A Geometry Loss Combination for 3D Human Pose Estimation

Ai Matsune Shichen Hu Guangquan Li Sihan Wen Xiantan Zhu Zhiming Tan*
Fujitsu R&D Center Co., Ltd.
{matsune.ai, hushichen, liguangquan, wensihan, zhuxiantan, zhmtan}@fujitsu.com

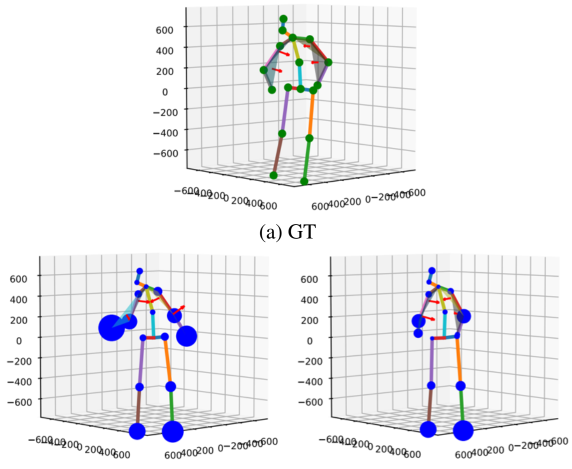
Abstract

Root-relative loss has formed the basis of 3D human pose estimation for many years. However, this point-to-point loss treats every keypoint separately and ignores internal connection information of the human body. This leads to illegal pose prediction, which humans cannot form in the real world. It also suffers from differences in estimation difficulty between keypoints. The farther the keypoint is from the torso, the less accurate it is. To address the above problems, this paper proposes geometry loss combination to utilize the geometric relationship between each keypoint fully. This loss combination consists of three loss functions: root-relative pose, bone length, and body part orientation. The previous two have already been used in prior works. Beyond them, we further develop a loss function called body part orientation loss for local body parts. Intuitively, the human body can be divided into three parts: the head, torso, and limbs. Based on this, we select the corresponding keypoints and create virtual planes for each body part. Experiments with different datasets and models demonstrate that our proposed method improves the prediction accuracy. We also achieve MPJPE of 65.0 on the 3DPW test set, which outperforms state-of-the-art methods.

1. Introduction

3D Human Pose Estimation (3DHPE) is a classic task in computer vision that aims to estimate 3D body joint coordinates from a given image or video. It has highly practical value with multiple applications in the fields of action recognition [6], action analysis [9], and human-robot interaction [40], *etc*.

Most of the 3DHPE methods treat every keypoints independently by using root-relative loss function, including heatmap-based [8, 37, 43] and regression-based methods. However, this commonly used point-to-point root-relative loss excludes any inner relationship between joints. Even if the distance between the ground truth (GT) and the prediction is extremely close, the predicted pose can even be illegal, which is impossible to be posed by a human in the real world. We also find that, for keypoints far from the torso, the further they are, the less accurate they are (see Fig. 1). This phenomenon is due to the complexity of the movement and the frequency of self-occlusion, which varies in difficulty of estimation depending on the keypoint. Therefore, using the geometric relationship between each keypoint is essential to limit the prediction results and reduce geometric ambiguity.



(a) GT

(b) Pred. (base)

(c) Pred. (base+ours)

Figure 1. Comparison between poses estimated with and without our proposed method. Red arrows indicate the orientation of the virtual plane formed by the three keypoints, and the size of the blue keypoints indicates the distance from GT.

Nevertheless, how to fully use the geometric relationship of keypoints draws little attention. Several works have tried different approaches to utilize the structural information of the human body. The straightforward idea is to use the bone length information. Sun [44] *et al*. analyze both 3D dataset and 2D dataset, then find that bone representation of poses is more stable and easier for learning. Compared to the root-relative loss, although bone length loss considers the relationship of the two adjacent keypoints, it neglects the geometry information of limbs that consist of bones. Dabral [7] *et al*. consider the illegal poses and defines an illegal-angle loss. However, this constraint only rejects the illegal pose prediction and does not focus on improving the prediction of ends of limbs.

To solve the above issues, we propose to combine three geometry loss functions: root-relative pose loss, bone length loss, and body part orientation loss. This combination allows the models to become aware of the movement and size of the body parts during the training process, which cannot be measured by mere distances between estimated points and GT points. Firstly, the root-relative pose loss calculates the distance to the GT of each joint based on the relative position of each joint to the pelvis joint. The bone length between two keypoints on the limbs is calculated as the second loss. Thirdly, the body part orientation is represented as the normal vector of the virtual planes, created from three significant keypoints belonging to the same body part. This body part orientation loss is first introduced to the 3DHPE field and can represent the geometry information of limbs.

Our proposed geometry loss combination can significantly improve the accuracy of keypoints at the end of limbs through the geometry information of the human body. We implement our proposed method on different models and evaluate on the standard benchmark datasets: Human3.6M [13], MPI-INF-3DHP [32], and 3DPW [49]. The results demonstrate that our proposed methods can estimate 3D human pose more accurately. Furthermore, we apply our methods to a large model and achieve State-Of-The-Art (SOTA) results on 3DPW multi-person benchmarks [49].

In summary, our main contributions are:

- We propose a body part orientation loss for 3DHPE. To the best of our knowledge, this is the first attempt to use surface normals of limb-formed local planes as a loss for 3D human pose estimation.
- We propose a geometry loss combination of root-relative pose, bone length, and body part orientation. These three components constrain the estimation from three aspects. This can improve the estimation of body parts that have considerable mobility.
- We implement our ideas based on different types of 3DHPE methods. Our proposed loss combination is general and applicable to other algorithms. We also outperform SOTA methods on the 3DPW benchmark.

2. Related Work

3D human pose estimation. 3D human pose estimation has been widely studied for many years. With the availability of large datasets [3,13,16,32,48] and models [10,24,41], the overall accuracy of pose estimation has improved recently. However, complex poses, such as heavily occluded, are still difficult to estimate, and the estimation is likely unnatural. We refer the readers to [5,28] for a detailed survey.

Loss functions for 3D human pose estimation. In recent years, researchers have been discussing how to properly supervise the human body's kinematic structure. The design of an appropriate loss function has attracted attention as a solution to this challenge.

Supervising the error distance of joints is an intuitive and fundamental approach in human pose estimation [26, 43, 46]. However, the point-to-point distance comparisons ignore the structural relationships between keypoints, essential to body composition. Sun [44] *et al.* statistically demonstrated that bone-based representations are more stable and suitable for training. The authors also proposed a compositional loss function that combines the L1 bone length loss and long-range joint loss, demonstrating the effectiveness of adding structure-aware supervision. Zhou [54] *et al.* proposed the geometric loss, which calculates the sum of the variance between the predicted bone length and the average bone length of the training dataset for each bone group. Pavllo *et al.* [38] introduced a soft constraint on the bone length by applying L2 loss, which incentivizes the plausible 3D pose estimation. Habibie [11] *et al.* also employed L2 bone length loss in addition to calculating joint loss.

While bone length is one of the essential factors in accurately representing the human pose, information about the orientation of each body part is also essential. Dabral [7] *et al.* introduced a combination of three structure-aware loss functions: illegal-angle loss, symmetry loss, and geometric loss [54]. Inspired by the anatomical fact that the human body has angular limitations in the range of motion of the limbs, the illegal-angle loss restricts the bending of the predicted limbs beyond 180 degrees.

Normal-based loss functions for depth estimation. Monocular depth estimation is one of the research directions for estimating the depth of an input image and has much in common with predicting the z direction in 3D pose estimation. Predicting the z direction is more complicated than predicting the xy direction because less information is available from the image. In order to achieve a geometrically consistent estimation, normal-based loss functions have been introduced. The surface normal [39] loss computes the L1 loss of the normal of the tangent plane of adjacent 3D points locally. The virtual normal loss [50] extends the surface normal globally, which computes the L1 loss of the virtual plane normal formed by randomly sampled 3D points.

3. Proposed Method

3.1. Geometry Loss Combination

The human body can be roughly divided into the head, torso, and limbs. As illustrated in Fig. 1, the joints far from the torso usually have worse results than those on the torso. This can be easily explained that the limbs have the largest range of motion and are the most difficult to predict, as discussed in [44]. To address this phenomenon, a restriction on the body parts is proposed to improve the prediction accuracy of these joints far from the torso. We then propose a geometric loss combination to improve the accuracy of those joints.

Our proposed geometric loss combination comprises three parts: root-relative pose loss (see Sec. 3.2), bone length loss (see Sec. 3.3), and body part orientation loss (see Sec. 3.4). The geometry loss combination L is formulated as:

$$L = w_{root} \cdot L_{root} + w_{bone} \cdot L_{bone} + w_{bpo} \cdot L_{bpo}, \quad (1)$$

where the L_{root} is the root-relative pose loss, L_{bone} is the bone length loss, and L_{bpo} is our proposed body part orientation loss. To balance the weighted loss into the same magnitude, the weights (w_{root}, w_{bone}, and w_{bpo}) are determined experimentally. In our experiments, we uniformly set w_{root} to 1.0.

3.2. Root-relative Pose loss

It is hard to accurately estimate the absolute joint position due to the dataset's wide variation of joint position. We choose the pelvis as the root joint and calculate the root-relative position of the other joints, yielding the loss function of root-relative pose L_{rel} as:

$$L_{rel} = \sum_{i=1}^{N} \left\| (\boldsymbol{J}_i^{pred} - \boldsymbol{J}_{root}^{pred}) - (\boldsymbol{J}_i^{gt} - \boldsymbol{J}_{root}^{gt}) \right\|_1 \quad (2)$$

where N is the total joint number, $\boldsymbol{J}_i$ denotes the ith joint, and $\boldsymbol{J}_{root}$ denotes the root joint. Here, we employ Manhattan Distance to calulate the distance.

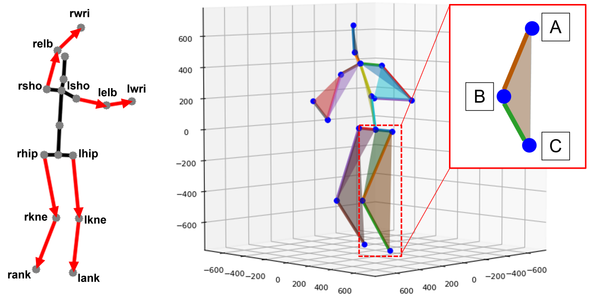
3.3. Bone Length Loss

Intuitively, the human skeleton contains hierarchical information limited by human biological structure. The distance between adjacent joints, called bone length, is relatively easier and more stable to learn than to regress joints directly [44].

We define the bone length for limbs (see Fig. 2a). For the arm, three joints are selected: shoulder, elbow, and wrist. Then, the distance between the adjacent joints is calculated. The bone length of the arm is defined as:

$$B_{up,1} = \|\boldsymbol{J}_{lsho} - \boldsymbol{J}_{lelb}\|_1 + \|\boldsymbol{J}_{rsho} - \boldsymbol{J}_{relb}\|_1, \quad (3)$$
$$B_{up,2} = \|\boldsymbol{J}_{lelb} - \boldsymbol{J}_{lwri}\|_1 + \|\boldsymbol{J}_{relb} - \boldsymbol{J}_{rwri}\|_1, \quad (4)$$

where $B_{up,1}$ stands for the Manhattan Distance between shoulder and elbow. Similarly, the $B_{up,2}$ represents the distance between the elbow and wrist.



(a) (b)

Figure 2. (a) An example of a human pose skeleton. Calculate the bone length of each red arrow. (b) Local planes that we define for each body part. The zoomed-in part shows the example point for calculating surface normal.

For the leg, we calculate the Manhattan Distance between hip and knee, also knee and ankle. The bone length of the leg is defined as:

$$B_{low,1} = \|\boldsymbol{J}_{rhip} - \boldsymbol{J}_{rkne}\|_1 + \|\boldsymbol{J}_{rhip} - \boldsymbol{J}_{rkne}\|_1, \quad (5)$$
$$B_{low,2} = \|\boldsymbol{J}_{lkne} - \boldsymbol{J}_{lank}\|_1 + \|\boldsymbol{J}_{rkne} - \boldsymbol{J}_{rank}\|_1. \quad (6)$$

Then, two arm parts are added up, same for the leg, denoted as L_{up} and L_{low}.

$$L_{up} = \|B_{up,1}^{gt} - B_{up,1}^{pred}\|_1 + \|B_{up,2}^{gt} - B_{up,2}^{pred}\|_1 \quad (7)$$
$$L_{low} = \|B_{low,1}^{gt} - B_{low,1}^{pred}\|_1 + \|B_{low,2}^{gt} - B_{low,2}^{pred}\|_1 \quad (8)$$

Finally, we define the bone length loss function as Eq. (9). The bone length is added to the loss function, and this structural information of the human body further restricts the position of the joints on the limbs.

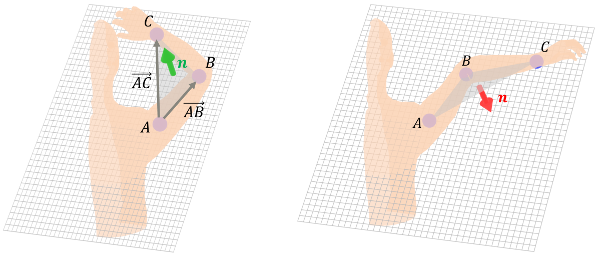
$$L_{bone} = mean(L_{up} + L_{low}) \quad (9)$$

3.4. Body Part Orientation Loss

Depth estimation is the most challenging part of estimating a 3D pose from a 2D image caused by the variation and self-occlusion of human limbs. The variations of human poses are typically represented by limbs rather than the torso. Hence, it is more difficult to predict the correct positions of arms and legs.

On the other hand, there is also plenty of effort into monocular depth prediction for 3D scene understanding [39]. Their research shows that geometric constraints can play a crucial role. The surface normal is the most used variable in point cloud data processing and depth estimation. It becomes a liable 3D cue for depth prediction from 2D images. Therefore, we apply these kinds of constraints to 3D human pose estimation.

The keypoints of the human body can also form many local planes, and the normal vectors of these planes represent



(a) Legal right-arm pose.

(b) Illegal pose.

Figure 3. For the right arm, wrist, elbow, and shoulder can form a triangle whose orientation is defined as $\overrightarrow{AB} \times \overrightarrow{AC}$. The direction of the plane normal in (a) points out the x-y plane, and the direction in (b) points in the x-y plane instead.

the orientations of this body part. We define six planes in total, as shown in Fig. 2b. Although root-relative pose and bone length make the prediction converge to GT, the limitation of joint mobility is not considered. As shown in Fig. 3, some directions of normal vectors are illegal in the typical human pose. To tackle this pose ambiguity, we apply body part orientation loss.

With three key points, our method can be applied to any part of the body. For each plane formed by the selected joints shown in Fig. 2b, the normal vector is calculated as:

$$\boldsymbol{n} = \frac{\overrightarrow{AB} \times \overrightarrow{AC}}{\|\overrightarrow{AB} \times \overrightarrow{AC}\|}, \tag{10}$$

where $\overrightarrow{AB}$ and $\overrightarrow{AC}$ refer to Fig. 2b, $\boldsymbol{n}$ stands for the normal vector of the local plane.

Each body part is represented by the keypoints that we selected. The loss function of body part orientation can be defined as:

$$L_{bpo} = \frac{1}{M}\left(\sum_{i=1}^{M}\left\|\boldsymbol{n}_i^{gt} - \boldsymbol{n}_i^{\text{pred}}\right\|_1\right), \tag{11}$$

where M denotes the number of triangles. By adding the L1 norm of the normal vector to the loss function, the model is more constrained in estimating the pose for the body parts far from the torso, further coping with the problem of having multiple solutions from 2D to 3D.

4. Experiments

To validate the effectiveness of the proposed geometry loss combination, we conduct four experiments: ablation studies (see Sec. 4.5), validation on different datasets (see Sec. 4.6), validation on different models (see Sec. 4.7), and comparison to the other SOTA methods (see Sec. 4.9).

4.1. Model Architecture

The following experiments mainly employ three models: MSH, MeTRAbs [43], and MeTRAbs+.

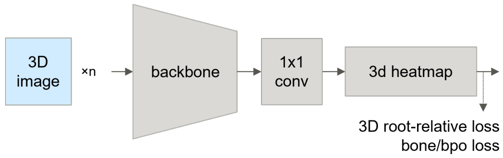


Figure 4. The network architecture of MSH.

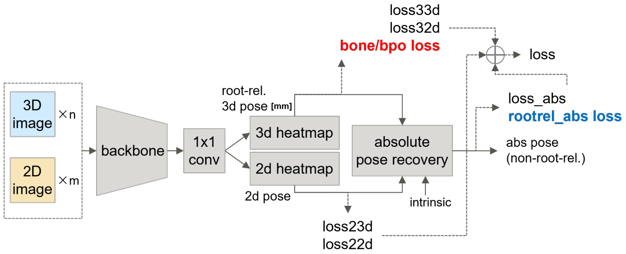


Figure 5. The network architecture of MeTRAbs. Red shows our additional loss functions. Blue shows our modification for MeTRAbs+. lossXYd: X denotes the type of GT annotation (2D/3D), and Y denotes the type of processed metric heatmap (2D/3D).

MSH for ablation studies (Sec. 4.5). First, we design a simple model called MSH (Metric Space Heatmap) to validate the effectiveness of the loss combination. The network architecture consists of only a backbone and a 3D metric heatmap, as shown in Fig. 4. Here, as our main focus is not proposing a new network architecture, we design a simple model for quick experiments.

MeTRAbs for generalization validation (Sec. 4.6, Sec. 4.7 and Sec. 4.8). Second, we re-implement MeTRAbs [43] for further experiments. The network architecture is shown in Fig. 5. MeTRAbs [43], an enhanced MeTRo [42] that combines 3D and 2D metric heatmaps, achieves first place in "3D Human Pose Estimation on 3D Poses in the Wild (3DPW) Challenge" [49].

MeTRAbs estimates root-relative pose (loss33d and loss32d) in metrics space through its 3D heatmap head and then recovers absolute pose (loss_abs) with the output of the 2D heatmap head and intrinsic camera matrix. In the implementation, we calculate the bone length loss (bone loss) and the body part orientation loss (bpo loss) of the root-relative pose in metric space.

MeTRAbs+ for SOTA comparisons (Sec. 4.9). Third, based on our experiments and [41], we make a few modifications to the original MeTRAbs. As described in Fig. 5, besides absolute pose error, we also calculate the root-relative error of absolute pose during training. Hereafter, this model is referred to as MeTRAbs+.

4.2. Datasets

Datasets for ablation studies. We conduct ablation studies on the largest standard benchmark Hu-

Dataset	#Examples	#Keypoints	Skeleton
3D-labeled data			
Human3.6M [13][1]	85K	24	SMPL
MPI-INF-3DHP [32]	76K	28	3DHP
Muco-3DHP [33][2]	220K	17	3DHP
CMU-Panoptic [16]	220K	19	COCO
AIST++ [23, 47]	211K	24	SMPL
3DOH50K [52]	50K	14	LSP
MADS [53]	33K	15(19)	MADS
HUMBI [51]	200K	19	COCO
AGORA [36]	59K	30	SMPL
SURREAL [48]	229K	24	SMPL
2D-labeled data			
COCO [27]	26K	17	
MPII [2]	23K	10	
LSP [15]	10K	14	
CrowdPose [20]	14K	14	
Totals			
3D Dozens-M [41][3]	10.8M (14 datasets)		
ours	1.38M (10 datasets)		
2D Dozens-M [41]	173K (4 datasets)		
ours	73K (4 datasets)		

[1] Annotations published by third parties [35] are used.
[2] Annotations published by third parties [46] are used.
[3] Medium dataset.

Table 1. Details of our 1.38M dataset.

man3.6M [13]. Four high-speed cameras are used to capture this dataset in the indoor studio. We use the setting of Protocol 2, in which subjects 1,5,6,7,8,9 are used as training sets while 11 are used as test sets.

Datasets for generalization validation. To further validate the generalization ability on different datasets of our loss combination, we also evaluate our methods on MPI-INF-3DHP [32]. 3DHP is captured by a commercial markerless motion capture system. 3DHP covers more complicated poses than Human3.6M. Besides, the test set of 3DHP contains three scenes: studio with green screen, studio without green screen, and outdoor.

3DPW [49] is the first in-the-wild dataset with accurate 3D annotation from inertial measurement units. 3DPW contains complex backgrounds and occlusions that are close to the real world. We use 3DPW as an unseen dataset to evaluate the generalization ability of real-world data.

Dataset combining. Sarandi [41, 43] *et al.* proved that combining large datasets could achieve high estimation performance. Unfortunately, they do not disclose the datasets and processing methods they use. We collect as many datasets as possible and create a large dataset including 1.38M examples (hereinafter called "1.38M dataset"). The details of 1.38M dataset is shown in Tab. 1.

Data augmentation. Following [43], we also apply the same data augmentations during training: geometric augmentations (scaling, rotation, translation, horizontal flip), synthetic occlusion, color distortion (brightness, contrast, hue, saturation), and background transformation.

Exp.	Backbone	Dataset	Batch	lr	GPU
Sec. 4.5	RV2	H36M	64	1×10^{-4}	TITAN X
Sec. 4.6, Sec. 4.8	RV1.5	H36M 3DHP	64	5×10^{-5}	RTX3090
Sec. 4.9	SV2	1.38M	18×2*	2×10^{-5}	RTX3090

* The batch is split and trained on two GPUs.

Table 2. Details of experimental settings (Batch=batch size, lr=learning rate, R=ResNet101, S=Swin).

As for background transformation, two datasets are considered. We use the same setting as [43] in Sec. 4.7, which uses INRIA Holidays [14]. In Sec. 4.9, we enhance the background augmentation by replacing INRIA Holidays [14] with the landscape-rich BG-20K [22] dataset and randomly cropping it to a size of 512×512 for each input image.

In addition, for learning the poses in different skeleton formats, the annotations are also copied and learned as SMPL [30] format annotations to facilitate learning of the SMPL format.

4.3. Implementation Details

Tensorflow 2.6.3 [1] is used for implementation. We adopt publicly released ResNet101 [12] and SwinV2 [29] for the backbone part. The weights are updated by AdamW [31] optimizer. According to the model's size, we experimentally select different batch sizes and initial learning rates for training each model. The learning rate is decayed exponentially in two parts as in [43]. For more specific settings, see Tab. 2.

4.4. Evaluation Metrics

To measure estimation performance in experiments, we use four evaluation metrics. Mean-Per-Joint-Position-Error (MPJPE) is a standard metric used in 3D pose estimation, which measures how well 3D human pose estimation performs based on the Euclidean Distance between predicted and GT body joints [13].

Procrustes-Aligned-Mean-Per-Joint-Position-Error (PA-MPJPE) is the MPJPE after rigid alignment by processing between the prediction pose and the ground truth pose to eliminate the effect of translation and rotation.

The Percentage of Correct Key points (PCK) is defined as the proportion of correct estimated points. The "correct" here refers to the distance between GT and prediction falling into a certain threshold [32]. The thresholds of PCK for 3DHP and 3DPW are 150mm and 50mm.

The Area Under Curve (AUC) calculates the average PCK through a range of threshold. For 3DHP, the thresh-

		MSH	+bone	+C1	+C2	+C1b	+C2b
root-relative		✓	✓	✓	✓	✓	✓
bone length		×	✓	×	×	✓	✓
bpo	sho-elb-wri	×	×	✓	✓	✓	✓
	neck-sho-elb	×	×	×	✓	×	✓
	hip-kne-ank	×	×	✓	✓	✓	✓
	pelv-hip-kne	×	×	×	✓	×	✓
MPJPE↓		61.3	57.8	59.6	58.8	**57.3**	58.0

Table 3. Ablation study of different loss combination on Human3.6M dataset (sho=sholder, elb=elbow, wri=wrist, kne=knee, ank=ankle). **Bold** denotes the best result.

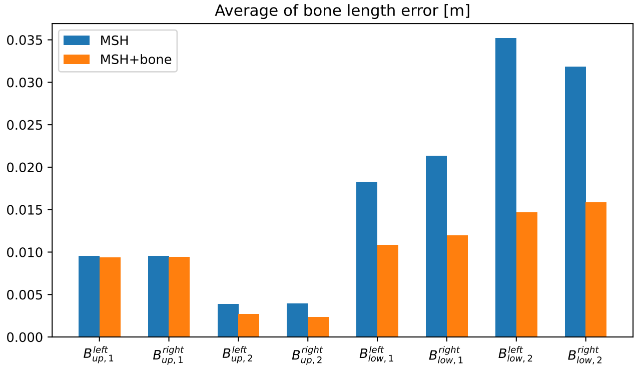


Figure 6. Effect of bone length loss on Human3.6M [13] dataset.

old range of AUC is 0mm to 150mm, and for 3DPW, the range is 0mm to 200mm.

4.5. Ablation Studies

We conduct ablation studies on two aspects: (1) the influence of the bone length loss; (2) the influence of the proposed body part orientation loss with the different combinations of planes. Here, we select the best results of each model to compare. Tab. 3 shows the overview of our ablation studies. This experiment set w_{bone} to 0.1 and w_{bpo} to 0.05.

Effect of the bone length loss. We compare the average bone length of each part on Human3.6M [13] dataset. Experimental results with and without the bone length loss are shown in Fig. 6. The result demonstrates that the estimated bone lengths are closer to the GT than the case without bone length loss. In particular, the impact on the lower body is greater than on the upper body. Also, MPJPE improved by 3.5, as shown in Tab. 3.

Effect of the body part orientation loss. We conduct experiments with two types of virtual plane combinations to demonstrate the validity and effect of the proposed body part orientation loss. (C1) one plane is considered on each side of the arms and legs; (C2) two planes are considered on both parts; see Tab. 3 for detailed plane combinations.

When the three loss functions (bone length loss, C1 loss, and C2 loss) are added independently to the relative root defect, they all positively affect the MSH model.

In the C1 vs. C1b and C2 vs. C2b comparisons, the scores with the addition of bone length loss exceeded without those. In particular, C1 improved MPJPE by 2.3. This indicates that adding not only bone length loss, but also body part posture loss provides a more accurate estimation of human body posture. In a comparison of C1b and C2b MPJPEs, the C1b score is 0.7 lower than the C2b score. This result suggests that more virtual planes are not necessarily better. This is because excessive constraints can confuse training.

Moreover, to analyze the impact of our proposed loss combination on each keypoint, we observe MPJPE for each keypoint as shown in Tab. 4. At the keypoints of the arms and legs (e.g., shoulders, wrists, ankles, knees, *etc.*) associated with the added virtual planes, MPJPE decreases significantly.

4.6. Validation on Different Datasets

We further implement our proposed method to MeTRAbs [43](see Fig. 5) and evaluated its accuracy on the Human3.6M [13] and 3DHP [32] datasets. This experiment set w_{bone} to 0.1 and w_{bpo} to 0.01.

The results are shown in Tab. 5a and Tab. 5b. On the Human3.6M dataset, MPJPE improves for the majority of activities. The average accuracy improvement is 0.7. On the 3DHP data set, 90% of activities show improvement in PCK, with improvements of 1.4 for PCK, 0.5 for AUC, and 2.0 for MPJPE total.

Visually, the estimated pose is more accurate with adding our proposed geometry loss combination, as illustrated in Fig. 7. In images where MeTRAbs is inaccurate in predicting limb orientation, our proposed method predicts limb orientation more accurately. The performance of both datasets demonstrates the effectiveness of the proposed methods. This also shows the generalization of our proposed geometry loss combination.

4.7. Validation on Different Models

Besides MeTRAbs [43], we also conduct experiments on ROMP [46] and Mesh Graphormer [26] to validate the generalization of our proposed geometry loss combination. ROMP and Mesh Graphormer are SMPL-based [30] models. ROMP predicts the SMPL map, which contains the 3D pose of the joints and the shape of the human mesh, and then regresses the joint locations through the SMPL model. Mesh Graphormer extracts the grid features and then feeds the tokenized features to a multi-layer Graphormer encoder to process a coarse mesh. After upsampling the coarse mesh, Mesh Graphormer predicts the joint locations and mesh vertices at the same time.

We conduct experiments with the settings following their

	rhip	rkne	rank	lhip	lkne	lank	tors	neck	head	htop	lsho	lelb	lwri	rsho	relb	rwri
MSH	**21.1**	57.2	97.7	**21.2**	56.5	103.9	**41.6**	50.5	**63.1**	**64.0**	62.9	**75.8**	**90.3**	61.4	80.7	94.0
MSH + C1b	23.5	**48.7**	**69.0**	21.5	**49.9**	**76.4**	42.6	**48.5**	65.2	68.3	**61.8**	78.2	90.4	**58.1**	**79.2**	**93.0**

Table 4. Quantitative comparision of MPJPE per keypoints (kne=knee, ank=ankle, tors=torso, sho=sholder, elb=elbow, wri=wrist). The initials r and l stand for right and left, respectively. **Bold** denotes the best result.

	Dir.	Dis.	Eat	Gre.	Phn.	Pose	Pur.	Sit	SitD	Sm.	Pht.	Wait	Walk	WD	WT	Avg↓
									MPJPE↓							
MeTRAbs (w/o GLC)	**51.3**	**53.7**	56.5	60.2	55.9	65.7	**60.6**	55.2	56.6	52.9	64.3	63.3	56.7	65.8	61.7	57.8
MeTRAbs (w GLC)	51.6	54.1	**55.8**	**59.3**	**55.3**	**65.5**	61.0	**54.8**	**54.9**	**52.3**	**63.0**	**61.8**	**55.4**	**63.8**	**60.1**	**57.1**

(a) Evaluation results on Human3.6M [13] dataset.

	Stand /Walk	Exer -cize	Sit on Chair	Cro./ reach	On floor	Sport	Misc.	Green screen	No gr.sc.	out-door	Total		
					PCK↑						PCK↑	AUC↑	MPJPE↓
MeTRAbs (w/o GLC)	91.8	88.6	87.8	87.8	**71.8**	90.7	88.4	92.2	85.6	81.6	87.3	49.1	87.4
MeTRAbs (w GLC)	**93.1**	**89.1**	**88.9**	**91.4**	71.7	**92.0**	**89.6**	**93.0**	**86.3**	**85.0**	**88.7**	**49.6**	**85.4**

(b) Evaluation results on MPI-INF-3DHP [32] dataset.

Table 5. Validation on different datasets. **Bold** denotes the best result. The MeTRAbs [43] result above is our re-trained version.

Method	GLC (ours)	MPJPE↓	
		Human3.6M [13]	3DPW [49]
ROMP [46]	×	–.–	75.7
	✓	–.–	**75.3**
Mesh Graphormer [26]	×	56.0	–.–
	✓	**55.0**	–.–
MeTRAbs [43]	×	57.8	69.7
	✓	**57.1**	**64.8**

Table 6. Ablation study of different models w and w/o our geometry loss combination (GLC=geometry loss combination). **Bold** denotes the best result.

papers [26,43,46]. The results are shown in Tab. 6. The proposed method shows improved performance for all models: 1.0 for Mesh Graphormer and 0.7 for MeTRAbs on the Human3.6M [13] dataset and 0.4 for ROMP and 4.9 for MeTRAbs on the 3DPW [49] dataset for MPJPE. This proves the effectiveness of our proposed method on different models.

4.8. Comparison to Prior Geometry Loss Functions

We compare our proposed methods to two existing loss functions: symmetry loss and illegal-angle loss [7]. These loss functions are most relevant to our proposed method because they provide constraints based on body geometry. We re-implement these two loss functions and train them on the Human3.6M [13] dataset with MeTRAbs. For a fair comparison, the L1 loss function is used as the loss function to calculate the symmetry of the left and right bone lengths. The results are shown in Tab. 7.

		E1	E2	E3	E4	E5	E6	E7
ours	root-relative	✓	✓	✓	✓	✓	✓	✓
	bone length	×	✓	×	✓	×	×	×
	bpo (C1)	×	×	✓	✓	×	×	×
[7]	symmetry	×	×	×	×	✓	×	✓
	illegal-angle	×	×	×	×	×	✓	✓
	MPJPE↓	57.8	57.5	**57.1**	**57.1**	58.4	57.4	57.6

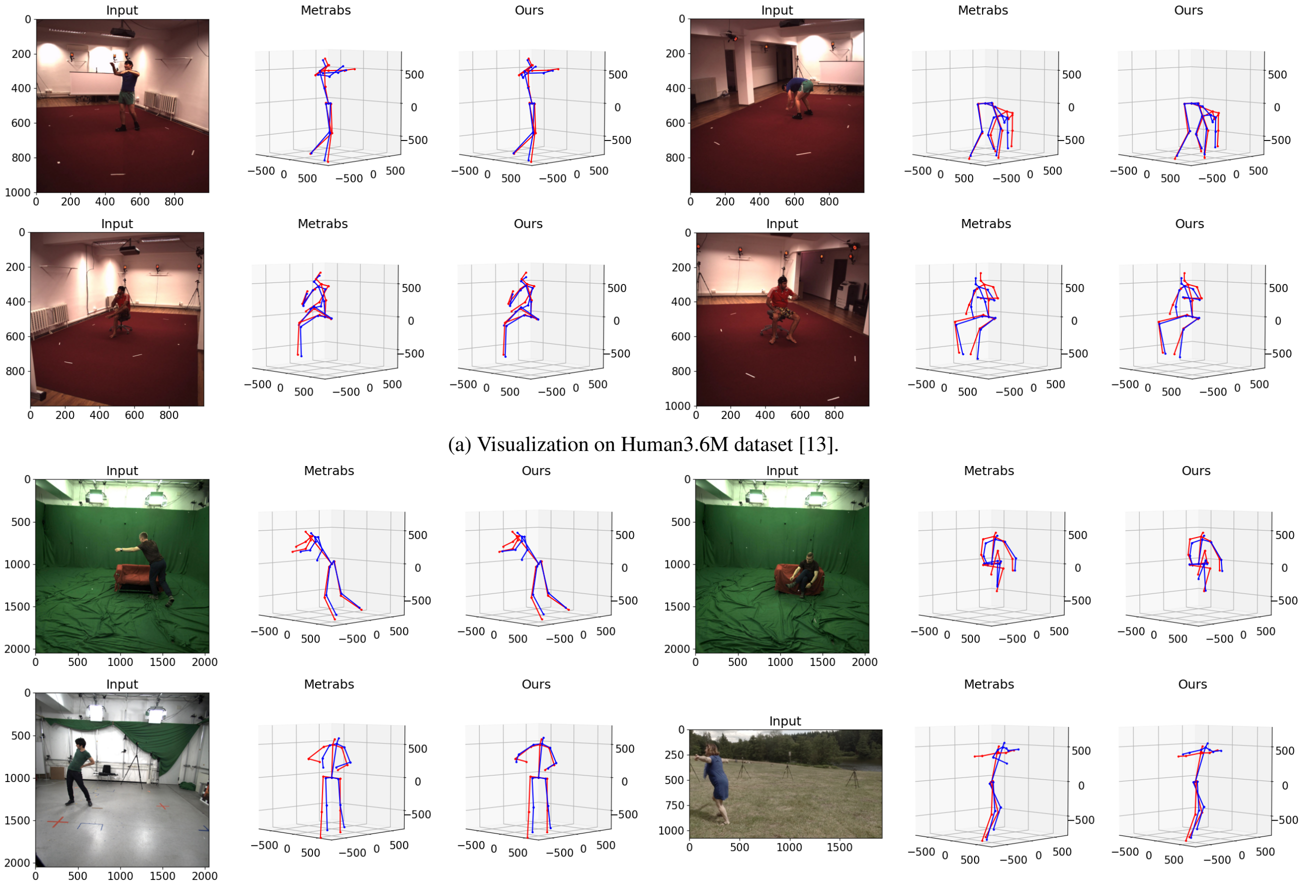
Table 7. Comparison to prior geometry loss functions. **Bold** denotes the best result.

We observe that our proposed method performs better than these methods. In the E2 vs. E5, our bone length loss is 0.3 MPJPE better than E1, while E5 is lower than E1. In the E3 vs. E6, our body part orientation loss achieves better 0.3 MPJPE than illegal-angle loss. In the E4 vs. E7, our proposed geometry loss combination achieves the best 57.1 MPJPE, while the combination of symmetry loss and illegal-angle loss achieves 57.6 MPJPE. This comparison further proves the validity of our proposed method.

4.9. Comparison to the State-of-the-Art Methods

We compare against the nine state-of-the-art 3D human pose estimation methods, such as CLIFF [24], DynaBOA [10], and Dozens-M [41] *etc*. As shown in Tab. 8, the comparison only uses the 3DPW [49] dataset for evaluation. This experiment set w_{bone} to 0.1 and w_{bpo} to 0.01.

In the protocol #PS, adding the proposed method improves MPJPE by about 1.5 over the case without adding it. The result of 65.0 MPJPE is also above the other methods. Besides, some methods [24, 26] use the training set of 3DPW for training or fine-tuning. In contrast, we did not



(a) Visualization on Human3.6M dataset [13].

(b) Visualization on MPI-INF-3DHP dataset [32].

Figure 7. Visualization results for the validation in Sec. 4.6. Red indicates the GT pose, blue indicates the predicted pose.

Method	PA-MPJPE↓	MPJPE↓
Protocol of SPIN [18] (#PS)		
SPIN [18]	59.2	96.9
HybrIK [21]	48.8	80.0
METRO [25]	47.9	77.1
Mesh Graphormer [26]	45.6	74.7
CLIFF [24]	43.0	69.0
Cha *et al.* [4]	**39.0**	66.0
DynaBOA [10]	40.4	65.5
MeTRAbs+ (w/o GLC)	48.0	66.5
MeTRAbs+ (w GLC)	47.4	**65.0**
Protocol of MeTRAbs [43] (#PM)		
MeTRAbs [43]	49.7	68.8
Dozens-M [41]	45.6	64.3
MeTRAbs+ (w/o GLC)	45.9	64.3
MeTRAbs+ (w GLC)	**45.2**	**62.8**

Table 8. Comparison to the state-of-the-art methods on 3DPW dataset. **Bold** denotes the best result.

use any data from 3DPW while achieving a better result.

In the protocol #PM, MeTRAbs+ trained with the proposed geometry loss combination obtains the best results with 62.8 MPJPE. Training on the smaller level of data, our results outperform the state-of-the-art Dozens-M [41].

These experimental results reveal that our proposed geometry loss combination improves the estimation performance. This also demonstrates the effectiveness and generality of our loss combination in the wild datasets.

5. Conclusion

We proposed a simple and effective geometry loss combination to improve the prediction for keypoints at the end of limbs. This combination contains three loss functions: root-relative loss, bone length loss and body part orientation loss. Experimental results proved the effectiveness and importance of geometry constrain for 3D human pose estimation. Furthermore, our proposed loss combination is general and can be easily applied to various related models. Extensive experiments demonstrate that our methods achieve state-of-the-art performance on the 3DPW dataset.

References

[1] Martín Abadi et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, et al. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, June 2014.

[3] Cristian Sminchisescu Catalin Ionescu, Fuxin Li. Latent structured models for human pose estimation. In *ICCV*, 2011.

[4] Junuk Cha, Muhammad Saqlain, GeonU Kim, Mingyu Shin, and Seungryul Baek. Multi-person 3d pose and shape estimation via inverse kinematics and refinement. In *ECCV*, pages 660–677, 2022.

[5] Yucheng Chen, Yingli Tian, and Mingyi He. Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding*, 192:102897, 2020.

[6] Mickael Cormier, Aris Clepe, Andreas Specker, et al. Where are we with human pose estimation in real-world surveillance? In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 591–601, 2022.

[7] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3d human pose from structure and motion. In *ECCV*, September 2018.

[8] Matteo Fabbri, Fabio Lanzi, Simone Calderara, et al. Compressed volumetric heatmaps for multi-person 3d pose estimation. In *CVPR*, June 2020.

[9] Mihai Fieraru, Mihai Zanfir, Silviu Cristian Pirlea, et al. Aifit: Automatic 3d human-interpretable feedback models for fitness training. In *CVPR*, pages 9914–9923, 2021.

[10] Shanyan Guan, Jingwei Xu, Michelle Zhang He, Yunbo Wang, Bingbing Ni, and Xiaokang Yang. Out-of-domain human mesh reconstruction via dynamic bilevel online adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):5070–5086, 2023.

[11] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Gerard Pons-Moll, and Christian Theobalt. In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In *CVPR*, June 2019.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[13] Catalin Ionescu, Dragos Papava, Vlad Olaru, et al. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014.

[14] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, pages 304–317, 2008.

[15] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, pages 12.1–12.11, 2010.

[16] Hanbyul Joo, Tomas Simon, Xulong Li, et al. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[17] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, et al. End-to-end recovery of human shape and pose. In *CVPR*, June 2018.

[18] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, October 2019.

[19] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, June 2019.

[20] Jiefeng Li, Can Wang, Hao Zhu, et al. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *CVPR*, June 2019.

[21] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *CVPR*, pages 3383–3393, June 2021.

[22] Jizhizi Li, Jing Zhang, Stephen J Maybank, et al. Bridging composite and real: Towards end-to-end deep image matting. *International Journal of Computer Vision*, 2022.

[23] Ruilong Li, Shan Yang, David A. Ross, et al. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*, pages 13401–13412, October 2021.

[24] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, pages 590–606, 2022.

[25] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, pages 1954–1963, June 2021.

[26] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *ICCV*, pages 12939–12948, October 2021.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, et al. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

[28] Wu Liu, Qian Bao, Yu Sun, et al. Recent advances of monocular 2d and 3d human pose estimation: A deep learning perspective. *ACM Comput. Surv.*, 55(4), nov 2022.

[29] Ze Liu, Han Hu, Yutong Lin, et al. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*, pages 12009–12019, June 2022.

[30] Matthew Loper, Naureen Mahmood, Javier Romero, et al. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015.

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2017.

[32] Dushyant Mehta, Helge Rhodin, Dan Casas, et al. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017.

[33] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, et al. Single-shot multi-person 3d pose estimation

from monocular rgb. In *3D Vision (3DV), 2018 Sixth International Conference on*. IEEE, sep 2018.

[34] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, et al. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Trans. Graph.*, 36(4), jul 2017.

[35] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Neuralannot: Neural annotator for 3d human mesh training sets. In *CVPRW*, pages 2299–2307, June 2022.

[36] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, et al. Agora: Avatars in geography optimized for regression analysis. In *CVPR*, pages 13468–13478, June 2021.

[37] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, et al. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *CVPR*, July 2017.

[38] Dario Pavllo, Christoph Feichtenhofer, David Grangier, et al. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, June 2019.

[39] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, et al. Geonet: Geometric neural network for joint depth and surface normal estimation. In *CVPR*, pages 283–291, 2018.

[40] Alessio Sampieri, Guido Maria D'Amely di Melendugno, Andrea Avogaro, et al. Pose forecasting in industrial human-robot collaboration. In *ECCV*, pages 51–69, 2022.

[41] István Sárándi, Alexander Hermans, and Bastian Leibe. Learning 3D human pose estimation from dozens of datasets using a geometry-aware autoencoder to bridge between skeleton formats. In *WACV*, 2023.

[42] István Sárándi, Timm Linder, Kai O. Arras, et al. Metric-scale truncation-robust heatmaps for 3D human pose estimation. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2020.

[43] István Sárándi, Timm Linder, Kai O. Arras, et al. MeTRAbs: metric-scale truncation-robust heatmaps for absolute 3D human pose estimation. *IEEE Transactions on Biometrics, Behavior, and Identity Science (T-BIOM)*, 3(1):16–30, 2021.

[44] Xiao Sun, Jiaxiang Shang, Shuang Liang, et al. Compositional human pose regression. In *ICCV*, Oct 2017.

[45] Xiao Sun, Bin Xiao, Fangyin Wei, et al. Integral human pose regression. In *ECCV*, pages 536–553, 2018.

[46] Yu Sun, Qian Bao, Wu Liu, et al. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, pages 11179–11188, October 2021.

[47] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, et al. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Proc. of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, Delft, Netherlands, Nov. 2019.

[48] Gul Varol, Javier Romero, Xavier Martin, et al. Learning from synthetic humans. In *CVPR*, July 2017.

[49] Timo von Marcard, Roberto Henschel, Michael Black, et al. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, sep 2018.

[50] Wei Yin, Yifan Liu, and Chunhua Shen. Virtual normal: Enforcing geometric constraints for accurate and robust depth prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7282–7295, 2022.

[51] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, et al. Humbi: A large multiview dataset of human body expressions. In *CVPR*, June 2020.

[52] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *CVPR*, June 2020.

[53] Weichen Zhang, Zhiguang Liu, Liuyang Zhou, et al. Martial arts, dancing and sports dataset. *Image Vision Comput.*, 61(C):22–39, may 2017.

[54] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: A weakly-supervised approach. In *ICCV*, Oct 2017.