Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision

Dushyant Mehta[1], Helge Rhodin[†], Dan Casas[††],
Oleksandr Sotnychenko[1], Weipeng Xu[1] and Christian Theobalt[1]

[1]Max Planck Institute For Informatics, Germany
[†]Ecole Polytechnique Fédérale de Lausanne, Switzerland
[††]Universidad Rey Juan Carlos, Spain

Abstract

We propose a CNN-based approach for 3D human body pose estimation from single RGB images, that addresses the issue of limited generalizability of models trained solely on the starkly limited publicly available 3D pose data. We propose novel CNN supervision techniques, using a regularization structure while training that extends the concept of multi-level skip connections, and leverage first and second order parent relationships along the skeletal kinematic tree to learn better representations. We introduce a new training set for human body pose estimation from monocular images of real humans, that has the ground truth captured with a multi-camera marker-less motion capture system. It complements existing corpora with greater diversity in pose, human appearance, clothing, occlusion, and viewpoints, and enables an increased scope of augmentation. We also contribute a new benchmark that covers outdoor and indoor scenes. We further combine it with transfer learning from 2D pose human pose prediction to achieve even better generalization, and improve over the state-of-the-art on standard benchmarks by more than 25%. We argue that the use of transfer learning of representations in tandem with algorithmic and data contributions is crucial for general progress along many different dimensions of the problem.

1. Introduction

We present a new method to estimate 3D articulated human pose from a single RGB image taken in a general environment. It has a notably higher accuracy than known state-of-the-art methods from the literature [10, 63, 36]. 3D human pose estimation from monocular RGB input is a timely and very challenging research problem that expands the scope of widely researched monocular 2D pose estimation. It has many practical applications in general scene understanding, and man-machine interaction. Our per-image setting differs from but is related to marker-less 3D motion capture methods that *track* articulated human poses from *multi-view* video sequences, often in very controlled scenes [72, 59, 60, 69, 7, 21, 61, 14]. Special RGB-D cameras enable real-time monocular pose estimation [55] or motion tracking [6], but often do not work in general scenes. In 2D joint detection and pose estimation from RGB, data-driven approaches using Convolutional Neural Networks (CNNs) have shown impressive results [16, 71, 67, 68, 26, 44, 12, 8, 42, 38, 23, 25, 13], outperforming previous hand crafted and model-based methods by a large margin [1, 5, 19].

Direct 3D pose regression from monocular RGB, however, remains challenging. A common approach is to lift 2D keypoints to 3D [73, 10, 70, 37, 77, 80, 76, 58, 57], but this requires computationally expensive iterative pose optimization which may be unstable under depth ambiguities. Though recent advances in direct CNN-based 3D regression show promise, utilizing different prediction space formulations [63, 36, 78] and incorporating additional constraints, e.g. [78, 65, 80, 75], they are far from the accuracy levels seen for 2D pose prediction.

The difficult nature of the problem aside, 3D pose prediction is further stymied by the lack of suitably large and diverse annotated 3D pose corpora, particularly due to the infeasibility of manual 3D body pose annotation, unlike 2D body pose data [53, 4, 31]. Existing datasets use marker-based motion capture for 3D annotation [28, 56], which restricts recording to skin-tight clothing, or markerless systems in a dome of hundreds of cameras [33], which enables diverse clothing but requires an expensive studio setup.

We ameliorate the lack of data through new neural network supervision techniques and data contributions:

First, in Section 3.2.1, we demonstrate the use of skip connections in CNNs as a training-time regularization structure to learn better representations from training data with limited appearance variation, leading to better generalization to in-the-wild data. In Section 3.2.2, we learn to select and fuse pose-dependent-constraints derived from the kinematic structure of the skeleton, and discuss this from the point of view of intermediate supervision. In Section 3.3 we derive a closed form solution to localize the global 3D position of the skeleton, and correct for perspective issues resulting from bounding-box cropping.

Second, in Section 4, we introduce the new MPI-INF-3DHP dataset of real humans with ground truth 3D anno-

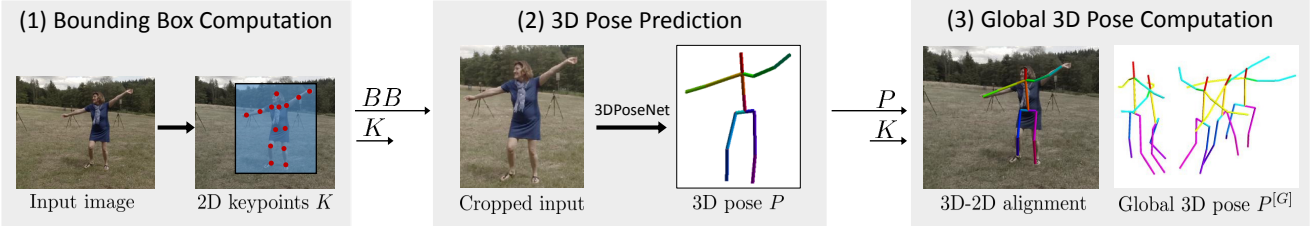

Figure 1. We infer 3D pose from single image in three stages: (1) extraction of the actor bounding box from 2D detections; (2) direct CNN-based 3D pose regression; and (3) global root position computation in original footage by aligning 3D to 2D pose.

tations from a state-of-the-art markerless motion capture system. It complements existing datasets with everyday clothing appearance, a large range of motions, interactions with objects, and more varied camera viewpoints. The data capture approach eases appearance augmentation to extend the captured variability, complemented with improvements to existing augmentation methods for enhanced foreground texture variation. This gives a further significant boost to the accuracy and generalizability of the learned models.

Third, in Section 5, we explore the use of transfer learning to leverage the highly relevant mid and high level features learned on the readily available in-the-wild 2D pose datasets [4, 32]. In addition to the aforementioned architectural improvements, this leads to further improved accuracy and generalizability compared to previous methods, without requiring more 3D training data.

The components of our method are thoroughly evaluated on existing test datasets, showing significant accuracy improvement on the state-of-the-art of more than 25%. Further we introduce a new test set, including sequences outdoors with accurate annotation, on which we demonstrate the generalization capability of the proposed method and validate the value of our new dataset.

2. Related Work

Both, learning based and model based approaches have been used for human body pose estimation from monocular images, with much of the recent progress coming through neural network based approaches. We review the various approaches, and discuss their relation with our work.

3D pose from 2D estimates Deep CNN architectures have dramatically improved 2D pose estimation [29, 42], and even run in real-time applications [71]. Graphical models [19, 1] maintain their merit for modeling multi-person relations [44]. 3D pose can be inferred from 2D pose through geometric and statistical priors [41, 62]. Optimization of the projection of a 3D human model to the 2D predictions is computationally expensive, but allows incorporation of various constraints. The ambiguity of 3D pose under projection can be countered with pose priors and inter-penetration constraints [10], sparsity assumptions [70, 77, 79], joint limits [17, 2], and temporal constraints [49]. Simo-Serra *et al*. [58] sample noisy 2D predictions to ambiguous 3D shapes, which they disambiguate using kinematic constraints, and improve discriminative 2D detection from likely 3D samples [57]. Li *et al*. look up the nearest neighbours in a learned joint embedding of human images and 3D poses [37] to estimate 3D pose from an image. We choose to use the geometric relations between the predicted 2D and 3D skeleton pose to infer the global subject position.

Estimating 3D pose directly Additional image information, e.g. on the front-back orientation of limbs, can be exploited by regressing 3D pose directly from the input image [63, 36, 78, 27]. Deep CNNs achieve state-of-the-art results [78, 64]. While CNNs dominate, regression forests have also been used to derive 3D *posebit descriptors* efficiently [46]. The input and output representations are important too. To localize the person, the input image is commonly cropped to the bounding box of the subject before 3D pose estimation [27]. Video input provides temporal cues, which translate to increased accuracy [65, 80]. The downside of conditioning on motion is the increased input dimensionality, and requires motion databases with sufficient motion variation, which are even harder to capture than pose data sets. In controlled conditions, fixed camera placement provides additional height cues [75]. Since monocular reconstruction is inherently scale-ambiguous, 3D joint positions relative to the pelvis, with normalized subject height are widely used as the output. To explicitly encode dependencies between joints, Tekin *et al*. [63] regressing to a high-dimensional pose representation, learned by an auto encoder. Li *et al*. [36] report that predicting positions relative to the parent joint of the skeleton improves performance, but we show that a pose-dependent combination of absolute and relative positions leads to further improvements. Zhou *et al*. [78] regress joint angles of a skeleton from single images, using a kinematic model.

Addressing the scarcity and limited appearance variability of datasets Learning based methods require large annotated dataset corpora. 3D annotation[27] is harder to obtain than 2D pose annotation. Some approaches treat 3D pose as a hidden variable, and use pose priors and projection to 2D to guide the training [11, 73]. Rogez *et al*. render mosaics of in-the-wild human pose images using projected mocap data [51]. Chen *et al*. [15] render textured

rigged human models, but still require domain adaptation to in-the-wild images for generalization. Our new dataset complements the existing datasets, through extensive appearance and pose variation, by using marker-less annotation and providing an increased scope for augmentation.

Transfer Learning [43] is commonly used in computer vision to leverage features and representations learned on one task to offset scarce data available for a related task. Low and/or mid-level CNN features can be shared also among unrelated tasks [54, 74]. Pretraining on ImageNet [52] is commonly used for weight initialization [26, 64]. We explore the use of low and mid-level features learned on in-the-wild 2D pose datasets for further improving the generalization of 3D pose prediction models.

3. CNN-based 3D Pose Estimation

Given an RGB image, we estimate the global 3D human pose $P^{[G]}$ in the camera coordinate system. We estimate the global positions of the joints of the skeleton depicted in Figure 2, accounting for the camera viewpoint, which goes beyond only estimating in a root-centered (pelvis) coordinate system, as is common in many previous works. Our algorithm consists of three steps, as illustrated in Figure 1. **(1)** the subject is localized in the frame with a 2D bounding box BB, computed from 2D joint heatmaps H, obtained with a CNN we call *2DPoseNet*; **(2)** the root-centered 3D pose P is regressed from the BB-cropped input with a second CNN termed *3DPoseNet*; and **(3)** global 3D pose coordinates $P^{[G]}$ and perspective correction are computed in closed form using 3D pose P, 2D joint locations K (extracted from H) and known camera calibration.

3.1. Bounding Box and 2D Pose Computation

We use the person detection method of Wei *et al*. [71] to get a localization heatmap, which, together with the original image, are used by our *2DPoseNet* to produce 2D joint location heatmaps H. The heat map maxima provide the most likely 2D joint locations K which are used to infer the person's bounding-box BB. See Figure 1 left. The 2D joint locations K are further used for global pose estimation in Section 3.3.

Our *2DPoseNet* is fully convolutional is trained on MPII [4] and LSP [32, 31] datasets. We use a CNN structure based on Resnet-101 [24] up to the filter banks at level 4, with striding and identity skip connections removed from level 5. For specifics of the network architecture and the training scheme, refer to the supplementary document.

3.2. 3D Pose Regression

The 3D pose CNN, termed *3DPoseNet*, is used to regress root-centered 3D pose P from a cropped RGB image, and makes use of our new CNN supervision techniques. Figure 3 depicts the main components of the method, detailed in the following sections.

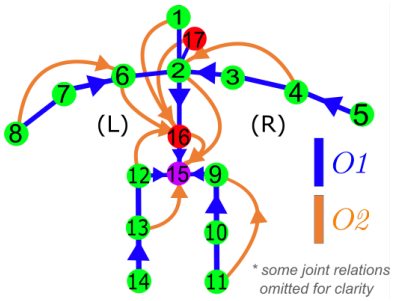

Figure 2. 3D pose, represented as a vector of 3D joint positions, is expressed variously as 1) P: relative to the root (joint #15), 2) O*1* (blue): relative to first order and, 3) O2 (orange): relative to second order parents in the kinematic skeleton hierarchy.

Network The *Base* network derives from Resnet-101 as well, and it is identical to *2DPoseNet* up to *res5a*. We remove the remaining layers from level 5. 3D prediction stubs comprised of a convolution layer $(k_{5\times5}, s_2)$ with 128 features and a final fully-connected layer that outputs the 3D joint locations are added on the top. Additionally we use intermediate supervision with heatmaps H and pose P. Refer to the supplementary for specifics of attachment points and loss weights.

3.2.1 Multi-level Corrective Skip Connections

The novelty of our method lies on the use multi-level skip connections [39] as a training-time regularization architecture that is not used during deployment, in contrast to vanilla skip-connections.

It is based on two insights regarding multi-level skip connections: First, the underlying notion of coarse and fine scaled information coming from different levels can be generalized beyond image-like predictions to regression of quantities such as the pose vector in our case. Second, the deepest level with its larger receptive field should intuitively contribute the overall pose. While connections from shallower levels, with their limited receptive fields, should only contribute corrections per joint based on local evidence. This is not guaranteed with regular skip connections.

To enforce the latter, we add an additional loss term at the output of the deepest contributor P_{deep} to the multi-level skip sum. Figure 3 shows a schematic overview. It forces the last stage of the core network to be the dominant contributor to the skip sum P_{sum}, leaving the skip connections to only contribute corrective terms to the "as good as possible" prediction at P_{deep}.

3.2.2 Multi-modal Pose Fusion

Formulating joint location prediction relative to a single local or global location is not always optimal. Existing litera-

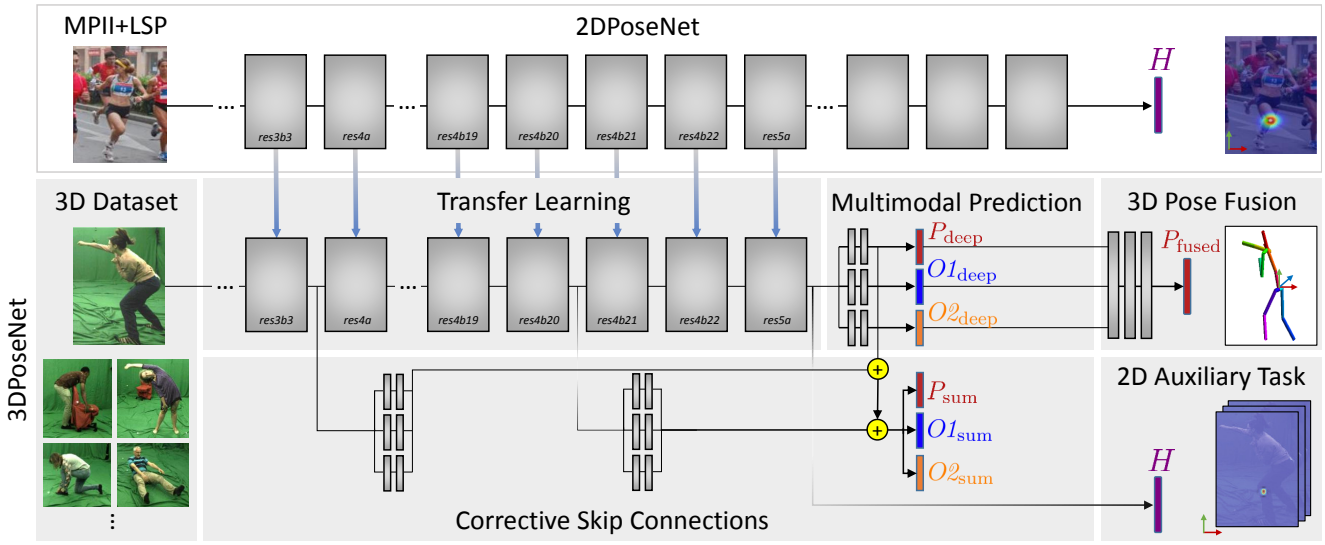

Figure 3. 3D pose Training overview. The main components are 1) regularization through corrective skip connections, and 2D pose prediction as auxiliary task, 2) Multi-modal 3D pose prediction and fusion, 3) a new marker-less 3D pose database with appearance augmentation, and 4) Transfer learning from features learned for 2D pose estimation.

ture [36] has observed that predicting joint locations relative to their direct kinematic parents (Order 1 parents) improves performance. Our experiments reveal that to not universally hold true. We find that depending on the pose and the visibility of the joints in the input image, the optimal relative joint for each joint's location prediction differs. We consider joint locations P relative to the root, $O1$ relative to Order 1 kinematic parents and $O2$ relative to Order 2 kinematic parents along the kinematic tree as the *three modes* of prediction, see Figure 2, and fuse them together.

For the joint set we consider, the kinematic relationships we propose are sufficient, as it puts at least one reference joint for each joint in the relatively low entropy torso [34]. We use three identical 3D prediction stubs attached to *res5a* for predicting the pose as P, $O1$ and $O2$, and for each we use corrective skip connections. These predictions are fed into a smaller network with three fully connected layers, to implicitly determine and fuse the better constraints per joint into the final prediction P_{fused}. The network has the flexibility to emphasize different combinations of constraints depending on the pose. Although this can be viewed as intermediate supervision with auxiliary tasks, the specific architecture used is the key to its efficacy.

3.3. Global Pose Computation

The BB-crop normalizes subject size and position, which frees 3D pose regression from having to localize the person in scale and image space, but loses global pose information. We propose a lightweight and efficient way to reconstruct the global 3D pose $P^{[G]} = (R|T)\,P_{\text{fused}}$ from pelvis-centered pose P_{fused}, the camera intrinsics, and K.

Perspective correction The BB cropping can be interpreted as using a virtual camera, rotated towards the crop center and its field of view covering the crop area. Since the 3DPoseNet only 'sees' the cropped input, its predictions live in this rotated view, leading to a consistent orientation error in P_{fused}. To compensate, we compute rotation R that rotates the virtual camera to the original view.

3D localization We seek the global translation T that aligns P_{fused} and K under perspective projection. We assume weak perspective projection, Π, and solve the linear least squares equation $\sum_i \|K^i - \Pi(T + P^i_{\text{fused}})\|^2$, where index i refers to each joint. This assumption yields global position

$$T = \frac{\sqrt{\sum_i \|P^i_{[xy]} - \bar{P}_{[xy]}\|^2}}{\sqrt{\sum_i \|K^i - \bar{K}\|^2}} \begin{pmatrix} \bar{K}_{[x]} \\ \bar{K}_{[y]} \\ f \end{pmatrix} - \begin{pmatrix} \bar{P}_{[x]} \\ \bar{P}_{[y]} \\ 0 \end{pmatrix}, \quad (1)$$

in terms of distances to the 3D mean $\bar{P}$ and 2D mean $\bar{K}$ over all joints. $P_{[xy]}$ is the x, y part of P_{fused} and single subscripts indicate the respective elements. Please see the supplemental document for the derivation and evaluation.

Our solution can be considered as a generalization of *Procrustes analysis* for projective alignment. It is different to *Perspective-n-Point* 6DOF rigid pose estimation [35], structure-from-motion, and from the convex approach of Zhou *et al*. [77], which require iterative optimization.

4. New Human Pose Dataset (MPI-INF-3DHP)

We propose a new dataset captured in a multi-camera studio with ground truth from commercial marker-less motion capture [66]. No special suits and markers are needed, allowing the capture of motions wearing everyday apparel, including loose clothing. In contrast to existing datasets, we record in green screen studio to allow automatic segmentation and augmentation. We recorded 8 actors (4m+4f),

performing 8 activity sets each, ranging from walking and sitting to complex exercise poses and dynamic actions, covering more pose classes than Human3.6M. Each activity set spans roughtly one minute. Each actor features 2 sets of clothing split across the activity sets. One clothing set is *casual everyday apparel*, and the other is *plain-colored* to allow augmentation.

We cover a wide range of viewpoints, with five cameras mounted at chest height with a roughly 15° elevation variation similar to the camera orientation jitter in other datasets [15]. Another five cameras are mounted higher and angled down 45°, three more have a top down view, and one camera is at knee height angled up. Overall, from all 14 cameras, we capture >1.3M frames, 500k of which are from the five chest high cameras. We make available both true 3D annotations, and a skeleton compatible with the "universal" skeleton of H3.6M

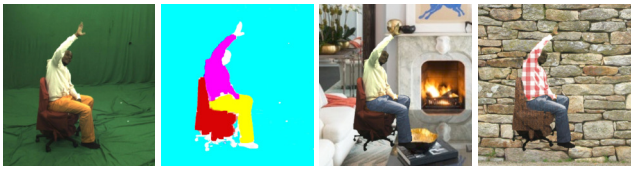

Figure 4. MPI-INF-3DHP dataset. We capture actors using a markerless multi-camera in a green screen studio (left), compute masks for different regions (center left) and augment the captured footage by compositing different textures to the background, chair, upper body and lower body areas, independently (center right and right).

Dataset Augmentation Although our dataset has more clothing variation than other datasets, the appearance variation is still not comparable to in-the-wild images. There have been several approaches proposed to enhance appearance variation. Pishchulin *et al*. warp human size in images with a parametric body model [45]. Images can be used to augment background of recorded footage [48, 15, 28]. Rhodin *et al*. [48] recolor plain-color shirts while keeping the shading details, using intrinsic image decomposition to separate reflectance and shading [40].

We provide chroma-key masks for the background, a chair/sofa in the scene, as well as upper and lower body segmentation for the plain-colored clothing sets. This provides an increased scope for foreground and background augmentation, in contrast to the marker-less recordings of Joo *et al*. [33]. For background augmentation, we use images sampled from the internet. For foreground augmentation, we use a simplified intrinsic decomposition. Since for plain colored clothing the intensity variation is solely due to shading, we use the average pixel intensity as a surrogate for the shading component. We composite cloth like textures with the pixel intensity of the upper body, lower body and chair marks independently, for a photo-realistic result. Figure 4 shows example captured and augmented frames.

Figure 5. Representative frames from MPI-INF-3DHP test set. We cover a variety of subjects with a diverse set of clothing and poses in 3 different settings: studio with green screen (left); studio without green screen (right); and outdoors (center).

Test Set We found the existing test sets for (monocular) 3D pose estimation to be restricted to limited settings due to the difficulty of obtaining ground truth labels in general scenes. *HumanEva* [56] and *Human3.6M* [28] are recorded indoors and test on similar looking scenes as the training set, the Human3D+ [15] test set was recorded with sensor suits that influence appearance and lacks global alignment, and the MARCONI set [17] is markerless through manual annotation, but shows mostly walking motions and multiple actors, which are not supported by most monocular algorithms. We create a new test set with ground truth annotations coming from a multi-view markerless motion capture system. It complements existing test sets with more diverse motions (Standing/Walking, Sitting/Reclining, Exercise, Sports (Dynamic Poses), On The Floor, Dancing/Miscellaneous), camera view-point variation, larger clothing variation (e.g. dress), and outdoor recordings from Robertini *et al*. [50] in unconstrained environments. This makes the test set suitable for testing the generalization of various methods. See Figure 5 for a representative sample. We use the "universal" skeleton for evaluation.

Alternate Metric In addition to the Mean Per Joint Position Error (MPJPE) widely used in 3D pose estimation, we concur with [28] and suggest an extension to 3D of the "Percentage of Correct Keypoints (PCK)" [68, 67] metric used for 2D Pose evaluation, as well as the "Area Under the Curve (AUC)" [26] computed for a range of PCK thresholds. These metrics are more expressive and robust than MPJPE, revealing individual joint mispredictions more strongly. We pick a threshold of 150mm, corresponding to roughly half of head size, similar what is used in MPII 2D Pose dataset. We propose evaluating on the common minimum set of joints across 2D and 3D approaches (Joints 1 to 14 in Figure 2), to ensure evaluation compatibility with existing approaches. Joints are grouped by bilateral symmetry (ankles, wrists, shoulders, etc...), and can be evaluated by scene setting or activity class.

5. Transfer Learning

We use the features learned with Resnet-101 from ImageNet to initialize both *2DPoseNet* and *3DPoseNet*. While this affords a faster convergence while training, there remains room for improved generalization beyond the gains

Table 1. Activity-wise results (MPJPE in mm) on Human3.6m [28]. Models trained on Human3.6m, with network weights initialized from Resnet101 (ImageNet) weights, or transferred from *2DPoseNet*. Evaluation with all 17 joints, on every 64th frame, with no rescaling of predictions to person specific skeleton, using GT Bounding boxes for crops.

	Direct	**Discuss**	**Eating**	**Greet**	**Phone**	**Posing**	**Purch.**	**Sitting**	**Sit Down**	**Smoke**	**Take Photo**	**Wait**	**Walk**	**Walk Dog**	**Walk Pair**	**Total**
Initialized with Resnet-101 (ImageNet) weights																
Base	98.98	100.14	86.07	101.83	101.34	96.74	94.89	125.28	158.31	100.21	112.49	99.57	83.39	109.61	95.79	104.32
Regular Skip	113.34	112.26	97.40	110.50	108.63	112.09	105.67	125.97	173.41	109.34	120.87	107.75	97.30	126.05	117.45	115.29
Corr. Skip	**92.57**	99.08	85.46	**95.43**	96.93	**89.56**	95.67	123.54	160.98	97.13	**107.56**	**93.86**	**76.99**	110.93	**88.73**	101.09
+ Fusion	93.80	99.17	84.73	95.60	94.48	89.40	93.15	119.94	154.61	95.94	106.09	94.13	77.25	108.82	87.38	99.79
Transfer Learning from 2DPoseNet weights																
Corr. + Fusion	59.69	69.74	60.55	68.77	76.36	59.05	75.04	96.19	122.92	70.82	85.42	68.45	54.41	82.03	59.79	74.14

from our architectural and dataset contributions. Due to the similarity of the tasks, features learned for 2D pose estimation on in-the-wild MPII and LSP training sets can be transferred to 3D pose estimation. We explore the consequences of this, thus far, un-utilized method of improving generalization, by transferring weights until level 4 from *2DPoseNet* to *3DPoseNet*.

There is a tradeoff to be made between the transferred features and learning new pertinent features. We achieve this through a learning rate discrepancy between the transferred layers and the new layers. The optimal ratio of learning rates is determined through validation. On *3DPoseNet*, with ImageNet features, the transferred layers' learning rate is scaled down by 10, while when using *2DPoseNet* features, it is scaled down by 1000.

6. Experiments and Evaluation

We evaluate the contributions proposed in the previous sections using the standard datasets *Human3.6M* and *HumanEva*, as well as our new MPI-INF-3DHP Test set. Additionally, we qualitatively observe the performance on LSP [31] and the CMU Panoptic [33] datasets, demonstrating robustness to general scenes. Refer to Figure 7. Also refer to the supplementary video for global 3D pose estimation results.

We evaluate the impact of training *3DPoseNet* on Human3.6m, and unaugmented and augmented variants of MPI-INF-3DHP. We only use Human3.6m compatible views from MPI-INF-3DHP. Further details are in the supplemental document.

6.1. Impact of New Supervision Methods

Multi-level corrective skip connections In Table 1 we compare a baseline method without any skip connections, a network with vanilla skip connections, and our proposed corrective skip regularization on Human3.6m test set. We observe that networks using vanilla skip connections perform markedly worse than the baseline, while corrective skip connections yield to more than 5mm improvement for

Table 2. Evaluation of our design choices on MPI-INF-3DHP test set by scene setting. *GS* indicates green screen background. All with perspective correction, and using ground truth bounding box crops.

3D Data	Network Arch.	Studio GS 3DPCK	Studio no GS 3DPCK	Outdoor 3DPCK	All 3DPCK	All AUC
Human 3.6m	Base	21.1	32.5	10.8	22.6	8.8
	Corr Skip	22.2	33.9	18.5	25.1	8.7
	+Fusion	22.3	34.2	20.0	26.0	9.5
Ours Unaug.	Base	73.6	42.9	19.5	49.0	23.3
	Corr Skip	66.9	38.2	27.9	46.8	20.9
	+Fusion	67.6	39.6	28.5	47.8	21.8
Ours Aug.	Base	77.2	59.5	48.7	63.7	31.1
	Corr. Skip	71.1	51.7	36.1	55.4	26.0
	+Fusion	73.5	53.1	37.9	57.3	28.0

7 classes of activities (marked in bold). The same models are evaluated on MPI-INF-3DHP test-set in Table 2. Additionally, corrective skip yields a significantly improved generalization to in-the-wild scenes, as evidenced by the ≈8% 3DPCK improvement on outdoor sequences.

With unaugmented MPI-INF-3DHP training data, we again see $\approx$ 8% 3DPCK improvement on the outdoor sequences (Table 2) using corrective skip. The regularization interpretation is further supported by the decrease in performance on the studio sequences, which appear similar to training data.

Augmented MPI-INF-3DHP training data has sufficient appearance variation for the the regularization effect of corrective skip to begin hurting performance, and the outdoor 3DPCK drops by $\approx$ 12% over the base network.

Intuitively, corrective skip scheme changes the loss landscape for the core network, allowing it devote resources at training time to correcting larger mispredictions while the skip connections handle the vast numbers of mostly correct predictions which together present a steeper gradient on the loss landscape. This is similar to adaptive re-weighting in AdaBoost [20], emphasizing under-represented and difficult poses. We verified that the effect is not due to a higher

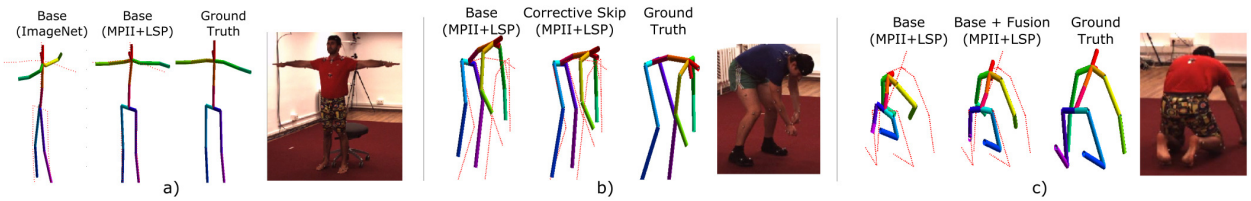

Figure 6. **a)** The better generalizability of features learned from 2D pose datasets allow the model to deal with unseen textures **b)** Multi-level corrective skip regularization is effective even with transfer learning, and allows the network to better tackle difficult poses such as bending and crouching **c)** Fusion of kinematic tree derived constraints helps the representation better handle poses with large self occlusions

Table 3. Evaluation on MPI-INF-3DHP Test set by scene setting. *GS* indicates green screen background. All have weights transferred from 2DPoseNet, have perspective correction applied, and use bounding box annotation, except where it is explicitly mentioned.

3D Dataset	Method	Studio GS 3DPCK	Studio no GS 3DPCK	Outdoor 3DPCK	All 3DPCK	All AUC
Human3.6m	Domain adapt.	44.1	42.6	35.2	41.4	17.7
Human3.6m	Corr. + Fusion	70.8	62.3	58.5	64.7	31.7
Ours Unaug.	Corr + Fusion	84.1	68.9	59.6	72.5	36.9
Ours.	Base + Fusion	82.8	68.0	62.3	72.4	36.9
Aug.	Corr. + Fusion	82.6	66.7	62.0	71.7	36.4
Ours Aug.	Corr. + Fusion	84.6	72.4	69.7	76.5	40.8
+	no persp. corr.	81.9	68.6	67.4	73.5	37.6
Human3.6m	with pred. BB	80.4	71.2	69.8	74.4	39.6

learning rate seen by the core network due to the added loss.

Multimodal prediction and fusion: The multi-modal fusion scheme yields noticeable improvement across all datasets tested in tables 1 and 2. Table 1 shows that the fusion scheme helps poses with large amounts of self occlusion (underlined), such as in Figure 6 c). The improvement is not simply due to additional training, and is less pronounced if predicting P, $O1$ and $O2$ with a single stub, even with more features in the fully connected layer. Refer to the supplementary document for details.

6.2. Benefits of MPI-INF-3DHP

Our dataset, even without augmentation, leads to a ≈9% 3DPCK improvement on outdoor scenes. However, our augmentation strategy is crucial for significantly improved generalization, as seen from the gains in 3DPCK across scene settings in Table 2, giving 63.7% 3DPCK overall.

6.3. Transfer Learning

Despite the improved generalization brought about by our dataset and supervision techniques (Table 2) for 3D pose estimation, it doesn't approach the level of performance seen for 2D pose estimation methods. Our *2DPoseNet* achieves 91.2 PCK and 66.3 AUC on the LSP test set, and 89.7 PCK and 61.3 AUC on MPII Single Person test set.

Our approach of transferring representations from *2DPoseNet* to *3DPoseNet* yields 64.7% 3DPCK on MPI-INF-3DHP test-set when trained with only Human3.6m data, compared to 63.7% 3DPCK of the model trained on our augmented training set without transfer learning. Combining our dataset and transfer learning leads to even better results at ≈ 72.5% 3DPCK, and further adding Human3.6m data leads to the best performing method with 76.5% 3DPCK. See Table 3.

The effect of the proposed supervision techniques is more subtle when using transfer learning. Refer to the supplementary document for details. Figure 6 b) shows a representative example of the nature of improvement. Multi-modal fusion helps with poses with a large degree of self-occlusion, as represented in 6 c).

In contrast to existing approaches countering data scarcity, transfer learning does not require complex dataset synthesis, yet exceeds the performance of Chen *et al*. [15] (with synthetic data and domain adaptation, 28.8% 3DPCK, after procrustes alignment) and our base model trained with the synthetic data of Rogez *et al*. [51] (21.7% 3DPCK). Our approach also performs better than simply doing domain adaptation [22] to in-the-wild data (Table 3). Refer to the supplementary for details.

6.4. Relevance of Other Method Components

Bounding box Computation On MPI-INF-3DHP testset, we additionally evaluate our best performing network using bounding boxes computed from *2DPoseNet*. The performance drops to 74.4%3DPCK from 76.5%3DPCK due to the additional difficulty (Table 3).

Perspective Correction Perspective correction also has a significant impact, without which, the performance drops to 73% 3DPCK from 76.5% (Table 3).

6.5. Quantitative Comparison to the State of the Art

Human3.6M Table 2 shows comparison of our method with existing methods, all trained on Human3.6m. Altogether, with our supervision contributions and transfer learning, we advance the state of the art on *Human3.6M* by more than 25% in terms of MPJPE (from 107.2mm of Zhou *et al*. [78] to 74.11mm). The improvement is not due to a deeper network, since our baseline network is on par with Zhou *et al*. Complementing *Human3.6M* with our aug-

Table 4. Comparison of results on Human3.6m [28] with the state of the art. Human3.6m, Subjects 1,5,6,7,8 used for training, and 9,11 used for testing. ^{S} = Scaled to test subject specific skeleton, computed from T-pose. ^{T}= Uses Temporal Information, $^{J14/J17}$ = Joint set evaluated, ^{A} = Uses Best Alignment To GT per frame, ^{Act} = Activitywise Training, $^{1/10/64}$ = Test Set Frame Sampling

Method	Total MPJPE (mm)
Deep Kinematic Pose[78] $^{\mathbf{J17},B}$	107.26
Sparse. Deep. [80] $^{T,\mathbf{J17},B,10,Act}$	113.01
Motion Comp. Seq. [65] $^{T,\mathbf{J17},B}$	124.97
LinKDE [28] $^{\mathbf{J17},B,Act}$	162.14
Du et al. [75] $^{T,\mathbf{J17},B}$	126.47
Rogez et al. [51] $^{(\mathbf{J13}),B,64}$	121.20
SMPLify [10] $^{\mathbf{J14},B,A,(First\ cam.)}$	82.3
Ours (with 2DPoseNet Transfer)	
Corr. Skip + Fusion $^{\mathbf{J17},B}$	74.11
Corr. Skip + Fusion $^{\mathbf{J17},B,\mathbf{S}}$	68.61
Corr. Skip + Fusion $^{\mathbf{J14},B,\mathbf{A}}$	54.59

mented MPI-INF-3DHP dataset further reduces the error to 72mm.

HumanEva The improvements on Human3.6m are confirmed with a 30.8 and 33.5 MPJPE score on the S1 Box and Walk sequences of HumanEva, after alignment. See supplemental document. **MPI-INF-3DHP** We also evaluated some of the existing best performing methods on our dataset. Deep Kinematic Pose [78], the state-of-the-art on Human3.6m, attains 13.8% 3DPCK overall. Our full model, without transfer learning, trained on Human3.6m attains 26% 3DPCK, and 64.7% 3DPCK with transfer learning.

7. Discussion

Our fully feed-forward regression-based method improves over the accuracy of the current state-of-the-art regression-based and model-based monocular 3D pose estimation methods by more than 25%. Our new method compensates perspective distortions due to cropping, and is the first to compute full global 3D pose in non-cropped images in closed form. Nonetheless, it has limitations. Estimating 3D pose from camera views starkly different from chest height positions is still a challenge for all methods. Partly this is because most training sets, also [15], have a strong bias towards chest height cameras. Our new dataset provides diverse view-points, which can support development towards viewpoint invariance in future methods. Similar to related approaches, our per-frame estimation exhibits some temporal noise on video sequences. In future, we will investigate integration with model-based temporal tracking to further increase accuracy and temporal smoothness.

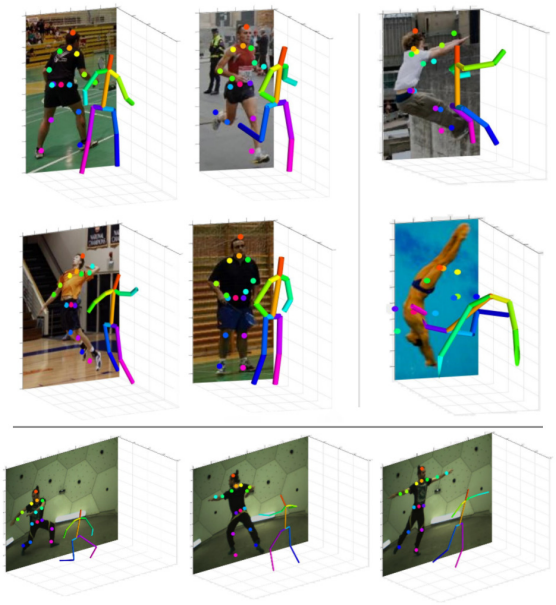

Figure 7. Qualitative evaluation on representative frames of the LSP test set. We succeed in challenging cases (left), with only few failure cases (right). The *Dance1* sequence of the *PanopticDataset* [33], is also well reconstructed (bottom).

Though, at less than 1 s per frame, our approach is much faster than model based methods which work offline in the order of minutes, there is scope for future improvement towards real-time.

We also show that joining forces with transfer learning, in conjunction with algorithmic and data contributions, will aide progress in 3D pose estimation in many different directions, such as overall accuracy and generalizability.

8. Conclusion

We have presented a fully feedforward CNN-based approach for monocular 3D human pose estimation that significantly outperforms state-of-the-art regression-based and model-based methods on established benchmarks [28, 56]. It uses enhanced CNN supervision techniques, re-purposing multi-level skip connections as a regularization structure, and using improved parent relationships in the kinematic chain. This, combined with a new dataset that includes a larger variety of real human appearances, activities and camera views, with improved augmentation potential, leads to significantly improved generalization to in-the-wild images. We show that using the easily accessible method of transferring representations learned from in-the-wild 2D pose data in tandem with our architectural and data contributions, helps us generalize and perform better than any of the existing approaches. Our method is also the first to efficiently extract global 3D position in non-cropped images, without brittle and time consuming iterative optimization.

Supplemental Document: Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision

This document accompanies the main paper, and the supplemental video.

1. Impact of Multi-level Corrective Skip Connections and Multi-modal Fusion With Transfer Learning

We discussed the improvements obtained with multi-level corrective skip regularization and multi-modal pose fusion schemes in Section 6.1. We further elaborate on the impact of these supervision techniques when used in conjunction with transfer learning from *2DPoseNet*. Using Corrective Skip connections for training significantly improves the predicted joint positions in complex poses, such as those covered under Crouching. Additionally, multi-modal fusion scheme leads to further improvements on top. This is depicted in Figure 1, showing the 3DPCK curves for a selection of joints on activities involving significant Crouching and Sitting on Human3.6m.

We also show that usual Skip Connections (used at training and test time) perform in the same ballpark as the core network without any skip connections. Also refer to Table 1 for activity-wise breakdown of the various design choices on Human3.6m, as well as Table 2 for per-activity comparison of our approach against other methods. In Figure 1, we see that elbow and wrist improve significantly, for both Crouching and Sitting activities, while the knee sees only minor improvement.

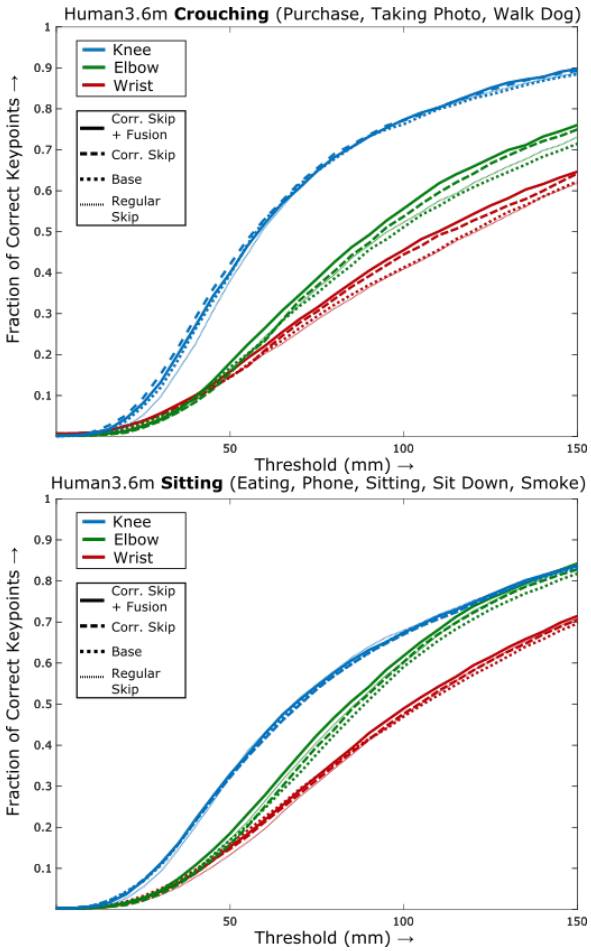

Figure 1. PCK improvement with our proposed supervision schemes, visualized for *Crouching* and *Sitting* activity groups on Human3.6m S9,11, for Knees, Elbows and Wrists. The network is fine-tuned on Human3.6m data, starting from 2DPoseNet weights.

2. Further Discussion of Design Choices Regarding Multi-modal Fusion

To demonstrate that the improvement seen due to the fusion scheme is not simply a result of finetuning, we compare the result of fusion with components successively removed. Using P, $O1$ and $O2$, we get an MPJPE of 74.49mm. On removing $O2$, the error increases to 74.77mm, and on removing both $O1$ and $O2$, the error increases to 75.27mm. The comparison here is without any multi-level corrective skip training.

For P, $O1$ and $O2$ to have different modes of mispredictions, the underlying feature set that they are computed from has to be as different as possible, because each is related to the other with a linear transform. We achieve some degree of decorrelation between the three by using 3 different prediction stubs, one each for P, $O1$ and $O2$ with a convolutional layer ($k_{5\times5}$, s_2) with 128 features followed by a fully-connected layer. If we replace the three stubs with a single stub with the convolutional layer having 256 features followed by a fully-connected layer, we get an MPJPE of 75.30mm after fusion, vs an MPJPE of 74.49mm from fusing the result of 3 prediction stubs. Both of these are without corrective-skip connections.

3. Further Discussion of Multi-level Corrective Skip

Since the proposed multi-level corrective skip scheme adds an additional loss at the last stage (X_{deep}, where X is P/$O1$/$O2$) of the network, it increases the effective learning rate seen by the core network. To verify that the improvements seen due to the proposed scheme are not caused by this difference in the effective learning rate, we trained a version of the *Base* network with loss weights as the sum of the loss weights for X_{deep} and X_{sum} specified in Table 6. We find that this network performs worse than the *Base* network (107.14mm vs 104.32mm MPJPE on Human3.6m), and does not approach the accuracy attained with multi-

Table 1. Activitywise results (MPJPE in mm) on Human3.6m [28], Subjects 9 and 11, with no rescaling of 'universal' skeleton to person specific skeleton. Network weights are initialized with the **weights from 2DPoseNet**, and evaluation is done with all 17 joints on every 64th frame, using GT Bounding boxes for crops.

	Direct	Discuss	Eating	Greet	Phone	Posing	Purch.	Sitting	Sit Down	Smoke	Take Photo	Wait	Walk	Walk Dog	Walk Pair	Total
Human3.6m																
Base	59.07	71.36	63.22	70.11	78.44	57.63	78.81	98.85	124.03	71.35	87.47	68.67	54.17	86.39	60.54	75.52
+ Fusion	58.56	70.06	62.62	69.68	77.47	57.01	76.83	97.46	121.86	70.23	86.20	68.46	53.93	84.20	60.06	74.49
1.5pt 1.5pt Regular Skip	60.57	73.23	62.24	71.31	78.25	60.44	77.79	97.90	124.20	71.09	87.89	69.91	56.56	87.19	62.87	76.21
Corr. Skip	60.09	70.06	60.76	69.39	77.19	59.07	75.49	96.62	122.72	71.26	86.02	69.11	55.16	83.11	60.77	74.65
+ Fusion	**59.69**	**69.74**	**60.55**	**68.77**	**76.36**	**59.05**	**75.04**	**96.19**	**122.92**	**70.82**	**85.42**	**68.45**	**54.41**	**82.03**	**59.79**	**74.14**
Our Aug. + Human3.6m																
Corr. Skip + Fusion	57.51	68.59	59.57	67.34	78.06	56.86	69.13	97.99	117.54	69.45	82.40	67.96	55.25	76.50	61.40	72.89

Table 2. Comparison of results on Human3.6m [28] with the state of the art. Human3.6m, Subjects 1,5,6,7,8 used for training. Subjects 9 and 11, all cameras used for testing. ^{S} = Scaled to test subject specific skeleton, computed using T-pose. ^{T}= Uses Temporal Information, ^{B}= Uses GT Bounding Box, $^{J14/J17}$ = Joint set evaluated, ^{A} = Uses Best Alignment To GT per frame, ^{Act} = Activitywise Training, $^{1/10/64}$ = Test Set Frame Sampling

	Direct	Discuss	Eat	Greet	Phone	Posing	Purchase	Sitting
Tekin et al[64]$^{\mathbf{J17},B}$	85.03	108.71	84.38	98.94	119.39	98.49	93.77	73.76
Deep Kine. Pose[78]$^{\mathbf{J17},B}$	91.83	102.41	96.95	98.75	113.35	90.04	93.84	132.16
Sparse. Deep. [80]$^{T,\mathbf{J17},B,10,Act}$	87.36	109.31	87.05	103.16	116.18	106.88	99.78	124.52
Motion Comp. [65]$^{T,\mathbf{J17},B}$	132.71	158.52	87.95	126.83	118.37	114.69	107.61	136.15
Tekin et al [63]$^{\mathbf{J17},B,Act}$	-	129.06	91.43	121.68	-	-	-	-
LinKDE [28]$^{\mathbf{J17},B,Act}$	132.71	183.55	132.37	164.39	162.12	150.61	171.31	151.57
Ours (with 2DPoseNet Transfer)								
Corr. Skip + Fusion $^{\mathbf{J17},B}$	59.72	69.48	60.89	68.66	76.56	58.88	78.65	90.86
Corr. Skip + Fusion $^{\mathbf{J17},B,S}$	52.55	63.85	55.44	62.27	71.80	52.62	72.22	86.22

	Sit Down	Smoke	Photo	Wait	Walk	Walk Dog	Walk Pair	Average
Tekin et al[64]$^{\mathbf{J17},B}$	170.40	85.08	95.65	116.91	62.08	113.72	94.83	100.08
Deep Kine. Pose[78]$^{\mathbf{J17},B}$	158.97	106.91	125.22	94.41	79.02	126.04	98.96	107.26
Sparse. Deep. [80]$^{T,\mathbf{J17},B,10,Act}$	199.23	107.42	143.32	118.09	79.39	114.23	97.70	113.01
Motion Comp. [65]$^{T,\mathbf{J17},B}$	205.65	118.21	185.02	146.66	65.86	128.11	77.21	125.28
Tekin et al [63]$^{\mathbf{J17},B,Act}$	-	-	162.17	-	65.75	130.53	-	-
LinKDE [28]$^{\mathbf{J17},B,Act}$	243.03	162.14	205.94	170.69	96.60	177.13	127.88	162.14
Ours (with 2DPoseNet transfer)								
Corr. Skip + Fusion $^{\mathbf{J17},B}$	125.17	71.15	85.69	68.85	54.01	82.63	60.01	74.11
Corr. Skip + Fusion $^{\mathbf{J17},B,S}$	120.64	66.03	79.84	63.97	48.92	76.77	53.69	**68.61**

level corrective skip scheme (101.09mm).

4. Global Pose Computation

4.1. 3D localization

In this section we describe a simple, very efficient method to compute the global 3D location T of a noisy 3D point set P with unknown global position, but known scaling and orientation, from its 2D projection estimate K in a camera with known intrinsics parameters (focal length f). We further assume that the point cloud spread in depth direction is negligible compared to its distance z_0 to the camera and approximate perspective projection of an object near position $(x_0, y_0, z_0)^\top$ with weak perspective projection (linearizing the pinhole projection model at z_0):

$$\begin{pmatrix} u \\ v \end{pmatrix} = \mathbf{\Pi} \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \text{ with } \mathbf{\Pi} = \begin{pmatrix} \frac{f}{z_0} & 0 & 0 \\ 0 & \frac{f}{z_0} & 0 \end{pmatrix}. \quad (1)$$

Estimates K and P are assumed to be noisy due to estimation errors. We find the optimal global position T in the least squares sense, by minimizing $T =$

$\arg\min_{(x,y,z)} E(x,y,z)$, with

$$E = \sum_i \|K^i - \Pi\left((x,y,z)^\top + P^i\right)\|^2$$
$$= \sum_i \|K^i - \frac{f}{z}\left((x,y)^\top + P^i_{[xy]}\right)\|^2, \quad (2)$$

where P^i and K^i denote the ith joint position in 2D and 3D, respectively, and $P^i_{[xy]}$ the xy component of P^i. It has partial derivative

$$\frac{\partial E}{\partial x} = \frac{2f}{z}\sum_i K^i_{[x]} + \frac{f}{z}\left(P^i_{[x]} - x\right), \quad (3)$$

where $P_{[x]}$ denotes the x part of P, and $\bar{P}$ the mean of P over all joints. Solving $\frac{\partial E}{\partial x} = 0$ gives the unique closed-form solutions $x = \bar{K}_{[x]}\frac{z}{f} - \bar{P}_{[x]}$ and equivalently $y = \bar{K}_{[y]}\frac{z}{f} - \bar{P}_{[y]}$, for $\frac{\partial E}{\partial y} = 0$.

Substitution of x and y in E and differentiating with respect to z yields

$$\frac{\partial E}{\partial z} = \frac{f\sum_i (K^i - \bar{K})^\top (P^i_{[xy]} - \bar{P}_{[xy]})}{z^2} + \frac{f^2 \sum_i \|P^i_{[xy]} - \bar{P}_{[xy]}\|^2}{z^3}. \quad (4)$$

Finally, solving $\frac{\partial E}{\partial z} = 0$ gives the depth estimate

$$z = f\frac{\sum_i \|P^i_{[xy]} - \bar{P}_{[xy]}\|^2}{\sum_i (K^i - \bar{K})^\top (P^i_{[xy]} - \bar{P}_{[xy]})} \approx f\frac{\sqrt{\sum_i \|P^i_{[xy]} - \bar{P}_{[xy]}\|^2}}{\sqrt{\sum_i \|K^i - \bar{K}\|^2}}, \quad (5)$$

where $(K^i - \bar{K})(P^i - \bar{P}) = \|K^i - \bar{K}\|\|P^i - \bar{P}\|\cos(\theta)$ is approximated for $\theta \approx 0$. This is a valid assumption in our case, since the rotation of 3D and 2D pose is assumed to be matching.

Evaluation on HumanEva: In addition to evaluating centered pose P, we evaluate the global 3D pose prediction $P^{[G]}$ on the widely used *HumanEva* motion capture dataset — *Box* and *Walk* sequences of *Subject 1* from the validation set. Note that we do not use any data from HumanEva for training. We significantly improve the state of the art for the Box sequence (82.1mm [10] vs 58.6mm). Results on the Walk sequence are of higher accuracy than Bogo *et al.* [10], but lower than the accuracy of Bo *et al.* [9] and Yasin *et al.* [73], who, however train on HumanEva [9] or use an example database dominated by walking motions [73]. Our skeletal structure does not match that of HumanEva, e.g. the head prediction has a consistent frontal offset and the hip is too wide. To compensate, we compute a linear map of dimension 14x14 (number of joints) that maps our joint positions as a linear combination to the HumanEva structure. The same mapping is applied at every frame, but is computed only once, jointly on the Box and Walk sequence, to limit the correction to global inconsistencies of the skeleton structure. This fine-tuned result is marked by $\sim$ in Table 3.

Table 3. Quantitative evaluation on *HumanEva-I* [56], on three metrics. For reference, we also show multi-view existing results. Our models use no data from HumanEva for training, while the other methods listed train/finetune on HumanEva-I. * = Does not use GT Bounding Box information. † = Translation alignment only. $\sim$ = trained on HumanEva-I.

		S1 Box			S1 Walk		
		$P^{[G]}$ (global)	P (align^{S,T})	P (align^{R,S,T})	$P^{[G]}$ (global)	P (align^{S,T})	P (align^{R,S,T})
	Our full model*	**129.5**	**69.4**	**58.6**	**145.6**	**79.2**	**67.1**
	w/o Persp. correct.*	133.9	79.4	58.6	147.7	83.6	67.2
Monocular	Bo *et al.* [9]*	-	-	-	-	54.8†	-
Monocular	Yasin *et al.* [73]*$\sim$	-	-	-	52.2	-	-
Monocular	Bogo *et al.* [10]$\sim$	-	-	82.1	-	-	73.3
Monocular	Akhter *et al.* [2]	-	-	165.5	-	-	186.1
Monocular	Ramakris. *et al.* [47]	-	-	151.0	-	-	161.8
Monocular	Zhou *et al.* [77]	-	112.5	-	-	100.0	-
Multi-view	Amin *et al.* [3]	**47.7**	-	-	**54.5**	-	-
Multi-view	Rhodin *et al.* [49]	59.7	-	-	74.9	-	-
Multi-view	Elhayek *et al.* [18]	60.0	-	-	66.5	-	-

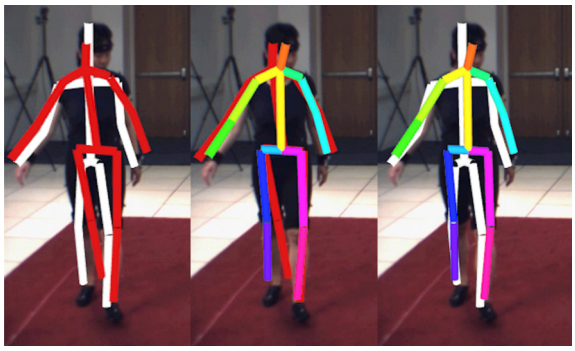

Figure 2. The predicted pose (red) is inaccurate for positions away from the camera center (left), compared against the ground truth (white). Perspective correction (colored) corrects the orientation (center) and is closer to the ground truth (right). Here tested on the walking sequence of HumanEva S1.

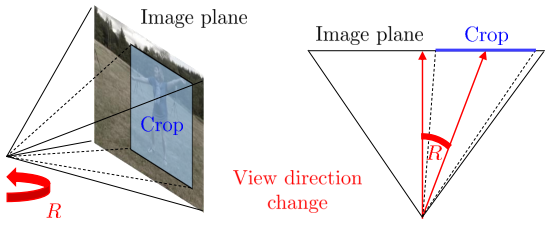

Figure 3. Sketch of the input image cropping and resulting change of field of view. The corresponding rotation R of the view direction is sketched in 2D on the right.

Table 4. Results of our 2DPoseNet on MPII Single Person Pose [4] dataset and LSP [31] 2D Pose datasets. * = Trained/Finetuned only on the corresponding training set

	MPII		LSP	
	$PCK_{h0.5}$	AUC	$PCK_{0.2}$	AUC
Our 2DPoseNet				
w Person Locali.	89.7	61.3	**91.2**	65.3
w/o Person Locali.	89.6	61.5	**91.2**	65.5
Stacked Hourgl.[42]	**90.9***	**62.9***	-	-
Bulat et al.[12]	89.7*	59.6*	90.7	-
Wei et al.[71]	88.5	61.4	90.5	65.4
DeeperCut [26]	88.5	60.8	90.1	**66.1**
Gkioxary et al [23]	86.1*	57.3*	-	-
Lifshitz et al. [38]	85.0	56.8	84.2	-
Belagiannis et al.[8]	83.9*	55.5*	85.1	-
DeepCut[44]	82.4	56.5	87.1	63.5
Hu&Ramanan [25]	82.4*	51.1*	-	-
Carreira et al. [13]	81.3*	49.1*	72.5*	-

4.2. Perspective correction

The 3DPoseNet predicts pose P in the coordinate system of the bounding box crop, which leads to inaccuracies, see Figure 2. The cropped image appears if as it was taken from a virtual camera with the same origin as the original camera, but with view direction to the crop center, see Figure 3. To map the reconstruction from the virtual camera coordinates to the original camera, we rotate P by the rotation R between the virtual and original camera. Since the existing training sets provide chest-height camera placements with the same viewpoint, the bias in vertical direction is already learned by the network. We apply perspective correction only in horizontal direction, where a change in cropping and yaw rotation of the person cannot be distinguished by the network. R is then the rotation around the camera up direction by the angle between the original and the virtual view direction, see Figure 3. On our MPI-INF-3DHPtest set perspective correction improves the PCK by 3 percent points. On HumanEva the improvement is up to 3 mm MPJPE, see Table 3. Using the vector from the camera origin to the centroid of 2D keypoints K as the virtual view direction was most accurate in our experiments. However, the crop center can be used instead. Opposed to the Perspective-n-Point algorithm applied by Zhou [80], any regression method that works on cropped images could immediately profit from this perspective correction, without computing 2D keypoint detections.

5. CNN Architecture and Training Specifics

5.1. 2DPoseNet

Architecture: The architecture derives from Resnet-101. It uses Resnet-101 structure as is till level 4. Since we are interested in predicting heatmaps, we remove striding at level 5. The number of features in the *res5a* module are halved. Identity skip connections are removed from *res5b* and *res5c*, and the number of features gradually tapered to 15 (heatmaps for 14 joints + root). There are two versions of 2DPoseNet: with and without additional person localization input. The former gets the person localization heatmap spliced in at *res3b3*. The latter is used for transferring weights to 3DPoseNet. Their performance on the 2D Pose benchmarks is almost identical. For 2DPoseNet, our results on MPII and LSP test sets approach that of the state of the art. See Table 4.

Table 5. Loss weight and learning rate, LR, taper scheme used for 2DPoseNet. 2DPoseNet also employes Multi-level Corrective Skip connections, and the heatmap H_{sum} is the sum of H_{deep} and the skip connections. Heatmaps H_{4b20} and H_{5a} are used for intermediate supervision.

Base LR	# Iter	Loss Weights ($\mathbf{w} \times L(H_{xx})$)			
		H_{sum}	H_{deep}	H_{4b20}	H_{5a}
0.050	60k	1.0	0.5	0.5	0.5
0.010	60k	1.0	0.4	0.1	0.1
0.005	60k	1.0	0.2	0.05	0.05
0.001	60k	1.0	0.2	0.05	0.05
6.6e-4	60k	1.0	0.1	0.005	0.005
0.0001	40k	1.0	0.01	0.001	0.001
2.5e-5	40k	1.0	0.001	0.0001	0.0001
0.0008	60k	1.0	0.0001	0.0001	0.0001
0.0001	40k	1.0	0.0001	0.0001	0.0001
3.3e-5	20k	1.0	0.0001	0.0001	0.0001

Intermediate Supervision: Additionally, we employ intermediate supervision at *res4b20* and *res5a*, treating the first 15 feature maps of the layers as the intermediate joint-location heatmaps. Further, we use a Multi-level Corrective Skip scheme, with skip connections coming from *res3b3* and *res4b22* through prediction stubs comprised of a 1×1 convolution with 20 feature maps followed by a 3×3 convolution with 15 outputs.

Training: For training, we use the Caffe [30] framework, with the AdaDelta solver with a momentum of 0.9 and weight decay rate of 0.005. We employ a batch size of 7, and use Euclidean Loss everywhere. For the Learning Rate and Loss Weight taper schema, refer to Table 5.

5.2. 3DPoseNet

Architecture: The core network is identical to *2DPoseNet* up to *res5a*. A 3D Prediction stub is attached on top, comprised of a 5×5 convolution layer with a stride of 2 and 128 features, followed by a fully-connected layer.

Multi-level Corrective Skip: We attach 3D prediction stubs to *res3b3* and *res4b20*, similar to the final prediction stub, but with 96 convolutional features instead of 128. The resulting predictions are added to P_{deep} to get P_{sum}. We add

Table 6. Loss weight and LR taper scheme used for 3DPoseNet. There is a difference in the number of iterations used when training with Human3.6m or MPI-INF-3DHP alone, v.s. when training with the two in conjunction. Part Labels PL are used only when training with H3.6m solely. Multi-level skip connections add up with X_{deep} to yield X_{sum}, where X is P or $O1$ $O2$.

Base LR	H3.6m/Our Batch = 5 #Epochs	H3.6m+Our Batch = 6 #Epochs	Loss Weights (w × L(A_{bb})) $X = P/O1/O2$ X_{4b5}	X_{4b20}	X_{deep}	X_{sum}	H	PL*
0.05	3 (45k)	2.4 (60k)	50	50	50	100	0.1	0.05
0.01	1 (15k)	1.2 (30k)	10	10	10	100	0.05	0.025
0.005	2 (30k)	1.2 (30k)	5	5	5	100	0.01	0.005
0.001	1 (15k)	0.6 (15k)	1	1	1	100	0.01	0.005
5e-4	2 (30k)	1.2 (30k)	0.5	0.5	0.5	100	0.005	0.001
1e-4	1 (15k)	0.6 (15k)	0.1	0.1	0.1	100	0.005	0.001

Table 7. Loss weight and LR taper scheme used for finetuning 3DPoseNet for Multi-modal Fusion scheme.

Base LR	H3.6m/Our Batch = 5 #Epochs	H3.6m+Our Batch = 6 #Epochs	Loss Weights (w × L(A_{bb})) P_{fused}
0.05	(1k)	(2k)	100
0.01	1 (15k)	0.8 (20k)	100
0.005	1 (15k)	0.8 (20k)	100
0.001	1 (15k)	0.8 (20k)	100

a loss term to P_{deep} in addition to the loss term at P_{sum}.

Multi-modal Fusion: We add prediction stubs for $O1$ and $O2$, similar to those for P. Note that the predictions for P, $O1$ and $O2$ are done with distinct stubs, and this slight decorrelation of predictions is important. These predictions are at a later finetuning step fed into three fully-connected layers, with 2k, 1k and 51 nodes respectively.

Intermediate Supervision: We use intermediate supervision at *4b5* and *res4b20*, using prediction stubs comprised of 7×7 convolution with a stride of 3 and 128 features, followed by a fully-connected layer predicting P, $O1$ and $O2$ as a single vector. Additionally, we predict joint location heatmaps and part-label maps using a 1×1 convolution layer after *res5a* as an auxiliary task. We don't use the part-label maps when training with MPI-INF-3DHP dataset.

Training: For training, the solver settings are similar to *2DPoseNet*, and we use Euclidean Loss everywhere. For transfer learning, we scale down the learning rate of the transferred layers by a factor determined by validation. For fine-tuning in the multi-modal fusion case, we similarly downscale the learning rate of the trained network by 10000 with respect to the three new fully-connected layers. For the learning rate and loss weight taper schema for both the main training and multi-modal fusion fine-tuning stages, refer to Tables 6 and 7. We use different training durations when using Human3.6M or MPI-INF-3DHP in isolation, v.s. when using both in conjunction. This is reflected in the aforementioned tables.

5.2.1 3D Pose Training Data

In the various experiments on *3DPoseNet*, for the datasets we consider, we select ≈37.5k frames for each, yielding ≈75k samples after scale augmentation at 2 scales (0.7 and 1.0). **Human3.6m** We use the H80k [27] subset of Human3.6m, and train with the "universal" skeleton, using S1,5,6,7,8 for training and S9,11 for testing. The predicted skeleton is not scaled to the test subject skeletons at test time.

MPI-INF-3DHP For our dataset, to maintain compatibility of view with Human3.6m and other datasets, we only pick the 5 chest high cameras for all 8 subjects, sampling frames such that at least one joint has moved by more than 200mm between selected frames. A random subset of these frames is used for training, to match the number of selected Human3.6m frames.

MPI-INF-3DHP Augmented The augmented version uses the same frames as the unaugmented MPI-INF-3DHP above, keeping ≈25% frames unaugmented, ≈40% with only BG and Chair augmentation, and the rest with full augmentation.

5.2.2 Domain Adaptation To In The Wild 2D Pose Data

We use a domain adaptation stub comprised of $conv_{3\times3,256}$, $conv_{3\times3,128}$, fc_{64} and fc_1 layers, and cross entropy domain classification loss. It uses Ganin et al's [22] gradient inversion approach. The domain adaptation stub is attached after $res4b22$ in the network. We found that directly starting out with $\lambda = -1$ performs better than gradually increasing the magnitude of λ with increasing iterations. We train on the Human3.6m training set, with 2D heatmap and part label prediction as auxiliary tasks. Images from MPII [4] and LSP [31, 32] training sets are used without annotations for learning better generalizable features. The generalizability is improved, as evidenced by the 41.4 3DPCK on MPI-INF-3DHPtest-set, but does not match up with the 64.7 3DPCK attained using transfer learning. Detailed results in main Table 3.

6. MPI-INF-3DHP Dataset

We cover a wide range of poses in our training and test sets, roughly grouped into various activity classes. A detailed description of the dataset is available in Section 4 of the main paper. In addition, Figure 4 samples the various different activity classes, augmentation and subjects represented in our dataset.

Similarly for the test set, we show a sample of the activities and the variety of subjects in Figure 5.

Figure 4. A sample of the activities, clothing, subjects as well as augmentation on MPI-INF-3DHP Trainig Set.

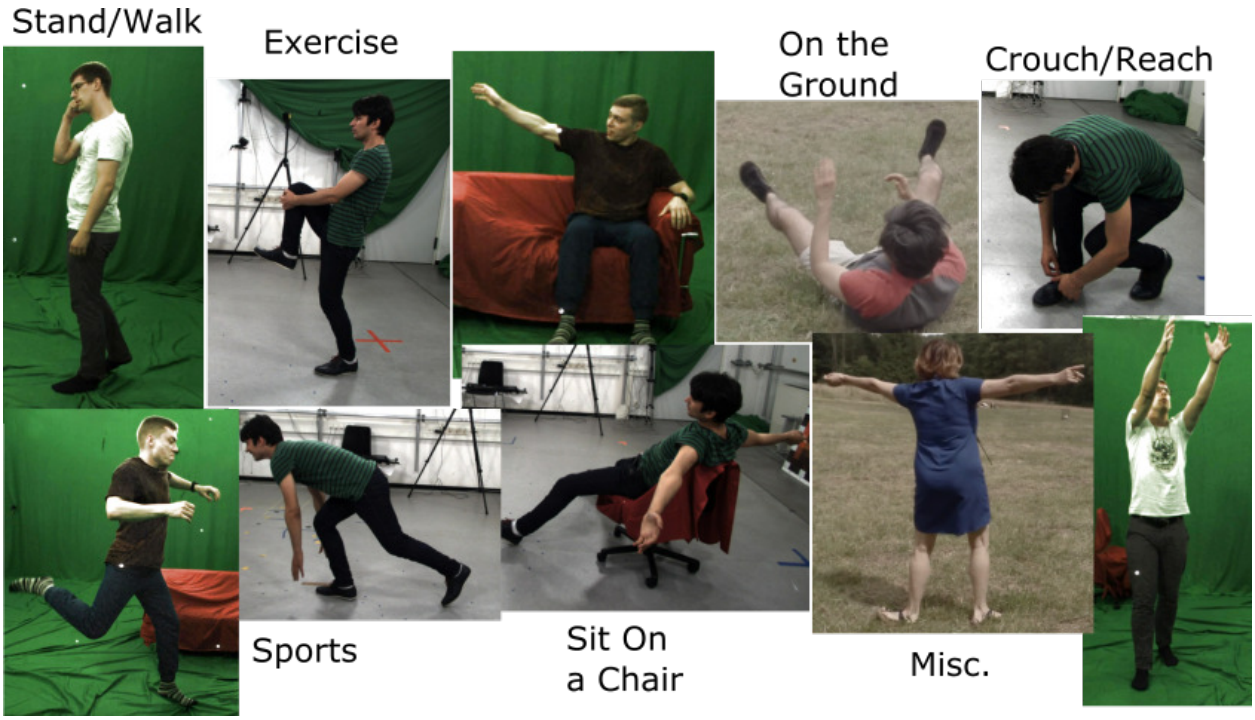

Figure 5. A sample of the activities and subjects in the test set of MPI-INF-3DHP

6.1. The Challenge of Learning Invariance to Viewpoint Elevation

In this paper, we only consider the cameras in the training set placed at chest-height, in part to be compatible with the existing datasets, and in part because viewpoint elevation invariance is a significantly more challenging problem. The existing benchmarks don't place emphasis on this. We will release an expanded version of our MPI-INF-3DHP test set with multiple camera viewpoint elevations, to complement the training data.

References

[1] A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28(1):44–58, 2006. 1, 2

[2] I. Akhter and M. J. Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1446–1455, 2015. 2, 11

[3] S. Amin, M. Andriluka, M. Rohrbach, and B. Schiele. Multi-view pictorial structures for 3D human pose estimation. In *BMVC*, 2013. 11

[4] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 1, 2, 3, 12, 13

[5] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1014–1021, 2009. 1

[6] A. Baak, M. Müller, G. Bharaj, H.-P. Seidel, and C. Theobalt. A data-driven approach for real-time full body pose reconstruction from a depth camera. In *IEEE International Conference on Computer Vision (ICCV)*, 2011. 1

[7] A. O. Balan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker. Detailed human shape and pose from images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007. 1

[8] V. Belagiannis and A. Zisserman. Recurrent human pose estimation. *arXiv preprint arXiv:1605.02914*, 2016. 1, 12

[9] L. Bo and C. Sminchisescu. Twin gaussian processes for structured prediction. In *International Journal of Computer Vision*, 2010. 11

[10] 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 *European Conference on Computer Vision (ECCV)*, 2016. 1, 2, 8, 11

[11] E. Brau and H. Jiang. 3D Human Pose Estimation via Deep Learning from 2D Annotations. In *International Conference on 3D Vision (3DV)*, 2016. 2

[12] A. Bulat and G. Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *European Conference on Computer Vision (ECCV)*, 2016. 1, 12

[13] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 12

[14] J. Chai and J. K. Hodgins. Performance animation from low-dimensional control signals. *ACM Transactions on Graphics (TOG)*, 24(3):686–696, 2005. 1

[15] 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 *International Conference on 3D Vision (3DV)*, 2016. 2, 5, 7, 8

[16] X. Chen and A. L. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1736–1744, 2014. 1

[17] A. Elhayek, E. Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt. MARCOnI - ConvNet-based MARker-less Motion Capture in Outdoor and Indoor Scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2016. 2, 5

[18] A. Elhayek, E. de Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt. Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3810–3818, 2015. 11

[19] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision (IJCV)*, 61(1):55–79, 2005. 1, 2

[20] Y. Freund, R. Schapire, and N. Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999. 6

[21] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel. Optimization and filtering for human motion capture. *International Journal of Computer Vision (IJCV)*, 87(1–2):75–92, 2010. 1

[22] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1180–1189, 2015. 7, 13

[23] G. Gkioxari, A. Toshev, and N. Jaitly. Chained predictions using convolutional neural networks. In *European Conference on Computer Vision (ECCV)*, 2016. 1, 12

[24] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[25] P. Hu, D. Ramanan, J. Jia, S. Wu, X. Wang, L. Cai, and J. Tang. Bottom-up and top-down reasoning with hierarchical rectified gaussians. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 12

[26] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision (ECCV)*, 2016. 1, 3, 5, 12

[27] C. Ionescu, J. Carreira, and C. Sminchisescu. Iterated second-order label sensitive pooling for 3d human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1661–1668, 2014. 2, 13

[28] 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 Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36(7):1325–1339, 2014. 1, 5, 6, 8, 10

[29] A. Jain, J. Tompson, Y. LeCun, and C. Bregler. Modeep: A deep learning framework using motion features for human pose estimation. In *Asian Conference on Computer Vision (ACCV)*, pages 302–315. Springer, 2014. 2

[30] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 675–678, 2014. 12

[31] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *British Machine Vision Conference (BMVC)*, 2010. doi:10.5244/C.24.12. 1, 3, 6, 12, 13

[32] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 2, 3, 13

[33] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, pages 3334–3342, 2015. 1, 5, 6, 8

[34] A. M. Lehrmann, P. V. Gehler, and S. Nowozin. A Non-parametric Bayesian Network Prior of Human Pose. In *IEEE International Conference on Computer Vision (ICCV)*, 2013. 4

[35] V. Lepetit and P. Fua. *Monocular model-based 3D tracking of rigid objects*. Now Publishers Inc, 2005. 4

[36] S. Li and A. B. Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *Asian Conference on Computer Vision (ACCV)*, pages 332–347, 2014. 1, 2, 3

[37] S. Li, W. Zhang, and A. B. Chan. Maximum-margin structured learning with deep networks for 3d human pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2848–2856, 2015. 1, 2

[38] I. Lifshitz, E. Fetaya, and S. Ullman. Human pose estimation using deep consensus voting. In *European Conference on Computer Vision (ECCV)*, 2016. 1, 12

[39] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 3

[40] A. Meka, M. Zollhöfer, C. Richardt, and C. Theobalt. Live intrinsic video. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 35(4):109:1–14, 2016. 5

[41] G. Mori and J. Malik. Recovering 3d human body configurations using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 28(7):1052–1062, 2006. 2

[42] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision (ECCV)*, 2016. 1, 2, 12

[43] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010. 3

[44] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 12

[45] L. Pishchulin, A. Jain, M. Andriluka, T. Thormählen, and B. Schiele. Articulated people detection and pose estimation: Reshaping the future. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3178–3185. IEEE, 2012. 5

[46] G. Pons-Moll, D. J. Fleet, and B. Rosenhahn. Posebits for monocular human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2337–2344, 2014. 2

[47] V. Ramakrishna, T. Kanade, and Y. Sheikh. Reconstructing 3d human pose from 2d image landmarks. In *European Conference on Computer Vision*, pages 573–586. Springer, 2012. 11

[48] H. Rhodin, C. Richardt, D. Casas, E. Insafutdinov, M. Shafiei, H.-P. Seidel, B. Schiele, and C. Theobalt. EgoCap: Egocentric Marker-less Motion Capture with Two Fisheye Cameras. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 2016. 5

[49] H. Rhodin, N. Robertini, D. Casas, C. Richardt, H.-P. Seidel, and C. Theobalt. General automatic human shape and motion capture using volumetric contour cues. In *European Conference on Computer Vision (ECCV)*, pages 509–526. Springer, 2016. 2, 11

[50] N. Robertini, D. Casas, H. Rhodin, H.-P. Seidel, and C. Theobalt. Model-based Outdoor Performance Capture. In *International Conference on Computer Vision (3DV)*, 2016. 5

[51] G. Rogez and C. Schmid. Mocap-guided data augmentation for 3d pose estimation in the wild. In *Advances in Neural Information Processing Systems*, pages 3108–3116, 2016. 2, 7, 8

[52] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 3

[53] B. Sapp and B. Taskar. Modec: Multimodal decomposable models for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 1

[54] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. In *Conference on Computer Vision and Pattern Recognition Workshops*, pages 806–813, 2014. 3

[55] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013. 1

[56] 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. *International Journal of Computer Vision (IJCV)*, 87(1-2):4–27, 2010. 1, 5, 8, 11

[57] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3634–3641, 2013. 1, 2

[58] E. Simo-Serra, A. Ramisa, G. Alenyà, C. Torras, and F. Moreno-Noguer. Single image 3d human pose estimation from noisy observations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2673–2680. IEEE, 2012. 1, 2

[59] C. Sminchisescu and B. Triggs. Covariance scaled sampling for monocular 3d body tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–447. IEEE, 2001. 1

[60] J. Starck and A. Hilton. Model-based multiple view reconstruction of people. In *IEEE International Conference on Computer Vision (ICCV)*, pages 915–922, 2003. 1

[61] C. Stoll, N. Hasler, J. Gall, H.-P. Seidel, and C. Theobalt. Fast articulated motion tracking using a sums of Gaussians body model. In *IEEE International Conference on Computer Vision (ICCV)*, pages 951–958, 2011. 1

[62] C. J. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 677–684, 2000. 2

[63] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua. Structured Prediction of 3D Human Pose with Deep Neural Networks. In *British Machine Vision Conference (BMVC)*, 2016. 1, 2, 10

[64] B. Tekin, P. Márquez-Neila, M. Salzmann, and P. Fua. Fusing 2D Uncertainty and 3D Cues for Monocular Body Pose Estimation. *arXiv preprint arXiv:1611.05708*, 2016. 2, 3, 10

[65] B. Tekin, A. Rozantsev, V. Lepetit, and P. Fua. Direct Prediction of 3D Body Poses from Motion Compensated Sequences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 8, 10

[66] The Captury. http://www.thecaptury.com/, 2016. 4

[67] 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 *Advances in Neural Information Processing Systems (NIPS)*, pages 1799–1807, 2014. 1, 5

[68] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1653–1660, 2014. 1, 5

[69] R. Urtasun, D. J. Fleet, and P. Fua. Monocular 3d tracking of the golf swing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 932–938, 2005. 1

[70] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao. Robust estimation of 3d human poses from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2361–2368, 2014. 1, 2

[71] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional Pose Machines. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 3, 12

[72] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland. Pfinder: real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 19(7):780–785, 1997. 1

[73] H. Yasin, U. Iqbal, B. Krüger, A. Weber, and J. Gall. A Dual-Source Approach for 3D Pose Estimation from a Single Image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 11

[74] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems (NIPS)*, pages 3320–3328, 2014. 3

[75] Y. Yu, F. Yonghao, Z. Yilin, and W. Mohan. Marker-less 3D Human Motion Capture with Monocular Image Sequence and Height-Maps. In *European Conference on Computer Vision (ECCV)*, 2016. 1, 2, 8

[76] F. Zhou and F. De la Torre. Spatio-temporal matching for human detection in video. In *European Conference on Computer Vision (ECCV)*, pages 62–77, 2014. 1

[77] X. Zhou, S. Leonardos, X. Hu, and K. Daniilidis. 3D shape estimation from 2D landmarks: A convex relaxation approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4447–4455, 2015. 1, 2, 4, 11

[78] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei. Deep kinematic pose regression. In *ECCV Workshop on Geometry Meets Deep Learning*, 2016. 1, 2, 7, 8, 10

[79] X. Zhou, M. Zhu, S. Leonardos, and K. Daniilidis. Sparse representation for 3d shape estimation: A convex relaxation approach. *arXiv preprint arXiv:1509.04309*, 2015. 2

[80] X. Zhou, M. Zhu, S. Leonardos, K. Derpanis, and K. Daniilidis. Sparseness Meets Deepness: 3D Human Pose Estimation from Monocular Video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2, 8, 10, 12