

# Enhancing 3D Human Pose Estimation with Bone Length Adjustment

Chih-Hsiang Hsu<sup>1</sup>  and Jyh-Shing Roger Jang<sup>1</sup> 

Graduate Institute of Networking and Multimedia, National Taiwan University,  
Taiwan  
[r11944034@ntu.edu.tw](mailto:r11944034@ntu.edu.tw)

**Abstract.** Current approaches in 3D human pose estimation primarily focus on regressing 3D joint locations, often neglecting critical physical constraints such as bone length consistency and body symmetry. This work introduces a recurrent neural network architecture designed to capture holistic information across entire video sequences, enabling accurate prediction of bone lengths. To enhance training effectiveness, we propose a novel augmentation strategy using synthetic bone lengths that adhere to physical constraints. Moreover, we present a bone length adjustment method that preserves bone orientations while substituting bone lengths with predicted values. Our results demonstrate that existing 3D human pose estimation models can be significantly enhanced through this adjustment process. Furthermore, we fine-tune human pose estimation models using inferred bone lengths, observing notable improvements. Our bone length prediction model surpasses the previous best results, and our adjustment and fine-tuning method enhance performance across several metrics on the Human3.6M dataset.

## 1 Introduction

3D human pose estimation aims to localize the 3D positions of human joints from monocular images or videos, holding significant implications for applications such as human-computer interaction, sports analysis, and medical diagnostics, due to its capacity to capture human motion. Presently, two-stage approaches dominate in 3D human pose estimation. These methods initially detect 2D keypoints from input images or videos and subsequently lift these 2D keypoints into 3D space, which is known as the 2D-to-3D lifting task.

The 2D-to-3D lifting task faces inherent challenges due to depth ambiguity: multiple 3D poses can project to the same 2D keypoints. A simple example is the flipping ambiguity described by [19], where finite possible 3D poses arise by flipping each bone forwards or backwards. This ambiguity intensifies when bone lengths are unknown or when 2D keypoints are inaccurate. Recognizing the importance of temporal information in resolving depth ambiguity, recent studies have employed recurrent neural networks (RNNs) [8, 12], temporal convolutions [2, 16], and transformer-based models [13, 27, 30] to extract temporal features.

Despite recent advancements, many existing methods overlook the natural structure of human poses. Studies have shown that focusing solely on minimizing per-joint errors independently neglects overall pose coherence. Addressing this issue, bone-based representations have been proposed [21]. Chen *et al.* [2] introduced a method to decompose human poses into bone lengths and directions, simplifying the pose estimation task. However, integrating physical constraints such as bone length consistency and body symmetry remains a challenge, with significant bone length errors observed in existing works.

Inspired by previous work [2], we propose RNN-based models to predict bone lengths. Our models leverage global information from all frames of a video, rather than short sequences. To enhance training effectiveness, we introduce a novel training time augmentation method using synthetic bone lengths generated by the statistical body shape model called SMPL [15]. This augmentation ensures that predicted bone lengths respect symmetry constraints and adhere to natural human body proportions. Unlike the method in [2] that randomly adjust bone lengths, potentially distorting body proportions, our method prioritizes realistic and accurate predictions.

We propose a novel adjustment method to enhance current state-of-the-art 2D-to-3D lifting models. This method preserves bone directions while replacing bone lengths, ensuring that the adjusted poses maintain anatomical correctness and achieve more precise joint positions. Finally, we fine-tune existing 2D-to-3D lifting models using bone length information. This fine-tuning process further improves model performance and can be applied to any 2D-to-3D lifting model, demonstrating its versatility and effectiveness.

The contributions of this work are threefold:

- We propose a bone length prediction model that effectively utilizes global information and synthetic augmentation to predict accurate bone lengths.
- We introduce a bone length adjustment method that enhances existing 2D-to-3D lifting models, ensuring realistic body shapes and accurate joint positions.
- We demonstrate the efficacy of fine-tuning existing models with predicted bone lengths, thereby improving their performance in 3D human pose estimation tasks.

## 2 Related Work

3D human pose estimation using deep learning methods can be categorized into two main approaches: the end-to-end approach and the two-stage approach.

The end-to-end approach predicts 3D human poses directly from RGB images. These methods [10, 11, 24] heavily rely on parametric 3D human shape models, such as SMPL [15]. They estimate parameterized pose, shape, and translation, which are then decoded via the SMPL model to obtain the human pose and shape.

In contrast, the two-stage approach first detects 2D positions of joints in an image, known as 2D keypoints. Techniques such as convolutional neural networks

(CNNs) generate heatmaps that indicate the probability of joint locations [3,20]. In the second stage, these 2D keypoints are used to estimate the 3D pose of the human figure, which is called 2D-to-3D lifting task. Currently, the two-stage approach tends to be more accurate than the end-to-end approach. End-to-end methods struggle with a lack of diversity in video data, especially variations in background, due to the complex requirements of 3D pose datasets, such as motion capture systems and high-speed cameras. This issue is mitigated in the two-stage approach since 2D keypoints can be manually labeled, and modern 2D keypoint detectors achieve high precision.

The 2D-to-3D lifting task is an ill-posed problem that involves predicting the additional depth dimension. Recent works have utilized temporal information by predicting the human pose of the central frame from a sequence of 2D keypoints in a video. The context information provides clues about the movement of the target, making the predicted poses more robust to noise. Hossain and Little [8] designed an LSTM sequence-to-sequence model to obtain temporally consistent 3D poses. Pavllo *et al.* [16] proposed a model based on dilated temporal convolution to capture long-term information and reduce computational overhead compared to LSTM models. Chen *et al.* [2] demonstrated that dividing the task into bone length and bone direction prediction yields better results.

Since pictorial information is often lost in the 2D-to-3D lifting task, some works have introduced image clues as input, such as the root depths of targets [4,14,26] and occlusion scores of keypoints [2,14]. Zhao *et al.* [28] exploited intermediate visual representations of 2D pose detectors, achieving promising results even when predicting from a single image.

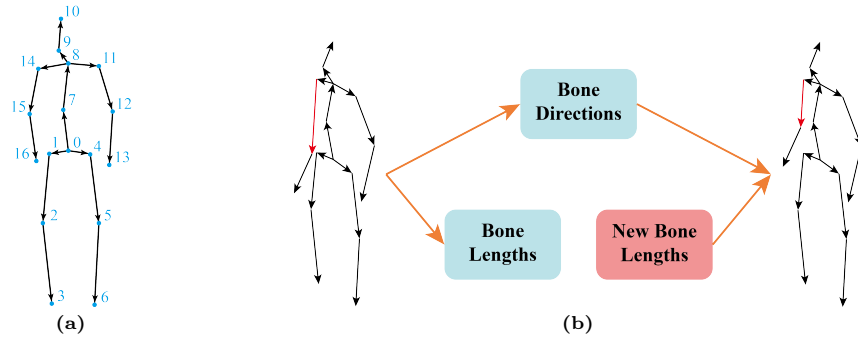
However, the consistency and accuracy of bone lengths have not been well considered in these approaches. Our research demonstrate that these models can be further improved by adjusting bone lengths, ensuring more accurate and realistic predictions.

### 3 Methods

Our work consists of two main components: bone length prediction and bone length adjustment. Section 3.1 details our bone length augmentation method, Section 3.2 discusses our model design for predicting bone lengths, and Section 3.3 explains how we apply our bone length prediction model to enhance 2D-to-3D lifting models.

#### 3.1 Bone Length Augmentation

**Augmentation.** We represent a human pose  $P = [\mathbf{p}_0 \cdots \mathbf{p}_{J-1}]^T \in \mathbb{R}^{J \times 3}$  with  $J$  3D joint positions as a tree structure, as shown in Figure 1a. The root joint is positioned on the pelvis and labeled as joint 0. For each joint  $\mathbf{p}_i \in \mathbb{R}^3$ , its parent is defined as the joint closer to the root (*e.g.*, joint 0 is the parent of joint 1). The pose can be decomposed into bone lengths  $L = [l_1 \cdots l_{J-1}]^T \in \mathbb{R}^{(J-1) \times 1}$  and bone directions  $D = [\mathbf{d}_1 \cdots \mathbf{d}_{J-1}]^T \in \mathbb{R}^{(J-1) \times 3}$  using the following equations:



**Fig. 1:** (a) The representation of a human pose with joint labels. (b) The overview of bone length replacement, which involves decomposing the pose into bone directions and bone lengths, and then substituting the original bone lengths with new ones.

$$\begin{aligned}
 l_i &= \|\mathbf{p}_i - \mathbf{p}_{\text{parent}(i)}\|_2, \quad i = 1, \dots, J-1 \\
 \mathbf{d}_i &= \frac{\mathbf{p}_i - \mathbf{p}_{\text{parent}(i)}}{l_i}, \quad i = 1, \dots, J-1
 \end{aligned} \tag{1}$$

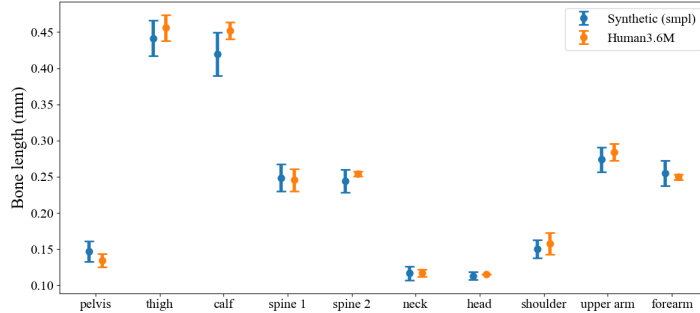
Here, vertices (joints) are labeled from 0 to  $J-1$  and the edges (bones) are labeled from 1 to  $J-1$ . Given  $L$  and  $D$ , the original pose  $P$  can be reconstructed.

In the augmentation process, we first decompose a pose  $P$  into bone lengths  $L$  and bone directions  $D$ . We then use new bone lengths  $L' = [l'_1 \cdots l'_{J-1}]^T$  and the original bone directions  $D$  to reconstruct a new pose  $\tilde{P} = [\tilde{\mathbf{p}}_1 \cdots \tilde{\mathbf{p}}_{J-1}]^T$ . The bone length replacement process is illustrated in Figure 1b. A random shift  $\mathbf{s} \in \mathbb{R}^3$  is added to the poses to enhance the augmentation. The final result is the augmented pose  $P' = [\mathbf{p}'_1 \cdots \mathbf{p}'_{J-1}]^T$ .

$$\begin{aligned}
 \mathbf{s} &\sim \mathcal{N}(\mu = 0, \sigma = 0.5) \\
 \mathbf{p}'_i &= \tilde{\mathbf{p}}_i + \mathbf{s}, \quad i = 1, \dots, J-1
 \end{aligned} \tag{2}$$

For a sequence of poses, the same random shift should be applied to preserve smoothness in the trajectory. Since we use 2D keypoints as the model input, we project the pose  $P'$  onto the 2D camera plane considering the camera intrinsic matrix (focal length and principal point), and both radial and tangential nonlinear lens distortion. Our model learns to predict  $L'$  from the projected 2D keypoints. The key to the augmentation process is generating reasonable bone lengths  $L'$ .

**Random Bone Lengths.** Before introducing our augmentation method, we briefly discuss the approach used in [2]. They randomly adjust bone lengths



**Fig. 2:** This error bar plot shows the means and the standard deviations of bone lengths in the Human3.6M dataset. Each mean value is represented by a dot, and the associated standard deviation is shown by the bars, indicating the variability around the mean bone lengths.

$L$  based on the average bone lengths in the training batch, denoted by  $\bar{L} = [\bar{l}'_1 \dots \bar{l}'_{J-1}]^T \in \mathbb{R}^{(J-1) \times 1}$ .

$$l'_i = l_i + r_i \bar{l}_i, \quad r_i \sim \mathcal{U}(-0.3, 0.3), \quad i = 1, \dots, J - 1 \quad (3)$$

The proportion varies between  $-30\%$  to  $30\%$ , which can lead to  $L'$  deviating from natural human anatomical structures. For example, this method could generate an unnaturally long forearm combined with a short upper arm. Additionally,  $L'$  might lack symmetry because each bone is adjusted independently by different random values. In our experiments, we ensure symmetry by applying identical random adjustments to corresponding bones on both sides of the body.

As shown in Figure 2, the variability of each bone length is different. For instance, subjects in the Human3.6M dataset [19] have similar lengths of forearms but differ in lengths of upper arms. Intuitively, we may randomly adjust the bone lengths by applying a normal distribution:

$$\begin{aligned}
 l'_i &\sim \mathcal{N}(l_i, \sigma_i), \quad i = 1, \dots, J - 1 \\
 l'_i &= l'_j, \quad \text{if they are the same body part on different side.}
 \end{aligned} \quad (4)$$

where the mean value is the  $i$ -th original length and  $\sigma_i$  denotes the standard deviation of the  $i$ -th bone length in the Human3.6M dataset. We also maintain the symmetry in this case.

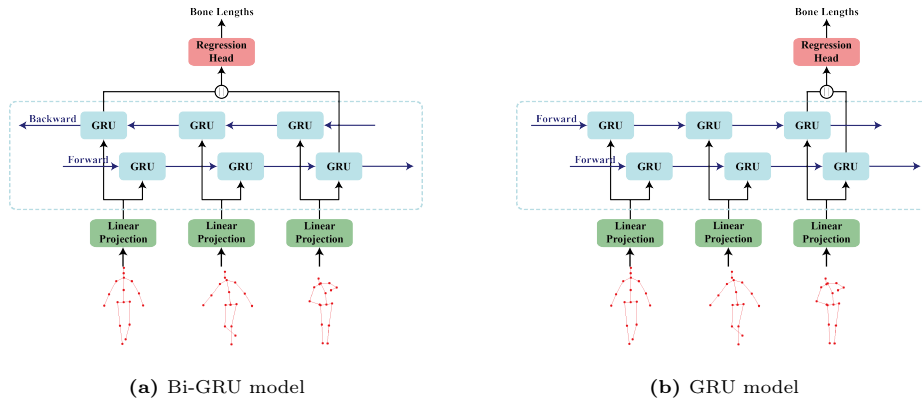
**Synthetic Bone Lengths.** SMPL [15] is a model that generates 3D human meshes from parameters. We use SMPL to randomly generate human meshes and then evaluate the bone lengths from these meshes. This ensures that the bone lengths are symmetric and reasonable, reflecting natural body shapes. To

evaluate bone lengths from meshes, we apply the joint regression matrix  $\mathcal{J}$  introduced in [5] to mesh coordinates  $M$ :

$$\tilde{L} = \mathcal{J}M \quad (5)$$

The 3D poses in the Human3.6M dataset are recorded using a marker-based motion capture system, where the position of each joint depends on the placement of the markers. Consequently, a single joint regression matrix cannot accurately describe the positions of the joints. When using a single joint regression matrix, the distribution of the regressed bone lengths differs from that of the Human3.6M dataset, as shown in Figure 2, leading to poor predictive ability. To mitigate the difference in data distribution, we align the mean value of the regressed bone lengths with the mean value in the Human3.6M dataset. After the alignment, we obtain the augmented bone lengths  $L'$ .

### 3.2 Bone Length Model



**Fig. 3:** The structures of our bone length prediction models. The input length is 3 for illustration.

The structures of our models are illustrated in Figure 3. Our primary model, depicted in Figure 3a, utilizes a single-layer bidirectional gated recurrent unit (Bi-GRU) [6]. This Bi-GRU model processes the entire sequence of 2D keypoints from a given video, leveraging both past and future information for improved prediction accuracy. However, due to its reliance on future data, the Bi-GRU model is not suitable for real-time online processing. To address this limitation, we also developed a GRU model, shown in Figure 3b, which updates bone lengths by processing the input keypoints frame by frame, making it suitable for online applications. In this section, we specifically introduce the Bi-GRU model.

During training, we slice the sequences of 2D keypoints into fixed-size segments for convenience. The input sequence of 2D keypoints is denoted by  $X =$

$[\mathbf{x}_0 \cdots \mathbf{x}_N] \in \mathbb{R}^{N \times (J \times 2)}$ , where  $N$  is the sequence length,  $J$  is the number of joints, and  $\mathbf{x}_t \in \mathbb{R}^{2J}$  is the flattened vector of 2D keypoints at frame  $t$ . A linear projection layer maps each  $\mathbf{x}_t$  to a higher dimension  $c$ .

$$\mathbf{x}'_t = W_p \mathbf{x}_t + \mathbf{b}_p. \quad (6)$$

where  $W_p$  is the weight matrix and  $\mathbf{b}_p$  is the bias vector in the linear projection layer.

The projected vectors  $\mathbf{x}'_t \in \mathbb{R}^c$  are then input to the GRU. The forward process at frame  $t$  can be written as

$$\mathbf{h}_t = \text{GRU}(X'_t, \mathbf{h}_{t-1}) \quad (7)$$

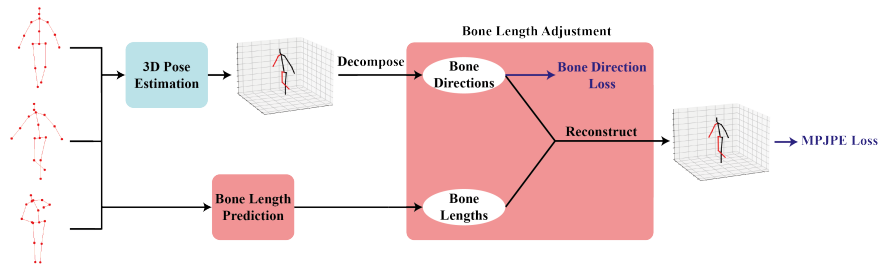
where  $\mathbf{h}_t \in \mathbb{R}^{c'}$  is the hidden state at frame  $t$  with hidden size  $c'$ , and the initial hidden state  $\mathbf{h}_0$  is a zero vector. The backward process is similar but processes  $X'_t$  in reverse order. We concatenate the final hidden states from the forward process and backward processes to obtain  $\mathbf{h} \in \mathbb{R}^{2c'}$ . The bone lengths  $L \in \mathbb{R}^{(J-1) \times 1}$  are then regressed from  $\mathbf{h}$  using the weight matrix  $W_R$ .

$$L = W_R \mathbf{h} \quad (8)$$

Our goal is to minimize the difference between the predicted bone lengths  $L$  and the groundtruth bone lengths  $\hat{L}$ . The loss function is defined by the mean absolute error:

$$\mathcal{L}_L = \frac{1}{J-1} \sum_{i=1}^{J-1} \|l_i - \hat{l}_i\|_1 \quad (9)$$

### 3.3 Bone Length Adjustment and Fine-tuning



**Fig. 4:** The overview of bone length adjustment. The 3D pose estimation is based on existing 2D-to-3D lifting models. The blue part is based on existing lifting models. Only the parameters in blue part are fine-tuned.

Figure 4 provides an overview of our bone length adjustment method. This technique is applied to the human poses predicted by existing 2D-to-3D lifting

models. The bone length adjustment involves replacing the bone lengths of the human poses with our predicted bone lengths, as illustrated in Figure 1b.

Given a sequence of 2D keypoints  $X$  and a lifting model, we first obtain predicted poses  $P$  from the lifting model. The sequence  $X$  is segmented to fit in the input requirements of the lifting model. We then decompose the poses  $P$  into bone lengths  $L$  and bone directions  $D$ . Concurrently, we use the entire sequence  $X$  to predict new bone lengths  $L'$  with our model. By combining the bone directions  $D$  from the lifting model and the bone lengths  $L'$  from our model, we generate the reconstructed poses  $P'$ . This process refines the poses, ensuring a more realistic body structure.

To evaluate the adjustment process, we use the Mean Per Joint Position Error (MPJPE) to measure the error between the reconstructed pose  $P'$  and the groundtruth pose  $\hat{P}$ :

$$\mathcal{L}_P = \frac{1}{J} \sum_{i=0}^{J-1} \|\mathbf{p}'_i - \hat{\mathbf{p}}_i\|_2 \quad (10)$$

**Fine-tuning.** In our adjustment process, bone lengths can also enhance the lifting models' ability to predict bone directions. We propose a fine-tuning method based on our adjustment process. The bone length prediction model is fixed during this process to prevent overfitting since the data without augmentation is used. We fix the weights of our bone length model and fine-tune the lifting models by minimizing the error in the predicted bone directions and the MPJPE of the reconstructed pose  $P'$ . The direction loss is defined as:

$$\mathcal{L}_D = \frac{1}{J-1} \sum_{i=1}^{J-1} \|\mathbf{d}_i - \hat{\mathbf{d}}_i\|_2 \quad (11)$$

where  $\mathbf{d}_i$  is the predicted direction and  $\hat{\mathbf{d}}_i$  is the groundtruth direction of the  $i$ -th bone. The total loss combines both the direction loss and the position error loss:

$$\mathcal{L} = \mathcal{L}_D + \mathcal{L}_P \quad (12)$$

## 4 Experimental setup

### 4.1 Dataset and Evaluation

We evaluate our method using the Human3.6M dataset [1, 9], which contains 3.6 million frames of 11 actors performing 15 diverse actions, recorded by four synchronized cameras at 50 Hz. Seven subjects are annotated with 3D poses. Following the standard protocol in prior works [2, 13, 16, 27, 30], we train our model on five subjects (S1, S5, S6, S7, S8) and test on two subjects (S9, S11), using a 17-joint skeleton. For bone length evaluation, we use the bone length error as described by Equation 9, comparing the predicted to the groundtruth



lengths. For human poses, we use two protocols: Protocol 1 measures the Mean Per Joint Position Error (MPJPE), the average Euclidean distance between the predicted and groundtruth joint positions, and Protocol 2 (P-MPJPE) reports the error after aligning the predicted poses with the groundtruth in terms of translation, rotation, and scale, enabling us to better evaluate the accuracy of the predicted body structure.

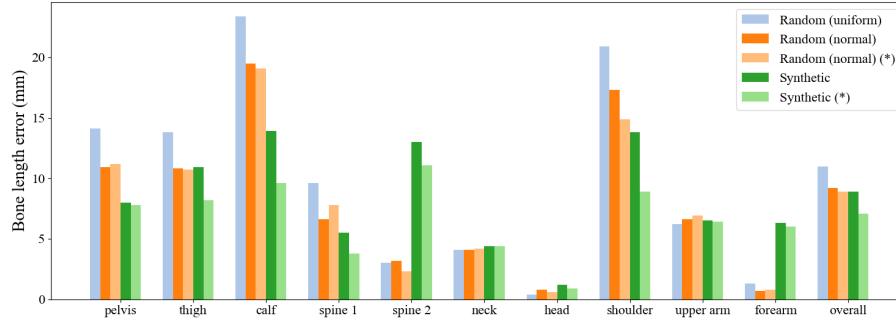
## 4.2 Implementation Details

**Bone Length Prediction.** We align the mean value of our synthetic bone lengths with the Human3.6M dataset. Our study evaluates five different methods for generating bone lengths, detailed in Section 3.1:

- **Random Uniform Distribution:** Randomly generated bone lengths from a uniform distribution.
- **Random Normal Distribution:** Bone lengths generated from a normal distribution using standard deviations in the training set.
- **Random Normal Distribution (Involving Test Data STD):** Bone lengths generated from a normal distribution using standard deviations in the entire dataset, involving test data statistics.
- **Synthetic:** Synthetic bone lengths aligned with the mean values in the training set.
- **Synthetic (Involving Test Data Mean Values):** Synthetic bone lengths aligned with the mean values in the entire dataset, involving test data statistics.

During training, we use a sequence length  $N = 512$  and utilize the entire sequence during testing. The projected dimension  $c$  is set to 256, and the hidden state dimension  $c'$  is set to 512. We train our models using the Adam optimizer with an exponentially decaying learning rate schedule. The initial learning rate is set to 0.0001, and it decays by a factor of 0.95 each epoch. The batch size is set to 256. We train our model with groundtruth 2D keypoints which are projected from synthetic poses and test our model with 2D keypoints predicted by the Cascaded Pyramid Network (CPN) [3].

**Fine-tuning.** For fine-tuning, we select Pavllo *et al.* [16] as our fine-tuning target. We configure the sequence length  $N$  to 243, and apply horizontal flip augmentation during both training and testing, following their settings. We fine-tune the model using the Adam optimizer and a batch-normalization momentum set to the final state 0.001. Similarly, we employ an exponentially decaying learning rate schedule, starting at 0.00004 with a decay factor of 0.95 per epoch. The batch size for fine-tuning is set to 1024, consistent with their work. We utilize 2D keypoints predicted by CPN for both training and testing phases. Finally, the horizontal flip augmentation is applied at train and test time, following previous works [2, 16, 30].



**Fig. 5:** The average bone length error comparison across all frames of the test set in Human3.6M. (\*) including test data statistics.

## 5 Results

In this section, we compare our results to recent works. Since our method involves adjusting model predictions, we reproduced the results by following the instructions provided by each work. This may result in slight variations due to differences in environments.

For diffusion models, they generate multiple hypotheses and select the best one to evaluate the error. As outlined in [18], there are four methods to select the best hypothesis. The reported errors in these models are based on selecting the joint-wise hypothesis closest to the ground truth, which is impractical in real-world applications. Instead, we applied a joint-level aggregation method to generate the final prediction without relying on the ground truth. Therefore, the errors reported here will differ from those claimed in the original works.

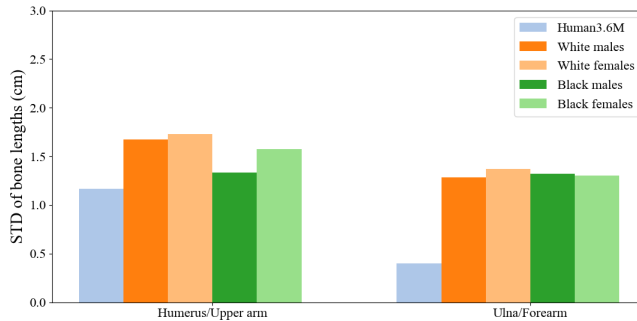
### 5.1 Bone Length Model

Figure 5 presents the outcomes of our Bi-GRU bone length model evaluation. Utilizing synthetic bone lengths during training time augmentation yields the lowest overall bone length error among all methods evaluated. The random uniform distribution method fail to generate bone lengths that adhere to natural human body proportions, resulting in poor performance. Conversely, synthetic methods demonstrate superior performance over random methods, even when not using test data statistics.

Table 1 shows the comparison of bone lengths. For the lifting models, the error is evaluated by decomposing the predicted poses into bone lengths. Our model achieves the state-of-the-art result when not using test data statistics. Additionally, the GRU model designed for online processing performs comparably to the Bi-GRU model. The synthetic method incorporating test data statistics notably outperforms all other results. Given that the training set comprises data from only five subjects, the statistics may not fully represent broader variations, leading to significant disparities between using and not using test data statistics.

**Table 1:** Quantitative comparison of bone length error. Best in bold and second best underlined. (\*) including test data statistics. (†) diffusion model. (-) bone length model only.

		Bone length error ↓ (mm)
Pavlo <i>et al.</i> [16]	CVPR'19	12.3
Chen <i>et al.</i> [2] (-)	TCSVT'21	10.3
Chen <i>et al.</i> [2]	TCSVT'21	8.9
Zheng <i>et al.</i> [30]	ICCV'21	10.8
Gong <i>et al.</i> [7] (†)	CVPR'23	<u>8.5</u>
Shan <i>et al.</i> [18] (†)	ICCV'23	10.6
Peng <i>et al.</i> [17] (†)	CVPR'24	10.9
Ours, Bi-GRU (synthetic) (-)		8.9
Ours, Bi-GRU (synthetic) (*) (-)		<b>7.1</b>
Ours, GRU (synthetic) (*) (-)		<b>7.1</b>



**Fig. 6:** Comparison between the standard deviation of real bone lengths and bone lengths in Human3.6M.

With a more comprehensive dataset, our approach could potentially circumvent such limitations.

Examining Figure 2, we observe that the standard deviations of certain bones, such as spine 2, neck, head, and forearm in the Human3.6M dataset are exceptionally small. As shown in Figure 6, anthropometric research [23] indicates similar standard deviations for lengths of the humerus (upper arm) and ulna (forearm). However, in Human3.6M, standard deviations for the lengths of the upper arm and forearm differ significantly. Additionally, the standard deviation of the upper arm closely matches the anthropometric result for the humerus, while the standard deviation of the lower arm is much smaller than the anthropometric result for the ulna. These discrepancies may arise from limitations in the transformation process from MoCap raw data to human poses, potentially

influenced by constraints on body shape or inaccuracies in marker placement within the Human3.6M dataset.

Even with dataset constraints, our synthetic method consistently outperforms random methods and the lifting models. We select the model trained with synthetic bone lengths using statistics of the test data as our final model.

## 5.2 Adjustment and Fine-tuning

**Table 2:** Quantitative comparison of the adjustment process on reconstruction error evaluated on Human3.6M under MPJPE and P-MPJPE. Best results of the same base model are in bold.

Base model	Bone length error ↓ (mm)	Bone length model	MPJPE ↓ (mm)	P-MPJPE ↓ (mm)
Pavlo <i>et al.</i> [16] CVPR'19	12.3	✗	46.8	36.5
		GRU	45.6	36.1
		Bi-GRU	<b>45.2</b>	<b>35.8</b>
Chen <i>et al.</i> [2] TCSVT'21	8.9	✗	44.2	35.0
		GRU	44.0	34.8
		Bi-GRU	<b>43.5</b>	<b>34.5</b>
Zheng <i>et al.</i> [30] ICCV'21	10.8	✗	44.3	34.6
		GRU	43.3	34.1
		Bi-GRU	<b>42.9</b>	<b>33.8</b>
Li <i>et al.</i> [13] CVPR'22	10.3	✗	43.0	34.5
		GRU	42.5	34.0
		Bi-GRU	<b>42.2</b>	<b>33.7</b>
Zhang <i>et al.</i> [27] CVPR'22	11.0	✗	40.9	32.7
		GRU	40.6	32.5
		Bi-GRU	<b>40.2</b>	<b>32.2</b>
Gong <i>et al.</i> [7] CVPR'23	8.5	✗	39.5	31.2
		GRU	39.4	31.1
		Bi-GRU	<b>39.0</b>	<b>30.8</b>
Tang <i>et al.</i> [22] CVPR'23	10.7	✗	41.8	32.9
		GRU	41.6	33.1
		Bi-GRU	<b>41.0</b>	<b>32.7</b>
Zhao <i>et al.</i> [29] CVPR'23	11.1	✗	45.2	35.6
		GRU	44.5	35.2
		Bi-GRU	<b>44.0</b>	<b>34.9</b>
Shan <i>et al.</i> [18] ICCV'23	10.6	✗	39.6	31.7
		GRU	38.9	31.2
		Bi-GRU	<b>38.6</b>	<b>31.0</b>
Peng <i>et al.</i> [17] CVPR'24	10.9	✗	40.2	32.2
		GRU	39.9	31.9
		Bi-GRU	<b>39.4</b>	<b>31.5</b>
Xu <i>et al.</i> [25] CVPR'24	12.2	✗	40.2	32.9
		GRU	40.1	32.5
		Bi-GRU	<b>39.6</b>	<b>32.2</b>

Table 2 illustrates the reconstruction error of existing lifting models before and after applying our adjustment method. Across all tested models, we observe consistent performance improvements under both protocol 1 (MPJPE) and protocol 2 (P-MPJPE) after adjustment with both GRU and Bi-GRU models. Our Bi-GRU model outperforms the GRU model, showing the advantage of utilizing future information. The degree of improvement correlates with the bone length error inherent in the original models. Models with larger initial bone length

errors, such as [16], demonstrate significant enhancement, achieving a 3% reduction in Protocol 1 error with the adjustment. In contrast, models like [2], which exhibit smaller bone length errors due to effective bone length prediction, show more modest improvements of around 1% under Protocol 1. Our adjustment effectively rectifies pose errors for all models under Protocol 2, which undergoes rigid alignment like scaling. This indicates that our predicted bone lengths possess better body proportions.

**Table 3:** Reconstruction error on Human3.6M before and after adjustment and fine-tuning with our Bi-GRU model. The top table shows the result under protocol 1. The bottom table shows the result under protocol 2. Best in bold and second best underlined. Unit: millimeter

Protocol 1	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Pavlo <i>et al.</i> [16]	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
Adjustment	41.9	<b>45.2</b>	42.1	44.3	46.1	<b>53.5</b>	43.4	<b>42.1</b>	55.5	64.0	45.5	42.3	47.1	32.1	32.9	45.2
Fine-tuning	<b>41.5</b>	<b>45.2</b>	<b>41.9</b>	<b>44.0</b>	<b>45.9</b>	53.6	<b>43.3</b>	42.3	<b>55.3</b>	<b>63.8</b>	<b>45.3</b>	<b>42.2</b>	<b>47.0</b>	<b>31.8</b>	<b>32.4</b>	<b>45.0</b>

Protocol 2	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Pavlo <i>et al.</i> [16]	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
Adjustment	32.8	<b>35.1</b>	<b>33.1</b>	36.2	<b>35.2</b>	42.0	33.8	<b>33.1</b>	<b>44.1</b>	<b>52.0</b>	36.5	<b>33.0</b>	<b>37.3</b>	25.0	27.0	35.8
Fine-tuning	<b>32.7</b>	35.3	<b>33.1</b>	<b>35.9</b>	35.3	<b>41.7</b>	<b>33.7</b>	33.3	<b>44.1</b>	<b>52.0</b>	<b>36.4</b>	<b>33.0</b>	<b>37.3</b>	<b>24.9</b>	<b>26.6</b>	<b>35.7</b>

Table 3 details the results of our fine-tuning process. We select Pavlo *et al.* [16] as the target lifting model and fine-tune it. We focus on comparing the result after adjustment and the result after fine-tuning. While the lifting model already performs well, fine-tuning demonstrates incremental improvements, particularly noticeable in dynamic actions like "Walk" (0.3 mm) and "Walk Together" (0.5 mm). This highlights the effectiveness of leveraging bone length cues to refine the model's predictions.

**Table 4:** Comparison on Parameters, frame per second (FPS), and MPJPE. The evaluation is performed without test-time augmentation.

Model	Frames	Parameters (M)	FPS	MPJPE (mm)
Pavlo <i>et al.</i> [16]	243	16.95	958	46.8
Chen <i>et al.</i> [2]	243	59.18	197	44.2
Zheng <i>et al.</i> [30]	81	9.60	379	44.3
Pavlo <i>et al.</i> [16] (with adjustment)	243	19.34	435	45.6
Chen <i>et al.</i> [2] (with adjustment)	243	61.57	154	44.0
Zheng <i>et al.</i> [30] (with adjustment)	81	11.99	252	43.3
Chen <i>et al.</i> [2] (bone length model)	-	8.56	715	-
Ours, GRU model	-	2.39	2097	-

In Table 4, we evaluate the inference efficiency across three scenarios: (1) the lifting models alone, (2) the lifting models with our adjustment process, and (3) the bone length models.

We measure the frames per second (FPS) for these models during real-time online processing, where each model predicts a single frame at a time. The horizontal flip augmentation is not applied in the evaluation. We repeat the inference step 10,000 times, simulating a test on a 10,000-frame video, using a single GeForce GTX 3060 Ti GPU. As our Bi-GRU model is unsuitable for online processing, we test using our GRU model instead. After applying our adjustment process, the MPJPE loss improves significantly with minimal overhead in model size and computation time. Although 16 with adjustment runs at half the FPS compared to without adjustment, it remains faster than the other models listed. Additionally, the complete human pose estimation includes 2D keypoint detection and 2D-to-3D lifting. The FPS of most 2D keypoint detection models is lower than 100. Thus our approach will not be the bottleneck.

For the bone length model, our approach updates bone length values faster than 2. Our model requires only the input of the new frame at each step, as past information is stored in the hidden state, whereas 2 needs to randomly select 50 frames from previous inputs to predict bone lengths. The FPS is limited by our adjustment process that we decompose poses into bone directions and reconstruct poses with inferred bone lengths.

In summary, our adjustment and fine-tuning methodologies enhance the robustness and accuracy of existing 3D lifting models, demonstrating their efficacy in improving pose estimation across different evaluation protocols and dynamic scenarios. In real-time online processing, our adjustments achieve competitive results with minimal efficiency overhead.

## 6 Conclusion

We introduced a novel approach to enhance 3D human pose estimation by integrating bone length prediction and adjustment methods. Using a single-layer bidirectional GRU with synthetic data augmentation, our bone length model achieved the lowest error on the Human3.6M dataset. The bone length adjustment technique significantly reduced MPJPE and P-MPJPE errors in 2D-to-3D lifting models. Fine-tuning further improved pose accuracy, particularly for dynamic actions. Additionally, our GRU model operates efficiently in real-time, demonstrating minimal overhead from the adjustment process.

Overall, our approach effectively enhances the anatomical accuracy of 3D human pose predictions, demonstrating significant improvements in error metrics and robustness across various models.

## References

1. Catalin Ionescu, Fuxin Li, C.S.: Latent structured models for human pose estimation. In: International Conference on Computer Vision (2011)

2. Chen, T., Fang, C., Shen, X., Zhu, Y., Chen, Z., Luo, J.: 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)
3. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 7103–7112 (2018)
4. Cheng, Y., Wang, B., Yang, B., Tan, R.T.: Graph and temporal convolutional networks for 3d multi-person pose estimation in monocular videos. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 35, pp. 1157–1165 (2021)
5. Choi, H., Moon, G., Lee, K.M.: Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*. pp. 769–787. Springer (2020)
6. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014)
7. Gong, J., Foo, L.G., Fan, Z., Ke, Q., Rahmani, H., Liu, J.: Diffpose: Toward more reliable 3d pose estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 13041–13051 (2023)
8. Hossain, M.R.I., Little, J.J.: Exploiting temporal information for 3d human pose estimation. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 68–84 (2018)
9. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: 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 (jul 2014)
10. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 7122–7131 (2018)
11. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 5253–5263 (2020)
12. Lee, K., Lee, I., Lee, S.: Propagating lstm: 3d pose estimation based on joint interdependency. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 119–135 (2018)
13. Li, W., Liu, H., Tang, H., Wang, P., Van Gool, L.: Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 13147–13156 (2022)
14. Liu, Q., Zhang, Y., Bai, S., Yuille, A.: Explicit occlusion reasoning for multi-person 3d human pose estimation. In: *European Conference on Computer Vision*. pp. 497–517. Springer (2022)
15. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* **34**(6), 248:1–248:16 (Oct 2015)
16. Pavlo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 7753–7762 (2019)
17. Peng, J., Zhou, Y., Mok, P.: Ktpformer: Kinematics and trajectory prior knowledge-enhanced transformer for 3d human pose estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 1123–1132 (2024)

18. Shan, W., Liu, Z., Zhang, X., Wang, Z., Han, K., Wang, S., Ma, S., Gao, W.: Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14761–14771 (2023)
19. Sminchisescu, C., Triggs, B.: Kinematic jump processes for monocular 3d human tracking. In: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings. vol. 1, pp. I–I. IEEE (2003)
20. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5693–5703 (2019)
21. Sun, X., Shang, J., Liang, S., Wei, Y.: Compositional human pose regression. In: Proceedings of the IEEE international conference on computer vision. pp. 2602–2611 (2017)
22. Tang, Z., Qiu, Z., Hao, Y., Hong, R., Yao, T.: 3d human pose estimation with spatio-temporal criss-cross attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4790–4799 (2023)
23. Trotter, M., Gleser, G.C.: Estimation of stature from long bones of american whites and negroes. *American journal of physical anthropology* **10**(4), 463–514 (1952)
24. Wei, W.L., Lin, J.C., Liu, T.L., Liao, H.Y.M.: Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13211–13220 (2022)
25. Xu, J., Guo, Y., Peng, Y.: Finepose: Fine-grained prompt-driven 3d human pose estimation via diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 561–570 (2024)
26. Zhan, Y., Li, F., Weng, R., Choi, W.: Ray3d: ray-based 3d human pose estimation for monocular absolute 3d localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13116–13125 (2022)
27. Zhang, J., Tu, Z., Yang, J., Chen, Y., Yuan, J.: 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. pp. 13232–13242 (2022)
28. Zhao, Q., Zheng, C., Liu, M., Chen, C.: A single 2d pose with context is worth hundreds for 3d human pose estimation. *Advances in Neural Information Processing Systems* **36** (2024)
29. Zhao, Q., Zheng, C., Liu, M., Wang, P., Chen, C.: Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8877–8886 (2023)
30. Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z.: 3d human pose estimation with spatial and temporal transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11656–11665 (2021)