

Compositional Human Pose Regression

Xiao Sun¹, Jiaxiang Shang¹, Shuang Liang^{2*}, Yichen Wei¹

¹Microsoft Research, ² Tongji University

{xias, yichenw}@microsoft.com, jiaxiang.shang@gmail.com, shuangliang@tongji.edu.cn

Abstract

Regression based methods are not performing as well as detection based methods for human pose estimation. A central problem is that the structural information in the pose is not well exploited in the previous regression methods. In this work, we propose a structure-aware regression approach. It adopts a reparameterized pose representation using bones instead of joints. It exploits the joint connection structure to define a compositional loss function that encodes the long range interactions in the pose. It is simple, effective, and general for both 2D and 3D pose estimation in a unified setting. Comprehensive evaluation validates the effectiveness of our approach. It significantly advances the state-of-the-art on Human3.6M [20] and is competitive with state-of-the-art results on MPII [3].

1. Introduction

Human pose estimation has been extensively studied for both 3D [20] and 2D [3]. Recently, deep convolutional neural networks (CNNs) have achieved significant progresses.

Existing approaches fall into two categories: detection based and regression based. Detection based methods generate a likelihood heat map for each joint and locate the joint as the point with the maximum value in the map. These heat maps are usually noisy and multi-mode. The ambiguity is reduced by exploiting the dependence between the joints in various ways. A prevalent family of state-of-the-art methods [11, 6, 31, 5, 47, 18] adopt a multi-stage architecture, where the output of the previous stage is used as input to enhance the learning of the next stage. These methods are dominant for 2D pose estimation [1]. However, they do not easily generalize to 3D pose estimation, because the 3D heat maps are too demanding for memory and computation.

Regression based methods directly map the input image to the output joints. They directly target at the task and they are general for both 3D and 2D pose estimation. Nevertheless, they are not performing as well as detection based

methods. As an evidence, only one method [6] in the 2D pose benchmark [1] is regression based. While they are widely used for 3D pose estimation [54, 30, 28, 43, 24, 42, 33], the performance is not satisfactory. A central problem is that they simply minimize the per-joint location errors *independently* but ignore the internal structures of the pose. In other words, joint dependence is not well exploited.

In this work, we propose a structure-aware approach, called *compositional pose regression*. It is based on two ideas. First, it uses bones instead of joints as pose representation, because the bones are more primitive, more stable, and easier to learn than joints. Second, it exploits the joint connection structure to define a *compositional loss function* that encodes long range interactions between the bones.

The approach is simple, effective and efficient. It only re-parameterizes the pose representation, which is the network output, and enhances the loss function, which relates the output to ground truth. It does not alter other algorithm design choices and is compatible with such choices, such as network architecture. It can be easily adapted into any existing regression approaches with little overhead for memory and computation, in both training and inference.

The approach is general and can be used for both 3D and 2D pose regression, indistinguishably. Moreover, 2D and 3D data can be easily mixed simultaneously in the training. *For the first time, it is shown that such directly mixed learning is effective.* This property makes our approach different from all existing ones that target at either 3D or 2D task.

The effectiveness of our approach is validated by comprehensive evaluation with a few new metrics, rigorous ablation study and comparison with state-of-the-art on both 3D and 2D benchmarks. Specifically, it advances the state-of-the-art on 3D Human3.6M dataset [20] by a large margin and achieves a record of 59.1 mm average joint error, about 12% relatively better than state-of-the-art. On 2D MPII dataset [3, 1], it achieves 86.4% (PCKh 0.5). It is the best-performing regression based method and on par with the state-of-the-art detection based methods. As a by-product, our approach generates high quality 3D poses for in the wild images, indicating the potential of our approach for transfer learning of 3D pose estimation in the wild.

*Corresponding author.

2. Related Work

Human pose estimation has been extensively studied for years. A complete review is beyond the scope of this work. We refer the readers to [29, 41] for a detailed survey.

The previous works are reviewed from two perspectives related to this work. First is how to exploit the joint dependency for 3D and 2D pose estimation. Second is how to exploit “in the wild” 2D data for 3D pose estimation.

3D Pose Estimation Some methods use two separate steps. They first perform 2D joint prediction and then reconstruct the 3D pose via optimization or search. There is no end-to-end learning. Zhou et al. [56] combines uncertainty maps of the 2D joints location and a sparsity-driven 3D geometric prior to infer the 3D joint location via an EM algorithm. Chen et al. [7] searches a large 3D pose library and uses the estimated 2D pose as query. Bogo et al. [4] fit a recently published statistical body shape model [27] to the 2D joints. Jahangiri et al. [21] generates multiple hypotheses from 2D joints using a novel generative model.

Some methods implicitly learn the pose structure from data. Tekin et al. [42] represents the 3D pose with an over-complete dictionary. A high-dimensional latent pose representation is learned to account for joint dependencies. Pavlakos et al. [34] extends the Hourglass [31] framework from 2D to 3D. A coarse-to-fine approach is used to address the large dimensionality increase. Li et al. [24] uses an image-pose embedding sub-network to regularize the 3D pose prediction.

Above works do not use prior knowledge in 3D model. Such prior knowledge is firstly used in [54, 55] by embedding a kinematic model layer into deep neural networks and estimating model parameters instead of joints. The geometric structure is better preserved. Yet, the kinematic model parameterization is highly nonlinear and its optimization in deep networks is hard. Also, the methods are limited for a fully specified kinematic model (fixed bone length, known scale). They do not generalize to 2D pose estimation, where a good 2D kinematic model does not exist.

2D Pose Estimation Before the deep learning era, many methods use graphical models to represent the structures in the joints. Pictorial structure model [13] is one of the earliest. There is a lot of extensions [23, 50, 36, 35, 51, 26, 9]. Pose estimation is formulated as inference problems on the graph. A common drawback is that the inference is usually complex, slow, and hard to integrate with deep networks.

Recently, the graphical models have been integrated into deep networks in various ways. Tompson et al. [46] firstly combine a convolutional network with a graphical model for human pose estimation. Ouyang et al. [32] joints feature extraction, part deformation handling, occlusion handling and classification all into deep learning framework. Chu et al. [10] introduce a geometrical transform kernels in CNN framework that can pass informations between differ-

ent joint heat maps. Both features and their relationships are jointly learned in a end-to-end learning system. Yang et al. [49] combine deep CNNs with the expressive deformable mixture of parts to regularize the output.

Another category of methods use a multi-stage architecture [11, 6, 31, 5, 47, 18, 14]. The results of the previous stage are used as inputs to enhance or regularize the learning of the next stage. Newell et al. [31] introduce an Stacked Hourglass architecture that better capture the various spatial relationships associated with the body. Chu et al. [11] further extend [31] with a multi-context attention mechanism. Bulat et al. [5] propose a detection-followed-by-regression CNN cascade. Wei et al. [47] design a sequential architecture composed of convolutional networks that directly operate on belief maps from previous stages. Gkioxari et al. [14] predict joint heat maps sequentially and conditionally according to their difficulties. All such methods learn the joint dependency from data, implicitly.

Different to all above 3D and 2D methods, our approach explicitly exploits the joint connection structure in the pose. It does not make further assumptions and does not involve complex algorithm design. It only changes the pose representation and enhances the loss function. It is simple, effective, and can be combined with existing techniques.

Leveraging in the wild 2D data for 3D pose estimation 3D pose capturing is difficult. The largest 3D human pose dataset Human3.6M [20] is still limited in that the subjects, the environment, and the poses have limited complexity and variations. Models trained on such data do not generalize well to other domains, such as in the wild images.

In contrast, in the wild images and 2D pose annotation are abundant. Many works leverage the 2D data for 3D pose estimation. Most of them consist of two separate steps.

Some methods firstly generate the 2D pose results (joint locations or heat maps) and then use them as input for recovering the 3D pose. The information in the 2D images is discarded in the second step. Bogo et al. [4] first use DeepCut [38] to generate 2D joint location, then fit with a 3D body shape model. Moreno et al. [30] use CPM [47] to detect 2D position of human joints, and then use these observations to infer 3D pose via distance matrix regression. Zhou et al. [57] use Hourglass [31] to generate 2D joint heat maps and then coupled with a geometric prior and Jahangiri et al. [21] also use Hourglass to predict 2D joint heat maps and then infer multiple 3D hypotheses from them. Wu et al. [48] propose 3D interpreter network that sequentially estimates 2D keypoint heat maps and 3D object structure.

Some methods firstly train the deep network model on 2D data and fine-tune the model on 3D data. The information in 2D data is partially retained by the pre-training, but not fully exploited as the second fine-tuning step cannot use 2D data. Pavlakos et al. [34] extends Hourglass [31] model for 3D volumetric prediction. 2D heat maps are used as in-

intermediate supervision. Tome et al. [44] extends CPM [47] to 3D by adding a probabilistic 3D pose model to the CPM.

Some methods train both 2D and 3D pose networks simultaneously by sharing intermediate CNN features [28, 33]. Yet, they use separate networks for 2D and 3D tasks.

Unlike the above methods, our approach treats the 2D and 3D data in the same way and combine them in a unified training framework. The abundant information in the 2D data is fully exploited during training. As a result, our method achieves strong performance on both 3D and 2D benchmarks. As a by-product, it generates plausible and convincing 3D pose results for in the wild images.

Some methods use synthetic datasets which are generated from deforming a human template model with known ground truth [8, 40]. These methods are complementary to the others as they focus on data augmentation.

3. Compositional Pose Regression

Given an image of a person, the pose estimation problem is to obtain the 2D (or 3D) position of all the K joints, $\mathcal{J} = \{\mathbf{J}_k | k = 1, \dots, K\}$. Typically, the coordinate unit is *pixel* for 2D and *millimeter (mm)* for 3D.

Without loss of generality, the joints are defined with respect to a *constant origin point* in the image coordinate system. For convenience, let the origin be \mathbf{J}_0 . Specifically, for 2D pose estimation, it is the top-left point of the image. For 3D pose estimation, it is the ground truth pelvis joint [54, 33].

For regression learning, normalization is necessary to compensate for the differences in magnitude of the variables. We use the standard normalization by subtraction of mean and division of standard deviation. For a variable var , it is normalized as

$$\tilde{var} = N(var) = \frac{var - \text{mean}(var^{gt})}{\text{std}(var^{gt})}. \quad (1)$$

The inverse function for *unnormailization* is

$$var = N^{-1}(\tilde{var}) = \tilde{var} \cdot \text{std}(var^{gt}) + \text{mean}(var^{gt}). \quad (2)$$

Note that both $\text{mean}(\cdot)$ and $\text{std}(\cdot)$ are constants and calculated from the ground truth training samples. The predicted output from the network is assumed already normalized. Both functions $N(\cdot)$ and $N^{-1}(\cdot)$ are parameter free and embedded in the network. For notation simplicity, we use \tilde{var} for $N(var)$.

3.1. Direct Joint Regression: A Baseline

Most previous regression based methods [6, 54, 33, 42, 43] directly minimize the squared difference of the predicted and ground truth joints. In experiments, we found that the absolute difference ($L1$ norm) performs better. In our *direct joint regression* baseline, the joint loss is

$$L(\mathcal{J}) = \sum_{k=1}^K \|\tilde{\mathbf{J}}_k - \tilde{\mathbf{J}}_k^{gt}\|_1. \quad (3)$$

Note that both the prediction and ground truth are normalized.

There is a clear drawback in loss Eq.(3). The joints are *independently* estimated. The joint correlation, or the internal structure in the pose, is not well exploited. For example, certain geometric constraints (*e.g.*, bone length is fixed) are not satisfied.

Previous works only evaluate the joint location accuracy. This is also limited because the internal structures in the pose are not well evaluated.

3.2. A Bone Based Representation

We show that a simple *reparameterization* of the pose is effective to address the above issues. As shown in Figure 1(left), a pose is structured as a tree. Without loss of generality, let pelvis be the the root joint \mathbf{J}_1 and tree edges be directed from the root to the end joints such as wrists and ankles. Let the function $\text{parent}(k)$ return the index of parent joint for k^{th} joint. For notation consistency, let the parent of the root joint \mathbf{J}_1 be the origin \mathbf{J}_0 , *i.e.*, $\text{parent}(1) = 0$.

Now, for k^{th} joint, we define its associated bone as a directed vector pointing from it to its parent,

$$\mathbf{B}_k = \mathbf{J}_{\text{parent}(k)} - \mathbf{J}_k. \quad (4)$$

The joints \mathcal{J} are defined in the global coordinate system. In contrast, bones $\mathcal{B} = \{\mathbf{B}_k | k = 1, \dots, K\}$ are more primitive and defined in the local coordinate systems. Representing the pose using bones brings several benefits.

Stability Bones are more stable than joints and easier to learn. Figure 1 (middle and right) shows that the standard deviation of bones is much smaller than that of their corresponding joints, especially for parts (ankle, wrist, head) far away from the root pelvis, in both 3D (Human 3.6M [20]) and 2D datasets (MPII [3]).

Geometric convenience Bones can encode the geometric structure and express the geometric constraints more easily than joints. For example, constraint of “*bone length is fixed*” involves one bone but two joints. Constraint of “*joint rotation angle is in limited range*” involves two bones but three joints. Such observations motivate us to propose new evaluation metrics for *geometric validity*, as elaborated in Section 5. Experiments show that bone based representation is better than joint based representation on such metrics.

Application convenience Many pose-driven applications only need the local bones instead of the global joints. For example, the local and relative “elbow to wrist” motion can sufficiently represent a “pointing” gesture that would be useful for certain human computer interaction scenarios.

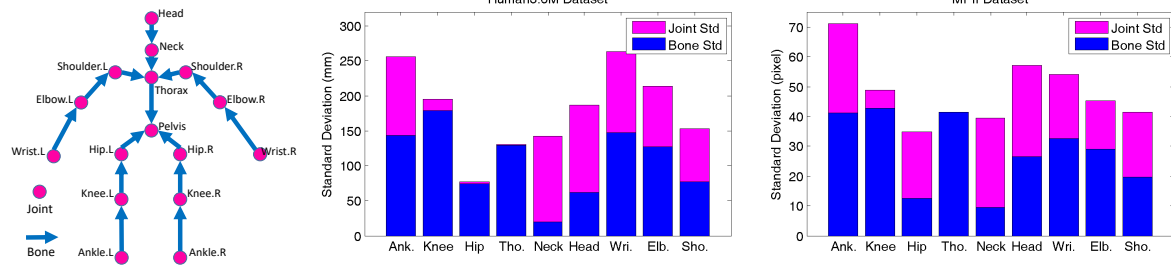


Figure 1. Left: a human pose is represented as either joints \mathcal{J} or bones \mathcal{B} . Middle/Right: standard deviations of bones and joints for the 3D Human3.6M dataset [20] and 2D MPII dataset [3].

3.3. Compositional Loss Function

Similar to the joint loss in Eq. (3), bones can be learnt by minimizing the bone loss function

$$L(\mathcal{B}) = \sum_{k=1}^K \|\tilde{\mathbf{B}}_k - \tilde{\mathbf{B}}_k^{gt}\|_1. \quad (5)$$

However, there is a clear drawback in this loss. As the bones are local and *independently* estimated in Eq. (5), the errors in the individual bone predictions would propagate along the skeleton and accumulate into large errors for joints at the far end. For example, in order to predict \mathbf{J}_{wrist} , we need to concatenate $\mathbf{B}_{wrist}, \mathbf{B}_{elbow}, \dots, \mathbf{B}_{pelvis}$. Errors in these bones will accumulate and affect the accuracy of \mathbf{J}_{wrist} in a random manner.

To address the problem, long range objectives should be considered in the loss. Long range errors should be balanced over the intermediate bones. In this way, bones are *jointly* optimized. Specifically, let \mathbf{J}_u and \mathbf{J}_v be two arbitrary joints. Suppose the path from \mathbf{J}_u to \mathbf{J}_v along the skeleton tree has M joints. Let the function $I(m)$ return the index of the m^{th} joint on the path, e.g., $I(1) = u, I(M) = v$. Note that M and $I(*)$ are constants but depend on u and v . Such dependence is omitted in the notations for clarity.

The long range, relative joint position $\Delta\mathbf{J}_{u,v}$ is the summation of the bones along the path, as

$$\begin{aligned} \Delta\mathbf{J}_{u,v} &= \sum_{m=1}^{M-1} \mathbf{J}_{I(m+1)} - \mathbf{J}_{I(m)} \\ &= \sum_{m=1}^{M-1} \text{sgn}(\text{parent}(I(m)), I(m+1)) \cdot N^{-1}(\tilde{\mathbf{B}}_{I(m)}). \end{aligned} \quad (6)$$

The function $\text{sgn}(*, *)$ indicates whether the bone $\mathbf{B}_{I(m)}$ direction is along the path direction. It returns 1 when $\text{parent}(I(m)) = I(m+1)$ and -1 otherwise. Note that the network predicted bone $\tilde{\mathbf{B}}(*)$ is normalized, as in Eq. (6). It is unnormalized via Eq. (2) before summation.

Eq.(6) is differentiable with respect to the bones. It is efficient and has no free parameters. It is implemented as a special *compositional* layer in the neural networks.

The ground truth relative position is

$$\Delta\mathbf{J}_{u,v}^{gt} = \mathbf{J}_u^{gt} - \mathbf{J}_v^{gt}. \quad (7)$$

Then, given a joint pair set \mathcal{P} , the *compositional loss function* is defined as

$$L(\mathcal{B}, \mathcal{P}) = \sum_{(u,v) \in \mathcal{P}} \|\tilde{\Delta\mathbf{J}}_{u,v} - \tilde{\Delta\mathbf{J}}_{u,v}^{gt}\|_1. \quad (8)$$

In this way, every joint pair (u, v) constrains the bones along the path from u to v . Each bone is constrained by multiple paths given a large number of joint pairs. The errors are better balanced over the bones during learning.

The joint pair set \mathcal{P} can be arbitrary. To validate the effectiveness of Eq.(8), we test four variants:

- $\mathcal{P}_{joint} = \{(u, 0) | u = 1, \dots, K\}$. It only considers the global joint locations. It is similar to joint loss Eq.(3).
- $\mathcal{P}_{bone} = \{(u, \text{parent}(u)) | u = 1, \dots, K\}$. It only considers the bones. It degenerates to the bone loss Eq.(5).
- $\mathcal{P}_{both} = \mathcal{P}_{joint} \cup \mathcal{P}_{bone}$. It combines the above two and verifies whether Eq.(8) is effective.
- $\mathcal{P}_{all} = \{(u, v) | u < v, u, v = 1, \dots, K\}$. It contains all joint pairs. The pose structure is fully exploited.

4. Unified 2D and 3D Pose Regression

All the notations and equations in Section 3 are applicable for both 3D and 2D pose estimation in the same way. The output pose dimension is either $3K$ or $2K$.

Training using mixed 3D and 2D data is straightforward. All the variables, such as joint \mathbf{J} , bone \mathbf{B} , and relative joint position $\Delta\mathbf{J}_{u,v}$, are decomposed into xy part and z part.

The loss functions can be similarly decomposed. For example, for *compositional loss function* Eq.(8), we have

$$L(\mathcal{B}, \mathcal{P}) = L_{xy}(\mathcal{B}, \mathcal{P}) + L_z(\mathcal{B}, \mathcal{P}). \quad (9)$$

The xy term $L_{xy}(*,*)$ is always valid for both 3D and 2D samples. The z term $L_z(*,*)$ is only computed for 3D samples and set to 0 for 2D samples. In the latter case, no gradient is back-propagated from $L_z(*,*)$.

Note that the xy part and z part variables have different dimensions. xy is in image coordinate frame and the unit is in pixel. z is in camera coordinate frame and the unit is metric (millimeters in our case). This is no problem. During training, they are appropriately normalized (Eq.(8), Eq.(1)) or unnormalized (Eq.(6), Eq.(2)). During inference, in order to recover the 3D metric coordinates, the xy part is back-projected into camera space using known camera intrinsic parameters and a perspective projection model.

Training We use the state-of-the-art ResNet-50 [16]. The model is pre-trained on ImageNet classification dataset [12]. The last fully connected layer is then modified to output $3K$ (or $2K$) coordinates and the model is fine-tuned on our target task and data. The training is the same for all the tasks (3D, 2D, mixed). SGD is used for optimization. There are 25 epoches. The base learning rate is 0.03. It drops to 0.003 after 10 epoches and 0.0003 after another 10 epoches. Mini-batch size is 64. Two GPUs are used. Weight decay is 0.0002. Momentum is 0.9. Batch-normalization [19] is used. Implementation is in Caffe [22].

Data Processing and Augmentation The input image is normalized to 224×224 . Data augmentation includes random translation($\pm 2\%$ of the image size), scale($\pm 25\%$), rotation(± 30 degrees) and flip. For MPII dataset, the training data are augmented by 20 times. For Human3.6M dataset, the training data are augmented by 4 times. For mixed 2D-3D task, each mini-batch consists of half 2D and half 3D samples, randomly sampled and shuffled.

5. Experiments

Our approach is evaluated on 3D and 2D human pose benchmarks. *Human3.6M* [20] is the largest 3D human pose benchmark. The dataset is captured in controlled environment. The image appearance of the subjects and the background is simple. Accurate 3D human joint locations are obtained from motion capture devices.

MPII [3] is the benchmark dataset for 2D human pose estimation. It includes about 25k images and 40k annotated 2D poses. 25k of them are for training and another 7k of the remaining are for testing. The images were collected from YouTube videos covering daily human activities with complex poses and image appearances.

5.1. Comprehensive Evaluation Metrics

For 3D human pose estimation, previous works [7, 44, 30, 57, 21, 28, 34, 52, 40, 4, 56, 43, 54] use the mean per joint position error (MPJPE). We call this metric *Joint Error*. Some works [52, 40, 7, 4, 30, 57] firstly align the predicted 3D pose and ground truth 3D pose with a rigid trans-

CNN prediction		loss function
Baseline	joints \mathcal{J}	$L(\mathcal{J})$, Eq.(3)
Ours (joint)	bones \mathcal{B}	$L(\mathcal{B}, \mathcal{P}_{joint})$, Eq.(8)
Ours (bone)		$L(\mathcal{B}, \mathcal{P}_{bone})$, Eq.(8)
Ours (both)		$L(\mathcal{B}, \mathcal{P}_{both})$, Eq.(8)
Ours (all)		$L(\mathcal{B}, \mathcal{P}_{all})$, Eq.(8)

Table 1. The baseline and four variants of our method.

formation using *Procrustes Analysis* [15] and then compute MPJPE. We call this metric *PA Joint Error*.

For 2D human pose estimation in MPII [3], Percentage of Correct Keypoints (PCK) metric is used for evaluation.

Above metrics only measures the accuracy of *absolute* joint location. They do not fully reflect the accuracy of internal structures in the pose. We propose three additional metrics for a comprehensive evaluation.

The first metric is the mean per bone position error, or *Bone Error*. It is similar to *Joint Error*, but measures the *relative* joint location accuracy. This metric is applicable for both 3D and 2D pose.

The next two are only for 3D pose as they measure the validity of 3D geometric constraints. Such metrics are important as violation of the constraints will cause physically infeasible 3D poses. Such errors are critical for certain applications such as 3D motion capture.

The second metric is the bone length standard deviation, or *Bone Std*. It measures the stability of bone length. For each bone, the standard deviation of its length is computed over all the testing samples of the same subject.

The third metric is the percentage of illegal joint angle, or *Illegal Angle*. It measures whether the rotation angles at a joint are physically feasible. We use the recent method and code in [2] to evaluate the legality of each predicted joint. Note that this metric is only for joints on the limbs and does not apply to those on the torso.

5.2. Experiments on 3D Pose of Human3.6M

For Human3.6M [20], there are two widely used evaluation protocols with different training and testing data split.

Protocol 1 Six subjects (S1, S5, S6, S7, S8, S9) are used in training. Evaluation is performed on every 64th frame of Subject 11's videos. It is used in [52, 40, 7, 30, 57]. *PA Joint Error* is used for evaluation.

Protocol 2 Five subjects (S1, S5, S6, S7, S8) are used for training. Evaluation is performed on every 64th frame of two subjects (S9, S11). It is used in [56, 43, 54, 7, 44, 30, 57, 21, 28, 34]. *Joint Error* is used for evaluation.

Ablation study. The *direct joint regression* baseline and four variants of our method are compared. They are briefly summarized in Table 1. As explained in Section 4, training can use additional 2D data (from MPII), optionally.

Training Data	Metric	Baseline	Ours (joint)	Ours(bone)	Ours (both)	Ours (all)
Human3.6M	Joint Error	102.2	103.3 \uparrow 1.1	104.6 \uparrow 2.4	95.2 \downarrow 7.0	92.4 \downarrow 9.8
	PA Joint Error	75.0	74.3 \downarrow 0.7	75.0 \downarrow 0.0	68.1 \downarrow 6.9	67.5 \downarrow 7.5
	Bone Error	65.5	63.5 \downarrow 2.0	62.3 \downarrow 3.2	59.1 \downarrow 6.4	58.4 \downarrow 7.1
	Bone Std	26.4	23.9 \downarrow 2.5	21.9 \downarrow 4.5	22.3 \downarrow 4.1	21.7 \downarrow 4.7
	Illegal Angle	3.7%	3.2% \downarrow 0.5	3.3% \downarrow 0.4	2.6% \downarrow 1.1	2.5% \downarrow 1.2
Human3.6M + MPII	Joint Error	64.2	62.9 \downarrow 1.3	63.8 \downarrow 0.4	60.7 \downarrow 3.5	59.1 \downarrow 5.1
	PA Joint Error	51.4	50.6 \downarrow 0.8	50.4 \downarrow 1.0	48.8 \downarrow 2.6	48.3 \downarrow 3.1
	Bone Error	49.5	49.3 \downarrow 0.2	47.4 \downarrow 2.1	47.2 \downarrow 2.3	47.1 \downarrow 2.4
	Bone Std	19.9	19.3 \downarrow 0.6	17.5 \downarrow 2.4	17.6 \downarrow 2.3	18.0 \downarrow 1.9

Table 2. Results of all methods under all evaluation metrics (the lower the better), with or without using MPII data in training. Note that the performance gain of all *Ours* methods relative to the *Baseline* method is shown in the subscript. The *Illegal Angle* metric for “Human3.6M+MPII” setting is not included because it is very good ($< 1\%$) for all methods.

Metric	Joint Error		PA Joint Error		Bone Error		Bone Std		Illegal Angle	
Method	BL	Ours (all)	BL	Ours (all)	BL	Ours (all)	BL	Ours (all)	BL	Ours (all)
Average	102.2	92.4 \downarrow 9.8	75.0	67.5 \downarrow 7.5	65.5	58.4 \downarrow 7.1	26.4	21.7 \downarrow 4.7	3.7%	2.5% \downarrow 1.2
Ankle(\rightarrow Knee)	94.5	88.5 \downarrow 6.0	81.5	75.8 \downarrow 5.7	81.2	74.1 \downarrow 7.1	32.9	32.0 \downarrow 0.9	-	-
Knee(\rightarrow Hip)	68.6	63.7 \downarrow 4.9	69.2	62.9 \downarrow 6.3	69.1	63.4 \downarrow 5.7	21.7	22.8 \uparrow 1.1	4.8%	3.8% \downarrow 1.0
Hip(\rightarrow Pelvis)	29.9	25.0 \downarrow 4.9	63.3	58.4 \downarrow 4.9	29.9	25.0 \downarrow 4.9	21.3	16.4 \downarrow 4.9	0.6%	0.6% \downarrow 0.0
Thorax(\rightarrow Pelvis)	97.2	90.1 \downarrow 7.1	30.7	28.1 \downarrow 2.6	97.2	90.1 \downarrow 7.1	28.0	26.7 \downarrow 1.3	-	-
Neck(\rightarrow Thorax)	104.3	96.4 \downarrow 7.9	36.7	35.5 \downarrow 1.2	22.2	22.9 \uparrow 0.7	12.4	11.7 \downarrow 0.7	2.2%	1.3% \downarrow 0.9
Head(\rightarrow Neck)	115.4	108.4 \downarrow 7.0	42.8	41.1 \downarrow 1.7	39.7	37.3 \downarrow 2.4	15.3	14.8 \downarrow 0.5	-	-
Wrist(\rightarrow Elbow)	181.9	163.0 \downarrow 18.9	130.2	115.2 \downarrow 15.0	102.6	89.0 \downarrow 13.6	40.6	30.6 \downarrow 10.0	-	-
Elbow(\rightarrow Shoulder)	168.8	146.9 \downarrow 21.9	115.8	97.6 \downarrow 18.2	96.9	81.4 \downarrow 15.5	27.6	21.4 \downarrow 6.2	8.5%	5.4% \downarrow 3.1
Shoulder(\rightarrow Thorax)	115.6	104.4 \downarrow 11.2	57.7	52.2 \downarrow 5.5	55.1	48.5 \downarrow 6.6	25.9	12.6 \downarrow 13.3	1.9%	0.8% \downarrow 1.1

Table 3. Detailed results on all joints for *Baseline* (BL) and *Ours (all)* methods, only trained on Human3.6M data (top half in Table 2). The relative performance gain is shown in the subscript. Note that the left most column shows the names for both the joint (and the bone).

We therefore tested two sets of training data: 1) only Human3.6M; 2) Human3.6M plus MPII.

Table 2 reports the results under Protocol 2, which is more commonly used. We observe several conclusions.

Using 2D data is effective. All metrics are significantly improved after using MPII data. For example, joint error is reduced from 102.2 to 64.2. This improvement should originate from the better learnt feature from the abundant 2D data. See the contemporary work [53] for more discussions. Note that adding 2D data in this work is simple and not considered as a main contribution. Rather, it is considered as a baseline to validate our regression approach.

Bone representation is superior than joint representation. This can be observed by comparing *Baseline* with *Ours (joint)* and *Ours (bone)*. They are comparable because they use roughly the same amount of supervision signals in the training. The two variants of ours are better on nearly all the metrics, especially the geometric constraint based ones.

Compositional loss is effective. When the loss function becomes better (*Ours (both)* and *Ours (all)*), further improvement is observed. Specifically, when trained only on Human3.6M, *Ours (all)* improves the *Baseline* by 9.8 mm (relative 9.6%) on joint error, 7.5 mm (relative 10%) on PA joint error, 7.1 mm (relative 10.8%) on bone error, 4.7 mm (relative 17.8%) on bone std, and 1.2% (relative 32.4%) on

illegal angle.

Table 3 further reports the performance improvement from *Ours (all)* to *Baseline* on all the joints (bones). It shows several conclusions. First, *limb joints are harder than torso joints and upper limbs are harder than lower limbs*. This is consistent as Figure 1 (middle). It indicates that the variance is a good indicator of difficulty and a per-joint analysis is helpful in both algorithm design and evaluation. Second, *our method significantly improves the accuracy for all the joints*, especially the challenging ones like wrist, elbow and ankle. Figure 2 shows the results on a testing video sequence with challenging arm motions. Our result is much better and more stable.

Comparison with the state-of-the-art There are abundant previous works. They have different experiment settings and fall into three categories. They are compared to our method in Table 4, 5, and 6, respectively.

The comparison is not completely fair due to the differences in the training data (when extra data are used), the network architecture and implementation. Nevertheless, two common conclusions validate that *our approach is effective and sets the new state-of-the-art in all settings by a large margin*. First, *our baseline is strong*. It is simple but already improves the state-of-the-art, by 3.9 mm (relative 7%) in Table 4, 2.7 mm (relative 4%) in Table 5, and

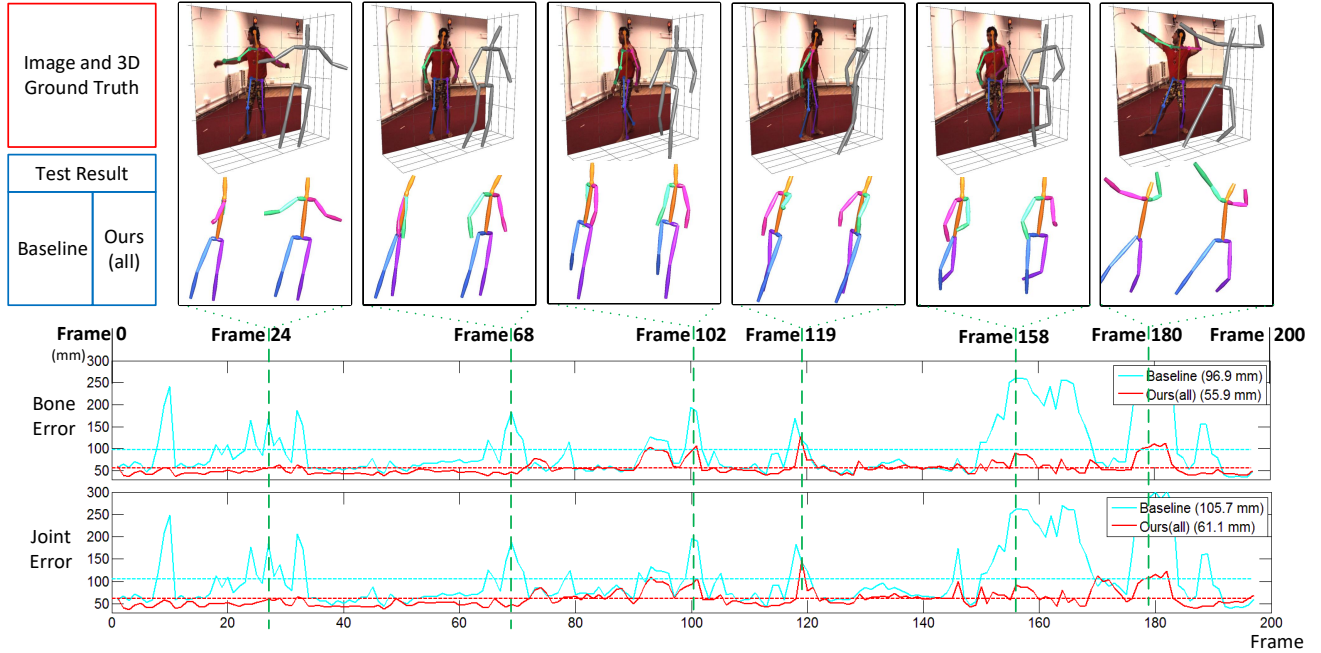


Figure 2. (best viewed in color) Errors of wrist joint/bone of *Baseline* and *Ours (all)* methods on a video sequence from Human3.6M S9, action Pose. The average error over the sequence is shown in the legends. For this action, the arms have large motion and are challenging. Our method has much smaller joint and bone error. Our result is more stable over the sequence. The 3D predicted pose and ground truth pose are visualized for a few frames. More video results are at <https://www.youtube.com/watch?v=c-hgHqVK90M>.

Method	Direction	Discuss	Eat	Greet	Phone	Pose	Purchase	Sit
Yasin[52]	88.4	72.5	108.5	110.2	97.1	81.6	107.2	119.0
Rogez[40]	-	-	-	-	-	-	-	-
Chen[7]	71.6	66.6	74.7	79.1	70.1	67.6	89.3	90.7
Moreno[30]	67.4	63.8	87.2	73.9	71.5	69.9	65.1	71.7
Zhou[57]	47.9	48.8	52.7	55.0	56.8	49.0	45.5	60.8
Baseline	45.2	46.0	47.8	48.4	54.6	43.8	47.0	60.6
Ours(all)	42.1	44.3	45.0	45.4	51.5	43.2	41.3	59.3
Method	SitDown	Smoke	Photo	Wait	Walk	WalkDog	WalkPair	Avg
Yasin[52]	170.8	108.2	142.5	86.9	92.1	165.7	102.0	108.3
Rogez[40]	-	-	-	-	-	-	-	88.1
Chen[7]	195.6	83.5	93.3	71.2	55.7	85.9	62.5	82.7
Moreno[30]	98.6	81.3	93.3	74.6	76.5	77.7	74.6	76.5
Zhou[57]	81.1	53.7	65.5	51.6	50.4	54.8	55.9	55.3
Baseline	79.0	54.5	56.0	46.7	42.2	51.0	47.9	51.4
Ours (all)	73.3	51.0	53.0	44.0	38.3	48.0	44.8	48.3

Table 4. Comparison with previous work on Human3.6M. Protocol 1 is used. Evaluation metric is averaged *PA Joint Error*. Extra 2D training data is used in all the methods. *Baseline* and *Ours (all)* use MPII data in the training. *Ours (all)* is the best and also wins in all the 15 activity categories.

5.1 mm (relative 4.8%) in Table 6. Therefore, it serves as a competitive reference. Second, *our method significantly improves the baseline, using exactly the same network and training*. Thus, the improvement comes from the new pose representation and loss function. It improves the state-of-the-art significantly, by 7 mm (relative 12.7%) in Table 4, 7.8 mm (relative 11.7%) in Table 5, and 14.9 mm (relative 13.9%) in Table 6.

Example 3D pose results are illustrated in Figure 3.

5.3. Experiments on 2D Pose of MPII

All leading methods on MPII benchmark [1] have sophisticated network architectures. As discussed in Section 2, the best-performing family of methods adopts a multi-stage architecture [11, 6, 31, 5, 47, 18, 14]. Our method is novel in the pose representation and loss func-

Method	Direction	Discuss	Eat	Greet	Phone	Pose	Purchase	Sit
Chen[7]	89.9	97.6	90.0	107.9	107.3	93.6	136.1	133.1
Tome[44]	65.0	73.5	76.8	86.4	86.3	68.9	74.8	110.2
Moreno[30]	69.5	80.2	78.2	87.0	100.8	76.0	69.7	104.7
Zhou[57]	68.7	74.8	67.8	76.4	76.3	84.0	70.2	88.0
Jahangiri[21]	74.4	66.7	67.9	75.2	77.3	70.6	64.5	95.6
Mehta[28]	57.5	68.6	59.6	67.3	78.1	56.9	69.1	98.0
Pavlakos[34]	58.6	64.6	63.7	62.4	66.9	57.7	62.5	76.8
Baseline	57.0	58.6	57.9	58.7	67.1	54.2	65.9	75.4
Ours(all)	52.8	54.8	54.2	54.3	61.8	53.1	53.6	71.7
Method	SitDown	Smoke	Photo	Wait	Walk	WalkDog	WalkPair	Avg
Chen[7]	240.1	106.7	139.2	106.2	87.0	114.1	90.6	114.2
Tome[44]	173.9	85.0	110.7	85.8	71.4	86.3	73.1	88.4
Moreno[30]	113.9	89.7	102.7	98.5	79.2	82.4	77.2	87.3
Zhou[57]	113.8	78.0	98.4	90.1	62.6	75.1	73.6	79.9
Jahangiri[21]	127.3	79.6	79.1	73.4	67.4	71.8	72.8	77.6
Mehta[28]	117.5	69.5	82.4	68.0	55.3	76.5	61.4	72.9
Pavlakos[34]	103.5	65.7	70.7	61.6	56.4	69.0	59.5	66.9
Baseline	98.1	66.2	71.1	58.4	51.4	65.2	56.8	64.2
Ours(all)	86.7	61.5	67.2	53.4	47.1	61.6	53.4	59.1

Table 5. Comparison with previous work on Human3.6M. Protocol 2 is used. Evaluation metric is averaged *Joint Error*. Extra 2D training data is used in all the methods. *Baseline* and *Ours (all)* use MPII data in the training. *Ours (all)* is the best and also wins in all the 15 activity categories.

Method	Direction	Discuss	Eat	Greet	Phone	Pose	Purchase	Sit
Zhou[56]	87.4	109.3	87.1	103.2	116.2	106.9	99.8	124.5
Tekin[43]	102.4	147.7	88.8	125.4	118.0	112.4	129.2	138.9
Xingyi[54]	91.8	102.4	97.0	98.8	113.4	90.0	93.8	132.2
Baseline	98.8	101.9	89.8	89.9	100.0	97.2	113.2	102.0
Ours(all)	90.2	95.5	82.3	85.0	87.1	87.9	93.4	100.3
Method	SitDown	Smoke	Photo	Wait	Walk	WalkDog	WalkPair	Avg
Zhou[56]	199.2	107.4	139.5	118.1	79.4	114.2	97.7	113.0
Tekin[43]	224.9	118.4	182.7	138.8	55.1	126.3	65.8	125.0
Xingyi[54]	159.0	106.9	125.2	94.4	79.0	126.0	99.0	107.3
Baseline	138.9	101.7	101.1	95.9	90.8	108.9	102.7	102.2
Ours(all)	135.4	91.4	94.5	87.3	78.0	90.4	86.5	92.4

Table 6. Comparison with previous work on Human3.6M. Protocol 2 is used. Evaluation metric is averaged *Joint Error*. No extra training data is used. *Ours (all)* is the best and wins in 12 out of 15 activity categories. Note that Tekin et al. [43] report more accurate results for "Walk" and "WalkPair", but their method uses the temporal context information in the video. Our method only runs on individual frames.

Stage	Metric	IEF*	joint	bone	both	all
0	Joint Error	29.7	27.2	27.8	27.5	27.2
	Bone Error	24.8	23.1	22.1	22.7	22.5
	PCKH 0.5	76.5%	79.3%	79.0%	79.2%	79.6%
1	Joint Error	25.0	23.8	25.2	23.0	22.8
	Bone Error	21.2	20.5	20.9	19.7	19.5
	PCKH 0.5	82.9%	84.1%	82.7%	84.9%	86.4%

Table 7. Results of the baseline and four variants of our method (see Table 1), in the two-stage IEF*.

tion. It is complementary to such sophisticated networks. In this experiment, it is integrated into the *Iterative Error Feedback* method (IEF) [6], which is the only regression based method in the family.

We implement a two stage baseline IEF, using ResNet-50

as the basic network in each stage. For reference, the original IEF [6] uses five stages with GoogLeNet for each stage. We denote our implementation as *IEF**. The two stages in *IEF** are then modified to use our bone based representation and compositional loss function. The training for all the settings remains the same, as specified in Section 4.

Ablation study Table 7 shows the results of *IEF** and our four variants. We observe the same conclusions as in Table 2. Both bone based representation and compositional loss function are effective under all metrics. In addition, both stages in *IEF** benefit from our approach.

Comparison with the state-of-the-art Table 8 reports the comparison result to state-of-the-art works on MPII. PCKH0.5 metric is used. Top section of Table 8 is detection based methods and bottom section is regression based. Ours

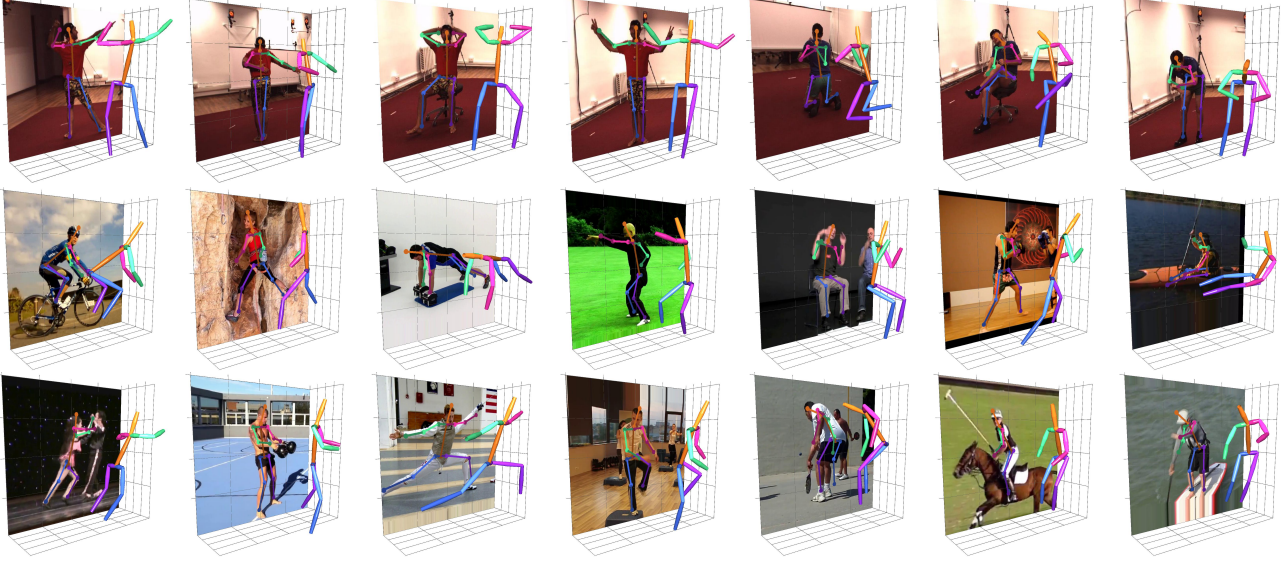


Figure 3. (best viewed in color) Examples of 3D pose estimation for Human3.6M (top row) and MPII (middle and bottom rows), using *Ours (all)* method in Table 5, trained with both 3D and 2D data. Note that the MPII 3D results are quite plausible and convincing.

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Pishchulin [37]	74.3	49.0	40.8	34.1	36.5	34.4	35.2	44.1
Tompson[46]	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6
Tompson[45]	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
Hu[17]	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
Pishchulin[38]	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
Lifshitz[25]	97.8	93.3	85.7	80.4	85.3	76.6	70.2	85.0
Gkioxary[14]	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1
Raf[39]	97.2	93.9	86.4	81.3	86.8	80.6	73.4	86.3
Insafutdinov[18]	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
Wei[47]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Bulat[5]	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
Newell[31]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.0
Chu[11]	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
Carreira(IEF)[6]	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
IEF*	96.3	92.6	83.1	74.6	83.7	74.1	71.4	82.9
Ours (all)	97.5	94.3	87.0	81.2	86.5	78.5	75.4	86.4

Table 8. Comparison to state-of-the-art works on MPII (top: detection based, bottom: regression based). PCKH 0.5 metric is used. Our approach significantly improves the baseline IEF and is competitive to other detection based methods.

(86.4%) produces significant improvement over the baseline (IEF*) and becomes the best regression based method. It is competitive to other detection based methods.

6. Conclusion

We show that regression based approach is competitive to the leading detection based approaches for 2D pose estimation once pose structure is appropriately exploited. Our approach is more potential for 3D pose estimation, where more complex structure constraints are critical.

Acknowledgement

This research work was supported by The National Science Foundation of China No. 61305091, and the Fundamental Research Funds for the Central Universities No. 2100219054.

References

- [1] MPII Leader Board. <http://human-pose.mpi-inf.mpg.de>. 1, 7
- [2] I. Akhter and M. J. Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1446–1455, 2015. 5
- [3] M. Andriluka, L. Pishchulin, P. Gehler, and B. 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*, pages 3686–3693, 2014. 1, 3, 4, 5
- [4] 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*, pages 561–578. Springer, 2016. 2, 5
- [5] A. Bulat and G. Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *European Conference on Computer Vision*, pages 717–732. Springer, 2016. 1, 2, 7, 9
- [6] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4733–4742, 2016. 1, 2, 3, 7, 8, 9

- [7] C.-H. Chen and D. Ramanan. 3d human pose estimation= 2d pose estimation+ matching. *arXiv preprint arXiv:1612.06524*, 2016. 2, 5, 7, 8
- [8] 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 *3D Vision (3DV), 2016 Fourth International Conference on*, pages 479–488. IEEE, 2016. 3
- [9] 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*, pages 1736–1744, 2014. 2
- [10] X. Chu, W. Ouyang, H. Li, and X. Wang. Structured feature learning for pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4715–4723, 2016. 2
- [11] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. Multi-context attention for human pose estimation. *arXiv preprint arXiv:1702.07432*, 2017. 1, 2, 7, 9
- [12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 5
- [13] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005. 2
- [14] G. Gkioxari, A. Toshev, and N. Jaitly. Chained predictions using convolutional neural networks. In *European Conference on Computer Vision*, pages 728–743. Springer, 2016. 2, 7, 9
- [15] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. 5
- [16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5
- [17] P. Hu and D. Ramanan. Bottom-up and top-down reasoning with hierarchical rectified gaussians. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5600–5609, 2016. 9
- [18] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepcrut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, pages 34–50. Springer, 2016. 1, 2, 7, 9
- [19] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 5
- [20] 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*, 36(7):1325–1339, 2014. 1, 2, 3, 4, 5
- [21] E. Jahangiri and A. L. Yuille. Generating multiple hypotheses for human 3d pose consistent with 2d joint detections. *arXiv preprint arXiv:1702.02258*, 2017. 2, 5, 8
- [22] 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. ACM, 2014. 5
- [23] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1465–1472. IEEE, 2011. 2
- [24] S. Li, W. Zhang, and A. B. Chan. Maximum-margin structured learning with deep networks for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2848–2856, 2015. 1, 2
- [25] I. Lifshitz, E. Fetaya, and S. Ullman. Human pose estimation using deep consensus voting. In *European Conference on Computer Vision*, pages 246–260. Springer, 2016. 9
- [26] C. Lindner, P. A. Bromiley, M. C. Ionita, and T. F. Cootes. Robust and accurate shape model matching using random forest regression-voting. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1862–1874, 2015. 2
- [27] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248, 2015. 2
- [28] D. Mehta, H. Rhodin, D. Casas, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation using transfer learning and improved cnn supervision. *arXiv preprint arXiv:1611.09813*, 2016. 1, 3, 5, 8
- [29] T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer vision and image understanding*, 81(3):231–268, 2001. 2
- [30] F. Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. *arXiv preprint arXiv:1611.09010*, 2016. 1, 2, 5, 7, 8
- [31] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016. 1, 2, 7, 9
- [32] W. Ouyang and X. Wang. Joint deep learning for pedestrian detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2056–2063, 2013. 2
- [33] S. Park, J. Hwang, and N. Kwak. 3d human pose estimation using convolutional neural networks with 2d pose information. In *Computer Vision–ECCV 2016 Workshops*, pages 156–169. Springer, 2016. 1, 3
- [34] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. *arXiv preprint arXiv:1611.07828*, 2016. 2, 5, 8
- [35] M. Pedersoli, A. Vedaldi, J. Gonzalez, and X. Roca. A coarse-to-fine approach for fast deformable object detection. *Pattern Recognition*, 48(5):1844–1853, 2015. 2
- [36] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595, 2013. 2
- [37] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Strong appearance and expressive spatial models for human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3487–3494, 2013. 9

- [38] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4929–4937, 2016. 2, 9
- [39] U. Rafi, I. Kostrikov, J. Gall, and B. Leibe. An efficient convolutional network for human pose estimation. In *BMVC*, volume 1, page 2, 2016. 9
- [40] 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. 3, 5, 7
- [41] 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. 2
- [42] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua. Structured prediction of 3d human pose with deep neural networks. *arXiv preprint arXiv:1605.05180*, 2016. 1, 2, 3
- [43] B. Tekin, A. Rozantsev, V. Lepetit, and P. Fua. Direct prediction of 3d body poses from motion compensated sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 991–1000, 2016. 1, 3, 5, 8
- [44] D. Tome, C. Russell, and L. Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. *arXiv preprint arXiv:1701.00295*, 2017. 3, 5, 8
- [45] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 648–656, 2015. 9
- [46] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in neural information processing systems*, pages 1799–1807, 2014. 2, 9
- [47] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016. 1, 2, 3, 7, 9
- [48] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman. Single image 3d interpreter network. In *European Conference on Computer Vision*, pages 365–382. Springer, 2016. 2
- [49] W. Yang, W. Ouyang, H. Li, and X. Wang. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *CVPR*, 2016. 2
- [50] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1385–1392. IEEE, 2011. 2
- [51] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2878–2890, 2013. 2
- [52] H. Yasin, U. Iqbal, B. Kruger, A. Weber, and J. 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*, pages 4948–4956, 2016. 5, 7
- [53] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *International Conference on Computer Vision*, 2017. 6
- [54] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei. Deep kinematic pose regression. In *Computer Vision–ECCV 2016 Workshops*, pages 186–201. Springer, 2016. 1, 2, 3, 5, 8
- [55] X. Zhou, Q. Wan, W. Zhang, X. Xue, and Y. Wei. Model-based deep hand pose estimation. *arXiv preprint arXiv:1606.06854*, 2016. 2
- [56] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4966–4975, 2016. 2, 5, 8
- [57] X. Zhou, M. Zhu, G. Pavlakos, S. Leonardos, K. G. Derpanis, and K. Daniilidis. Monocap: Monocular human motion capture using a cnn coupled with a geometric prior. *arXiv preprint arXiv:1701.02354*, 2017. 2, 5, 7, 8