sensors

MDPI

Article

Whole-Body 3D Pose Estimation Based on Body Mass Distribution and Center of Gravity Constraints

Fan Wei [1], Guanghua Xu [1,2,3,4,*], Qingqiang Wu [1], Penglin Qin [1], Leijun Pan [1] and Yihua Zhao [1]

[1] School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China; wf3117370016@stu.xjtu.edu.cn (F.W.); wuqingqiang@xjtu.edu.cn (Q.W.); qplqplqpl@stu.xjtu.edu.cn (P.Q.); panleijun@stu.xjtu.edu.cn (L.P.); zyh111@stu.xjtu.edu.cn (Y.Z.)
[2] The State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an 710054, China
[3] The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, China
[4] State Industry-Education Integration Center for Medical Innovations, Xi'an Jiaotong University, Xi'an 710049, China
* Correspondence: ghxu@xjtu.edu.cn

Abstract

Estimating the 3D pose of a human body from monocular images is crucial for computer vision applications, but the technique remains challenging due to depth ambiguity and self-occlusion. Traditional methods often suffer from insufficient prior knowledge and weak constraints, resulting in inaccurate 3D keypoint estimation. In this paper, we propose a method for whole-body 3D pose estimation based on a Transformer architecture, integrating body mass distribution and center of gravity constraints. The method maps the pose to the center of gravity position using the anatomical mass ratio of the human body and computes the segment-level center of gravity using the moment synthesis method. A combined loss function is designed to enforce consistency between the predicted keypoints and the center of gravity position, as well as the invariance of limb length. Extensive experiments on the Human 3.6M WholeBody dataset demonstrate that the proposed method achieves state-of-the-art performance, with a whole-body mean joint position error (MPJPE) of 44.49 mm, which is 60.4% lower than the previous Large Simple Baseline method. Notably, it reduces the body part keypoints' MPJPE from 112.6 to 40.41, showcasing the enhanced robustness and effectiveness to occluded scenes. This study highlights the effectiveness of integrating physical constraints into deep learning frameworks for accurate 3D pose estimation.

Keywords: deep learning; 3D pose estimation; whole-body pose; Transformer; body mass distribution; center of gravity constraints

check for updates

Academic Editors: Yun Zhang, Sam Kwong, Xu Long and Tiesong Zhao

Received: 22 May 2025
Revised: 24 June 2025
Accepted: 24 June 2025
Published: 25 June 2025

Citation: Wei, F.; Xu, G.; Wu, Q.; Qin, P.; Pan, L.; Zhao, Y. Whole-Body 3D Pose Estimation Based on Body Mass Distribution and Center of Gravity Constraints. *Sensors* **2025**, *25*, 3944. https://doi.org/10.3390/s25133944

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

1.1. Background and Challenges

As a hot topic in the intersection of computer vision and artificial intelligence, human 3D pose estimation has an extremely important research status and broad application prospects. As one of the challenging problems in computer vision, it has received in-depth attention from researchers for some time. As basic research work, human pose estimation has a promoting effect on many downstream studies, including human–computer interaction [1], scene modeling [2], motion analysis [3], smart medical care [4], and virtual reality [5].

However, accurately estimating the 3D pose of the human body from a 2D image is a challenging task. Depth ambiguity under a monocular view is a major obstacle [6]. In a 2D image, only the projection positions of body keypoints at a certain angle are included, and the corresponding depth information is lost. However, since the 2D pose is a projection of the 3D pose, the same 2D pose may correspond to multiple 3D poses [7], which makes the mapping from the 2D pose to the 3D pose uncertain. The self-occlusion problem also makes pose estimation difficult. When some parts of the human body are occluded by other parts, it is difficult for vision-based methods to obtain accurate information about the occluded parts [8], thus affecting the accuracy of pose estimation. Moreover, 3D pose estimation faces many problems, such as complex backgrounds [9], illumination changes [10], and scale changes [11].

The existing methods for estimating the 3D body pose on the basis of monocular color images usually adopt two-stage solutions [12–18]. These methods first use an off-the-shelf 2D detector [16,19–22] to estimate the 2D pose of the human body from RGB images and then use deep learning methods [23–25] to lift the 2D pose in order to obtain accurate 3D pose estimation results. These methods make full use of the high precision and generalization ability of current 2D pose estimation technology and provide relatively accurate basic information for subsequent 3D pose estimation. Moreover, the two-stage methods have a relatively simple network structure and high computational efficiency, which makes them easy to implement in practical applications. However, these methods are highly dependent on the accuracy of 2D pose estimation results. When the keypoint detection is inaccurate or the keypoint position cannot be identified due to occlusion, the error is eliminated in the subsequent coordinate lifting process, thereby affecting the final 3D pose estimation results.

In addition, existing methods for estimating the whole-body 3D pose of the human body generally estimate the keypoint positions directly, lacking constraints on the keypoint positions, resulting in a loss of accuracy in the keypoint position estimation. When part of the human body is occluded, the direct regression 3D coordinate methods use the model to learn complex mapping relationships directly from the images. Due to the lack of information about the occluded part, the model has difficulty capturing the complete posture features. The method of first regressing and then lifting the coordinates cannot accurately detect the occluded keypoints, which seriously affects the final 3D posture estimation results. In addition, traditional methods do not effectively constrain the estimated positions of keypoints and make insufficient use of prior knowledge.

1.2. Related Work

1.2.1. Transformer Structure

In the task of 3D human pose estimation, traditional methods can be roughly divided into two categories according to the network model used: convolutional neural network-based methods [2,7,26–31] and Transformer-based methods [25,32–40]. Convolutional neural networks can efficiently capture underlying visual features such as edge textures in images, while the weights and local connection characteristics of convolution operations make their complexity and number of parameters much lower than the Transformer structure. Moreover, convolutional neural networks expand the receptive field by stacking multiple layers of convolution, which makes it difficult to directly model the dependencies of distant joints, resulting in weak global dependency modeling capabilities. The self-attention mechanism of the Transformer structure can directly calculate the dependencies between any two keypoints and has strong global correlation modeling capabilities. Zheng et al. [37] proposed a pure Transformer framework that uses the 2D joint coordinates of multiple consecutive frames as inputs, uses a network structure to extract cross-frame

temporal features and single-frame spatial features, and encodes the keypoint positions in time and space, thereby solving the depth blur and occlusion problems in video sequences. Most of the subsequent Transformer-based methods have improved upon this method. Li et al. [32] proposed a multi-hypothesis Transformer structure, which models multiple hypotheses for the same pose and selects the one with the highest credibility as the final result. Zhang et al. [34] proposed a mixed spatiotemporal Transformer model that uses the Transformer structure to alternately extract the temporal and spatial features of keypoints.

In this work, we propose using the Transformer structure to model the position encoding of keypoints, use the change in the center of gravity position and posture to establish a mapping relationship, and use the mass distribution of body parts to constrain the estimated position of each keypoint, so that the model can better understand the structural limitations of the human pose, thereby improving the authenticity and effectiveness of the pose estimation results.

1.2.2. Whole-Body 3D Pose Estimation

Existing whole-body 3D human pose estimation methods can be divided into end-to-end methods [41–46] and multistage methods. End-to-end methods directly regress 3D coordinates from 2D joint points and build a general whole-body pose estimation model on the basis of data-driven explicit priors. Wen et al. [41] used a semantic graph attention network to model the semantic relationships between joints and used the distance information between the joints as auxiliary supervision to enhance the rationality of the pose. Zhu et al. [42] proposed the fine annotation of the keypoints of the whole body and combined the Transformer structure to model video sequence dependencies and geometric constraints to improve the consistency of body and hand movements. Semet et al. [45] divided the human body into multiple parts, modeled different parts independently, and introduced a temporal attention mechanism to extract the motion patterns of adjacent frames in the feature extraction process of spatiotemporal sequences, improving the adaptability of the model to complex scenes.

The multistage method [47–53] divides the keypoints of the whole body into body parts, faces, and hands, estimates them separately, and gradually refines them through a cascade network. Rong et al. [47] proposed dividing the human body into three independent modules and performed posture regression separately. Moreover, the output of each module is integrated through coordinate transformation and optimization fitting to ensure the rationality of the movement. Weinzaepfrl et al. [49] proposed an integrated framework based on knowledge distillation, which divides keypoints into body, hand, and face, builds expert models for each keypoint for training, and unifies them into the global model. This method is robust under complex backgrounds, occlusions, and lighting changes, but the model has a large number of parameters, and the quality of the local expert model directly affects the performance of the global model.

In this work, we propose a whole-body 3D pose estimation method based on weight distribution and center of gravity constraints. Based on the human body structure, we use the network to establish the mapping relationship between different postures and center of gravity positions and use the body mass distribution relationship to constrain the estimated positions of keypoints. Considering the ambiguity of depth information, multiple encoding methods are proposed to encode the keypoint positions, and the most similar result is selected from multiple estimation results as the final prediction result to improve the estimation accuracy of the whole-body 3D pose.

1.3. Research Gap and Our Contributions

Existing methods directly regress joint positions without enforcing anatomical plausibility, resulting in physically unreasonable poses, such as limb penetration or body proportion distortion under occlusion or depth ambiguity. At the same time, existing methods make insufficient use of prior information and do not fully consider the biomechanical properties of human kinematics when estimating poses, resulting in occlusion sensitivity.

To bridge these gaps, we propose a method for estimating the whole-body 3D pose on the basis of the body mass distribution and center of gravity constraints. Based on the limb connection relationship and the Transformer structure, the positions of the keypoints are encoded, and a mapping relationship between the different postures and center of gravity positions is established. The estimated positions of the keypoints are constrained by the mass relationship of each body segment so that the model can better understand the physical limitations of human poses, thereby improving the authenticity and effectiveness of the pose estimation results.

Our contributions to the whole-body 3D pose estimation include the following aspects:

1. We propose a new method for whole-body 3D pose estimation on the basis of the body mass distribution and center of gravity position constraints, and our method achieves the SOTAs on a widely used dataset.
2. Based on the observation of different postures and weight changes in the human body and the research of existing methods, the changing rules of the center of gravity of the human body under different postures are fully considered, and the prior information in the process of estimating the human body posture is increased in the network, which effectively solves the problem of the inaccurate estimation of keypoint positions due to occlusion in the posture estimation process.
3. On the basis of the mass distribution relationship of human body segments, we add body mass distribution constraints when the network is used to estimate the whole-body poses, establish a mapping relationship between the center of gravity position and the human body pose, and use two different methods to encode the keypoint position information, thereby improving the accuracy of pose estimation.

2. Methodology

Inspired by the above methods, we propose a method for estimating the whole-body 3D pose on the basis of the body segment mass distribution and center of gravity constraints. We use the Transformer structure to model the relationships between human keypoints and encode the keypoint positions in two different ways. We use the body segment mass distribution relationship to estimate the center of gravity position and use the network to model the mapping relationship between the center of gravity position and the pose. Figure 1 shows the overall framework of our proposed method. The input is 2D keypoints from the detector, bone vector calculation and bone position calculation are used as constraints, the Transformer with the cross-attention mechanism is the main network framework, the weight is estimated through mass distribution, and a loss function based on keypoint distribution and the center of gravity position is designed.

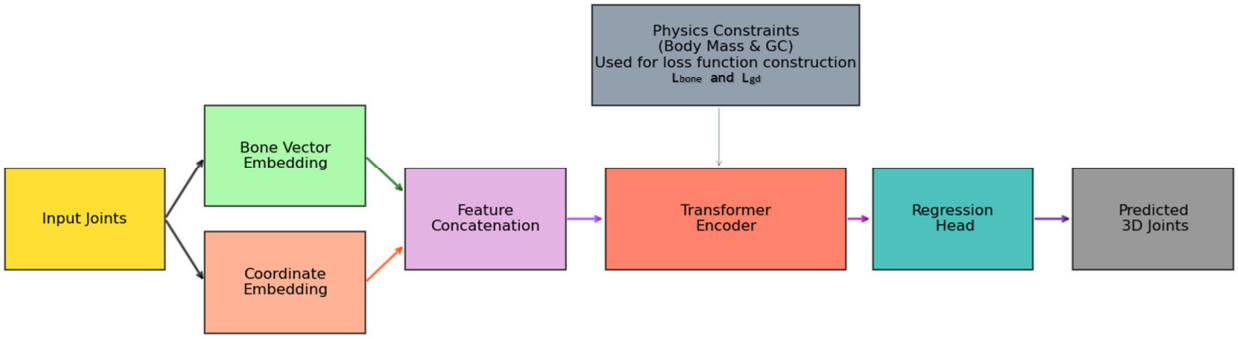


Figure 1. Overview framework of the proposed method. Input joints: 2D keypoints from detector → 3.1. datasets and evaluation metrics; bone vector embedding → 2.1 keypoint position encoding based on the human body structure; coordinate embedding → 2.2 body mass distribution and center of gravity constraints; Transformer with regression head → 3.2. implementation details; loss functions with physics constraints (L, L_{gd}, L_{bone}) → 2.3 loss function.

2.1. Keypoint Position Encoding Based on the Human Body Structure

Under the camera imaging principle of the pinhole imaging model [54], the three-dimensional coordinates $P(x, y, z)$ of a point in the scene are related to the projection of its pixel coordinate point $J(u, v)$ in the image, as expressed in Equation (1):

$$\begin{cases} \frac{x}{u} = \frac{z}{f_x} \\ \frac{y}{v} = \frac{z}{f_y} \end{cases} \tag{1}$$

In (1), f_x and f_y represent the focal length of the camera lens. The camera imaging principle based on the pinhole imaging model is shown in Figure 2.

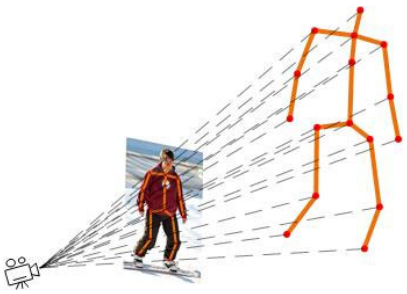


Figure 2. Schematic diagram of imaging principle based on the pinhole model.

According to the principle of constant limb length, the two joint points have the following relationship, as expressed in Equation (2):

$$\begin{cases} \sqrt{(\Delta_x)^2 + (\Delta_y)^2 + (\Delta_z)^2} = L \\ \Delta_x = u_1 \cdot \frac{z_1}{f_x} - u_2 \cdot \frac{z_1 + \Delta_z}{f_x} \\ \Delta_y = v_1 \cdot \frac{z_1}{f_y} - v_2 \cdot \frac{z_1 + \Delta_z}{f_y} \end{cases} \tag{2}$$

In (2), Δ_z represents the vertical coordinate position error of keypoints J_1 and J_2. According to the above equation, since the length of the human limbs remains unchanged, the positions of the keypoints in different coordinate systems can be converted to each other via the above equation.

When a coordinate system is used to represent the position of keypoints, due to the limitations of image pixels and resolution, precision loss occurs when the keypoint positions in the real world are mapped to the image pixel coordinate system, resulting in deviations in the real keypoint positions. To reduce this deviation, this method uses a bone vector

method to represent the keypoint positions. First, the keypoints of the human body are modeled according to the human body structure and simplified into a tree structure, with the keypoint at the human pelvis as the root node and other keypoints expanding outward from the root node [55]. The simplified human body model is shown in Figure 3.

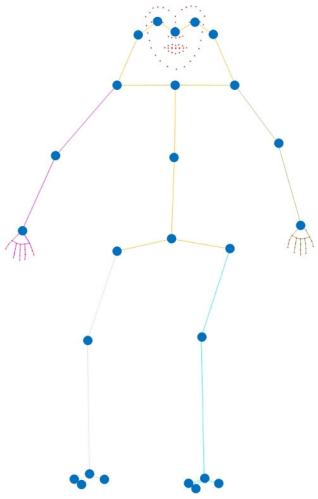


Figure 3. Simplified tree diagram model of whole-body keypoints.

As shown in Figure 3, the human body is connected by multiple joints and bones, which can be regarded as a tree structure with the center of the pelvis as the starting point. On the basis of the above connection relationship, this method uses the bone vector representation to encode the position of each keypoint. Through the bone vector representation, each keypoint can be regarded as the endpoint of a vector, and the bones connecting adjacent keypoints can be represented by a vector. The vector not only defines the length of the bone but also determines the direction and position of the bone in three-dimensional space, thereby fully describing the posture of the human body.

The position of the elbow joint is taken as an example. In coordinate notation, its coordinates are expressed as $E(x_e, y_e, z_e)$, and the coordinates of the adjacent shoulder joint and wrist joint are expressed as $S(x_s, y_s, z_s)$ and $W(x_w, y_w, z_w)$, respectively. In the bone vector representation, the bone vector from the elbow to the shoulder is represented as $\vec{ES} = (x_s - x_e, y_s - y_e, z_s - z_e)$, and the bone vector from the elbow to the wrist is represented as $\vec{EW} = (x_w - x_e, y_w - y_e, z_w - z_e)$. Moreover, the length of the upper arm between the shoulder joint and the elbow joint can be expressed in Equation (3):

$$L_{ua} = \left|\vec{ES}\right| \tag{3}$$

In (3), $\vec{SE}$ is the bone vector from the elbow to the shoulder. And the length of the forearm between the elbow joint and the wrist joint can be expressed in Equation (4):

$$L_{fa} = \left|\vec{EW}\right| \tag{4}$$

In (4), $\vec{EW}$ is the bone vector from the elbow to the wrist. So the angle of the elbow joint can be expressed in Equation (5):

$$\theta_e = \arccos \frac{\vec{ES} \cdot \vec{EW}}{L_{ua} \cdot L_{fa}} \tag{5}$$

In (5), $\overrightarrow{ES}$ and $\overrightarrow{EW}$ are the bones vectors from the elbow to the shoulder and wrist, respectively. And the L_{ua} and L_{fa} are the lengths of the corresponding bone vector.

We use the Transformer structure to encode the two representation methods into feature vectors. For the encoding of the coordinate point representation method, each coordinate point p_i is linearly embedded as expressed in Equation (6):

$$e_i^p = W_p p_i + b_p \tag{6}$$

In (6), by adding a learnable positional encoding, its topological order can be preserved. For the encoding of the bone vector representation method, the length and direction of each bone vector encoder are as expressed in Equation (7):

$$e_i^b = W_b\left(\frac{b_i}{\|b_i\|_2} \oplus \log(\|b_i\|_2 + \varepsilon)\right) + b_b \tag{7}$$

Moreover, a dual-branch cross-attention structure is used to model the coordinate point modality and the bone vector modality to establish the global spatial dependency between joints and the local motion constraints of the bone chain to reduce posture ambiguity.

2.2. Body Mass Distribution and Center of Gravity Constraints

To make the keypoint estimation results more consistent with human kinematics, this method divides the human body into reasonable segments and accurately calculates the mass distribution of each segment. The body segments are divided according to the body parts as shown in Figure 4.

Table 1. Relative Mass Distribution of each body segment.

Body Segment	Gender	Relative Quality/%	Body Segment	Gender	Relative Quality/%
Neck	M	8.62	Upper Arm	M	2.43
	F	8.20		F	2.66
Upper Torso	M	16.82	Forearm	M	1.25
	F	16.35		F	1.14
Lower Torso	M	27.23	Hand	M	0.64
	F	27.48		F	0.42
Thigh	M	14.19	Foot	M	1.48
	F	14.10		F	1.24
Calf	M	3.67			
	F	4.43			

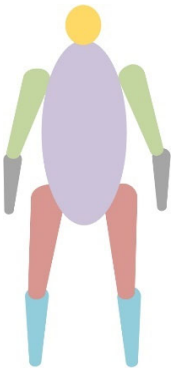


Figure 4. Schematic diagram of body part segmentation based on segment mass relationships. Segment division aligns with mass coefficients in Table 1, following [56].

According to the structural characteristics of human anatomy, the human body is divided into multiple main segments, such as the head, trunk, upper limbs, and lower limbs. According to the "Inertial Parameters of Adult Human Body [56]", we establish the connection between the keypoint position and the mass of the body segment and assign relative mass coefficients to each body segment to regard the human body model as a real physical model. The relative mass distribution of the "Inertial Parameters of the Adult Human Body" is shown in Table 1.

The center of gravity is the balance point of the body's mass distribution. Corresponding to the structure of the human body, the position of the center of gravity also changes at any time. Since each part of the human body can be regarded as a rigid body, the position of the center of gravity can be obtained by the algebraic sum of the moments of each part. Therefore, we use the moment synthesis method to determine the center of gravity position and establish a mapping relationship between the different postures of the human body and the center of gravity position.

In the human anatomy model, the human body is divided into different parts. The part close to the heart is called the proximal segment, and the part far from the heart is called the distal segment. Each part is regarded as an independent rigid body model. Taking the right upper arm as an example, the proximal end of the segment is the right shoulder keypoint, and the distal end of the segment is the right elbow keypoint. The center of gravity position is shown in Figure 5.

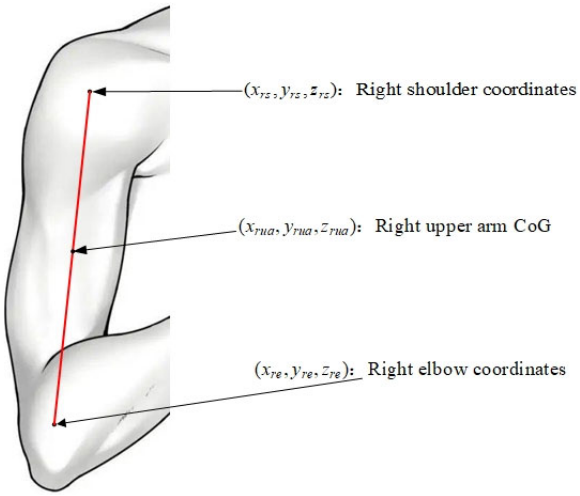


Figure 5. The center of gravity position of the right upper arm based on the moment synthesis method.

Multiply the right shoulder joint coordinates by the corresponding proximal coefficient and the right elbow joint coordinates by the corresponding distal coefficient to obtain the three-dimensional center of gravity coordinates of the right upper arm as expressed in Equation (8):

$$\begin{cases} x_{rua} = x_{rs} \cdot l_p + x_{re} \cdot l_d \\ y_{rua} = y_{rs} \cdot l_p + y_{re} \cdot l_d \\ z_{rua} = z_{rs} \cdot l_p + z_{re} \cdot l_d \end{cases} \tag{8}$$

In (8), $(x_{rua}, y_{rua}, z_{rua})$ represents the three-dimensional center of gravity coordinates of the right upper arm; (x_{rs}, y_{rs}, z_{rs}) and (x_{re}, y_{re}, z_{re}) represent the coordinate positions of the right shoulder joint and the right elbow joint, respectively. l_p and l_d represent the proximal coefficient and distal coefficient of the right upper arm segment, respectively. For the human body, the center of gravity position is the algebraic sum of the center of gravity

positions of each body segment. That is, the three-dimensional center of gravity coordinate position of the human body as a whole is expressed in Equation (9):

$$\begin{cases} X_{CoG} = \sum_i k_i x_i \\ Y_{CoG} = \sum_i k_i y_i \\ Z_{CoG} = \sum_i k_i z_i \end{cases} \tag{9}$$

In (9), $(X_{CoG}, Y_{CoG}, Z_{CoG})$ represents the three-dimensional center of gravity coordinate position of the human body as a whole, (x_i, y_i, z_i) represents the three-dimensional center of gravity coordinate position of each body segment, and k represents the mass coefficient of the body segment. To obtain the accurate center of gravity position, this paper obtains the corresponding body segment mass coefficient on the basis of the relative mass distribution of each body segment in the "Inertial Parameters of Adult Human Body", as shown in Table 2.

Table 2. Relative positions of the center of mass of each body segment.

Body Segment	Gender	L_{es}	L_{ex}	Body Segment	Gender	L_{es}	L_{ex}
Neck	M	46.9	53.1	Upper Arm	M	47.8	52.2
	F	47.3	52.7		F	46.7	53.3
Upper Torso	M	53.6	46.4	Forearm	M	42.4	57.6
	F	49.3	50.7		F	45.3	54.7
Lower Torso	M	40.3	59.7	Hand	M	36.6	63.4
	F	44.6	55.4		F	34.9	65.1
Thigh	M	45.3	54.7	Foot	M	48.6	51.4
	F	44.2	55.8		F	45.1	54.9
Calf	M	39.3	60.7	Overall Centroid	M	43.8	56.2
	F	42.5	57.5		F	44.5	55.5

L_{es} and L_{ex} represent the percentage of the upper and lower dimensions of the centroid of each body segment to the total length of the body segment, respectively.

2.3. Loss Function

To train and optimize the model, it is necessary to design the corresponding loss function. In the traditional method of human 3D pose estimation, the commonly used loss function is the average joint position error loss function, and its functional form is expressed in Equation (10):

$$L = \frac{1}{N}\sum_{i=1}^{N} \|p_i - \hat{p}_i\|_2 \tag{10}$$

In Equation (10), p_i and $\hat{p}_i$ represent the predicted position and ground truth position of the keypoints, respectively. N represents the total number of keypoints.

This method uses the weight distribution relationship and the center of gravity position to constrain the position of the keypoint estimation. Because of the mapping relationship between different postures and the center of gravity position, we also designed the corresponding loss function as expressed in Equation Equation (11):

$$L_{gd} = \frac{1}{N}\sum_{i=1}^{N} \|g_i - \hat{g}_i\|_2 \tag{11}$$

In Equation (11), N represents the number of all keypoints, g_i represents the distance from the predicted keypoint to the predicted center of gravity, and $\hat{g}_i$ represents the distance from the ground truth keypoint to the ground truth center of gravity.

Moreover, the invariance of the limb length is used as a constraint to design the loss function. In the bone vector representation method, for a certain keypoint, two adjacent keypoints and their angles are used to represent the current keypoint, so the consistency of the limb length and the cosine similarity of the angle are used as the measurement criteria, as expressed in Equation Equation (12):

$$L_{bone} = \frac{1}{N-1}\sum_{i=1}^{N-1}(1 - \frac{b_i \cdot \hat{b}_i}{\|b_i\|\left\|\hat{b}_i\right\|}) + \lambda\left\||b_i| - \left|\hat{b}_i\right|\right\|_2 \quad (12)$$

In (12), N represents the total number of keypoints, b_i represents the predicted limb vector, and $\hat{b}_i$ represents the ground truth limb vector. Therefore, the loss function used in this paper is expressed in Equation (13):

$$Loss_{total} = L + L_{gd} + L_{bone} \quad (13)$$

In this work, the weighted sum of multiple loss functions is used as the total loss function to participate in network training, and different weight parameters are used during the training process.

3. Experiments

3.1. Datasets and Evaluation Metrics

In this work, we use the Human 3.6M WholeBody [42] public dataset for network training and verification. The Human 3.6M WholeBody dataset is an annotation extension of the Human 3.6M [57] dataset. It follows the keypoint layout used in the COCO-WholeBody [58] dataset and focuses on the 3D body pose annotation and estimation tasks. On the basis of the original annotation, the detailed keypoint annotations of the face, hands, and feet are added. The Human 3.6M WholeBody dataset consists of 133 pairs of 2D and 3D whole-body keypoint annotations from a set of 100,000 images in the Human3.6 M dataset. The training set used in this paper contains 80,000 images and 2D and 3D pose triplets; the test set contains 20,000 images and 2D and 3D pose triplets.

To verify the accuracy of the pose estimation results, we use the commonly used evaluation indicator, the mean joint position error (MPJPE). This indicator refers to the average Euclidean distance error between the predicted joint points and the actual joint points, and its formula is expressed in Equation (14):

$$MPJPE = \frac{1}{N}\sum_{i=1}^{N}\|p_i - \hat{p}_i\| \quad (14)$$

In (14), N is the number of keypoints, and p_i and $\hat{p}_i$ are the predicted coordinates and ground truth coordinates of the i-th joint point, respectively.

3.2. Implementation Details

We used NVIDIA RTX2080Ti GPUs to train and validate the network on PyTorch 1.13.1. We used an off-the-shelf 2D keypoint detector [44] and normalized the coordinates centered on the pelvis. We performed data augmentation on the input data, including random rotation ($\pm 15^\circ$), scaling (0.8–1.2), and flipping. These operations preserve anatomical constraints (e.g., limb length) while expanding the training distribution. The deep learning model configuration includes the Transformer architecture and dual-branch cross-attention. The Transformer structure has 6 layers, 8 attention heads, and a hidden layer dimension of 256. The dual-branch cross-attention has four layers for feature fusion. We used the AdamW optimizer to train the network for 210 epochs with a batch size of 64 and an

initial learning rate of 0.001. We reduced the learning rate at the 170th and 200th epochs, respectively, and the learning rate decay strategy was cosine annealing.

3.3. Performance Comparison

Results on the Human 3.6M WholeBody dataset. Table 3 shows the prediction accuracy of different body part keypoints in the test set. Following the 2D-to-3D lifting method, we used the existing 2D whole-body keypoint pose detector and the detected 2D pose results as input to predict the 3D pose of the whole-body keypoints. As shown in Table 3, our proposed method achieves the best results in 3D pose estimation for all parts' keypoints. Among them, the MPJPE index for the whole-body keypoints reached 22.07, which is 20.5% higher than that of the current best method. Moreover, the MPJPE index of our proposed method for the estimation of facial keypoints reached 6.09, which is 41.67% higher than that of the current best method. This is because when we perform pose estimation, we establish a mapping between different poses and center of gravity positions and use the body mass distribution to constrain the estimated position of the keypoints, thereby reducing the error between the estimated keypoints and the actual keypoint positions. Moreover, the model achieves a computational speed of 28 FPS on an NVIDIA RTX 2080Ti, with a computational cost of 12.5 GFLOPs and 38M parameters, meeting the near-real-time rendering requirement (30 FPS). This is faster than JointFormer [40] (22 FPS) and SemGAN [41] (19 FPS). This shows the real-time performance of the proposed model, providing a basis for further online applications.

Table 3. Results for whole-body 3D pose estimation on the H3WB dataset.

Method	WholeBody	Body	Face	Hands
Simple Baseline [42]	125.4	125.7	24.6	42.5
Large Simple Baseline [21]	112.3	112.6	14.6	31.7
JointFormer [40]	88.3	84.9	17.8	43.7
3D-LFM [59]	64.13	60.83	10.44	28.22
SemGAN [41]	47.87	45.39	15.95	27.77
Ours	44.49	40.41	6.09	22.07

3.4. Ablation Study

To validate the contributions of the components of our proposed method and the corresponding loss functions, we designed ablation experiments on the Human 3.6M WholeBody dataset.

3.4.1. Effects of the Body Mass Distribution and Center of Gravity Position Constraints

To verify the influence of the proposed method on the whole-body 3D pose estimation on the basis of the body mass distribution and the center of gravity position constraints, we selected the methods of adding only the body mass distribution, only the center of gravity position constraints, and both for the keypoints and designed an ablation experiment on the Human 3.6M WholeBody dataset. For the sake of fairness, the other network parameters were set exactly the same, and the ablation experiment results are shown in Table 4.

Table 4. Effects of adding body mass distribution and center of gravity position constraints.

Constraints	WholeBody	Body	Face	Hands
body mass distribution constraints	67.50	59.24	10.10	29.41
center of gravity position constraints	58.34	50.76	8.32	25.73
body mass distribution and center of gravity constraints	44.49	40.41	6.09	22.07

Table 4 shows that for the estimation of whole-body keypoints, the MPJPE result of adding only the body mass distribution constraints is 67.5, whereas the MPJPE result of adding the body mass distribution and the center of gravity position is 44.49, which is an improvement of 34.08%. This further shows that by encoding the keypoint information via the physical structure of the human body and the center of gravity position, and using the body mass distribution relationship and the center of gravity position to estimate the keypoint position, the accuracy and robustness of the keypoint position estimation can be improved. At the same time, it can be seen from the table that the interaction between the weight and center of mass constraints is crucial: weight distribution can enhance the consistency of limb lengths, while center of mass constraints can ensure the rationality of global posture. As shown in Table 4, combining the two can reduce the MPJPE by 34.08% compared with the constraints based on weight alone, which shows that the center of mass provides a global physical prior that can complement the local constraints based on weight.

3.4.2. Impact of the Loss Function

Table 5 compares the results under different loss functions. We compared the effects of different combinations of loss functions on the final pose estimation results. As shown in the table, when the center of gravity position loss function is added, the MPJPE result for whole-body keypoints is 51.75, which is 13.3% greater than the result when only the L loss function is used. When the body mass distribution loss function is added, the MPJPE result is 17.45% greater than the result when only the L loss function is used. The results once again verify that adding the body mass distribution and center of gravity position constraints contributes greatly to the 3D whole-body pose estimation.

Table 5. Ablation study results for loss function.

Loss Function	WholeBody	Body	Face	Hands
L	59.71	47.69	15.35	31.93
L_{gd}	51.75	44.21	13.29	25.61
L_{bone}	49.29	45.57	9.90	26.61
$Loss_{total}$	44.49	40.41	6.09	22.07

3.5. Discussion

This paper proposes a new method for 3D whole-body pose estimation. The network is used to establish the mapping relationship between the center of gravity position of the human body and different poses, and different body segment masses are added as constraints to improve the estimation accuracy of the whole-body keypoints. In addition, the proposed method achieved encouraging results.

Table 1 reports the comparison results of our method with other SOTAs. We use the results of an existing 2D whole-body keypoint pose detector as input. Compared with the existing SOTAs, our proposed method has better estimation results for keypoints in different parts of the body and obtains the smallest average error. This is because we use the network to establish the mapping relationship between the center of gravity position and different poses and add the body mass distribution to constrain the estimated position of the keypoints. Compared with the SemGAN [41] method, our method uses limited human structure relationships to constrain the estimated position of the keypoints when performing pose estimation, so the prior information is not fully utilized. Our method exploits the human body structure as prior information, encodes the keypoint locations in two ways, and constrains the estimated locations of keypoints using body segment mass distribution and the center of gravity. A comparison of the results of the ablation experiment also illustrates this point. The Human 3.6M WholeBody dataset follows the official split:

80,000 training images and 20,000 test images, which aligns with prior works [42,57]. Cross-validation is not performed due to the dataset's standardized protocol, but the model's generalizability is validated by outperforming SOTAs on the test set. Future work will explore cross-dataset validation on the MPII human pose. Figure 6 shows some visual comparison results. It can be seen from the figure that, in addition to the small gap between the simple pose estimation result and the true value, the proposed method can still obtain a relatively reasonable 3D pose estimation result in the case of partial occlusion. This shows the effectiveness of the proposed method and also indicate the direction for subsequent methods.

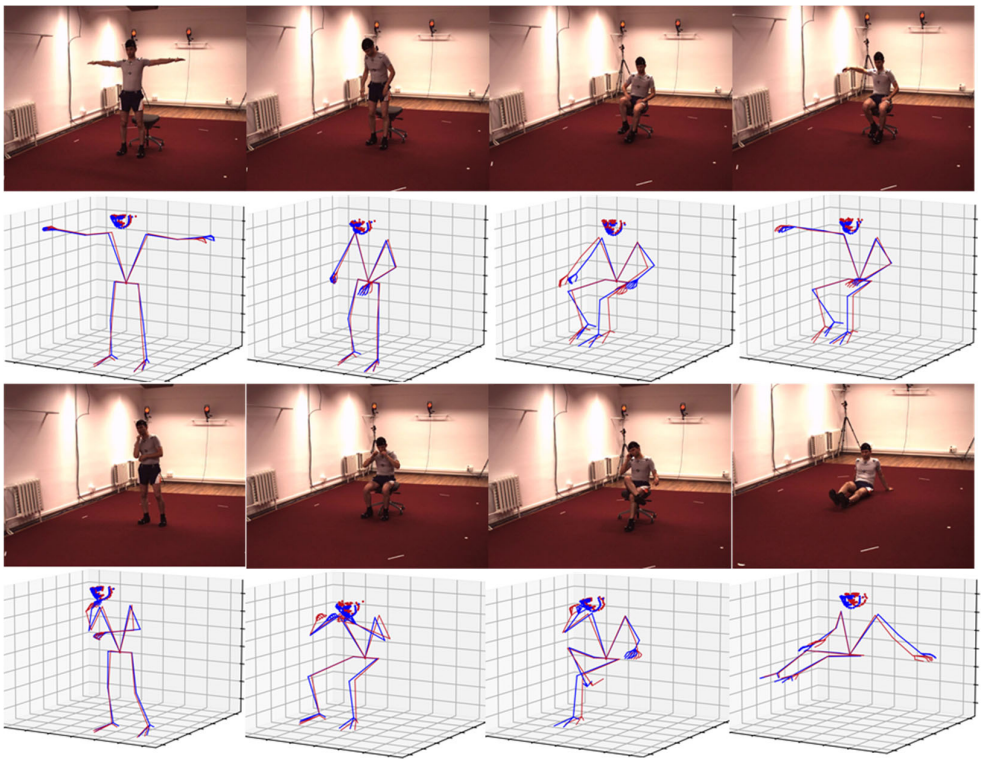


Figure 6. Visualization results of whole-body 3D pose estimation. The blue line represents the ground truth, and the red line represents the estimated result of our method.

In order to quantitatively analyze the improvement of the method in 3D pose estimation, this paper compares the estimation results of different poses with the existing methods on the Human 3.6M dataset, and the results are shown in Table 6. As can be seen from the table, compared with the existing methods, the method proposed in this paper has a significant improvement in the estimation of simple actions. For example, compared with the JointFormer method, the estimation accuracy of the method in walking action is improved by 29.2%, and the estimation accuracy of the method in eating action is improved by 27.9%. This is due to the addition of center of gravity position constraints and body segment mass constraints in the process of pose estimation. In some complex actions, especially in the presence of occlusion, this method can still obtain relatively reliable results. As shown in Figure 6, in the visual comparison results of this method and the true value, the red line representing the predicted result still remains close to the blue line representing the true result in the occlusion scene.

Table 6. Quantitative comparison on the human 3.6M whole-body dataset.

Methods	Dir.	Disc.	Eat.	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.
PoseFormer [37]	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
JointFormer [40]	45.0	48.8	46.6	49.4	53.2	60.1	47.0	46.7	59.6	67.1	51.2	47.1	53.8	39.4	42.4
Ours	41.4	38.1	33.6	41.7	42.6	50.2	42.3	40.9	57.0	61.2	43.5	46.6	43.4	27.9	30.6

One limitation of this method is that it relies on the Human 3.6M dataset, and its generalization ability to natural scene datasets needs further verification. In wild scenes, pose estimation performance may degrade due to extreme occlusions, unseen camera angles, and low-resolution inputs. Future work will explore domain adaptation techniques to enhance its robustness in different environments. Currently, our method can only estimate the keypoints defined in the dataset in a supervised manner. Moreover, 3D pose estimation of the whole-body keypoints of a single frame cannot meet the requirements of online use. In future work, we plan to achieve the real-time estimation of full-body keypoints and further improve the accuracy of full-body keypoint pose estimation. At the same time, we will develop unsupervised/semi-supervised learning for real-time deployment and integrate motion blur enhancement techniques to enhance the robustness of the model in dynamic scenes. We will also consider exploring multimodal fusion (e.g., RGB-D) to achieve cross-domain generalization and combine it with instance segmentation for multi-person pose estimation in extended crowded scenes.

4. Conclusions

In this paper, we propose a new method to estimate the whole-body 3D pose on the basis of the mass distribution of body segments and the center of gravity position constraints. This method uses a network model to encode the position information of the whole-body keypoints of the human body, establishes a mapping relationship between different postures of the human body and the center of gravity position, and constrains the estimated position of the keypoints with the mass distribution of each body segment of the human body to obtain a reliable position estimate of the whole-body keypoints of the human body. We conducted experiments on the widely used Human 3.6M WholeBody dataset. The experimental results show that this method reached a new SOTA level, and the estimation error of the keypoints of different parts improved.

Author Contributions: Conceptualization, F.W., G.X. and Q.W.; Methodology, F.W., G.X., Q.W. and P.Q.; Software, F.W., Q.W., L.P. and Y.Z.; Validation, F.W., P.Q. and L.P.; Formal analysis, F.W., G.X., Q.W. and P.Q.; Investigation, F.W. and L.P.; Resources, G.X., Q.W. and Y.Z.; Data curation, L.P. and Y.Z.; Writing—original draft, F.W.; Writing—review & editing, F.W., G.X. and Q.W.; Visualization, P.Q. and Y.Z.; Supervision, G.X., P.Q. and L.P.; Project administration, P.Q.; Funding acquisition, G.X. and Q.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key R&D Project in Shaanxi Province grant number 2024SF-ZDCYL-01-07, the Fundamental Research Funds for the Central Universities grant number xzd012024011, National Science Foundation of China grant number 62402375, Basic research funding within the school grant number xzy012024025.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Huo, R.; Gao, Q.; Qi, J.; Ju, Z. 3d human pose estimation in video for human-computer/robot interaction. In *International Conference on Intelligent Robotics and Applications*; Springer: Singapore, 2023; pp. 176–187.
2. Grabner, A.; Roth, P.M.; Lepetit, V. 3d pose estimation and 3d model retrieval for objects in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3022–3031.
3. Liang, S.; Zhang, Y.; Diao, Y.; Li, G.; Zhao, G. The reliability and validity of gait analysis system using 3D markerless pose estimation algorithms. *Front. Bioeng. Biotechnol.* **2022**, *10*, 857975. [CrossRef] [PubMed]
4. Divya, R.; Peter, J.D. Smart healthcare system-a brain-like computing approach for analyzing the performance of detectron2 and PoseNet models for anomalous action detection in aged people with movement impairments. *Complex Intell. Syst.* **2022**, *8*, 3021–3040. [CrossRef]
5. Wu, M.-Y.; Ting, P.-W.; Tang, Y.-H.; Chou, E.-T.; Fu, L.-C. Hand pose estimation in object-interaction based on deep learning for virtual reality applications. *J. Vis. Commun. Image Represent.* **2020**, *70*, 102802. [CrossRef]
6. Zhang, S.; Wang, C.; Dong, W.; Fan, B. A survey on depth ambiguity of 3D human pose estimation. *Appl. Sci.* **2022**, *12*, 10591. [CrossRef]
7. Park, S.; Hwang, J.; Kwak, N. 3D human pose estimation using convolutional neural networks with 2D pose information. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8–10 and 15–16, 2016, Proceedings, Part III 14*; Springer: Cham, Swizerland, 2016; pp. 156–169.
8. Liang, C.-J.; Lundeen, K.M.; McGee, W.; Menassa, C.C.; Lee, S.; Kamat, V.R. A vision-based marker-less pose estimation system for articulated construction robots. *Autom. Constr.* **2019**, *104*, 80–94. [CrossRef]
9. Hofmann, M.; Gavrila, D.M. Multi-view 3D human pose estimation in complex environment. *Int. J. Comput. Vis.* **2012**, *96*, 103–124. [CrossRef]
10. Xu, Y.; Roy-Chowdhury, A. Inverse compositional estimation of 3d pose and lighting in dynamic scenes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1300–1307.
11. Savarese, S.; Fei-Fei, L. 3D generic object categorization, localization and pose estimation. In Proceedings of the 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007; pp. 1–8.
12. Wu, J.; Trivedi, M.M. A two-stage head pose estimation framework and evaluation. *Pattern Recognit.* **2008**, *41*, 1138–1158. [CrossRef]
13. Pessanha Santos, N.; Lobo, V.; Bernardino, A. Two-stage 3D model-based UAV pose estimation: A comparison of methods for optimization. *J. Field Robot.* **2020**, *37*, 580–605. [CrossRef]
14. Wang, Y.-K.; Cheng, K.-Y. 3D Human pose estimation by an annealed two-stage inference method. In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 535–538.
15. Zhang, Y.; Li, S.; Long, P. 3D human pose estimation in motion based on multi-stage regression. *Displays* **2021**, *69*, 102067. [CrossRef]
16. Liu, W.; Bao, Q.; Sun, Y.; Mei, T. Recent advances of monocular 2D and 3D human pose estimation: A deep learning perspective. *ACM Comput. Surv.* **2022**, *55*, 1–41. [CrossRef]
17. Chen, L.; Cao, W.; Zheng, B.; Wu, M.; Pedrycz, W.; Hirota, K. Two-Stage Representation Refinement Based on Convex Combination for 3D Human Poses Estimation. *IEEE Trans. Artif. Intell.* **2024**, *5*, 6500–6508. [CrossRef]
18. Wang, Y.-K.; Cheng, K.-Y. A two-stage Bayesian network method for 3D human pose estimation from monocular image sequences. *EURASIP J. Adv. Signal Process.* **2010**, *2010*, 761460. [CrossRef]
19. Simo-Serra, E.; Quattoni, A.; Torras, C.; Moreno-Noguer, F. A joint model for 2d and 3d pose estimation from a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3634–3641.
20. Rogez, G.; Weinzaepfel, P.; Schmid, C. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 1146–1161. [CrossRef]
21. Martinez, J.; Hossain, R.; Romero, J.; Little, J.J. A simple yet effective baseline for 3d human pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2640–2649.
22. Andriluka, M.; Roth, S.; Schiele, B. Monocular 3d pose estimation and tracking by detection. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 623–630.
23. Mehrizi, R.; Peng, X.; Xu, X.; Zhang, S.; Li, K. A Deep Neural Network-based method for estimation of 3D lifting motions. *J. Biomech.* **2019**, *84*, 87–93. [CrossRef]
24. Mehraban, S.; Qin, Y.; Taati, B. Evaluating Recent 2D Human Pose Estimators for 2D-3D Pose Lifting. In Proceedings of the 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG), Istanbul, Turkiye, 27–31 May 2024; pp. 1–5.
25. Li, W.; Liu, H.; Ding, R.; Liu, M.; Wang, P. Lifting transformer for 3d human pose estimation in video. *arXiv* **2021**, arXiv:2103.14304.

26. Grinciunaite, A.; Gudi, A.; Tasli, E.; Den Uyl, M. Human pose estimation in space and time using 3d cnn. In *European Conference on Computer Vision*; Springer: Cham, Swizerland, 2016; pp. 32–39.
27. Ghezelghieh, M.F.; Kasturi, R.; Sarkar, S. Learning camera viewpoint using CNN to improve 3D body pose estimation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 685–693.
28. Kim, Y.; Kim, D. A CNN-based 3D human pose estimation based on projection of depth and ridge data. *Pattern Recognit.* **2020**, *106*, 107462. [CrossRef]
29. Mahendran, S.; Ali, H.; Vidal, R. 3d pose regression using convolutional neural networks. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 2174–2182.
30. Ge, L.; Liang, H.; Yuan, J.; Thalmann, D. 3d convolutional neural networks for efficient and robust hand pose estimation from single depth images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1991–2000.
31. Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; Theobalt, C. Monocular 3d human pose estimation in the wild using improved cnn supervision. In Proceedings of the 2017 International Conference on 3D Vision (3DV), Qingdao, China, 10–12 October 2017; pp. 506–516.
32. 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, New Orleans, LA, USA, 18–24 June 2022; pp. 13147–13156.
33. Zhao, W.; Wang, W.; Tian, Y. Graformer: Graph-oriented transformer for 3d pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 20438–20447.
34. 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, New Orleans, LA, USA, 18–24 June 2022; pp. 13232–13242.
35. Shuai, H.; Wu, L.; Liu, Q. Adaptive multi-view and temporal fusing transformer for 3d human pose estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 4122–4135. [CrossRef]
36. Hampali, S.; Sarkar, S.D.; Rad, M.; Lepetit, V. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11090–11100.
37. 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, Montreal, BC, Canada, 11–17 October 2021; pp. 11656–11665.
38. Hassanin, M.; Khamiss, A.; Bennamoun, M.; Boussaid, F.; Radwan, I. Crossformer: Cross spatio-temporal transformer for 3d human pose estimation. *arXiv* **2022**, arXiv:2203.13387. [CrossRef]
39. Li, W.; Liu, H.; Ding, R.; Liu, M.; Wang, P.; Yang, W. Exploiting temporal contexts with strided transformer for 3d human pose estimation. *IEEE Trans. Multimed.* **2022**, *25*, 1282–1293. [CrossRef]
40. Lutz, S.; Blythman, R.; Ghosal, K.; Moynihan, M.; Simms, C.; Smolic, A. Jointformer: Single-frame lifting transformer with error prediction and refinement for 3d human pose estimation. In Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; pp. 1156–1163.
41. Wen, S.; Zhu, X.; Tan, Z. 3D WholeBody Pose Estimation based on Semantic Graph Attention Network and Distance Information. *arXiv* **2024**, arXiv:2406.01196.
42. Zhu, Y.; Samet, N.; Picard, D. H3wb: Human3. In 6m 3d wholebody dataset and benchmark. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 20166–20177.
43. Yang, Z.; Zeng, A.; Yuan, C.; Li, Y. Effective whole-body pose estimation with two-stages distillation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 4210–4220.
44. Jiang, T.; Xie, X.; Li, Y. RTMW: Real-time multi-person 2D and 3D whole-body pose estimation. *arXiv* **2024**, arXiv:2407.08634.
45. Samet, N.; Rommel, C.; Picard, D.; Valle, E. PAFUSE: Part-based Diffusion for 3D Whole-Body Pose Estimation. *arXiv* **2024**, arXiv:2407.10220.
46. Xu, Y.; Zhang, J.; Zhang, Q.; Tao, D. Vitpose++: Vision transformer for generic body pose estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *46*, 1212–1230. [CrossRef]
47. Rong, Y.; Shiratori, T.; Joo, H. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *arXiv* **2020**, arXiv:2008.08324.
48. Moon, G.; Choi, H.; Lee, K.M. Accurate 3D hand pose estimation for whole-body 3D human mesh estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 2308–2317.

49. Weinzaepfel, P.; Brégier, R.; Combaluzier, H.; Leroy, V.; Rogez, G. Dope: Distillation of part experts for whole-body 3d pose estimation in the wild. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*; Springer: Cham, Swizerland, 2020; pp. 380–397.
50. Joo, H.; Simon, T.; Sheikh, Y. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8320–8329.
51. Romero, J.; Tzionas, D.; Black, M.J. Embodied hands: Modeling and capturing hands and bodies together. *arXiv* **2022**, arXiv:2201.02610. [CrossRef]
52. Xiang, D.; Joo, H.; Sheikh, Y. Monocular total capture: Posing face, body, and hands in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10965–10974.
53. Zhang, M.; Gao, Q.; Lai, Y.; Hu, J.; Zhang, X.; Ju, Z. HR-GCN: 2D-3D Whole-body Pose Estimation with High-Resolution Graph Convolutional Network From a Monocular Camera. *IEEE Sens. J.* **2025**. [CrossRef]
54. Benedetti, L.R.; Izumi, N.; Khan, S.; Kyrala, G.; Landen, O.; Ma, T.; Nagel, S.; Pak, A. Simplified model of pinhole imaging for quantifying systematic errors in image shape. *Appl. Opt.* **2017**, *56*, 8719–8731. [CrossRef]
55. Chen, T.; Fang, C.; Shen, X.; Zhu, Y.; Chen, Z.; Luo, J. Anatomy-aware 3d human pose estimation with bone-based pose decomposition. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 198–209. [CrossRef]
56. *GB/T 17245-2004*; Inertial Parameters of Adult Human Body. National Standards of the People's Republic of China: Beijing, China, 2004.
57. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 1325–1339. [CrossRef]
58. Jin, S.; Xu, L.; Xu, J.; Wang, C.; Liu, W.; Qian, C.; Ouyang, W.; Luo, P. Whole-body human pose estimation in the wild. In *European Conference on Computer Vision*; Springer: Cham, Swizerland, 2020; pp. 196–214.
59. Dabhi, M.; Jeni, L.A.; Lucey, S. 3d-lfm: Lifting foundation model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 10466–10475.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.