Anatomy and Geometry Constrained One-Stage Framework for 3D Human Pose Estimation

Xin Cao[1,2] and Xu Zhao[1,2](✉)

[1] Department of Automation, Shanghai Jiao Tong University, Shanghai, China
[2] Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China
{xinc1024,zhaoxu}@sjtu.edu.cn

Abstract. Although significant progress has been achieved in monocular 3D human pose estimation, the correlation between body parts and cross-view geometry consistency have not been well studied. In this work, to fully explore the priors on body structure and view-relationship for 3D human pose estimation, we propose an anatomy and geometry constrained one-stage framework. First of all, we define a kinematic structure model in deep learning framework which represents the joint positions in a tree-structure model. Then we propose bone-length and bone-symmetry losses based on the anatomy prior, to encode the body structure information. To further explore the cross-view geometry information, we introduce a novel training mechanism for multi-view consistency constraints, which effectively reduces unnatural and implausible estimation results. The proposed approach achieves state-of-the-art results on both Human3.6M and MPI-INF-3DHP data sets.

1 Introduction

Human pose estimation is a fundamental task in computer vision and has been studied for decades. It refers to estimating human anatomical key points or parts and supports many applications, such as human-computer interaction, video surveillance, augmented reality, sports performance analysis and so forth [1].

In traditional way, some approaches try to learn a concise low-dimensional embedding [2] of high-dimensional 3D pose structure space to solve this problem. Pictorial structure model [3] is another representative way to model body structure, where the joints and their relations are represented as vertexes and edges respectively in a non-circular graph. Actually, tree-structured model is the

This work has been supported in part by the funding from NSFC (61673269, 61273285) and the project funding of the Institute of Medical Robotics at Shanghai Jiao Tong University.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-69525-5_14) contains supplementary material, which is available to authorized users.

H. Ishikawa et al. (Eds.): ACCV 2020, LNCS 12622, pp. 227–243, 2021.
https://doi.org/10.1007/978-3-030-69525-5_14

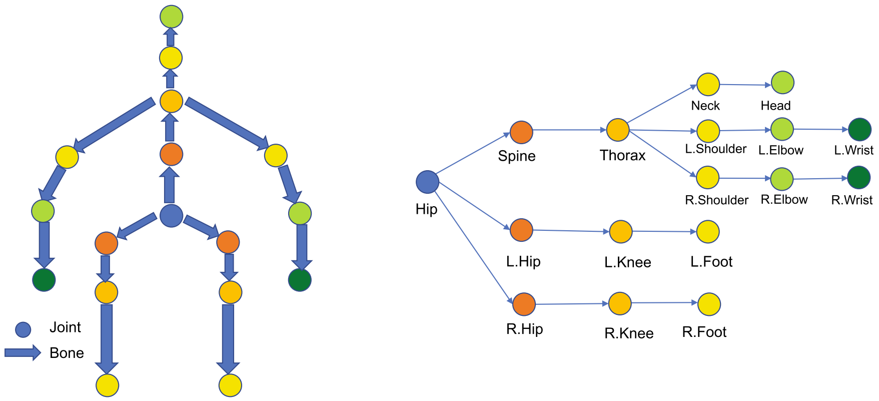

Fig. 1. Representation of human joints and bones in a body tree structure, different colors indicate different hierarchy levels. (Color figure online)

most popular pose representation and had been well studied in the traditional methods. For example, Yub et al. [4] proposed a kinematic tree pose estimation method along with the RTW expectation, where the joint positions are determined sequentially according to the typical skeletal topology.

Recently with the development of Deep Learning (DL) and the emergence of large scale 3D pose data sets [2,5], many state-of-the-art methods [6–10] have been proposed for 3D human pose estimation. These methods can be simply divided into two categories. In the first category, 3D pose positions are directly regressed from raw images. While in the second category, usually a well-trained 2D pose network is used to estimate 2D joint positions, and then a following 2D-3D lifting network is used to further acquire 3D poses.

Despite the remarkable progress that has been achieved, we argue that most of the existing 3D human pose estimation methods in DL framework treated the body joints independently and overlooked the structure information and the correlation between body parts. Seldom method utilizes kinematic structure information. Besides, current popular 3D human pose data sets like Human3.6M [2] and MPI-INF-3DHP [11] are captured in multi-view settings. However, the geometry information that can be extracted via multi-view consistency constrains have not been well studied yet.

To address the above mentioned issues, we propose an anatomy and geometry constrained one-stage framework which impose the anatomy prior and fully explore the geometry relationship for 3D human pose estimation.

In fact, the human body is like a tree structure. Suppose that the hip joint is a root node, according to the distance to the root joint, we can define six hierarchical levels of human body joints as illustrated in Fig. 1. Usually the motion range of the leaf node joint is larger than its parent node joint, so it is more difficult to estimate its 3D joint locations. Therefore it is intuitive to infer the location of the child node joint as a dependent valuable of its parent node joint's location and hence get more plausible results. To this end, we first define a **kinematic structure model** in a deep learning framework which represents human joint positions by the root joint position and each joints' transformation matrices relative to their parent joints. In this way, we can obtain the position of a joint by multiplying the transformation matrix to its parent joint along the way to the root joint following a kinematic tree. To simplify the learning process, we further decompose the transformation matrices into rotation angles and translation parameters.

To avoid the error accumulation with the expansion of the kinematic tree, following the idea of Newell et al. [12], we adopt **bone-length loss** [13], which usually is reliable information and acts as intermediate supervision during the training process. In addition, we bring up a novel **bone-symmetry loss** function based on the symmetry of human's left/right parts to penalize the inequality between the left/right limbs. These two loss functions are both based on body anatomy prior and with which some implausible results are effectively removed.

Besides, in order to fully explore the cross-view geometry relationship, we propose a **multi-view consistency constraints** algorithm to study the latent pose representation. Pose estimation results of the same person from different camera views are mapped to a latent space to encode the pose information and then the similarity between them is computed. In this way, the model is required to output the same pose representation for multi-view inputs, which implicitly explores the geometry information from different views and strengthen the generalization ability of the model.

Our contributions can be summarized as follows:

- We propose a one-stage deep learning framework with anatomy-aware kinematic structure model, by which human body structure information and anatomy prior can be captured effectively.
- We show that adding multi-view consistency constraints into the one-stage framework is able to explore the geometry relationship and reduce implausible results for 3D human pose estimation.
- Quantitative and qualitative experiments are conducted on public 3D human pose estimation data-sets, and the results demonstrate the effectiveness of our proposed method.

2 Related Work

Here we will briefly review the two main streams of 3D pose estimation solutions, the *one-stage* methods and the *two-stage* ones.

2.1 One Stage Methods

One-stage methods usually directly regresses 3D pose positions from raw images [6–9, 13, 14]. According to the final representation of human pose, this method can be further divided into regression-based and detection-based sub-categories.

Regression based methods directly map the input image space to the output joint positions. Li et al. [15] proposed a multi-task learning task which simultaneously conducted joint point regression and joint point detection tasks. Tekin et al. [16] introduced an auto-decoder model to learn a high-dimensional latent pose representation and account for joint dependencies. After that, the latent representation was mapped back to the original pose space using the decoder. The regression based method usually obtained unsatisfactory performance, because mapping raw image to the pose space is a highly non-linear process and ignores the spatial relationship between body parts. Besides, the detection based methods regard the human pose estimation problem as a detection problem and usually output a heatmap for each joint. Pavlakos et al. [17] proposed a fine discretization of the 3D space around the human body subject and trained a convNet to predict per voxel likelihoods for each joint. To improve the initial estimation positions, they used a coarse-to-fine scheme to further improve the results. In order to overcome the quantization error of the argmax operation, Sun et al. [7] proposed an integral regression method to take the expectation of the heatmap as the output 3D joint locations.

Due to the lack of large-scale in-the-wild 3D human pose datasets, there are also some researches for weakly-supervised and unsupervised 3D human pose estimation. Zhou et al. [6] used mixed 2D and 3D labels in a deep neural network which presented a two-stage cascaded structure. 2D datasets does not have 3D labels but with diverse in-the-wild images, and hence acted as weak labels for 3D pose estimation. Yang et al. [8] proposed a multi-source discriminator to distinguish the predicted 3D pose from the ground truth. Rhogin et al. [18] introduced multi-view constraints as weak supervision and trained the system to predict the same pose from all views.

As for the kinematic related works, Mount et al. [19] defined the kinematics or forward kinematic as the problem of determining where a point is transformed as a result of the rotations associated with individual joints. For the deep learning based method, Zhou et al. [20] developed a new layer to realize the non-linear forward kinematics in human hand pose estimation and obtained geometrically valid results. Zhou et al. [21] introduced the kinematic structure model for 3D human pose estimation and demonstrated its effectiveness. In fact, our method takes the inspiration from this work but have several improvements.

- In the work of [21], the root joint is simply fixed at the origin point and the bone length is set as the average of the training subject with a global scale. However, it will reduce the generality of the method and lead to intrinsic errors because scale is unknown for the test phase. Therefore, we add the root joint position and bone length as learnable parameters in the network and optimized with the training data.
- We introduce body bone length and symmetry loss which is able to express the anatomy prior and also act as intermediate supervision to avoid the accumulation of errors in the kinematic tree.

2.2 Two Stage Methods

While the two-stage methods usually first used a well-trained 2D pose network to estimate 2D pose positions, then trained a 2D-3D lifting network to further acquire 3D joint positions [10,22–26]. Thanks to the available of large scale 2D human pose datasets, these methods were able to acquire accurate 2D pose results and focused on the 2D-to-3D mapping process.

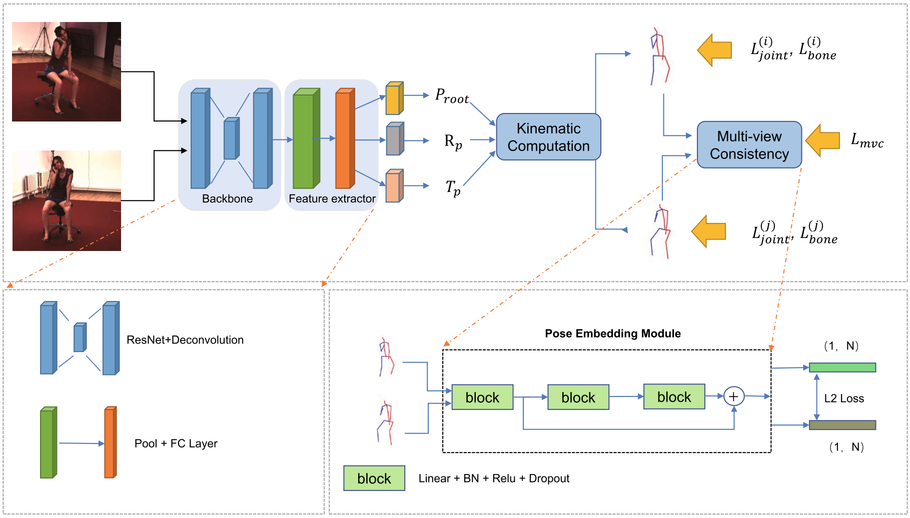

Fig. 2. Diagram of the proposed anatomy and geometry constrained framework. During training, a pair of images of different views (I_i, I_j) taken from the same person are sent to the network and acquire the corresponding root joint locations, rotation parameters and translation parameters. Then with the kinematic computation module we can compute the joint locations with the predefined process. The output of the network is optimized by the joint loss, body bone loss and multi-view consistency loss which is implemented by the pose embedding module. During inference, the network takes a single image as input and output its 3D joint locations.

Chen et al. [22] proposed to first generate a 2d pose result and then estimate its depth by matching to a library of 3d pose. Martinez et al. [23] used several residual blocks to learn the mapping from 2D joints to 3D joints, and found that a large portion of the error of modern deep 3d pose estimation systems stems from their visual analysis. Li et al. [10] proposed a multi-modal mixture density network to generate multiple hypothesis of the 3D pose for 2D joints. Drover et al. [24] utilized an adversarial framework to impose a prior on the 3D structure which is learned solely from their random 2D projections. Zhao et al. [27] proposed a semantic graph convolutional network to infer 3D joint locations from 2D joints.

3 Method

Given a cropped image $I \in \mathbb{R}^{W \times H \times 3}$, we aim to learn a mapping function θ, such that $\theta(I) = P_K$, where $P_K \in \mathbb{R}^{3 \times K}$ is the estimated position of K joints. We assume that x, y are in image pixel coordinates while z is the relative depth value to the root joint in camera coordinates.

The diagram of our proposed anatomy and geometry constrained framework is illustrated in Fig. 2. In this section, we will first introduce the kinematic computation process and the kinematic structure model. Then we will explain the bone-length and bone-symmetry loss as well as the multi-view consistency constraints. Finally we will demonstrate the total loss function for training.

3.1 Kinematic Computation

A human body is composed of joints and bones. Following a kinematic tree structure, we can reach any position of body joints from root joint with body structure information.

Suppose that the hip joint is a root node in a tree structure, we can categorize the body joints into different hierarchies according to the distance to the root joint. For example, the left hip, right hip and spine joint can be considered as the second hierarchy because they can reach the root joint without passing any other joints. Following this idea, we can define six hierarchy levels of human body joints as illustrated in Fig. 1, the circle dots indicate the 17 body joints while different colors demonstrate different hierarchy levels in a tree structure. Besides, we use arrows to define bone structures which start from parent joint and heads to its child joint.

Following the idea of [21], we can obtain the position of a child joint with its parent joint's position and the corresponding rotation and translation matrix in Eq. 1, where $\mathbf{R}_p \in \mathbb{R}^{3 \times 3}$ and $\mathbf{T}_p \in \mathbb{R}^{3 \times 3}$ indicate the rotation and translation matrices and $P_{parent} \in \mathbb{R}^{3 \times 1}$ is the coordinate of the parent joint. In a similar fashion, the position P_k of joint k can be represented following the path from the root joint to itself, where $\mathcal{P}_{(k)}$ indicates the set of parent joints along the way, and P_{root} is the position of the root joint.

For example, if we want to calculate the 3D positions of the left wrist (lw), we need to calculate the transformation matrix from the root joint to itself. With the predefined Eq. 2, we can set $\mathcal{P}_{(k)} = \{$left hip (lh), spine (sp), left shoulder (ls), left elbow (le), left wrist (lw)$\}$. Then the location of the left wrist joint can be calculated with Eq. 3.

$$P_{child} = (\mathbf{R}_p \cdot \mathbf{T}_p) \cdot P_{parent} \tag{1}$$

$$P_k = (\prod_{a \in \mathcal{P}_{(k)}} \mathbf{R}_a \cdot \mathbf{T}_a) \cdot P_{root} \tag{2}$$

$$P_{lw} = (\mathbf{R}_{lw} \cdot \mathbf{T}_{lw}) \cdot (\mathbf{R}_{le} \cdot \mathbf{T}_{le}) \cdot (\mathbf{R}_{ls} \cdot \mathbf{T}_{ls}) \cdot (\mathbf{R}_{sp} \cdot \mathbf{T}_{sp}) \cdot (\mathbf{R}_{lh} \cdot \mathbf{T}_{lh}) \cdot P_{root} \tag{3}$$

3.2 Kinematic Structure Model

According to the kinematic computation process, we first design a kinematic structure model in a deep learning framework which is illustrated in Fig. 2. Multi-view input images are first passed to a shared backbone network to extract representative contextual features. Then we use a feature extractor to output root joint positions $P_{root} \in \mathbb{R}^{3\times 1}$, rotation parameters $R_p \in \mathbb{R}^{3\times K}$, and translation parameters $T_p \in \mathbb{R}^{K}$. For the $i^{(th)}$ joint, suppose its translation parameter is l_i and the rotation parameter is α_i, β_i and γ_i, we can acquire the translation matrix and rotation matrix with the following equation. Finally together with the root joint position, Eq. 2 is applied to obtain the 3D joint positions. All the defined computing process are differentiable which allowed our model to be trained end-to-end.

$$\mathbf{T}_i = \begin{pmatrix} 1 & 0 & l_i \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \mathbf{R}_i = \begin{pmatrix} \cos\alpha_i & -\sin\alpha_i & 0 \\ \sin\alpha_i & \cos\alpha_i & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \cos\beta_i & 0 & \sin\beta_i \\ 0 & 1 & 0 \\ -\sin\beta_i & 0 & \cos\beta_i \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\gamma_i & -\sin\gamma_i \\ 0 & \sin\gamma_i & \cos\gamma_i \end{pmatrix}$$

3.3 Bone-Length and Bone-Symmetry Loss

One drawback of the kinematic structure model is that joint errors may accumulate following the kinematic tree. Apart from joint position loss which is usually adopted in current 3D pose estimation methods, we also calculate the difference between the predicted bone length and the ground truth which can be considered as intermediate supervisions [12]. For the $k^{(th)}$ joint, we define its parent joint's index as $parent(k^{(th)})$, then the associated bone can be represented as Eq. 4. In general, bone representation is more stable and able to cover geometry constraints [13].

$$\mathcal{B}_k = P_{parent(k^{(th)})} - P_{(k^{(th)})} \tag{4}$$

Besides, considering that human body is a symmetry structure. As illustrated in Fig. 3, we define four groups of symmetry bones and make a statistic for the symmetry bone length errors in Human3.6m dataset [2]. Then, we devise a loss function to penalize the inequality between the predefined symmetry left and right parts. These two loss functions are both based on body structure

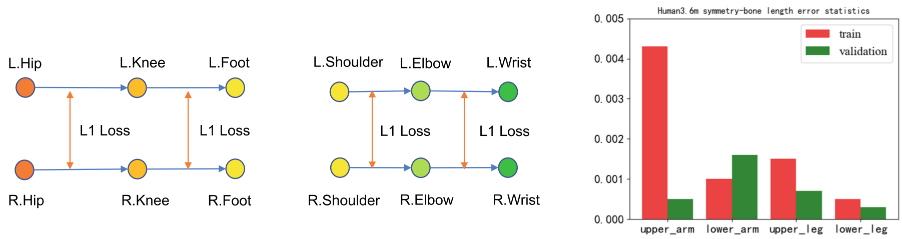

Fig. 3. Definition of the bone-symmetry loss and the statistic results in Human3.6m dataset

information and therefore able to implicitly impose the anatomy prior into the kinematic structure model.

The total bone loss is defined in Eq. 5, here $\mathcal{L}_p = \{$left lower/upper arm, left lower/upper leg$\}$, $\mathcal{R}_p = \{$right lower/upper arm, right lower/upper leg$\}$.

$$L_{bone} = \sum_{k=1}^{K-1}(\| \hat{\mathcal{B}}_k - \mathcal{B}_k{}^{gt} \|_2) + \sum_{i=1}^{4}(\| \mathcal{B}_{\hat{\mathcal{L}_p(i)}} - \mathcal{B}_{\hat{\mathcal{R}_p(i)}} \|_2) \tag{5}$$

3.4 Multi-view Consistency Constraints

To further explore the geometry relationship between multiple views, we propose a pose embedding module to represent the latent pose information. To this end, pose estimation results of the same person at the same time across multiple views are mapped to a latent space, then we computed the similarity between the cross-view encoding results. To be specific, suppose the 3D joint estimation of image i and image j are $P_K^i \in \mathbb{R}^{3\times K}$ and $P_K^j \in \mathbb{R}^{3\times K}$. Here we use a residual block to encode the joint locations into geometry latent vectors $g^{(i)} \in \mathbb{R}^{1\times N}$ and $g^{(j)} \in \mathbb{R}^{1\times N}$, where N is the length of the latent vector. Then we apply a L2 loss to compute the similarity between two geometry latent representations.

The general idea behind this computation is that P_K^i and P_K^j should be the same pose representation under global world coordinate, and then mapped to the corresponding camera coordinate. In this way, the consistency constraints will enforce the network to output the same pose embedding results across multiple views and effectively filter out implausible predictions. Moreover, our training mechanism doesn't need camera extrinsic parameters and can be implemented to any multi-view datasets. In a word, it is a multi-view and self-supervised method during training, and a single view method during inference.

The loss function is defined in Eq. 6, where $\mathcal{F}$ is the multi-view structure information encoding function, i and j indicate different camera views of the same person, and P^i and P^j represent joint positions from camera view i and j.

$$L_{mvc} = \| \mathcal{F}(P^i) - \mathcal{F}(P^j) \|_2 \tag{6}$$

3.5 Loss Function

Our total loss function includes the joint loss, body bone loss and multi-view consistency loss. Here joint loss L_{joint} is defined as the smooth L_1 loss between the predicted and ground truth joint positions.

$$L_{joint} = \sum_{i=1}^{K} \begin{cases} \frac{1}{2}(\hat{P_{i(th)}} - P_{i(th)}{}^{gt})^2 & if \mid \hat{P_{i(th)}} - P_{i(th)}{}^{gt} \mid < 1 \\ \mid \hat{P_{i(th)}} - P_{i(th)}{}^{gt} \mid -0.5 & otherwise \end{cases} \tag{7}$$

And the total loss function is in Eq. 8, where λ and β are loss weights for body bone loss and multi-view consistency loss, M is the total training samples and N is the total multi-view training sample pairs.

$$L = \sum_{i=1}^{M} \frac{1}{M}(L_{joint}(i) + \lambda L_{bone}(i)) + \sum_{j=1}^{N} \frac{1}{N} \beta L_{mvc}(j) \tag{8}$$

4 Experiment

In this part, we first introduce the datasets and evaluation metrics, then we will provide the implementation details and augmentation operations during training. And next we will show the ablation studies as well as comparisons with the state-of-the-art results.

4.1 Datasets

We conduct quantitative experiments on two publicly available 3D human pose estimation datasets: Human3.6M dataset and MPI-INF-3DHP dataset. To demonstrate the generality of the proposed model, we also provide some qualitative results on the MPII dataset, which is a challenging public outdoor 2D human pose dataset.

Human3.6M. [2] is one of the most publicly used dataset in 3D human pose estimation. It captures 3.6 million images and there are 11 subjects performing daily activities from 4 camera views in a lab environment. The 3D ground truth is obtained by the motion capture system and the camera intrinsic and extrinsic parameters are also provided.

Following the standard, we use subject 1, 5, 6, 7, 8 for training and and evaluate on every 64^{th} frame for subject 9 and 11. The evaluation metric is the mean per joint position error (MPJPE) between the estimated and the ground-truth joint positions after aligning the root position.

MPI-INF-3DHP. [5] is a recently released dataset which includes both indoor and outdoor scenes. Following the common practice, we use the 3D Percentage of Correct Keypoints (3DPCK@150 mm) and Area Under Curve (AUC) as the evaluation metrics.

MPII. [28] is a 2D dataset which provides 22k in-the-wild dataset, we demonstrate the qualitative results on this dataset to reveal the generality of our proposed method.

4.2 Implementation Details

We use ResNet-50 followed by three deconvolution layers as our backbone network to extract representative features. For the multi-view consistency module,

Table 1. Ablation study for MPJPE on Human3.6M dataset.

Model	Bone loss	Multi-view consistency loss	MPJPE (mm)
Kinematic Model	×	×	62.01
	✓	×	58.69
	×	✓	57.33
	✓	✓	**56.18**

*Here bone loss indicates bone-length and bone-symmetry loss

Table 2. Ablation study for joint location errors on Human3.6M dataset.

Method	LShoulder	LElbow	LWrist	RShoulder	RElbow	Rwrist
baseline	64.8	75.9	96.5	64.9	81.8	102.2
baseline+bone	62.2	75.2	94.4	**57.2**	**69.3**	93.8
baseline+bone+mvc	**61.3**	**67.7**	**85.7**	58.2	70.9	**89.5**

*Here bone and mvc indicate bone loss and multi-view consistency loss

we adopt a residual block [23] followed by a fully connected layer as the mapping function. As for the total loss function weights, we set $\lambda = 1\mathrm{e}{-3}$ and $\beta = 1\mathrm{e}{-4}$ after cross-validation experiments. For the Human3.6M dataset, we randomly select two views from the total four views and form six pairs of multi-view inputs. In the same way, there are four pairs of multi-view inputs for the MPI-INF-3DHP dataset.

Input images are cropped with the ground truth bbox to extract the human body region and then resized to 256 × 256. Augmentation of random rotate and flip are used for both datasets. Besides we utilize synthetic occlusions [29] to make the network robust to occluded joints. For the MPI-INF-3DHP dataset, we also apply clothing and background augmentation.

During training, we use batch size of 64. Each model was trained for 140 epochs with an initial learning rate of 1e−3 which dropped at steps 90 and 120. The ADAM optimizer is used for all the training steps. Our code was implemented with PyTorch [30] and the proposed model was trained for 8 h with 2 Nvidia 1080Ti GPUS.

4.3 Experiment Results on Human3.6M Dataset

Ablation Study. We conduct an ablation study to explore the contribution of the proposed anatomy prior and the multi-view consistency constraints.

The experiment results are demonstrated on Table 1. We notice that multi-view consistency loss had a very noticeable impact on the precision, and improving the MPJPE by 4.7 mm. This finding implies that including multi-view constraints is able to learn the geometry information and produce more reliable results.

Table 3. Ablation study for bone length error in Human3.6m datasets. Here LU indicates left upper, LL indicates left lower, RU indicates right upper and RL indicates right lower.

Method	LU arm	LL arm	RU arm	RL arm	LU leg	LL leg	RU leg	RL leg
baseline	10.8	17.3	10.9	17.2	16.0	13.8	14.4	13.1
baseline+bone	**10.2**	**15.5**	10.2	20.1	**12.9**	11.3	**12.5**	10.9
baseline+bone+mvc	10.6	15.7	**9.2**	**14.6**	13.9	**9.6**	13.0	**8.9**

Table 4. Ablation study for bone symmetry error on Human3.6M dataset.

Method	Upper arm	Lower arm	Upper leg	Lower leg
baseline	11.1	21.9	14.8	11.8
baseline+bone-loss	11.0	22.9	11.0	11.4
baseline+bone-loss+mvc-loss	**8.6**	**14.4**	**9.7**	**10.7**

Also, bone loss was shown to be of considerable benefit to the result with a decrease of MPJPE by 3.3 mm, proving the effectiveness of the proposed anatomy prior constraints.

Discussion. To further analyse the effect of the proposed anatomy and geometry constraints, we also calculate the high-hierarchy joint errors together with the bone-length and bone-symmetry errors in Human3.6M dataset. The quantitative results are illustrated in Table 2, 3 and 4.

Here we choose six joints which are far from the root joint and analyse the effect of the proposed two loss functions. We notice that for the right elbow and right wrist joint, adding bone-length and bone-symmetry loss significantly reduced the joint location errors, which indicated that bone information was able to express the anatomy prior and reduce the location errors of the high-hierarchy joints.

As for the bone length and bone symmetry errors, the results are illustrated in Table 3. Similarly introducing bone loss and multi-view consistency loss will eventually lead to smaller errors especially for the legs.

Visualization Results. In Fig. 4, we present some hard examples on the Human3.6m datasets, and make a qualitative comparison of the visualization results. These pictures include some occluded and severely deformed actions like Sitting down and Lying. Our baseline kinematic structure model is already able to output plausible results. After adding bone loss and multi-view consistency loss, some details will be further refined. For example, in the fourth picture, introducing the anatomy prior and multi-view consistency supervision will revise the unreasonable leg positions of the baseline model. More details are outlined by the green circles.

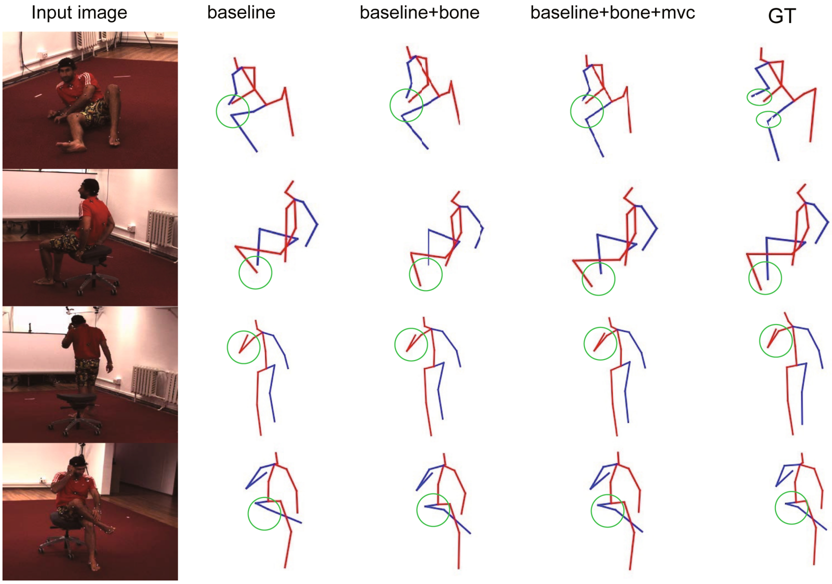

Fig. 4. Some visualization results on Human3.6m dataset. Here bone indicates bone-length and bone-symmetry loss and mvc indicates multi-view consistency loss. (Color figure online)

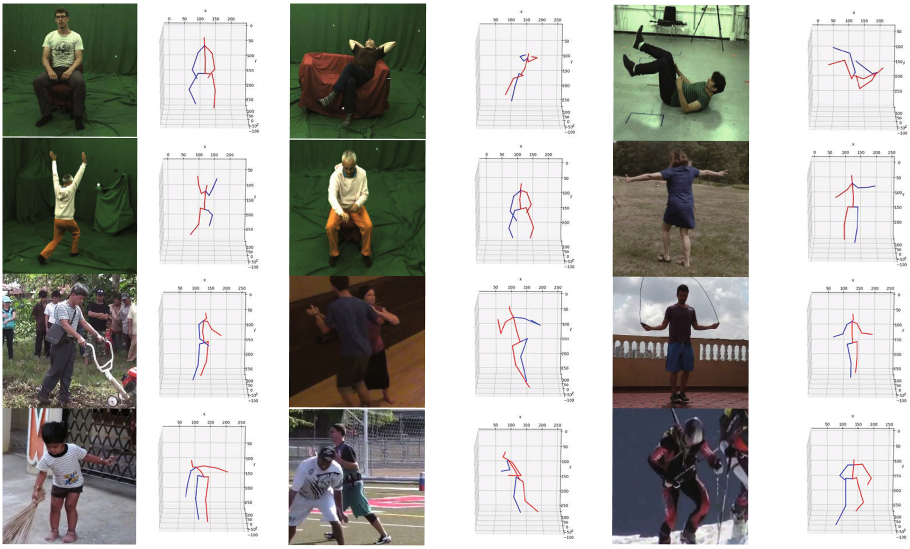

Fig. 5. Qualitative results. The first and second rows show results on MPI-INF-3DHP dataset, while the third and forth rows demonstrate results on MPII datasets.

Table 5. Comparison with the state-of-the-art on Human3.6M dataset.

Method	Dir	Disc	Eat	Greet	Phone	Photo	Pose	Purch	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkT	Avg
Chen et al. [22]	89.9	97.6	90.0	107.9	107.3	139.2	93.6	136.1	133.1	240.1	106.7	106.2	114.1	87.0	90.6	114.2
Zhou et al. [21]	91.8	102.4	96.9	98.7	113.3	125.2	90.0	93.8	132.1	158.9	106.9	94.4	126.0	79.0	98.9	107.2
Rogez et al. [31]	76.2	80.2	75.8	83.3	92.2	105.7	79.0	71.1	105.9	127.1	88.0	83.7	86.6	64.9	84.0	87.7
Pavlakos et al. [17]	67.4	71.9	66.7	69.1	72	77	65	68	83	96.5	71.7	65.8	74.9	59.1	63.2	71.9
Zhou et al. [6]‡	54.8	60.7	58.2	71.4	62	65.5	53.8	55.6	75.2	111.6	64.1	66	51.4	63.2	55.3	64.9
Martinez et al. [23]	51.8	56.2	58.1	59	69.5	78.4	55.2	58.1	74	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Yang et al. [8]‡	51.5	58.9	50.4	57.0	62.1	65.4	49.8	52.7	69.2	85.2	57.4	58.4	43.6	60.1	47.7	58.6
Sun et al. [7]‡	46.5	48.1	**49.9**	**51.1**	**47.3**	**43.2**	**45.9**	57	77.6	**47.9**	54.9	**46.9**	**37.1**	49.8	**41.2**	**49.8**
PVH-TSP [32]§	92.7	85.9	72.3	93.2	86.2	101.2	75.1	78.0	83.5	94.8	85.8	82.0	114.6	94.9	79.7	87.3
Trumble et al. [33]§	**41.7**	**43.2**	52.9	70.0	64.9	83.0	57.3	63.5	**61.0**	95.0	70.0	62.3	66.2	53.7	52.4	62.5
Ours	51.4	53.0	52.4	66.6	52.9	57.1	46.6	**47.5**	62.9	92.5	**53.0**	62.5	52.3	**41.6**	45.2	56.1

*Here § indicates multi-view methods, ‡ indicates methods training with extra 2D Pose Datasets. A lower value is better for MPJPE. The results are taken from corresponding papers. The best results are marked in bold while the second best approach is underlined.

Comparison with the State-of-the-Art. In Table 5 we compare the prediction results of our proposed 3D pose models with current state-of-the-art methods. Recall that our models are trained in a multi-view setting without any camera information and tested in single view, to make a fair comparison, we also consider results from multi-view input approaches.

Even though our model is trained with only 3D datasets, our method still outperform most of the benchmark results and is only inferior to the model by Sun et al. [7] which is a high-memory required heatmap-based model and trained with extra 2D dataset. Besides, we also make a comparison of the per-action joint errors and demonstrate the effectiveness of the proposed model.

Table 6. Comparison with the state-of-the-art results on MPI-INF-3DHP dataset.

Method	3DPCK	AUC
VNect [11]	**76.6**	**40.4**
Mehta et al. [34]	75.2	37.8
SPIN [35]	76.4	36.1
Ching-Hang Chen et al. [36]	71.1	36.3
LCR-Net [31]	59.6	27.6
Zhou et al. [6]	69.2	32.5
Ours	72.0	37.3

*A higher value is better for 3DPCK and AUC. The results are taken from corresponding papers.

We notice that in some actions like purchasing, photoing and sitting where extensive movements may appear in body parts, our model got the lowest error. This finding provides compelling empirical evidence for the benefits of our proposed bone loss and multi-view consistency constraints which effectively encodes geometry structure information and thus reduces implausible results.

4.4 Experiment Results on MPI-INF-3DHP Dataset

MPI-INF-3DHP Dataset contains a mixture of indoor and outdoor scenes in test set, and we also evaluate our method on this challenging dataset. As can be seen from Table 6, we achieved a comparable result of 72.0 in 3DPCK and 37.3 in AUC, which indicated the strong robustness of our proposed model.

Besides, we provide some visualization results on the test set in Fig. 5. Even in some severely movable action and unseen outdoor scenes, our model still provides satisfactory results.

4.5 Qualitative Results on MPII Dataset

To demonstrate the cross-domain generalization ability of our proposed model, we also test our method on the MPII dataset. Note that our model is only trained on the Human3.6m dataset which contains constrained data in indoor environment. Since the ground truth 3D pose results are not available, we only give qualitative results on the third and forth rows of Fig. 5. We can see that our model can output plausible results and generalize well on unseen scenes.

5 Conclusion

In this paper, we propose an anatomy and geometry constrained one-stage framework which imposes the anatomy prior and fully explore the geometry relationship for 3D human pose estimation. We first define a kinematic structure model in a deep learning framework, then we introduce bone loss which utilizes bone-length and bone-symmetry property to capture the anatomy prior. In addition, we show that adding a multi-view consistency constraints during training can improve the performance and reduce implausible results. We conduct quantitative experiments on two 3D benchmark datasets and achieve state-of-the-art results.

References

1. Sarafianos, N., Boteanu, B., Ionescu, B., Kakadiaris, I.A.: 3D human pose estimation: a review of the literature and analysis of covariates. Comput. Vis. Image Underst. **152**, 1–20 (2016)
2. 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. **36**, 1325–1339 (2014)
3. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. Int. J. Comput. Vis. **61**, 55–79 (2005)
4. Yub Jung, H., Lee, S., Seok Heo, Y., Dong Yun, I.: Random tree walk toward instantaneous 3D human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2467–2474 (2015)
5. Mehta, D., et al.: Monocular 3D human pose estimation in the wild using improved CNN supervision. In: 2017 Fifth International Conference on 3D Vision (3DV). IEEE (2017)
6. 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 IEEE International Conference on Computer Vision, pp. 398–407 (2017)
7. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11210, pp. 536–553. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01231-1_33
8. 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 IEEE Conference on Computer Vision and Pattern Recognition, pp. 5255–5264 (2018)
9. Rhodin, H., Salzmann, M., Fua, P.: Unsupervised geometry-aware representation for 3D human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 750–767 (2018)
10. Li, C., Lee, G.H.: Generating multiple hypotheses for 3D human pose estimation with mixture density network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9887–9895 (2019)
11. Mehta, D., et al.: VNect: real-time 3D human pose estimation with a single RGB camera. ACM Trans. Graph. **36**, 44 (2017)
12. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
13. Sun, X., Shang, J., Liang, S., Wei, Y.: Compositional human pose regression. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2602–2611 (2017)

14. 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 IEEE Conference on Computer Vision and Pattern Recognition, pp. 10905–10914 (2019)
15. Li, S., Chan, A.B.: 3D human pose estimation from monocular images with deep convolutional neural network. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9004, pp. 332–347. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16808-1_23
16. Tekin, B., Katircioglu, I., Salzmann, M., Lepetit, V., Fua, P.: Structured prediction of 3D human pose with deep neural networks. arXiv preprint arXiv:1605.05180 (2016)
17. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3D human pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7025–7034 (2017)
18. Rhodin, H., et al.: Learning monocular 3D human pose estimation from multi-view images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8437–8446 (2018)
19. Mount, D.M.: CMSC 425 game programming (2013)
20. Zhou, X., Wan, Q., Zhang, W., Xue, X., Wei, Y.: Model-based deep hand pose estimation. arXiv preprint arXiv:1606.06854 (2016)
21. Zhou, X., Sun, X., Zhang, W., Liang, S., Wei, Y.: Deep kinematic pose regression. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9915, pp. 186–201. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49409-8_17
22. Chen, C.H., Ramanan, D.: 3D human pose estimation = 2D pose estimation + matching. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
23. 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, pp. 2640–2649 (2017)
24. Drover, D., Rohith, M.V., Chen, C.-H., Agrawal, A., Tyagi, A., Huynh, C.P.: Can 3D pose be learned from 2D projections alone? In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018. LNCS, vol. 11132, pp. 78–94. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11018-5_7
25. Wandt, B., Rosenhahn, B.: RepNet: weakly supervised training of an adversarial reprojection network for 3D human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7782–7791 (2019)
26. Akhter, I., Black, M.J.: Pose-conditioned joint angle limits for 3D human pose reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1446–1455 (2015)
27. Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N.: Semantic graph convolutional networks for 3D human pose regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3425–3435 (2019)
28. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2D human pose estimation: new benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, pp. 3686–3693 (2014)
29. Sárándi, I., Linder, T., Arras, K.O., Leibe, B.: How robust is 3D human pose estimation to occlusion? arXiv preprint arXiv:1808.09316 (2018)
30. Paszke, A., et al.: Pytorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, pp. 8024–8035 (2019)

31. Rogez, G., Weinzaepfel, P., Schmid, C.: LCR- Net: localization-classification-regression for human pose. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
32. Trumble, M., Gilbert, A., Malleson, C., Hilton, A., Collomosse, J.: Total capture: 3D human pose estimation fusing video and inertial sensors. In: British Machine Vision Conference 2017 (2017)
33. Trumble, M., Gilbert, A., Hilton, A., Collomosse, J.: Deep autoencoder for combined human pose estimation and body model upscaling. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11214, pp. 800–816. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01249-6_48
34. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Theobalt, C.: Single-shot multi-person 3D pose estimation from monocular RGB. In: 2018 International Conference on 3D Vision (3DV) (2018)
35. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2252–2261 (2019)
36. Chen, C.H., Tyagi, A., Agrawal, A., Drover, D., Stojanov, S., Rehg, J.M.: Unsupervised 3D pose estimation with geometric self-supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5714–5724 (2019)