

Skeleton Transformer Networks: 3D Human Pose and Skinned Mesh from Single RGB Image

Yusuke Yoshiyasu, Ryusuke Sagawa, Ko Ayusawa, Akihiko Murai

¹ National Institute of Advanced Industrial Science and Technology (AIST)

² CNRS-AIST JRL (Joint Robotics Laboratory), UMI3218/RL,

{yusuke-yoshiyasu, ryusuke.sagawa, k.ayusawa, a.murai}@aist.go.jp

Abstract. In this paper, we present Skeleton Transformer Networks (SkeletonNet), an end-to-end framework that can predict not only 3D joint positions but also 3D angular pose (bone rotations) of a human skeleton from a single color image. This in turn allows us to generate skinned mesh animations. Here, we propose a two-step regression approach. The first step regresses bone rotations in order to obtain an initial solution by considering skeleton structure. The second step performs refinement based on heatmap regressor using a 3D pose representation called cross heatmap which stacks heatmaps of xy and zy coordinates. By training the network using the proposed 3D human pose dataset that is comprised of images annotated with 3D skeletal angular poses, we showed that SkeletonNet can predict a full 3D human pose (joint positions and bone rotations) from a single image in-the-wild.

Keywords: Convolutional neural networks, 3D human pose, skeleton.

1 Introduction

Estimating 3D human pose from a single image is an important yet very challenging problem in computer vision, where applications range from surveillance to robotics. Recent work has shown that convolutional neural networks (ConvNets) can detect 2D joint positions accurately. The key to achieving accurate predictions is to represent 2D joint locations as heatmaps and iteratively refines them gradually by incorporating context information [16,3]—early approaches [25] on the other hand directly perform regression of the 2D joint coordinates using ConvNets, which is a difficult problem that needs to model a highly-nonlinear mapping from an input image to real values. Recent techniques, such as OpenPose [5], can robustly detect 2D joints of multiple people in an image.

In contrast to its 2D counterpart, the progress of 3D human pose detection has been relatively slow. The main challenges regarding to 3D human pose estimation is as follows:

3D pose representation To predict 3D joint locations using ConvNets, 3D pose representation used is critical, which affects prediction accuracy. Previous approaches have shown that regression of a 3D pose using heatmaps (e.g., volumetric [18] and 2D heatmaps + depth [28]) leads to accurate 3D joint predictions.

On the other hand, regression of joint angles using ConvNets [29,10] has not been successful so far in contributing to accurate 3D joint localizations, because they are difficult to learn with ConvNets due to their high non-linearity. From the application point of view, such as computer animation and biomechanics, it is desirable to predict not only 3D joint locations but also angular pose of skeleton, e.g., joint angles or segment rotations.

Data scarcity Compare to 2D human pose dataset, 3D human pose dataset is smaller in size. This is because obtaining a 3D human pose dataset where the images paired with 3D joint annotations requires time and effort. In particular, annotations of 3D angular skeletal poses are difficult to obtain. One common way to achieve this is to use a motion capture (MoCap) system and RGB video cameras at the same time. However, such dataset are usually limited to a small variety of subjects—for example, Human 3.6M dataset [9], which is the most common dataset for 3D human pose, is limited to around 10 subjects. Consequently, it is difficult to learn sufficient visual features from 3D pose dataset solely to localize 2D/3D joints accurately.

Skeletal structure has been incorporated into 3D human pose estimation as a form of constraints or subspace. In biomechanics and robotics, forward and inverse kinematics have been well-studied and are used to generate human pose from MoCap by controlling joint angles of a skeleton. Previous approaches in the computer vision field estimated 3D human pose from 2D key points using a statistical model and enforcing constraints such as segment length [21], joint limit [1] and symmetry. In computer graphics, human skeletal pose is often represented using linear or affine transformation matrices, as can be seen for example in linear blend skinning for character animation.

In this paper, we propose skeleton transformer networks (SkeletonNet) for 3D human pose detection which respects skeletal structure while attaining 3D joint prediction accuracy. SkeletonNet combines and benefits from the two paradigms, skeleton and heatmap representations. SkeletonNet first regresses bone rotations to an input image in order to have an initial solution which is not precisely accurate but considers skeletal structure. Starting from this initial solution, the second step refines it using ConvNet heatmap regressor. This strategy allows us to recover reasonably accurate predictions of full 3D human poses (joint positions and bone rotations) from a single in-the-wild image. To contribute to solving the data scarcity problem, we also construct a dataset where 3D angular skeletal poses are annotated on in-the-wild images, based on a human-validation approach. By fusing the proposed dataset and Mocap dataset captured under a controlled environment (Human3.6M), SkeletonNet can predict a full 3D pose (joint positions and bone rotation) from a single image in-the-wild. In addition, experimental results showed that SkeletonNet outperforms previous approaches based on joint angles [10,29] in terms of MPJPE joint position accuracy.

The contributions of this work is summarized as follows:

- We propose an end-to-end ConvNet framework for predicting a full 3D human pose (joint positions and bone rotations).

- We propose a bone rotation regressor which predicts 3D human pose using 3×3 transformation matrices. To make arbitrarily linear transformations into rotations, we propose a Gram Schmidt orthogonalization layer. This combination is the key to learning angular pose accurately with using ConvNets.
- We propose a 3D human pose representation called cross heatmap for accurate 3D joint detection. This representation combines two heat maps, one for representing 2D joints in image space (xy space) and the other for zy space. The benefit of this representation is that it is more efficient than the volumetric heatmaps [18], while accurately predicting 3D joint positions when trained using Mocap-video dataset (Human 3.6M) and in-the-wild dataset (MPII with 3D pose annotations).
- We built a 3D human pose dataset in-the-wild which includes annotations of 3D bone rotations.

2 Related work

Estimating 2D joint positions using ConvNets Recent work has shown that the detection of 2D joint positions can be done very accurately using convolutional neural networks (ConvNets). Toshev et al. [25] first proposed a method based on ConvNets for detecting human pose i.e., 2D key points representing joint locations from a single image. Tompson et al. instead represented joint locations in images using 2D heat maps so that it can avoid complicated a nonlinear mapping that goes from an image to xy pixel coordinates. The recent techniques, such as the stacked hourglass network [16] and its variants [27,7,6], accurately predict 2D joint positions by iteratively refining 2D heatmaps.

Predicting 3D joint positions The early approaches predicts 3D joint positions from key points [20]. These approaches assume that the almost perfect 2D key points are already extracted from an image. Li et al. [13] first used ConvNets to directly regress 3D human joints with an image. There are two main reasons for the improvements on accuracy of 3D human pose detection. First, the recent approach combines multiple data sources to increase the 3D pose dataset [15,28]. Second, the recent techniques make use of more natural 3D joint representation. For example, Pavlakos et al. [18] uses a volumetric heatmap representation, which can avoid regressing the real values in a highly nonlinear manner. Other methods first predicts 2D joints with heatmaps and then regress 3D joint positions or depths from them. Tome et al. [24] have tried to iteratively update 3D joints represented as a weighted combination of PCA basis that is constructed from 3D MoCap dataset.

Exploiting skeletal structure and predicting angular pose In biomechanics, robotics and computer animation fields, inverse kinematics has been well-studied and used to generate human pose from MoCap by controlling joint angles. Previous approaches [20,4] estimated 3D human pose from 2D key points by combining a statistical model and constraints such as joint limit [1], segment length [20] and symmetry. Some methods perform regression of joint angles or

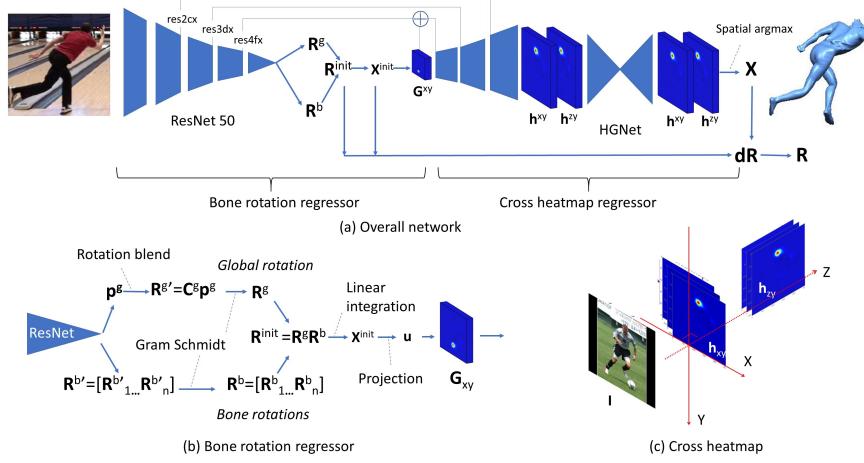


Fig. 1. Network architecture of SkeletonNet. (a) SkeletonNet is a two-step regressor. (b) The first part performs regression of bone rotations. (c) The second part detects heatmaps of xy and zy spaces.

axis angles [29,10] to estimate angular skeletal pose using ConvNets but the high nonlinearity prevents them from accurate prediction of joint locations.

Weakly supervision and predictions from in-the-wild images Recent works tackle the data scarcity problem by using both 3D human pose dataset captured in the experimental room and 2D human pose dataset captured in a wide range of environment [28,15,23]. Sun et al. [23] used compositional loss function that is defined by integrating 2D positions and depths by properly normalizing the regression target values. Zhou et al. [28] took a weak supervised approach and used bone length constraint when 3D information (depths) is not available. The use of this approach not only enables 3D joint predictions from in-the wild images but also performance boost in joint prediction accuracy. We go further by building a 3D human pose dataset in-the-wild which includes annotations of 3D bone rotations.

3 Skeleton Transformer Networks

Skeleton Transformer Networks (SkeletonNet) is based on deep ConvNets, which predicts not only 3D joint positions but also rotations of body segments. Our approach is based on an end-to-end two-step regression approach. The network architecture of SkeletonNet is depicted in Fig. 1 (a). The first step is called *bone rotation regressor* that performs regression of angular skeletal pose using 3×3 rotation matrices to provide an initial solution. The resulting pose is not precisely accurate but it respects skeletal structure, without having to produce large errors such as left and right confusions. The second step, *cross heatmap*

regressor, starts from this initial solution and refine the joint predictions using heatmap regression. We benefit from the two different paradigms, i.e., skeleton and ConvNets, to achieve accuracy and preservation of structure.

3.1 Bone rotation regressor

Bone rotation regressor predicts bone rotations of a human skeleton. We achieve this by solving two simpler problems separately: predicting 1) a global rotation and 2) transformations of bone segments relative to the root. The idea behind this strategy is that the global orientation of the body has some discrete patterns e.g., sit, stand and lie, which can be effectively solved as a classification problem. On the other hand, bone rotations have more continuous distributions within some range, which can be effectively modeled as a regression problem.

To predict a global rotation, we convert the rotation estimation problem into a classification problem. Specifically, we first cluster the training dataset into 200 clusters based on its global rotations with k -means clustering. The samples within the same cluster is put in the same class. ConvNets is trained to output rotation class probabilities \mathbf{p}^g using Softmax. We use a classification loss (cross entropy) for supervision:

$$\mathcal{L}_{\text{RotG}} = L_{\text{cls}}(\mathbf{p}^g, \bar{\mathbf{p}}^g)$$

where L_{cls} is the log loss and $\bar{\mathbf{p}}^g$ is the one-hot class label of global rotation. From the output probabilities \mathbf{p}^g , we obtain a 3×3 global transformation matrix by blending cluster centers \mathbf{C}^g , $\mathbf{R}^{g'} = \mathbf{C}^g \mathbf{p}^g$. We tried linear blending of axis angles but found that blending matrices works better.

Since bone rotations relative to the root have more continuous distributions than global rotation, regression is more suitable in this case than classification. Bone rotations $\mathbf{R}^{b'} = [\mathbf{R}_1^{b'} \dots \mathbf{R}_n^{b'}]$, where n is the number of bones, are thus predicted directly using a 3×3 rotation matrix (9 parameters). The loss for bone rotations is defined using the mean squared error (MSE) loss as:

$$\mathcal{L}_{\text{RotB}} = \sum_i^n \|\text{vec}(\mathbf{R}_i^{b'}) - \text{vec}(\bar{\mathbf{R}}_i^b)\|_2^2$$

where $\text{vec}(\cdot)$ makes a matrix to a vector and $\bar{\mathbf{R}}^b$ is the ground truth bone rotations. Note that regression of rotation matrices demands more memory spaces because they need more parameters than other rotation representation, such as Euler angles, quaternions and axis angles. However, for 3D human pose detection, we have under 20 joints to predict, which means that the additional costs are almost negligible. The down side of Euler angles and quaternions are their nonlinearities and ambiguities (sign flips for quaternions and periodical angle jumps for Euler angles), which is difficult to use as the regression targets—we could not train a network properly using Euler angles as supervisions as reported in [29].

Gram Schmidt orthogonalization layer The problem of the above strategy is that it does not guarantee to produce orthonormal matrices. This means the resulting skeleton is deformed in an undesirable way with scales and shears. To solve this issue, we propose the Gram Schmidt (GS) orthogonalization layer which performs GS to make transformations into rotations. GS requires elemental functions only, such as dot product, subtraction and division, which is differentiable and can be relatively easily incorporated into ConvNets. We input global transformation \mathbf{R}^g' and bone transformations $\mathbf{R}^{b'}$ into the GS layer to make transformations to rotations, obtaining \mathbf{R}^g and \mathbf{R}^b .

Once transformations are orthonormalized, we multiply a global rotation \mathbf{R}^g and bone rotations \mathbf{R}^b in order to obtain the absolute bone rotations. Finally, 3D joint positions are computed by applying these absolute rotations to the original bone vectors in the rest pose and performing linear integration to add up bone vectors from the root (Fig. 1 (b)).

3.2 Cross heatmap regressor

We propose cross heatmap regressor which is used for refining the 3D joints obtained using bone rotation regressor. In the current design, cross heatmap regressor stacks xy and zy heatmaps (Fig. 1 (c)) because they sufficiently cover xyz coordinates and the variance of human joint locations in zy coordinates are usually larger than that of zx space.

To integrate bone rotation regressor and cross heatmap regressor, we project 3D joint positions obtained in Sec 3.1 into the image plane (xy plane). Here, we did not estimate a camera pose and scale explicitly. Instead we scale the xy coordinates to 90% of the width of the first upsampling layer, which is 16 pixels. From the projected 2D joints, 2D Gaussian maps [18] are obtained and, after convolutions, they are summed up with the feature maps from bone rotation regressor to serve as approximate positions of 2D joints for cross heatmap regressor. Note that all of these process are differentiable, which can be optimized using back propagation.

Once the feature maps are up-sampled three times to the size of 64×64 , a single stack of hourglass module [16] is used to compute cross heatmaps. The cross heatmap representation concatenates two heatmaps, one for representing 2D joints in image space (xy space) \mathbf{h}^{xy} and the other for the zy space, \mathbf{h}^{zy} . For training we use the MSE loss as:

$$\mathcal{L}_{\text{hm}} = \sum_i^m \sum_{j,k} \|\mathbf{h}_{(j,k)}^{xy} - \bar{\mathbf{h}}_{(j,k)}^{xy}\|_2^2 + \sum_i^m \sum_{j,k} \|\mathbf{h}_{(j,k)}^{zy} - \bar{\mathbf{h}}_{(j,k)}^{zy}\|_2^2$$

where $\bar{\mathbf{h}}^{xy}$ and $\bar{\mathbf{h}}^{zy}$ are the ground truth heatmaps for xy and zy spaces. In addition, m is the number of joints. The benefit of this representation is that it is more compact and efficient than the volumetric heatmaps, while maintaining accuracy. To extract xyz coordinates from cross heatmaps in an end-to-end manner, we use spatial argmax layers similar to those proposed in [14,12]. Finally, we compute rotations \mathbf{dR} that align bone vectors obtained using bone rotation

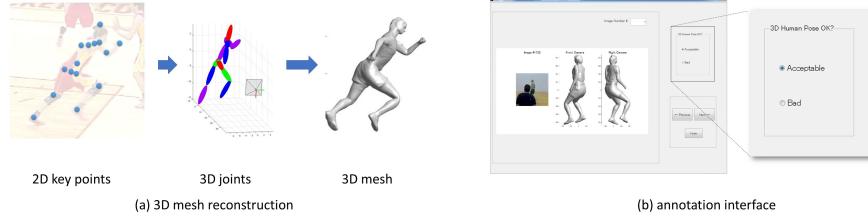


Fig. 2. 3D pose annotation system. (a) Given 2D key points as inputs, we compute 3D joints using the projected matching pursuit (PMP) technique. Next, skeleton rotations are obtained using a skeleton transformation optimization method. 3D mesh is obtained from rotations and positions. (b) With the annotation tool, the annotator is just needed to judge whether the 3D pose is acceptable or not by comparing the real image and the rendered image created from the mesh.

regressor with that of the final predicted positions. They are multiplied with the predicted absolute bone rotations \mathbf{R}^{init} to make them consistent with the final joint positions \mathbf{x} . This way, we can exploit accurate heatmaps to refine rotations, avoiding a difficult regression problem of nonlinear angle parameters. Now that the loss for the final positions \mathbf{x} and rotations \mathbf{R} are defined as:

$$\mathcal{L}_{\text{pos}} = \sum_i^m \|\mathbf{x}_i - \bar{\mathbf{x}}_i\|_2^2, \quad \mathcal{L}_{\text{Rot}} = \sum_i^n \|\text{vec}(\mathbf{R}_i) - \text{vec}(\bar{\mathbf{R}}_i)\|_2^2$$

where and $\bar{\mathbf{x}}_i$ and $\bar{\mathbf{R}}_i$ are the ground truth labels of positions and rotations.

With the final positions \mathbf{x} and rotations \mathbf{R} , linear blend skinning can be done to produce a 3D mesh. Note that we perform skinning outside the network but this process can also be done within the network in an end-to-end manner, as this is a linear matrix multiplication.

3.3 Loss function

For supervision, a standard cross entropy loss and MSE loss is used for comparing the predictions and ground truth labels of the global rotation class probability, bone rotations and cross heatmaps. In total, we minimize the loss function of the form:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{RotG}} + \alpha \mathcal{L}_{\text{RotB}} + \beta \mathcal{L}_{\text{Rot}} + \gamma \mathcal{L}_{\text{pos}} + \lambda \mathcal{L}_{\text{hm}}$$

where $\mathcal{L}_{\text{RotG}}$, $\mathcal{L}_{\text{RotB}}$, \mathcal{L}_{Rot} , \mathcal{L}_{pos} and \mathcal{L}_{hm} are a cross entropy loss for global rotation and MSE loss for bone rotations \mathbf{R}^b , final rotations \mathbf{R} , final positions \mathbf{x} and cross heatmaps (\mathbf{h}^{xy} and \mathbf{h}^{zy}), respectively. In addition, α , β , γ and λ are the respective weights.

4 In-the-wild 3D Human Pose Dataset

To build 3D human pose dataset that have variations in clothes and background, we annotate 3D positions and bone rotations on MPII human pose dataset as shown in Fig. 2. To this end, we followed the human verification approach [17]. Here a 3D human pose is obtained from 2D key points using a previous technique, which will then be sorted by human annotators whether each result is acceptable or not. This is in spirit similar to the concurrent work of [11]. The aim for constructing our dataset is not to provide accurate 3D pose annotations but to obtain reasonable ones on in-the-wild images, including joint positions and rotations, such that they can be used for training our bone rotation regressor. In fact, the annotators are instructed to judge whether the pose is ‘acceptable’ or ‘bad’, e.g., if the global rotation of the resulting pose looks deviating from the true pose more than 30 deg, the pose is ‘bad’. As a consequence, our dataset have more 3D pose annotations than [11], possibly at the cost of accuracy.

To obtain 3D joint positions from 2D key points, we use the projected matching pursuit (PMP) approach [20]. This method calculates a camera pose, scale and 3D joint positions as a combination of PCA basis that is constructed from Mocap database. From the resulting 3D joint positions, rotations of bones are obtained based on a method which is conceptually similar non-rigid surface deformation techniques [22]. Specifically, the skeleton in the rest shape is fitted to the PMP result by balancing the rigidity of bones, the smoothness between bone rotations and the position constraints to attract the skeleton to them. The initial rotations are obtained from local coordinate frames, which are defined in a similar manner using [1]. Reconstructing one model from 2D key points takes approx. 1min.

We also designed a simple annotation tool Fig. 2 (b) to simplify the process of 3D pose annotations. With this tool, human annotators are just required to decide a 3D pose is acceptable or not. In addition, we obtained skin meshes from the 3D pose using linear blend skinning and showed rendered images. By visualizing a skin mesh, it makes the annotators’ decisions significantly easier and quicker. From among the images in MPII dataset, we extracted those with all 16 joints are inside the image region, which was approximately 20000 poses. After annotations, we were able to collect 10291 images with 3D pose annotations. Note that we remove ‘bad’ poses and do not use them in the training. It took about 2-3 hours for an annotator to process 1000 images.

5 Experiments

5.1 Dataset and evaluation protocols

MPII [2] This dataset contains in-the-wild images for 2D human pose estimation, which includes 25k training images and 3k validation images. Those images are annotated with 2D joint locations and bounding boxes. In Section 4, we constructed a 3D pose dataset on top of MPII dataset by annotating 3D joints to images. This dataset is used in training.

Human3.6M Human 3.6M dataset [9] is used in training and testing. Human 3.6M dataset is a large scale dataset for 3D human pose detection. This dataset contains 3.6 million images of 15 everyday activities, such as walking, sitting and making a phone call, which is performed by 7 professional actors. 3D positions of joint locations captured by motion capture (Mocap) systems are also available in the dataset. In addition, 2D projections of those 3D joint locations into images are available. The images are taken from four different views. As with previous researches, we down-sampled the video from 50fps to 10fps in order to reduce redundancy in video frames. We followed the same evaluation protocol used in previous approaches [18,28] for evaluation, where we use 5 subjects (S1, S5, S6, S7, S8) for training and the rest 2 subjects (S9, S11) for testing. We used the 3D model of an actor provided in Human3.6 to generate mesh animations but this could be replaced with any 3D models.

The error metric used is called mean per joint position error (MPJPE) in *mm*. In the evaluation protocol, the position of the root joints is aligned with the ground truth but the global orientation is kept as is. Following [28] the output joint positions from ConvNets is scaled so that the sum of all 3D bone lengths is equal to that of a canonical average skeleton. This is done by:

$$\mathbf{P}_j = (\mathbf{P}_j^{\text{pred}} - \mathbf{P}_0^{\text{pred}}) \cdot l^{\text{ave}} / l^{\text{pred}} \quad (1)$$

where $\mathbf{P}_j^{\text{pred}}$ is the predicted position, \mathbf{P}_0 is the root position, l^{pred} is the sum of skeleton length of the predicted skeleton and l^{ave} is the average of sum of skeleton length for all the training subjects in Human 3.6M dataset.

We also evaluated the method with the error measure called the reconstruction error, where, before calculating the error, the result is aligned to the ground truth with a similarity transformation.

5.2 Baselines

Four baseline methods are implemented to conduct ablation studies. We trained the first three networks using Human3.6M dataset only and the fourth one with both Human3.6M and MPII dataset.

Rotation regress (Rot reg) only This only uses bone rotation regressor. The position is computed from linear integration of bone rotations.

Heatmap (HM) only This on the other hand only uses cross heatmap regressor.

Rot reg + HM This method is our proposed method which combines bone rotation regressor and cross heatmap regressor.

All This is our proposed method trained using Human 3.6M and MPII dataset with 3D annotations obtained using the method presented in Sec 4.

5.3 Implementation and training detail

Our method is implemented using MatConvNet toolbox [26]. For the bone rotation regressor, we use ResNet50 [8] as the base network, which is pre-trained on

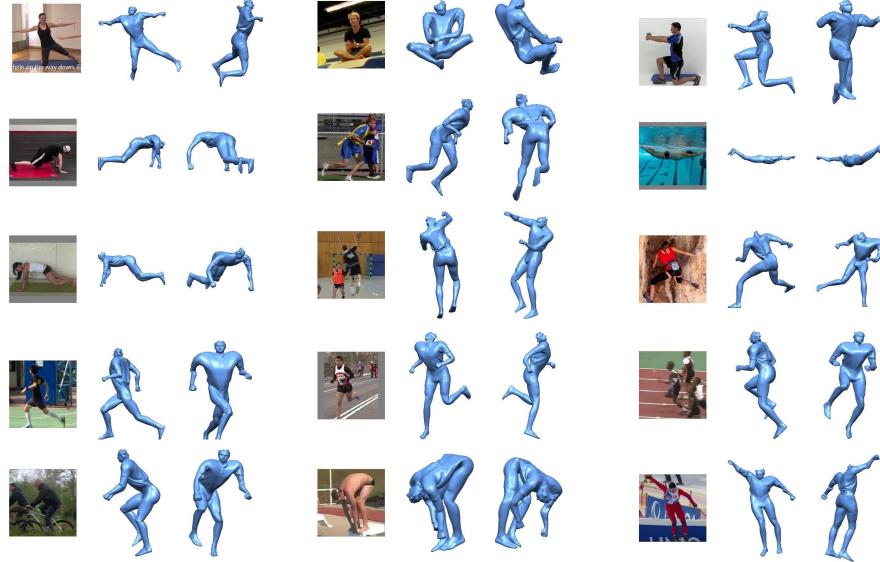


Fig. 3. Some results on in-the-wild images.

the ImageNet dataset. A single up-sampling layer and a single hourglass module is followed by the base network to predict heatmaps. We also use skip connections to connect skeleton regression layers and up-sampling layers. We used a skeleton with 16 joints and 15 segments each of them have 9 rotational parameters. The definition of joints is same as that of MPII dataset. Training a whole model takes about 1 day using three NVIDIA Quadro P6000 graphics cards with 24 GB memory. The batch size is 30 for each GPU. As for augmentation, we used left/right flip only—no scaling and rotation augmentation is used. We trained a model with SGD for 70 epochs, starting from the learning rate of 0.001 and decreasing it to 0.00001. During test time, a single forward pass of our network is approx. 0.12 sec, which means the performance of our method is approx. 8-9 fps.

We set the parameters in the loss function as $\alpha = \beta = \gamma = 0.1$ and $\lambda = 0.001$. When training with both Human 3.6M and MPII datasets, we randomly selected the images from both dataset such that half of the batch is filled with Human 3.6M and the other half by MPII dataset, following [28]. Since the 3D annotations of MPII are not accurate, we use them for supervising bone rotation regressor only. Thus, when training images are from MPII, we do not back propagate gradients from the losses that include the final 3D pose, \mathbf{h}^{zy} , \mathbf{x} and \mathbf{R} , (i.e., L_{pos} , L_{Rot} and the right term of L_{hm}) to the network. On the other hand, when the training sample is from Human3.6M dataset, which has accurate 3D annotations, we minimize all the losses.

5.4 Results

Qualitative results In Figs. 3 and 6, we show the result of our 3D pose prediction method. Our method is able to predict 3D joint positions and bone orientations reasonably accurately even on in-the-wild-images. Because our method can predict bone rotations of human body skeletons, we can produce mesh animations from 3D joint positions and rotations using linear blend skinning. Note that the rotations of hands and feet are not supervised.

Comparisons to other state-of-the-art We have compared our techniques against other state-of-the-art in Table 1. Our technique is comparable with other state-of-the-art in terms of MPJPE. Volumetric heatmaps [18] can achieve MPJPE approx. 71 mm. However, it got worse results when including MPII (MPJPE 78 mm) with their decoupled structure, whereas we are around MPJPE 70 mm. Also, compared with [18] with two stacks of hourglass networks, cross heatmap is more compact, which requires 1/32 of memory spaces to store.

Compared with the previous techniques that predict angular poses [29,10], SkeletonNet is more accurate. In fact, MPJPE of our result is 69.9 mm, whereas that of Kanazawa et al. [10] is 87.97 mm. In Table 2, we also compared the reconstruction error with previous approaches. Our technique outperforms previous techniques that iteratively optimizes joint angles [4] and perform regression of joint angles [19]. The benefit of SkeletonNet is, in addition to estimating 3D joint positions relatively accurately, we can predict 3D bone rotations, which is useful in animating a human body mesh or possibly predicting dynamics such as joint torques.

Table 1. Comparisons to other state-of-the-art. MPJPE [mm] is used for error metric.

	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases
Zhou et al. [29]	91.8	102.4	97.0	98.8	113.4	90.0	93.8	132.2
Tome et al. [24]	64.98	73.47	76.82	86.43	86.28	110.67	68.93	74.79
Mehta et al. [15]	59.69	69.74	60.55	68.77	76.36	85.42	59.05	75.04
Pavlakos et al. [18]	67.38	71.95	66.70	69.07	71.95	76.97	65.03	68.30
Ours (All)	63.33	71.59	61.39	70.40	69.90	83.17	62.98	68.77
	Sitting	SittingDown	Smoking	Waiting	WalkDog	Walking	WalkPair	Average
Zhou et al. [29]	159.0	106.9	125.2	94.4	79.0	126.0	99.0	107.3
Tome et al. [24]	110.19	172.91	84.95	85.78	86.26	71.36	73.14	88.39
Mehta et al. [15]	96.19	122.92	70.82	68.45	54.41	82.03	59.79	74.14
Pavlakos et al. [18]	83.66	96.51	71.74	65.83	74.89	59.11	63.24	71.90
Ours (All)	76.81	98.90	68.24	67.45	73.74	57.72	57.13	69.95

Comparisons between baselines In Table 3 and Fig. 4, we show comparisons between the baselines. As can be seen from Fig. 4 a, the result of bone rotation regressor preserves skeletal structure, but the joint positions are not accurate

Table 2. Comparison of reconstruction errors on Human 3.6M dataset.

Zhou et al. [30]	Bogo et al. [4]	Lassner et al. [11]	Pavlakos et al. [19]	Ours
106.7	82.3	80.7	75.9	61.4

Table 3. Comparisons between baselines. MPJPE [mm] is used for error metric.

Rot reg only	Heatmap only	Rot reg + HM	All
112.43	128.55	87.05	69.95

enough. With only heatmaps, however, skeletal structure is sometimes destructed e.g., by the left and right flips. By combining our bone rotation and cross heatmap regressor, a more accurate result is produced, while preserving skeletal structure. Note that in previous work this kind of confusions are remedied by incorporating recurrence [3] or using many stacks of a hourglass module [16]. By training with both Human 3.6M and MPII, we get the best result (Table 3). In addition, we found that annotating 3D rotations is important for reconstructing human poses from in-the-wild images (Fig. 4 b). Thus, the key to our improvements in MPJPE is the use of cross heatmap and the use of MPII dataset in training. Even when MPII dataset is not provided for training, SkeletonNet can predict reasonably accurately 3D human pose by exploiting the combination of skeletal structure and heatmaps.

Comparisons between rotation representation We have also compared the results of bone rotation regressor by changing its rotation representation. Specifically, we tested the network that 1) regresses axis angles but indirectly supervised with rotation matrices (AA), 2) regresses axis angles but supervised with relative joint rotation matrices and converts them back to the absolute space using forward kinematics, which is equivalent as SMPL [4] (FKAA), 3) regresses absolute rotations (AbsRotReg), 4) regresses rotations without the GS layer (w/o GS), 5) regresses a global rotation (GlobalReg), 6) classifies a global rotation (GlobalClass) and 7) is same as GlobalClass but aligns rotations with heatmaps (All), respectively. The networks are trained with Human 3.6M, except for All that was trained with both MPII and Human 3.6M. To compare the rotation prediction accuracy, we compute relative rotations between the ground truth and predicted bone rotations, convert them to axis angles and take the norms in degrees, which reflects all three DoFs of rotations.

In Table 4, Global Rot. Err. indicates the error of global rotations. Bone Rot. Err. indicates the average error of bone rotations relative to the root. As shown in Table 4, the proposed method based on the GS layer, which classifies a global rotation, is the best in terms of MPJPE accuracy. AbsRotReg is also high in accuracy but it produces bone rotations with its determinant of -1, which collapse skeletal structure. The method based on axis angle tends to produce large errors probably because of their high non-linearity, requiring an iterative

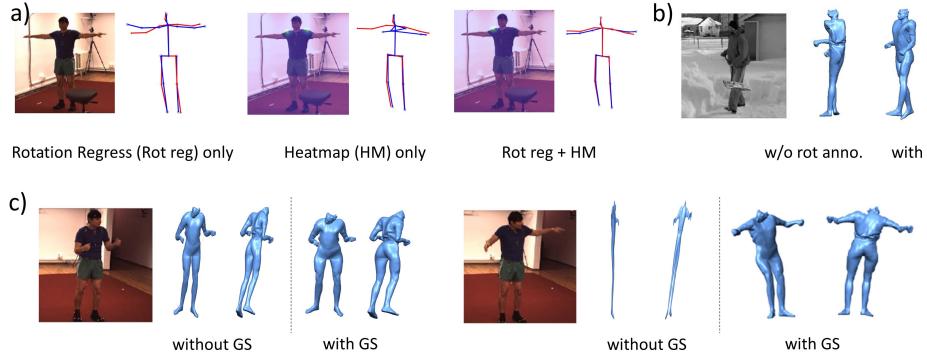


Fig. 4. Comparisons of baselines. a) Bone rotation regressor preserves skeletal structure but the joint positions are not accurate enough. With only heatmaps, skeletal structure is sometimes destructed e.g., with the left and right flips. By combining both, it produces more accurate result while preserving structure. b) The rotation annotation is important for reconstructing a pose from in the wild images. c) With the proposed Gram Schmidt (GS) orthogonalization layer, undesirable deformations such as shears and scalings are removed.

process [10] or a more informative geometric loss, e.g., the one using differences between silhouettes [19]. In summary, our method can benefit from the use of 3×3 rotation matrices, which can probably be modeled as simpler functions than other angle representations, which is more friendly to ConvNets to learn with. As reported in [29], we could not train a network properly using Euler angles as supervisions, where training and validation losses remained high. In contrast, SkeletonNet can model subtle pose appearances due to e.g., medial and lateral rotations around segment axes by providing supervisions on both rotations and positions. With joint position supervisions only and no rotational supervisions, it is possible to get reasonable results in joint position predictions [29] but is difficult to obtain good results for bone orientations.

Table 4. Comparison of rotation representation.

	AA	FKAA	AbsRotReg	w/o GS	GlobalReg	GlobalClass	All
MPJPE	175.06	197.44	114.70	124.06	119.18	112.42	69.95
Global Rot. Err.	30.46	35.81	18.83	21.28	21.64	20.81	12.93
Bone Rot. Err.	37.34	44.77	—	28.72	29.08	29.46	21.94

Failure cases and limitations In Fig. 5, we show failure cases. Our technique fails when there are large self-occlusions and occlusions by objects or other humans. In addition, our network are currently designed for the single-person de-



Fig. 5. Failure cases.

tection and thus fails when multiple humans exist in the image. Since we scale a skeleton, we are not be able to model absolute bone lengths. Cross heatmap regressor possesses the ability to alter relative bone lengths but our method have generalization issues when the body type is extremely different from the original skeleton, e.g., prediction of small children’s poses. Also our network does not take into account hand and foot orientations.

6 Conclusion

We have presented SkeletonNet, a novel end-to-end 3D human pose detection technique from a single image. The first step regresses bone segment rotations to obtain an initial solutions without large errors by considering skeleton structure. The second step performs refinement based on heatmap regressor that is based on the representation called cross heatmap which stacks heatmaps of xy and zy coordinates. This combination allows us to predict bone orientations and joint positions accurately, which may provide useful information to applications like animation and biomechanics. We also presented a 3D human pose dataset constructed by adding 3D rotational annotations to publicly-available 2D human pose dataset.

In future work, we would like to address monocular detections of other human body properties, such as body shape, body weight, contact forces and joint forces/torques. We are also interested in generative adversarial networks (GAN) to improve pose prediction results using an unsupervised manner based on the image dataset that does not have 3D annotations.

Acknowledgment

I would like to thank Rie Nishihama and CNRS-AIST JRL members for supporting us constructing 3D human pose dataset. This work was partly supported by JSPS Kakenhi No. 17K18420 and No. 18H03315.

References

1. I. Akhter and M. J. Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *CVPR*, pages 1446–1455, 2015.

2. M. Andriluka, L. Pishchulin, P. Gehler, and S. Bernt. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, June 2014.
3. V. Belagiannis and A. Zisserman. Recurrent human pose estimation. *arXiv preprint arXiv:1605.02914*, 2016.
4. F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, pages 561–578. Springer, 2016.
5. Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.
6. Y. Chen, C. Shen, X. Wei, L. Liu, and J. Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. *CoRR*, abs/1705.00389, 2017.
7. X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. Multi-context attention for human pose estimation. *CoRR*, abs/1702.07432, 2017.
8. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
9. C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2014.
10. A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.
11. C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *CVPR*, July 2017.
12. S. Levine, C. Finn, T. Darrell, and P. Abbeel. End-to-end training of deep visuo-motor policies. *CoRR*, abs/1504.00702, 2015.
13. S. Li and A. B. Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *ACCV*, pages 332–347. Springer, 2014.
14. D. C. Luvizon, H. Tabia, and D. Picard. Human pose regression by combining indirect part detection and contextual information. *CoRR*, abs/1710.02322, 2017.
15. D. Mehta, H. Rhodin, D. Casas, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation using transfer learning and improved cnn supervision. *arXiv preprint arXiv:1611.09813*, 2016.
16. A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499. Springer, 2016.
17. D. P. Papadopoulos, J. R. R. Uijlings, F. Keller, and V. Ferrari. We don’t need no bounding-boxes: Training object class detectors using only human verification. *CoRR*, abs/1602.08405, 2016.
18. G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. *CoRR*, abs/1611.07828, 2016.
19. G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *CVPR*, 2018.
20. V. Ramakrishna, T. Kanade, and Y. Sheikh. Reconstructing 3d human pose from 2d image landmarks. *ECCV*, pages 573–586, 2012.
21. E. Simo-Serra, A. Ramisa, G. Alenyà, C. Torras, and F. Moreno-Noguer. Single image 3d human pose estimation from noisy observations. In *CVPR*, pages 2673–2680. IEEE, 2012.
22. O. Sorkine and M. Alexa. As-rigid-as-possible surface modeling. In *Proceedings of the Fifth Eurographics Symposium on Geometry Processing*, SGP ’07, pages 109–116, Aire-la-Ville, Switzerland, Switzerland, 2007. Eurographics Association.

23. X. Sun, J. Shang, S. Liang, and Y. Wei. Compositional human pose regression. *arXiv preprint arXiv:1704.00159*, 2017.
24. D. Tome, C. Russell, and L. Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. *arXiv preprint arXiv:1701.00295*, 2017.
25. A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, pages 1653–1660, 2014.
26. A. Vedaldi and K. Lenc. Matconvnet – convolutional neural networks for matlab. In *Proceeding of the ACM Int. Conf. on Multimedia*, 2015.
27. W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang. Learning feature pyramids for human pose estimation. *CoRR*, abs/1708.01101, 2017.
28. X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Weakly-supervised transfer for 3d human pose estimation in the wild. *arXiv preprint arXiv:1704.02447*, 2017.
29. X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei. Deep kinematic pose regression. In *ECCV Workshops*, pages 186–201. Springer, 2016.
30. X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *CVPR*, pages 4966–4975, 2016.

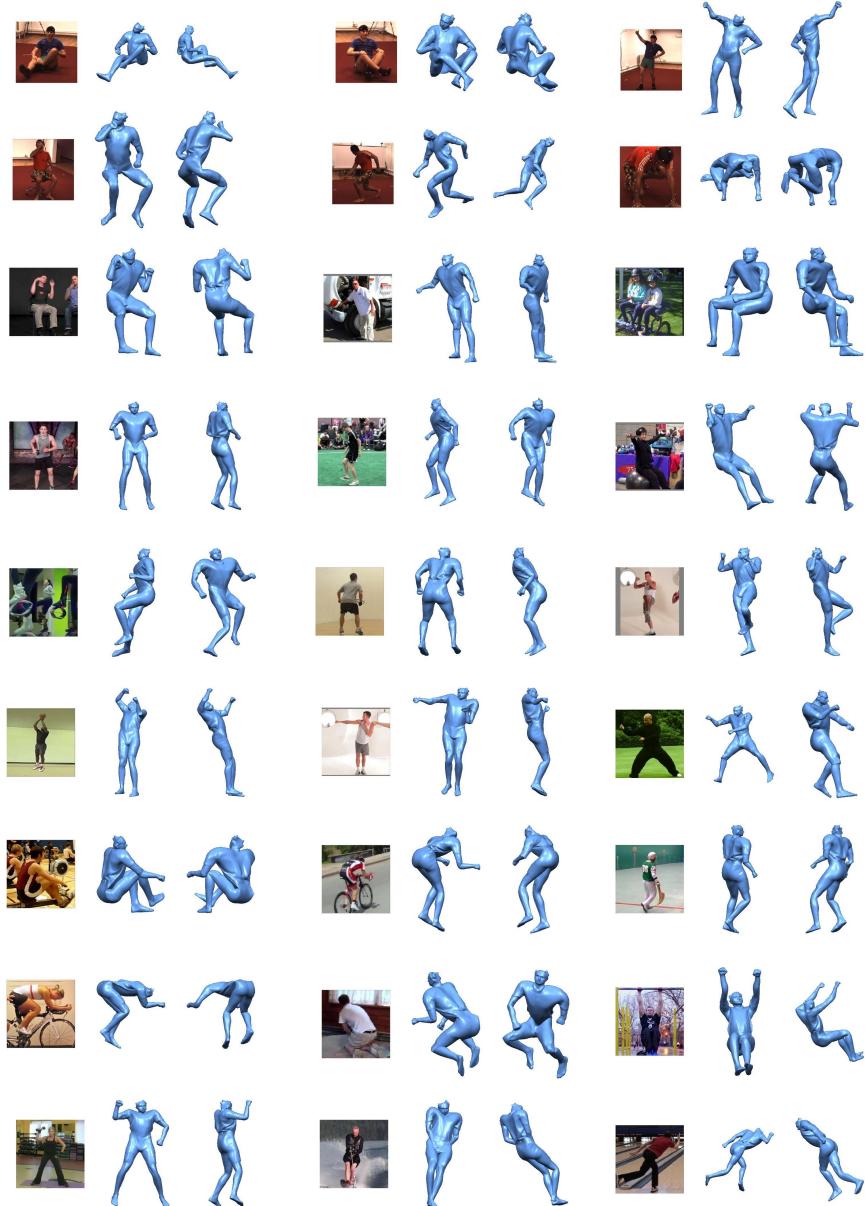


Fig. 6. More results.