

HEMlets Pose: Learning Part-Centric Heatmap Triplets for Accurate 3D Human Pose Estimation

Kun Zhou¹, Xiaoguang Han², Nianjuan Jiang¹, Kui Jia³, and Jiangbo Lu^{1*}

¹Shenzhen Cloudream Technology Co., Ltd. ²The Chinese University of Hong Kong (Shenzhen)

³South China University of Technology *Corresponding email: jiangbo.lu@gmail.com

Abstract

Estimating 3D human pose from a single image is a challenging task. This work attempts to address the uncertainty of lifting the detected 2D joints to the 3D space by introducing an intermediate state - Part-Centric Heatmap Triplets (HEMlets), which shortens the gap between the 2D observation and the 3D interpretation. The HEMlets utilize three joint-heatmaps to represent the relative depth information of the end-joints for each skeletal part. In our approach, a Convolutional Network (ConvNet) is first trained to predict HEMlets from the input image, followed by a volumetric joint-heatmap regression. We leverage on the integral operation to extract the joint locations from the volumetric heatmaps, guaranteeing end-to-end learning. Despite the simplicity of the network design, the quantitative comparisons show a significant performance improvement over the best-of-grade method (about 20% on Human3.6M). The proposed method naturally supports training with “in-the-wild” images, where only weakly-annotated relative depth information of skeletal joints is available. This further improves the generalization ability of our model, as validated by qualitative comparisons on outdoor images.

1. Introduction

Human pose estimation from a single image is an important problem in computer vision, because of its wide applications, e.g., video surveillance and human-computer interaction. Given an image containing a single person, 3D human pose inference aims to predict 3D coordinates of the human body joints. Recovering 3D information of human poses from a single image faces several challenges. The challenges are at least three folds: 1) reasoning 3D human poses from a single image is by itself very challenging due to the inherent ambiguities; 2) being a regression problem, existing approaches have not achieved a good balance between representation efficiency and learning effectiveness; 3) for “in-the-wild” images, both 3D capturing and manual labeling require a lot of efforts to obtain high-quality 3D

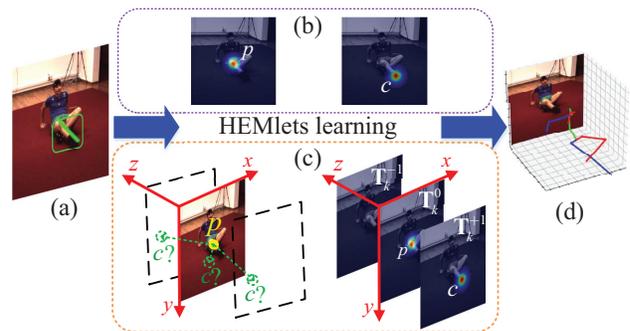


Figure 1. Overview of the HEMlets-based 3D pose estimation. (a) Input RGB image. Our algorithm encodes (b) the 2D locations for the joints p and c , but also (c) their relative depth relationship for each skeletal part $\bar{p}\bar{c}$ into HEMlets. (d) Output 3D human pose.

annotations, making the training data extremely scarce.

For 2D human pose estimation, almost all best performing methods are detection based [16, 11, 34]. Detection-based approaches essentially divide the joint localization task into local image classification tasks. The latter is easier to train, because it effectively reduces the feature and target dimensions for the learning system [28]. Existing 3D pose estimation methods often use detection as an intermediate supervision mechanism as well. A straightforward strategy is to use volumetric heatmaps to represent the likelihood map of each 3D joint location [20]. Sun *et al.* [28] further proposed a differentiable soft-argmax operator that unifies the joint detection task and the regression task into an end-to-end training framework. This significantly improves the state-of-the-art 3D pose estimation accuracy.

In this work, we propose a novel effective intermediate representation for 3D pose estimation - *Part-Centric Heatmap Triplets (HEMlets)* (as shown in Fig. 1). The key idea is to polarize the 3D volumetric space around each distinct skeletal part, which has the two end-joints kinematically connected. Different from [19], our relative depth information is represented as three polarized heatmaps, corresponding to the different state of the local depth ordering of the part-centric joint pairs. Intuitively, HEMlets encodes

the co-location likelihoods of pairwise joints in a dense per-pixel manner with the coarsest discretization in the depth dimension. Instead of considering arbitrary joint pairs, we focus on kinematically connected ones as they possess semantic correspondence with the input image, and are thus a more effective target for the subsequent learning. In addition, the encoded relative depth information is strictly local for the part-centric joint pairs and suffers less from potential inconsistent data annotations.

The proposed network architecture is shown in Fig. 3. A ConvNet is first trained to learn the HEMlets and 2D joint heatmaps, which are then fed together with the high-level image features to another ConvNet to produce a volumetric heatmap for each joint. We leverage on the soft-argmax regression [28] to obtain the final 3D coordinates of each joint. Significant improvements are achieved compared to the best competing methods quantitatively and qualitatively. Most notably, our HEMlets method achieves a record MPJPE of 39.9mm on Human3.6M [9], yielding about 20% improvement over the best-of-grade method [28].

The merits of the proposed method lie in three aspects:

- **Learning strategy.** Our method takes on a progressive learning strategy, and decomposes a challenging 3D learning task into a sequence of easier sub-tasks with mixed intermediate supervisions, i.e., 2D joint detection and HEMlets learning. HEMlets is the key bridging and learnable component leading to 3D heatmaps, and is much easier to train and less prone to overfitting. Its training can also take advantage of existing labeled datasets of relative depth ordering [19, 25].
- **Representation power.** HEMlets is based on 2D per-joint heatmaps, but extends them by a couple of additional heatmaps to encode local depth ordering in a dense per-pixel manner. It builds on top of 2D heatmaps but unleashes the representation power, while still allowing leveraging the soft-argmax regression [28] for end-to-end learning.
- **Simple yet effective.** The proposed method features a simple network architecture design, and it is easy to train and implement. It achieves state-of-the-art 3D pose estimation results validated by the evaluations over all standard benchmarks.

2. Related Work

In this section, we review the approaches that are based on deep ConvNets for 3D human pose estimation.

2.1. Direct encoder-decoder

With the powerful feature extraction capability of deep ConvNets, many approaches [12, 29, 18] learn end-to-end *Convolutional Neural Networks* (CNNs) to infer human

poses directly from the images. Li and Chen [12] are the first who used CNNs to estimate 3D human pose via a multi-task framework. Tekin *et al.* [29] designed an auto encoder to model the joint dependencies in a high-dimensional feature space. Park *et al.* [18] proposed fusing 2D joint locations with high-level image features to boost the estimation of 3D human pose. However, these single-stage methods are limited by the availability of 3D human pose datasets and cannot take advantage of large-scale 2D pose datasets that are vastly available.

2.2. Transition with 2D joints

To avoid collecting 2D-3D paired data, a large number of works [23, 37, 35, 13, 7, 25] decompose the task of 3D pose estimation into two independent stages: 1) firstly inferring 2D joint locations using well-studied 2D pose estimation methods, such as [37, 23]; 2) and then learning a mapping to lift them into the 3D space. These approaches mainly focus on tackling the second problem. For example, a simple fully connected residual network is proposed by Martinez *et al.* [13] to directly recover 3D human pose from its 2D projection. Fang *et al.* [7] considered prior knowledge of human body configurations and proposed human pose grammar, leading to better recovery of the 3D pose from only 2D joint locations. Yang *et al.* [35] adopted an adversarial learning scheme to ensure the anthropometrical validity of the output pose and further improved the performance. Recently, by involving a reprojection mechanism, the proposed method in [33] shows insensitivity to overfitting and accurately predicts the result from noisy 2D poses. Though promising results have been achieved by these two-stage methods, a large gap exists between the 3D human pose and its 2D projections due to inherent ambiguities.

2.3. 3D-aware intermediate states

To further bridge the gap between the 2D image and the target 3D human pose under estimation, some recent works [20, 25, 19, 28] proposed to involve 3D-aware states for intermediate supervisions. Namely, a network is firstly trained to map the input image to these 3D-aware states, and then another network is trained to convert those states to the 3D joint locations. Finally, these two networks are combined and optimized jointly. A volumetric representation for 3D joint-heatmaps is proposed in [20], with which the 3D pose is regressed in a coarse-to-fine manner. However, regressing a probability grid in the 3D space globally is also a very challenging task. It usually suffers from quantization errors for the joint locations. To address this issue, Sun *et al.* [28] exploited a soft-argmax operation and proposed an end-to-end training scheme for the 3D volumetric regression, achieving by far the best performance on 3D pose estimation. Inspired by [21] that the relative depth ordering across joints is helpful for resolving pose ambiguities,

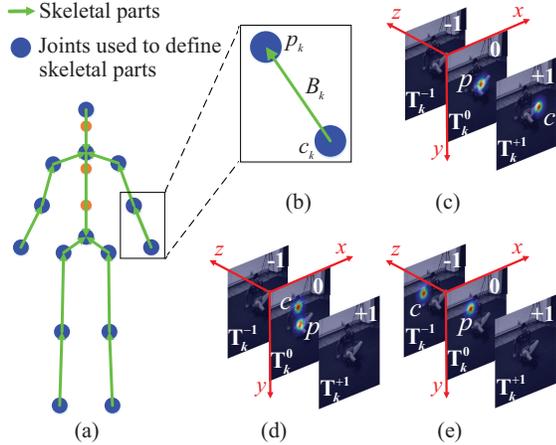


Figure 2. Part-centric heatmap triplets $\{\mathbf{T}_k^{-1}, \mathbf{T}_k^0, \mathbf{T}_k^{+1}\}$ where p and c are the parent joint and the child joint. (a, b) Joints and skeletal parts. We locate the parent joint p of the k -th skeletal part B_k at the zero polarity heatmap \mathbf{T}_k^0 (c-e). The child joint c is located, according to relative depth of p and c , in the positive (c), zero (d) and negative polarity heatmap (e), respectively.

Pavlakos *et al.* [19] adopted a ranking loss for pairwise ordinal depth to train the 3D human pose predictor explicitly. A similar scheme of relative depth supervision is utilized in the work of [23]. Forward-or-Backward Information (FBI), proposed in [25], is another kind of relative depth information but focuses more on the bone orientations.

In this work, we propose HEMlets, a novel representation that encodes both 2D joint locations and the part-centric relative depth ordering simultaneously. Experiments justify that this representation reaches by far the best balance between representation efficiency and learning effectiveness.

2.4. “In-the-wild” adaptation

All the aforementioned approaches are mainly trained on the datasets collected under indoor settings, due to the difficulty of annotating 3D joints for “in-the-wild” images [3]. Thus, many strategies are developed to make domain adaptation. By exploiting graphics techniques, previous works [32, 5] have synthesized a large “faked” dataset mimicking real images. Though these data benefit 3D pose estimation, they are still far from realistic, making the applicability limited. Recently, both Pavlakos *et al.* [19] and Shi *et al.* [25] proposed to label the relative depth relationship across joints instead of the exact 3D joint coordinates. This weak annotation scheme not only makes building large-scale “in-the-wild” datasets feasible, but also provides 3D-aware information for training the inference model in a weakly-supervised manner. With HEMlets representation, we can readily use these weakly annotated “in-the-wild” data for domain adaptation.

3. HEMlets Pose Estimation

We propose a unified representation of heatmap triplets to model the local information of body skeletal parts, i.e., kinematically connected joints, whereas the corresponding 2D image coordinates and relative depth ordering are considered. By such a representation, images annotated with relative depth ordering of skeletal parts can be treated equally with images annotated with 3D joint information. While the latter is usually very scarce, the former is relatively easy to obtain [25, 19]. In this section, we first present the proposed part-centric heatmap triplets and its encoding scheme. Then, we elaborate a simple network architecture that utilizes the part-centric heatmap triplets for 3D human pose estimation.

3.1. Part-centric heatmap triplets

We divide the full body skeleton consisting of $N = 18$ joints into $K = 14$ parts as shown in Fig. 2(a). Specifically, we use B to denote the set of skeletal parts, where $B = \{B_1, B_2, \dots, B_K\}$. For each part, we denote the two associated joints as (p, c) , with p being the parent node and c being the child node. The relative depth ordering, denoted as $r(z_p, z_c)$, can be then described as a tri-state function [19, 25]:

$$r(z_p, z_c) = \begin{cases} 1 & z_p - z_c > \epsilon \\ 0 & |z_p - z_c| < \epsilon \\ -1 & z_p - z_c < -\epsilon \end{cases}, \quad (1)$$

where ϵ is used to adjust the sensitivity of the function to the relative depth difference. The absolute depths of the two joints p and c are denoted by z_p and z_c , respectively.

We argue that directly using the discretized label as an intermediate state for learning the 3D pose from a 2D joint heatmap, as was done in [19, 25], is not as effective. Since this abstraction tends to lose some important features encoded in the joints’ spatial domain. Instead of elevating the problem straight away to the 3D volumetric space, we utilize an intermediate representation of the 3D-aware relationship of the parent joint p_k and the child joint c_k of a skeletal part B_k . Provided with the supervision signals, we define polarized target heatmaps where a pair of normalized Gaussian peaks corresponding to the 2D joint locations are placed accordingly across three heatmaps (see Figure 2). We term them as the *negative polarity heatmap* \mathbf{T}_k^{-1} , the *zero polarity heatmap* \mathbf{T}_k^0 and the *positive polarity heatmap* \mathbf{T}_k^{+1} with respect to the function value in Eq. (1). The parent joint p_k is always placed in the zero polarity heatmap \mathbf{T}_k^0 . The child joint c_k will appear in the negative/positive polarity heatmap, if its depth is larger/smaller than that of the parent joint p_k (i.e., $|r(z_p, z_c)| \neq 0$). Both parent and

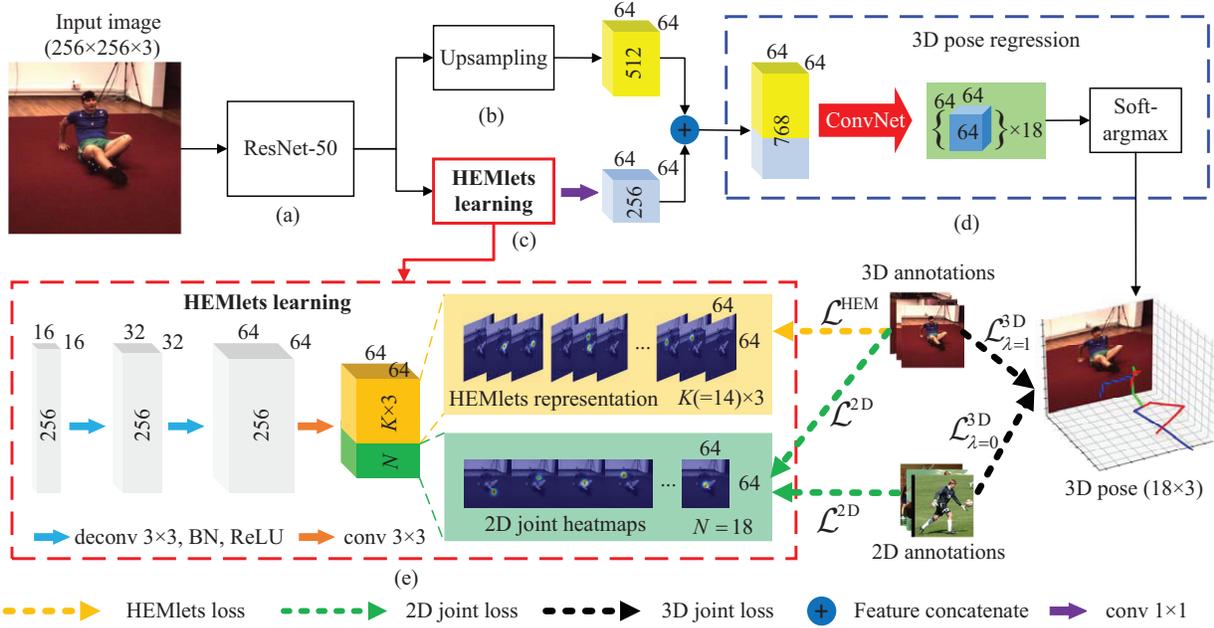


Figure 3. The network architecture of our proposed approach. It consists of four major modules: (a) A ResNet-50 backbone for image feature extraction. (b) A ConvNet for image feature upsampling. (c) Another ConvNet for HEMlets learning and 2D joint detection. (d) A 3D pose regression module adopting a soft-argmax operation for 3D human pose estimation. (e) Details of the HEMlets learning module. “Feature concatenate” denotes concatenating the feature maps from the HEMlets learning branch and the upsampling branch together.

child joints are co-located in the zero polarity heatmap if their depths are roughly the same (i.e., $r(z_p, z_c) = 0$).

Formally, we denote the heatmap triplets of the skeletal part B_k as the stacking of three heatmaps $\mathbf{T}_k^{-1}, \mathbf{T}_k^0, \mathbf{T}_k^{+1}$:

$$\mathbf{T}_k = \text{Stack}[\mathbf{T}_k^{-1}, \mathbf{T}_k^0, \mathbf{T}_k^{+1}]. \quad (2)$$

Given 3D groundtruth coordinates of all joints, we can readily compute the heatmap triplets of each skeletal part. For easy reference, we shall refer to the part-centric heatmap triplets \mathbf{T}_k as *HEMlets*, and use it afterwards.

Discussions. Here we provide some understandings of HEMlets from a few perspectives. First, different from a joint-specific 2D heatmap that models the detection likelihood for each intended joint on the (x, y) plane, HEMlets models part-centric pairwise joints’ co-location likelihoods on the (x, y) plane simultaneously with their ordinal depth relations. This helps to learn geometric constraints (e.g., bone lengths) implicitly. Second, by augmenting a 2D heatmap to a triplet of heatmaps, HEMlets learns and evaluates the co-location likelihood for a pair of connected joints (p, c) by the joint probability distribution $P(x_p, y_p, x_c, y_c, r(z_p, z_c))$ in a locally-defined volumetric space. In contrast, Pavlakos *et al.* [19] relaxed the learning target and marginalized the 3D probability distributions independently for the (x, y) plane i.e., $P(x_p, y_p), P(x_c, y_c)$ and the z -dimension, with the latter supervised independently by $r(z_p, z_c)$ based on a ranking loss. Third, by ex-

plotting the available supervision signals to a larger extent, HEMlets brings the benefit of making the knowledge more explicitly expressed and easier to learn, and bridges the gap in learning the 3D information from a given 2D image.

3.2. 3D pose inference

Network architecture. We employ a fully convolutional network to predict the 3D human pose as illustrated in Figure 3. A ResNet-50 [8] backbone architecture is adopted for basic feature extraction. One of the two upsampling branches is used to learn the HEMlets and the 2D heatmaps of skeletal joints, and the other one is used to perform upsampling of the learned features to the same resolution as the output heatmaps. Both HEMlets and the 2D joint heatmaps are then encoded jointly by a 2D convolutional operation to form a latent global representation. Finally these global features are joined with the convolutional features extracted from the original image to predict a 3D feature map for each joint. We perform a soft-argmax operation [28] to aggregate information in the 3D feature maps to obtain the 3D joint estimations.

HEMlets loss. Let us denote with \mathbf{T}^{gt} the groundtruth HEMlets of all skeletal parts and with $\hat{\mathbf{T}}$ the corresponding prediction. We use a standard L_2 distance between \mathbf{T}^{gt} and $\hat{\mathbf{T}}$ to compute the HEMlets loss as follows:

$$\mathcal{L}^{\text{HEM}} = \|(\mathbf{T}^{\text{gt}} - \hat{\mathbf{T}}) \odot \mathbf{A}\|_2^2, \quad (3)$$

where \odot denotes an element-wise multiplication, and \mathbf{A} is a binary tensor to mask out missing annotations.

Auxiliary 2D joint loss. As HEMlets essentially contains heatmap responses of 2D joint locations, we adopt a heatmap-based 2D joint detection scheme to facilitate HEMlets prediction. The L_2 loss of 2D joint prediction is computed as:

$$\mathcal{L}^{2D} = \sum_{n=1}^N \|\mathbf{H}_n^{\text{gt}} - \hat{\mathbf{H}}_n\|_2^2, \quad (4)$$

where \mathbf{H}_n^{gt} is the groundtruth 2D heatmap of the n -th 2D joint and $\hat{\mathbf{H}}_n$ is the corresponding network prediction.

Soft-argmax 3D joint loss. To avoid quantization errors and allow end-to-end learning, Sun *et al.* [28] suggested a soft-argmax regression for 3D human pose estimation. Given learned volumetric features \mathbf{F}_n of size $(h \times w \times d)$ for the n -th joint, the predicted 3D coordinates are given as:

$$[\hat{x}_n, \hat{y}_n, \hat{z}_n] = \int_{\mathbf{v}} \mathbf{v} \cdot \text{Softmax}(\mathbf{F}_n), \quad (5)$$

where \mathbf{v} denotes a voxel in the volumetric feature space of \mathbf{F}_n . For robustness, we employ the L_1 loss for the regression of 3D joints. Specifically, the loss is defined as:

$$\mathcal{L}_{\lambda}^{3D} = \sum_{n=1}^N (|x_n^{\text{gt}} - \hat{x}_n| + |y_n^{\text{gt}} - \hat{y}_n| + \lambda |z_n^{\text{gt}} - \hat{z}_n|), \quad (6)$$

where the groundtruth 3D position of the n -th joint is given as $(x_n^{\text{gt}}, y_n^{\text{gt}}, z_n^{\text{gt}})$. We use the same 2D and 3D mixed training strategy in [28] ($\lambda \in \{0, 1\}$): λ in Eq. (6) is set to 1 when the training data is from 3D datasets, and $\lambda = 0$ when the data is from 2D datasets.

Training strategy. For HEMlets prediction, We combine \mathcal{L}^{HEM} and \mathcal{L}^{2D} for the intermediate supervision. The loss function is defined as:

$$\mathcal{L}^{\text{int}} = \mathcal{L}^{\text{HEM}} + \mathcal{L}^{2D}. \quad (7)$$

By using \mathcal{L}^{HEM} and \mathcal{L}^{2D} jointly as supervisions, we allow training the network using images with 2D joint annotations and 3D joint annotations. By 3D joint annotation, we refer to annotations with exact 3D joint coordinates or relative depth ordering between part-centric joint pairs.

The end-to-end training loss \mathcal{L}^{tot} is defined by combining \mathcal{L}^{int} with $\mathcal{L}_{\lambda}^{3D}$:

$$\mathcal{L}^{\text{tot}} = \alpha * \mathcal{L}^{\text{int}} + \mathcal{L}_{\lambda}^{3D}, \quad (8)$$

where $\alpha = 0.05$ in all our experiments.

Implementation details. We implement our method in PyTorch. The model is trained in an end-to-end manner using both images with 3D annotations (e.g., Human3.6M [9] or HumanEva-I [26]), and 2D annotations (MPII [1]). In our experiments, we adopt an adaptive value of ϵ in Eq. (1) for each skeletal part: $\epsilon_k = 0.5\|B_k\|$ ($\|B_k\|$ is the 3D Euclidean distance between the two end joints of the skeletal part B_k). The training data is further augmented with rotation ($\pm 30^\circ$), scale (0.75–1.25), horizontal flipping (with a probability of 0.5) and color distortions. By using a batch size of 64, a learning rate of 0.001 and Adam optimization, the training took 100K iterations to converge. It took about a few days (2–4) with four NVIDIA GTX 1080 GPUs to train the model.

4. Experiments

We perform quantitative evaluation on three benchmark datasets: Human3.6M [9], HumanEva-I [26] and MPI-INF-3DHP [14]. Ablation study is conducted to evaluate our design choices. We demonstrate that the proposed method shows superior generalization ability to in-the-wild images.

4.1. Datasets and evaluation protocols

Human3.6M. Human3.6M [9] contains 3.6 million RGB images captured by a MoCap System in an indoor environment, in which 7 professional actors were performing 15 activities such as walking, eating, sitting, making a phone call and engaging in a discussion, etc. We follow the standard protocol as in [13, 20], and use 5 subjects (S1, S5, S6, S7, S8) for training and the rest 2 subjects (S9, S11) for evaluation (referred to as Protocol #1). Some previous works reported their results with 6 subjects (S1, S5, S6, S7, S8, S9) used for training and only S11 for evaluation [36, 28, 6] (referred to as Protocol #2). Despite *not* using S9 as training data, we compare our results with these methods.

HumanEva-I. HumanEva-I [26] is one of the early datasets for evaluating 3D human poses. It contains fewer subjects and actions compared to Human3.6M. Following [2], we train a single model on the training sequences of Subject 1, 2 and 3, and evaluate on the validation sequences.

MPI-INF-3DHP. This is a recent 3D human pose dataset which includes both indoor and outdoor scenes [14]. Without using its training set, we evaluate our model trained from Human3.6M only on the test set. The results are reported using the 3DPCK and the AUC metric [1, 14, 19].

Evaluation metric. We follow the standard steps to align the 3D pose prediction with the groundtruth by aligning the position of the central hip joint, and use the *Mean Per-Joint Position Error* (MPJPE) between the groundtruth and the prediction as evaluation metrics. In some prior

Protocol #1	Direct	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SittingD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
LinKDE <i>et al.</i> [9]	132.7	183.6	132.3	164.4	162.1	205.9	150.6	171.3	151.6	243.0	162.1	170.7	177.1	96.6	127.9	162.1
Tome <i>et al.</i> . [31]	65.0	73.5	76.8	86.4	86.3	110.7	68.9	74.8	110.2	173.9	85.0	85.8	86.3	71.4	73.1	88.4
Rogez <i>et al.</i> [22]	76.2	80.2	75.8	83.3	92.2	105.7	79.0	71.7	105.9	127.1	88.0	83.7	86.6	64.9	84.0	87.7
Tekin <i>et al.</i> [30]	54.2	61.4	60.2	61.2	79.4	78.3	63.1	81.6	70.1	107.3	69.3	70.3	74.3	51.8	74.3	69.7
Martinez <i>et al.</i> [13]	53.3	60.8	62.9	62.7	86.4	82.4	57.8	58.7	81.9	99.8	69.1	63.9	67.1	50.9	54.8	67.5
Fang <i>et al.</i> [7]	50.1	54.3	57.0	57.1	66.6	73.3	53.4	55.7	72.8	88.6	60.3	57.7	62.7	47.5	50.6	60.4
Pavlakos <i>et al.</i> [19]	48.5	54.4	54.4	52.0	59.4	65.3	49.9	52.9	65.8	71.1	56.6	52.9	60.9	44.7	47.8	56.2
Sáráandi <i>et al.</i> [24]	51.2	58.7	51.7	53.4	56.8	59.3	50.7	52.6	65.5	73.2	56.8	51.4	56.6	47.0	42.4	55.8
Sun <i>et al.</i> [28]	47.5	47.7	49.5	50.2	51.4	55.8	43.8	46.4	58.9	65.7	49.4	47.8	49.0	38.9	43.8	49.6
Ours	34.4	42.4	36.6	42.1	38.2	39.8	34.7	40.2	45.6	60.8	39.0	42.6	42.0	29.8	31.7	39.9
Protocol #2	Direct	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SittingD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Nie <i>et al.</i> [17]	90.1	88.2	85.7	95.6	103.9	92.4	90.4	117.9	136.4	98.5	103.0	94.4	86.0	90.6	89.5	97.5
Chen <i>et al.</i> . [4]	53.3	46.8	58.6	61.2	56.0	58.1	41.4	48.9	55.6	73.4	60.3	45.0	76.1	62.2	51.1	57.5
Martinez <i>et al.</i> [13]	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	69.4	49.2	45.0	49.5	38.0	43.1	47.7
Fang <i>et al.</i> [7]	38.2	41.7	43.7	44.9	48.5	55.3	40.2	38.2	54.5	64.4	47.2	44.3	47.3	36.7	41.7	45.7
Pavlakos <i>et al.</i> . [19]	34.7	39.8	41.8	38.6	42.5	47.5	38.0	36.6	50.7	56.8	42.6	39.6	43.9	32.1	36.5	41.8
Yang <i>et al.</i> [35]	26.9	30.9	36.3	39.9	43.9	47.4	28.8	29.4	36.9	58.4	41.5	30.5	29.5	42.5	32.2	37.7
Ours	29.1	34.9	29.9	32.6	31.2	32.3	27.0	33.3	37.6	45.9	32.2	31.5	34.5	22.9	25.9	32.1
PA MPJPE	Direct	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SittingD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Yasin <i>et al.</i> [36]	88.4	72.5	108.5	110.2	97.1	81.6	107.2	119.0	170.8	108.2	142.5	86.9	92.1	165.7	102.0	108.3
Sun <i>et al.</i> [28]	36.9	36.2	40.6	40.4	41.9	34.9	35.7	50.1	59.4	40.4	44.9	39.0	30.8	39.8	36.7	40.6
Dabral <i>et al.</i> [6]	28.0	30.7	39.1	34.4	37.1	44.8	28.9	32.2	39.3	60.6	39.3	31.1	37.8	25.3	28.4	36.3
Ours	21.6	27.0	29.7	28.3	27.3	32.1	23.5	30.3	30.0	37.7	30.1	25.3	34.2	19.2	23.2	27.9

Table 1. Quantitative comparisons of the mean per-joint position error (MPJPE) on Human3.6M [9] under Protocol #1 and Protocol #2, as well as using PA MPJPE as the evaluation metric. Similar to most of the competing methods (e.g., [28, 19, 35, 6, 30, 7]), our models were trained on the Human3.6M dataset and used also the extra MPII 2D pose dataset [1].

works [36, 28, 6], the pose prediction was further aligned with the groundtruth via a rigid transformation. The resulting MPJPE is termed as *Procrustes Aligned* (PA) MPJPE.

4.2. Results and comparisons

Human3.6M. We compare our method against state-of-the-art under three protocols, and the quantitative results are reported in Table 1. As can be seen, our method outperforms all competing methods on all action subjects for the protocols used. It is worth mentioning that our approach makes considerable improvements on some challenging actions for 3D pose estimation such as *Sitting* and *Sitting Down*. Thanks to HEMlets learning, our method demonstrates a clear advantage for handling complicated poses.

With a simple network architecture and little parameter tuning, we produce the most competitive results compared to previous works with carefully designed networks powered by e.g., adversarial training schemes or prior knowledge. On average, we improve the 3D pose prediction accuracy by 20% than that reported in Sun *et al.* [28] under Protocol #1. We also report our performance using PA MPJPE as the evaluation metric, and compare with these methods that make use of S9 as additional training data. We still outperform all of them across all action subjects, even *without* utilizing S9 for training.

Approach	Walking			Jogging			Avg
	S1	S2	S3	S1	S2	S3	
Simo-Serra <i>et al.</i> [27]	65.1	48.6	73.5	74.2	46.6	32.2	56.7
Moreno-Noguer <i>et al.</i> [15]	19.7	13.0	24.9	39.7	20.0	21.0	26.9
Martinez <i>et al.</i> [13]	19.7	17.4	46.8	26.9	18.2	18.6	24.6
Fang <i>et al.</i> [7]	19.4	16.8	37.4	30.4	17.6	16.3	22.9
Pavlakos <i>et al.</i> [19]	18.8	12.7	29.2	23.5	15.4	14.5	18.3
Ours	13.5	9.9	17.1	24.5	14.8	14.4	15.2

Table 2. Detailed results on the validation set of HumanEva-I [14].

HumanEva-I. With the same network architecture where *only* the HumanEva-I dataset is used for training, our results are reported in Table 2 under the popular protocol [27, 15, 13, 7, 19]. Different from these approaches [19, 15, 13, 7] which used extra 2D datasets (e.g., MPII) or pre-trained 2D detectors (e.g., CPM [34]), our method still outperforms previous approaches.

MPI-INF-3DHP. We evaluate our method on the MPI-INF-3DHP dataset using two metrics, the PCK and AUC. The results are generated by the model we trained for Human3.6M. In Table 3, we compare with three recent methods which are not trained on this dataset. Our result of “Studio GS” is one percentage lower than [19]. But our method outperforms all these methods with particularly large mar-

Approach	Studio	Studio	Outdoor	All	All
	GS	no GS			
	3DPCK	3DPCK	3DPCK	3DPCK	AUC
Mehta <i>et al.</i> [14]	70.8	62.3	58.8	64.7	31.7
Zhou <i>et al.</i> [37]	71.1	64.7	72.7	69.2	32.5
Pavlakos <i>et al.</i> [19]	76.5	63.1	77.5	71.9	35.3
ours	75.6	71.3	80.3	75.3	38.0

Table 3. Detailed results on the test set of MPI-INF-3DHP [14]. No training data from this dataset was used to train our model.

Method	Supervision	H3.6M #1	H3.6M #1*
Baseline	\mathcal{L}_λ^{3D}	47.1	55.3
w/ 2D heatmaps	$\mathcal{L}_\lambda^{3D} + \mathcal{L}^{2D}$	44.2	49.9
w/ HEMlets	$\mathcal{L}_\lambda^{3D} + \mathcal{L}^{HEM}$	42.6	46.0
Full	$\mathcal{L}_\lambda^{3D} + \mathcal{L}^{HEM} + \mathcal{L}^{2D}$	39.9	45.1

Table 4. Ablative study on the effects of alternative intermediate supervision evaluated on Human3.6M using Protocol #1. The last column * reports the results using only the Human3.6M dataset for training (without using the extra MPII 2D pose dataset).

gins for the ‘‘Outdoor’’ and ‘‘Studio no GS’’ sequences.

4.3. Ablation study

We study the influence on the final estimation performance of different choices made in our network design and the training procedure.

Alternative intermediate supervision. First, We examine the effectiveness of using HEMlets supervision. We evaluate the model trained without any intermediate supervision (Baseline), with 2D heatmap supervision only, with HEMlets supervision only, and with both 2D heatmap supervision and HEMlets supervision (Full). All of these design variants are evaluated with the same experimental setting (including training data, network architecture and \mathcal{L}_λ^{3D} loss definition) under Protocol #1 on Human3.6M.

The detailed results are presented in Table 4. Using 2D heatmaps supervision for training, the prediction error is reduced by 3.0mm compared to the baseline. The HEMlets supervision provided 1.7mm lower mean error compared to the 2D heatmaps supervision. This validates the effectiveness of the intermediate supervision. By combining all these choices, our approach using HEMlets with 2D heatmap supervision achieves the lowest error. Without using the extra MPII 2D pose dataset, we repeated this study. Similar conclusions can still be drawn. But the gap between w/ HEMlets (excluding \mathcal{L}^{2D} , 46.0mm) and Full (45.1mm) shrinks, suggesting the strength of the HEMlets representation in encoding both 2D and (local) 3D information.

To further illustrate the effectiveness of HEMlets representation, we provide a visual comparison in Fig. 4. Though the 2D joint errors of the two estimations are quite close, the

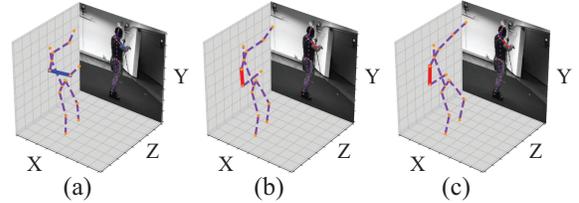


Figure 4. An example image with the detected joints overlaid and shown from a novel view, using different methods: (a) $\mathcal{L}_\lambda^{3D} + \mathcal{L}^{2D}$ (2D error: 15.2; 3D joint error: **81.3mm**). (b) $\mathcal{L}_\lambda^{3D} + \mathcal{L}^{2D} + \mathcal{L}^{HEM}$ (2D error: 13.0; 3D error: **41.2mm**). (c) Ground-truth. HEMlets learning helps fixing local part errors, see blue in (a) vs. red in (b).

method with HEMlets learning significantly improves the 3D joint estimation result and fixes the gross limb errors.

Regarding the runtime, tested on a NVIDIA GTX 1080 GPU, our full model (with a total parameter number of 47.7M) takes 13.3ms for a single forward inference, while the baseline model (with 34.3M parameters) takes 8.5ms.

Variants of HEMlets. We next experimented with some variants of HEMlets on Human3.6M and MPII 2D pose datasets. In the first variant, we use five-state heatmaps, referred to as *5s-HEM*, where the child joint is placed to different layers of the heatmaps according to the angle of the associated skeletal part with respect to the imaging plane. Specifically, we define the five states corresponding to the $(-90^\circ, -60^\circ)$, $(-60^\circ, -30^\circ)$, $(-30^\circ, 30^\circ)$, $(30^\circ, 60^\circ)$ and $(60^\circ, 90^\circ)$ range, respectively. In the second variant, we place a pair of joints in the negative and positive polarity heatmaps respectively according to their depth ordering (i.e., the closer/farther joint will appear in the positive/negative polarity heatmap. If their depths are roughly the same, they are co-located in the zero polarity heatmap. We refer to this variant as *2s-HEM*. We trained 5s-HEM, 2s-HEM and HEMlets with the Human3.6M dataset only. A comparison on the validation loss is given in Fig. 6. The other two variants produce inferior convergence compared to HEMlets under the same experiment setting.

Augmenting datasets. Many state-of-the-art approaches use a mixed training strategy for 3D human pose estimation. In addition to exploiting Human3.6M and MPII datasets, we study the effect of using augmenting datasets such as Ordinal [19] and FBI [25] for training. Firstly, we adapt the annotations of Ordinal and FBI datasets to the required form of HEMlets. Then we train our model using different combinations of these additional datasets. The comparisons on the MPI-INF-3DHP dataset [14] are reported in Table 5. We find augmenting datasets slightly increase the 3DPCK score for the trained model. Interestingly, training with FBI annotations attains a better 3DPCK score than Ordinal annotations. We suspect this is due to the amount of manual annotation errors related to different annotation schemes.

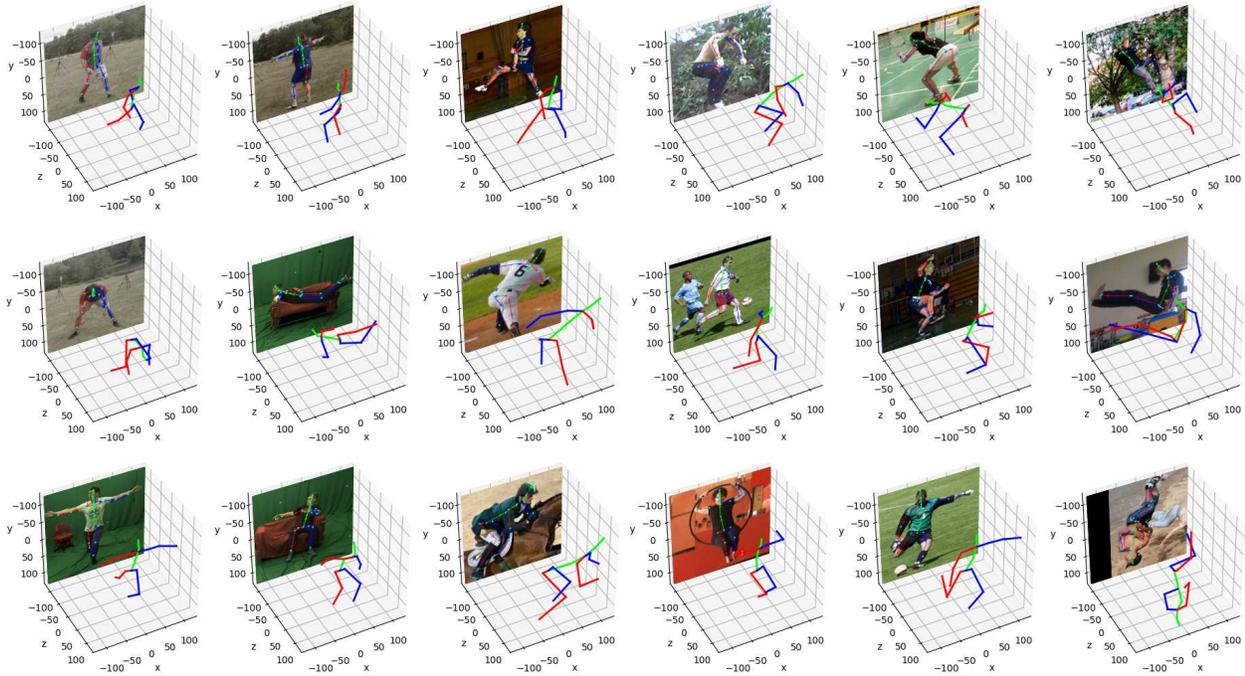


Figure 5. Qualitative results on different validation datasets: the first two columns are from the test dataset of 3DHP [14]. The other columns are from Leeds Sports Pose (LSP) [10]. Our approach produces visually correct results even on challenging poses (last column).

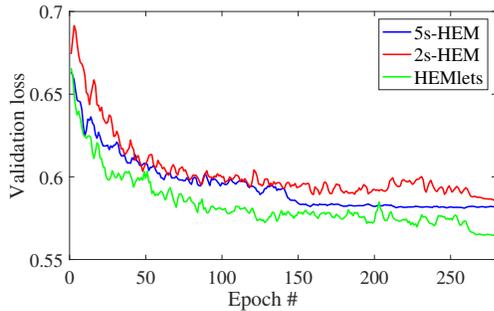


Figure 6. The validation loss of 5s-HEM, 2s-HEM and HEMlets, respectively. All are trained with the Human3.6M dataset.

Generalization. For an evaluation of in-the-wild images from Leeds Sports Pose (LSP) [10] and the validation set of MPI-INF-3DHP [14], we list some visual results predicted by our approach. As shown in Fig. 5, even for challenging data (e.g., self-occlusion, upside-down), our method yields visually correct pose estimations for these images.

5. Conclusion

In this paper, we proposed a simple and highly effective HEMlets-based 3D pose estimation method from a single color image. HEMlets is an easy-to-learn intermediate representation encoding the relative forward-or-backward depth relation for each skeletal part’s joints, together with their spatial co-location likelihoods. It is proved very helpful to bridge the input 2D image and the output 3D pose

Dataset	3DPCK
Base	75.3
w/ Ordinal [19]	76.1
w/ FBI [25]	76.9
w/ FBI [25] + Ordinal [19]	76.5

Table 5. Evaluation of 3DPCK scores by adding different augmenting datasets that provide relative depth ordering annotations. Base denotes using the base datasets (Human3.6M and MPII).

in the learning procedure. We demonstrated the effectiveness of the proposed method tested over the standard benchmarks, yielding a relative accuracy improvement of about 20% over the best-of-grade method on the Human3.6M benchmark. Good generalization ability is also witnessed for the presented approach. We believe the proposed HEMlets idea is actually general, which may potentially benefit other 3D regression problems e.g., scene depth estimation.

Acknowledgements. This work is supported in part by the National Natural Science Foundation of China (Grant No.: 61771201), the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (Grant No.: 2017ZT07X183), the Pearl River Talent Recruitment Program Innovative and Entrepreneurial Teams in 2017 (Grant No.: 2017ZT07X152), the Shenzhen Fundamental Research Fund (Grants No.: KQTD2015033114415450 and ZDSYS201707251409055), and Department of Science and Technology of Guangdong Province Fund (2018B030338001).

References

- [1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3686–3693, 2014.
- [2] Liefeng Bo and Cristian Sminchisescu. Twin gaussian processes for structured prediction. *International Journal of Computer Vision*, 87(1-2):28, 2010.
- [3] Lubomir Bourdev and Jitendra Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1365–1372, 2009.
- [4] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation = 2d pose estimation + matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7035–7043, 2017.
- [5] Wenzheng Chen, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Synthesizing training images for boosting human 3d pose estimation. In *3D Vision (3DV)*, pages 479–488. IEEE, 2016.
- [6] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3d human pose from structure and motion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 668–683. Springer, 2018.
- [7] Hao-Shu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. Learning pose grammar to encode human body configuration for 3d pose estimation. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.
- [8] Golnaz Ghiasi and Charless C Fowlkes. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2385–2392, 2014.
- [9] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014.
- [10] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *British Machine Vision Conference (BMVC)*, 2010.
- [11] Lipeng Ke, Ming-Ching Chang, Honggang Qi, and Siwei Lyu. Multi-scale structure-aware network for human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 713–728. Springer, 2018.
- [12] Sijin Li and Antoni 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.
- [13] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2640–2649, 2017.
- [14] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved CNN supervision. In *3D Vision (3DV)*, pages 506–516. IEEE, 2017.
- [15] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1561–1570, 2017.
- [16] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 483–499. Springer, 2016.
- [17] Bruce Xiaohan Nie, Ping Wei, and Song-Chun Zhu. Monocular 3d human pose estimation by predicting depth on joints. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3467–3475, 2017.
- [18] Sunghoon Park, Jihye Hwang, and Nojun Kwak. 3d human pose estimation using convolutional neural networks with 2d pose information. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 156–169. Springer, 2016.
- [19] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7307–7316, 2018.
- [20] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1263–1272, 2017.
- [21] Gerard Pons-Moll, David J Fleet, and Bodo Rosenhahn. Posebits for monocular human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2337–2344, 2014.
- [22] Grégory Rogez and Cordelia Schmid. Mocap-guided data augmentation for 3d pose estimation in the wild. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 3108–3116, 2016.
- [23] Matteo Ruggero Ronchi, Oisín Mac Aodha, Robert Eng, and Pietro Perona. It’s all relative: Monocular 3d human pose estimation from weakly supervised data. In *British Machine Vision Conference (BMVC)*.
- [24] István Sáráncsi, Timm Linder, Kai O Arras, and Bastian Leibe. How robust is 3d human pose estimation to occlusion? In *IROS Workshop - Robotic Co-workers 4.0*, 2018.
- [25] Yulong Shi, Xiaoguang Han, Nianjuan Jiang, Kun Zhou, Kui Jia, and Jiangbo Lu. FBI-pose: Towards bridging the gap between 2d images and 3d human poses using forward-or-backward information. *arXiv preprint arXiv:1806.09241*, 2018.
- [26] Leonid Sigal, Alexandru O Balan, and Michael J Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(1-2):4, 2010.
- [27] Edgar Simo-Serra, Ariadna Quattoni, Carme Torras, and Francesc Moreno-Noguer. A joint model for 2d and 3d

- pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3634–3641, 2013.
- [28] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 529–545. Springer, 2018.
- [29] Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Structured prediction of 3d human pose with deep neural networks. In *British Machine Vision Conference (BMVC)*, 2016.
- [30] Bugra Tekin, Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3941–3950, 2017.
- [31] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2500–2509, 2017.
- [32] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4627–4635, 2017.
- [33] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7782–7791, 2019.
- [34] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4732, 2016.
- [35] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5255–5264, 2018.
- [36] Hashim Yasin, Umar Iqbal, Bjorn Kruger, Andreas Weber, and Juergen Gall. A dual-source approach for 3d pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4948–4956, 2016.
- [37] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 398–407, 2017.