Learning 3D Human Pose from Structure and Motion

Rishabh Dabral[1(✉)], Anurag Mundhada[1], Uday Kusupati[1], Safeer Afaque[1], Abhishek Sharma[2], and Arjun Jain[1]

[1] Indian Institute of Technology Bombay, Mumbai, India
{rdabral,ajain}@cse.iitb.ac.in, {anuragmundhada,udaykusupati}@iitb.ac.in
[2] Gobasco AI Labs, Lucknow, India
abhsharayiya@gmail.com

Abstract. 3D human pose estimation from a single image is a challenging problem, especially for in-the-wild settings due to the lack of 3D annotated data. We propose two anatomically inspired loss functions and use them with a weakly-supervised learning framework to jointly learn from large-scale in-the-wild 2D and indoor/synthetic 3D data. We also present a simple temporal network that exploits temporal and structural cues present in predicted pose sequences to temporally harmonize the pose estimations. We carefully analyze the proposed contributions through loss surface visualizations and sensitivity analysis to facilitate deeper understanding of their working mechanism. Jointly, the two networks capture the anatomical constraints in static and kinetic states of the human body. Our complete pipeline improves the state-of-the-art by 11.8% and 12% on Human3.6M and MPI-INF-3DHP, respectively, and runs at 30 FPS on a commodity graphics card.

1 Introduction

Accurate 3D human pose estimation from monocular images and videos is the key to unlock several applications in robotics, human computer interaction, surveillance, animation and virtual reality. These applications require *accurate* and *real-time* 3D pose estimation from monocular image or video under challenging variations of clothing, lighting, view-point, self-occlusions, activities, background clutter etc. [32,33]. With the advent of recent advances in deep learning, compute hardwares and, most importantly, large-scale *real-world* datasets (ImageNet [31], MS COCO [20], CityScapes [10] etc.), computer vision systems have witnessed dramatic improvements in performance. Human-pose estimation has also benefited from synthetic and real-world datasets such as MS COCO [20], MPII Pose [3], Human3.6M [6,14], MPI-INF-3DHP [22], and SURREAL [37]. Especially, 2D pose prediction has witnessed tremendous improvement due to

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-01240-3_41) contains supplementary material, which is available to authorized users.

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11213, pp. 679–696, 2018.
https://doi.org/10.1007/978-3-030-01240-3_41

large-scale in-the-wild datasets [3,20]. However, 3D pose estimation still remains challenging due to severely under-constrained nature of the problem and absence of any real-world 3D annotated dataset.

A large body of prior art either directly regresses for 3D joint coordinates [17, 18,34] or infers 3D from 2D joint-locations in a two-stage approach [19,22,24,41, 43]. These approaches perform well on synthetic 3D benchmark datasets, but lack generalization to the real-world setting due to the lack of 3D annotated in-the-wild datasets. To mitigate this issue, some approaches use synthetic datasets [9, 37], green-screen composition [22,23], domain adaptation [9], transfer learning from intermediate 2D pose estimation tasks [17,22], and joint learning from 2D and 3D data [34,41]. Notably, joint learning with 2D and 3D data has shown promising performance in-the-wild owing to large-scale real-world 2D datasets. We seek motivation from the recently published joint learning framework of Zhou et al. [41] and present a novel structure-aware loss function to facilitate training of Deep ConvNet architectures using both 2D and 3D data to accurately predict the 3D pose from a single RGB image. The proposed loss function is applicable to 2D images during training and ensures that the predicted 3D pose does not violate anatomical constraints, namely joint-angle limits and left-right symmetry of the human body. We also present a simple learnable temporal pose model for pose-estimation from videos. The resulting system is capable of jointly exploiting the structural cues evident in the static and kinetic states of human body.

Our proposed structure-aware loss is inspired by anatomical constraints that govern the human body structure and motion. We exploit the fact that certain body-joints cannot bend beyond an angular range; e.g. the knee (elbow) joints cannot bend forward (backward). We also make use of left-right symmetry of human body and penalize unequal corresponding pairs of left-right bone lengths. Lastly, we also use the bone-length ratio priors from [41] that enforce certain pairs of bone-lengths to be constant. It is important to note that the illegal-angle and left-right symmetry constraints are complementary to the bone-length ratio prior, and we show that they perform better too. We present the visualization of the loss surfaces of the proposed losses to facilitate a deeper understanding of their workings. The three aforementioned structure losses are used to train our *Structure-Aware PoseNet*. Joint-angle limits and left-right symmetry have been used previously in the form of optimization functions [1,4,13]. To the best of our knowledge we are the first ones to exploit these two constraints, in the form of differentiable and tractable loss functions, to train ConvNets directly. Our structure-aware loss function outperforms the published state-of-the-art in terms of Mean-Per-Joint-Position-Error (MPJPE) by 7% and 2% on Human3.6M and MPI-INF-3DHP, respectively.

We further propose to learn a temporal motion model to exploit cues from sequential frames of a video to obtain anatomically coherent and smoothly varying poses, while preserving the realism across different activities. We show that a moving-window fully-connected network that takes previous N poses performs extremely well at capturing temporal as well as anatomical cues from pose sequences. With the help of carefully designed controlled experiments we

show the temporal and anatomical cues learned by the model to facilitate better understanding. We report an additional 7% improvement on Human3.6M with the use of our temporal model and demonstrate real-time performance of the full pipeline at 30 fps. Our final model improves the published state-of-the-art on Human3.6M [14] and MPI-INF-3DHP [22] by 11.8% and 12%, respectively.

2 Related Work

This section presents a brief summary of the 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 [32] for a detailed review of the literature.

ConvNet Architectures: Most existing ConvNet based approaches either directly regress 3D poses from the input image [17,34,42,43] or infer 3D from 2D pose in a two-stage approach [19,23,24,35,41]. Some approaches make use of volumetric-heatmaps [27], some define a pose using bones instead of joints [34], while the approach in [23] directly regresses for 3D location maps. The use of 2D-to-3D pipeline enables training with large-scale in-the-wild 2D pose datasets [3,20]. A few approaches use statistical priors [1,43] to lift 2D poses to 3D. Chen et al. [7] and Yasin et al. [40] use a pose library to retrieve the nearest 3D pose given the corresponding 2D pose prediction. Recent ConvNet based approaches [23,27,30,34,41,43] have reported substantial improvements in real-world setting by pre-training or joint training of their 2D prediction modules, but it still remains an open problem.

Utilizing Structural Information: The structure of the human skeleton is constrained by fixed bone lengths, joint angle limits, and limb inter-penetration constraints. Some approaches use these constraints to infer 3D from 2D joint locations. Akhter and Black [1] learn pose-dependent joint angle limits for lifting 2D poses to 3D via an optimization problem. Ramakrishna et al. [28] solve for anthropometric constraints in an activity-dependent manner. Recently, Moreno [24] proposed to estimate the 3D inter-joint distance matrix from 2D inter-joint distance matrix using a simple neural network architecture. These approaches do not make use of rich visual cues present in images and rely on the predicted 2D pose that leads to sub-optimal results. Sun et al. [34] re-parameterize the pose presentation to use bones instead of joints and propose a structure-aware loss. But, they do not explicitly seek to penalize the feasibility of inferred 3D pose in the absence of 3D ground-truth data. Zhou et al. [41] introduce a weakly-supervised framework for joint training with 2D and 3D data with the help of a geometric loss function to exploit the consistency of bone-length ratios in human body. We further strengthen this weakly-supervised setup with the help of joint-angle limits and left-right symmetry based loss functions for better training. Lastly, there are methods that recover both shape and pose from a 2D image via a mesh-fitting strategy. Bogo et al. [4] penalize body-part

inter-penetration and illegal joint angles in their objective function for finding SMPL [21] based shape and pose parameters. These approaches are mostly offline in nature due to their computational requirements, while our approach runs at 30 fps.

Utilizing Temporal Information: Direct estimation of 3D pose from disjointed images leads to temporally incoherent output with visible jitters and varying bone lengths. 3D pose estimates from a video can be improved by using simple filters or temporal priors. Mehta et al. [23] propose a real-time approach which penalizes acceleration and depth velocity in an optimization step after generating 3D pose proposals using a ConvNet. They also smooth the output poses with the use of a tunable low-pass filter [5] optimized for interactive systems. Zhou et al. [43] introduce a first order smoothing prior in their temporal optimization step. Alldieck et al. [2] exploit 2D optical flow features to predict 3D poses from videos. Wei et al. [38] exploit physics-based constraints to realistically interpolate 3D motion between video keyframes. There have also been attempts to learn motion models. Urtasun et al. [36] learn activity specific motion priors using linear models while Park et al. [26] use a motion library to find the nearest motion given a set of 2D pose predictions followed by iterative fine-tuning. The motion models are activity-specific whereas our approach is generic. Recently, Lin et al. [19] used recurrent neural networks to learn temporal dependencies from the intermediate features of their ConvNet based architecture. In a similar attempt, Coskun et al. [11] use LSTMs to design a Kalman filter that learns human motion model. In contrast with the aforementioned approaches, our temporal model is simple yet effectively captures short-term interplay of past poses and predicts the pose of the current frame in a temporally and anatomically consistent manner. It is generic and does not need to be trained for activity-specific settings. We show that it learns complex, non-linear inter-joint dependencies over time; e.g. it learns to refine wrist position, for which the tracking is least accurate, based on the past motion of elbow and shoulder joints.

3 Background and Notations

This section introduces the notations used in this article and also provides the required details about the weakly-supervised framework of Zhou et al. [41] for joint learning from 2D and 3D data.

A 3D human pose $P = \{p_1, p_2, \ldots, p_k\}$ is defined by the positions of $k = 16$ body joints in Euclidean space. These joint positions are defined relative to a root joint, which is fixed as the pelvis. The input to the pose estimation system could be a single RGB image or a continuous stream of RGB images $I = \{\ldots, I_{i-1}, I_i\}$. The i^{th} joint p_i is the coordinate of the joint in a 3D Euclidean space i.e. $p_i = (p_i^x, p_i^y, p_i^z)$. Throughout this article inferred variables are denoted with a $\tilde{*}$ and ground-truth is denoted with a $\hat{*}$, therefore, an inferred joint will be denoted as $\tilde{p}$ and ground-truth as $\hat{p}$. The 2D pose can be expressed with only the x, y-coordinates and denoted as $p^{xy} = (p^x, p^y)$; the depth-only joint location is denoted as $p^z = (p^z)$. The i^{th} training data from a 3D annotated dataset consists

of an image I_i and corresponding joint locations in 3D, $\hat{P}_i$. On the other hand, the 2D data has only the 2D joint locations, $\hat{P}_i^{xy}$. Armed with these notations, below we describe the weakly-supervised framework for joint learning from [41].

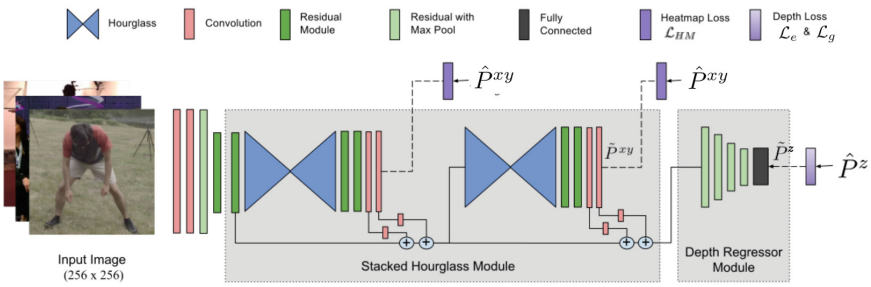


Fig. 1. A schematic of the network architecture. The stacked hourglass module is trained using the standard Euclidean loss $\mathcal{L}_{HM}$ against ground truth heatmaps. Whereas, the depth regressor module is trained on either $\mathcal{L}_{3D}^z$ or $\mathcal{L}_{2D}^z$ depending on whether the ground truth depth $\hat{P}^z$ is available or not.

Due to the absence of in-the-wild 3D data, the pose estimation systems learned using the controlled or synthetic 3D data fail to generalize well to in-the-wild settings. Therefore, Zhou et al. [41] proposed a weakly-supervised framework for joint learning from both 2D and 3D annotated data. Joint learning exploits the 3D data for depth prediction and the in-the-wild 2D data for better generalization to real-world scenario. The overall schematic of this framework is shown in Fig. 1. It builds upon the stacked hourglass architecture [25] for 2D pose estimation and adds a depth-regression sub-network on top of it. The stacked hourglass is trained to output the 2D joint locations, $\tilde{P}^{xy}$ in the image coordinate with the use of standard Euclidean loss between the predicted and the ground-truth joint-location heatmaps, please refer to [25] for more details. The depth-regression sub-network, a series of four residual modules [12] followed by a fully connected layer, takes a combination of different feature maps from stacked hourglass and outputs the depth of each joint i.e. $\tilde{P}^z$. Standard Euclidean loss $\mathcal{L}_e(\tilde{P}^z, \hat{P}^z)$ is used for the 3D annotated data-sample. On the other hand, a weak-supervision in the form of a geometric loss function, $\mathcal{L}_g(\tilde{P}^z, \hat{P}^{xy})$, is used to train with a 2D-only annotated data-sample. The geometric loss acts as a regularizer and penalizes the pose configurations that violate the consistency of bone-length ratio priors. Please note that the ground-truth xy-coordinates, $\hat{P}^{xy}$, with inferred depth, $\tilde{P}^z$ are used in $\mathcal{L}_g$ to make the training simple.

The geometric loss acts as an effective regularizer for the joint training and improves the accuracy of 3D pose estimation under controlled and in-the-wild test conditions, but it ignores certain other *strong* anatomical constraints of the human body. In the next section, we build upon the discussed weakly-supervised framework and propose a novel structure-aware loss that captures richer anatom-

ical constraints and provides stronger weakly-supervised regularization than the geometric loss.

4 Proposed Approach

This section introduces two novel anatomical loss functions and shows how to use them in the weakly-supervised setting to train with 2D annotated data-samples. Next, the motivation and derivation of the proposed losses and the analyses of the loss surfaces is presented to facilitate a deeper understanding and highlight the differences from the previous approaches. Lastly, a learnable temporal motion model is proposed with its detailed analysis through carefully designed controlled experiments.

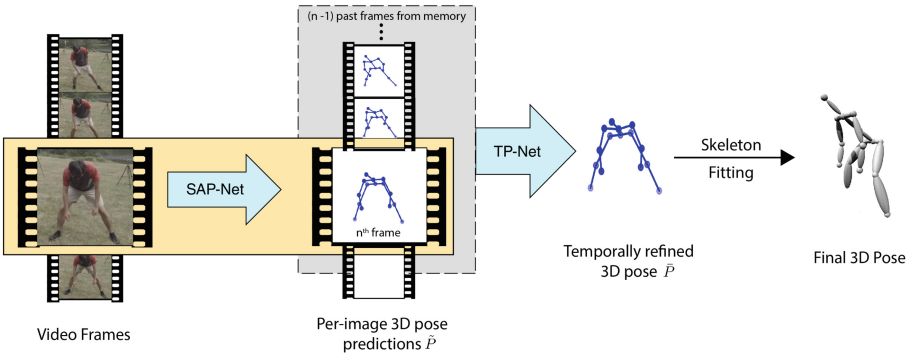


Fig. 2. Overall pipeline of our method: We sequentially pass the video frames to a ConvNet that produces 3D pose outputs (one at a time). Next, the prediction is temporally refined by passing a context of past N frames along with the current frame to a temporal model. Finally, skeleton fitting may be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

Figure 2 shows our complete pipeline for 3D pose estimation. It consists of

1. **Structure-Aware PoseNet or SAP-Net:** A single-frame based 3D pose-estimation system that takes a single RGB image I_i and outputs the inferred 3D pose $\tilde{P}_i$.
2. **Temporal PoseNet or TP-Net:** A learned temporal motion model that can take a continuous sequence of inferred 3D poses $\{\ldots, \tilde{P}_{i-2}, \tilde{P}_{i-1}\}$ and outputs a temporally harmonized 3D pose $\bar{P}_i$.
3. **Skeleton fitting:** Optionally, if the actual skeleton information of the subject is also available, we can carry out a simple skeleton fitting step which preserves the directions of the bone vectors.

4.1 Structure-Aware PoseNet or SAP-Net

SAP-Net uses the network architecture shown in Fig. 2, which is taken from [41]. This network choice allows joint learning with both 2D and 3D data in weakly-supervised fashion as described in Sect. 3. A 3D annotated data-sample provides strong supervision signal and drives the inferred depth towards a unique solution. On the other hand, weak-supervision, in the form of anatomical constraints, imposes penalty on invalid solutions, therefore, restricts the set of solutions. Hence, the stronger and more comprehensive the set of constraints, the smaller and better the set of solutions. We seek motivation from the discussion above and propose to use loss functions derived from joint-angle limits and left-right symmetry of human body in addition to bone-length ratio priors [41] for weak-supervision. Together, these three constraints are stronger than the bone-length ratio prior only and lead to better 3D pose configurations. For example, bone-length ratio prior will consider an elbow bent backwards as valid, if the bone ratios are not violated, but the joint-angle limits will invalidate it. Similarly, the symmetry loss eliminates the configurations with asymmetric left-right halves in the inferred pose. Next we describe and derive differentiable loss functions for the proposed constraints.

Illegal Angle Loss ($\mathcal{L}_a$): Most body joints are constrained to move within a certain angular limits only. Our illegal angle loss, $\mathcal{L}_a$, encapsulates this constraint for the knee and elbow joints and restricts their bending beyond 180°. For a given 2D pose P^{xy}, there exist multiple possible 3D poses and $\mathcal{L}_a$ penalizes the 3D poses that violate the knee or elbow joint-angle limits. To exploit such constraints, some methods [1,8,13] use non-differentiable functions to infer the legality of a pose. Unfortunately, the non-differentiability restricts their direct use in training a neural network. Other methods resort to representing a pose in terms of rotation matrices or quaternions for imposing joint-angle limits [1,38] that affords differentiability, but, makes it difficult to use in-the-wild 2D data (MPII). Therefore, this formulation is non-trivial when representing poses in terms of joint-positions, which are a more natural representation for ConvNets.

Our novel formulation of illegal-angle discovery resolves the ambiguity involved in differentiating between the internal and external angle of a joint for a 3D joint-location based pose representation. Using our formulation and keeping in mind our the requirement of differentiability, we formulate $\mathcal{L}_a$ to be used directly as a loss function. We illustrate our formulation with the help of Fig. 3, and explain its derivation 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 $\mathbf{v^r_{sn}} = P^r_s - P_n$, $\mathbf{v^r_{es}} = P^r_e - P^r_s$ and $\mathbf{v^r_{we}} = P^r_w - P^r_e$ as the collar-bone, upper-arm and the lower-arm, respectively (See Fig. 3). Now, $\mathbf{n^r_s} = \mathbf{v^r_{sn}} \times \mathbf{v^r_{es}}$ is the normal to the plane defined by the collar-bone and the upper-arm. For the elbow joint to be legal, $\mathbf{v^r_{we}}$ must have a positive component in the direction of $\mathbf{n^r_s}$, i.e. $\mathbf{n^r_s} \cdot \mathbf{v^r_{we}}$ must be positive. We do not incur any penalty when the joint angle is legal and define $E^r_e = \min(\mathbf{n^r_s} \cdot \mathbf{v^r_{we}}, 0)$ 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)

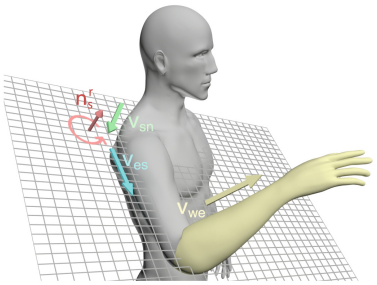


Fig. 3. Illustration of Illegal Angle loss: For the elbow joint angle to be legal, the lower-arm must project a positive component along $\mathbf{n_s^r}$ (normal to collarbone-upper arm plane), i.e. $\mathbf{n_s^r} \cdot \mathbf{v_{we}} \geq 0$. Note that we only need 2D annotated data to train our model using this formulation.

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. $\mathcal{L}_a$ can now be defined as:

$$\mathcal{L}_a = -E_e^r e^{-E_e^r} + E_e^l e^{E_e^l} + E_k^r e^{E_k^r} - E_k^l e^{-E_k^l} \tag{1}$$

All the terms in the loss are functions of bone vectors which are, in turn, defined in terms of the inferred pose. Therefore, $\mathcal{L}_a$ is differentiable. Please refer to the supplementary material for more details.

Symmetry Loss ($\mathcal{L}_s$): It is simple yet heavily constrains the joint depths, especially when the inferred 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

$$\mathcal{L}_s = \sum_{b \in \mathcal{B}} ||BL_b - BL_{C(b)}||_2 \tag{2}$$

where $C(.)$ indicates the corresponding left side bone.

Finally, our structure-aware loss $\mathcal{L}_{SA}^z$ is defined as weighted sum of illegal-angle loss $\mathcal{L}_a^z$, symmetry-loss $\mathcal{L}_s^z$ and geometric loss $\mathcal{L}_g^z$ from [41] -

$$\mathcal{L}_{SA}^z(\tilde{P}^z, \hat{P}^{xy}) = \lambda_a \mathcal{L}_a(\tilde{P}^z, \hat{P}^{xy}) + \lambda_s \mathcal{L}_s(\tilde{P}^z, \hat{P}^{xy}) + \lambda_g \mathcal{L}_g(\tilde{P}^z, \hat{P}^{xy}) \tag{3}$$

Loss Surface Visualization: Here we take help of local loss surface visualization to appreciate how the proposed losses are pushing invalid configurations towards their valid counterparts. In order to obtain the loss surfaces we take a random pose P and vary the (x_{le}, z_{le}) coordinates of left elbow over an XZ grid while keeping all other joint locations fixed. Then, we evaluate $\mathcal{L}_{SA}^z$ at different (x, z) locations in the XZ grid to obtain the loss, which is plotted as surfaces in Fig. 4. We plot loss surfaces with only 2D-location loss,

2D-location+symmetry loss, 2D-location+symmetry+illegal angle loss and 3D-annotation based Euclidean loss to show the evolution of the loss surfaces under different anatomical constraints. From the figure it is clear that both the symmetry loss and illegal angle loss morph the loss surface to facilitate moving away from illegal joint configurations.

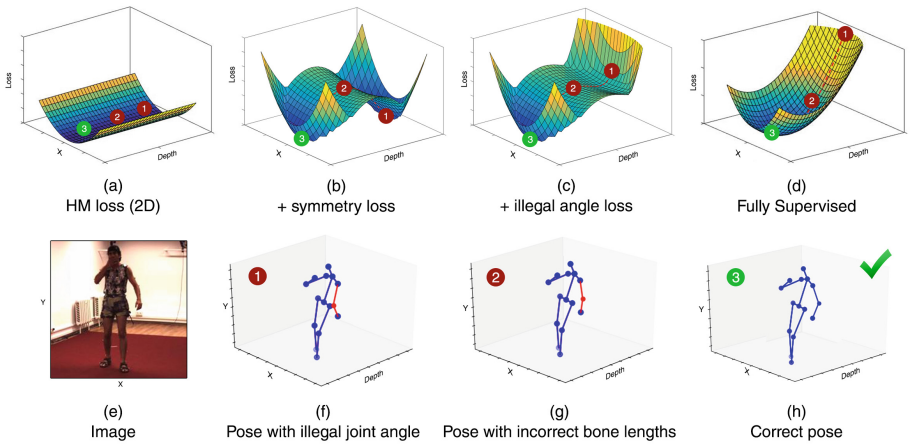
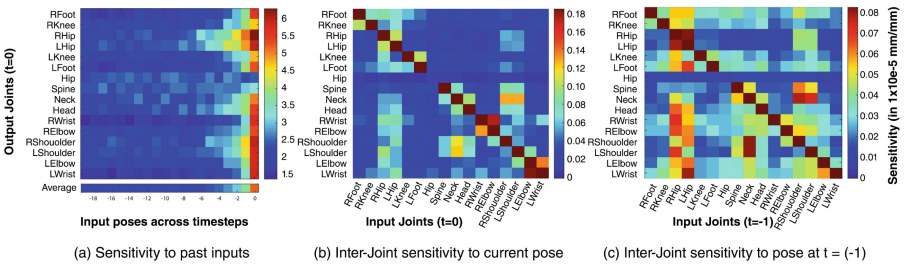


Fig. 4. Loss Surface Evolution: Plots (a) to (d) show the local loss surfaces for (a) 2D-location loss. (b) 2D-location+symmetry loss (c) 2D-location+symmetry+illegal angle loss and (d) full 3D-annotation Euclidean loss. The points (1), (2) and (3) highlighted on the plots are the corresponding 3D poses shown in (f), (g) and (h), with (3) being the ground-truth depth. 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 symmetry is lost. Pose (3) is correct. We can see that without the angle loss, the loss at (1) and (3) are equal and we cannot discern between the two points.

4.2 Temporal PoseNet or TP-Net

In this section we propose to learn a temporal pose model, referred as Temporal PoseNet, to exploit the temporal consistency and motion cues present in video sequences. Given independent pose estimates from SAP-Net, we seek to exploit the information from a set of adjacent pose-estimates $\mathbf{P_{adj}}$ to improve the inference for the required pose P. We propose to use a simple two-layer, 4096 hidden neurons, fully-connected network with ReLU non-linearity that takes a fixed number, $N = 20$, of adjacent poses as inputs and outputs the required pose $\bar{P}$. The adjacent pose vectors are simply flattened and concatenated in order to make a single vector that goes into the TP-Net and it is trained using standard L_2 loss from the ground-truth pose. Despite being extremely simple in nature, we show that it outperforms a more complex variant such as RNNs, see Table 4. Why? We believe it happens because intricate human motion has increasing variations possible with increasing time window, which perhaps makes



(a) Sensitivity to past inputs (b) Inter-Joint sensitivity to current pose (c) Inter-Joint sensitivity to pose at t = (-1)

Fig. 5. (a) The variation of sensitivity in output pose w.r.t to the perturbations in input poses of TP-Net for 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. The self correlations (diagonal elements) are an order of magnitude larger and the colormap range has been capped to better display. (Color figure online)

additional information from too far in the time useless or at least difficult to utilize. Therefore, a dense network with a limited context can effectively capture the useful consistency and motion cues.

In order to visualize the temporal and structural information exploited by TP-Net we carried out a simple sensitivity analysis in which we randomly perturbed the joint locations of P_t that is t time-steps away from the output of TP-Net $\bar{P}$ and plot the sensitivity for time-steps $t = -1$ to $t = -19$ for all joints in Fig. 5(a). We can observe that poses beyond 5 time-steps (or 200 ms time-window) does not have much impact on the predicted pose. Similarly, Fig. 5(b) shows the structural correlations the model has learned just within the current frame. TP-Net learns to rely on the locations of hips and shoulders to refine almost all the other joints. We can also observe that the child joints are correlated with parent joints, for e.g. the wrists are strongly correlated with elbows, and the shoulders are strongly correlated with the neck. Figure 5(c) shows the sensitivity to the input pose at $t = -1$. Here, the correlations learned from the past are weak, but exhibit a richer pattern. The sensitivity of the child joints extends further upwards into the kinematic chain, e.g., the wrist shows higher correlations with elbow, shoulder and neck, for the $t = -1$ frame. Therefore, we can safely conclude that TP-Net learns complex structural and motion cues despite being so simple in nature. We hope this finding would be useful for future research in this direction.

Since TP-Net takes as input a fixed number of adjacent poses, we can choose to take all the adjacent poses before the required pose, referred to as *online* setting, or we can choose to have $N/2 = 10$ adjacent poses on either side of required pose, referred to as *semi-online* setting. Since our entire pipeline runs at 30 fps, even semi-online setting will run at a lag of 10 fps only. From Fig. 5 we observe that TP-Net can learn complex, non-linear inter-joint dependencies over time - for e.g. it learns to refine wrist position, for which the tracking is least accurate, based on the past motion of elbow and shoulder joints.

4.3 Training and Implementation Details

While training the SAP-Net, both 2D samples, from MPII2D, and 3D samples, from either of the 3D datasets, were consumed in equal proportion in each iteration with a minibatch size of 6. In the *first stage* we obtain a strong 2D pose estimation network by pre-training the hourglass modules of SAP-Net on MPII and Human3.6 using SGD as in [25]. Training with weakly-supervised losses require a warm start [44], therefore, in the *second stage* we train the 3D depth module with only 3D annotated data-samples for 240k iterations so that it learns to output reasonable poses before switching on weak-supervision. In the *third stage* we train SAP-Net with $\mathcal{L}_g$ and $\mathcal{L}_a$ for 160k iterations with $\lambda_a = 0.03$, $\lambda_g = 0.03$ with a learning-rate of $2.5e{-}4$. Finally, in the *fourth stage* we introduce the symmetry loss, $\mathcal{L}_f$ with $\lambda_s = 0.05$ and learning-rate $2.5e{-}5$.

TP-Net was trained using Adam optimizer [16] for 30 epochs using the pose predictions generated by fully-trained SAP-Net. In our experiments, we found that a context of $N = 20$ frames yields the best improvement on MPJPE (Fig. 5) and we use that in all our experiments. It took approximately two days to train SAP-Net and one hour to train TP-Net using one NVIDIA 1080 Ti GPU. SAP-Net runs at an average testing time of 20 ms per image while TP-Net adds negligible delay (<1 ms).

5 Experiments

In this section, we present ablation studies, quantitative results on Human3.6M and MPI-INF-3DHP datasets and comparisons with previous art, and qualitative results on MPII 2D and MS COCO datasets. We start by describing the datasets used in our experiments.

Human3.6M has 11 subjects performing different indoor actions with ground-truth annotations captured using a marker-based MoCap system. We follow [35] and evaluate our results under (1) *Protocol 1* that uses Mean Per Joint Position Error (MPJPE) as the evaluation metric w.r.t. root relative poses and (2) *Protocol 2* that uses Procrustes Aligned MPJPE (PAMPJPE) which is MPJPE calculated after rigid alignment of predicted pose with the ground truth. As is common, we evaluate the results on every fifth frame.

MPI-INF-3DHP (test) dataset is a recently released dataset of 6 test subjects with different indoor settings (green screen and normal background) and 2 subjects performing in-the-wild that makes it more challenging than Human3.6M, which only has a single indoor setting. We follow the evaluation metric proposed in [22] and report Percentage of Correct Keypoints (PCK) within *150* mm range and Area Under Curve (AUC). Like [41], we assume that the global scale is known and perform skeleton retargeting while training to account for the difference of joint definitions between Human3.6M and MPI-INF-3DHP datasets. Finally, skeleton fitting is done as an optional step to fit the pose into a skeleton of known bone lengths.

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.

Method	Direction	Discuss	Eat	Greet	Phone	Pose	Purchase	Sit
Zhou [43]	68.7	74.8	67.8	76.4	76.3	84.0	70.2	88.0
Jahangiri [15]	74.4	66.7	67.9	75.2	77.3	70.6	64.5	95.6
Lin [19]	58.0	68.2	63.2	65.8	75.3	61.2	65.7	98.6
Mehta [22]	57.5	68.6	59.6	67.3	78.1	56.9	69.1	98.0
Pavlakos [27]	58.6	64.6	63.7	62.4	66.9	57.7	62.5	76.8
Zhou [41]	54.8	60.7	58.2	71.4	62.0	53.8	55.6	75.2
Sun [34]	52.8	54.8	54.2	54.3	61.8	53.1	53.6	71.7
Ours (SAP-Net)	46.9	53.8	47.0	52.8	56.9	45.2	48.2	68.0
Ours (TP-Net)	**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
Zhou [43]	113.8	78.0	78.4	89.1	62.6	75.1	73.6	79.9
Jahangiri [15]	127.3	79.6	79.1	73.4	67.4	71.8	72.8	77.6
Lin [19]	127.7	70.4	93.0	68.2	50.6	72.9	57.7	73.1
Mehta [22]	117.5	69.5	82.4	68.0	55.3	76.5	61.4	72.9
Pavlakos [27]	103.5	65.7	70.7	61.6	56.4	69.0	59.5	66.9
Zhou [41]	111.6	64.1	65.5	66.0	51.4	63.2	55.3	64.9
Sun [34]	**86.7**	61.5	67.2	53.4	47.1	61.6	53.4	59.1
Ours (SAP-Net)	94.0	55.7	63.6	51.6	40.3	55.4	44.3	55.5
Ours (TP-Net)	87.3	**51.7**	**61.4**	**48.5**	**37.6**	**52.2**	**41.9**	**52.1**

Table 2. Comparative evaluation of our model on Human 3.6M using Protocol 2. The models were trained only on Human3.6M and MPII 2D datasets.

Method	Direct.	Discuss	Eat	Greet	Phone	Pose	Purch.	Sit	Sit Down	Smoke	Photo	Wait	Walk	Walk Dog	Walk Pair	Avg
Yasin [40]	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 [29]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	88.1
Chen [7]	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 [24]	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 [43]	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 [34]	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(SAP-Net)	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 (TP-Net)	**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**

2D Datasets: MS-COCO and MPII are in-the-wild 2D pose datasets with no 3D ground truth annotations. Therefore, we show qualitative results for both of them in Fig. 6. Despite lack of depth annotation, our approach generalizes well and predicts valid 3D poses under background clutter and significant occlusion.

Table 3. Ablation of different loss terms on Human3.6M using Protocol 1.

Method	MPJE
Zhou w/o $\mathcal{L}_g$ [41]	65.69
+ Geometry loss	64.90
Baseline	58.50
Baseline + $\mathcal{L}_s$	58.30
Baseline + $\mathcal{L}_a$	57.70
Baseline + $\mathcal{L}_g$	58.30
Baseline + $\mathcal{L}_g$ + $\mathcal{L}_a$	56.20
Baseline + $\mathcal{L}_g$ + $\mathcal{L}_a$ + $\mathcal{L}_s$	55.51
Baseline + $\mathcal{L}_g$ + $\mathcal{L}_a$ + $\mathcal{L}_s$ + TP-Net	**52.10**
Baseline + $\mathcal{L}_g$ + $\mathcal{L}_a$ + $\mathcal{L}_s$ + Bi-TP-Net	**51.10**

Table 4. Comparison of different temporal models considered with varying context sizes. LSTM nets model the entire past context till time t. Bidirectional networks take half contextual frames from the future and half from the past.

Model	Number of input frames		
	4	10	20
LSTM	-	-	54.05
Bi-LSTM	53.86	53.72	53.65
TP-Net	53.0	52.24	52.1
Bi-TP-Net	52.4	51.36	**51.1**

5.1 Quantitative Evaluations

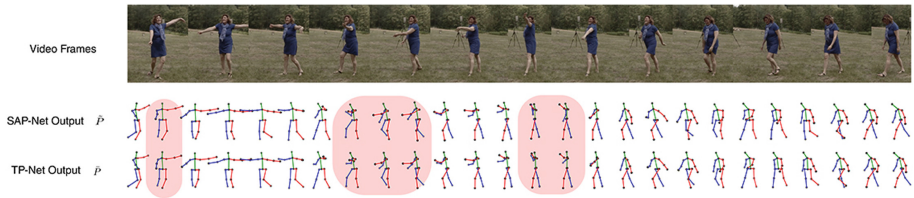
We evaluate the outputs of the three stages of our pipeline and show improvements at each stage.

1. **Baseline:** We train the same network architecture as SAP-Net but with only the fully supervised losses i.e. 2D heatmap supervision and $\mathcal{L}^e$ for 3D data only.
2. **SAP-Net:** Trained with the proposed structure-aware loss following Sect. 4.3.
3. **TP-Net:** Trained on the outputs of SAP-Net from video sequences (see Sect. 4.3).
4. **Skeleton Fitting (optional):** We fit a skeleton based on the subject's bone lengths while preserving the bone vector directions obtained from the 3D pose estimates.

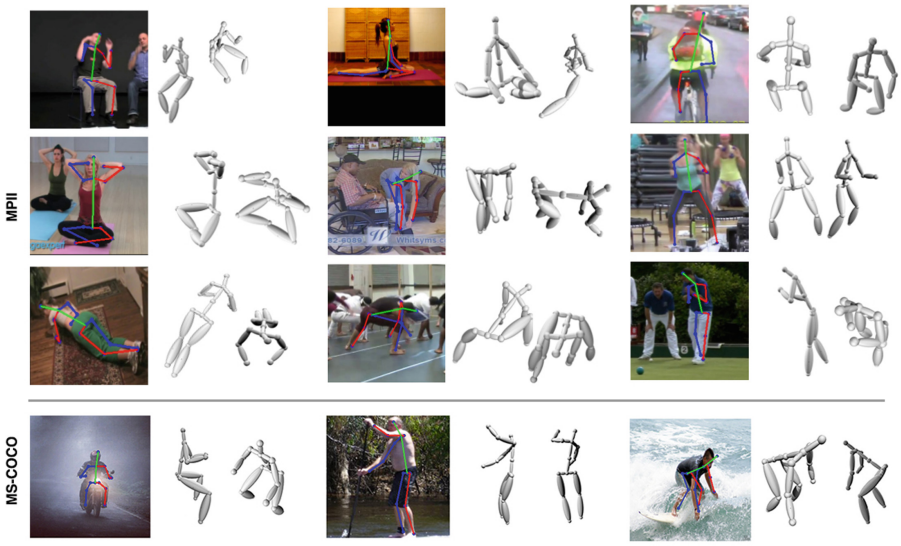
Below, we conduct ablation study on SAP-Net and report results on the two datasets.

SAP-Net Ablation Study: In order to understand the effect of individual anatomical losses, we train SAP-Net with successive addition of geometry $\mathcal{L}_g^z$, illegal-angle $\mathcal{L}_a^z$ and symmetry $\mathcal{L}_s^z$ losses and report their performance on Human3.6M under *Protocol 1* in Table 3. We can observe that the incorporation of illegal-angle and symmetry losses to geometry loss significantly improves the performance while geometry loss does not offer much improvement even over the baseline. Similarly, TP-Net offers significant improvements over SAP-Net and the *semi-online* variant of TP-Net (Bi-TP-Net) does even better than TP-Net.

Evaluations on Human3.6M: We show significant improvement over the state-of-the-art and achieve an MPJPE of 55.5 mm with SAP-Net which is further improved by TP-Net to 52.1 mm. Tables 1 and 2 present a comparative analysis of our results under *Protocol 1* and *Protocol 2*, respectively. We outperform



(a) Qualitative results of TP-Net and SAP-Net on a video sequence



(b) Qualitative results of SAP-Net on MPII and MS-COCO

Fig. 6. (a) Comparison of our temporal model TP-Net with SAP-Net on a video. The highlighted poses demonstrate the ability of TP-Net to learn temporal correlations, and smoothen and refine pose estimates from SAP-Net. (b) Qualitative results of SAP-Net on some images from MPII and MS-COCO datasets, from multiple viewpoints.

other competitive approaches by significant margins leading to an improvement of 12%.

Evaluations on MPI-INF-3DHP: The results from Table 5 show that we achieve slightly worse performance in terms of PCK and AUC but much better performance in terms of MPJPE, improvement of 12%, as compared to the current state-of-the-art. It is despite the lack of data augmentation through green-screen compositing during training.

5.2 Structural Validity Analysis

This section analyzes the validity of the predicted 3D poses in terms of the anatomical constraints, namely left-right symmetry and joint-angle limits. Ide-

Table 5. Results on MPI-INF-3DHP dataset. Higher PCK and AUC are desired while a lower MPJPE is better. Note that unlike [22,23], the MPI-INF-3DHP training dataset was not augmented.

Method	PCK	AUC	MPJPE
Mehta [22]	75.7	39.3	117.6
Mehta [23]	76.6	**40.4**	124.7
Ours	**76.7**	39.1	**103.8**

Table 6. Evaluating our models on (i) symmetry - mean L_1 distance in mm between left/right bone pairs (upper half), and (ii) the standard deviation (in mm) of bone lengths across all video frames (lower half) on MPI-INF-3DHP dataset.

Bone	Zhou [41]	SAP-Net	TP-Net
Upper arm	37.8	$25.8_{\downarrow 31.7\%}$	$\mathbf{23.9}_{\downarrow 36.7\%}$
Lower arm	50.7	$\mathbf{32.1}_{\downarrow 36.7\%}$	$33.9_{\downarrow 33.1\%}$
Upper leg	43.4	$27.8_{\downarrow 35.9\%}$	$\mathbf{24.8}_{\downarrow 42.8\%}$
Lower leg	47.8	$38.2_{\downarrow 20.1\%}$	$\mathbf{29.2}_{\downarrow 38.9\%}$
Upper arm	–	49.6	**39.8**
Lower arm	–	66.0	**48.3**
Upper leg	–	61.3	**48.8**
Lower leg	–	68.8	**48.3**

ally, the corresponding left-right bone pairs should be of similar length; therefore, we compute the mean L_1 distance in mm between the corresponding left-right bone pairs on MPI-INF-3DHP dataset and present the results in the upper half of Table 6. For fairness of comparison, we evaluate on model trained only on Human3.6M. We can see that SAP-Net, trained with symmetry loss, significantly improves the symmetry as compared to the system in [41] which uses bone-length ratio priors and TP-Net offers further improvements by exploiting the temporal cues from adjacent frames. It shows the importance of explicit enforcement of symmetry. Moreover, it clearly demonstrates the effectiveness of TP-Net in implicitly learning the symmetry constraint. The joint-angle validity of the predicted poses is evaluated using [1] and we observe only 0.8% illegal non-torso joint angles as compared to 1.4% for [41].

The lower-half of Table 6 tabulates the standard deviation of bone lengths in mm across frames for SAP-Net and TP-Net. We can observe that TP-Net reduces the standard deviation of bone-length across the frames by 28.7%. 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*. Finally, we present some qualitative results in Fig. 6 and in the supplementary material to show that TP-Net effectively corrects the jerks in the poses predicted by SAP-Net.

6 Conclusion

We proposed two anatomically inspired loss functions, namely illegal-angle and symmetry loss. We showed them to be highly effective for training weakly-supervised ConvNet architectures for predicting valid 3D pose configurations

from a single RGB image in-the-wild setting. We analyzed the evolution of local loss surfaces to clearly demonstrate the benefits of the proposed losses. We also proposed a simple, yet surprisingly effective, sliding-window fully-connected network for temporal pose modeling from a sequence of adjacent poses. We showed that it is capable of learning semantically meaningful short-term temporal and structure correlations. Temporal model was shown to significantly reduce jitters and noise from pose prediction for video sequences while taking <1 ms per inference. Our complete pipeline improved the published state-of-the-art by 11.8% and 12% on Human3.6M and MPI-INF-3DHP, respectively while running at 30 fps on NVIDIA Titan 1080Ti GPU.

Acknowledgement. This work is supported by Mercedes-Benz Research & Development India (RD/0117-MBRDI00-001).

References

1. Akhter, I., Black, M.J.: Pose-conditioned joint angle limits for 3D human pose reconstruction. In: CVPR (2015)
2. Alldieck, T., Kassubeck, M., Wandt, B., Rosenhahn, B., Magnor, M.: Optical flow-based 3D human motion estimation from monocular video. In: Roth, V., Vetter, T. (eds.) GCPR 2017. LNCS, vol. 10496, pp. 347–360. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66709-6_28
3. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2D human pose estimation: New benchmark and state of the art analysis. In: CVPR (2014)
4. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: automatic estimation of 3D human pose and shape from a single image. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 561–578. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_34
5. Casiez, G., Roussel, N., Vogel. D.: 1 filter: a simple speed-based low-pass filter for noisy input in interactive systems. In: SIGCHI (2012)
6. Sminchisescu, C., Ionescu, C., Li, F.: Latent structured models for human pose estimation. In: ICCV (2011)
7. Chen, C.-H., Ramanan, D.: 3D human pose estimation = 2D pose estimation + matching. In: CVPR (2017)
8. Chen, J., Nie, S., Ji, Q.: Data-free prior model for upper body pose estimation and tracking. IEEE Trans. Image Process. **22**, 4627–4639 (2013)
9. Chen, W., et al.: Synthesizing training images for boosting human 3D pose estimation. In: 3DV (2016)
10. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
11. Coskun, H., Achilles, F., DiPietro, R., Navab, N., Tombari, F.: Long short-term memory Kalman filters: recurrent neural estimators for pose regularization. In: ICCV (2017)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
13. Herda, L., Urtasun, R., Fua, P.: Hierarchical implicit surface joint limits for human body tracking. Comput. Vis. Image Underst. **99**, 189–209 (2005)

14. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE TPAMI **36**, 1325–1339 (2014)
15. Jahangiri, E., Yuille, A.L.: Generating multiple diverse hypotheses for human 3D pose consistent with 2D joint detections. In: ICCV (2017)
16. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: ICLR (2015)
17. Li, S., Chan, A.B.: 3D human pose estimation from monocular images with deep convolutional neural network. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9004, pp. 332–347. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16808-1_23
18. Li, S., Zhang, W., Chan, A.B.: Maximum-margin structured learning with deep networks for 3D human pose estimation. In: ICCV (2015)
19. Lin, M., Lin, L., Liang, X., Wang, K., Cheng, H.: Recurrent 3D pose sequence machines. In: CVPR (2017)
20. Lin, T., et al.: Microsoft COCO: common objects in context. arXiv preprint arXiv:1405.0312 (2014)
21. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: a skinned multi-person linear model. ACM Trans. Graph. **34**, 248 (2015)
22. Mehta, D., et al.: Monocular 3D human pose estimation in the wild using improved CNN supervision. In: 3DV (2017)
23. Mehta, D.: VNect: real-time 3D human pose estimation with a single RGB camera. ACM ToG **36**, 44 (2017)
24. Moreno-Noguer, F.: 3D human pose estimation from a single image via distance matrix regression. In: CVPR (2017)
25. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
26. Park, M.J., Choi, M.G., Shinagawa, Y., Shin, S.Y.: Video-guided motion synthesis using example motions. ACM ToG **25**, 1327–1359 (2006)
27. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3D human pose. In: CVPR (2017)
28. Ramakrishna, V., Kanade, T., Sheikh, Y.: Reconstructing 3D human pose from 2D image landmarks. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7575, pp. 573–586. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33765-9_41
29. Rogez, G., Schmid, C.: MoCap-guided data augmentation for 3D pose estimation in the wild. In: NIPS (2016)
30. Rogez, G., Weinzaepfel, P., Schmid, C.: LCR-Net: localization-classification-regression for human pose. In: CVPR (2017)
31. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. ArXiv e-prints (2014)
32. Sarafianos, N., Boteanu, B., Ionescu, B., Kakadiaris, I.A.: 3D human pose estimation: a review of the literature and analysis of covariates. Comput. Vis. Image Underst. **152**, 1–20 (2016)
33. Sminchisescu, C., Triggs, B.: Estimating articulated human motion with covariance scaled sampling. Int. J. Robot. Res. **22**, 371–391 (2003)
34. Sun, X., Shang, J., Liang, S., Wei, Y.: Compositional human pose regression. In: ICCV (2017)
35. Tome, D., Russell, C., Agapito, L.: Lifting from the deep: convolutional 3D pose estimation from a single image. In: CVPR (2017)

36. Urtasun, R., Fleet, D.J., Fua, P.: Temporal motion models for monocular and multiview 3D human body tracking. Comput. Vis. Image Underst. **104**, 157–177 (2006)
37. Varol, G., et al.: Learning from synthetic humans. In: CVPR (2017)
38. Wei, X., Chai, J.: VideoMocap: modeling physically realistic human motion from monocular video sequences. ACM ToG **29**, 42 (2010)
39. Nie, B.X., Wei, P., Zhu, S.-C.: Monocular 3D human pose estimation by predicting depth on joints. In: ICCV, October 2017
40. Yasin, H., Iqbal, U., Kruger, B., Weber, A., Gall, J.: A dual-source approach for 3D pose estimation from a single image. In: CVPR (2016)
41. Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Towards 3D human pose estimation in the wild: a weakly-supervised approach. In: ICCV (2017)
42. Zhou, X., Sun, X., Zhang, W., Liang, S., Wei, Y.: Deep kinematic pose regression. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9915, pp. 186–201. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49409-8_17
43. Zhou, X., Zhu, M., Derpanis, K., Daniilidis, K.: Sparseness meets deepness: 3D human pose estimation from monocular video. In: CVPR (2016)
44. Zhou, Z.-H.: A brief introduction to weakly supervised learning. Natl. Sci. Rev. **5**, 44–53 (2017)