Supplemental Document: In the Wild Human Pose Estimation using Explicit 2D Features and Intermediate 3D Representations

1. Details of the loss functions used for training

Here we describe the loss functions used to train our neural network.

Given an input image $\mathbf{I} \in \mathbb{R}^{w \times h \times 3}$, the extractor network f_{RGB} will predict the features $\mathbf{F}_{3D}$ which consist of the explicit 2D pose features $\mathbf{h}_{2D}$ and additional pose cues $\mathbf{d}$ as feature maps. The predicted 2D pose features are defined as 2D per-joint heatmaps [2]

$$\begin{aligned} \mathbf{h}_{2D} &= (\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_K), \\ \text{where } \mathbf{m}_k &\in \mathbb{R}^{\frac{w}{s} \times \frac{h}{s}}. \end{aligned}$$

where $s = 16$ is the heatmap down-sampling factor.

Similarly, the ground truth heatmaps are defined as

$$\begin{aligned} \mathbf{h}^{GT} &= (\mathbf{m}_1^{GT}, \mathbf{m}_2^{GT}, \ldots, \mathbf{m}_K^{GT}), \\ \text{where } \mathbf{m}_k^{GT} &\in \mathbb{R}^{\frac{w}{s} \times \frac{h}{s}}. \end{aligned}$$

To train the 2D pose features, we minimize the difference between the predicted 2D joint heatmaps and the ground truth maps using an L2 loss

$$\mathcal{L}_{2Dheatmap} = \sum_{k=1}^{K} b_k \parallel \mathbf{m}_k - \mathbf{m}_k^{GT} \parallel_2^2, \tag{1}$$

where $b_k \in \{0, 1\}$ is a binary mask to ensure that the objective is not evaluated if the annotation of a particular joint is not available.

The latent features $\mathbf{F}_{3D}$ are then used to predict the 3D pose $\mathbf{P}_{3D} \in \mathbb{R}^{21 \times 3}$ by the sub-network f_{3D}. Given a 3D pose annotation $\mathbf{P}^{GT} \in \mathbb{R}^{21 \times 3}$, the 3D joint position loss is calculated as follows

$$\mathcal{L}_{3Dpose} = \parallel \mathbf{P}_{3D} - \mathbf{P}^{GT} \parallel_2^2 . \tag{2}$$

Given that $Parent(\mathbf{J}_k)$ is the position of the parent of a joint $\mathbf{J}_k \in \mathbb{R}^3$ in the kinematic chain, when the 3D joint position ground truth $\mathbf{J}_k^{GT} \in \mathbb{R}^3$ is available, the bone during training is defined as

$$\begin{aligned} \mathcal{L}_{bone} = \sum_{k=1}^{K} \parallel \Big(Parent(\mathbf{J}_k) - \mathbf{J}_k \Big) \\ - \Big(Parent(\mathbf{J}_k^{GT}) - \mathbf{J}_k^{GT} \Big) \parallel_2^2 . \end{aligned} \tag{3}$$

On the other hand, if we train on data for which only 2D joint annotations, but no 3D annotations are available, then we instead only compare the bone length magnitude between the predicted joint with a bone length $\mathbf{J}_k^S$ randomly selected from a training annotation

$$\begin{aligned} \mathcal{L}_{bone} = \sum_{k=1}^{K} \Big\| \parallel Parent(\mathbf{J}_k) - \mathbf{J}_k \parallel_2^2 \\ - \parallel Parent(\mathbf{J}_k^S) - \mathbf{J}_k^S \parallel_2^2 \Big\|_2^2 . \end{aligned} \tag{4}$$

Finally, given a predicted 2D pose from the projection layer $\mathbf{p}_{2D}$ (Eq. 2 in the main document) and its corresponding ground truth 2D joint coordinates in the image space $\mathbf{p}_{2D}^{GT}$, the projection loss is defined as

$$\mathcal{L}_{2Dpose} = \parallel \mathbf{p}_{2D} - \mathbf{p}_{2D}^{GT} \parallel_2^2 . \tag{5}$$

The final training loss can be expressed as

$$\begin{aligned} \mathcal{L}_{all} = \lambda_{2Dheatmap}\mathcal{L}_{2Dheatmap} + \lambda_{3Dpose}\mathcal{L}_{3Dpose} \\ + \lambda_{bone}\mathcal{L}_{bone} + \lambda_{2Dpose}\mathcal{L}_{2Dpose}, \end{aligned} \tag{6}$$

where $\lambda_{3Dpose} = 10$, $\lambda_{2Dheatmap} = 0.1$, $\lambda_{2Dpose} = 10$. $\lambda_{bone} = 10$ if the bone direction is considered (i.e. 3D pose annotations are given) and $\lambda_{bone} = 100$ if we only estimate the bone length scalar (i.e. only 2D annotations are given).

2. Additional comparisons on MPI-INF-3DHP

At some point in the past, the authors of MPI-INF-3DHP released a correction to the ground truth annotations of a subset of two their six test sequences. For all our tests, we used the corrected data.

Their very first version of the test set contained small errors on the in-studio sequences with general, i.e. no green screen, background (test subject 3 and 4, meaning sequences labelled **No GS** in the paper and this document).

On these sequences, before correction, the annotations were temporally misaligned by one or two frames.

It is hard for us to say what previous paper we compared against may have unknowingly used the uncorrected subset of sequences.

For our tests to be as transparent and fair as possible, we therefore also provide a comparison on the subset of 4 out of 6 MPI-INF-3DHP test sequences (**GS** and **Outdoors**) that were always correct.

Method	3D training data	PCK GS	PCK Outdoor	PCK All	AUC All	MPJPE All
Mehta *et al.* [1]	H3.6M + MPI-INF-3DHP	84.6	69.7	78.8	-	-
Mehta *et al.* [1]	H3.6M	70.8	58.5	66.0	-	-
Zhou *et al.* [3]	H3.6M	71.1	72.7	71.7	-	-
Ours (*unscaled*)	H3.6M + MPI-INF-3DHP	87.8	73.8	82.3	45.3	91.4
Ours (*unscaled*)	H3.6M	74.6	64.0	70.5	36.3	128.7
Ours (*glob. scaled*)	H3.6M (sampled at 5 fps)	75.4	66.9	72.1	37.2	125.5
Ours (*glob. scaled*)	H3.6M + MPI-INF-3DHP	88.0	74.8	82.9	45.6	91.8
Ours (*glob. scaled*)	H3.6M	75.2	65.3	71.4	36.9	131.4
Ours (*glob. scaled*)	H3.6M (sampled at 5 fps)	75.8	67.9	72.8	37.8	128.6
Ours (*Procrustes*)	H3.6M + MPI-INF-3DHP	94.9	84.0	90.7	58.0	66.1
Ours (*Procrustes*)	H3.6M	85.9	78.8	83.2	46.6	91.1
Ours (*Procrustes*)	H3.6M (sampled at 5 fps)	86.2	78.0	83.0	47.5	89.6

Table 1: Comparison on the subset of MPI-INF-3DHP test sequences that was not corrected at some point by the authors of MPI-INF-3DHP (GS and Outdoors). All here refers to the average on this subset of sequences. Unless stated otherwise, all H3.6M training data mentioned in this table use H80K samples.

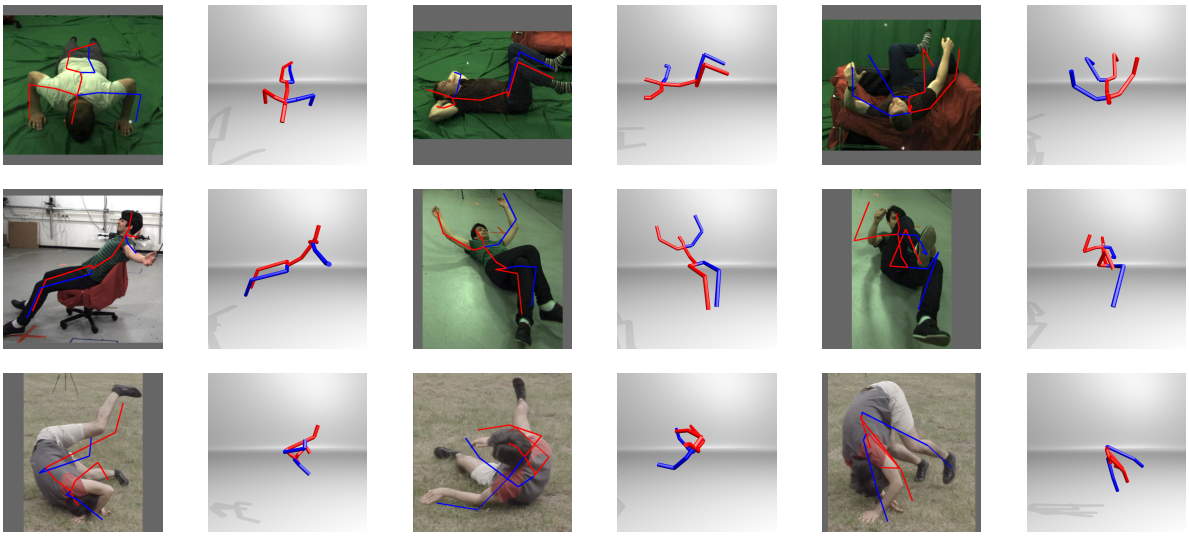

Figure 1: Prediction failure examples by our proposed method on different scenes. Each row from top to bottom represents studio green screen, studio non-green screen, and outdoor scenarios respectively.

Table 1 shows the comparison for methods trained on both H3.6M and MPI-INF-3DHP using the mentioned subset for testing. We include methods that in their original papers reported the respective results on the subsets of test sequences, too. Again, all evaluations of our method in the main paper and here are performed with the corrected annotations. Our proposed method is also state-of-the-art when tested on this subset of sequences.

We also show the activity-wise performance of our method tested on MPI-INF-3DHP in Table 2 respectively. Our method achieves a very high 3D PCK of more than 80% on almost all categories, except for the on-the-floor activities (60.7%), which are in general also challenging for other methods.

3. Failure cases

We show additional pose prediction failure cases on different scenes (studio green screen, studio without green screen, outdoor) in Figure 1.

4. Additional results

We show additional qualitative results on the MPI-INF-3DHP test set in Figure 2 and the LSP test set in Figure 3. Our approach captures even difficult 3D poses well from a single color image. For more qualitative results please refer to the accompanying video.

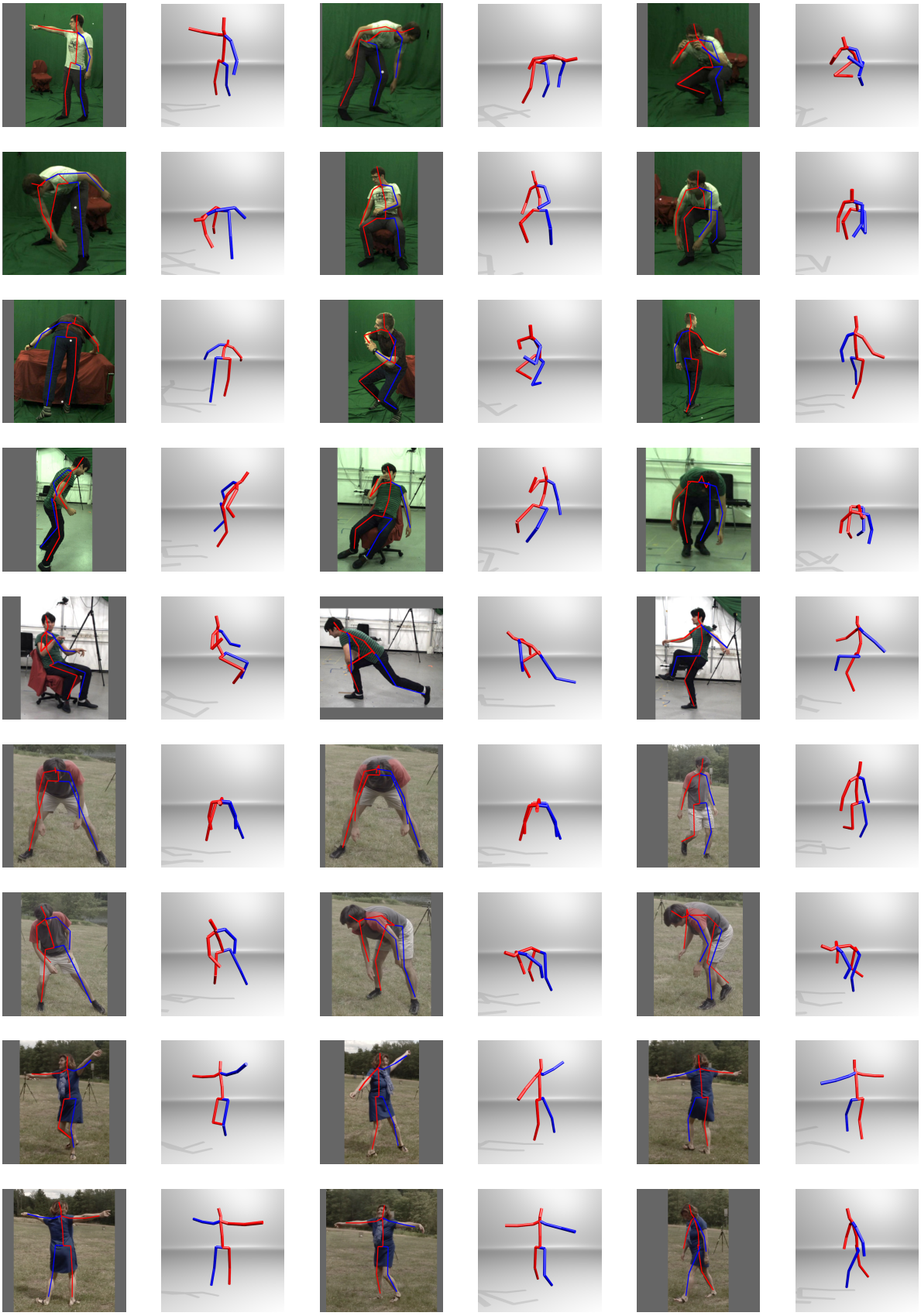

Figure 2: Additional qualitative examples of applying our method to the MPI-INF-3DHP test set.

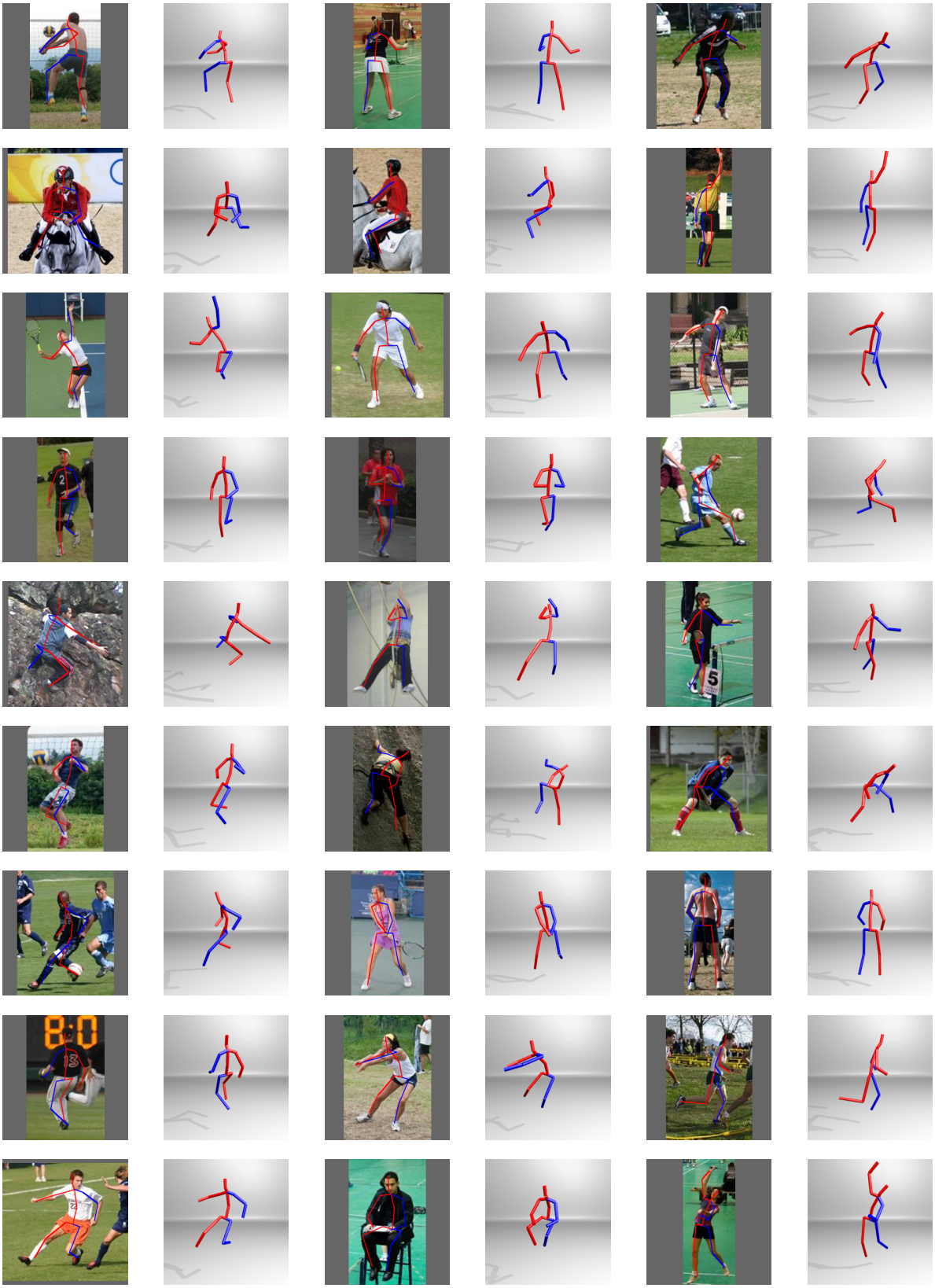

Figure 3: Additional qualitative examples of applying our method on the LSP test set.

Action	PCK									AUC
	Head	Neck	Shou	Elbow	Wrist	Hip	Knee	Ankle	Total	
Standing/Walking	93.2	100.0	99.6	89.8	74.3	100.0	90.0	77.3	89.7	51.2
Exercising	91.3	98.2	98.2	87.6	75.6	100.0	77.6	65.5	85.6	47.2
Sitting	81.7	92.8	91.8	76.7	65.1	99.8	75.8	63.9	80.0	43.7
Reaching/Crouching	76.6	91.1	91.3	83.3	78.0	98.7	84.2	73.2	84.6	47.6
On The Floor	62.8	83.9	78.9	54.7	40.9	94.6	53.9	28.6	60.7	28.5
Sports	90.0	99.2	98.7	84.9	67.8	100.0	90.6	72.4	87.0	49.3
Miscellaneous	80.8	96.8	95.3	71.3	53.8	100.0	86.5	66.9	80.4	43.4
All	82.3	94.9	93.7	78.0	64.5	99.3	81.2	65.5	81.5	44.7

Table 2: Activity-wise 3D PCK of our method on the MPI-INF-3DHP test set. Our method achieved more than 80% 3D PCK in most actions except for the challenging on-the-floor examples (60.7% 3D PCK).

References

[1] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. 2

[2] J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, Cambridge, MA, USA, 2014. MIT Press. 1

[3] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Towards 3d human pose estimation in the wild: A weakly-supervised approach. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2