762, 2019. 1, 2, 6, 7, 8
- [24] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3433–3441, 2017. 2
- [25] Peter J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984. 2, 8
- [26] Peter J Rousseeuw and Annick M Leroy. *Robust Regression and Outlier Detection*. John Wiley & Sons, 2005. 2
- [27] Wenkang Shan, Haopeng Lu, Shanshe Wang, Xinfeng Zhang, and Wen Gao. Improving robustness and accuracy via relative information encoding in 3d human pose estimation. In *ACM Int. Conf. Multimedia*, pages 3446–3454, 2021. 7

- [28] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In *Eur. Conf. Comput. Vis.*, pages 461–478. Springer, 2022. [7](#)
- [29] Yiyuan She and Art B Owen. Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, 106(494):626–639, 2011. [2](#)
- [30] Aljoscha Smolic and Jens-Rainer Ohm. Robust global motion estimation using a simplified m-estimator approach. In *Proceedings of the IEEE International Conference on Image Processing*, pages 868–871. IEEE, 2000. [2](#)
- [31] Christoph Studer, Patrick Kuppinger, Graeme Pope, and Helmut Bölcskei. Recovery of sparsely corrupted signals. *IEEE Transactions on Information Theory*, 58(5):3115–3130, 2012. [2](#)
- [32] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.* [2](#)
- [33] Zhenhua Tang, Zhaofan Qiu, Yanbin Hao, Richang Hong, and Ting Yao. 3d human pose estimation with spatio-temporal criss-cross attention. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4790–4799, 2023. [1](#), [2](#), [7](#)
- [34] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2500–2509, 2017. [2](#)
- [35] John Wright and Yi Ma. Dense error correction via  $\ell_1$ -minimization. *IEEE Transactions on Information Theory*, 56(7):3540–3560, 2010. [2](#)
- [36] John Wright, Allen Y Yang, A Ganesh, SS Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009. [2](#)
- [37] Bruce XB Yu, Zhi Zhang, Yongxu Liu, Sheng-hua Zhong, Yan Liu, and Chang Wen Chen. Gla-gcn: Global-local adaptive graph convolutional network for 3d human pose estimation from monocular video. In *Int. Conf. Comput. Vis.*, pages 8818–8829, 2023. [6](#), [7](#), [8](#)
- [38] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution vision transformer for dense predict. *Adv. Neural Inform. Process. Syst.*, 34:7281–7293, 2021. [2](#)
- [39] Yu Zhan, Fenghai Li, Renliang Weng, and Wongun Choi. Ray3d: ray-based 3d human pose estimation for monocular absolute 3d localization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13116–13125, 2022. [1](#), [2](#), [7](#)
- [40] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Jun-song Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13232–13242, 2022. [1](#), [2](#), [6](#), [7](#), [8](#)
- [41] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8877–8886, 2023. [7](#)
- [42] Jianan Zhen, Qi Fang, Jiaming Sun, Wentao Liu, Wei Jiang, Hujun Bao, and Xiaowei Zhou. Smap: Single-shot multi-person absolute 3d pose estimation. In *Eur. Conf. Comput. Vis.*, pages 550–566. Springer, 2020. [2](#)
- [43] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11656–11665, 2021. [7](#)
- [44] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Int. Conf. Comput. Vis.*, pages 15085–15099, 2023. [2](#), [7](#)
- [45] Abdelhak M Zoubir, Visa Koivunen, Yacine Chakhchoukh, and Michael Muma. Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts. *IEEE Signal Processing Magazine*, 29(4):61–80, 2012. [2](#)