symmetry

MDPI

Article

Absolute 3D Human Pose Estimation Using Noise-Aware Radial Distance Predictions

Inho Chang, Min-Gyu Park, Je Woo Kim and Ju Hong Yoon *

Korea Electronics Technology Institute, Seongnam-si 13488, Republic of Korea
* Correspondence: jhyoon@keti.re.kr

Abstract: We present a simple yet effective pipeline for absolute three-dimensional (3D) human pose estimation from two-dimensional (2D) joint keypoints, namely, the 2D-to-3D human pose lifting problem. Our method comprises two simple baseline networks, a 3D conversion function, and a correction network. The former two networks predict the root distance and the root-relative joint distance simultaneously. Given the input and predicted distances, the 3D conversion function recovers the absolute 3D pose, and the correction network reduces 3D pose noise caused by input uncertainties. Furthermore, to cope with input noise implicitly, we adopt a Siamese architecture that enforces the consistency of features between two training inputs, i.e., ground truth 2D joint keypoints and detected 2D joint keypoints. Finally, we experimentally validate the advantages of the proposed method and demonstrate its competitive performance over state-of-the-art absolute 2D-to-3D pose-lifting methods.

Keywords: absolute 3D human pose estimation; 2D-to-3D human pose lifting; distance prediction

Citation: Chang, I.; Park, M.-G.; Kim, J.W.; Yoon, J.H. Absolute 3D Human Pose Estimation Using Noise-Aware Radial Distance Predictions. *Symmetry* **2023**, *15*, 25. https://doi.org/10.3390/sym15010025

Academic Editors: Jan Awrejcewicz and Zhixun Su

Received: 1 November 2022
Revised: 7 December 2022
Accepted: 14 December 2022
Published: 22 December 2022

1. Introduction

The 3D human pose estimation is one of the most actively researched areas in computer vision, and it plays an essential role in a broad number of applications, from AR and VR experiences to motion pictures. However, the accurate absolute-scale 3D human pose is conventionally and even currently obtained using marker-based vision systems [1,2] that require various devices and sensors, including optical markers and multiple IR cameras in a controlled environment. This complicated system makes general use burdensome. In contrast, recent advances in deep learning have significantly improved the performance of marker-less 3D human pose estimation methods that typically exploit video sequences, images, two-dimensional (2D) joints [3], or noisy 3D joints [4] as inputs. Since the monocular 3D pose estimation problem inherently suffers from predicting the 3D human pose on an absolute scale, most studies relax this problem to predict root relative distance by defining the pelvis as the origin or root node. This is referred to as either a root-relative or person-centric 3D human pose estimation problem, whereas the absolute 3D pose estimation problem is referred to as a camera-centric 3D human pose problem.

Compared to the root-relative 3D human pose, the absolute 3D human pose is beneficial for real-world applications, such as surveillance systems and autonomous vehicles, where real-scale human motion and body size are crucial. For example, the groundbreaking work of [5] solves the absolute 3D human pose estimation problem by separately predicting the real-scale distance of a root joint and root-relative 3D human pose and then integrating them into a camera-centric 3D human pose. Similarly, several studies also estimated real-scale 3D poses from a single image or video sequence [6–8]. Since they do not use 2D pose information as input, we call them direct absolute 3D pose-estimation approaches. Direct 3D pose estimation has two advantages: it is possible to use a large amount of image information and an end-to-end trainable architecture. However, they require pairs of images and their corresponding 3D annotations, which are expensive and time-consuming.

In addition, a direct method is typically trained in an end-to-end manner using a single cost function. Therefore, once the target environment changes, a new dataset that contains a new image, and its corresponding 3D pose label is required.

Owing to recent impressive advances in 2D pose estimation [9–11], several studies have decomposed the 3D pose estimation problem into a 2D pose estimation and 2D-to-3D lifting problems. 2D-to-3D lifting methods [12–18] have proven their practical efficiency and generality, and show competitive performance compared to the direct method. Many 2D-to-3D lifting methods are typically built on a simple regression network [12] and are more flexible and easier to use than direct methods. For instance, different datasets can be utilized for training a 2D pose prediction network and 2D-to-3D pose lifting network. Indeed, we can easily generate pairs of labeled data by augmenting 3D joints with viewpoint changes or various poses [19] and reprojecting them onto image coordinates. To the best of our knowledge, compared to the root-relative methods, the absolute 2D-to-3D pose lifting method or root depth estimation has not been received much attention except for [18,20–22].

In contrast to 2D-to-3D human pose lifting, the direct method predicts a 3D human pose from an image without 2D joint keypoints. The direct method [23–27] is suitable for recovering 3D human poses because it extracts both contextual and pose information from an image. However, acquiring supervised data, that is, ground truth 2D and 3D human pose labels aligned with images, is a demanding process in practice and is not easy to apply to different environments universally.

Since the absolute 2D-to-3D pose-lifting method relies solely on 2D joints, the input information is very limited compared to that of direct methods. In addition, the impact of noise in 2D joints significantly degrades the pose-prediction accuracy, as demonstrated in [12]. We address these two critical issues while keeping the network as compact as possible. The contributions of our study are summarized as follows.

- The previous methods directly perceive absolute 3D joints from 2D joint keypoints as input i.e., $f : \mathbb{R}^2 \rightarrow \mathbb{R}^3$, but our method predicts both one-dimensional root distance and root-relative distance of other joints using two regression networks i.e., $f_R, f_r : \mathbb{R}^2 \rightarrow \mathbb{R}^1$. Afterwards, a simple 3D conversion function computes an initial absolute 3D pose using the input and the predicted distances. Thanks to lower-dimensional output, we can efficiently reduce the number of network parameters while keeping the performance competitive with state-of-the-art methods. This is because the radial distance prediction indirectly avoids perspective projection errors by disentangling the depth ambiguity caused by horizontal and vertical components.
- We propose to use a correction network to reduce the absolute 3D pose uncertainties, affected by input noise. When the ground truth of 2D joint keypoints are available, we adopt a Siamese architecture [28] that shares the network for noisy and clean inputs. By applying the feature consistency for different inputs, we reduce the influence of detection noise implicitly. Contrarily to our approach, the existing study [20] augments training data using synthesized errors generated from the error statistics.
- Inspired by [29–31], we also exploit relationships between joints in loss functions to regularize the problem. We design two loss functions, i.e., bone length symmetry and directional consistency between adjacent joints.

2. Related Work

We review 3D pose estimation methods that use a 2D pose or RGB image as the input. Furthermore, the methods are divided into two categories: absolute and root-relative pose estimation approaches. Most methods solve the 3D pose estimation problem using the root-relative method, and a handful of studies predict real-scale 3D joints.

2.1. Absolute 2D-to-3D Pose Lifting

The absolute-scale approach estimates a real-scale 3D human pose using a 2D pose as the input. Pavllo et al. [32] used temporal information with dilated convolutions over 2D keypoint trajectories to estimate the 3D trajectory of the root joint. Chang et al. [20]

proposed a simple cascade approach that combines a 2D pose detector and 2D-to-3D pose lifter. They normalized the input using the principal point of the camera instead of the real depth and rebuilt it to the canonical root depth. The root depth generates the final absolute depth multiplied by the focal length. Furthermore, to handle the noise of detected 2D joints, they augmented the ground truth 2D pose with synthetic errors from the error statistics of 2D pose estimation in training. The authors of [18] decomposed the absolute pose estimation problem into two sub-problems, root-relative pose estimation and root localization, with decoupling camera parameter and keypoints. Additionally, they employed temporal key-point motion information to help resolve the 3D pose ambiguity caused by occlusion. The authors of [21] proposed MonoLoco that robustly estimates absolute root depth using 2D joints as input. They further improved MonoLoco by solving the problem using a spherical coordinate system [22]. This disentangles the depth ambiguity from the horizontal and vertical components, i.e., x and y, and alleviates errors caused by perspective projections. [18] tackles absolute pose estimation by converting pixel space input to 3D normalized ray space, which makes it robust to changes in camera intrinsic parameters.

2.2. Root-Relative 2D-to-3D Pose Lifting

The root-relative approach estimates the normalized-scale 3D human pose using a 2D pose or an RGB image as the input. The authors of [12] proposed a simple baseline for 2D-to-3D pose lifting, which is the most efficient method for estimating the root-relative depth of each joint. Furthermore, it uses only two fully connected blocks while demonstrating good performance. The authors of [15] proposed using ordinal depth as additional supervision for CNN training. The authors of [17] introduced static and dynamic hyper-graphs to represent a human body for 3D pose estimation. Several studies have used human body structures to improve 3D pose accuracy, such as universal bone lengths, limitation of joint angles, and limb interpenetration constraints. The authors of [29] used pose-dependent joint angle limits for 2D-3D liter through optimization problem. The authors of [33] suggested multi-view pose augmentation from the 2D pose in the single view and estimated 3D poses using graph convolution networks [34]. The authors of [30] presented a distance matrix method that estimates a 3D Euclidean distance matrix from a 2D Euclidean distance matrix using a simple regression network. The authors of [31] applied re-parameterized pose representation, which uses the joint connection structure. The authors of [14] enforced high-level constraints over pose using the human body grammar model. The authors of [35] encode relative positional and temporal enhanced representations, and this approach achieves excellent root-relative 3D pose accuracy. The authors of [36] adopt a pure Transformer to capture human joint correlations and temporal dependencies.

Direct methods directly estimate the 3D pose from an RGB image input. Various representations of these outputs exist, such as 3D coordinates [37], volumetric heatmaps [24], and bone-based representations [31]. Moreover, several studies leveraged temporal information to further improve and smooth the 3D pose from continuous video frames [38,39]. However, both direct and sequential methods, constructing images, and the corresponding 3D pose labels or sequential label data are impractical.

3. Methodology

Inspired by the simple baseline method [12], we propose a simple yet effective absolute 2D-to-3D pose lifting pipeline that effectively exploits the limited information of 2D poses. To keep the network size compact and alleviate perspective projection errors, our method predicts the absolute distance of joints rather than directly estimating the absolute 3D pose from 2D joint keypoints. This section introduces the proposed 2D-to-3D pose-lifting method and explains the loss functions in detail.

3.1. Problem Formulation

We formulate the absolute 2D-to-3D human pose-lifting problem as

$$\hat{\mathbf{P}} = f(\mathbf{p}, \theta), \tag{1}$$

where $f : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ is a lifting function, θ indicates trainable parameters, and $\mathbf{p} = [p_1, ..., p_N]$ denotes the detected 2D joint keypoints as input with the number of keypoints N. The lifting function predicts the 3D pose, $\hat{\mathbf{P}}$. To train θ, we minimize the following function:

$$\min_{\theta} \mathcal{L}(f(\mathbf{p}, \theta), \mathbf{P}), \tag{2}$$

which minimizes the discrepancy between the predicted 3D joints $\hat{\mathbf{P}}$ and the ground truth 3D joints $\mathbf{P}$. A detailed description is provided in Section 3.4.

3.2. Absolute 3D Pose from Distances

Inspired by previous studies [18,21,22], we propose an absolute 3D human pose estimation with a calibrated camera. Therefore, we normalize the input using the camera intrinsic matrix $\mathbf{K}$ by

$$\bar{\mathbf{p}} = \mathbf{K}^{-1}[\mathbf{p}^\top, \mathbf{1}_N^\top]^\top, \tag{3}$$

where a vector of the 2D keypoints and its normalization are denoted by

$$\mathbf{p} = [p_1, ..., p_N]^\top, \;\; \bar{\mathbf{p}} = [\bar{p}_1, ..., \bar{p}_N]^\top, \tag{4}$$

where the ith 2D keypoint is represented by $p_i = [x_i, y_i]^\top$ and $\bar{p}_i = [\bar{x}_i, \bar{y}_i]^\top$, respectively. The normalization procedure prevents the overfitting of a specific camera [21].

Other absolute 2D-to-3D human pose-lifting methods directly predict 3D human poses based on a function $f : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ [18,20,21]. By contrast, our method predicts a one-dimensional radial distance for each joint from a normalized 2D keypoint $f : \mathbb{R}^2 \rightarrow \mathbb{R}^1$. This effectively reduces the number of network parameters. The radial distance of 3D joints is defined by

$$\boldsymbol{\text{æ}} = [\rho_1, ..., \rho_N]^\top, \;\; \rho_i = \sqrt{X_i^2 + Y_i^2 + Z_i^2}, \tag{5}$$

where N is the number of joints, and i-th radial distance ρ_i is computed by $\mathbf{P}$ of i-th joint, X_i, Y_i, and Z_i. With the radial distances, we can recover the 3D human pose using the normalized 2D keypoints. We can formulate the absoulte 3D human pose $P_i \in \mathbf{P}$ with the normalized keypoints $\bar{p}_i \in \bar{\mathbf{p}}$ and predicted radial distance $\hat{\rho}_i \in \hat{\rho}$ by

$$P_i = \psi_t(\bar{p}_i, \, \rho_i) = Z_i[\bar{x}_i, \bar{y}_i, 1]^\top, \;\; Z_i = \frac{\rho_i}{\sqrt{\bar{x}_i^2 + \bar{y}_i^2 + 1}}, \tag{6}$$

Since the normalized 2D keypoints $\bar{\mathbf{p}}$ are given, we only need to estimate the radial distance to get the absolute 3D pose.

3.3. Proposed Pipeline

The proposed framework is built on the simple baseline [12] as the backbone. Similar to the work of Moon et al. [5], we adopt two backbone networks in our framework, as shown in Figure 1. They predict the root-relative distance, i.e., relative distances, of joints, and the absolute root distance, i.e., root distances, respectively. Subsequently, the correction network further minimizes the predicted 3D pose errors caused by the noise of the detected 2D joint keypoints. We recover the absolute 3D pose $\hat{\mathbf{P}}$ using

$$\hat{\mathbf{P}} = \bar{\mathbf{P}} + f_c(\bar{\mathbf{P}}), \tag{7}$$

where $\bar{\mathbf{P}}$ is the initial absolute 3D pose with N joints and the correction network $f_c : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ predicts a residual to refine the initial 3D pose. We compute the initial absolute 3D pose $\bar{\mathbf{P}}$ from a predicted radial distance vector ρ and normalized keypoints $\bar{\mathbf{p}}$ using (6).

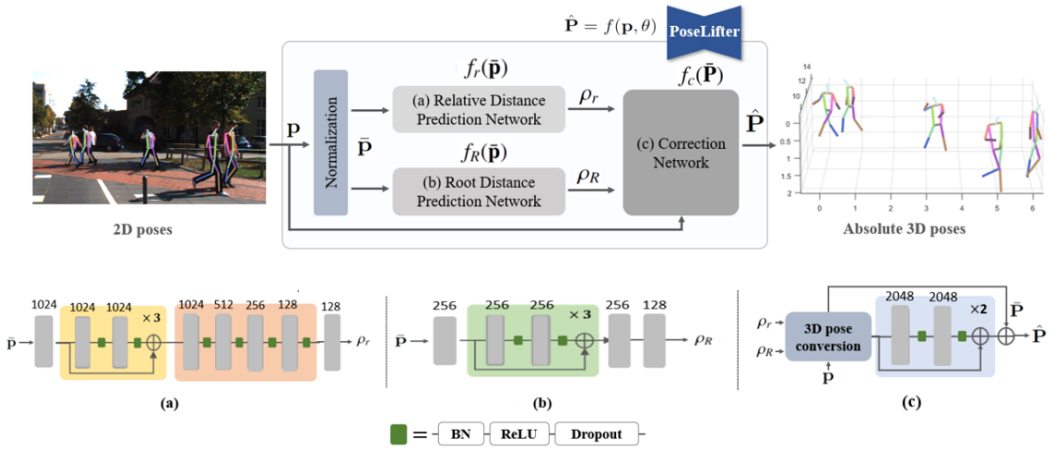

Figure 1. Proposed framework $\hat{\mathbf{P}} = f(\mathbf{p}, \theta)$. We normalize the input 2D pose to prevent the method from over-fitting to a specific camera. We adopt the simple baseline network [12] for the root distance prediction $f_R(\bar{\mathbf{p}})$, the relative distance prediction $f_r(\bar{\mathbf{p}})$ in (9), and the correction network $f_c(\bar{\mathbf{P}})$ in (7). The 3D pose conversion in (6) computes an absolute 3D pose $\bar{\mathbf{P}}$ with the predicted distances and corresponding 2D joints. The correction network finally predicts 3D pose residuals to reduce errors caused by noise of the 2D joint keypoints. The bottom three blocks represent (**a**) Relative Distance Prediction Network, (**b**) Root Distance Prediction Network, and (**c**) Correction Network.

To predict the radial distance vector, we first decompose the distances into root and relative distances, as shown in Figure 1 as follows:

$$\rho = [\rho_1, ..., \rho_N]^\top = \mathbf{1}_N \rho_R + \rho_r, \tag{8}$$

where we use a vector of $\mathbf{1}_N \in \mathbb{R}^N$ to represent the radial distances of joints as a vector, which is decomposed into a root distance $\rho_R \in \mathbb{R}^1$ and a relative distance vector $\rho_r \in \mathbb{R}^N$. These are estimated using normalized keypoints as inputs by

$$\rho_R = f_R(\bar{\mathbf{p}}), \;\; \rho_r = f_r(\bar{\mathbf{p}}). \tag{9}$$

where two regression networks, i.e., f_R and f_r, predict the root distance ρ_R and relative distance vector ρ_r, respectively. Then, we obtain a vector of absolute radial distance based on (8) and finally recover the absolute 3D human pose $\hat{\mathbf{P}}$ via (7).

3.4. Training via Siamese Architecture

To deal with noisy 2D joints, we adopted a Siamese architecture to train the functions $f_R(\cdot)$, $f_r(\cdot)$, and $f_c(\cdot)$ in Section 3.3 when labeled 2D poses are available, as shown in Figure 2.

We design a loss function that combines the pose errors $\mathcal{L}_\mathbf{P}$, distance errors $\mathcal{L}_\rho$, joint relation constraints $\mathcal{L}_C$, and feature consistency $\mathcal{L}_\mathcal{F}$,

$$\mathcal{L} = w_\rho \mathcal{L}_\rho + w_P \mathcal{L}_\mathbf{P} + w_C \mathcal{L}_C + w_F \mathcal{L}_\mathcal{F}, \tag{10}$$

where the coefficient w indicates the control parameters. Distance loss measures the errors between the predicted joint distances $\hat{\rho}$ and label distances ρ.

$$\mathcal{L}_\rho = \|\hat{\rho} - \rho\|_2. \tag{11}$$

The pose loss measures errors between the predicted 3D poses $\hat{\mathbf{P}}$ and the label 3D poses $\mathbf{P}$.

$$\mathcal{L}_{\mathbf{P}} = \|\hat{\mathbf{P}} - \mathbf{P}\|_2. \tag{12}$$

Here, the distance and pose losses are defined in the Euclidean coordinates.

We formulate the constraint with bone length symmetry and the directional constraint of the adjacent joints as follows:

$$\mathcal{L}_C = \sum_B \|b_i - b_j\|_2 + \sum_{i=1}^{8} \|\hat{\mathbf{d}}_i - \mathbf{d}_i\|_2, \tag{13}$$

where we denote a set of joint bone length indices as $(i, j) \in B$ and a set of directional vectors of joints as $\mathbf{d}_i \in \mathcal{D}$ for the ground truth and $\hat{\mathbf{d}}_i \in \hat{\mathcal{D}}$ for the prediction. The details of their representation are shown in Figure 3.

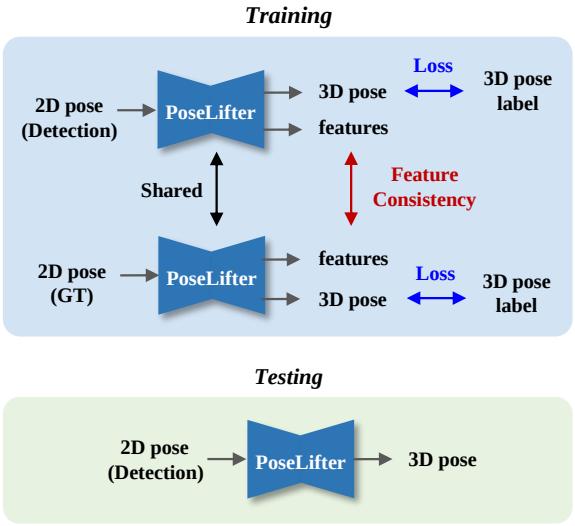

Figure 2. Siamese architecture for noise awareness. We adopt a Siamese architecture for training the proposed pose lifting method when ground truth 2D poses, i.e., 2D joint keypoints, are available, and in this case, we can utilize the feature consistency loss. If the ground truth 2D pose is unavailable, we only use the upper part in this figure for the training without the feature consistency loss. For testing, we use detected 2D poses as input.

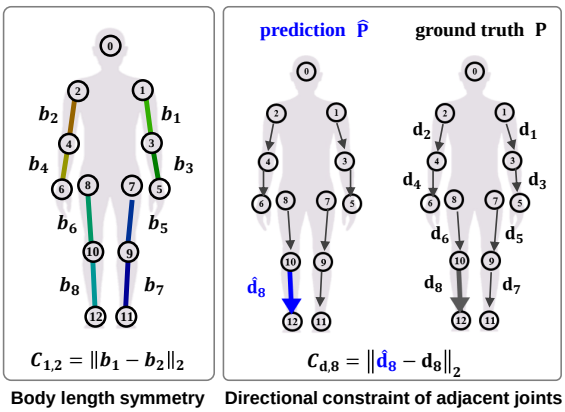

Figure 3. Examples of body length symmetry and directional constraint. Body length symmetry is self loss term which minimizing difference between left and right bone length based on human body symmetric characteristic. Left arm and leg, right arm and leg are used. Directional constraint limits bone vector between prediction and ground-truth. All body parts are used.

To understand the noise implicitly, we introduce the feature consistency formulated by

$$\mathcal{L}_{\mathcal{F}} = \|\mathcal{F}_A^d - \mathcal{F}_A^g\|_2 + \|\mathcal{F}_R^d - \mathcal{F}_R^g\|_2 + \|\mathcal{F}_O^d - \mathcal{F}_O^g\|_2, \tag{14}$$

where the superscript d denotes the features generated by the detected inputs, and g denotes the features generated by the labeled inputs. We illustrate the extraction of these features from the networks in Figure 1. Feature $\mathcal{F}_O$ is the output of the layer before the last layer of the correction network. Similarly, $\mathcal{F}_A$ and $\mathcal{F}_R$ are the outputs of backbone networks. The feature consistency loss forces the features of the detected noisy 2D inputs to become similar to those of the labeled inputs. Therefore, we can achieve a similar accuracy for 3D poses from noisy inputs over 3D poses from labeled inputs.

4. Experimental Results

In this section, we evaluate the proposed method using different experimental setups and compare its performance with that of state-of-the-art (SOTA) methods. In particular, we analyze the advantages of the proposed architecture in detail over previous similar works that also use a simple baseline network as a backbone, and show the effectiveness of noise awareness via a Siamese architecture. Finally, the ablation study highlights the merits of feature consistency, relational constraints, and the correction network for absolute 2D-to-3D human pose lifting.

4.1. Implementation

We run the training procedure for 50 epochs using the Adam optimizer [40] and set the learning rate and batch size to 0.001 and 2048, respectively. ReLU is selected as the nonlinear activation function. The dropout rate is set to 0.5, except for the linear blocks shown in Figure 1c, which is set as 0.2. The weights of the linear layers are initialized by Kaiming normal initialization [41]. The loss weights in Section 3.4 are empirically selected, where w_ρ is 1, w_P is 1, w_F is 0.01, and w_C is 0.1. For the Human3.6M dataset, we train the network with GT and 2D detection inputs from a Cascaded Pyramid Network (CPN) [9]. For KITTI dataset [42] for we use 2D detected inputs from OpenPifPaf [10].

4.2. Datasets

We evaluate our method qualitatively and quantitatively on four publicly available 3D human pose datasets: Human3.6M [43], MPI-INF-3DHP [23], MuCo-3DHP, and MuPoTs-3D [44]. Furthermore, we evaluate the root depth estimation using the KITTI dataset for comparison with a recent study by [22]. The Human3.6M is one of the largest human 3D pose estimation datasets. It consists of 3.6 million human poses and ground truth annotations of 2D and 3D poses and camera parameters. According to [12], we use five subjects, i.e., 1, 5, 6, 7, and 8, for training and two subjects, i.e., 9 and 11, for testing in the Human3.6M dataset. The MPI-INF-3DHP (3DHP) dataset is a recently released large-scale 3D pose dataset that contains 1.3 million images with various actions. The 3DHP dataset consists of six test subjects with different indoor environments and two subjects with in-the-wild settings. We use all samples in the test set that are captured in indoor scenes with and without a green background to validate the accuracy of the proposed method. The MuCo-3DHP dataset is an indoor multiperson dataset for training that is composed of a single-person 3DHP dataset. MoPoTS-3D is a synthetic multiperson 3D pose dataset for evaluation. It consists of 8000 frames covering five indoor and 15 outdoor settings. The ground-truth 3D poses are captured using a multi-view markerless motion capture system. The KITTI dataset contains 7481 training images along with camera calibration files, and 5000 instances provided by [22] are used for training and testing.

4.3. Evaluation Metrics

We adopt two evaluation metrics for quantitative evaluations: the Absolute Mean Per Joint Position Error (Abs-MPJPE) in millimeters and Mean Root Position Error (MRPE) [18]. For ablation study, we use Mean Per Joint Position Error (MPJPE) and Procrustes Aligned

MPJPE (P-MPJPE), which computes MPJPE after rigid-body alignment of the estimated pose and the ground truth pose. For MPI-3DHP and MuPoTS-3D, we apply the percentage of correct keypoints (PCK) with a threshold of 150 mm and the area under curve (AUC) for a range of PCK thresholds, as provided by [23]. For the KITTI dataset, we also adopt the absolute average localization precision (ALP) proposed in [21]. Similar to PCK, ALP represents the percentage of correctness by computing the ratio of distance errors below a certain threshold. For Abs-MPJPE, MRPE, MPJPE, and P-MPJPE, a lower value is better, and vice versa for PCK, AUC, and ALP. Abs-MPJPE and MRPE are absolute distance evaluation metrics, whereas the others are root-relative distance evaluation metrics.

4.4. Ablation Study

We evaluated the proposed method with various configurations and summarized the results in Table 1 to analyze the impact of different losses, network architecture, and input noise. In Table 1, the latter "B" represents the proposed method without the correction network trained with only the distance and pose losses in (11) and (12). Relational constraints improve root-relative accuracy. Feature consistency improves the performance of root distance accuracy more than the relational constraints because it implicitly reduces the detection noise with a Siamese architecture. When we use feature consistency and relational constraints together, both the root-relative and root distances are enhanced. Moreover, if we also use the correction network, we can improve MRPE by 4.0 mm, MPJPE by 2.1 mm, and P-MPJPE by 2.1 mm compare to without correction network. We can observe that the correction network accurately estimates the noise of the predicted 3D pose according to the results of the full configurations.

Table 1. An ablation study: "B" the proposed method without the correction network and other constraints. "RC" relational constraints. "F" feature consistency. "C" correction network. Here, we use the detected 2D joint keypoints as input, obtained by CPN.

Conf.	Input	MRPE	MPJPE	P-MPJPE
B	CPN	109.4	64.4	48.9
+RC	CPN	109.3	59.3	45.7
+F	CPN	107.3	60.6	46.3
+RC+F	CPN	107.2	58.8	45.2
+RC+F+C	CPN	103.2	56.7	43.1

4.5. Evaluation on Different Parameter Sizes

Similar to our method, [20] used the simple baseline method [12] as the backbone. The difference is that they directly estimate the absolute 3D pose $\hat{\mathbf{P}}$ based on the function $f : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ from the detected 2D joints. However, our method separately estimates the root distance ρ_R and root-relative joint distances ρ_r using two networks $f_R : \mathbb{R}^2 \rightarrow \mathbb{R}^1$ and $f_r : \mathbb{R}^2 \rightarrow \mathbb{R}^1$, respectively, which significantly reduces the number of parameters and the output solution space. The number of parameters in [20] is approximately 67×10^6. By contrast, that of our best full model is approximately 13.75×10^6 as shown in Table 2. Furthermore, our method is more accurate in terms of MRPE. This implies that our noise-aware distance prediction accurately predicts the absolute 3D pose while maintaining network size efficiency. Specifically, we adopt a backbone network of root distance prediction from [21] and relative distance prediction from [12] with the same number of parameters, which is approximately 5.16×10^6. In the experiments, we test different correction networks by changing the size of the linear layers by 512, 1024, and 2048, and obtain the best accuracy with a size of 2048.

Table 2. MRPE along different number of network parameters on Human3.6M. "Root" root distance prediction network. "Relative" relative distance prediction network. "Correct" correction network. "Full" full model. The best performance is highlighted in **Bold**. PoseLifter [20], PoseFormer [36], and Ray3D [18]. All MRPE results of other methods are obtained from [18].

Method	MRPE	# of Parameters (M = $\times 10^6$)			
		Root	Relative	Correct	Full
PoseLifter	135.1	-	-	-	67.0M
PoseFormer	127.7	-	-	-	18.2M
Ray3D	105.0	-	-	-	45.8M
Ours w/o C	107.2	0.24M	4.92M	-	5.2M
Ours (L = 512)	105.8	0.24M	4.92M	0.57M	5.8M
Ours (L = 1024)	104.5	0.24M	4.92M	2.20M	7.4M
Ours (L = 2048)	**103.2**	0.24M	4.92M	8.59M	13.8M

4.6. Quantitative and Qualitative Results

We explain the quantitative evaluation results of the proposed method using the various evaluation metrics mentioned in Section 4.3. In the experiments, we use the detected 2D joints as input, obtained by the CPN detector [9] as in other studies [18,21]. Furthermore, to demonstrate our absolute 2D-to-3D human pose-lifting performance more intuitively, we show some qualitative results in Figures 4 and 5.

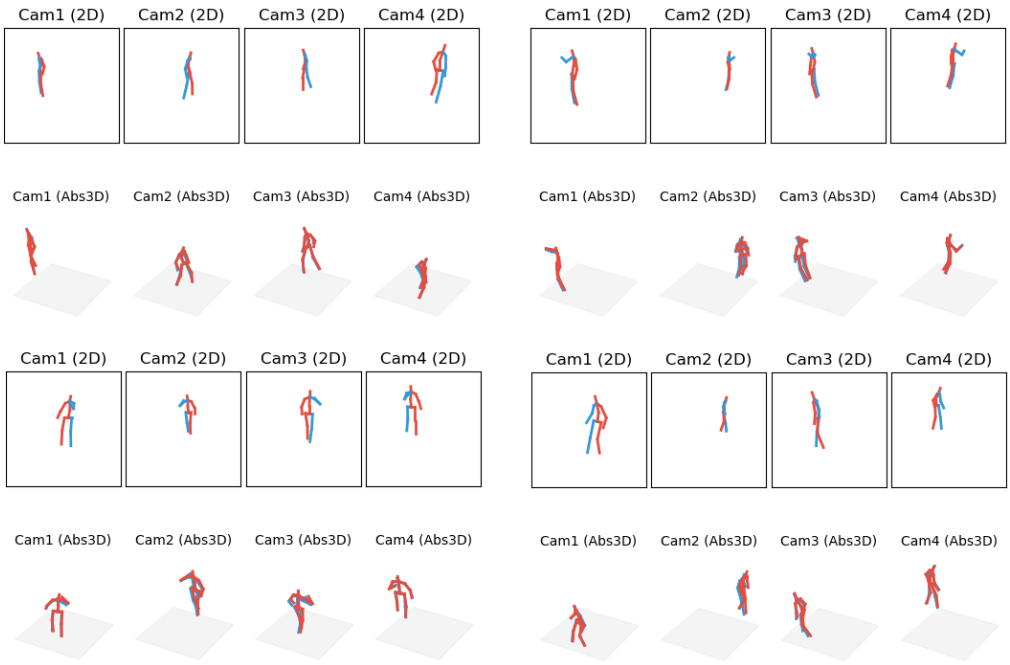

Figure 4. Absolute 2D-to-3D animated example output on the Human3.6M test set. Cam(2D): 2D CPN input from each camera-view. Cam(Abs3D): Absolute 3D lifting output, red and blue denote ground truth and prediction results.

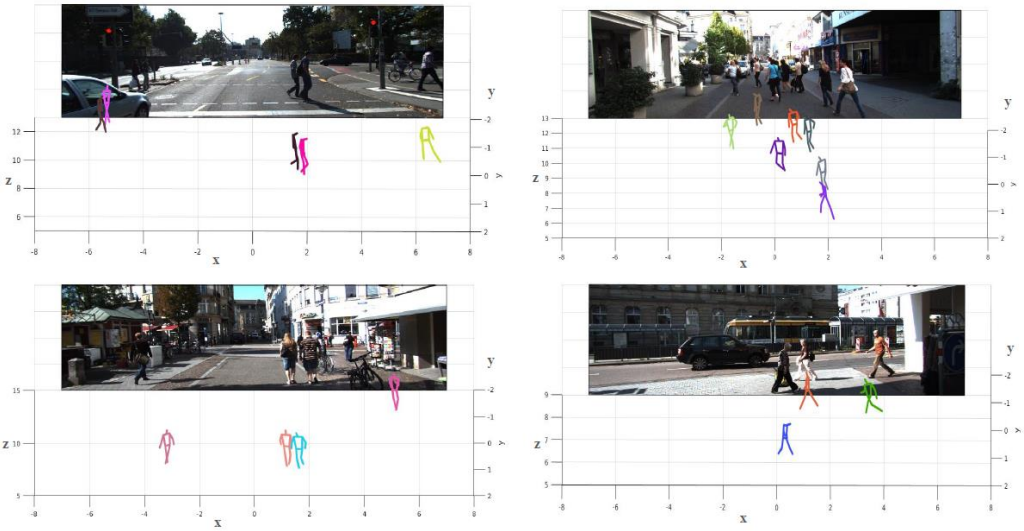

Figure 5. Results of absolute 2D-to-3D human pose lifting on KITTI. The values in the graphs are in meters.

4.6.1. Human3.6M

Table 3 shows the Abs-MPJPE and MRPE results for Human3.6M and Figure 4 shows the absolute 2D-to-3D lifting visual results from each camera. We use the detected 2D joints obtained from the CPN detector as input and separate the results based on the number of input frames i.e., $f = 1$ and $f = 9$. The results of other methods are obtained from [18]. We observe that our method achieves competitive performance over SOTA methods. Specifically, Abs-MPJPE and MRPE surpass [20] by 28.8 mm and 31.9 mm, respectively. Furthermore, compared with the method proposed in a recent study [18], our method has a 2.5 mm gap in Abs-MPJPE and 2.8 mm better performance in MRPE with a single frame case. Our method can achieve competitive accuracy because it estimates the root joint distance and root-relative joint distances with a more reduced solution space than the solution space that directly 2D-to-3D human pose lifting methods solve. In addition, the correction network effectively reduces the uncertainties of the initial absolute 3D pose caused by the noise of 2D joint keypoints in the 3D conversion, as shown in Table 1.

Table 3. Quantitative evaluation results under Abs-MPJPE and MRPE on Human3.6M: the best (red) and the second best (blue).

Abs-MPJPE	**Direct**	**Discuss**	**Eating**	**Greet**	**Phone**	**Photo**	**Pose**	**Purch.**	**Sit**	**SitD.**	**Smoke**	**Wait**	**Walk**	**WalkD.**	**WalkT.**	**Avg**
[35] (f = 9)	143.2	133.2	143.9	142.7	110.9	151.4	125.9	98.4	136.4	273.4	127.5	138.9	126.8	107.3	116.0	138.4
[36] (f = 9)	112.6	137.1	117.6	145.8	113.0	166.0	125.5	113.8	128.8	245.7	122.7	144.8	125.0	118.9	129.3	136.5
[32] (f = 9)	128.9	125.4	124.4	138.2	108.2	155.5	116.6	101.1	135.8	287.6	128.6	130.9	122.1	101.6	110.7	134.4
[18] (f = 9)	92.9	97.4	139.8	118.6	113.8	105.9	84.5	74.9	148.6	165.7	116.6	113.9	98.2	83.6	87.9	109.5
[20] (f = 1)	140.9	113.2	139.9	148.2	122.0	155.3	121.5	121.1	170.0	267.6	139.2	142.9	146.4	132.1	135.2	146.4
[18] (f = 1)	80.1	100.8	123.8	125.5	110.7	111.8	96.1	99.3	129.4	176.3	106.8	129.2	120.4	109.1	106.6	115.1
Ours (f = 1)	105.4	114.6	99.1	123.1	100.9	145.0	106.8	96.7	115.5	193.7	113.9	121.2	104.1	118.1	106.2	117.6
MRPE	Direct	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	Walk	WalkD.	WalkT.	Avg
[35] (f = 9)	139.1	124.5	129.9	133.1	99.2	141.4	116.3	93.5	124.0	265.9	118.4	131.3	117.1	100.4	109.2	129.6
[36] (f = 9)	104.7	134.7	103.9	137.4	99.6	154.6	119.8	108.9	108.2	233.7	111.1	141.1	116.2	117.9	123.8	127.7
[32] (f = 9)	124.2	115.9	111.0	127.3	97.6	141.9	105.7	96.4	122.0	276.5	119.6	123.3	111.3	94.0	101.6	124.6
[18] (f = 9)	83.7	86.8	128.9	104.8	109.3	91.6	75.0	65.2	143.9	150.5	108.6	105.7	88.4	73.9	77.8	99.6
[20] (f = 1)	134.7	102.3	126.9	135.7	109.9	138.5	110.7	110.9	170.0	252.4	128.4	133.9	139.4	121.6	124.4	135.1
[18] (f = 1)	67.3	91.7	113.6	111.8	104.5	96.3	85.8	94.6	124.4	161.7	97.6	119.5	110.9	100.9	94.8	105.0
Ours (f = 1)	86.7	100.1	82.9	107.3	91.2	126.6	89.1	91.4	97.6	172.0	101.3	109.2	93.9	103.9	94.0	103.2

4.6.2. MPI-3DHP

Table 4 describes the PCK, AUC, and MRPE performance on the 3DHP dataset. Our method and [20] estimate the absolute-scale 3D pose, whereas other methods predict the root-relative 3D pose. Similar to the results for Human3.6M, our method shows competitive accuracy compared to the SOTA methods and achieves the best performance in cross-validation evaluation, for which we used Human3.6M for training and MPI for testing. This result indicates that our method is suitable for various applications as long as 2D joint keypoints are given as input. Generality is a distinct benefit of keypoint-based 3D pose prediction, which is less sensitive to environmental changes and camera specifications than image-based 3D pose prediction.

Table 4. Quantitative comparison on MPI-3DHP.

Method	Training	PCK_{rel}	AUC_{rel}	MRPE
Yang [45]	H36M+MPII	69.0	32.0	-
Zhou [46]	H36M+MPII	69.2	32.5	-
Martinez [12]	H36M	42.5	17.0	-
Mehta [23]	H36M	64.7	31.7	-
Luo [47]	H36M	65.6	33.2	-
Habibie [27]	H36M	70.4	36.0	-
Ci [48]	H36M	74.0	36.7	-
Liu [17]	H36M	74.9	37.5	-
Chang [20]	H36M	76.5	40.2	421.3
Ours	H36M	**77.0**	**43.2**	**280.3**
Ours	MPI	92.3	61.0	192.7

4.6.3. KITTI

Table 5 summarizes the performance of the camera-to-root 3D localization results on the KITTI dataset and Figure 5 shows the visualized results of the absolute 2D-to-3D human pose lifting. For fair comparisons, we use 2D joint keypoints as inputs, as provided by [22]. We use the detected 2D poses provided by [22] and calculate ALP on three distance thresholds, <0.5 m, <1 m and <2 m, as in [22]. MonoLoco estimates the 3D location in Cartesian coordinates, whereas MonoLoco++ predicts the 3D location in spherical coordinates. Our method predicts the camera-to-root distance, which is a scalar value without the azimuthal and polar angles. Although we do not consider the angle values in the prediction, our approach is more accurate than that of MonoLoco. Moreover, as proved in [22], radial distance prediction indirectly disentangles the noise of the 2D joint position. Similarly, MonoLoco++ shows a better performance because it effectively utilizes the epistemic and aleatoric uncertainties in the spherical coordinate, especially for the case of <0.5 m. However, because our network aims to predict a simple scalar value, i.e., the radial distance, it shows a more robust performance according to ALP on <1 m and <2 m.

Table 5. Quantitative comparison on KITTI. The best results are highlighted with **Bold**.

Method	Training	ALP [%] ↑		
		<0.5 m	**<1 m**	**<2 m**
MonoLoco [21] (ICCV'19)	KITTI	25.3	43.4	60.5
MonoLoco++ [22] (TITS'21)	KITTI	**37.4**	53.2	63.6
Ours	KITTI	31.3	**54.7**	**79.6**

4.7. Limitations

One potential drawback of our approach is that the ground truth considers nonoccluded situations. Therefore, occluded joints can degrade the accuracy of 3D pose prediction. Although our approach can alleviate depth ambiguities compared to existing monocular methods, it shows poorer performance than multi-view or temporal approaches. Our network shows weakness in the case of side and rear views or complex postures, such as sitting and sitting-down. Nevertheless, our approach shows better performance than the existing monocular-based approaches. The human pose dataset consists mainly of a front view, which familiarizes the network with the front position. For this reason, it is more difficult for the network to lift the rear of the view than the front. Furthermore, the sitting behavior involves only a small fraction of the dataset, making it difficult to determine the optimal output. Figure 6 shows the failures of our network. In addition, the singular case exists when a single 2D pose is given as input, for example, all joints are at the same distance, which is a common limitation for all 2D joint-based approaches.

Figure 6. Inaccurate cases on the Human3.6M test set. In the top row, the second row shows the sitting position action and the third row, the last row, shows the results of the photo action.

5. Conclusions

We have proposed a simple and effective absolute 2D-to-3D human pose-lifting framework consisting of two baseline networks followed by a pose-correction network. The two baseline networks predicted both the absolute root distance and root-relative distance simultaneously, and the correction network refined the predicted results. In addition, our framework has a Siamese architecture that imposes feature consistency between the predicted features, which improves the robustness of our framework under noisy input. We experimentally demonstrated the advantages of the proposed method with state-of-the-art absolute 2D-to-3D pose-lifting methods on public benchmark datasets. Our method achieved competitive results with fewer trainable parameters, verifying its effectiveness.

Author Contributions: Conceptualization, I.C., M.-G.P. and J.H.Y.; Methodology, M.-G.P. and J.H.Y.; Formal analysis, M.-G.P.; Resources, J.W.K.; Writing—original draft, I.C. and J.H.Y.; Writing—review & editing, I.C., M.-G.P. and J.H.Y.; Visualization, I.C.; Supervision, J.H.Y.; Project administration, J.W.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Human3.6M: http://vision.imar.ro/human3.6m/description.php, MPI-INF-3DHP: https://vcai.mpi-inf.mpg.de/3dhp-dataset/ (accessed on 31 October 2022), MuCo-3DHP and MuPoTs-3D https://vcai.mpi-inf.mpg.de/projects/SingleShotMultiPerson/ (accessed on 31 October 2022), and KITTI: https://www.cvlibs.net/datasets/kitti/ (accessed on 31 October 2022).

Acknowledgments: This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2021-0-00193, Development of photorealistic digital human creation and 30fps realistic rendering technology).

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

References

1. Optitrack. Available online: https://www.optitrack.com/ (accessed on 1 January 2022).
2. Qualisys. Available online: https://www.qualisys.com/ (accessed on 1 January 2022).
3. Liu, W.; Bao, Q.; Sun, Y.; Mei, T. Recent Advances in Monocular 2D and 3D Human Pose Estimation: A Deep Learning Perspective. *arXiv* **2021**, arXiv:2104.11536
4. Wu, Y.; Ma, S.; Zhang, D.; Sun, J. 3D Capsule Hand Pose Estimation Network Based on Structural Relationship Information. *Symmetry* **2020**, *12*, 1636. [CrossRef]
5. Moon, G.; Chang, J.Y.; Lee, K.M. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In Proceedings of the ICCV, Seoul, Republic of Korea, 27 October–2 November 2019.
6. Lin, J.; Lee, G.H. HDNet: Human depth estimation for multi-person camera-space localization. In Proceedings of the ECCV, Seattle, WA, USA, 13–19 June 2020.
7. 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, Palo Alto, CA, USA, 22–24 March 2021.
8. Cheng, Y.; Wang, B.; Yang, B.T.; Tan, R. Monocular 3D Multi-Person Pose Estimation by Integrating Top-Down and Bottom-Up Networks. In Proceedings of the CVPR, Nashville, TN, USA, 20–25 June 2021.
9. Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; Sun, J. Cascaded pyramid network for multi-person pose estimation. In Proceedings of the CVPR, Salt Lake City, UT, USA, 18–23 June 2018.
10. Kreiss, S.; Bertoni, L.; Alahi, A. PifPaf: Composite fields for human pose estimation. In Proceedings of the CVPR, Long Beach, CA, USA, 15–20 June 2019.
11. Cao, Z.; Hidalgo Martinez, G.; Simon, T.; Wei, S.; Sheikh, Y.A. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 172–186. [CrossRef]
12. Martinez, J.; Hossain, R.; Romero, J.; Little, J.J. A Simple yet effective baseline for 3d human pose estimation. In Proceedings of the ICCV, Venice, Italy, 22–29 October 2017.
13. Tome, D.; Russell, C.; Agapito, L. Lifting from the deep: Convolutional 3d pose estimation from a single image. In Proceedings of the CVPR, Honolulu, HI, USA, 21–26 July 2017.
14. Fang, H.S.; Xu, Y.; Wang, W.; Liu, X.; Zhu, S.C. Learning pose grammar to encode human body configuration for 3d pose estimation. In Proceedings of the AAAI, New Orleans, LA, USA, 2–7 February 2018.
15. Wang, M.; Chen, X.; Liu, W.; Qian, C.; Lin, L.; Ma, L. DRPose3D: Depth ranking in 3d human pose estimation. In Proceedings of the IJCAI, Stockholm, Sweden, 13–19 July 2018.
16. Li, Y.; Li, K.; Jiang, S.; Zhang, Z.; Huang, C.; Da X., R.Y. Geometry-driven self-supervised method for 3D human pose estimation. In Proceedings of the AAAI, New York, NY, USA, 7–12 February 2020.
17. Liu, S.; Lv, P.; Zhang, Y.; Fu, J.; Cheng, J.; Li, W.; Zhou, B.; Xu, M. Semi-dynamic hypergraph neural network for 3d pose estimation. In Proceedings of the IJCAI, Yokohama, Japan, 7–15 January 2020.
18. Zhan, Y.; Li, F.; Weng, R.; Choi, W. Ray3D: Ray-based 3D human pose estimation for monocular absolute 3D localization. In Proceedings of the CVPR, New Orleans, LA, USA, 19–20 June 2022.
19. Li, S.; Ke, L.; Pratama, K.; Tai, Y.; Tang, C.; Cheng, K.T. Cascaded deep monocular 3d human pose estimation with evolutionary training data. In Proceedings of the CVPR, Seattle, WA, USA, 16–18 June 2020.
20. Chang, J.Y.; Moon, G.; Lee, K.M. PoseLifter: Absolute 3d human pose lifting network from a single noisy 2D human pose. *arXiv* **2019**. arXiv:1910.12029
21. Bertoni, L.; Kreiss, S.; Alahi, A. MonoLoco: Monocular 3d pedestrian localization and uncertainty estimation. In Proceedings of the ICCV, Seoul, Republic of Korea, 27 October–2 November 2019.
22. Bertoni, L.; Kreiss, S.; Alahi, A. Perceiving Humans: From Monocular 3D Localization to Social Distancing. *IEEE Trans. Intell. Trans. Sys.* **2021**, *23*, 7401–7418.. [CrossRef]
23. 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 3DV, Qingdao, China, 10–12 October 2017.
24. Pavlakos, G.; Zhou, X.; Derpanis, K.G.; Daniilidis, K. Coarse-to-fine volumetric prediction for single-image 3D human pose. In Proceedings of the CVPR, Honolulu, HI, USA, 21–26 July 2017.

25. Tekin, B.; Rozantsev, A.; Lepetit, V.; Fua, P. Direct prediction of 3d body poses from motion compensated sequences. In Proceedings of the CVPR, Las Vegas, NV, USA, 27–30 June 2016.
26. Sun, X.; Xiao, B.; Wei, F.; Liang, S.; Wei, Y. Integral human pose regression. In Proceedings of the ECCV, Munich, Germany, 8–14 September 2018.
27. Habibie, I.; Xu, W.; Mehta, D.; Pons-Moll, G.; Theobalt, C. In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In Proceedings of the CVPR, Long Beach, CA, USA, 15–20 June 2019.
28. Chopra, S.; Hadsell, R.; LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. In Proceedings of the CVPR, Washington, DC, USA, 20–26 June 2005.
29. Akhter, I.; Black, M.J. Pose-conditioned joint angle limits for 3D human pose reconstruction. In Proceedings of the CVPR, Boston, MA, USA, 7–12 June 2015.
30. Moreno-Noguer, F. 3d human pose estimation from a single image via distance matrix regression. In Proceedings of the CVPR, Honolulu, HI, USA, 21–26 July 2017.
31. Sun, X.; Shang, J.; Liang, S.; Wei, Y. Compositional human pose regression. In Proceedings of the ICCV, Venice, Italy, 22–29 October 2017.
32. Pavllo, D.; Feichtenhofer, C.; Grangier, D.; Auli, M. 3D Human Pose Estimation in Video With Temporal Convolutions and Semi-Supervised Training. In Proceedings of the CVPR, Long Beach, CA, USA, 15–20 June 2019.
33. Sun, J.; Wang, M.; Zhao, X.; Zhang, D. Multi-View Pose Generator Based on Deep Learning for Monocular 3D Human Pose Estimation. *Symmetry* **2020**, *12*, 1116. [CrossRef]
34. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907
35. Shan, W.; Lu, H.; Wang, S.; Zhang, X.; Gao, W. Improving robustness and accuracy via relative information encoding in 3D human pose estimation. In Proceedings of the ACM MM, New York, NY, USA, 20–24 October 2021.
36. 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 ICCV, Montreal, QC, Canada, 10–17 October 2021.
37. Li, S.; Chan, A.B. 3d human pose estimation from monocular images with deep convolutional neural network. In Proceedings of the ACCV, Singapore, 1–5 November 2014.
38. Zhou, X.; Zhu, M.; Leonardos, S.; Derpanis, K.; Daniilidis, K. Sparseness meets deepness: 3d human pose estimation from monocular video. In Proceedings of the CVPR, Las Vegas, NV, USA, 27–30 June 2016.
39. Arnab, A.; Doersch, C.; Zisserman, A. Exploiting temporal context for 3D human pose estimation in the wild. In Proceedings of the CVPR, Long Beach, CA, USA, 15–20 June 2019.
40. Kingma, D.P.; Jimmy, B. Adam: A Method for Stochastic Optimization. In Proceedings of the ICLR, San Diego, CA, USA, 7–9 May 2015.
41. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the ICCV, Santiago, Chile, 7–13 December 2015.
42. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The kitti dataset. *Int. J. Rob. Res.* **2013**, *32*, 1231–1237. [CrossRef]
43. 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.* **2014**, *36*, 1325–1339. [CrossRef] [PubMed]
44. Mehta, D.; Sotnychenko, O.; Mueller, F.; Xu, W.; Sridhar, S.; Pons-Moll, G.; Theobalt, C. Single-shot multi-person 3d pose estimation from monocular rgb. In Proceedings of the 3DV, Verona, Italy, 5–8 September 2018.
45. Yang, W.; Ouyang, W.; Wang, X.; Ren, J.; Li, H.; Wang, X. 3d human pose estimation in the wild by adversarial learning. In Proceedings of the CVPR, Salt Lake City, UT, USA, 18–23 June 2018.
46. Zhou, X.; Huang, Q.; Sun, X.; Xue, X.; Wei, Y. Towards 3d human pose estimation in the wild: A weakly-supervised approach. In Proceedings of the ICCV, Venice, Italy, 22–29 October 2017.
47. Luo, C.; Chu, X.; Yuille, A. Orinet: A fully convolutional network for 3d human pose estimation. *arXiv* **2018**, arXiv:1811.04989.
48. Ci, H.; Wang, C.; Ma, X.; Wang, Y. Optimizing network structure for 3d human pose estimation. In Proceedings of the CVPR, Long Beach, CA, USA, 15–20 June 2019.