Ordinal Depth Supervision for 3D Human Pose Estimation

Georgios Pavlakos[1], Xiaowei Zhou[2], Kostas Daniilidis[1]
[1] University of Pennsylvania [2] Zhejiang University

Abstract

Our ability to train end-to-end systems for 3D human pose estimation from single images is currently constrained by the limited availability of 3D annotations for natural images. Most datasets are captured using Motion Capture (MoCap) systems in a studio setting and it is difficult to reach the variability of 2D human pose datasets, like MPII or LSP. To alleviate the need for accurate 3D ground truth, we propose to use a weaker supervision signal provided by the ordinal depths of human joints. This information can be acquired by human annotators for a wide range of images and poses. We showcase the effectiveness and flexibility of training Convolutional Networks (ConvNets) with these ordinal relations in different settings, always achieving competitive performance with ConvNets trained with accurate 3D joint coordinates. Additionally, to demonstrate the potential of the approach, we augment the popular LSP and MPII datasets with ordinal depth annotations. This extension allows us to present quantitative and qualitative evaluation in non-studio conditions. Simultaneously, these ordinal annotations can be easily incorporated in the training procedure of typical ConvNets for 3D human pose. Through this inclusion we achieve new state-of-the-art performance for the relevant benchmarks and validate the effectiveness of ordinal depth supervision for 3D human pose.

1. Introduction

Human pose estimation has been one of the most remarkable successes for deep learning approaches. Leveraging large-scale datasets with extensive 2D annotations has immensely benefited 2D pose estimation [55, 34, 31, 61], semantic part labeling [11, 57] and multi-person pose estimation [15, 30, 9]. In contrast, the complexity of collecting images with corresponding 3D ground truth has constrained 3D human pose datasets in small scale [19] or strictly in studio settings [42, 16]. The goal of this paper is to demonstrate that in the absence of accurate 3D ground truth, end-to-end learning can be competitive by using weaker supervision in the form of ordinal depth of the joints (Figure 1).

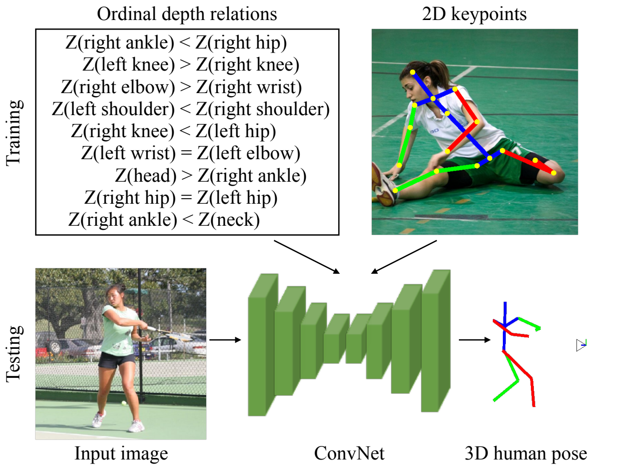

Ordinal depth relations
2D keypoints
Training
Z(right ankle) < Z(right hip)
Z(left knee) > Z(right knee)
Z(right elbow) > Z(right wrist)
Z(left shoulder) < Z(right shoulder)
Z(right knee) < Z(left hip)
Z(left wrist) = Z(left elbow)
Z(head) > Z(right ankle)
Z(right hip) = Z(left hip)
Z(right ankle) < Z(neck)
Testing
Input image
ConvNet
3D human pose

Figure 1: Summary of our approach. In the absence of accurate 3D ground truth we propose the use of ordinal depth relations (closer-farther) of the human body joints for end-to-end training of 3D human pose estimation systems.

Aiming to boost end-to-end discriminative approaches, different techniques attempt to augment the training data. Synthetic examples can be produced in abundance [13, 53], but there is no guarantee that they come from the same distribution as natural images. Multi-view systems for accurate capture of 3D ground truth can work outdoors [26], but they need to be synchronized and calibrated, so data collection is not practical and hard to scale. These limitations have favored reconstruction approaches, e.g., [5, 25], which employ reliable 2D pose detectors and recover 3D pose in a subsequent step using the 2D joint estimates. Unfortunately, even in the presence of perfect 2D correspondences, the final 3D reconstruction can be erroneous. This 2D-to-3D reconstruction ambiguity is mainly attributed to the binary ordinal depth relations of the joints (closer-farther) [45]. Leveraging image-based evidences, such as occlusion and shading, can largely resolve the ambiguity, yet this information is discarded by reconstruction approaches.

Motivated by the particular power of ordinal depth relations at resolving reconstruction ambiguities and the fact that this information can be acquired by human annotators, we propose to use ordinal depth relations to train ConvNets for 3D human pose estimation. Since humans can easily perceive pose [24] and they are better at estimating ordinal depth than explicit metric depth [50], annotators can provide pairwise ordinal depth relations for a wide range of imaging conditions, activities, and viewpoints. We develop on the idea of ordinal relations demonstrating their flexibility and effectiveness in a variety of settings: 1) we use them to predict directly the depths of joints, 2) we combine them with 2D keypoint annotations to predict 3D poses, 3) we demonstrate how they can be incorporated within a volumetric representation of 3D pose [32]. In every case, the weak supervision signal provided by these ordinal relations leads to a competitive performance compared to fully supervised approaches that employ the actual 3D ground truth. Additionally, to motivate the use of ordinal depth relations for human pose, we provide ordinal depth annotations for two popular 2D human pose datasets, LSP [18] and MPII [2]. This extension allows us to provide quantitative and qualitative evaluation of our approach in non-studio settings. Simultaneously, these ordinal annotations for in-the-wild images can be easily incorporated in the training procedure of typical ConvNets for 3D human pose leading to new state-of-the-art results for the standard benchmarks of Human3.6M and HumanEva-I. These performance benefits underline the effectiveness of ordinal depth supervision for human pose problems and provide motivation for further exploration using the available annotations.

Our contributions can be summarized as follows:

- We propose the use of ordinal depth relations of human joints for 3D human pose estimation to bypass the need for accurate 3D ground truth.
- We showcase the flexibility of the ordinal relations by incorporating them in different network settings, where we always achieve competitive results to training with the actual 3D ground truth.
- We augment two popular 2D pose datasets (LSP and MPII) with ordinal depth annotations and demonstrate the applicability of the proposed approach to 3D pose estimation in non-studio conditions.
- We include our ordinal annotations in the training procedure of typical ConvNets for 3D human pose and exemplify their effectiveness by achieving new state-of-the-art results on the standard benchmarks.

2. Related work

Since the literature on 3D human pose estimation is vast, here we discuss works closely related to our approach and refer the interested reader to Sarafianos *et al.* [41] for a recent survey on this topic.

Reconstruction approaches: A long line of approaches follows the reconstruction paradigm by employing 2D pose detectors to localize 2D human joints and using these locations to estimate plausible 3D poses [10, 17]. Zhou *et al.* [67, 68] use 2D heatmaps from a 2D pose ConvNet to reconstruct 3D pose in a video sequence. Bogo *et al.* [5] fit a statistical model of 3D human shape to the predicted 2D joints. Alternatively, a network can also handle the step of lifting 2D estimates to 3D poses [56, 28, 51]. Notably, Martinez *et al.* [25] achieve state-of-the-art results with a simple multilayer perceptron that regresses 3D joint locations, given 2D keypoints as input. Despite the success of this paradigm, it comes with important drawbacks. No image-based evidence is used during the reconstruction step, the result is too reliant on an imperfect 2D pose detector and even for perfect 2D correspondences, the 3D estimate might fail because of the reconstruction ambiguity. In contrast, by using ordinal depth relations we can leverage rich image-based information during estimation, without relinquishing the accuracy of reconstruction approaches, which can also be integrated in our framework (Section 3.4).

Discriminative approaches: Discriminative approaches are orthogonal to the reconstruction paradigm since they estimate the 3D pose directly from the image. Prior work uses ConvNets to regress the coordinates of the 3D joints [22, 47, 48, 51, 44, 26], to regress 3D heatmaps [32], or to classify each image in the appropriate pose class [39, 40]. The main critique of these end-to-end approaches is that images with corresponding 3D ground truth are required for training. Our work attempts to relax this important constraint, by training with weak 3D information in the form of ordinal depth relations for the joints and 2D keypoints. Weak supervision was also used in recent work [64] by constraining the lengths of the predicted limbs. However, we argue that our supervision does not simply constraint the output of the network, but also provides novel information for in-the-wild images and further enhances training.

Generating training examples: The limited availability of 3D ground truth for training 3D human pose ConvNets has also been addressed in various ways in recent works. The most straightforward solution is to use graphics to augment the training data [13, 53, 26]. Differently, Rogez and Schmid [39] propose a collage approach by composing human parts from different images to produce combinations with known 3D pose. In both cases though, most examples do not reach the detail and variety level that in-the-wild images have. Mehta *et al.* [26] record multiple views outdoors and estimate accurate 3D ground truth for every view. However, multi-view systems need to be synchronized and calibrated, so large-scale data collection is not trivial.

3D annotations: Prior works have also relied on humans to perceive and annotate 3D properties that are lost through the projection of a 3D scene on a 2D image. Bell *et al.* [3] and Chen *et al.* [12] annotate the ordinal relations for the apparent depth of pixels in the image. In the work of Xiang *et al*. [59, 58], humans align 3D CAD models with single images to provide viewpoint information. Concerning 3D human pose annotations, the famous poselets work from Bourdev and Malik [6] uses an interactive tool for annotators to adjust the 3D pose, making the procedure laborious. Maji *et al*. [23] provide 3D annotations for human pose, but only in the form of yaw angles for head and torso. The idea of ordinal depth relations is also explored by Pons-Moll *et al*. [35] where attributes regarding the relative 3D position of the body parts are included in their posebits database. Different to them, we provide annotations by humans for a much larger set of images (i.e., more than 15k images with our annotations compared to 1k for the posebits dataset), and instead of exploring an extensive set of pose attributes, we propose a cleaner training scheme that requires only 2D keypoint locations and ordinal depth relations. In recent work, Lassner *et al*. [21] estimate proposals of 3D human shape fits for single images which are accepted or rejected by annotators. Despite the rich ground truth in case of a good fit, many automatic proposals are of low quality, leading to many discards. Our work aims for a more balanced solution where 3D annotations have a weaker form, but the task is easy for humans, so that they can provide annotations on a large scale for practically any available image.

Ordinal relations: There is a long history for learning from ordinal relations, outside the field of computer vision, with particular interest in the area of information retrieval, where many algorithms for learning-to-rank have been developed [7, 8, 46]. In the context of computer vision, previous works have used relations to learn apparent depth [69, 12] or reflectance [29, 63] of a scene. We share a common motivation with these approaches in the sense that ordinal relations are easier for humans to annotate, compared to metric depth or absolute reflectance values.

3. Technical approach

In this section we present our proposed approach for different settings of 3D human pose estimation. First, in Section 3.1 we predict only the depths of the human joints, relying on ordinal depth relations and a ranking loss for training. Then, in Section 3.2 we combine the ordinal relations with 2D keypoint annotations to predict the 3D pose coordinates. In Section 3.3 we explore the incorporation of ordinal relations within a volumetric representation for 3D human pose [32]. Finally, Section 3.4 presents the extension of the previous networks with a component designed to encode a geometric 3D pose prior.

3.1. Depth prediction

Our initial goal is to establish the training procedure such that we can leverage ordinal depth relations to learn to predict the depths of human joints. This is the simplest case, where instead of explicitly predicting the 3D pose, we only predict depth values for the joints.

Let us represent the human body with N joints. For each joint i we want to predict its depth z_i. The provided data are in the form of pairwise ordinal depth relations. For a pair of joints (i, j), we denote the ordinal depth relation as $r_{(i,j)}$ taking the value:

- $+1$, if joint i is closer than j,
- -1, if joint j is closer than i,
- 0, if their depths are roughly the same.

The ConvNet we use for this task takes the image as input and predicts N depth values z_i, one for each joint. Given the $r_{(i,j)}$ relation and assuming that the ConvNet is producing the depth estimates z_i and z_j for the two corresponding joints, the loss for this pair is:

$$\mathcal{L}_{i,j} = \begin{cases} \log\left(1 + \exp(z_i - z_j)\right), & r_{(i,j)} = +1 \\ \log\left(1 + \exp(-z_i + z_j)\right), & r_{(i,j)} = -1 \\ (z_i - z_j)^2, & r_{(i,j)} = 0. \end{cases} \qquad (1)$$

This is a differentiable ranking loss expression, which has similarities with early works on the learning-to-rank literature [7] and was also adopted by [12] for apparent depth estimation. Intuitively, it enforces a large margin between the values z_i and z_j if one of them has been annotated as closer than the other, otherwise it enforces them to be equal. Denoting with $\mathcal{I}$ the set of pairs of joints that have been annotated with an ordinal relation, the complete expression for the loss takes the form:

$$\mathcal{L}_{rank} = \sum_{(i,j)\in\mathcal{I}} \mathcal{L}_{i,j}. \qquad (2)$$

An interesting property of this loss is that we do not require the relations for all pairs of joints to be available during training. The loss can be computed based only on the subset of pairs that have been annotated. Additionally, the relations do not have to be consistent, i.e., no strict global ordering is required. Instead, the ConvNet is allowed to learn a consensus from the provided relationships by minimizing the incurred loss. This is a helpful property in case there are ambiguities in the annotations.

3.2. Coordinate prediction for 3D pose

Our initial ConvNet only predicts the depths of the human joints. To enable full 3D pose reconstruction, we additionally need to precisely localize the corresponding joints on the image. Given the ConvNet used in the previous section, the most natural extension is to enrich its output by predicting the 2D coordinates of the joints as well. Thus, we predict $2N$ additional values which correspond to the pixel coordinates $\boldsymbol{w} = (x, y)$ of each joint. We consider this combination of 2D keypoints with ordinal depth as a form

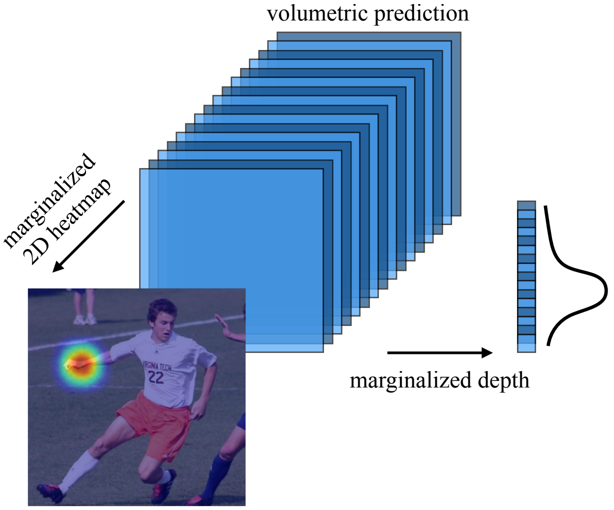


Figure 2: Visualization of the volumetric output for an individual joint. The predictions are volumetric, but in the absence of accurate 3D ground truth, the supervision is applied independently on the 2D image plane and the depth dimension. The marginalized likelihoods are computed by means of sum-pooling operations.

of *weak 3D information* and we refer to the corresponding ConvNet as the *weakly supervised version*.

Let us denote with $\boldsymbol{w}_n$ the ground truth 2D location for joint n, and with $\hat{\boldsymbol{w}}_n$ the corresponding ConvNet prediction. Assuming the availability of 2D keypoint annotations, the familiar $\mathcal{L}_2$ regression loss can be applied:

$$\mathcal{L}_{keyp} = \sum_{n=1}^{N} \|\boldsymbol{w}_n - \hat{\boldsymbol{w}}_n\|_2^2. \tag{3}$$

By combining the ranking loss for the values z_n and the regression loss for the keypoint coordinates $\boldsymbol{w}_n$, we can train the ConvNet end-to-end: $\mathcal{L} = \mathcal{L}_{rank} + \lambda \mathcal{L}_{keyp}$, where the value $\lambda = 100$ is used for our experiments.

3.3. Volumetric prediction for 3D pose

Apart from direct regression of the 3D pose coordinates, recent work has investigated the use of a volumetric representation for 3D human pose [32]. In this case, the space around the subject is discretized, and the ConvNet predicts per-voxel likelihoods for every joint in the 3D space. The training target for the volumetric space is a 3D Gaussian centered at the 3D location of each joint. However, without explicit 3D ground truth, supervising the same volume is not trivial. To demonstrate the general applicability of ordinal relations, we adapt this representation, to make it compatible with ordinal depth supervision as well.

To bypass the seemingly complex issue, we propose to preserve the volumetric structure of the output, but decompose the supervision a) in the 2D image plane and b) the z dimension (depth), as presented in Figure 2. Precisely, for every joint n, the ConvNet predicts score maps Ψ_n, which can be transformed to a probability distribution, by applying a *softmax* operation σ. So, the joint n is located in position $\boldsymbol{u} = (x, y, z)$ with probability $p(\boldsymbol{u}|n) = \sigma[\Psi_n]_{\boldsymbol{u}}$. The marginalized probability distribution in the 2D plane is:

$$p(x, y|n) = \sum_z p(\boldsymbol{u}|n), \tag{4}$$

and can be computed efficiently as a sum-pooling operation across all the slices of the volume. This operation is equivalent to adopting a weak perspective camera model. Similarly, the marginalized probability distribution for the depth dimension is:

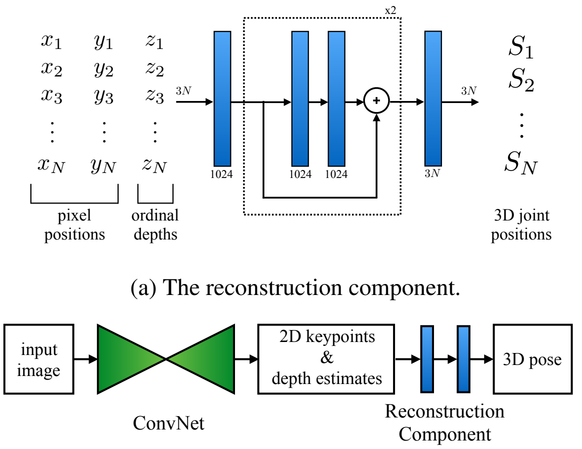
$$p(z|n) = \sum_{x,y} p(\boldsymbol{u}|n), \tag{5}$$

and can again be computed as a sum-pooling operation across all the pixels of a slice. This decomposition has the advantage that even if we do not have complete 3D ground truth, we can still supervise the ConvNet. The 2D image plane (values of equation 4) and the depth dimension (values of equation 5) are supervised independently, but they are connected by the underlying volumetric representation which enforces the 3D consistency. Our loss function takes the form: $\mathcal{L} = \mathcal{L}_{rank} + \lambda \mathcal{L}_{heat}$. The loss for the z-dimension, $\mathcal{L}_{rank}$, is the same ranking loss as before (equation 2), where we recover depth for each joint by taking the mean value of the estimated soft distribution: $z_n = \sum_z z p(z|n)$. For the x-y dimensions, the target for each keypoint is a heatmap with a Gaussian centered around its ground truth location and $\mathcal{L}_{heat}$ is an $\mathcal{L}_2$ loss between the predicted and the ground truth heatmaps [52, 33].

We stress here that the alterations presented up to this point refer only to the supervision type, without interfering with the network architecture. This allows most of the state-of-the-art discriminative ConvNets [64, 44, 48, 32] to be used as-is, and be complemented with the proposed ordinal depth supervision when 3D ground truth is not available.

3.4. Integration with a reconstruction component

The strength of the aforementioned networks is that they leverage image-based information to resolve the single-view depth ambiguities and produce depth estimates z_n that respect the ordinal depths of the human joints. However, the predicted depth values do not typically match the exact metric depths of the joints, since no full 3D pose example has been used to train the networks. This motivates us to enhance the architecture with our proposed *reconstruction* component, which takes as input the estimated 2D keypoints $\boldsymbol{w}_n$ and the ordinal depth estimates z_n, for all joints n, and reconstructs the 3D pose, $S \in \mathbb{R}^{n \times 3}$. This input-output relation is presented in Figure 3a. Conveniently, for



(a) The reconstruction component.

(b) Integration of the reconstruction component.

Figure 3: (a) The reconstruction component is a multi-layer perceptron with two bilinear units [25]. The input is the concatenation of the pixel locations of the joints (x_i, y_i), and the ordinal depths z_i, while the output is the 3D pose coordinates S_i. (b) Integration of the reconstruction module in the full framework. The ConvNet of Section 3.2 or 3.3 estimates 2D keypoint locations and depths which are used by the reconstruction module to predict a coherent 3D pose.

the training of this component we require only MoCap data, which are available in abundance. During training, we simply project each 3D pose skeleton to the 2D image plane. To simulate the input, we use the projected 2D joint locations and a noisy version of the depths of the joints, such that the majority of their ordinal relations are preserved, while their values might not necessarily match the actual depth. Denoting with $\hat{S}_i$ the output 3D joints of the ConvNet and with S_i the joints of the 3D pose that was used to generate the input, our supervision is an $\mathcal{L}_2$ loss:

$$\mathcal{L}_{3D} = \sum_{n=1}^{N} \|S_n - \hat{S}_n\|_2^2. \tag{6}$$

This module can be easily incorporated in an end-to-end framework by using as input the output of the ConvNet from Section 3.2 or Section 3.3. This is presented schematically in Figure 3b. The benefit from employing such a reconstruction module is demonstrated empirically in Section 4.

4. Empirical evaluation

This section concerns the empirical evaluation of the proposed approach. First, we present the benchmarks that we employed for quantitative and qualitative evaluation. Then, we provide some essential implementation details of the approach. Finally, quantitative and qualitative results are presented on the selected datasets.

4.1. Datasets

We employed two standard indoor benchmarks, Human3.6M [16] and HumanEva-I [42], along with a recent dataset captured in indoor and outdoor conditions, MPI-INF-3DHP [26, 27]. Additionally, we extended two popular 2D human pose datasets, Leeds Sports Pose dataset (LSP) [18] and MPII human pose dataset (MPII) [2] with ordinal depth annotations for the human joints.

Human3.6M: It is a large-scale dataset captured in an indoor environment that contains multiple subjects performing typical actions like "Eating" and "Walking". Following the most popular protocol (e.g., [67]), we train using subjects S1,S5,S6,S7, and S8 and test on subjects S9 and S11. The original videos are downsampled from 50fps to 10fps to remove redundancy. A single model is trained for all actions. Results are reported using the mean per joint error and the reconstruction error, which allows a Procrustes alignment of the prediction with the ground truth.

HumanEva-I: It is a smaller scale dataset compared to Human3.6M, including fewer users and actions. We follow the typical protocol (e.g., [4]), where the training sequences of subjects S1, S2 and S3 are used for training and the validation sequences of the same subjects are used for testing. We train a single model for all actions and users, and we report results using the reconstruction error.

MPI-INF-3DHP: It is a recent dataset that includes both indoor and outdoor scenes. We use it exclusively for evaluation, without employing the training data, to demonstrate robustness of the trained model under significant domain shift. Following the typical protocol ([26, 64]), results are reported using the PCK3D and the AUC metric.

LSP + MPII Ordinal: Leeds Sports Pose [18] and MPII human pose [2] are two of the most widely used benchmarks for 2D human pose. Here we extend both of them, offering ordinal depth annotations for the human joints. For LSP we annotate all the 2k images, while for MPII we annotate the subset of 13k images used by Lassner *et al.* [21].

Annotators were presented with a pair of joints for each image and answered which joint was closer to the camera. The option "ambiguous/hard to tell" was also offered. We considered 14 joints, excluding thorax and spine joints of MPII, which are often not used for training (e.g., [55]). The questions for each image were continued until a global ordering could be inferred for all the joints. By enforcing a global ordering we conveniently do not encounter any contradicting annotations. More importantly though, this approach significantly decreased annotation time. If the relative questions had to be answered for all joints, then we would require $\binom{14}{2} = 91$ questions for each image. In contrast, with the procedure we followed, we could get a global ordering with roughly 17 questions per image in the mean case. This resulted in 5 times faster annotation time. Additionally, we observed that annotators were much more ef-

Architecture		Supervision	Avg error
depth prediction		ordinal supervision	84.24
		direct regression	80.23
coordinate regression		weakly supervised	115.08
		fully supervised [32]	112.41
volume regression	one hourglass	weakly supervised	89.93
		fully supervised [32]	85.82
	two hourglasses	weakly supervised	79.03
		fully supervised [32]	69.77

Table 1: Effect of training with the actual 3D ground truth, versus employing weaker ordinal depth supervision on Human3.6M. The results are mean per joint errors (mm).

ficient when they were asked continuously about a specific pair of joints, instead of changing the pair of focus. As a result, we created groups of 50 images containing questions about the same pair of joints. This way we could get annotations at a rate of 3.5 secs per question, meaning that in total the procedure required roughly 1 minute per image.

We clarify that our goal for this dataset is to provide a novel information source (ordinal depth) for in-the-wild images. We do not use it for evaluation, since it is not a mm level accuracy benchmark like Human3.6M or HumanEva-I. Furthermore, the goal is not to conduct a computational study concerning the level of accuracy that humans perceive 3D poses as this has been already examined in the past [24]. In contrast, we use these annotations to demonstrate that: a) they can boost performance of 3D human pose estimation for standard benchmarks, and b) they assist our ConvNets to proper generalize and make them applicable in non-studio conditions, or in cases with significant domain shift.

4.2. Implementation details

For the ConvNets that predict 2D keypoints and/or depths, we follow the hourglass design [31]. When the output is in coordinate form (Sections 3.1 and 3.2), we use one hourglass with a fully connected layer in the end, while when we have volumetric target (Section 3.3), we use two hourglasses (unless stated otherwise). For comparisons with the state-of-the-art, we follow a mixed training strategy combining images with 3D ground truth from the respective dataset (Human3.6M or HumanEva-I), with LSP+MPII Ordinal images. For the LSP+MPII Ordinal examples, the loss is computed based on the human annotations (weak supervision), while for the respective dataset examples, the loss is computed based on the known ground truth (full supervision). We train the network with a batch size of 4, learning rate set to 2.5e-4, and using rmsprop for the optimization. Augmentation for rotation (±30°), scale (0.75-1.25) and flipping (left-right) is also used. The duration of the training depends on the size of the dataset (300k iterations for Human3.6M data only, 2.5M iterations for mixed Human3.6M and LSP+MPII Ordinal data, 1.5M iterations for mixed HumanEva-I and LSP+MPII Ordinal data). For the reconstruction component (Section 3.4), we follow the design of [25]. We train the network with a batch size of 64, learning rate set to 2.5e-4, we use rmsprop for the optimization, and the training lasts for 200k iterations.

				Avg error
Human3.6M				71.9
Human3.6M	+ 2D keyp			66.6
Human3.6M	+ 2D keyp	+ Ord		62.1
Human3.6M	+ 2D keyp		+ Rec	59.1
Human3.6M	+ 2D keyp	+ Ord	+ Rec	**56.2**

Table 2: Ablative study on Human3.6M demonstrating the effect of incorporating additional data sources in the training procedure (2D keypoints and ordinal depth relations), as well as integrating a rconstruction component. The numbers are mean per joint errors (mm).

	PCK3D	AUC
Human3.6M	17.1	6.3
Human3.6M + 2D keyp	44.3	19.8
Human3.6M + 2D keyp + Ord	**71.9**	**35.3**

Table 3: Ablative study on MPI-INF-3DHP demonstrating that supervision through our ordinal annotations is important for proper generalization.

4.3. Ablative studies

Ordinal supervision: First, we examine the effect of using ordinal depth supervision versus employing the actual 3D groudtruth for training. For this part, we focus on Human3.6M which is a large scale benchmark and provides 3D ground truth to perform the quantitative comparison. To define the ordinal depth relations, the depth values for each pair of joints are considered. If they differ less than 100mm, then the corresponding relation is set to $r = 0$ (similar depth). Otherwise, it is set to $r = \pm 1$, depending on which joint is closer. Since for this comparison we want to focus on the form of supervision, this is the only set of experiments that uses ordinal depth relations inferred from 3D ground truth. For the remaining evaluations, all ordinal depth relations were provided by human annotators.

Following the analysis of Section 3, we explore three different prediction schemes, i.e., depth prediction, coordinate regression and volume regression. For each one of them, we compare a version where ordinal supervision is used, versus

	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SitingD	Smoke	Wait	WalkD	Walk	WalkT	Avg
Tekin *et al.* [49] (CVPR'16)	102.4	147.2	88.8	125.3	118.0	182.7	112.4	129.2	138.9	224.9	118.4	138.8	126.3	55.1	65.8	125.0
Zhou *et al.* [67] (CVPR'16)	87.4	109.3	87.1	103.2	116.2	143.3	106.9	99.8	124.5	199.2	107.4	118.1	114.2	79.4	97.7	113.0
Du *et al.* [14] (ECCV'16)	85.1	112.7	104.9	122.1	139.1	135.9	105.9	166.2	117.5	226.9	120.0	117.7	137.4	99.3	106.5	126.5
Zhou *et al.* [65] (ECCVW'16)	91.8	102.4	96.7	98.8	113.4	125.2	90.0	93.8	132.2	159.0	107.0	94.4	126.0	79.0	99.0	107.3
Chen *et al.* [10] (CVPR'17)	89.9	97.6	90.0	107.9	107.3	139.2	93.6	136.1	133.1	240.1	106.7	106.2	114.1	87.0	90.6	114.2
Tome *et al.* [51] (CVPR'17)	65.0	73.5	76.8	86.4	86.3	110.7	68.9	74.8	110.2	173.9	85.0	85.8	86.3	71.4	73.1	88.4
Rogez *et al.* [40] (CVPR'17)	76.2	80.2	75.8	83.3	92.2	105.7	79.0	71.7	105.9	127.1	88.0	83.7	86.6	64.9	84.0	87.7
Pavlakos *et al.* [32] (CVPR'17)	67.4	71.9	66.7	69.1	72.0	77.0	65.0	68.3	83.7	96.5	71.7	65.8	74.9	59.1	63.2	71.9
Nie *et al.* [60] (ICCV'17)	90.1	88.2	85.7	95.6	103.9	103.0	92.4	90.4	117.9	136.4	98.5	94.4	90.6	86.0	89.5	97.5
Tekin *et al.* [48] (ICCV'17)	54.2	61.4	60.2	61.2	79.4	78.3	63.1	81.6	70.1	107.3	69.3	70.3	74.3	51.8	74.3	69.7
Zhou *et al.* [64] (ICCV'17)	54.8	60.7	58.2	71.4	62.0	65.5	53.8	55.6	75.2	111.6	64.2	66.1	51.4	63.2	55.3	64.9
Martinez *et al.* [25] (ICCV'17)	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Ours	**48.5**	**54.4**	**54.4**	**52.0**	**59.4**	**65.3**	**49.9**	**52.9**	**65.8**	**71.1**	**56.6**	**52.9**	**60.9**	**44.7**	**47.8**	**56.2**

Table 4: Detailed results on Human3.6M [16]. Numbers are mean per joint errors (mm). The results of all approaches are obtained from the original papers. We outperform all other approaches across the table.

	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SitingD	Smoke	Wait	WalkD	Walk	WalkT	Avg
Akhter & Black [1]* (CVPR'15)	199.2	177.6	161.8	197.8	176.2	186.5	195.4	167.3	160.7	173.7	177.8	181.9	176.2	198.6	192.7	181.1
Ramakrishna *et al.* [38]* (ECCV'12)	137.4	149.3	141.6	154.3	157.7	158.9	141.8	158.1	168.6	175.6	160.4	161.7	150.0	174.8	150.2	157.3
Zhou *et al.* [66]* (CVPR'15)	99.7	95.8	87.9	116.8	108.3	107.3	93.5	95.3	109.1	137.5	106.0	102.2	106.5	110.4	115.2	106.7
Bogo *et al.* [5] (ECCV'16)	62.0	60.2	67.8	76.5	92.1	77.0	73.0	75.3	100.3	137.3	83.4	77.3	86.8	79.7	87.7	82.3
Moreno-Noguer [28] (CVPR'17)	66.1	61.7	84.5	73.7	65.2	67.2	60.9	67.3	103.5	74.6	92.6	69.6	71.5	78.0	73.2	74.0
Pavlakos *et al.* [32] (CVPR'17)	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
Martinez *et al.* [25] (ICCV'17)	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
Ours	**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**

Table 5: Detailed results on Human3.6M [16]. Numbers are reconstruction errors. The results of all approaches are obtained from the original papers, except for (*), which were obtained from [5]. We outperform all other approaches across the table.

employing the actual 3D ground truth for training. The detailed results are presented in Table 1. Interestingly, in all cases, the weaker ordinal supervision signal is competitive and achieves results very close to the fully supervised baseline. The gap increases only when we employ more powerful architectures, i.e., the volume regression case with two hourglass components. In fact, in this case the average error is already very low (below 80mm), and one would expect that for even lower prediction errors, the highly accurate 3D ground truth would be necessary for training.

Improving 3D pose detectors: After the sanity check that ordinal supervision is competitive to training with the full 3D ground truth, we explore using ordinal depth annotations provided by humans, to boost the performance of a standard ConvNet for 3D human pose [32]. As detailed in Section 4.2, we follow a mixed training strategy, leveraging Human3.6M images with 3D ground truth and LSP+MPII Ordinal images with our annotations. Data augmentation using natural images with 2D keypoint annotations is a standard practice [48, 26, 36, 64, 44], but here we also consider the effect of our ordinal depth supervision. Optionally, the reconstruction component can be used at the end of the network, helping with coherent 3D pose prediction. The detailed results of the ablative study are presented in Table 2.

Unsurprisingly, using more training examples improves performance. The supervision with 2D keypoints is helpful (line 2), however the addition of our ordinal depth supervision provides novel information to the network and further improves the results (line 3). The refinement step using the reconstruction module (lines 4 and 5) is also beneficial, and helps providing coherent 3D pose results. In fact, the last line corresponds to state-of-the-art results for this dataset, which we discuss in more detail in Section 4.4.

Robustness to domain shift: Besides boosting current state-of-the-art models, we ultimately aspire to use our ordinal supervision for better generalization of the trained models so that they are applicable for in-the-wild images. To demonstrate this potential, we test our approach on the MPI-INF-3DHP dataset. This dataset is not considered exactly in-the-wild, but has a significant domain shift compared to Human3.6M. The complete results for this ablative experiment are presented in Table 3. Interestingly, the model trained only on Human3.6M data (line 1) has embarrassing performance, because of heavy overfitting. Using additional in-the-wild images with 2D keypoints (line 2) is helpful, but from inspection of the results, the benefit comes mainly from better 2D pose estimates, while depth prediction is generally mediocre. The best generalization comes after incorporating also the ordinal depth supervision (line 3), elevating the model to state-of-the-art results.

	Walking			Jogging			
	S1	S3	S3	S1	S2	S3	Avg
Radwan *et al.* [37]	75.1	99.8	93.8	79.2	89.8	99.4	89.5
Wang *et al.* [54]	71.9	75.7	85.3	62.6	77.7	54.4	71.3
Simo-Serra *et al.* [43]	65.1	48.6	73.5	74.2	46.6	32.2	56.7
Bo *et al.* [4]	46.4	30.3	64.9	64.5	48.0	38.2	48.7
Kostrikov *et al.* [20]	44.0	30.9	41.7	57.2	35.0	33.3	40.3
Yasin *et al.* [62]	35.8	32.4	41.6	46.6	41.4	35.4	38.9
Moreno-Noguer [28]	19.7	13.0	**24.9**	39.7	20.0	21.0	26.9
Pavlakos *et al.* [32]	22.1	21.9	29.0	29.8	23.6	26.0	25.5
Martinez *et al.* [25]	19.7	17.4	46.8	26.9	18.2	18.6	24.6
Ours	**18.8**	**12.7**	29.2	**23.5**	**15.4**	**14.5**	**18.3**

Table 6: Results on the HumanEva-I [42] dataset. Numbers are reconstruction errors (mm). The results of all approaches are obtained from the original papers.

4.4. Comparison with state-of-the-art

Human3.6M: We use for evaluation the same ConvNet with the previous section, which follows a mixed training strategy and includes the reconstruction component. The detailed results in terms of mean per joint error and reconstruction error are presented in Tables 4 and 5 respectively. Our complete approach achieves state-of-the-art results across all actions and metrics, with relative error reduction over 10% on average. Since most other works (e.g., [51, 64, 48, 25]) also use in-the-wild images with 2D keypoints for supervision, most of the improvement for our approach comes from augmenting training with ordinal depth relations for these examples. In particular, the error decrease with respect to previous work is more significant for challenging actions like Sitting Down, Photo or Sitting, with a lot of self-occlusions and rare poses. This benefit can be attributed to the greater variety of the LSP+MPII Ordinal images not just in terms of appearance (this also benefits the other approaches), but mainly in terms of 3D poses which are observed from our ConvNet in a weak 3D form.

HumanEva-I: The ConvNet architecture remains the same, where HumanEva-I and LSP+MPII Ordinal images are used for mixed training. The reconstruction component is trained only on HumanEva-I MoCap. Our results are presented in Table 6 and show important accuracy benefit over previous approaches. On average, the relative error reduction is again over 10%, which is a solid improvement considering the numbers for this dataset have mostly saturated.

MPI-INF-3DHP: For MPI-INF-3DHP, we report results using the same ConvNet we trained for Human3.6M, with Human3.6M and LSP+MPII Ordinal images. In Table 7 we compare with two recent baselines which are not trained on this dataset, and we outperform them, with particularly large margin for the Outdoor sequence.

Approach	Studio GS	Studio no GS	Outdoor	All	All
	3DPCK	3DPCK	3DPCK	3DPCK	AUC
Mehta *et al.* [26]	70.8	62.3	58.8	64.7	31.7
Zhou *et al.* [64]	71.1	**64.7**	72.7	69.2	32.5
Ours	**76.5**	63.1	**77.5**	**71.9**	**35.3**

Table 7: Detailed results on the test set of MPI-INF-3DHP [26]. The results for all approaches are taken from the original papers. No training data from this dataset have been used for training by any method.

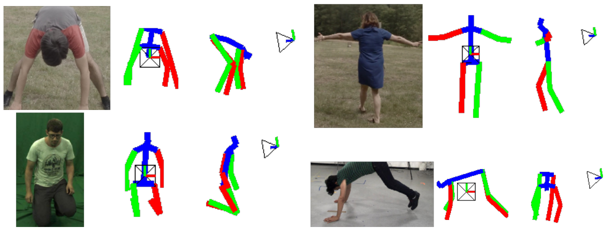


Figure 4: Typical qualitative results from MPI-INF-3DHP, from the original and a novel viewpoint.

4.5. Qualitative evaluation

In Figure 4 we have collected a sample of 3D pose output for our approach, focusing on MPI-INF-3DHP, since it is the main dataset that we evaluate without touching the training data. A richer collection of success and failure examples is included in the supplementary material.

5. Summary

The goal of this paper was to present a solution for training end-to-end ConvNets for 3D human pose estimation in the absence of accurate 3D ground truth, by using a weaker supervision signal in the form of ordinal depth relations of the joints. We investigated the flexibility of these ordinal relations by incorporating them in recent ConvNet architectures for 3D human pose and demonstrated competitive performance with their fully supervised versions. Furthermore, we extended the MPII and LSP datasets with ordinal depth annotations for the human joints, allowing us to present compelling results for non-studio conditions. Finally, these annotations were incorporated in the training procedure of recent ConvNets for 3D human pose, achieving state-of-the-art results in the standard benchmarks.

Project Page: https://www.seas.upenn.edu/~pavlakos/projects/ordinal

Acknowledgements: We gratefully appreciate support through the following grants: NSF-IIP-1439681 (I/UCRC), ARL RCTA W911NF-10-2-0016, ONR N00014-17-1-2093, DARPA FLA program and NSF/IUCRC.

References

[1] I. Akhter and M. J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *CVPR*, 2015. 7

[2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 2, 5

[3] S. Bell, K. Bala, and N. Snavely. Intrinsic images in the wild. *ACM Transactions on Graphics (TOG)*, 33(4):159, 2014. 2

[4] L. Bo and C. Sminchisescu. Twin Gaussian processes for structured prediction. *IJCV*, 87(1-2):28–52, 2010. 5, 8

[5] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, 2016. 1, 2, 7

[6] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3D human pose annotations. In *ICCV*, 2009. 3

[7] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *ICML*, 2005. 3

[8] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *ICML*, 2007. 3

[9] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *CVPR*, 2017. 1

[10] C.-H. Chen and D. Ramanan. 3D human pose estimation = 2D pose estimation + matching. In *CVPR*, 2017. 2, 7

[11] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, 2016. 1

[12] W. Chen, Z. Fu, D. Yang, and J. Deng. Single-image depth perception in the wild. In *NIPS*, 2016. 2, 3

[13] W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen. Synthesizing training images for boosting human 3D pose estimation. In *3DV*, 2016. 1, 2

[14] Y. Du, Y. Wong, Y. Liu, F. Han, Y. Gui, Z. Wang, M. Kankanhalli, and W. Geng. Marker-less 3D human motion capture with monocular image sequence and height-maps. In *ECCV*, 2016. 7

[15] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, B. Andres, and B. Schiele. Articulated multi-person tracking in the wild. In *CVPR*, 2016. 1

[16] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *PAMI*, 36(7):1325–1339, 2014. 1, 5, 7

[17] E. Jahangiri and A. L. Yuille. Generating multiple hypotheses for human 3D pose consistent with 2D joint detections. In *ICCVW*, 2017. 2

[18] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 2, 5

[19] V. Kazemi, M. Burenius, H. Azizpour, and J. Sullivan. Multi-view body part recognition with random forests. In *BMVC*, 2013. 1

[20] I. Kostrikov and J. Gall. Depth sweep regression forests for estimating 3D human pose from images. In *BMVC*, 2014. 8

[21] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *CVPR*, 2017. 3, 5

[22] S. Li and A. B. Chan. 3D human pose estimation from monocular images with deep convolutional neural network. In *ACCV*, 2014. 2

[23] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR*, 2011. 3

[24] E. Marinoiu, D. Papava, and C. Sminchisescu. Pictorial human spaces: A computational study on the human perception of 3D articulated poses. *IJCV*, 119(2):194–215, 2016. 2, 6

[25] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3D human pose estimation. In *ICCV*, 2017. 1, 2, 5, 6, 7, 8

[26] D. Mehta, H. Rhodin, D. Casas, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *3DV*, 2017. 1, 2, 5, 7, 8

[27] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. VNect: Real-time 3D human pose estimation with a single RGB camera. *ACM Transactions on Graphics*, 36, 2017. 5

[28] F. Moreno-Noguer. 3D human pose estimation from a single image via distance matrix regression. In *CVPR*, 2017. 2, 7, 8

[29] T. Narihira, M. Maire, and S. X. Yu. Learning lightness from human judgement on relative reflectance. In *CVPR*, 2015. 3

[30] A. Newell, Z. Huang, and J. Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *NIPS*, 2017. 1

[31] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 1, 6

[32] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *CVPR*, 2017. 2, 3, 4, 6, 7, 8

[33] T. Pfister, J. Charles, and A. Zisserman. Flowing convnets for human pose estimation in videos. In *ICCV*, 2015. 4

[34] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele. DeepCut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, 2016. 1

[35] G. Pons-Moll, D. J. Fleet, and B. Rosenhahn. Posebits for monocular human pose estimation. In *CVPR*, 2014. 3

[36] A.-I. Popa, M. Zanfir, and C. Sminchisescu. Deep multitask architecture for integrated 2D and 3D human sensing. In *CVPR*, 2017. 7

[37] I. Radwan, A. Dhall, and R. Goecke. Monocular image 3D human pose estimation under self-occlusion. In *ICCV*, 2013. 8

[38] V. Ramakrishna, T. Kanade, and Y. Sheikh. Reconstructing 3D human pose from 2D image landmarks. In *ECCV*, 2012. 7

[39] G. Rogez and C. Schmid. MoCap-guided data augmentation for 3D pose estimation in the wild. In *NIPS*, 2016. 2

[40] G. Rogez, P. Weinzaepfel, and C. Schmid. LCR-Net: Localization-classification-regression for human pose. In *CVPR*, 2017. 2, 7

[41] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris. 3D human pose estimation: A review of the literature and analysis of covariates. *CVIU*, 152:1–20, 2016. 2

[42] L. Sigal, A. O. Balan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87(1-2):4–27, 2010. 1, 5, 8

[43] E. Simo-Serra, A. Quattoni, C. Torras, and F. Moreno-Noguer. A joint model for 2D and 3D pose estimation from a single image. In *CVPR*, 2013. 8

[44] X. Sun, J. Shang, S. Liang, and Y. Wei. Compositional human pose regression. In *ICCV*, 2017. 2, 4, 7

[45] C. J. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. In *CVPR*, 2000. 1

[46] M. Taylor, J. Guiver, S. Robertson, and T. Minka. SoftRank: optimizing non-smooth rank metrics. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 77–86. ACM, 2008. 3

[47] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua. Structured prediction of 3D human pose with deep neural networks. In *BMVC*, 2016. 2

[48] B. Tekin, P. Marquez Neila, M. Salzmann, and P. Fua. Learning to fuse 2D and 3D image cues for monocular body pose estimation. In *ICCV*, 2017. 2, 4, 7, 8

[49] B. Tekin, A. Rozantsev, V. Lepetit, and P. Fua. Direct prediction of 3D body poses from motion compensated sequences. In *CVPR*, 2016. 7

[50] J. T. Todd and J. F. Norman. The visual perception of 3-D shape from multiple cues: Are observers capable of perceiving metric structure? *Perception & Psychophysics*, 65(1):31–47, 2003. 2

[51] D. Tome, C. Russell, and L. Agapito. Lifting from the deep: Convolutional 3D pose estimation from a single image. In *CVPR*, 2017. 2, 7, 8

[52] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, 2014. 4

[53] G. Varol, J. Romero, X. Martin, N. Mahmood, M. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *CVPR*, 2017. 1, 2

[54] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao. Robust estimation of 3D human poses from a single image. In *CVPR*, 2014. 8

[55] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016. 1, 5

[56] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman. Single image 3D interpreter network. In *ECCV*, 2016. 2

[57] F. Xia, P. Wang, L.-C. Chen, and A. L. Yuille. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In *ECCV*, 2016. 1

[58] Y. Xiang, W. Kim, W. Chen, J. Ji, C. Choy, H. Su, R. Mottaghi, L. Guibas, and S. Savarese. ObjectNet3D: A large scale database for 3D object recognition. In *ECCV*, 2016. 3

[59] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond PASCAL: A benchmark for 3D object detection in the wild. In *WACV*, 2014. 3

[60] B. Xiaohan Nie, P. Wei, and S.-C. Zhu. Monocular 3D human pose estimation by predicting depth on joints. In *ICCV*, 2017. 7

[61] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang. Learning feature pyramids for human pose estimation. In *ICCV*, 2017. 1

[62] H. Yasin, U. Iqbal, B. Krüger, A. Weber, and J. Gall. A dual-source approach for 3D pose estimation from a single image. In *CVPR*, 2016. 8

[63] T. Zhou, P. Krähenbühl, and A. A. Efros. Learning data-driven reflectance priors for intrinsic image decomposition. In *ICCV*, 2015. 3

[64] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Towards 3D human pose estimation in the wild: A weakly-supervised approach. In *ICCV*, 2017. 2, 4, 5, 7, 8

[65] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei. Deep kinematic pose regression. In *ECCVW*, 2016. 7

[66] X. Zhou, M. Zhu, S. Leonardos, and K. Daniilidis. Sparse representation for 3D shape estimation: A convex relaxation approach. *PAMI*, 2016. 7

[67] X. Zhou, M. Zhu, S. Leonardos, K. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3D human pose estimation from monocular video. In *CVPR*, 2016. 2, 5, 7

[68] X. Zhou, M. Zhu, G. Pavlakos, S. Leonardos, K. G. Derpanis, and K. Daniilidis. MonoCap: Monocular human motion capture using a CNN coupled with a geometric prior. *PAMI*, 2018. 2

[69] D. Zoran, P. Isola, D. Krishnan, and W. T. Freeman. Learning ordinal relationships for mid-level vision. In *ICCV*, 2015. 3