Structure-Aware and Temporally Coherent 3D Human Pose Estimation

Rishabh Dabral Anurag Mundhada Uday Kusupati Safeer Afaque Arjun Jain

Department of Computer Science, IIT Bombay

{rdabral@cse, anuragmundhada@, udaykusupati@cse, safeer@cse, ajain@cse}.iitb.ac.in

Abstract

Deep learning methods for 3D human pose estimation from RGB images require a huge amount of domain-specific labeled data for good in-the-wild performance. However, obtaining annotated 3D pose data requires a complex motion capture setup which is generally limited to controlled settings. We propose a semi-supervised learning method using a structure-aware loss function which is able to utilize abundant 2D data to learn 3D information. Furthermore, we present a simple temporal network which uses additional context present in pose sequences to improve and temporally harmonize the pose estimates. Our complete pipeline improves upon the state-of-the-art by 11.8%, and works at 30 FPS on a commodity graphics card.

1. Introduction

Estimating human motion from images and videos is the key to unlocking several applications in robotics, human computer interaction, surveillance and human sensing. It is an active and challenging area of research owing to its ill-constrained nature, self-occlusions, background and viewpoint invariance, etc [1, 2]. In recent years, advances in deep learning techniques, better hardware and, most importantly, large datasets [3, 4, 5, 6, 7] have facilitated significant improvement in the state-of-the-art. In this paper, we present a Convolutional Neural Network (ConvNet) based approach that estimates 3D human poses from monocular images or videos in real-time.

For estimating 3D human pose, most state-of-the-art methods either perform direct regression of joint coordinates [8, 9, 10] or infer 3D pose from 2D heatmaps in a pipelined approach [5, 11, 12, 13, 14]. While these methods have impressive performance on 3D benchmark datasets, most of them do not generalize well to in-the-wild images. This can be attributed to a lack of annotated 3D human-pose datasets in *natural settings*, as opposed to the abundant in-the-wild 2D human-pose datasets [15, 16] available. To mitigate this issue, past approaches have proposed using synthetic datasets [7, 6], green-screen compositing [5, 17], domain adaptation [7], transfer learning from intermediate 2D pose estimation tasks [5, 8], and mixed learning of 2D and 3D pose [14, 10].

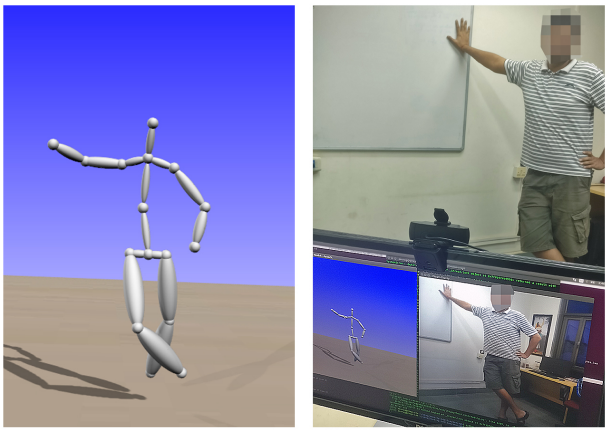


Figure 1. We propose a real-time approach for 3D human pose estimation from RGB images and video, that beats all previously published benchmarks on the standard datasets.

Mixed learning of 2D and 3D pose, in particular, leads to better lower level features, which leads to a significant increase in both accuracy and generalizability. This is on account of the availability of large 2D pose datasets in highly diverse settings. Zhou et al. [14] proposed a semi-supervised loss which exploits bone length ratio priors to provide depth supervision. We build on their approach and introduce a new, structure-aware loss function, that enables us to obtain depth supervision from 2D data.

We incorporate information about joint angle limits and bone length constraints, using three regularization terms in the loss function. In the first term we penalize illegal extension of the knee and elbow joints. Human elbows and knees are constrained to move within a limited angle of flexion while extensions are physically not possible. Penalizing such poses pushes the ConvNet away from the space of implausible poses. In the second regularization term, we penalize asymmetric lengths of left/right bone pairs. We also use the geometric regularization proposed in [14]. These three terms together provide highly effective depth supervi-

sion in the absence of 3D-labeled data, and guide the model towards better local minima. We validate the effectiveness of our approach by quantitative as well as qualitative evaluations (Section 4). Using our structure-aware loss function alone, we improve the Mean Per Joint Position Error (MPJPE) over the state-of-the-art on Human 3.6M dataset by 6%, and MPI-INF-3DHP test set by 7.7%.

Additionally, we exploit additional cues present in video frames to predict temporally coherent poses. Interactive applications require smooth movements which are typically ensured by using simple low-pass filters applied to the final output, at the cost of introducing lag and making the motion seem 'uncanny'. We show that a simple, fully-connected network at the output of our ConvNet performs extremely well at modeling temporal as well as structural correlations of the body joints, and also results in temporally harmonized output. Using our temporal model, we report an additional 7% improvement on Human3.6M and 2% on MPI-INF-3DHP and demonstrate real-time performance of the combined pipeline at 30fps. Our final model achieves an MPJPE of *52.1mm* on Human3.6M [3] and *103.8mm* on MPI-INF-3DHP [5] test sets which is 11.8% and 12% better than the state-of-the-art, respectively.

In summary, our contribution through this work is twofold: *Firstly*, we significantly improve the state-of-the art by introducing a semi-supervised, structure-aware loss function, that enables learning of depth information from 2D labeled data. *Secondly*, we show that a simple, fully-connected network is highly effective at learning structural and temporal correlations and improves and stabilizes the output stream of poses.

2. Related Work

We review past work related to Human Pose Estimation from three viewpoints: (1) ConvNet architectures and training strategies, (2) Utilizing structural constraints of human bodies, and (3) 3D pose estimation from video. The reader is referred to [2] for a detailed review of the literature.

ConvNet architectures: Most existing ConvNet based approaches either directly regress 3D poses from the input image [10, 8, 18, 13] or infer 3D pose from 2D pose in a pipelined approach [19, 14, 17, 11, 12]. Some approaches utilize novel formulation of the output space like volumetric heatmaps [20], defining a pose using bones instead of joints [10], or regressing location maps [17]. Using a 2D-to-3D pipeline enables the use of rich 2D pose datasets. A few approaches use statistical priors [13][2012 papers] to lift 2D poses to 3D. Chen et al. [21] and Yasin et al. [22] use a pose library to retrieve the nearest 3D pose given a 2D pose prediction. Recent ConvNet based approaches [17, 23, 14, 10, 13, 20] have reported substantial improvements in in-the-wild performances by using the more diverse 2D pose datasets to pre-train or jointly train

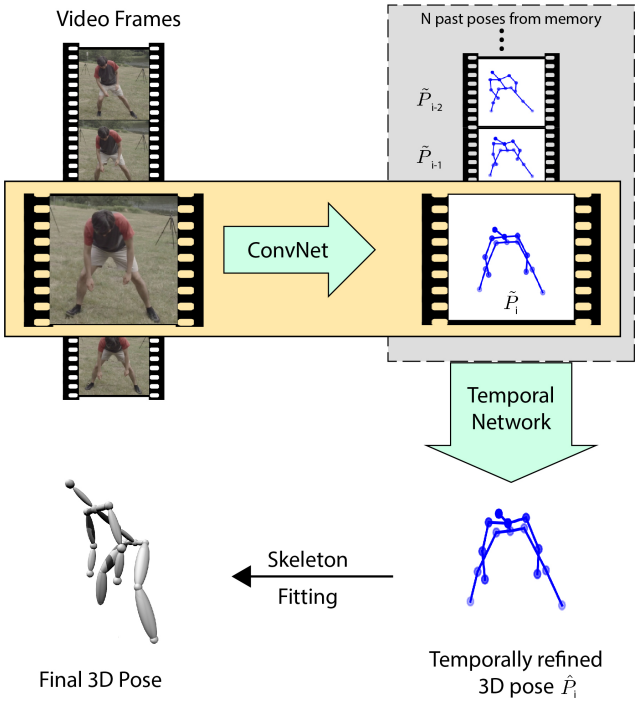


Figure 2. Overall pipeline of our method: We sequentially pass the video frames to a ConvNet that produces 3D pose outputs (one at a time). Next, the prediction is temporally refined by passing a context of past N frames along with the current frame to a temporal model. Finally, skeleton fitting may be performed as an optional step depending upon the application requirement.

their 2D prediction modules.

Utilizing structural information: The structure of the human skeleton is constrained by fixed bone lengths, joint angle limits, and limb interpenetration constraints. Bogo et al. [24] penalize body-part interpenetration and illegal joint angles in their objective function for finding SMPL [25] based pose-parameters. Akhter and Black [26] learn pose-dependent joint angle limits for lifting 2D poses to 3D. Sun et al. [10] propose a structure-aware loss function that operates on bone based pose parameterization. Zhou et al. [14] introduce a geometric loss function that ensures the consistency of bone-length ratios in the pose predictions of their ConvNet model. In our work, we use joint angle limits and bone lengths in our loss function for semi-supervised learning.

Utilizing temporal information: Direct estimation of 3D pose from disjointed images leads to temporally incoherent output with visible jitter and varying bone lengths. 3D pose estimates from a video can be improved by using simple filters or temporal priors. Mehta et al. [17] propose a real-time approach which penalizes acceleration and depth velocity in an optimization step after generating pose proposals using a ConvNet. They also smoothen the output poses by using a filter optimized for interactive sys-

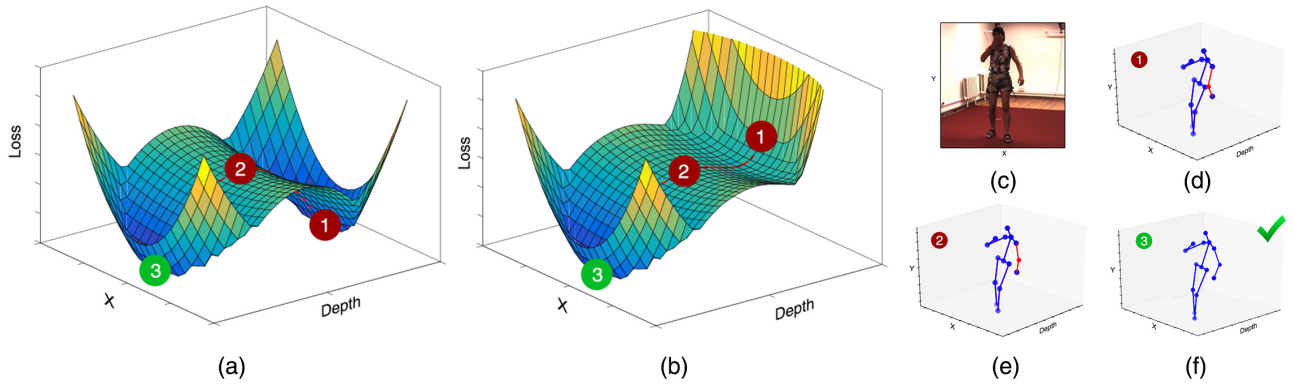


Figure 3. Evolution of the semi-supervised loss L_z^{2D} with the addition of the illegal angle penalty, for the left elbow. The loss surfaces are obtained using the $\lambda_a \mathcal{L}_a$ and $\lambda_s \mathcal{L}_s$ terms from Equation 2, by varying the x-coordinate and depth of the elbow joint while keeping y fixed. (a) shows the loss function surface formed by 2D mean square error and bone-length regularization. (b) shows the transformed loss function after adding angle regularization. (c) is the input image, while (d), (e) and (f) are the poses corresponding to locations (1), (2) and (3) in the two loss surfaces. The illegal angle penalty increases the loss for pose (1), which has the elbow bent backwards. Pose (2) has a legal joint angle, but the bone length regularizer contributes to its high loss. Pose (3) is correct. Without the angle loss, loss at (1) and (3) are equal and we cannot discern between the two points.

tems [27]. Zhou et al. [13] introduce a first order smoothening prior in their temporal optimization step. Alldieck et al. [28] exploit 2D optical flow features to predict 3D poses from videos. Wei et al. [29] exploit physics-based constraints to realistically interpolate 3D motion between video keyframes. There have also been attempts to *learn* a motion model for human motion. Urtasun et al. [30] learn a strong activity specific motion prior using linear models. Park et al. [31] use a motion library to find the nearest motion given a set of 2D pose predictions and then iteratively fine-tune the retrieved motion. Recently, Lin et al. [12] used Long-Short Term Memory (LSTM) networks to learn temporal dependencies from the intermediate features of their ConvNet based architecture. In a similar attempt, Coskun et al. [32] propose to use LSTMs to design a Kalman filter that learns a motion model for humans. In contrast, we use a simple yet effective model that captures short-term interplay of past poses and consequently stitches the output frames in a temporally consistent manner.

3. 3D Pose Estimation Pipeline

We define a 3D human pose $P = \{p_1, p_2, \ldots, p_K\}$ by the positions of $K = 16$ body joints in Euclidean space. These joint positions are defined relative to a root joint, which we fix as the pelvis. The input is a stream of RGB images $I = \{\ldots, I_{i-1}, I_i\}$ from a live stream or a video.

Fig. 2 shows the schematic of our pipeline. We reuse the 3D pose network design of Zhou et al. [14], which is itself based on the popular stacked hourglass network [33]. We call this network *UnitPoseNet* and train it with our proposed semi-supervised loss. The input images are sequentially fed to UnitPoseNet which outputs a 3D pose estimate $\tilde{P}_i$ corresponding to the i^{th} input image I_i. $\tilde{P}_i$ is further refined by our temporal model which we call *TimePoseNet*, that takes the pose estimates of the *past* N frames along with $\tilde{P}_i$ to generate a temporally harmonized pose $\hat{P}_i$. Finally, a simple skeleton fitting step which preserves the directions of the bone vectors is performed optionally. We train UnitPoseNet on Human3.6M, MPI-INF-3DHP (3D) and MPII (2D) datasets, and TimePoseNet on Human3.6M and MPI-INF-3DHP datasets where 3D-pose annotated videos are available. The next two sections explain these models in detail. The subscript i is dropped while discussing UnitPoseNet for better readability.

3.1. UnitPoseNet

We use the highly popular stacked hourglass network architecture [33] along with the depth regressor extension suggested in [14] (ref. supplementary material). This network choice allows us to mix 3D and 2D data while training and use our structure-aware loss for semi-supervised learning of depth from images with only 2D annotated data. The stacked hourglass module outputs 2D heatmaps for each joint which are used to infer 2D joint positions $\tilde{P}_{xy}$. The depth regressor module uses the intermediate features of the hourglass to regress the depths of the body joints $\tilde{P}_z$. It consists of a series of four residual modules [34] followed by a fully connected layer at the end. We trained the architecture from scratch using our semi-supervised approach.

3.1.1 Loss Function

The stacked hourglass network is always trained using Euclidean loss between the predicted heatmaps and the ground-truth heatmaps, generated using a Gaussian ker-

nel around the ground truth [35]. The depth regressor is trained using two different losses depending on the availability of 2D or 3D labels. When 3D labels are available (Human3.6M and MPI-INF-3DHP), we use the Euclidean loss:

$$\mathcal{L}_z^{3D}(\tilde{P}_z, P_z^{gt}) = \lambda_d ||\tilde{P}_z - P_z^{gt}||_2 \quad (1)$$

In the absence of 3D annotated data, we use our structure-aware loss $\mathcal{L}_z^{2D}$ to obtain depth supervision. $\mathcal{L}_z^{2D}$ requires only the 2D ground truth (x and y) to supervise the depth regressor. We define two regularization terms - the Illegal Angle Loss $\mathcal{L}_a$ and the Symmetry Loss $\mathcal{L}_s$ which are defined using the ground truth 2D annotations P_{xy}^{gt} and the predicted depth $\tilde{P}_z$. For both the losses, we define a joint as $J = \{P_x^{gt}, P_y^{gt}, \tilde{P}_z\}$. These terms, along with Zhou et al's [14] geometric regularizer, $\mathcal{L}_g$, generate a loss surface as shown in Fig. 3. The depth regressor's loss function $\mathcal{L}_z^{2D}$ for 2D annotated data is formulated as:

$$\begin{aligned}\mathcal{L}_z^{2D}(\tilde{P}_z, P_{xy}^{gt}) = \lambda_a \mathcal{L}_a(\tilde{P}_z, P_{xy}^{gt}) + \lambda_s \mathcal{L}_s(\tilde{P}_z, P_{xy}^{gt}) \\ + \lambda_g \mathcal{L}_g(\tilde{P}_z, P_{xy}^{gt})\end{aligned} \quad (2)$$

Illegal Angle Loss ($\mathcal{L}_a$): Most body joints are constrained to move within certain angle limits within their degrees of freedom. For given 2D joint coordinates, there exist multiple possible 3D joint configurations. However, only a few of these 3D poses are legal. $\mathcal{L}_a$ encapsulates this prior for the knee and elbow joints, specifically that knee and elbow joints cannot flex to angles greater than $180°$.

Inferring the joint angles is non-trivial from a set of joint positions. Also, once illegal angles are identified, appropriating losses and gradients for the implicated joints is non-trivial. We illustrate our formulation in Fig. 4, and explain it here for the right elbow joint. Subscripts n, s, e, w, k denote neck, shoulder, elbow, wrist and knee joints in that order, and superscripts l and r represent left and right body side, respectively. We define $\mathbf{v_{sn}^r} = J_s^r - J_n$, $\mathbf{v_{es}^r} = J_e^r - J_s^r$ and $\mathbf{v_{we}^r} = J_w^r - J_e^r$ as the collar-bone, upper-arm and the lower-arm, respectively (See Fig. 4). Now, $\mathbf{n_s^r} = \mathbf{v_{sn}^r} \times \mathbf{v_{es}^r}$ is the normal to the plane defined by the collar-bone and the upper-arm. For the elbow joint to be legal, $\mathbf{v_{we}^r}$ must have a positive component in the direction of $\mathbf{n_s^r}$, i.e. $\mathbf{n_s^r} \cdot \mathbf{v_{we}^r}$ must be positive. We do not incur any penalty when the joint angle is legal and define $E_e^r = \min(\mathbf{n_s^r} \cdot \mathbf{v_{we}^r}, 0)$ as a measure of implausibility. Note that this case is opposite for the right knee and left elbow joints (as shown by the right hand rule) and requires E_k^r and E_e^l to be positive for the illegal case. We exponentiate E to strongly penalize large deviations beyond legality. Our angle loss can now be defined as:

$$\mathcal{L}_a = -E_e^r e^{-E_e^r} + E_e^l e^{E_e^l} + E_k^r e^{E_k^r} - E_k^l e^{-E_k^l} \quad (3)$$

All the terms of the loss are functions of bone vectors which are defined in terms of the output poses. The loss is, therefore, differentiable and the reader is referred to supplementary material for more details.

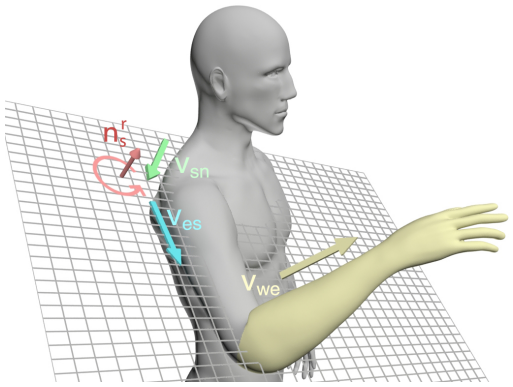


Figure 4. Illustration of Illegal Angle loss: For the elbow joint angle to be legal, the lower-arm must project a positive component along $\mathbf{n_s^r}$ (normal to collarbone-upperarm plane) , i.e. $\mathbf{n_s^r} \cdot \mathbf{v_{we}} \geq 0$. Note that we only need 2D annotated data to train our model using this formulation.

Symmetry Loss ($\mathcal{L}_s$): We further extend our semi-supervised approach by introducing a bone length symmetry term $\mathcal{L}_s$. The loss is simple yet heavily constrains the joint depths, especially when depth is ambiguous due to occlusions. $\mathcal{L}_s$ is defined as the difference in lengths of left/right bone pairs. Let $\mathcal{B}$ be the set of all the bones on right half of the body except torso and head bones. Also, let BL_b represent the bone-length of bone b. We define L_s as

$$\mathcal{L}_s = \sum_{b \in \mathcal{B}} ||BL_b - BL_{C(b)}||_2 \quad (4)$$

where $C(.)$ indicates the corresponding left side bone. For example, for the case when b is the right upper arm $C(b)$ would be the left upper arm.

3.2. TimePoseNet

The poses predicted by any model which works independently on each image are prone to noise and also not temporally coherent. After experimenting with several architectures for a temporal refiner like RNNs, LSTMs and Recurrent Encoder-Decoders, we settled on using a fully-connected network because of its higher effectiveness and simplicity. Let $\tilde{P}_i$ and $\hat{P}_i$ denote the i^{th} ConvNet and Temporal outputs, respectively. We train a fully-connected network TimePoseNet to refine a *coarse* input pose estimate $\tilde{P}_i$. The network takes as input previous N 3D poses from the memory, along with $\tilde{P}_i$. We flatten the input to $\mathbb{R}^{(N+1)\times K\times 3}$ and map it to an output in $\mathbb{R}^{K\times 3}$. The network consists of one hidden layer with 4096 neurons and is trained using the standard L_2 loss. The loss can be written as

$$\mathcal{L}_t(\tilde{P}_{i-N}, \tilde{P}_{i-N+1}, \ldots, \tilde{P}_i) = ||\hat{P}_i - P_i^{gt}||_2 \quad (5)$$

Method	Direction	Discuss	Eat	Greet	Phone	Pose	Purchase	Sit
Chen [21]	89.9	97.6	90.0	107.9	107.3	93.6	136.1	133.1
Tome [19]	65.0	73.5	76.8	86.4	86.3	68.9	74.8	110.2
Moreno [11]	69.5	80.2	78.2	87.0	100.8	76.0	69.7	104.2
Zhou [13]	68.7	74.8	67.8	76.4	76.3	84.0	70.2	88.0
Jahangiri [36]	74.4	66.7	67.9	75.2	77.3	70.6	64.5	95.6
Lin [12]	58.0	68.2	63.2	65.8	75.3	61.2	65.7	98.6
Mehta [5]	57.5	68.6	59.6	67.3	78.1	56.9	69.1	98.0
Pavlakos [20]	58.6	64.6	63.7	62.4	66.9	57.7	62.5	76.8
Zhou [14]	54.8	60.7	58.2	71.4	62.0	53.8	55.6	75.2
Sun [10]	52.8	54.8	54.2	54.3	61.8	53.1	53.6	71.7
Ours(UnitPoseNet)	46.9	53.8	47.0	52.8	56.9	45.2	48.2	68.0
Ours(TimePoseNet)	**44.8**	**50.4**	**44.7**	**49.0**	**52.9**	**43.5**	**45.5**	**63.1**
Method	SitDown	Smoke	Photo	Wait	Walk	WalkDog	WalkPair	Avg
Chen [21]	240.1	106.7	139.2	106.2	87.0	114.1	90.5	114.2
Tome [19]	173.9	85.0	110.7	85.8	71.4	86.3	73.1	88.4
Moreno [11]	113.9	89.7	102.7	98.7	79.2	82.4	77.2	87.3
Zhou [13]	113.8	78.0	78.4	89.1	62.6	75.1	73.6	79.9
Jahangiri [36]	127.3	79.6	79.1	73.4	67.4	71.8	72.8	77.6
Lin [12]	127.7	70.4	93.0	68.2	50.6	72.9	57.7	73.1
Mehta [5]	117.5	69.5	82.4	68.0	55.3	76.5	61.4	72.9
Pavlakos [20]	103.5	65.7	70.7	61.6	56.4	69.0	59.5	66.9
Zhou [14]	111.6	64.1	65.5	66.0	51.4	63.2	55.3	64.9
Sun [10]	**86.7**	61.5	67.2	53.4	47.1	61.6	53.4	59.1
Ours(UnitPoseNet)	94.0	55.7	63.6	51.6	40.3	55.4	44.3	55.5
Ours(TimePoseNet)	87.3	**51.7**	**61.4**	**48.5**	**37.6**	**52.2**	**41.9**	**52.1**

Table 1. Comparative evaluation of our model on Human 3.6 following Protocol 1. The evaluations were performed on subjects 9 and 11 using ground truth bounding box crops and the models were trained only on Human3.6 and MPII 2D pose datasets.

The model learns correlations between the pose predictions of the previous N frames and the current frame (see Fig. 7), and outputs the best possible candidate solution for the current frame. It also implicitly learns to extract best estimates of bone lengths in its output pose (Table. 4). It is worth noting that since we only use the past context and a simple architecture, the latency introduced by our TimePoseNet model is negligible.

3.3. Training and Implementation details

UnitPoseNet We describe our training strategy for training UnitPoseNet on Human3.6M and a similar strategy is used for the MPI-INF-3DHP dataset. We use the publicly available code in [14] based on Torch7 [37] and train the model using Stochastic Gradient Descent optimization. At every epoch, samples from both MPII 2D and Human3.6M are consumed in equal proportion. We train the model in four stages: In the *first* stage, the stacked hourglass module is trained using 2D data for 60 epochs as in [33]. In the *second* stage, the depth regressor module is trained with only 3D data with $\lambda_d = 0.1$, $\lambda_a = 0$ and $\lambda_s = 0$ for another 20 epochs. The last two stages sequentially add the structure-aware loss terms to the model for 30 epochs each. We introduce the Illegal Angle loss $\mathcal{L}_a$ with $\lambda_a = 0.005$ in the *third* stage and Symmetry loss $\mathcal{L}_s$ with $\lambda_s = 0.005$ in the *fourth* stage, for finetuning. Learning rate was set to $2.5e-4$ for Stages 1-3, and $2.5e-5$ for Stage 4.

TimePoseNet was trained in a single stage using Adam optimizer for 30 epochs using coarse predictions generated by fully-trained UnitPoseNet. A sequence of 20 frames were randomly chosen from the entire dataset, and the network was trained to predict the last pose. In our experiments, we found that a context of $N = 19$ past frames yields the best improvement on MPJPE (Fig. 7) and we use that in all our experiments.

Both the models were trained on one NVIDIA 1080 Ti GPU. It took approximately two days to train UnitPoseNet and one hour to train TimePoseNet. UnitPoseNet runs at an average testing time of $20ms$ per image while TimePoseNet adds negligible delay (<1ms).

4. Experiments

4.1. Experimental Setup

In this section, we quantitatively evaluate our results on Human3.6M and MPI-INF-3DHP datasets. We also provide

Method	Stand/ Walk	Exercise	Sit on Chair	Crouch/ Reach	On the Floor	Sports	Misc.	Avg. PCK	Avg. AUC	MPJPE (mm)
Mehta [5]	86.6	75.3	**74.8**	73.7	52.2	82.1	77.5	75.7	39.3	117.6
Mehta [17]	87.7	**77.4**	74.7	72.9	51.3	**83.3**	80.1	76.6	**40.4**	124.7
Ours* (MPI-INF-3DHP)	84.2	70.4	71.9	74.5	57.2	73.9	69.7	72.3	34.8	116.31
Baseline	81.2	67.9	73.0	74.6	53.6	74.0	64.7	70.3	34.1	121.3
+ UnitPoseNet	88.1	72.2	75.7	**79.0**	**60.13**	78.5	70.7	75.0	37.3	108.5
+ TimePoseNet	87.2	73.1	75.1	77.8	55.3	80.2	72.9	75.2	36.2	106.3
+ Skeleton Fitting	**89.1**	75.1	73.6	77.9	49.2	79.3	**80.8**	**76.7**	39.1	**103.8**

Table 2. Comparison of our method on MPI-INF-3DHP dataset. The table shows activity-wise PCK results along with the average AUC and MPJPE. Higher PCK and AUC are desired while a lower MPJPE is desirable. The Baseline and the subsequent models were trained using MPI-INF-3DHP and Human3.6M datasets. Note that unlike [5, 17], the MPI-INF-3DHP training dataset was not augmented. Ours* represents the full model including the skeleton fitting step trained on only MPI-INF-3DHP dataset.

qualitative evaluations on MPII 2D human pose dataset.

Human3.6M: It consists of 11 subjects performing a range of actions in an indoor room setting. The ground truth annotations are captured using a marker-based MoCap system. Following the literature, we follow two standard protocols to train and evaluate our results.

In *Protocol 1*, Subjects S1, S5, S6, S7 and S8 are used for training while Subjects S9 and S11 are used for testing. All the videos are downsampled from *50 fps* to *10 fps*. The evaluation metric used is Mean Per Joint Position Error (MPJPE), which is the average Euclidean distance in *mm* between the predicted and ground truth joint positions after aligning the positions of root joints. This protocol is followed in [10, 14, 20, 12, 13, 11].

In *Protocol 2*, Subjects S1, S5, S6, S7, S8 and S9 are used for training and the testing is performed on subject S11. The error metric used is Procrustes Aligned MPJPE (PA MPJPE) which is the MPJPE calculated after rigidly aligning the predicted pose with the ground truth. The evaluations are performed on every 64^{th} frame of the videos. This protocol is followed in [21, 22, 11, 38]. In both the protocols, all the cameras and activities are considered for evaluation.

MPI-INF-3DHP (test) dataset: It is a recently released dataset consisting of 6 test subjects including 2 subjects performing in-the-wild. Of the 4 indoor sequences, 2 have green screen (GS) background and 2 have normal backgrounds. MPI-INF-3DHP provides a more challenging testing ground than Human3.6M, which has just one type of indoor setting. The annotations are captured by a markerless MoCap system. The evaluation metric proposed in [5] is Percentage of Correct Keypoints (PCK) within *150mm* range and Area Under Curve (AUC). Like [14], we assume that the global scale is known. We do skeleton retargeting while training to account for the difference of joint definitions between Human3.6M and MPI-INF-3DHP datasets.

MPII 2D dataset: It is a 2D pose dataset for which 3D ground truth annotations are not available. We show qualitative results in Fig. 6.

4.2. Quantitative Evaluations

We evaluate the outputs of the three stages of our pipeline and show improvements at each stage.

Baseline: We train the same network architecture as UnitPoseNet with only the supervised loss $\mathcal{L}_z^{2D}$ with both 2D and 3D data for our baseline.

UnitPoseNet: This model is trained on the joint losses detailed in Section 3.1.1. This model is trained as per the training strategy mentioned in Section 3.3.

TimePoseNet: TimePoseNet is applied to the outputs of UnitPoseNet in this step.

Skeleton Fitting: In this step, we fit a skeleton based on the subject's bone lengths while preserving the joint angles.

Evaluations on Human3.6M: Our quantitative evaluations on Human 3.6M dataset show significant improvement over the state-of-the-art. We achieve an MPJPE of $55.5mm$ with our single image model UnitPoseNet. This is further improved by the temporal model TimePoseNet that reduces the MPJPE down to $52.1mm$. Table 1 presents a comparative analysis of our results following *Protocol 1*. We also achieve state-of-the-art results following *Protocol 2* evaluations as is shown in Table 3.

Evaluations on MPI-INF-3DHP: MPI-INF-3DHP is a more challenging dataset as it contains in-the-wild scenarios with more complex actions. Our overall results are competitive with the current state-of-the-art [17] and better in outdoor settings. Table 2 tabulates the activity-wise results on MPI-INF-3DHP. We achieve a PCK of 75.0% with 37.3% AUC using the angle and symmetry losses. Our complete method achieves a PCK of 76.73% with 39.1% AUC.

4.2.1 Plausibility Analysis

Structural plausibility: Length of each bone is fixed, and joints are constrained to allow movement along certain degrees of freedom with limits on how far they can bend. We evaluate plausibility of the predicted poses on two metrics - symmetry and percentage of illegal angles. Results are shown in Table 4. We observe only 0.8% illegal non-

Method	Direct.	Discuss	Eat	Greet	Phone	Pose	Purch.	Sit	Sit Down	Smoke	Photo	Wait	Walk	Walk Dog	Walk Pair	Avg
Yasin [22]	88.4	72.5	108.5	110.2	97.1	91.6	107.2	119.0	170.8	108.2	142.5	86.9	92.1	165.7	102.0	108.3
Rogez [38]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	88.1
Chen [21]	71.6	66.6	74.7	79.1	70.1	67.6	89.3	90.7	195.6	83.5	93.3	71.2	55.7	85.9	62.5	82.7
Nie [39]	62.8	69.2	79.6	78.8	80.8	72.5	73.9	96.1	106.9	88.0	86.9	70.7	71.9	76.5	73.2	79.5
Moreno [11]	67.4	63.8	87.2	73.9	71.5	69.9	65.1	71.7	98.6	81.3	93.3	74.6	76.5	77.7	74.6	76.5
Zhou [13]	47.9	48.8	52.7	55.0	56.8	49.0	45.5	60.8	81.1	53.7	65.5	51.6	50.4	54.8	55.9	55.3
Sun [10]	42.1	44.3	45.0	45.4	51.5	43.2	41.3	59.3	73.3	51.0	53.0	44.0	38.3	48.0	44.8	48.3
Ours(UnitPoseNet)	32.8	36.8	42.5	38.5	42.4	35.4	34.3	53.6	66.2	46.5	49.0	34.1	30.0	42.3	39.7	42.2
Ours (TimePoseNet)	**28.0**	**30.7**	**39.1**	**34.4**	**37.1**	**28.9**	**31.2**	**39.3**	**60.6**	**39.3**	**44.8**	**31.1**	**25.3**	**37.8**	**28.4**	**36.3**

Table 3. Comparative evaluation of our model on Human 3.6M using Protocol 2. The evaluations were performed on subject 11 using the ground truth bounding box crops and trained on subjects 1,5,6,7 and 8. The models were trained only on Human3.6M and MPII 2D.

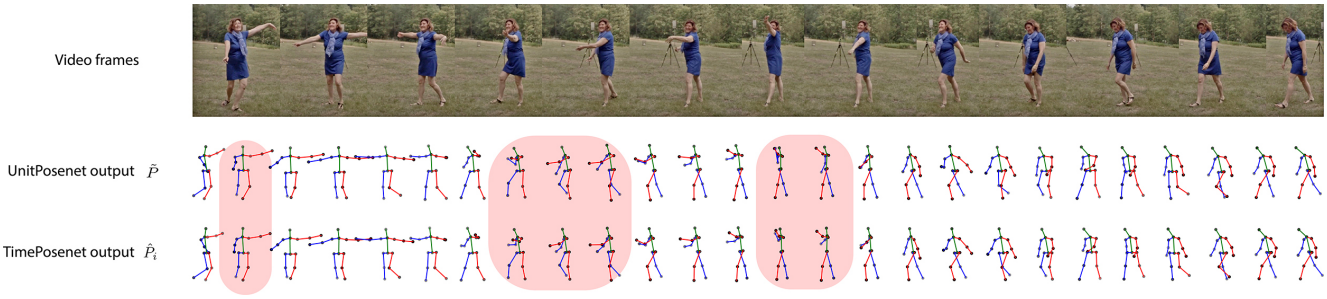


Figure 5. Comparison of our temporal model TimePoseNet with UnitPoseNet on a video. The highlighted poses demonstrate the ability of TimePoseNet to learn temporal correlations, and smoothen and refine pose estimates from UnitPoseNet.

torso joint angles by following the evaluation in [26] with our final model.

Temporal Plausibility: We perform both qualitative and quantitative evaluations for determining the effectiveness of TimePoseNet as a temporal model. The qualitative evaluations in Fig. 5 and the supplementary material clearly indicate that TimePoseNet effectively corrects jerks in the poses predicted by UnitPoseNet. We compare the standard deviation of bone lengths across frames with and without the temporal model in Table 4. Introducing the temporal model reduces the average standard deviation of bone lengths by 28.7%, indicating that our outputs are more stable and consistent. It is also worth noting that we do not use any additional filter (moving average, 1 Euro, etc.) which introduces lag and makes the motion look uncanny. TimePoseNet learns a smart filter implicitly and produces smooth output.

5. Discussion

5.1. Obtaining supervision from structural constraints

Since the task of 3D pose estimation from monocular images is underconstrained and highly non-linear, it is very easy for a model to get trapped in suboptimal local minima for natural images. Some of these minima can be rejected using structural priors of the human body. Sminchisescu et al. [40] show that this may be achieved by forcing the optimization state to climb uphill towards the transition state (saddle point) at the *peak* of the energy surface. Though we do not use an optimization approach, our structure-aware loss transforms the loss function such that the saddle points are mitigated, and gradient descent can work effectively. Fig. 3 shows how the local loss function changes in a semi-supervised setting as we exploit the joint angle limit and bone length constraint.

Bone	Baseline	Ours (UnitPoseNet)	Ours (TimePoseNet)	Percentage Improvement
Upper arm	43.9	37.6	**30.3**	31.0%
Lower arm	49.7	44.6	**39.7**	20.1%
Upper leg	35.1	33.9	**25.9**	26.2%
Lower leg	49.2	47.6	**33.2**	32.5%
Upper arm	55.0	49.6	**39.8**	27.6%
Lower arm	69.1	66.0	**48.3**	30.1%
Upper leg	65.6	61.3	**48.8**	25.6%
Lower leg	70.5	68.8	**48.3**	31.5%

Table 4. Evaluating our models on (i) symmetry - mean L1 distance in mm between left/right bone pairs (upper half), and (ii) the standard deviation of bone lengths across all video frames (lower half) on MPI-INF-3DHP dataset. The temporal model substantially improves symmetry and stabilizes variations in bone lengths.

5.2. Learning structural and temporal correlations

In this section, we investigate how effectively the fully connected network has learned *structural* and *temporal* correlations from the additional information from previous

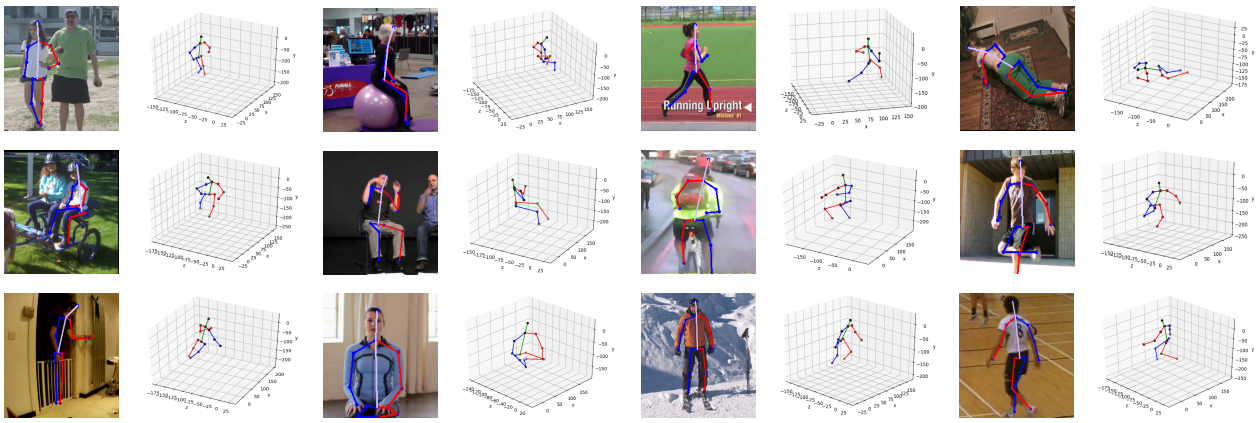
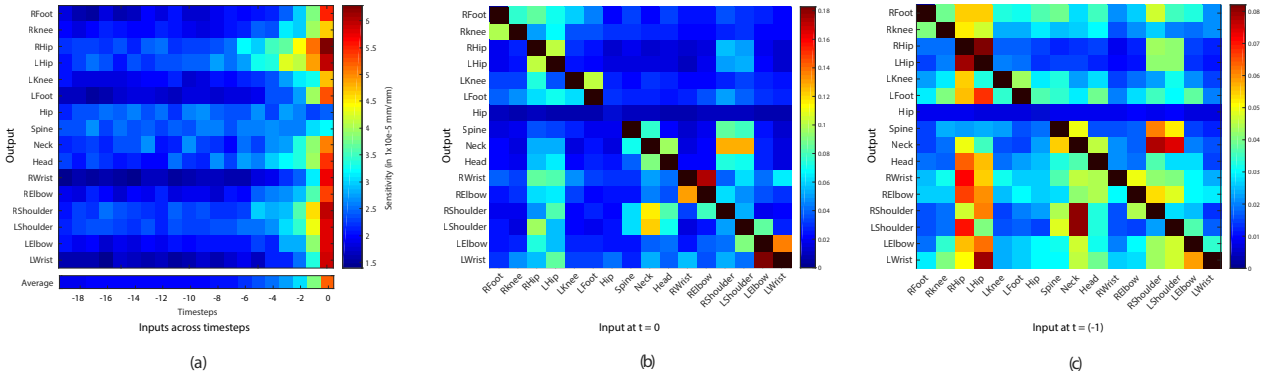


Figure 6. Qualitative results of UnitPosenet on the MPII 2D dataset. Our model predicts structurally plausible poses.



(a) (b) (c)

Figure 7. (a) The sensitivity of the output pose to perturbations in the input pose of the temporal model decreases from t=0 to t=-19. (b) Strong structural correlations are learned from the pose input at t=0 frame. (c) Past frames show smaller but more complex structural correlations. For example, the output *wrist* position is more strongly correlated to the *shoulder* position from t=-1 than from t=0. The self correlations of the joints (diagonal) are an order of magnitude larger and the colormap range has been capped to better display other correlations.

frames. We measure the sensitivity S of each joint of the output pose to random perturbations in the positions of each joint of the input $N + 1$ poses. Each input to the temporal model is separately perturbed with unit Gaussian noise, and the Euclidean distance between the original output and the perturbed output is averaged over the inputs corresponding to each frame. The analysis is shown in Fig. 7. We also verify that the norms of the corresponding weights of the fully-connected layer's input vary similarly.

The temporal model learns motion from the past frames as shown in Fig. 7 (a) with decreasing strength with time. The remaining frames contribute to the model inferring better bone lengths estimates and structure. Fig. 7 (b) shows the structural correlations the model has learned just within the current frame. The model learns to rely on the positions of hips and shoulders to refine almost all the other joints. We can also observe that child joints are correlated with parent joints, for eg. the wrists are correlated strongly with elbows, and the shoulders are strongly correlated with the neck. Fig. 7 (c) shows the sensitivity to the input pose at t = -1. Correlations learned from the past are weaker, but show a richer pattern. The sensitivity of the child joints extends further upwards into the kinematic chain, eg. the wrist shows higher correlations with elbow, shoulder and neck, for the t = -1 frame.

6. Conclusion

In this paper, we proposed a method that advances the state-of-the-art on 3D pose estimation on images as well as videos. We showed that anthropometric constraints can be leveraged for semi-supervised learning of 3D pose from 2D labeled images. We also demonstrated that a simple, fully-connected network is highly effective in modeling pose sequences and learns short-term temporal and structural correlations. In the future, we will work on further improving in-the-wild performance without additional data.

References

[1] C. Sminchisescu and B. Triggs, "Estimating articulated human motion with covariance scaled sampling," *The International Journal of Robotics Research*, 2003. 1

[2] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris, "3d human pose estimation: A review of the literature and analysis of covariates," *Computer Vision and Image Understanding*, 2016. 1, 2

[3] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE TPAMI*, 2014. 1, 2

[4] C. S. Catalin Ionescu, Fuxin Li, "Latent structured models for human pose estimation," in *ICCV*, 2011. 1

[5] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3d human pose estimation in the wild using improved cnn supervision," in *3DV*, 2017. 1, 2, 5, 6

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

[7] 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

[8] S. Li and A. B. Chan, "3d human pose estimation from monocular images with deep convolutional neural network," in *ACCV*, 2014. 1, 2

[9] S. Li, W. Zhang, and A. B. Chan, "Maximum-margin structured learning with deep networks for 3d human pose estimation," in *ICCV*, 2015. 1

[10] X. Sun, J. Shang, S. Liang, and Y. Wei, "Compositional human pose regression," in *ICCV*, 2017. 1, 2, 5, 6, 7

[11] F. Moreno-Noguer, "3d human pose estimation from a single image via distance matrix regression," in *CVPR*, 2017. 1, 2, 5, 6, 7

[12] M. Lin, L. Lin, X. Liang, K. Wang, and H. Cheng, "Recurrent 3d pose sequence machines," in *CVPR*, 2017. 1, 2, 3, 5, 6

[13] X. Zhou, M. Zhu, K. Derpanis, and K. Daniilidis, "Sparseness meets deepness: 3d human pose estimation from monocular video," in *CVPR*, 2016. 1, 2, 5, 6, 7

[14] 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. 1, 2, 3, 4, 5, 6

[15] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *CVPR*, 2014. 1

[16] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *arXiv preprint arXiv:1405.0312*, 2014. 1

[17] 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," in *ACM ToG*, 2017. 1, 2, 6

[18] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei, "Deep kinematic pose regression," *arXiv preprint arXiv:1609.05317*, 2016. 2

[19] D. Tome, C. Russell, and L. Agapito, "Lifting from the deep: Convolutional 3d pose estimation from a single image," in *CVPR*, 2017. 2, 5

[20] 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, 5, 6

[21] C.-H. Chen and D. Ramanan, "3d human pose estimation = 2d pose estimation + matching," in *CVPR*, 2017. 2, 5, 6, 7

[22] H. Yasin, U. Iqbal, B. Kruger, A. Weber, and J. Gall, "A dual-source approach for 3d pose estimation from a single image," in *CVPR*, 2016. 2, 6, 7

[23] G. Rogez, P. Weinzaepfel, and C. Schmid, "Lcr-net: Localization-classification-regression for human pose," in *CVPR*, 2017. 2

[24] 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. 2

[25] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: a skinned multi-person linear model," *ACM Trans. Graph.*, 2015. 2

[26] I. Akhter and M. J. Black, "Pose-conditioned joint angle limits for 3d human pose reconstruction," in *CVPR*, 2015. 2, 7

[27] G. Casiez, N. Roussel, and D. Vogel, "1 filter: a simple speed-based low-pass filter for noisy input in interactive systems," in *SIGCHI*, ACM, 2012. 2

[28] T. Alldieck, M. Kassubeck, B. Wandt, B. Rosenhahn, and M. Magnor, "Optical flow-based 3d human motion estimation from monocular video," in *GCPR*, 2017. 3

[29] X. Wei and J. Chai, "Videomocap: Modeling physically realistic human motion from monocular video sequences," *ACM ToG*, 2010. 3

[30] R. Urtasun, D. J. Fleet, and P. Fua, "Temporal motion models for monocular and multiview 3d human body tracking," *Computer vision and image understanding*, 2006. 3

[31] M. J. Park, M. G. Choi, Y. Shinagawa, and S. Y. Shin, "Video-guided motion synthesis using example motions," *ACM ToG*, 2006. 3

[32] H. Coskun, F. Achilles, R. DiPietro, N. Navab, and F. Tombari, "Long short-term memory kalman filters: Recurrent neural estimators for pose regularization," in *ICCV*, 2017. 3

[33] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *ECCV*, 2016. 3, 5

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016. 3

[35] A. Jain, J. Tompson, M. Andriluka, G. W. Taylor, and C. Bregler, "Learning human pose estimation features with convolutional networks," in *ICLR*, 2014. 3

[36] E. Jahangiri and A. L. Yuille, "Generating multiple diverse hypotheses for human 3d pose consistent with 2d joint detections," in *ICCV*, 2017. 5

[37] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A matlab-like environment for machine learning," in *BigLearn, NIPS Workshop*, 2011. 5

[38] G. Rogez and C. Schmid, "Mocap-guided data augmentation for 3d pose estimation in the wild," in *NIPS*, 2016. 6, 7

[39] B. Xiaohan Nie, P. Wei, and S.-C. Zhu, "Monocular 3d human pose estimation by predicting depth on joints," in *ICCV*, Oct 2017. 7

[40] C. Sminchisescu and B. Triggs, "Building roadmaps of local minima of visual models," *ECCV*, 2002. 7