It's all Relative: Monocular 3D Human Pose Estimation from Weakly Supervised Data

Matteo Ruggero Ronchi*, Oisin Mac Aodha*, Robert Eng, Pietro Perona
{mronchi,macaodha,reng,perona}@caltech.edu
California Institute of Technology

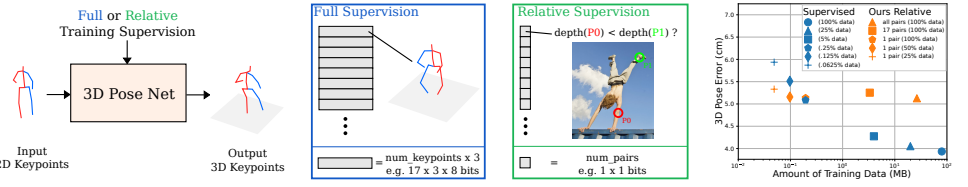
Abstract

We address the problem of 3D human pose estimation from 2D input images using only weakly supervised training data. Despite showing considerable success for 2D pose estimation, the application of supervised machine learning to 3D pose estimation in real world images is currently hampered by the lack of varied training images with associated 3D poses. Existing 3D pose estimation algorithms train on data that has either been collected in carefully controlled studio settings or has been generated synthetically. Instead, we take a different approach, and propose a 3D human pose estimation algorithm that only requires relative estimates of depth at training time. Such training signal, although noisy, can be easily collected from crowd annotators, and is of sufficient quality for enabling successful training and evaluation of 3D pose. Our results are competitive with fully supervised regression based approaches on the Human3.6M dataset, despite using significantly weaker training data. Our proposed approach opens the door to using existing widespread 2D datasets for 3D pose estimation by allowing fine-tuning with noisy relative constraints, resulting in more accurate 3D poses.

1 Introduction

Reasoning about the pose of humans in images and videos is a fundamental problem in computer vision and robotics. To ensure that future autonomous systems are safe to interact with, they need to be able to understand not only the positions but also the poses of the people around them. Recent success in 2D pose estimation has been driven by larger, more varied, labeled datasets. While laborious, it is possible for human annotators to click on the 2D locations of different body parts to generate such training data. Unfortunately, in the case of 3D pose estimation, it is much more challenging to acquire large amounts of training data containing people in real world settings with their corresponding 3D poses. This lack of large scale training data makes it difficult to both train deep models for 3D pose estimation and to evaluate the performance of existing methods in situations where there are large variations in scene types and poses. As a result, researchers have resorted to various alternative methods for collecting 3D pose training data - including motion capture, synthetic datasets, video, and multi-camera setups. In this work, we argue that instead of using additional hardware to acquire full 3D ground truth training data from closed settings, Fig. 1 (b), we can make use of human annotated relative depth information from images in the wild, Fig. 1 (c).

*These authors contributed equally to this work.



(a) Lifting 3D Pose Estimation (b) Full 3D Labels (c) Relative Labels (d) Train Data vs Error

Figure 1: (a) Lifting based methods take a set of 2D keypoints as input and predict their 3D position. (b) This is achieved using ground truth 3D poses during training. (c) We show that weak supervision specifying the relative depth of as little as one pair of keypoints per image, can be used to train effective 3D pose estimation algorithms. (d) Even with small amounts of training data our model still predicts accurate poses compared to using full 3D supervision.

Our main contributions are: (1) a loss for 3D pose estimation of articulated objects that can be trained on sparse and easy to collect relative depth annotations with performance comparable to the state of the art, (2) an empirical evaluation of the ability of crowd annotators to provide relative depth supervision in the context of human poses, and (3) a dataset of relative joint depth annotations that can be used for both training and evaluation purposes.

2 Related Work

2D Pose Estimation: Current state of the art methods for 2D human keypoint estimation are based on deep networks trained on large quantities of supervised data [6, 11, 28, 45, 50]. In addition to more sophisticated network architectures, a large driver in the improved prediction accuracy of these approaches is the increase in the size and complexity of datasets that contain images with corresponding keypoint annotations indicating the locations, in pixels, of specific body parts *e.g.* [2, 14, 18, 38]. While suitable for 2D pose estimation, by and large, most existing 2D datasets do not contain any supervised signal for 3D pose estimation. In this work we show that existing 2D pose datasets can indeed be used for 3D pose estimation by augmenting them with relative depth annotations collected from crowd annotators.

3D Pose Estimation: There exist two main categories of methods for 3D pose estimation: (1) end-to-end models and (2) lifting based approaches. The first set of models take a 2D image as input, where a person detector has first been run on the image, and then produce an estimate of the individual's 3D pose as output. This is achieved by learning to regress the 3D keypoint locations during training, either as a set of 3D coordinates [17] or as volumetric heat maps [30]. These methods assume the availability of a training set of 2D images paired with corresponding 3D annotations. To further constrain the problem, it is possible to enforce a strong prior on the predictions in the form of a parameterized model of human body shape [3, 20, 42, 44]. While this ensures realistic looking outputs [16], it can be limiting if the prior is not flexible enough to cover the full range of plausible poses.

As an alternative, lifting based approaches take a set of predicted 2D keypoints as input and *lift* them into 3D. The rise in popularity of these methods is driven by two factors: (1) the 2D location of keypoints is a strong cue regarding their 3D configuration and (2) the limited number of 'in the wild' datasets featuring *paired* 2D images with 3D poses. A variety of lifting approaches have been proposed that either frame the problem as one of regression [24, 27], or use data driven retrieval [2, 35, 52], dictionary based reconstruction [1], or generative adversarial learning [51].

Instead of requiring full 3D pose information for each individual in an input image, we propose a method that only needs a small amount of sparse data indicating the relative depth of different body parts. This results in high quality predictions with as little as one relative depth constraint per pose at training time.

3D Pose Training Data: A major factor holding back progress in 3D pose estimation is the lack of *in the wild* datasets featuring images with ground truth 3D pose information. Most existing 3D pose datasets feature single individuals captured in controlled studio settings [13, 40] and are challenging to acquire due to the need for specialized equipment such as motion capture cameras and markers. Setups with multiple cameras make it easier to capture small numbers of interacting people [15, 26], but require multiple synchronized cameras in confined spaces to produce accurate depth. Depth cameras can be used to generate 3D training data [39], but are usually limited to the indoors. Other alternatives use additional equipment such as inertial sensors [32] or passive markers [49]. The main limitation of these setups is that it is very difficult to also capture 2D images that cover all the variation in appearance that one would encounter in real world, non-studio, settings.

One technique to amplify studio captured motion capture data is to use computer graphics techniques to generative potentially unlimited amounts of synthetic training data [9, 46]. Synthetic training data has been successful for low-level vision tasks such as depth and optical flow estimation [25]. However, rendering realistic environments featuring people interacting with others and their surroundings is a challenging problem. Furthermore, even if it is possible to successfully generate plausible scenes, these methods are still limited by the variation in pose and the range of subject interactions that are present in the input motion capture data. Different graphical user interfaces have been explored to allow crowd annotators to annotate 3D pose information in existing image datasets. Examples include, manually configuring 3D skeletons [29] or providing coarse body part orientation information [2, 21, 22]. These can be very laborious tasks and take a large amount of time per image.

Relative Depth Supervision: Using ordinal relations between a sparse set of point pairs as a weak form of supervision has been previously explored in the context of dense depth estimation in monocular images [8, 54]. While this is not comparable to metric ground truth depth, it enables the easy collection of data that can be used for both training and evaluating monocular depth estimation. This type of data has also been used for 3D pose estimation [4, 41]. However, unlike previous work that require complete annotations of all joint pairs to infer the 3D pose, we use only sparse relative annotations. Our annotations, Fig. 1 (c), are binary labels specifying the relative distance between keypoints and the camera. While also [33] crowdsources sparse pairwise pose constraints, they use them as a replacement for 2D keypoints, and in their case the relative relationships are in the coordinate frame of the person in the image. We show that depth annotations relative to the camera are easy to collect and can be combined with existing 2D keypoint annotations for improving 3D pose estimation.

Parallel to our work, [31] also explored training 3D pose estimation models using relative depth constraints as a form of supervisory signal. However, they use multiple ordinal annotations per image as a form of data augmentation for 3D supervised baselines, while we use as little as one relative pair per image and focus on the setting where no 3D supervised ground truth is available, and do not perform calibration at test time. Finally, we also conduct a detailed user study evaluating how accurate human annotators are at providing relative depth annotations in the context of human pose.

3 Method

Our goal is to predict the 3D pose of an individual depicted in an input image. We represent pose in 2D as a set of coordinates $\mathbf{p} \in \mathbb{R}^{2\times J}$, where each element $\mathbf{p}_j = [u_j, v_j]$ is a row vector that encodes the location, in pixels, of one of J different joints. For each $\mathbf{p}$, we aim to infer its position in 3D $\mathbf{P} \in \mathbb{R}^{3\times J}$, where each entry specifies the location of the joint j in 3D, $\mathbf{P}_j = [x_j, y_j, z_j]$. In this work we take inspiration from lifting based approaches, Fig. 1 (a), and attempt to learn the parameters of a function $f : \mathbb{R}^{2\times J} \rightarrow \mathbb{R}^{3\times J}$, that maps 2D input keypoints to their estimated 3D position, where $f(\mathbf{p}) = \hat{\mathbf{P}}$. We parametrize f as a neural network, where the input joint positions $\mathbf{p}$ can come from the output of a 2D human pose estimation algorithm *e.g.* [6, 28, 50].

3.1 Supervised 3D Pose Estimation

Given a set of N input 2D keypoints and their corresponding ground truth 3D pose one could use a supervised loss to train f

$$\mathcal{L}_{sup}(\hat{\mathbf{P}}, \mathbf{P}) = ||\hat{\mathbf{P}} - \mathbf{P}||_2. \tag{1}$$

This is the approach taken in [24], where a neural network, f, is trained to project the input coordinates $\mathbf{p}$ into 3D. While they only need to infer the missing z_j values for each 2D keypoint, their model predicts each $[x_j, y_j, z_j]$ coordinate, making the approach more robust to small errors in the input locations.

3.2 3D Pose Estimation with Relative Constraints

As noted earlier, acquiring large quantities of varied ground truth 3D pose data is challenging. Instead, we opt to use much weaker supervision in the form of depth ordering labels that describe the relative distance to the camera for a pair of keypoints, see Fig. 1 (c).

We assume that we have access to a set of crowdsourced relative annotations for an image i, $\mathcal{A}^i = \{(j_1, k_1, r_1), (j_2, k_2, r_2), ..., (j_A, k_A, r_A)\}$, where each annotation $\{j, k, r\}$ is a tuple specifying the joints j and k and their estimated relative depth $r \in \{-1, 1, 0\}$. The number of specified pairwise constraints, A, can be different for every image, and varies between one and $\binom{J}{2}$, when the ordinal supervision is provided for every pair of keypoints. The value $r = -1$ indicates that $z_j < z_k + \varepsilon$ (joint j is closer to the camera compared to k), while $r = 1$ specifies that $z_k < z_j + \varepsilon$. If the distance between the two keypoints is below a certain tolerance ε, then $r = 0$. In practice, this corresponds to the case in which human annotators cannot disambiguate the relative position of two keypoints. For all the experiments, except the ablation study in Table 1, we explore the setting where $r \in \{-1, 1\}$.

Similar to [8], we use a pairwise ranking loss to encourage our model to predict the correct depth ordering of a 3D keypoint pair

$$\mathcal{L}_{rel}(\hat{\mathbf{P}}, \mathcal{A}) = \sum_{(j,k,r)\in\mathcal{A}} \begin{cases} \log(1 + \exp(-r\hat{d}_{jk})), & r = -1, +1 \\ ||\hat{d}_{jk}||_2, & r = 0, \end{cases} \tag{2}$$

where $\hat{d}_{jk} = \lambda(\hat{z}_j - \hat{z}_k)$, with $\hat{z}_j$ is the predicted depth from our network for keypoint j and λ controls the strength of the loss. In practice, we found that it is important to normalize the range of depth values $\hat{d}$ to ensure numerical stability [47]. This is achieved by scaling by the

mean absolute depth difference across each minibatch during training. We also constrain our 3D predictions so they are centered at a root joint that is encouraged to remain at the origin

$$\mathcal{L}_{root}(\hat{\mathbf{P}}) = ||\hat{\mathbf{P}}_{root}||_2. \tag{3}$$

This controls the size of the output space, as the network does not have to model all possible poses at all possible distances from the camera.

The above ranking loss only encourages the relative distances to the camera to be respected for each keypoint pair, in essence constraining the z values. To force the correct location in both x and y, we use an image reprojection loss

$$\mathcal{L}_{proj}(\hat{\mathbf{P}}, \mathbf{p}, \mathbf{v}, s) = \sum_j ||v_j(\Pi\hat{\mathbf{P}}_j^\top - \mathbf{p}_j)||_2, \tag{4}$$

where $v_j \in \{0,1\}$ is a visibility flag and Π is a projection matrix. When no camera intrinsic information is available $\Pi = \left[\begin{smallmatrix} s & 0 & 0 \\ 0 & s & 0 \end{smallmatrix}\right]$ *i.e.* scaled orthographic projection. Here, in addition to predicting the 3D pose, our network also learns to predict the scaling parameter s for each input pose. If the ground truth focal lengths are available during training $\Pi = \left[\begin{smallmatrix} f_x & 0 & 0 \\ 0 & f_y & 0 \end{smallmatrix}\right]$ and $\tilde{\mathbf{P}}_j = [x_j/(z_j+s), y_j/(z_j+s), 1]$. Now the network's scaling parameter, s, has a different interpretation, where it is used to predict the distance from the camera to the center of the person in 3D, and the reprojection loss becomes

$$\mathcal{L}_{proj}(\tilde{\mathbf{P}}, \mathbf{p}, \mathbf{v}, s) = \sum_j ||v_j(\Pi\tilde{\mathbf{P}}_j^\top - \mathbf{p}_j)||_2. \tag{5}$$

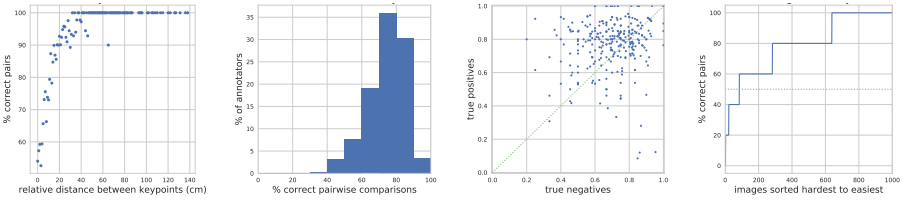
Even with the above terms, 3D pose estimation from 2D inputs is heavily underconstrained as many different 3D pose configurations can respect both the relative depth constrains and the reprojection loss. To further constrain the problem we include one additional geometric loss that enforces weak prior knowledge related to the ratio between the lengths of the different limbs. We assume we are given an input skeleton $\mathcal{B} = \{(b_1^1, b_1^2, l_1), (b_2^1, b_2^2, l_2), ..., (b_B^1, b_B^2, l_B)\}$, consisting of B 'bones' (*i.e.* limbs), where each entry (b^1, b^2, l) specifies the indices of the keypoint pair that are the endpoints for that particular limb, and its length l. The limb length loss then measures the difference in length between the predicted limb and the predefined reference length,

$$\mathcal{L}_{skel}(\hat{\mathbf{P}}) = \sum_{(b^1,b^2,l)\in\mathcal{B}} ||\text{len}(\hat{\mathbf{P}}, b^1, b^2) - l||_2, \tag{6}$$

where $\text{len}(\hat{\mathbf{P}}, j, k) = ||\hat{\mathbf{P}}_j - \hat{\mathbf{P}}_k||_2$. In practice we do not minimize the difference between the absolute bone lengths but instead normalize the predicted and reference bones by fixing one of the limbs to be unit length, in effect constraining their ratios as in [48, 53]. The skeleton loss also implicitly enforces symmetry between the corresponding left and right body limbs.

Our final loss $\mathcal{L}$ is the combination of the above four terms with additional weighting hyperparameters to optionally control the influence of each component

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{root}(\hat{\mathbf{P}}^i) + \alpha\mathcal{L}_{rel}(\hat{\mathbf{P}}^i, \mathcal{A}^i) + \beta\mathcal{L}_{proj}(\hat{\mathbf{P}}^i, \mathbf{p}^i, \mathbf{v}^i, s^i) + \gamma\mathcal{L}_{skel}(\hat{\mathbf{P}}^i). \tag{7}$$



(a) Accuracy vs Depth (b) Annotator Accuracy (c) Annotator Skill (d) Image Difficulty

Figure 2: (a-c) Analysis of human performance in the relative depth annotation task measured on 1,000 randomly selected images from Human3.6M. Annotators always perform better than chance and are more than 90% accurate when the distance between pairs is larger than 20 cm. (d) Sorting of the images based on their difficulty.

Figure 3: Examples of the supervision collected for the LSP dataset. The names of the keypoints are written on top and the green keypoint is annotated to be closer to the camera compared to the red one. Images are sorted left to right from most to least confident based on the annotator's agreement, and the last two examples illustrate particularly challenging cases where the keypoints are at a similar distance to the camera.

4 Human Relative Depth Annotation Performance

Our model for 3D pose estimation makes use of relative depth annotations at training time. In order to use this type of supervision, it is important to understand how accurately can humans provide these labels. This is in contrast to the study carried out in [23], which investigates the ability of humans to observe and physically reenact a target 3D pose. However, one important detail for our analysis, missing from [23], is a measure of the quality of the relative joint annotations that can be collected via a crowd sourcing platform. We performed an evaluation using participants recruited via Mechanical Turk and randomly selected 1,000 images from the Human3.6M dataset [13], as it features ground truth depth. A snapshot of the user interface for the task is provided in the supplementary material. For each annotation task, the crowd workers were presented with an image from the dataset along with two randomly selected keypoints and were instructed to imagine themselves looking through the camera and report which of the two keypoints appeared closer to them. We decided to force annotators to choose from one of the two possibilities and did not provide a 'same distance' option for ambiguous situations, as those cases can be inferred by looking at the disagreement between the workers. For each of the 1,000 images, we collected five random pairs of keypoints, ensuring that five different annotators labeled the same keypoints and image combination. In total, this resulted in 25,000 individual annotations collected from 284 annotators, with an average of 88 annotations each. We merged each of the five votes per keypoint pair using the crowd annotation system of [5], resulting in a single predicted label per pair. We found this to perform slightly better than majority voting, with the added benefit of providing a probabilistic label.

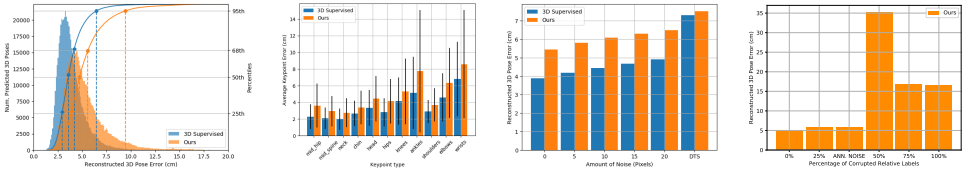
Model					3D Pose Error (mm)	
	Reprojection Type	**Skeleton**	**Amount Train**	**Distance Tolerance**	**Scale**	**Procrustes**
	Perspective	H36 Avg	1 Pair	No	75.50	53.24
	Scaled Orthographic	Generic Avg	1 Pair	No	76.46	58.29
	Scaled Orthographic	H36 Subject	1 Pair	No	69.75	51.89
	Scaled Orthographic	H36 Avg	All Pairs (136)	No	70.00	51.84
	Scaled Orthographic	H36 Avg	1 Pair	Yes (100mm)	67.40	50.56
Ours Relative	Scaled Orthographic	H36 Avg	1 Pair	No	70.30	52.21
3D Supervised	-	-	-	-	47.51	38.41

Table 1: Results for different variants of our model on Human3.6M. The last two columns show the performance at test time when an optimal re-scaling (*Scale*) or a full rigid alignment (*Procrustes*) is performed for each predicted pose based on the available ground truth. The row (*Ours Relative*) shows the default parameters adopted in our implementation.

We observed a bias in the annotations, due to the fact that crowd workers tend to not factor in the forward lean of the camera when making their predictions. Correcting for this bias during evaluation results in 73% of the total raw annotations being correct, and a 79% accuracy of the relative orderings after merging the five repetitions using [5]. Note that fixing this bias requires ground truth depth data, and is a meaningful operation only if done to evaluate more precisely annotators' performance. In fact, in real world scenarios our models will learn this annotation bias. A more detailed exploration of the lean bias, and its correction, is presented in the supplementary material.

The plots in Fig. 2 have been computed after the removal of the bias, and quantitatively explore how accurately did the crowd workers perform in the described task. In Fig. 2 (a) we see that for keypoint pairs that are separated by more than 20 cm our merged predictions are correct over 90% of the time, where random guessing is 50%. Furthermore, Fig. 2 (b), while only a small number of workers annotated over 90% of the pairs correctly, the vast majority tends to perform better than random guessing. In Fig. 2 (c) we observe that the rate of true positives versus true negatives for every annotator is fairly symmetric, indicating that workers are equally good at providing the correct answer independently of a keypoint being in front or behind another one. Some image and keypoint combinations are more challenging than others, and in Fig. 2 (d) we sort the images from hardest to easiest based on the percentage of keypoint pairs that are correctly annotated. For over two thirds of the images, four out of the five pairs are correctly annotated. Importantly, the cases where annotators have trouble predicting the correct keypoint order, by and large, tend to be ambiguous pairs where the actual ground truth distances between the keypoints are small. These results indicate that human annotators can indeed provide high quality weak 3D pose information in the form of relative depth annotations, which can be used as supervision for our methods.

Using the same protocol described above we collected annotations for all 2,000 images in the Leeds Sports Pose (LSP) dataset [14], which features a much larger variation in camera viewpoint and pose compared to Human3.6M. Again, we selected five random keypoint pairs per image, with five repeats, resulting in a total of 50,000 annotations. Annotations were performed by 348 annotators who provided an average of 144 labels each. Example annotations after merging the five responses the using [5] can be seen in Fig. 3. Unlike Human3.6M, there is no ground truth depth data available for LSP, so to evaluate the quality of the crowd annotations two of the authors independently annotated the same subset of 500 keypoint pairs. Agreement between the two was 84%, where the majority of disagreements occurred in ambiguous cases. For the set of pairs where the two annotators agreed, the merged crowd annotations were the same 90.2% of the time. These results are consistent with the performance on Human3.6M, despite the larger variation in poses and camera viewpoints.



(a) Overall Error (b) Per Keypoint Error (c) 2D Keypoint Noise (d) 3D Annotation Noise

Figure 4: (a-b) Histogram of the errors on the Human3.6M dataset (lines and dots respectively show the cumulative sum and percentiles) with a breakdown over individual keypoints. (c-d) Our method is robust to increasingly strong perturbations on the 2D input keypoint locations at test time and noise on the relative depth labels during training.

5 3D Pose Estimation Results

We use a similar underlying network to [24] for all our experiments and report the 3D pose estimation performance of our model on the Human3.6M [13] and LSP [14] datasets in the following sections. A more thorough description of our architecture and implementation details are available in the supplementary material, along with several qualitative examples.

5.1 Human3.6M

As noted in [7], many different evaluation protocols have been defined for Human3.6M, making it very challenging to comprehensively compare to other methods. We opt to use protocol #2 from the state-of-the-art [24], as it is the model most similar to ours, and has been used by several recent approaches. Here, training is performed on subjects [1, 5, 6, 7, 8] and the test set consists of all frames and cameras for subjects [9, 11]. Some baselines, evaluate at 10fps which we observe to make no difference to the test scores. As in [24], at test time, Procrustes alignment is performed between each prediction and its corresponding test example. With the exception of the results in Table 2, where we average across actions, everywhere else we report results by averaging across all frames. Finally, unless specified, we use 17 2D keypoints as input and predict their corresponding 3D locations.

We explored different configurations of our model at training time, by varying the type of reprojection loss adopted, the proportions of the input skeleton, the amount of training data, and the tolerance parameter. We describe our most interesting findings below, and report a quantitative evaluation in Table 1. *Reprojection*: Using the ground truth focal lengths in Eqn. 5 performs worse than Eqn. 4, suggesting that it is harder for the network to predict an accurate value for the scale parameter s when it represents distance. *Skeleton*: Using less accurate limb length ratios (obtained from [41]) in Eqn. 6 hurts performance, but there is no significant improvement from using subject specific ratios as compared to the training set mean. *Amount Train*: Increasing the amount of keypoint pairs annotated with relative depth does not alter the results. However, this effect is likely due to the high redundancy present in Human3.6M. This is consistent with the findings of Fig. 1 (d), where we report the error against the percentage of training images used. *Distance Tolerance*: We observed that setting the depth tolerance to 100 mm, results in a noticeable increase in performance. This is because using $r = 0$ in Eqn. 2 forces the network to constrain the exact depth value of a pair of predicted keypoints, as opposed to their relative order. Some of these configurations are more realistic than others when using crowd provided annotations on an 'in the wild' image dataset. For this reason, we denote as 'Ours Relative' in Table 1 the method characterized by the most realistic assumptions and use it for the remaining experiments.

	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SitingD	Smoke	Wait	WalkD	Walk	WalkT	Avg
Sanzari et al. [32]	48.8	56.3	96.0	84.8	96.5	105.6	66.3	107.4	116.9	129.6	97.8	65.9	130.5	92.9	102.2	93.2
Rogez et al. [28]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	71.6
Kanazawa et al. [14]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	58.1
Pavlakos et al. [24]	47.5	50.5	48.3	49.3	50.7	55.2	46.1	48.0	61.1	78.1	51.1	48.3	52.9	41.5	46.4	51.9
Tekin et al. [36]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	50.1
Fang et al. [10]	38.2	41.7	43.7	44.9	48.5	55.3	40.2	38.2	54.5	64.4	47.2	44.3	47.3	36.7	41.7	45.7
Hossain et al. [9] (T)	36.9	37.9	42.8	40.3	46.8	46.7	37.7	36.5	48.9	52.6	45.6	39.6	43.5	35.2	38.5	42.0
Pavlakos et al. [24] (E)	34.7	39.8	41.8	38.6	42.5	47.5	38.0	36.6	50.7	56.8	42.6	39.6	43.9	32.1	36.5	41.8
Martinez et al. [22] 17j GT/GT	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	37.1
Martinez et al. [22] 17j SH/SH	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	69.4	49.2	45.0	49.5	38.0	43.1	47.7
3D Supervised 17j GT/GT	31.6	36.6	34.3	37.2	37.5	43.6	36.8	34.0	42.3	50.2	38.2	38.6	40.9	33.7	36.6	38.1
Ours Relative 17j GT/GT	45.7	45.5	48.3	53.3	49.3	56.7	42.6	48.5	64.1	80.7	48.6	50.4	51.5	51.7	51.4	52.6
Ours Relative 17j GT/GT (N)	48.2	48.2	57.6	55.4	52.9	60.9	44.1	54.2	70.6	107.9	52.5	54.6	54.6	54.7	54.9	58.1
3D Supervised 16j GT/GT	31.6	36.8	35.3	38.0	37.7	44.9	36.4	34.1	43.6	50.1	38.5	38.4	41.6	33.9	36.9	38.5
Ours Relative 16j GT/GT	45.2	45.7	48.8	53.8	50.6	58.8	43.0	47.7	66.1	85.3	51.1	52.0	52.4	53.7	52.6	53.8
3D Supervised 16j GT/SH	56.4	69.9	60.2	67.8	80.6	73.6	59.2	78.7	80.7	105.2	72.7	67.4	83.1	75.2	77.8	73.9
Ours Relative 16j GT/SH	60.9	64.2	67.3	71.7	80.8	76.9	55.3	68.2	92.2	114.7	74.4	66.8	72.7	73.5	73.1	74.2

Table 2: 3D pose reconstruction error in mm on Human3.6M using protocol #2 as defined in Sec. 5.1. GT and SH are ground truth and detected 2D input keypoints. GT/SH indicates trained on GT, tested on SH. j is the number of joints used for testing. (T) represents the use of temporal information, (E) represents the use of extra training data, and (N) represents the use of noise in the relative annotations. '3D Supervised' is our re-implementation of [24].

Fig. 4 summarizes the overall performance and robustness to noise of our model. In Fig. 4(a) we show a histogram of the pose errors on the test set both for our method and [24]. The mode and median of the two curves are less than 15mm from each other. However, our method suffers from more catastrophic errors, as can be seen in the longer tail. This is due to the fact that even when respecting all the relative depth labels we do not fully constrain their absolute depth, and is also observable by looking at the breakdown of the error over keypoint type visualized in Fig. 4(b). As one might expect, body extremities such as ankles and wrists show a larger error (and deviation) since they are the least rigid joints. Fig. 4(c) shows the degradation in performance when adding noise to the 2D input keypoints at *test* time. In the first five bars we add a Gaussian noise $\mathcal{N}(0, \sigma^2)$ with increasingly high variance (up to 20 pixels) and observe a smooth degradation of performance. The rightmost bar shows the case in which the outputs of a keypoint detector [28] are used. Interestingly, here our method is closer to the performance of [24]. We hypothesize this behavior is due to our reprojection loss at training time, which encourages our method to find plausible 3D poses that respect the input poses, making it more robust to slight changes in the distribution of the input keypoints. Finally in Fig. 4(d) we demonstrate that our model is also robust to noise in the relative depth labels during *training*. Performance is mostly unchanged when up to 30% of the labels are randomly flipped. The third bar corresponds to the amount of noise obtained from simulated crowd annotators, regressed from Fig. 2(a). This is of interest, as it shows performance with noise comparable to what we would expect to collect in the wild. The worst performance is obtained when the labels are randomly flipped, and improves for cases in which the amount of noise is larger than 50%, as the model is able to exploit structure that is still present in the data, but produces poses that are flipped back to front.

Finally, in Table 2 we compare our model to existing fully 3D supervised approaches. Even with significantly less training data, and without any architecture exploration, we still perform competitively compared to recent supervised methods. When available, 3D ground truth is a very powerful training signal, but our results show that relative depth data can still be used at the expense of some accuracy at test time. Our model is robust to using noisy predicted 2D keypoints at test time, again with a minor decrease in performance. Example 3D predicted poses on Human3.6M can be seen in the supplementary material.

Figure 5: Predicted 3D poses on LSP. Fine-tuning (FT) on LSP significantly improves the quality of our predictions especially for images containing uncommon poses and viewpoints that are not found in Human3.6M, such as those visualized in the last row.

5.2 Leeds Sports Pose Dataset

Finally, we perform experiments on the LSP dataset which, unlike Human3.6M, does not contain ground truth 3D poses. As a result, we train and test our model using the merged crowd annotations collected via Mechanical Turk. Pre-training on Human3.6M both the 3D supervised method [24] and our relative model using one relative comparison per input pose, results in a test error of 34.3% and 35.2% respectively. Here, test error refers to the percentage of relative labels incorrectly predicted compared to the merged crowd annotations. Our model uses only weak training data but still performs comparably to the fully supervised baseline. However, if we fine-tune our model on the LSP training set, the error reduces to 25.6%. Example predicted 3D poses are displayed in Fig. 5. This shows that we can successfully make use of noisy training annotations, resulting in improved output 3D poses.

6 Conclusion

We presented a weakly supervised approach for 3D human pose estimation. We showed that sparse constraints that indicate the relative depth of pairs of keypoints can be used as a training signal, resulting in competitive results at a fraction of the amount of training data. Unlike most existing approaches that require ground truth 3D poses, our method can be applied to legacy image collections, as only the input 2D keypoints and relative depth annotations are required. This opens the door to using existing datasets for 3D pose estimation in the wild.

Large scale annotation is time consuming and expensive, even when only collecting weak supervision. In future, we plan to investigate efficient, active learning based, approaches for collecting annotations *e.g.* [19]. Current state-of-the-art 2D pose estimation algorithms perform best on single humans in isolation and their performance deteriorates when there are large numbers of occluded keypoints and closely interacting people [36]. Including weak 3D information for multiple interacting individuals may help resolve some of these ambiguities.

Acknowledgements We would like to thank Google for their gift to the Visipedia project and Amazon Web Services (AWS) for Research Credits.

References

[1] Ijaz Akhter and Michael J Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *CVPR*, 2015.

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

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM Transactions on Graphics (TOG)*, 2005.

[4] Lubomir Bourdev and Jitendra Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009.

[5] Steve Branson, Grant Van Horn, and Pietro Perona. Lean crowdsourcing: Combining humans and machines in an online system. In *CVPR*, 2017.

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[7] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *CVPR*, 2017.

[8] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *NIPS*, 2016.

[9] Wenzheng Chen, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Synthesizing training images for boosting human 3d pose estimation. In *3DV*, 2016.

[10] Haoshu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. Learning pose grammar to encode human body configuration for 3d pose estimation. In *AAAI*, 2018.

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[12] Mir Rayat Imtiaz Hossain and James J Little. Exploiting temporal information for 3d pose estimation. *arXiv:1711.08585*, 2017.

[13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *PAMI*, 2014.

[14] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010.

[15] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, 2015.

[16] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018.

[17] Sijin Li and Antoni B Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *ACCV*, 2014.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[19] Buyu Liu and Vittorio Ferrari. Active learning for human pose estimation. In *ICCV*, 2017.

[20] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 2015.

[21] Subhransu Maji. Large scale image annotations on amazon mechanical turk. *EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2011-79*, 2011.

[22] Subhransu Maji, Lubomir Bourdev, and Jitendra Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR*, 2011.

[23] Elisabeta Marinoiu, Dragos Papava, and Cristian Sminchisescu. Pictorial human spaces: A computational study on the human perception of 3d articulated poses. *IJCV*, 2016.

[24] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017.

[25] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016.

[26] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3DV*, 2017.

[27] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *CVPR*, 2017.

[28] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

[29] Quan Nguyen and Michael Kipp. Annotation of human gesture using 3d skeleton controls. In *LREC*, 2010.

[30] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *CVPR*, 2017.

[31] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. *CVPR*, 2018.

[32] Gerard Pons-Moll, Andreas Baak, Juergen Gall, Laura Leal-Taixe, Meinard Mueller, Hans-Peter Seidel, and Bodo Rosenhahn. Outdoor human motion capture using inverse kinematics and von mises-fisher sampling. In *ICCV*, 2011.

[33] Gerard Pons-Moll, David J Fleet, and Bodo Rosenhahn. Posebits for monocular human pose estimation. In *CVPR*, 2014.

[34] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. LCR-Net: Localization-Classification-Regression for Human Pose. In *CVPR*, 2017.

[35] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *arXiv:1803.00455*, 2018.

[36] Matteo Ruggero Ronchi and Pietro Perona. Benchmarking and error diagnosis in multi-instance pose estimation. In *ICCV*, 2017.

[37] Marta Sanzari, Valsamis Ntouskos, and Fiora Pirri. Bayesian image based 3d pose estimation. In *ECCV*, 2016.

[38] Ben Sapp and Ben Taskar. Modec: Multimodal decomposable models for human pose estimation. In *CVPR*, 2013.

[39] Tianmin Shu, Michael S Ryoo, and Song-Chun Zhu. Learning social affordance for human-robot interaction. *IJCAI*, 2016.

[40] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 2010.

[41] Camillo J Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *CVIU*, 2000.

[42] Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Structured prediction of 3d human pose with deep neural networks. In *BMVC*, 2016.

[43] Bugra Tekin, Pablo Marquez Neila, Mathieu Salzmann, and Pascal Fua. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *ICCV*, 2017.

[44] Denis Tome, Christopher Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. *CVPR*, 2017.

[45] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, 2014.

[46] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017.

[47] Chaoyang Wang, Jose Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. *arXiv:1712.00175*, 2017.

[48] Chunyu Wang, Yizhou Wang, Zhouchen Lin, Alan L Yuille, and Wen Gao. Robust estimation of 3d human poses from a single image. In *CVPR*, 2014.

[49] Robert Wang, Sylvain Paris, and Jovan Popović. Practical color-based motion capture. In *SIGGRAPH/Eurographics Symposium on Computer Animation*, 2011.

[50] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016.

[51] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. *arXiv:1803.09722*, 2018.

[52] Hashim Yasin, Umar Iqbal, Bjorn Kruger, Andreas Weber, and Juergen Gall. A dual-source approach for 3d pose estimation from a single image. In *CVPR*, 2016.

[53] 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*, 2017.

[54] Daniel Zoran, Phillip Isola, Dilip Krishnan, and William T Freeman. Learning ordinal relationships for mid-level vision. In *ICCV*, 2015.

7 Supplementary Material

7.1 Implementation Details

We use the same fully connected network architecture as [24] for all experiments. We use the one stage version of the model as we did not observe any significant loss in performance compared the two stage version, see Fig. 6. To predict the scale parameter s used in our reprojection losses we add an additional fully connected layer to the output of the penultimate set of layers and apply a sigmoid non-linearity to its output. In the relative depth loss we set $\lambda = 2.5$. We set the weighting hyperparameters α, β, and γ in the main loss to 1. All the results in the main paper are reported We train all models for 25 epochs, as we observe that our model does not tend to benefit from additional training time. For our relative model we center the input 2D keypoints by setting the root location to (0,0). We did not perform this centering for the supervised baseline as we found that it hurt performance, but we did center the 3D coordinates in a similar fashion. As in [24], we clip the gradients to one during training. Training time is less than five minutes for one epoch for our relative model.

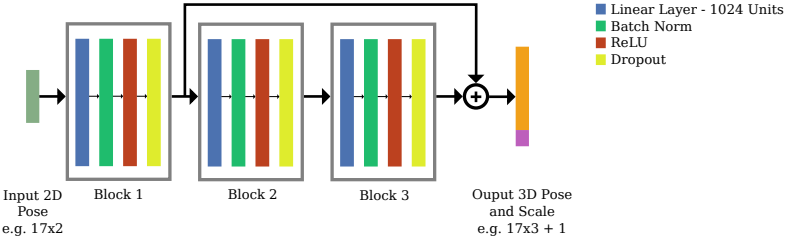
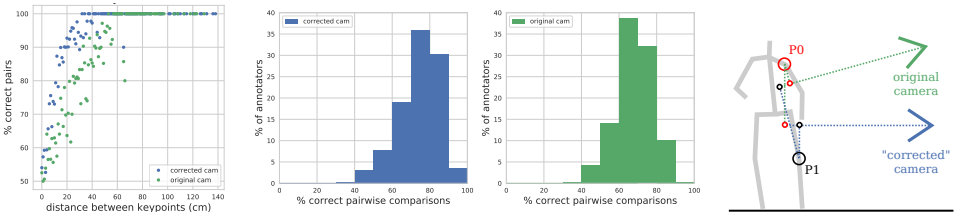


Figure 6: Network architecture. We use a similar architecture to [24] but include scale prediction at the end of the network.

7.2 Human Annotation Performance

In this section we provide some additional discussion of the results of our user study on Human3.6M [13]. As mentioned in the main paper, we observed that annotators tended to estimate depth in images by correcting for the orientation of the camera. In Fig. 7 (c) we see an illustration of this effect. Here, from the perspective of the input camera (green) keypoint 'P0' is closer than 'P1'. In practice, even though they see an image of the scene taken from the perspective of the green camera, annotators seemingly correct for the orientation of the camera and 'imagine' the distance of the scene from the perspective of the blue camera. While this change in camera position is subtle, it affects the relative ordering of the points as 'P1' is now closer to the camera. We hypothesize that this is a result of the annotator imagining themselves in the same pose as the individual in the image and then estimating the distance to the camera in a Manhattan world sense. Without correcting for this effect 67% of the provided pairwise annotations are correct, but when this is taken into account then accuracy increases to 73%. We correct for the bias by forcing the camera to be upright when computing the scene depth. The results before and after applying this correction and annotator accuracies can be viewed in Figs. 7 (a) and (b). This effect is likely to be exacerbated in Human3.6M as there are only four different camera viewpoints in the entire dataset and they are all facing downwards. We expect this to be less of an issue for datasets that fea-



(a) Accuracy vs Distance (b) Annotator Accuracy (c) Camera Correction

Figure 7: Human relative depth annotation performance on 1,000 images selected from the Human3.6M dataset [13]. (a-b) Without correcting for the orientation of the camera, annotators perform worse (green lines). (c) The green camera represents the input view and the blue is the upright orientated view as perceived by our annotators. If the camera is orientated upwards when performing the evaluation the relative depths are a better match to the annotator provided labels.

ture a larger variation in camera positions relative to the subject of interest as the dominant ground plane will have less of a biasing effect.

Fig. 11 depicts an example task from our user interface that was shown to annotators. The first time annotators performed our task they were presented with a short tutorial that included sample images and were instructed on how to use the interface and given feedback when they guessed the incorrect depth ordering. For each task, we also included a short delay before annotators could select their response to encourage them to pay attention to the input image when performing the task. Example annotations from Human3.6M [13] can be seen in Fig. 10. Unsurprisingly, keypoint pairs that have larger relative distances are easier to annotation. For these examples the ground truth accuracies are computed with respect to the corrected ground truth. Example 3D predicted poses on Human3.6M can be seen in Fig. 8.

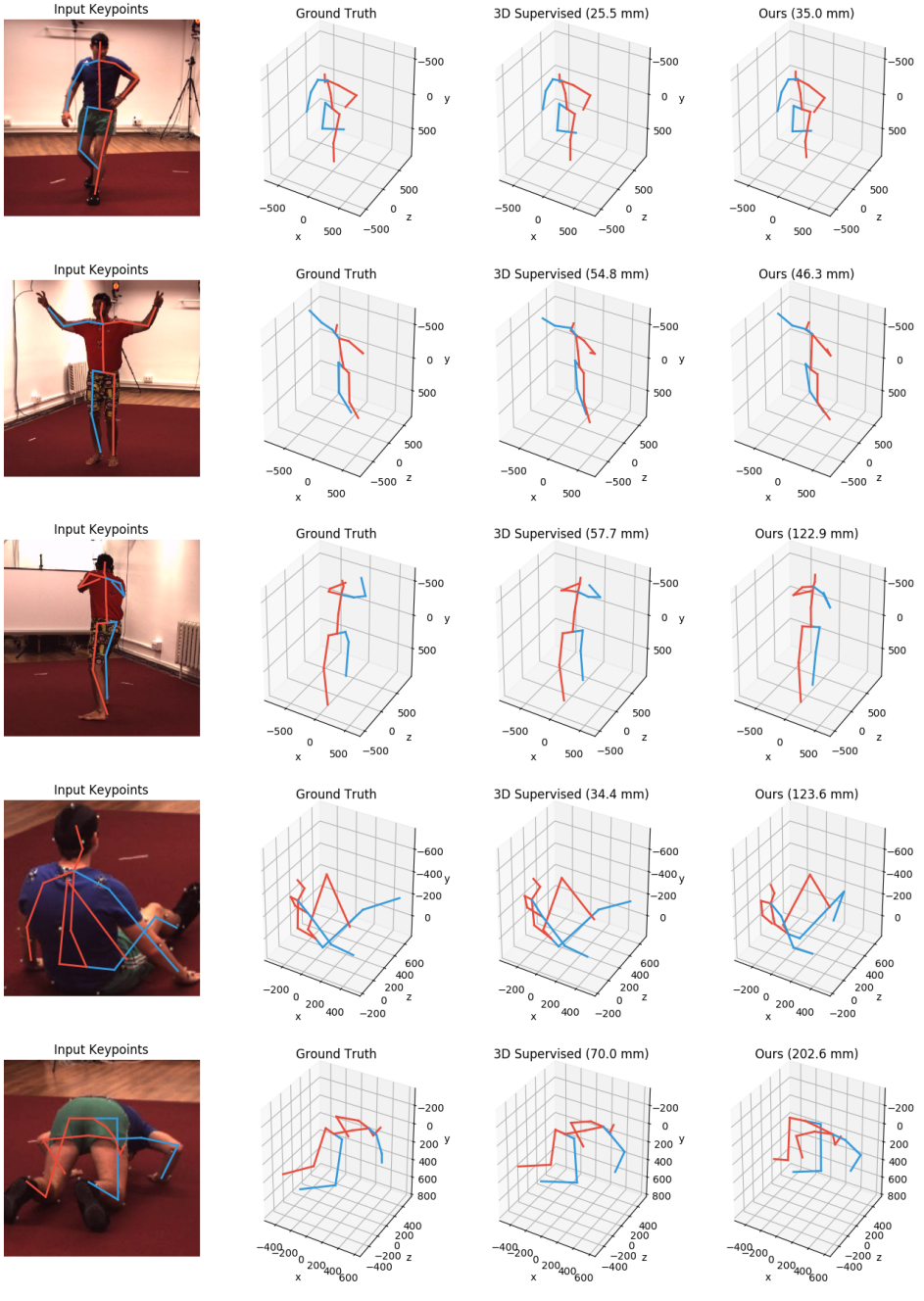
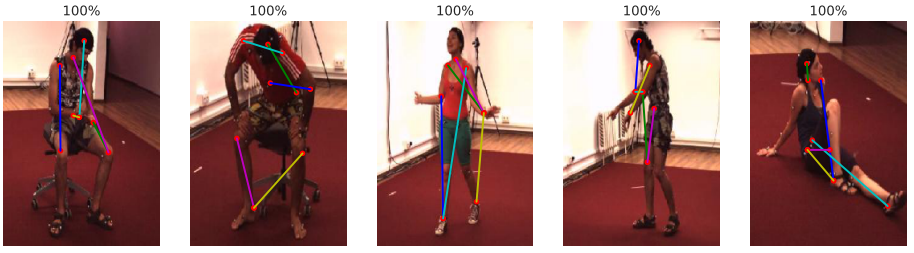
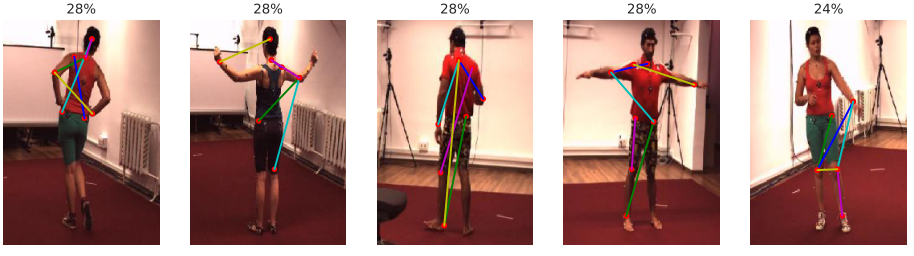


Figure 8: Test time predictions on Human3.6M. Despite using much weaker training data our relative model produces sensible results for most input poses. Both the supervised and our approach are depicted after rigid alignment, with the pose error displayed on top.

Figure 9: Predicted 3D poses on LSP. Fine-tuning (FT) on LSP significantly improves the quality of our predictions especially for images containing uncommon poses and viewpoints that are not found in Human3.6M.



(a) Easy keypoint pairs.



(b) Difficult keypoint pairs.

Figure 10: Example Human3.6M keypoint pairs. Colored lines link pairs that were shown to annotators. The numbers on top represent the raw accuracy of the crowd provided labels before merging. (a) Easy pairs where no incorrect annotations were made i.e. each of the five annotators annotated all five pairs correctly. (b) Difficult examples where the randomly selected pairs tend to be at a similar distance to the camera, resulting in lower performance.

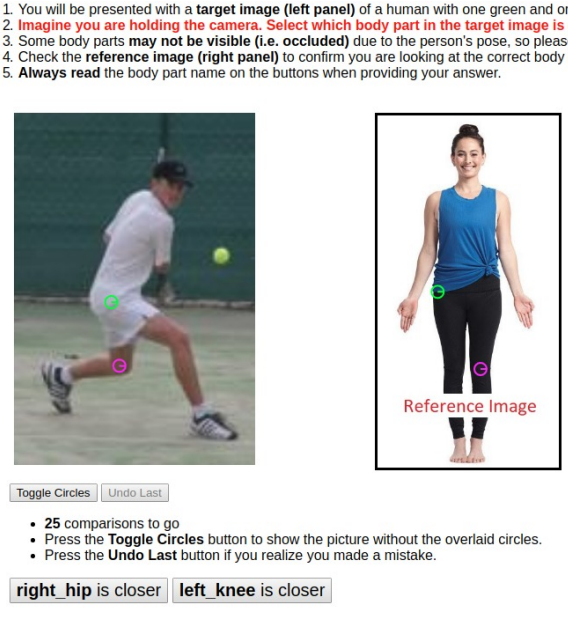


Figure 11: Our user interface. The annotator's goal is the determine the relative depth of the highlighted keypoints in the left image. The reference image on the right highlights the same keypoints and helps in situations where they are occluded.