Single-Shot Multi-Person 3D Pose Estimation From Monocular RGB

Dushyant Mehta[1,2], Oleksandr Sotnychenko[1,2], Franziska Mueller[1,2],
Weipeng Xu[1,2], Srinath Sridhar[3], Gerard Pons-Moll[1,2], Christian Theobalt[1,2]

[1] MPI For Informatics   [2] Saarland Informatics Campus   [3] Stanford University

Abstract

We propose a new single-shot method for multi-person 3D pose estimation in general scenes from a monocular RGB camera. Our approach uses novel occlusion-robust pose-maps (ORPM) which enable full body pose inference even under strong partial occlusions by other people and objects in the scene. ORPM outputs a fixed number of maps which encode the 3D joint locations of all people in the scene. Body part associations [8] allow us to infer 3D pose for an arbitrary number of people without explicit bounding box prediction. To train our approach we introduce MuCo-3DHP, *the first large scale training data set showing real images of sophisticated multi-person interactions and occlusions. We synthesize a large corpus of multi-person images by compositing images of individual people (with ground truth from mutli-view performance capture). We evaluate our method on our new challenging 3D annotated multi-person test set* MuPoTs-3D *where we achieve state-of-the-art performance. To further stimulate research in multi-person 3D pose estimation, we will make our new datasets, and associated code publicly available for research purposes.*

1. Introduction

Single-person pose estimation, both 2D and 3D, from monocular RGB input is a challenging and widely studied problem in vision [4, 3, 33, 34, 7, 11, 28, 37]. It has many applications, *e.g.*, in activity recognition and content creation for graphics. While methods for 2D multi-person pose estimation exist [43, 17, 8, 37], most 3D pose estimation methods are restricted to a single un-occluded subject. Natural human activities take place with multiple people in cluttered scenes hence exhibiting not only self-occlusions of the body, but also strong inter-person occlusions or occlusions by objects.

This makes the under-constrained problem of inferring 3D pose (of all subjects) from monocular RGB input even harder and leads to drastic failure of existing single person 3D pose estimation methods.

Recent work approaches this more general 3D multi-person pose estimation problem by decomposing it into multiple single-person instances [31, 49], often with significant redundancy in the decomposition [49]. The single person predictions are post-processed to filter, refine and fuse the predictions into a coherent estimate. Bottom-up joint multi-person reasoning remains largely unsolved, and the multi-person 3D pose estimation lacks appropriate performance benchmarks.

We propose a new *single shot* CNN-based method to estimate multi-person 3D pose in general scenes from monocular input. We call our method single shot since it reasons about all people in a scene jointly in a single forward pass, and does not require explicit bounding box proposals by a separate algorithm as a pre-processing step [49, 40]. The latter may fail under strong occlusions and may be expensive to compute in dense multi-person scenes. Our fully-convolutional method jointly infers 2D and 3D joint locations using our new occlusion-robust pose-map (ORPM) formulation. ORPM enables multi-person 3D pose estimates under strong (self-)occlusions by incorporating redundancy in the encoding, while using a fixed number of outputs regardless of the number of people in the scene. Our subsequent hierarchical read-out strategy starts with a base pose estimate, and is able to refine the estimate based on which joints of a person are visible, leading to robust 3D pose results.

To train our CNN we introduce a new multi-person 3D pose data set *MuCo-3DHP*. While there are several single-person datasets with 3D joint annotations, there are no annotated multi-person datasets containing large corpora of real video recordings of human–human interaction with large person and background diversity. Important advances in this direction have been made by Joo *et al.* [24] using a multi-camera studio setup but background diversity remains limited.
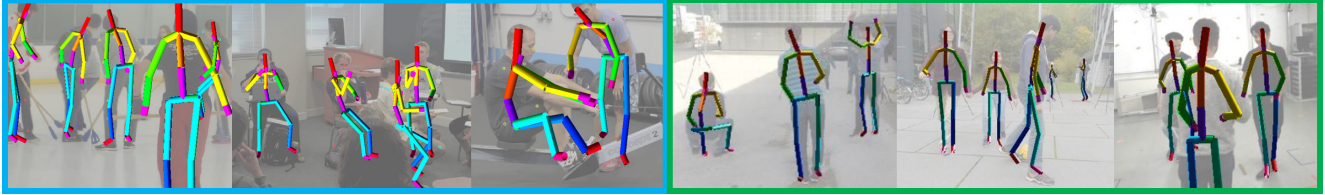
Figure 1. Qualitative results of our approach shown on MPII 2D [3] dataset (blue), as well as our new *MuPoTS-3D* evaluation set (green). Our pose estimation approach works for general scenes, handling occlusions by objects or other people. Note that our 3D pose predictions are root-relative, and scaled and overlaid only for visualization.

Some prior work creates 3D annotated multi-person images by using 2D pose data augmented with 3D poses from motion capture datasets [49], or by finding 3D consistency in 2D part annotations from multi-view images recorded in a studio [53]. To create training data with much larger diversity in person appearance, camera view, occlusion and background, we transform the MPI-INF-3DHP single-person dataset [33] into the first multi-person set that shows images of real people in complex scenes. *MuCo-3DHP* is created by compositing multiple 2D person images with ground-truth 3D pose from multi-view marker-less motion capture. Background augmentation and shading-aware foreground augmentation of person appearance enable further data diversity. To validate the generalizability of our approach to real scenes, and since there are only very few annotated multi-person test sets [12] showing more than two people, we contribute a new multi-person 3D test set, *MuPoTS-3D*. It features indoor and outdoor scenes, challenging occlusions and inter-actions, varying backgrounds, more than two persons, and ground truth from commercial marker-less motion capture. All datasets will be made publicly available. In summary, we contribute:

- A CNN-based single-shot *multi-person pose estimation method* based on a novel multi-person 3D pose-map formulation to jointly predict 2D and 3D joint locations of all persons in the scene. Our method is tailored for scenes with occlusion by objects or other people.
- The first *multi-person dataset* of real person images with 3D ground truth that contains complex inter-person occlusions, motion, and background diversity.
- A real *in-the-wild test set* for evaluating multi-person 3D pose estimation methods that contains diverse scenes, challenging multi-person interactions, occlusions, and motion.

Our method achieves state-of-the-art performance on challenging multi-person scenes where single-person methods completely fail. Although designed for the much harder multi-person task, it performs competitively on single-person test data.

2. Related Work

We focus on most directly related work estimating the pose of *multiple people in 2D* or a *single person in 3D* from monocular RGB input. [50] provide a more comprehensive review.

Multi-Person 2D Pose Estimation: A common approach for multi-person 2D pose estimation is to first detect single persons and then 2D pose [44, 13, 55, 19, 40]. Unfortunately, these methods fail when the detectors fail—a likely scenario with multiple persons and strong occlusions. Hence, a body of work first localizes the joints of each person with CNN-based detectors and then find the correct association between joints and subjects in a post-processing step [43, 16, 36, 8].

Single-Person 3D Pose Estimation: Existing monocular single-person 3D pose methods show good performance on standard datasets [18, 51, 59, 41, 28, 27]. However, since many methods train a discriminative predictor for 3D poses [5], they often do not generalize well to natural scenes with varied poses, appearances, backgrounds and occlusions. This is due to the fact that most 3D datasets are restricted to indoor setups with limited backgrounds and appearance. The advent of large real world image datasets with 2D annotations made 2D pose estimation in the wild remarkably accurate. However, annotating images with 3D pose is much harder with many recent work focusing on leveraging 2D image datasets for 3D human pose estimation or multi-view settings [24]. Additional annotations to these 2D image datasets allow some degree of 3D reasoning, either through body joint depth ordering constraints [42] or dense shape correspondences [14]. Some works split the problem in two: first estimate 2D joints and then lift them to 3D [58, 54, 9, 66, 35, 69, 1, 52, 20, 32, 6, 26, 38, 57, 2], *e.g.*, by database matching, neural network regression, or fitting the SMPL body model [30]. Some works integrate SMPL within the CNN to exploit 3D and 2D annotations in an end-to-end fashion [39, 25, 42].

Other work leverages the features learned by a 2D pose estimation CNN for 3D pose estimation. For example, [60] learn to merge features from a 2D and 3D joint prediction network. Another approach is to train

Figure 2. Examples from our **MuCo-3DHP** dataset, created through compositing MPI-INF-3DHP [33] data. (Top) composited examples without appearance augmentation, (bottom) with BG and clothing augmentation. The last two columns show rotation and scale augmentation, and truncation with the frame boundary.

a network with separate 2D and 3D losses for the different data sources [46, 68, 56, 65, 31]. Some approaches jointly reason about 2D and 3D pose with multi-stage belief maps [62]. The advantage of such methods is that they can be trained end to end. A simpler yet very effective approach is to refine a network trained for 2D pose estimation for the task of 3D pose estimation [34, 33]. A major limitation of methods that rely on 2D joint detections directly or on bounding boxes is that they easily fail under body occlusion or with incorrect 2D detections, both of which are common in multi-person scenes. In contrast, our approach is more robust to occlusions since a base 3D body pose estimate is available even under significant occlusion. [34] showed that 3D joint prediction works best when the receptive field is centered around the joint of interest. We build upon this insight to refine the base body pose where 2D joint detections are available.

Multi-Person 3D Pose Estimation: To our knowledge, only [49] tackle multi-person 3D pose estimation from single images.[1] They first identify bounding boxes likely to contain a person using [47]. Instead of a direct regression to pose, the bounding boxes are classified into a set of K-poses similar to [45]. These poses are scored by a classifier and refined using a regressor. The method implicitly reasons using bounding boxes and produces multiple proposals per subject that need to be accumulated and fused. However, performance of their method under large person-person occlusions is unclear. In contrast, our approach produces multi-person 2D joint locations and 3D pose maps in a single shot, from which the 3D pose can be inferred even under severe person-person occlusion.

3D Pose Datasets: Existing pose datasets are either for a single person in 3D [18, 51, 63, 64, 33] or multi-person with only 2D pose annotations [3, 29]. Of

the two exceptions, the MARCOnI dataset [12] features 5 sequences but contains only 2 persons simultaneously, and there are no close interactions. The other is the Panoptic dataset [24] which has a limited capture volume, pose and background diversity. There is work on generating synthetic images [48, 10] from mocap data, however the resulting images are not plausible. We choose to leverage the person segmentation masks available in MPI-INF-3DHP [33] to generate annotated multi-person 3D pose images of real people at scale through compositing using the available segmentation masks. Furthermore, we captured a 3D benchmark dataset featuring multiple closely interacting persons which was annotated by a video-based multi-camera motion capture system.

3. Multi-Person Dataset

Generating data by combining in-the-wild multi-person 2D pose data [3, 29] and multi-person multi-view motion capture for 3D annotation would be a straightforward extension of previous (single-person) approaches [33]. However, multi-person 3D motion capture under strong occlusions and interactions is challenging even for commercial systems, often requiring manual pose correction constraining 3D accuracy. Hence, we merely employ purely multi-view marker-less motion capture to create the 20 sequences of *MuPoTs-3D*, the first expressive in-the-wild multi-person 3D pose benchmark. For the much larger training set *MuCo-3DHP*, we resort to a new compositing and augmentation scheme that leverages the single-person image data of real people in MPI-INF-3DHP[33] to composite an arbitrary number of multi-person interaction images under user control, with 3D pose annotations.

3.1. MuCo-3DHP: Compositing-Based Training Set

The MPI-INF-3DHP [33] single-person 3D pose dataset provides marker-less motion capture based annotations for real images of 8 subjects, each captured with 2 clothing sets, using 14 cameras at different elevations. We build upon these person segmentation masks to create per-camera composites with 1 to 4 subjects, with frames randomly selected from the 8×2 sequences available per camera. Since we have ground-truth 3D skeleton pose for each video subject in the same space, we can composite in a 3D-aware manner resulting in correct depth ordering and overlap of subjects. We refer to this composited training set as the **Mu**ltiperson **Co**mposited **3D H**uman **P**ose dataset (see Fig. 2 for examples). The compositing process results in plausible images covering a range of simulated inter-person overlap and activity scenarios. Furthermore, user-control over the desired pose and occlu-

---

[1]A second approach [67] was published in the review period.

Figure 3. Examples from our *MuPoTS-3D* evaluation set. Ground truth 3D pose reference and joint occlusion annotations are available for up to 3 subjects in the scene. The set covers a variety of scene settings, activities and clothing.

sion distribution, and foreground/background augmentation using the masks provided with MPI-INF-3DHP is possible (see supplementary document for more details). Even though the synthesized composites may not simulate all the nuances of human-human interaction fully, we observe that our approach trained on this data generalizes well to real world scenes in the test set.

3.2. MuPoTS-3D: Diverse Multi-Person 3D Test Set

We also present a new filmed (not composited) multi-person test set comprising 20 general real world scenes with ground-truth 3D pose for up to three subjects obtained with a multi-view marker-less motion capture system [61]. Additionally, per joint occlusion annotations are available. The set covers 5 indoor and 15 outdoor settings, with trees, office buildings, road, people, vehicles, and other stationary and moving distractors in the background. Some of the outdoor footage also has challenging elements like drastic illumination changes, and lens flare. The indoor sequences use 2048×2048px footage at 30fps, and outdoor sequences use 1920×1080px GoPro footage at 60fps. The test set consists of >8000 frames, split among the 20 sequences, with 8 subjects, in a variety of clothing styles, poses, interactions, and activities. Notably, the test sequences do not resemble the training data, and include real interaction scenarios. We call our new test set **Mu**ltiperson **Po**se **T**est **S**et in **3D** (*MuPoTS-3D*).

Evaluation Metric: We use the robust *3DPCK* evaluation metric proposed in [33]. It treats a joint's prediction as correct if it lies within a 15cm ball centered at the ground-truth joint location, and is evaluated for the common minimum set of 14 joints marked in green in Fig. 5. We report the *3DPCK* numbers per sequence, averaged over the subjects for which GT reference is available, and additionally report the performance breakdown for occluded and un-occluded joints.

The relative robustness of 3DPCK over MPJPE[18] is also useful to offset the effect of jitter that arises in all non-synthetic annotations, including ours. For completeness, we also report the MPJPE error for predictions matched to an annotated subject.

4. Method

At the core of our approach is a novel formulation which allows us to estimate the pose of multiple people in a scene even under strong occlusions with a single forward pass of a fully convolutional network. Our method builds upon the *location-maps* formulation [34] that links 3D pose inference more strongly to image evidence by inferring 3D joint positions at the respective 2D joint pixel locations.We first recap the location-map formulation before describing our approach.

Location-Maps [34]: A location-map is a joint specific feature channel storing the 3D coordinate x, y, or z at the joint 2D pixel location. For every joint, three location-maps, as well as a 2D pixel location heatmap are estimated. The latter encodes the 2D pixel location of the joint as a confidence map in the image plane. The 3D position of a joint can be read out from its location-map at the 2D pixel location of the joint, as shown in Fig. 4. For an image of size $W \times H$, $3n$ location-maps of size $W/k \times H/k$ are used to store the 3D location of all n joints, where k is a down-sampling factor. During training, the L_2 loss between the ground truth and the estimated location-map is minimized in the area around the joint's 2D pixel location. Although this simple location-map formulation enables full 3D pose inference, it has several shortcomings. First, it assumes that all joints of a person are fully visible, and breaks down under partial occlusion, which is common in general scenes. Second, efficient extension to multiple people is not straightforward. Introducing separate location-maps per person requires dynamically changing the number of outputs.

Occlusion-Robust Pose-Maps (ORPMs): We propose a novel occlusion-robust formulation that has a fixed number of outputs regardless of the number of people in the scene, while enabling pose read-outs for strongly occluded people. Our key insight is the incorporation of *redundancy* into the location-maps. We represent the body by decomposing it into torso, four limbs, and head (see Fig. 5). Our occlusion-robust pose-maps (ORPMs) support multiple levels of redundancy: (1) they allow the read-out of the complete *base pose* $\mathbf{P} \in \mathbb{R}^{3 \times n}$ at one of the torso joint locations (neck or pelvis), (2) the base pose (which may not capture the full extent of articulation) can be further refined by reading out the head and individual limb poses where 2D detections are available, and (3) the complete limb

pose can be read out at any 2D joint location of that limb. Together, these ensure that a complete and as articulate as possible pose estimate is available even in the presence of heavy occlusions of the body (see Fig. 5, and Fig. 2,3 in the supplementary document). In addition, the redundancy in ORPMs allows to encode the pose of multiple partially overlapping persons without loss of information, thus removing the need for a variable number of output channels. See Fig. 4.

Naïve Redundancy: The naïve approach to introduce redundancy by allowing full pose read-out at all body joint locations breaks down for interacting and overlapping people, leading to supervision and *read-out* conflicts in all location-map channels. Our selective introduction of redundancy restricts these conflicts to pose-map channels of similar limbs, *i.e.*, wrist of one person in the proximity of a knee of another person cannot cause read-out conflicts because their pose is encoded in their respective pose-maps. If the complete pose was encoded at each joint location, there would be conflicts for each pair of proximate joints across people. We now formally define ORPMs (Sec. 4.1) and explain the inference process (Sec. 4.2).

4.1. Formulation

Given a monocular RGB image $\mathcal{I}$ of size $W \times H$, we seek to estimate the 3D pose $\mathcal{P} = \{\mathbf{P}_i\}_{i=1}^m$ for each of the m persons in the image. Here, $\mathbf{P}_i \in \mathbb{R}^{3 \times n}$ describes the 3D locations of the $n = 17$ body joints of person i. The body joint locations are expressed relative to the parent joints as indicated in Fig. 5 and converted to pelvis-relative locations for evaluation. We first decompose the body into pelvis, neck, head, and a set of limbs: $L = \{\{\text{shoulder}_s, \text{elbow}_s, \text{wrist}_s\}, \{\text{hip}_s, \text{knee}_s, \text{ankle}_s\} \mid s \in \{\text{right}, \text{left}\}\}$. The 3D locations of the joints are then encoded in the occlusion-robust pose-maps denoted by $\mathcal{M} = \{\mathbf{M}_j\}_{j=1}^n$, where $\mathbf{M}_j \in \mathbb{R}^{W \times H \times 3}$. In contrast to simple location-maps, the ORPM $\mathbf{M}_j$ stores the 3D location of joint j not only at this joint's 2D pixel location $(u,v)_j$ but at a set of 2D locations $\rho(j) = \{(u,v)_{\text{neck}}, (u,v)_{\text{pelvis}}\} \cup \{(u,v)_k\}_{k \in limb(j)}$, where:

$$limb(j) = \begin{cases} l, & \text{if } \exists l \in L \text{ with } j \in l \\ \{\text{head}\}, & \text{if } j = \text{head} \\ \emptyset, & \text{otherwise} \end{cases} . \quad (1)$$

Note that—since joint j of all persons i is encoded in $\mathbf{M}_j$—it can happen that read-out locations coincide for different people, *i.e.*, $\rho_{i1}(j) \cap \rho_{i2}(j) \neq \emptyset$. In this case, $\mathbf{M}_j$ contains information about the person closer to the camera at the overlapping locations. However, due to our built-in redundancy in the ORPMs, a pose
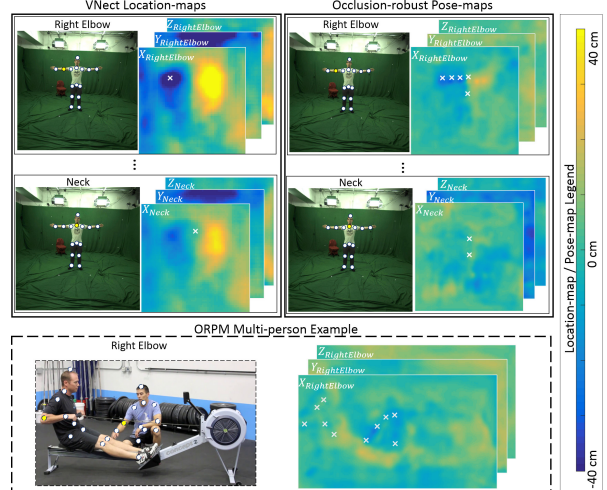

Figure 4. Multiple levels of selective redundancy in our Occlusion-robust Pose-map (ORPM) formulation. VNect Location-maps [34] (left) only support readout at a single pixel location per joint type. ORPMs (middle) allow the complete body pose to be read out at torso joint pixel locations (neck, pelvis). Further, each individual limb's pose can be read out at all 2D joint pixel locations of the respective limb. This translates to read-out of each joint's location being possible at multiple pixel locations in the joint's location map. The example at the bottom shows how 3D locations of multiple people are encoded into the same map per joint and no additional channels are required.

estimate for the partially occluded person can still be obtained at other available read-out locations.

To estimate where the pose-maps can be read out, we make use of 2D joint *heatmaps* $\mathcal{H} = \{\mathbf{H}_j \in \mathbb{R}^{W \times H}\}_{j=1}^n$ predicted by our network. Additionally, we estimate *part affinity fields* $\mathcal{A} = \{\mathbf{A}_j \in \mathbb{R}^{W \times H \times 2}\}_{j=1}^n$ which represent a 2D vector field pointing from a joint of type j to its parent [8]. This facilitates association of 2D detections in the heatmaps (and hence read-out locations for the ORPMs) to person identities and enables per-person read-outs when multiple people are present. Note that we predict a fixed number of maps (n heatmaps, $3n$ pose-maps, and $2n$ part affinity fields) in a single forward pass irrespective of the number of persons in the scene, jointly encoding the 2D and 3D pose for all subjects, *i.e.*, our network is *single-shot*.

4.2. Pose Inference

Read-out of 3D pose of multiple people from ORPMs starts with inference of 2D joint locations $\mathcal{P}^{2D} = \{\mathbf{P^{2D}}_i\}_{i=1}^m$ with $\mathbf{P^{2D}}_i = \{(u,v)_j^i\}_{j=1}^n$ and joint detection confidences $\mathcal{C}^{2D} = \{\mathbf{C^{2D}}_i \in \mathbb{R}^n\}_{i=1}^m$ for each person i in the image. Explicit 2D joint-to-person association is done with the predicted heatmaps $\mathcal{H}$ and part affinity fields $\mathcal{A}$ using the approach of Cao *et al.* [8].

| a. Un-occluded | b. Occluded extremity | c. Occluded limb | d. Occluded extremity | e. Proximity conflict |

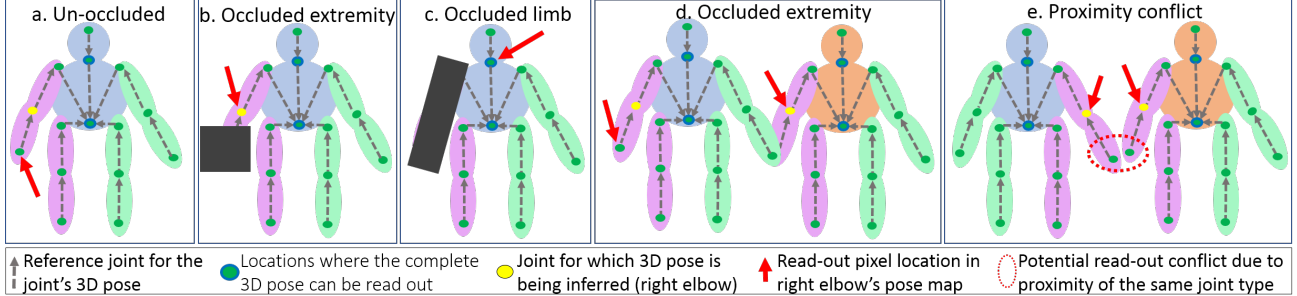| ↕ Reference joint for the joint's 3D pose | ● Locations where the complete 3D pose can be read out | ● Joint for which 3D pose is being inferred (right elbow) | ▲ Read-out pixel location in right elbow's pose map | ⬭ Potential read-out conflict due to proximity of the same joint type |

Figure 5. Example of the choice of read-out pixel location for right elbow pose under various scenarios. First the complete body pose is read out at one of the torso locations. a.) If the limb extremity is un-occluded, the pose for the entire limb is read out at the extremity (wrist), b.) If the limb extremity is occluded, the pose for the limb is read out at the joint location further up in the joint hierarchy (elbow), c.) If the entire limb is occluded, we retain the base pose read out at one of the torso locations (neck), d.) Read-out locations indicated for inter-person interaction, e.) If two joints of the same type (right wrist here) overlap or are in close proximity, limb pose read-out is done at a safer isolated joint further up in the hierarchy.

Next, we use the 2D joint locations $\mathcal{P}^{2D}$ and the joint detection confidences $\mathcal{C}^{2D}$ in conjunction with ORPMs $\mathcal{M}$ to infer the 3D pose of all persons in the scene.

Read-Out Process: By virtue of the ORPMs we can read out 3D joint locations at select multiple pixel locations as described above. We define *extremity joints*: the wrists, the ankles, and the head. The neck and pelvis 2D detections are usually reliable, these joints are most often not occluded and lie in the middle of the body. Therefore, we start reading the full base pose at the neck location. If the neck is *invalid* (as defined below) then the full pose is read at the pelvis instead. If both of these joints are invalid, we consider this person as not visible in the scene and we do not predict the person's pose. While robust, full poses read at the pelvis and neck tend to be closer to the average pose in the training data. Hence, for each limb, we continue by reading out the limb pose at the extremity joint. Note again that the *complete limb* pose can be accessed at any of that limb's 2D joint locations. If the extremity joint is valid, the limb pose replaces the corresponding elements of the base pose. If the extremity joint is invalid however, we walk up the kinematic chain and check the other joints of this limb for validity. If all joints of the limb are invalid, the base pose cannot be further refined. This read-out procedure is illustrated in Fig. 5 and algorithmically described in the supplementary document.

2D Joint Validation: We declare a 2D joint location $\mathbf{P^{2D}}_i^j = (u,v)_j^i$ of person i as *valid read-out location* iff (1) it is un-occluded, *i.e.*, has confidence value higher than a threshold t_C, and (2) it is sufficiently far ($\geq t_D$) away from all read-out locations of joint j of other individuals:

$$valid(\mathbf{P^{2D}}_i^j) \Leftrightarrow \mathbf{C^{2D}}_i^j > t_C \ \wedge \ ||a - \mathbf{P^{2D}}_i^j||_2 \geq t_D$$
$$\forall \bar{i} = [1{:}m], \ \bar{i} \neq i. \ \forall a \in \rho_{\bar{i}}(j). \qquad (2)$$

Our ORPM formulation together with the occlusion-aware inference strategy with limb refinement enables us to obtain accurate poses even for strongly occluded body parts while exploiting all available information if individual limbs are visible. We validate our performance on occluded joints and the importance of limb refinement on our new test set (see Sec. 5).

4.3. Network and Training Details

Our network has ResNet-50 [15] as the core, after which we split it into two—a *2DPose+Affinity* stream and a *3DPose* stream. The core network and the first branch are trained on MS-COCO [29] and the second branch is trained with MPI-INF-3DHP or MuCo-3DHP as per the scenario. Training and architectural specifics are in the supplementary document.

The *2DPose+Affinity* stream predicts the 2D heatmaps $\mathcal{H}_{COCO}$ for the MS-COCO body joint set, and part affinity fields $\mathcal{A}_{COCO}$. The *3DPose* stream predicts 3D ORPMs $\mathcal{M}_{MPI}$ as well as 2D heatmaps $\mathcal{H}_{MPI}$ for the MPI-INF-3DHP [33] joint set, which has some overlap with the MS-COCO joint set. For pose read-out locations as described previously, we restrict ourselves to the common minimum joint set between the two, indicated by the circles in Fig. 5.

Loss: The 2D heatmaps $\mathcal{H}_{COCO}$ and $\mathcal{H}_{MPI}$ are trained with per-pixel $L2$ loss comparing the predictions to the reference which has unit peak Gaussians with a limited support at the ground truth 2D joint locations, as is common. The part affinity fields $\mathcal{A}_{COCO}$ are similarly trained with a per-pixel $L2$ loss, using the framework made available by Cao *et al.* [8]. While training ORPMs with our *MuCo-3DHP*, per joint type j, for all subjects i in the scene, a per-pixel $L2$ loss is enforced in the neighborhood of all possible read-out locations $\rho_i(j)$. The loss is weighted by a limited support Gaussian centered at the read-out location.

Table 1. Sequence-wise evaluation of our method and LCR-net[49] on multi-person 3D pose test set *MuPoTS-3D*. We report both (a) the overall accuracy (3DPCK), and (b) accuracy only for person annotations matched to a prediction

		TS1	TS2	TS3	TS4	TS5	TS6	TS7	TS8	TS9	TS10	TS11	TS12	TS13	TS14	TS15	TS16	TS17	TS18	TS19	TS20	Total
a.)	LCR-net	67.7	49.8	53.4	59.1	67.5	22.8	43.7	49.9	31.1	78.1	50.2	51.0	51.6	49.3	56.2	66.5	65.2	62.9	66.1	59.1	53.8
	Ours	**81.0**	**59.9**	**64.4**	**62.8**	**68.0**	**30.3**	**65.0**	**59.2**	**64.1**	**83.9**	**67.2**	**68.3**	**60.6**	**56.5**	**69.9**	**79.4**	**79.6**	**66.1**	**66.3**	**63.5**	**65.0**
b.)	LCR-net	69.1	**67.3**	54.6	61.7	**74.5**	25.2	48.4	**63.3**	**69.0**	78.1	53.8	52.2	60.5	60.9	59.1	70.5	76.0	70.0	**77.1**	81.4	62.4
	Ours	**81.0**	64.3	**64.6**	**63.7**	73.8	**30.3**	**65.1**	60.7	64.1	**83.9**	**71.5**	**69.6**	**69.0**	**69.6**	**71.1**	**82.9**	**79.6**	**72.2**	76.2	**85.9**	**69.8**

5. Results and Discussion

The main goal of our method is *multi-person* 3D pose estimation in general scenes, which exhibits specific and more difficult challenges than single-person pose estimation. However, we validate the usefulness of our ORPM formulation on the single-person pose estimation task as well. To validate our approach, we perform extensive experiments on our proposed multi-person test set *MuPoTS-3D*, as well as two publicly available single-person benchmarks, namely *Human3.6m* [18] and *MPI-INF-3DHP* [33]. Fig. 1 presents qualitative results of our method showcasing the ability to handle complex in-the-wild scenes with strong inter-person occlusion. An extensive collection of qualitative results is provided in the supplementary document and video.

5.1. Comparison with Prior Art

For sequences with strong occlusion, we obtain much better results than the state-of-the-art. For unoccluded sequences our results are comparable to methods designed for single person. We outperform the only other multi-person method (LCR-net [49]) quantitatively and qualitatively on both single-person and multi-person tasks. For fairness of comparison, in all evaluations, we re-target the predictions from LCR-net [49] to a skeleton with bone-lengths matching the ground truth.

Multi-Person Pose Performance: We use our proposed *MuPoTS-3D* (see Sec. 3.2) to evaluate multi-person 3D pose performance in general scenes for our approach and LCR-net [49]. In addition, we evaluate VNect [34] on images cropped with the ground truth bounding box around the subject. We evaluate for all subjects that have 3D pose annotations available. If an annotated subject is missed by our method, or by LCR-net, we consider all of its joints to be incorrect in the 3DPCK metric. Table 1(a) reports the 3DPCK metric for all 20 sequences when taking all available annotations into account. Our method performs significantly better than LCR-net for most sequences, while being comparable for a few, yielding an overall improved performance of 65.0 3DPCK vs 53.8 3DPCK for LCR-net. We provide a joint-wise breakdown of the overall accuracy in the supplementary document.

Overall, our approach detects 93% of the annotated subjects, whereas LCR-net was successful for 86%. This is an additional indicator of performance. Even ignoring the undetected annotated subjects, our approach outperforms LCR-net in terms of 3D pose error (69.8 vs 62.4 3DPCK, and 132.5 vs 146 mm MPJPE).

VNect is evaluated on ground truth crops of the subjects, and therefore it operates at a 100% detection rate. In contrast, we do not use ground truth crops, and missed detections by our method count as all joints wrong. Despite this our method achieves better accuracy (65.0 vs 61.1 3DPCK, 30.1 vs 27.6 AUC).

Single-Person Pose Performance: On the MPI-INF-3DHP dataset (see Table 2) we compare our method trained on MPI-INF-3DHP (single-person) and MuCo-3DHP (multi-person) to three single-person methods—VNect, Zhou *et al.* [68], Mehta *et al.* [33]— and LCR-net as the only other multi-person approach. Our method trained on multi-person data (73.4 3DPCK) performs marginally worse than our single-person version (75.2 3DPCK) due to the effective loss in network capacity when training on harder data. Nevertheless, both our versions consistently outperform Zhou *et al.* (69.2 3DPCK) and LCR-net (59.7 3DPCK) over all metrics. LCR-net predictions have a tendency to be conservative about the extent of articulation of limbs as shown in Fig. 6. In comparison to VNect (76.6 3DPCK) and Mehta *et al.* (75.7 3DPCK), our method (75.2 3DPCK) achieves an average accuracy that is on par. Our approach outperforms existing methods by ≈3-4 3DPCK for activities which exhibit significant self- and object-occlusion like *Sit on Chair* and *Crouch/Reach*. For the full activity-wise breakdown, see the supplemental document.

For detailed comparisons on Human3.6m [18], refer to the supplementary document. Our approach at 69.6mm MPJPE performs ≈17mm better than LCR-net [49] (87.7mm), and outperforms the VNect location-map [34] (80.5mm) formulation by ≈10mm. Our results are comparable to the recent state-of-the-art results of Pavlakos *et al.* [41] (67.1mm), Martinez *et al.* [32] (62.9mm), Zhou *et al.* [68] (64.9mm), Mehta *et al.* [33] (68.6mm) and Tekin *et al.* [60] (70.81mm), and better than the recent results from Nie *et al.* [38] (79.5mm) and Tome *et al.* [62] (88.39mm).
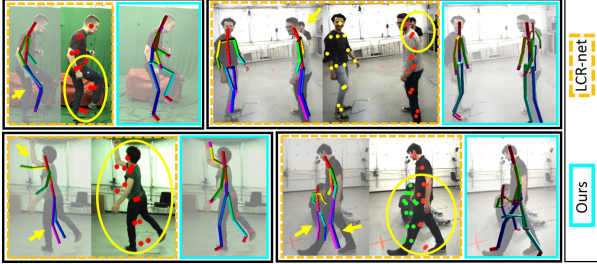
Figure 6. Qualitative comparison of LCR-net [49] and our method. LCR-net output is limited in the extent of articulation of limbs, tending towards neutral poses. LCR-net also has more detection failures under significant occlusion.

Occlusion evaluation: To demonstrate the occlusion robustness of our method, we create synthetic random occlusions on the MPI-INF-3DHP test set. The synthetic occlusions cover about 14% of the joints. Our single-person and multi-person variants both outperform VNect for occluded joints (62.8 vs 64.0 vs 53.2 3DPCK) by a large margin, and are comparable for un-occluded joints (67.0 vs 71.0 vs 69.4 3DPCK). Note again that the single-person variant is trained on similar data as VNect and the occlusion robustness is inherent to the formulation. See the supplemental document for a more detailed breakdown by test sequence.

We use the per-joint occlusion annotations from *MuPoTS-3D* to further assess LCR-net and our approach under occlusion. Considering both self- and inter-personal occlusions, $\approx$23.7 % of the joints of all subjects are occluded. Our method is more robust than LCR-net on both occluded (48.7 vs 42 3DPCK) and un-occluded (70.0 vs 57.5 3DPCK) joints.

5.2. Ablative Analysis

We validate the improvement provided by our limb refinement strategy on MPI-INF-3DHP test set. We empirically found that the base pose read out at one of the torso joints tends towards the mean pose of the training data. Hence, for poses with significant articulation of the limbs, the base pose does not provide an accurate pose estimate, especially for the end effectors. On the other side, the limb poses read out further down in the kinematic chain, *i.e.*, closer to the extremities, include more detailed articulation information for that limb. Our proposed read-out process exploits this fact, significantly improving overall pose quality by limb refinement when limbs are available. Table 2 shows that the benefit of the full read-out is consistent over all metrics and valid for our method independent of whether it is trained on single-person (MPI-INF-3DHP) or multi-person data (MuCu-3DHP), with a $\approx$10 3DPCK advantage over torso read-out. See Fig. 2,3 in the supplementary document.

Table 2. Comparison of results on MPI-INF-3DHP [33] test set. We report the *Percentage of Correct Keypoints measure in 3D* (@150mm) for select activities, and the total 3DPCK and the Area Under the Curve for all activities. Complete activity-wise breakdown in the supplementary document

Method	Sit	Crouch	Total	
	PCK	PCK	PCK	AUC
VNect [34]	74.7	72.9	**76.6**	**40.4**
LCR-net [49]	58.5	69.4	59.7	27.6
Zhou et al.[68]	60.7	71.4	69.2	32.5
Mehta et al.[33]	74.8	73.7	75.7	39.3
Our Single-Person (Torso)	69.1	68.7	65.6	32.6
Our Single-Person (Full)	**77.8**	**77.5**	75.2	37.8
Our Multi-Person (Torso)	64.6	65.8	63.6	31.1
Our Multi-Person (Full)	75.9	73.9	73.4	36.2

6. Limitations and Future Work

As discussed in Section 4, we handle overlapping joints of the same type by only supervising the one closest to the camera. However, when joints of the same type are in close proximity (but not overlapping) the ground-truth ORPM for those may transition sharply from one person to the other, which are hard to regress and may lead to inaccurate predictions. One possible way to alleviate the issue is to increase the resolution of the output maps. Another source of failures is when 2D joints are mis-predicted or mis-associated. Furthermore, we have shown accurate root-relative 3D pose estimation, but estimating the relative sizes of people is challenging and remains an open problem for future work. While the compositing based *MuCo-3DHP* covers many plausible scenarios, further investigation into capturing/generating true person-person interactions at scale would be an important next step.

7. Conclusion

Multi-person 3D pose estimation from monocular RGB is a challenging problem which has not been fully addressed by previous work. Experiments on single-person and multi-person benchmarks show that our method, relying on a novel occlusion robust pose formulation (ORPM), works well to estimate the 3D pose even under strong inter-person occlusions and human–human interactions better than previous approaches. The method has been trained on our new multi-person dataset (*MuCo-3DHP*) synthesized at scale from existing single-person images with 3D pose annotations. Our method trained on this dataset generalizes well to real world scenes shown in our *MuPOTS-3D* evaluation set. We hope further investigation into monocular multi-person pose estimation would be spurred by the proposed training and evaluation data sets.

Supplementary Document: Single-Shot Multi-Person 3D Pose Estimation From Monocular RGB

1. Read-out Process

An algorithmic description of the read-out process is provided in Alg. 1.

Algorithm 1 3D Pose Inference

1: Given: $\mathcal{P}^{2D}$, $\mathcal{C}^{2D}$, $\mathcal{M}$
2: **for all** $i \in (1..m)$ **do**
3:    **if** $\mathbf{C}_i^{2D}[k] > thresh$, $k \in \{pelvis, neck\}$ **then**
4:       Person i is detected
5:       **for all** joints $j \in (1..n)$ **do**
6:          $rloc = \mathbf{P}_i^{2D}[k]$
7:          $\mathbf{P}_i[:,j] = $ READLOCMAP(j, rloc)
8:       **for all** limbs $l \in \{arm_l, arm_r, leg_l, leg_r, head\}$ **do**
9:          **for** $j = $ GETEXTREMITY(l); $j \notin \{pelvis, neck\}$; $j = parent(j)$ **do**
10:             **if** ISVALIDREADOUTLOC(i, j) **then**
11:                REFINELIMB(l, $\mathbf{P}_i^{2D}[j]$)
12:                **break**
13:    **else**
14:       No person detected
15: **function** GETEXTREMITY(limb l)
16:    **if** $l = leg_s$ **then return** $ankle_s$
17:    **else**
18:       **if** $l = arm_s$ **then return** $wrist_s$
19:       **else return** $head$
20: **function** READLOCMAP(joint j, 2DLocation rloc)
21:    $rloc = rloc/locMap\_scale\_factor$
22:    **return** $\mathbf{M}_j[rloc]$
23: **function** REFINELIMB(limb l, 2DLocation rloc)
24:    **for all** joints $b \in$ limb l **do**
25:       $\mathbf{P}_i[:,b] = $ READLOCMAP(b, rloc)
26: **function** ISVALIDREADOUTLOC(person i, joint j)
27:    **if** $(\mathbf{C}_i^{2D}[j] > 0)$ **then**
28:       **return** ISISOLATED(i,j)
29:    **else**
30:       **return** 0
31: **function** ISISOLATED(person i, joint j)
32:    $isol = 1$
33:    **for all** persons $\bar{i} \in (1..m), \bar{i} \neq i$ **do**
34:       **for all** 2DLocations $a \in \rho_{\bar{i}}(j)$ **do**
35:          **if** $||a - \mathbf{P}_i^{2D}[j]||_2 < isoThresh$ **then**
36:             $isol = 0$
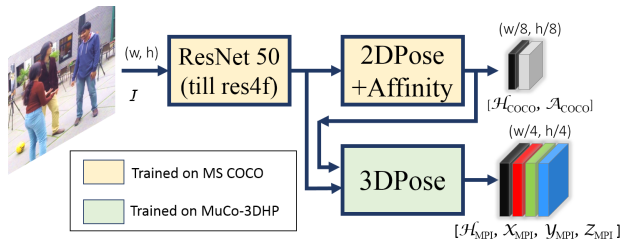37:   **break**
38:    **return** $isol$

Figure 1. The network architecture with *2DPose+Affinity* branch predicting the 2D *heatmaps* $\mathcal{H}_{COCO}$ and *part affinity maps* $\mathcal{A}_{COCO}$ with a spatial resolution of $(W/8, H/8)$, and *3DPose* branch predicting 2D *heatmaps* $\mathcal{H}_{MPI}$ and ORPMs $\mathcal{M}_{MPI}$ with a spatial resolution of $(W/4, H/4)$, for an input image with resolution (W, H).

2. Network Details

2.1. Architecture

A visualization if our network architecture using the web-based visualization tool *Netscope* can be found at: http://ethereon.github.io/netscope/#/gist/069a592125c78fbdd6eb11fd45306fa0.

2.2. Data

We use 12 out of the 14 available camera viewpoints (using only 1 of the 3 available top down views) in MPI-INF-3DHP [33] training set, and create 400k composite frames of MuCo-3DHP, of which half are without appearance augmentation. For training, we crop around the subject closest to the camera, and apply rotation, scale, and bounding-box jitter augmentation. Since the data was originally captured in a relatively restricted space, the likelihood of there being multiple people visible in the crop around the main person is high. The combination of scale augmentation, bounding-box jitter, and cropping around the subject closest to the camera results in many examples with truncation from the frame boundary, in addition to the inter-person occlusions occurring naturally due to the compositing.

2.3. Training

We train our network using the Caffe [21] framework. The core network's weights were initialized with those trained for 2D body pose estimation on MPI [3] and LSP [22, 23] datasets as done in [33]. The core network and the *2DPose + Affinity* branch are trained for multi-person 2D pose estimation using the framework provided by Cao et al. [8]. We use the AdaDelta solver, with a momentum of 0.9 and weight decay multiplier of 0.005, and a batch size of 8. We train for 640k iterations with a cyclical learning rate ranging from 0.1 to 0.000005. The *3DPose* branch is trained with the core network and *2DPose + Affinity* branch weights
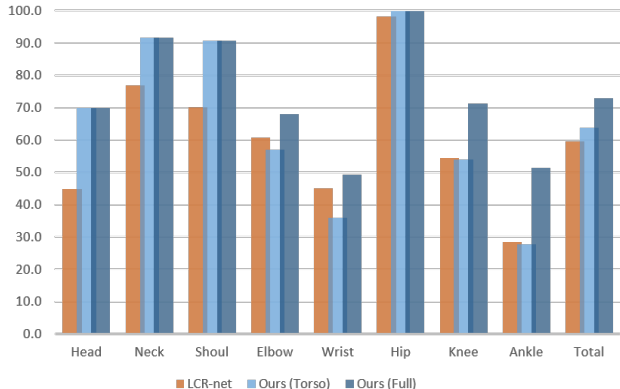
Figure 2. Joint-wise accuracy comparison of our method and LCR-net [49] on the single person MPI-INF-3DHP test set. 3D Percentage of Correct Keypoints (@150mm) as the vertical axis. LCR-net predictions were mapped to the ground truth bone lengths for fairness of comparison.

frozen. We use a batch size of 6 and train for 360k iterations with a cyclical learning rate ranging from 0.1 to 0.000001. We empirically found that training the part affinity fields and occlusion-robust pose-maps at lower resolution (see Fig. 1) leads to better results.

3. Joint-wise Analysis

Figure 2 shows joint-wise accuracy comparison of our approach with LCR-net [49] on the single person MPI-INF-3DHP test set. For limb joints (elbow, wrist, knee, ankle) LCR-net performs comparably or better than our torso-only readout, but our full readout performs significantly better. See Figure 3.

Figure 5 shows joint-wise accuracy comparison of our approach with LCR-net on our proposed multi-person 3D pose test set. We see that our approach obtains a better accuracy for all joint types for most sequences, only performing worse than LCR-net for a select few joint types on certain sequences (Test-Seq18,19,20).

4. Evaluation on Single-person Test Sets

Here we provide a detailed comparison against other methods for single-person 3D pose estimation. Evaluation on Human3.6m is in Table 1, and on MPI-INF-3DHP test set in Table 2. We additionally provide comparisons with the VNect location-maps trained on our training setup, which includes the 2D pretraining, and the 3D pose samples.

Table 3 provides a sequencewise breakdown for the synthetic occlusion experiment on MPI-INF-3DHP test set wherein through randomly placed occlusions ≈14% of the joints are occluded. This doesn't account for self-occlusions.

Figure 3. Qualitative comparison of LCR-net [49] and our method. LCR-net predictions are limited in terms of the extent of articulation of limbs, tending towards neutral poses. For our method, the base pose read out at the torso is similarly limited in terms of degree of articulation of limbs, and our full read-out addresses the issue.

Figure 4. Examples from our MuPoTS-3D evaluation set. Ground truth 3D pose reference and joint occlusion annotations are available for up to 3 subjects in the scene (shown here for the frame on the top right). The set covers a variety of scene settings, activities and clothing.

Table 1. Comparison of results on Human3.6m [18], for single un-occluded person. Human3.6m, subjects 1,5,6,7,8 used for training. Subjects 9 and 11, all cameras used for testing. Mean Per Joint Postion Error reported in mm

	Direct	Disc.	Eat	Greet	Phone	Pose	Purch.	Sit.
Pavlakos et al [41]	60.9	67.1	61.8	62.8	67.5	58.8	64.4	79.8
Mehta et al [33]	52.5	63.8	55.4	62.3	71.8	52.6	72.2	86.2
Tome et al [62]	65.0	73.5	76.8	86.4	86.3	69.0	74.8	110.2
Chen et al [9]	89.9	97.6	90.0	107.9	107.3	93.6	136.1	133.1
Moreno et al [35]	67.5	79.0	76.5	83.1	97.4	74.6	72.0	102.4
Zhou et al [68]	54.8	60.7	58.2	71.4	62.0	53.8	55.6	75.2
Martinez et al [32]	51.8	56.2	58.1	59.0	69.5	55.2	58.1	74.0
Tekin et al [60]	53.9	62.2	61.5	66.2	80.1	64.6	83.2	70.9
Nie et al [38]	62.8	69.2	79.6	78.8	80.8	72.5	73.9	96.1
VNect [34]	62.6	78.1	63.4	72.5	88.3	63.1	74.8	106.6
LCR-net [49]	76.2	80.2	75.8	83.3	92.2	79.9	71.7	105.9
VNect (with our setup)	65.52	78.8	64.8	75.0	85.2	66.4	88.1	110.2
Our Single-Person	58.2	67.3	61.2	65.7	75.82	62.2	64.6	82.0

	Sit Down	Smk.	Photo	Wait	Walk	Walk Dog	Walk Pair	Avg.
Pavlakos et al [41]	92.9	67.0	72.3	70.0	54.0	71.0	57.6	67.1
Mehta et al [33]	120.6	66.0	79.8	64.0	48.9	76.8	53.7	68.6
Tome et al [62]	173.9	84.9	110.7	85.8	71.4	86.3	73.1	88.4
Chen et al [9]	240.1	106.6	139.2	106.2	87.0	114.0	90.5	114.2
Moreno et al [35]	116.7	87.7	100.4	94.6	75.2	87.8	74.9	85.6
Zhou et al [68]	111.6	64.1	65.5	66.0	63.2	51.4	55.3	64.9
Martinez et al [32]	94.6	62.3	78.4	59.1	49.5	65.1	52.4	62.9
Tekin et al [60]	107.9	70.4	79.4	68.0	52.8	77.8	63.1	70.8
Nie et al [38]	106.9	88.0	86.9	70.7	71.9	76.5	73.2	79.5
VNect [34]	138.7	78.8	93.8	73.9	55.8	82.0	59.6	80.5
LCR-net [49]	127.1	88.0	105.7	83.7	64.9	86.6	84.0	87.7
VNect (with our setup)	155.9	82.0	95.2	76.8	59.7	94.1	64.3	84.3
Our Single-Person	93.0	68.8	84.5	65.1	57.6	72.0	63.6	69.9

Ours

	Head	Neck	Shoul	Elbow	Wrist	Hip	Knee	Ankle	Total
TestSeq1	96.8	100.0	96.6	78.0	50.9	99.8	81.0	62.3	81.0
TestSeq2	63.9	85.9	68.3	54.9	47.6	75.6	55.3	42.8	59.9
TestSeq3	79.9	91.9	90.5	56.5	46.8	98.9	42.0	30.2	64.4
TestSeq4	73.1	82.7	76.5	56.6	49.3	97.0	50.7	31.4	62.8
TestSeq5	56.5	82.0	79.7	66.2	65.7	85.5	67.3	42.0	68.0
TestSeq6	6.5	32.1	35.5	12.7	10.9	99.4	23.3	10.8	30.3
TestSeq7	66.6	97.8	81.5	47.0	21.0	98.7	63.0	61.9	65.0
TestSeq8	55.2	71.6	65.8	57.2	45.2	96.1	44.4	42.8	59.2
TestSeq9	67.2	84.1	81.5	30.1	25.7	100.0	62.7	73.1	64.1
TestSeq10	98.2	100.0	100.0	63.2	52.1	100.0	95.4	77.8	83.9
TestSeq11	66.0	92.7	84.1	56.0	44.1	89.2	73.8	43.9	67.2
TestSeq12	46.1	73.1	76.2	73.4	66.8	97.1	64.4	40.8	68.3
TestSeq13	58.5	77.9	74.5	48.6	38.5	84.0	68.9	41.3	60.6
TestSeq14	47.5	73.3	69.7	43.8	38.4	79.8	62.1	41.0	56.5
TestSeq15	62.3	91.2	84.7	58.3	42.6	97.1	77.4	52.3	69.9
TestSeq16	72.9	87.8	86.1	82.3	80.1	92.9	81.9	51.9	79.4
TestSeq17	74.4	73.8	78.0	78.1	61.3	96.5	91.0	78.6	79.6
TestSeq18	54.8	73.8	77.1	73.1	44.2	87.6	74.6	41.5	66.1
TestSeq19	44.9	78.4	79.4	55.2	54.1	84.4	77.7	37.9	64.3
TestSeq20	50.8	73.5	71.7	62.5	58.6	73.0	69.0	47.5	63.5

LCR-net

	Head	Neck	Shoul	Elbow	Wrist	Hip	Knee	Ankle	Total
TestSeq1	73.1	81.8	74.4	61.3	41.9	97.0	74.9	47.1	67.7
TestSeq2	54.6	69.3	53.9	43.5	31.7	69.3	48.6	39.3	49.8
TestSeq3	71.2	81.0	56.4	35.8	29.6	95.1	49.1	31.7	53.4
TestSeq4	57.6	80.7	63.2	55.1	48.3	94.7	52.0	31.3	59.1
TestSeq5	68.0	84.9	72.1	70.5	60.5	84.7	65.1	43.2	67.5
TestSeq6	6.3	26.8	18.6	17.1	12.8	84.8	7.6	2.3	22.8
TestSeq7	34.4	66.8	38.8	31.5	22.9	87.6	49.1	25.6	43.7
TestSeq8	56.3	70.0	52.1	44.6	35.4	75.7	46.8	31.7	49.9
TestSeq9	34.9	41.9	34.1	20.0	15.9	45.1	32.8	31.4	31.1
TestSeq10	80.5	97.6	98.0	53.1	31.9	100.0	87.0	87.9	78.1
TestSeq11	23.3	43.0	39.4	41.5	44.1	88.2	57.4	27.4	50.2
TestSeq12	24.2	41.5	39.2	61.1	65.1	97.1	41.2	20.6	51.0
TestSeq13	42.0	62.2	51.5	48.2	37.4	78.5	57.1	36.5	51.6
TestSeq14	36.6	63.2	50.7	39.9	29.2	80.7	57.5	37.4	49.3
TestSeq15	39.4	72.8	59.1	44.6	34.4	91.4	73.6	34.6	56.2
TestSeq16	48.1	67.8	65.0	78.6	68.3	93.1	67.0	35.4	66.5
TestSeq17	44.1	77.7	75.7	68.8	58.9	85.4	59.7	46.8	65.2
TestSeq18	53.7	89.9	83.5	63.6	42.5	89.9	60.8	28.4	62.9
TestSeq19	69.6	80.0	67.2	62.4	53.0	81.1	73.9	50.2	66.1
TestSeq20	70.5	48.8	49.5	67.3	61.9	72.6	64.4	38.0	59.1

Difference

	Head	Neck	Shoul	Elbow	Wrist	Hip	Knee	Ankle	Total
TestSeq1	23.6	18.2	22.3	16.7	9.0	2.7	6.1	15.2	13.3
TestSeq2	9.4	16.5	14.4	11.4	15.9	6.3	6.7	3.5	10.2
TestSeq3	8.7	10.8	34.2	20.6	17.2	3.7	-7.0	-1.5	11.0
TestSeq4	15.4	2.0	13.3	1.5	0.9	2.3	-1.3	0.1	3.6
TestSeq5	-11.5	-2.9	7.6	-4.3	5.2	0.9	2.2	-1.1	0.5
TestSeq6	0.2	5.3	16.9	-4.4	-1.8	14.6	15.7	8.5	7.5
TestSeq7	32.2	30.9	42.7	15.5	-1.8	11.1	13.8	36.3	21.3
TestSeq8	-1.1	1.6	13.7	12.6	9.8	20.4	-2.4	11.1	9.3
TestSeq9	32.2	42.2	47.4	10.2	9.8	54.9	29.8	41.6	33.0
TestSeq10	17.7	2.4	2.0	10.2	20.2	0.0	8.5	-10.2	5.8
TestSeq11	42.7	49.7	44.7	-5.5	0.0	1.0	16.4	16.5	17.0
TestSeq12	21.9	31.5	37.0	12.3	1.7	0.0	23.2	20.3	17.3
TestSeq13	16.4	15.7	23.0	0.3	1.1	5.5	11.8	4.8	8.9
TestSeq14	10.9	10.2	19.0	3.8	9.3	-0.9	4.6	3.6	7.1
TestSeq15	22.9	18.3	25.6	13.8	8.2	5.7	3.7	17.8	13.6
TestSeq16	24.8	20.0	21.1	3.7	11.8	-0.2	14.9	16.5	12.9
TestSeq17	30.3	-3.9	2.2	9.3	2.3	11.1	31.3	31.7	14.5
TestSeq18	1.1	-16.1	-6.3	9.5	1.7	-2.4	13.8	13.1	3.1
TestSeq19	-24.7	-1.5	12.2	-7.2	1.0	3.2	3.8	-12.4	-1.8
TestSeq20	-19.6	24.7	22.1	-4.9	-3.3	0.4	4.6	9.5	4.4

Figure 5. Comparison of our method and LCR-net [49] on our proposed multi-person test set, here visualized as joint-wise breakdown of PCK for all 20 sequences, as well as the difference in accuracy between our method and LCR-net. LCR-net predictions were mapped to the ground truth bone lengths for fairness of comparison.

Table 2. Comparison of our method against the state of the art on single person MPI-INF-3DHP test set. All evaluations use ground-truth bounding box crops around the subject. We report the *Percentage of Correct Keypoints measure in 3D* (@150mm), and the Area Under the Curve for the same, as proposed by MPI-INF-3DHP. We additionally report the Mean Per Joint Position Error in mm. Higher PCK and AUC is better, and lower MPJPE is better.

Network	Stand/ Walk	Exercise	Sit On Chair	Crouch/ Reach	On the Floor	Sports	Misc.	Total		
	PCK	PCK	PCK	PCK	PCK	PCK	PCK	PCK	AUC	MPJPE(mm)
VNect [34]	**87.7**	**77.4**	74.7	72.9	51.3	**83.3**	**80.1**	76.6	40.4	124.7
LCR-net [49]	70.5	56.3	58.5	69.4	39.6	57.7	57.6	59.7	27.6	158.4
Zhou et al.[68]	85.4	71.0	60.7	71.4	37.8	70.9	74.4	69.2	32.5	137.1
Mehta et al.[33]	86.6	75.3	74.8	73.7	52.2	82.1	77.5	75.7	39.3	117.6
Ours Single-Person (Torso)	75.0	64.8	69.1	68.7	48.6	70.0	60.6	65.6	32.6	142.8
Ours Single-Person (Full)	83.8	75.0	<u>77.8</u>	**77.5**	<u>55.1</u>	80.4	72.5	75.2	37.8	122.2
Ours Multi-Person (Torso)	73.7	63.7	64.6	65.8	44.7	69.5	60.2	63.6	31.1	146.8
Ours Multi-Person (Full)	82.0	74.5	75.9	<u>73.9</u>	51.6	79.0	71.8	73.4	36.2	126.3
VNect (our train. setup)	85.7	75.4	**78.6**	72.3	**60.2**	81.8	73.4	75.8	38.9	120.1

Table 3. Testing occlusion robustness of our method through synthetic occlusions on MPI-INF-3DHP single person test set. The synthetic occlusions cover about 14% of the evaluated joints overall. We report the *Percentage of Correct Keypoints measure in 3D* (@150mm) overall, as well as split by occlusion. Higher PCK.

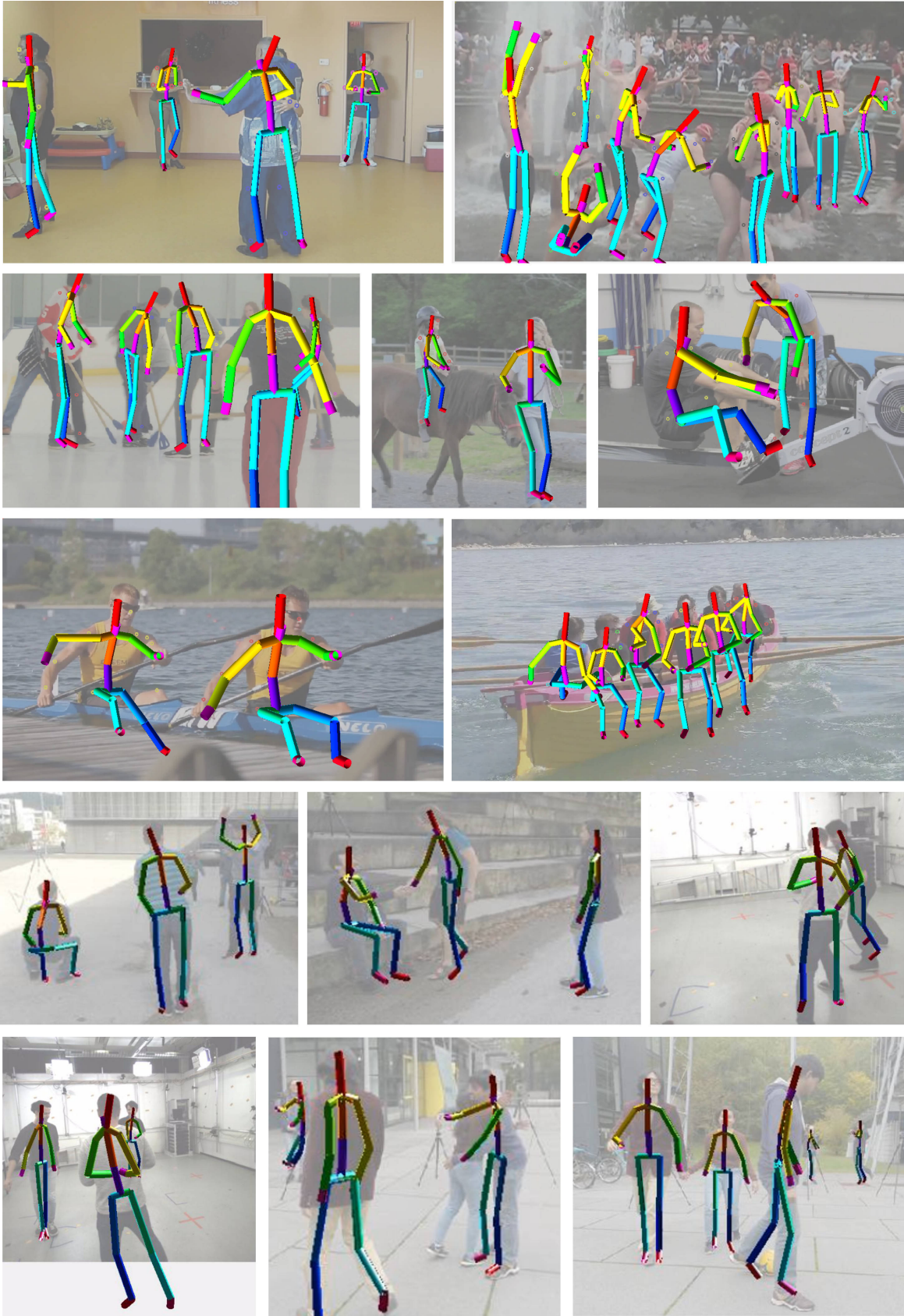|	Seq1	Seq2	Seq3	Seq4	Seq5	Seq6	Total
	PCK	PCK	PCK	PCK	PCK	PCK	PCK
Overall							
Ours Multi-Person	78.7	70.0	71.9	65.2	61.4	60.7	69.0
Ours Single-Person	80.9	72.8	72.6	65.7	62.5	65.8	71.1
VNect [34]	80.1	72.4	72.4	61.5	50.2	69.8	69.4
VNect (our train. setup)	79.3	74.4	72.2	67.2	55.7	64.6	70.4
Occluded Subset of Joints							
Ours Multi-Person	73.3	66.5	55.0	56.5	45.1	64.9	62.8
Ours Single-Person	74.9	63.2	59.0	54.2	48.0	68.4	64.0
VNect [34]	61.4	54.5	47.6	36.4	30.5	66.2	53.2
VNect (our train. setup)	69.6	61.9	49.0	50.8	43.5	63.4	59.2
Un-occluded Subset of Joints							
Ours Multi-Person	79.9	70.5	73.7	66.2	64.6	59.5	70.0
Ours Single-Person	82.1	74.0	74.1	67.0	65.3	65.1	72.2
VNect [34]	83.9	74.6	75.0	64.4	54.0	70.9	72.1
VNect (our train. setup)	81.3	76.0	74.6	69.0	58.1	64.8	72.2

Figure 6. More qualitative results of our approach on MPI 2D pose dataset [3] and our proposed MuPoTS-3D test set.

References

[1] 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] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2018.

[3] 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.

[4] 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.

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

[6] 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.

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

[8] Z. Cao, T. Simon, S. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[9] C.-H. Chen and D. Ramanan. 3d human pose estimation = 2d pose estimation + matching. In *CVPR 2017-IEEE Conference on Computer Vision & Pattern Recognition*, 2017.

[10] 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.

[11] 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.

[12] 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.

[13] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik. Using k-poselets for detecting people and localizing their keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3582–3589, 2014.

[14] R. A. Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild.

[15] 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.

[16] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, B. Schiele, and S. I. Campus. Arttrack: Articulated multi-person tracking in the wild. In *Proc. of CVPR*, 2017.

[17] 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.

[18] 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.

[19] U. Iqbal and J. Gall. Multi-person pose estimation with local joint-to-person associations. In *European Conference on Computer Vision Workshops*, pages 627–642. Springer, 2016.

[20] E. Jahangiri and A. L. Yuille. Generating multiple diverse hypotheses for human 3d pose consistent with 2d joint detections. In *IEEE International Conference on Computer Vision (ICCV) Workshops (PeopleCap)*, 2017.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Regonition (CVPR)*, 2018.

[26] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[27] 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.

[28] 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.

[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[30] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015.

[31] D. C. Luvizon, D. Picard, and H. Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2018.

[32] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[33] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*, 2017.

[34] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. volume 36, 2017.

[35] F. Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *CVPR 2017-IEEE Conference on Computer Vision & Pattern Recognition*, 2017.

[36] A. Newell and J. Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

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

[38] B. X. Nie, P. Wei, and S.-C. Zhu. Monocular 3d human pose estimation by predicting depth on joints. In *IEEE International Conference on Computer Vision*, 2017.

[39] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *International Conf. on 3D Vision*, 2018.

[40] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy. Towards accurate multi-person pose estimation in the wild. *arXiv preprint arXiv:1701.01779*, 2017.

[41] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *CVPR 2017-IEEE Conference on Computer Vision & Pattern Recognition*, 2017.

[42] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *CVPR*, 2018.

[43] 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.

[44] 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.

[45] 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.

[46] A.-I. Popa, M. Zanfir, and C. Sminchisescu. Deep multitask architecture for integrated 2d and 3d human sensing. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[47] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[48] 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.

[49] G. Rogez, P. Weinzaepfel, and C. Schmid. Lcr-net: Localization-classification-regression for human pose. In *CVPR 2017-IEEE Conference on Computer Vision & Pattern Recognition*, 2017.

[50] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris. 3d human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding*, 152:1–20, 2016.

[51] 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.

[52] 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.

[53] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[54] C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3d human tracking. In *Com-

puter Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, volume 1, pages I–I. IEEE, 2003.

[55] M. Sun and S. Savarese. Articulated part-based model for joint object detection and pose estimation. In *IEEE International Conference on Computer Vision*, pages 723–730. IEEE, 2011.

[56] X. Sun, J. Shang, S. Liang, and Y. Wei. Compositional human pose regression. *arXiv preprint arXiv:1704.00159*, 2017.

[57] J. K. V. Tan, I. Budvytis, and R. Cipolla. Indirect deep structured learning for 3d human body shape and pose prediction. In *BMVC*, volume 3, page 6, 2017.

[58] 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.

[59] 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.

[60] B. Tekin, P. Márquez-Neila, M. Salzmann, and P. Fua. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

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

[62] D. Tome, C. Russell, and L. Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[63] M. Trumble, A. Gilbert, C. Malleson, A. Hilton, and J. Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *Proceedings of 28th British Machine Vision Conference*, pages 1–13, 2017.

[64] T. von Marcard, G. Pons-Moll, and B. Rosenhahn. Human pose estimation from video and imus. *Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1533–1547, Jan. 2016.

[65] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang. 3d human pose estimation in the wild by adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, 2018.

[66] 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.

[67] A. Zanfir, E. Marinoiu, and C. Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes–the importance of multiple scene constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2148–2157, 2018.

[68] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Towards 3d human pose estimation in the wild: A

weakly-supervised approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 398–407, 2017.

[69] 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.