

RepNet: Weakly Supervised Training of an Adversarial Reprojection Network for 3D Human Pose Estimation

Bastian Wandt and Bodo Rosenhahn
 Leibniz Universität Hannover
 Hannover, Germany

wandt@tnt.uni-hannover.de

Abstract

This paper addresses the problem of 3D human pose estimation from single images. While for a long time human skeletons were parameterized and fitted to the observation by satisfying a reprojection error, nowadays researchers directly use neural networks to infer the 3D pose from the observations. However, most of these approaches ignore the fact that a reprojection constraint has to be satisfied and are sensitive to overfitting. We tackle the overfitting problem by ignoring 2D to 3D correspondences. This efficiently avoids a simple memorization of the training data and allows for a weakly supervised training. One part of the proposed reprojection network (RepNet) learns a mapping from a distribution of 2D poses to a distribution of 3D poses using an adversarial training approach. Another part of the network estimates the camera. This allows for the definition of a network layer that performs the reprojection of the estimated 3D pose back to 2D which results in a reprojection loss function.

Our experiments show that RepNet generalizes well to unknown data and outperforms state-of-the-art methods when applied to unseen data. Moreover, our implementation runs in real-time on a standard desktop PC.

1. Introduction

Human pose estimation from monocular images is a very active research field in computer vision with many applications *e.g.* in movies, medicine, surveillance, or human-computer interaction. Recent approaches are able to infer 3D human poses from monocular images in good quality [27, 8, 21, 23, 28, 32, 24, 20, 19, 31]. However, most recent methods use neural networks that are straightforwardly trained with a strict assignment from input to output data *e.g.* [27, 8, 21, 23, 28, 32, 24, 19]. This leads to surprisingly impressive results on similar data but usually the generalization to unknown motions and camera positions is problem-

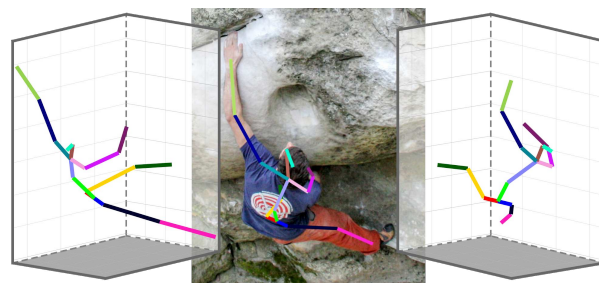


Figure 1. Our network predicts 3D human poses from noisy 2D joint detections. We use weakly supervised adversarial training without 2D to 3D point correspondences. Our critic networks enforces a plausible 3D pose while a reprojection layer projects the 3D pose back to 2D. Even strong deformations and unusual camera poses can be reconstructed.

atic. This paper presents a method to overcome this problem by using a neural network trained with a weakly supervised adversarial learning approach. We relax the assumption that a specific 3D pose is given for every image in the training data by training a discriminator network –widely used in generative adversarial networks (GAN) [9]– to learn a distribution of 3D human poses. A second neural network learns a mapping from a distribution of detected 2D keypoints (obtained by [25]) to a distribution of 3D keypoints which are valid 3D human poses according to the discriminator network. From the generative adversarial network point of view this can be seen as the generator network. To force the generator network to generate matching 3D poses to the 2D observations we propose to add a third neural network that predicts camera parameters from the input data. The inferred camera parameters are used to reproject the estimated 3D pose back to 2D which gives this framework its name: **Reprojection Network (RepNet)**. Fig. 2 shows an overview of the proposed network. Additionally, to further enforce kinematic constraints we propose to employ an easy to calculate and implement descriptor for joint lengths and angles inspired by the kinematic chain space (KCS) of Wandt et al. [41].

In contrast to other works our proposed method is very robust against overfitting to a specific dataset. This claim is reinforced by our experiments where the network can even infer human poses and camera positions that are not in the training set. Even if there are strong deformations or unusual camera poses our network achieves good results as can be seen in the rock climbing image in Fig. 1. This leads to our conclusion that the discriminator network does not *memorize* all poses from the training set but learns a meaningful manifold of feasible human poses. As we will show the inclusion of the KCS as a layer in the discriminator network plays an important role for the quality of the discriminator.

We evaluate our method on the three datasets Human3.6M [13], MPI-INF-3DHP [21], and Leeds Sports Pose (LSP) [16]. On all the datasets our method achieves state-of-the-art results and even outperforms most supervised approaches. Furthermore, the proposed network can predict a human pose in less than 0.1 milliseconds on standard hardware which allows to build a real-time pose estimation system when combining it with state-of-the-art 2D joint detectors, such as OpenPose [5].

The code will be made available.

Summarizing, our contributions are:

- An adversarial training method for a 3D human pose estimation neural network (RepNet) based on a 2D reprojection.
- Weakly supervised training without 2D-3D correspondences and unknown cameras.
- Simultaneous 3D skeletal keypoints and camera pose estimation.
- A layer encoding a kinematic chain representation that includes bone lengths and joint angle informations.
- A pose regression network that generalizes well to unknown human poses and cameras.

2. Related Work

The most relevant approaches related to our work can be roughly divided into two categories. The first group consists of optimization-based approaches where a 3D human body model is deformed such that it satisfies a reprojection error. The second group contains the most recent approaches that try to estimate 3D poses directly from images or detected keypoints.

2.1. Reprojection Error Optimization

Early works on human pose estimation from single images date back to Lee and Chen [18] in 1985. They use

known bone lengths and a binary decision tree to reconstruct a human pose. Some authors [15, 11, 6] propose to search for 3D poses in large pose databases that explain the 2D observations the best. To compress the knowledge from these databases a widely used method is to learn an overcomplete dictionary of 3D human poses either using principal component analysis (PCA) or another dictionary learning method. Commonly the best linear combination of bases obtained by a principal component analysis is optimized [7, 43, 49, 50]. To constrain the optimization several priors are proposed, such as joint angle limits [1], physical plausibility [46], or anthropometric regularization [30, 33, 42]. Other works enforce temporal coherence in video sequences [40, 2, 41, 46] or use additional sensors [37, 39, 38].

2.2. Direct Inference using Neural Networks

Recently, many researchers focus on directly regressing the 3D pose from image data or from 2D detections using deep neural networks. Several works try to build an end-to-end system which extracts the 3D pose from the image data [27, 8, 21, 23, 28, 32, 19, 17, 26, 29, 36, 45]. Moreno-Noguer [24] learns a mapping from 2D to 3D distance matrices. Martinez et al. [20] train a deep neural network on 2D joint detections to directly infer the 3D human pose. They trained their network to achieve an impressive performance on the benchmark dataset Human3.6M [13]. However, the network has significantly more parameters than poses in the training set of Human3.6M which could indicate a simple memorization of the training set. Although our proposed pose estimation network has a similar number of parameters our experiments indicate that overfitting is avoided by our adversarial training approach. Hossain et al. [31] extended the approach of [20] by using a recurrent neural network for sequences of human poses. The special case of weak supervision is rarely considered. Kanazawa et al. [17] propose a method that can also be trained without 2D to 3D supervision. In contrast to our approach they use the complete image information to train an end-to-end model to reconstruct a volumetric mesh of a human body. Yang et al. [45] train a multi-source discriminator network to build an end-to-end model.

3. Method

The basic idea behind the proposed method is that 3D poses are regressed from 2D observations by learning a mapping from the input distribution (2D poses) to the output distribution (3D poses).

In standard generative adversarial network (GAN) training [9] a generator network learns a mapping from an input distribution to the an output distribution which is rated by another neural network, called discriminator network. The discriminator is trained to distinguish between real samples

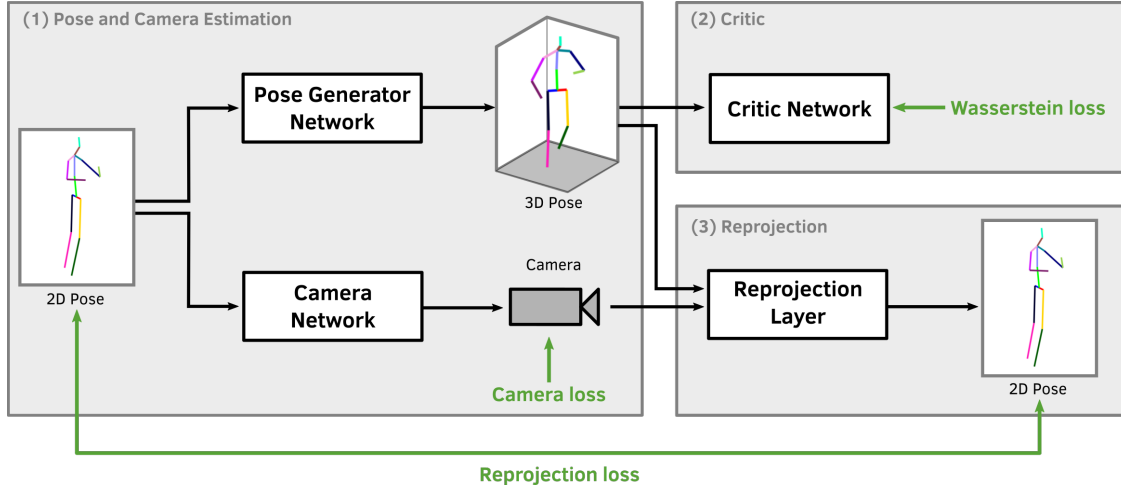


Figure 2. The proposed adversarial training structure for RepNet consist of three parts: a pose and camera estimation network (1), a critic network (2), and a reprojection network (3). There are losses (green) for the critic, the camera, and the reprojection.

from a database and samples created from the generator network. When training the generator to create samples that the discriminator predicts as real samples the discriminator parameters are fixed. The generator and the discriminator are trained alternately and therefore compete with each other until they both converge to a minimum.

In standard GAN training the input is sampled from a gaussian or uniform distribution. Here, we assume that the input is sampled from a distribution of 2D observations of human poses. Adopting the Wasserstein GAN naming [3] we call the discriminator *critic* in the following. Without knowledge about camera projections the network produces random, yet feasible human 3D poses. However, these 3D poses are very likely the incorrect 3D reconstructions of the input 2D observations. To obtain matching 2D and 3D poses we propose a camera estimation network followed by a reprojection layer. As shown in Fig. 2 the proposed network consists of three parts: The pose and camera estimation network (1), the critic used in the adversarial training (2), and the reprojection part (3). The critic and the complete adversarial model are trained alternately as described above.

3.1. Pose and Camera Estimation

The pose and camera estimation network splits into two branches, one for regression of the pose and the other for the camera estimation. In the following $\mathbf{X} \in \mathbb{R}^{3 \times n}$ denotes a 3D human pose where each column contains the xyz -coordinates of a body joint. In the neural network this matrix is written as a $3n$ dimensional vector. Correspondingly, if n joints are reconstructed the input of the pose and camera estimation network is a $2n$ dimensional vector containing the coordinates of the detected joints in the image.

The pose estimation part consists of two consecutive

residual blocks, where each block has two hidden layers of 1000 densely connected neurons. For the activation functions we use leaky ReLUs [12] which produced the best results in our experiments. The last layer outputs a $3n$ dimensional vector which contains the 3D pose and can be reshaped to \mathbf{X} . The camera estimation branch has a similar structure as the pose estimation branch with the output being a 6 dimensional vector containing the camera parameters. Here, we use a weak perspective camera model that can be defined by only six variables. To obtain the camera matrix the output vector is reshaped to $\mathbf{K} \in \mathbb{R}^{2 \times 3}$.

3.2. Reprojection Layer

The reprojection layer takes the output pose \mathbf{X} of the 3D generator network and the camera \mathbf{K} of the camera estimation network. The reprojection into 2D coordinate space can then be performed by

$$\mathbf{W}' = \mathbf{K}\mathbf{X}, \quad (1)$$

where \mathbf{W}' is called the *2D reprojection* in the following. This allows for the definition of a reprojection loss function

$$\mathcal{L}_{rep}(\mathbf{X}, \mathbf{K}) = \|\mathbf{W} - \mathbf{K}\mathbf{X}\|_F, \quad (2)$$

where \mathbf{W} is the input 2D pose observation matrix which has the same structure as \mathbf{W}' . $\|\cdot\|_F$ denotes the Frobenius norm. Note that the reprojection layer is a single layer which only performs the reprojection and does not have any trainable parameters. To deal with occlusions columns in \mathbf{W} and \mathbf{X} that correspond to not detected joints can be set to zero. This means they will have no influence on the value of the loss function. The missing joints will then be hallucinated by the pose generator network according to the critic network. In fact, the stacked hourglass network that produces the 2D joint detections [25] that we use as the input

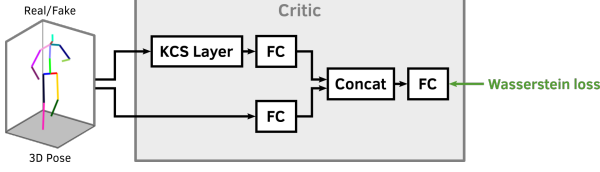


Figure 3. Network structure of the critic network. In the upper path the 3D pose is transformed into the KCS matrix and fed into a fully connected (FC) network. The lower path is build from multiple FC layers. The feature vectors of both paths are concatenated and fed into another FC layer which outputs the critic value.

does not predict the spine joint. We therefore set the corresponding columns in \mathbf{W} and \mathbf{X} to zero in all our experiments.

3.3. Critic Network

The complete network in Fig. 2 is trained alternately with the critic network. The loss on the last layer of the critic is a Wasserstein loss function [3]. The obvious choice of a critic network is a fully connected network with a structure similar to the pose regression network. However, such networks struggle to detect properties of human poses such as kinematic chains, symmetry and joint angle limits. Therefore, we follow the idea of Wandt et al. [41] and add their *kinematic chain space* (KCS) into our model. We develop a KCS layer with a successive fully connected network which is added in parallel to the fully connected path. These two paths in the critic network are merged before the output layer. Fig. 3 shows the network structure of the critic.

The KCS matrix is a representation of a human pose containing joint angles and bone lengths and can be computed by only two matrix multiplications. A bone \mathbf{b}_k is defined as the vector between the r -th and t -th joint

$$\mathbf{b}_k = \mathbf{p}_r - \mathbf{p}_t = \mathbf{X}\mathbf{c}, \quad (3)$$

where

$$\mathbf{c} = (0, \dots, 0, 1, 0, \dots, 0, -1, 0, \dots, 0)^T, \quad (4)$$

with 1 at position r and -1 at position t . Note that the length of the vector \mathbf{b}_k has the same direction and length as the corresponding bone. By concatenating b bones a matrix $\mathbf{B} \in \mathbb{R}^{3 \times b}$ can be defined as

$$\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_b). \quad (5)$$

This leads to a matrix $\mathbf{C} \in \mathbb{R}^{j \times b}$. The matrix \mathbf{B} is calculated by concatenating the corresponding vectors \mathbf{c} . It follows

$$\mathbf{B} = \mathbf{X}\mathbf{C}. \quad (6)$$

Multiplying \mathbf{B} with its transpose we compute the so called

KCS matrix

$$\Psi = \mathbf{B}^T \mathbf{B} = \begin{pmatrix} l_1^2 & \cdot & \cdot & \cdot \\ \cdot & l_2^2 & \cdot & \cdot \\ \cdot & \cdot & \ddots & \cdot \\ \cdot & \cdot & \cdot & l_b^2 \end{pmatrix}. \quad (7)$$

Because each entry in Ψ is an inner product of two bone vectors the KCS matrix has the bone lengths on its diagonal and a (scaled) angular representation on the other entries. In contrast to an euclidean distance matrix [24] the KCS matrix Ψ is easily calculated by two matrix multiplications. This allows for an efficient implementation as an additional layer. By giving the discriminator network an additional feature matrix it does not need to learn joint lengths computation and angular constraints on its own. In fact, in our experiments it was not possible to achieve an acceptable symmetry between the left and right side of the body without the KCS matrix. Section 4.1 shows how the 3D reconstruction benefits from adding the additional KCS layer. In our experiments there was no difference between adding convolutional layers or fully connected layers after the KCS layer. In the following we will use two fully connected layers, each containing 100 neurons, after the KCS layer. Combined with the parallel fully connected network this leads to the critic structure in Fig. 3.

3.4. Camera

Since the camera estimation sub-network in Fig. 2 can produce any 6-dimensional vector we need to force the network to produce matrices describing weak perspective cameras. If the 3D poses and the 2D poses are centered at their root joint the camera matrix \mathbf{K} projects \mathbf{X} to \mathbf{W}' according to Eq. 1. A weak perspective projection matrix \mathbf{K} has the property

$$\mathbf{K}\mathbf{K}^T = s^2 \mathbf{I}_2, \quad (8)$$

where s is the scale of the projection and \mathbf{I}_2 is the 2×2 identity matrix. Since the scale s is unknown we derive a computationally efficient method of calculating s . The scale s equals to the largest singular value (or the ℓ_2 -norm) of \mathbf{K} . Both singular values are equal. Since the trace of $\mathbf{K}\mathbf{K}^T$ is the sum of the squared singular values

$$s = \sqrt{\text{trace}(\mathbf{K}\mathbf{K}^T)/2}. \quad (9)$$

The loss function can now be defined as

$$\mathcal{L}_{cam} = \left\| \frac{2}{\text{trace}(\mathbf{K}\mathbf{K}^T)} \mathbf{K}\mathbf{K}^T - \mathbf{I}_2 \right\|_F, \quad (10)$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Note that only one matrix multiplication is necessary to compute the quadratic scale.

3.5. Data Preprocessing

The camera estimation network infers the parameters of the weak perspective camera. That means the camera matrix contains a rotational and a scaling component. To avoid ambiguities between the camera and 3D pose rotation all the rotational and scaling components from the 3D poses are removed. This is done by aligning every 3D pose to a template pose. We do this by calculating the ideal rotation and scale for the corresponding shoulder and hip joints via procrustes alignment. The resulting transformation is applied to all joints.

Depending on the persons size in the image the 2D joint detections can have arbitrary scale. To remove the scale component we divide each 2D pose vector by its standard deviation. Note that using this scaling technique the same person can have different sized 2D pose representations depending on the camera and 3D pose. However, the value for all possible 2D poses is constrained. The remaining scale variations are compensated by the cameras scale component. In contrast to *e.g.* [20] we do not need to know the mean and standard deviation of the training set. This allows for an easy transfer of our method to a different domain of 2D poses.

3.6. Training

We implemented the Improved Wasserstein GAN training procedure of [10]. In our experience this results in better and faster convergence compared to the traditional Wasserstein GAN [3] and standard GAN training [9] using binary cross entropy or similar loss functions. We use an initial learning rate of 0.001 with exponential decay every 10 epochs.

4. Experiments

We perform experiments on the three datasets Human3.6M [13], MPI-INF-3DHP [21], and LSP [16]. Human3.6M is the largest benchmark dataset containing images temporally aligned to 2D and 3D correspondences. Unless otherwise noted we use the training set of Human3.6M for training our networks. To show quantitative results on unseen data we evaluate our method on MPI-INF-3DHP. For unusual poses and camera angles subjective results are shown on LSP. Matching most comparable methods we use stacked hourglass networks [25] for 2D joint estimations from the input images in most of the experiments.

4.1. Quantitative Evaluation on Human3.6M

In the literature there are two main evaluation protocols on the Human3.6M dataset using subjects 1, 5, 6, 7, 8 for training and subject 9, 11 for testing. Both protocols calculate the *mean per joint positioning error* (MPJPE), i.e. the mean euclidean distance between the reconstructed and

the ground truth joint coordinates. Protocol-I computes the MPJPE directly whereas protocol-II first employs a rigid alignment between the poses. For a sequence the MPJPE's are summed and divided by the number of frames.

Table 1 shows the results of protocol-I without a rigid alignment. The rotation of the pose relative to the camera can be directly calculated from the camera matrix estimated by the camera regression network. Rotating the reconstructed pose in the world frame of the dataset gives the final 3D pose. Table 2 shows the results of protocol-II using a rigid alignment before calculating the error. The row *RepNet-noKCS* shows the errors without using the KCS layer. It can be seen that the additional KCS layer in the discriminator significantly improves the pose estimation. We are aware of the fact that our method will not be able to outperform supervised methods trained to perform exceptionally well on Human3.6M, such as [20] and [19]. Instead, in this section we show that even if we ignore the 2D-3D correspondences and train weakly supervised our network achieves comparable results to supervised state-of-the-art methods and is even better than most of them. Comparing to weakly supervised approaches [44, 35] we outperform the best by about 30% on protocol-II. For subjective evaluation the 1500th frame for every motion can be seen in Fig. 4. For comparability we show the same frame from every motion sequence from the same viewing angle. Even difficult poses, for instance sitting cross-legged, are reconstructed well.

In our opinion, although widely used on Human3.6M, the euclidean distance is not the only metric that should be considered to evaluate the performance of a human pose estimation system. Since there are some single frames that cannot be reconstructed well and can be seen as outliers we also calculate the median of the MPJPE over all frames. Additionally, we calculate the *percentage of correctly positioned keypoints* (PCK3D) as defined by [21] in Table 3.

In the following section we will show that although we do not improve on all supervised state-of-the-art methods directly trained on Human3.6M our approach outperforms every other known method on MPI-INF-3DHP without additional training.

4.2. Quantitative Evaluation on MPI-INF-3DHP

Our main contribution is a neural network that infers even unseen human poses while maintaining a meaningful 3D pose. We compare our method against several state-of-the-art approaches. Table 4 shows the results for different metrics. We clearly outperform every other method without having trained our model on this specific dataset. Even approaches trained on the training set of MPI-INF-3DHP perform worse than ours. This shows the generalization capability of our network. The row *RepNet 3DHP* is the result when training on the training set of MPI-INF-3DHP.

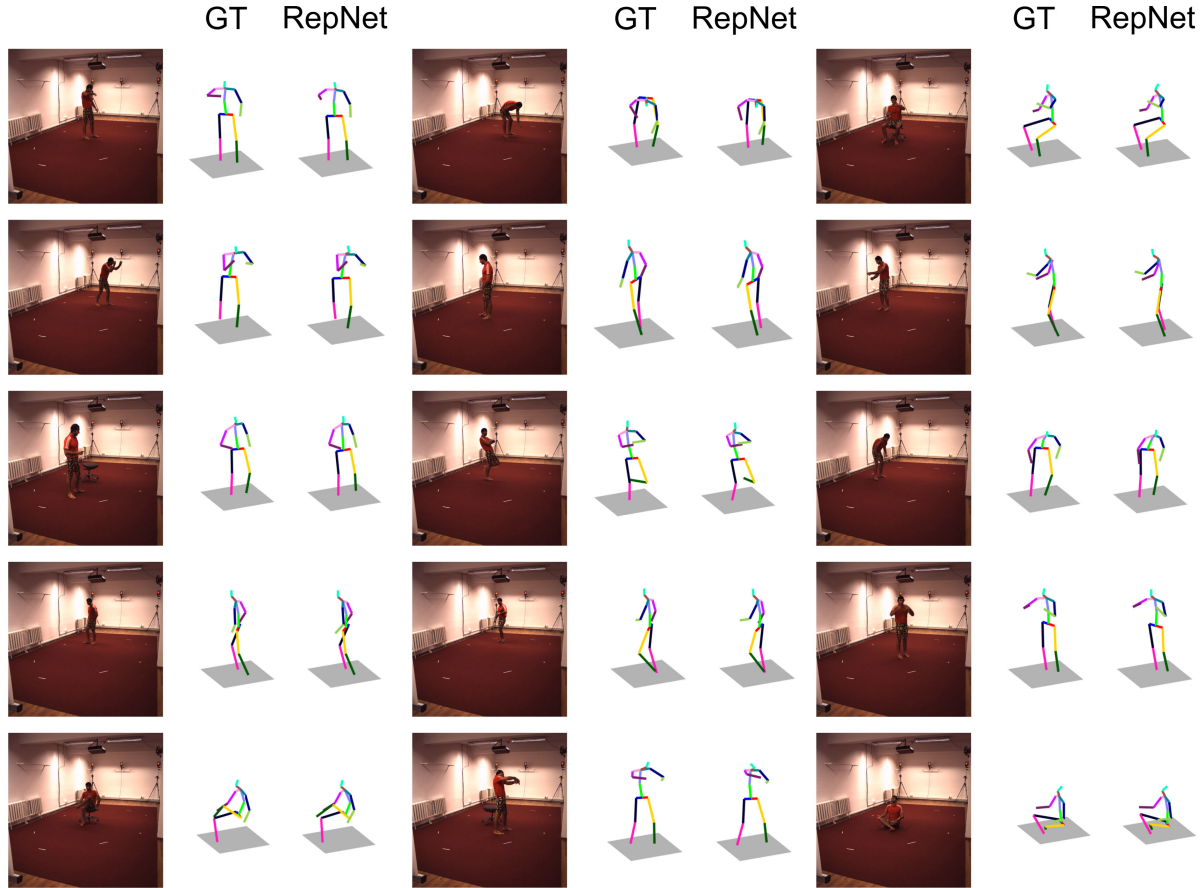


Figure 4. One example reconstruction for every motion from the test set of Human3.6M. The left 3D skeleton is the ground truth (GT) and the right shows our reconstruction (RepNet). Even difficult poses such as crossed legs or sitting on the floor are reconstructed well.

Table 1. Results for the reconstruction of the Human3.6M dataset compared to other state-of-the-art methods following *Protocol-I* (no rigid alignment). All numbers are taken from the referenced papers. For comparison the row *RepNet+2DGT* shows the error when using the ground truth 2D labels. The column *WS* denotes weakly supervised approaches. Note that there are no results available for other weakly supervised works.

Protocol-I	WS	Direct.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD	Smoke	Wait	Walk	WalkD	WalkT	Avg.
LinKDE [14]		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
Tekin et al. [34]		102.4	147.2	88.8	125.3	118.0	182.7	112.4	129.2	138.9	224.9	118.4	138.8	126.3	55.1	65.8	125.0
Zhou et al. [50]		87.4	109.3	87.1	103.2	116.2	143.3	106.9	99.8	124.5	199.2	107.4	118.1	114.2	79.4	97.7	113.0
Du et al. [8]		85.1	112.7	104.9	122.1	139.1	135.9	105.9	166.2	117.5	226.9	120.0	117.7	137.4	99.3	106.5	126.5
Park et al. [27]		100.3	116.2	90.0	116.5	115.3	149.5	117.6	106.9	137.2	190.8	105.8	125.1	131.9	62.6	96.2	117.3
Zhou et al. [48]		91.8	102.4	96.7	98.8	113.4	125.2	90.0	93.8	132.2	159.0	107.0	94.4	126.0	79.0	99.0	107.3
Luo et al. [19]		68.4	77.3	70.2	71.4	75.1	86.5	69.0	76.7	88.2	103.4	73.8	72.1	83.9	58.1	65.4	76.0
Pavlakos et al. [28]		67.4	71.9	66.7	69.1	72.0	77.0	65.0	68.3	83.7	96.5	71.7	65.8	74.9	59.1	63.2	71.9
Zhou et al. [47]		54.8	60.7	58.2	71.4	62.0	65.5	53.8	55.6	75.2	111.6	64.2	66.1	63.2	51.4	55.3	64.9
Martinez et al. [20]		53.3	60.8	62.9	62.7	86.4	82.4	57.8	58.7	81.9	99.8	69.1	63.9	50.9	67.1	54.8	67.5
RepNet (Ours)	✓	77.5	85.2	82.7	93.8	93.9	101.0	82.9	102.6	100.5	125.8	88.0	84.8	72.6	78.8	79.0	89.9
RepNet+2DGT (Ours)	✓	50.0	53.5	44.7	51.6	49.0	58.7	48.8	51.3	51.1	66.0	46.6	50.6	42.5	38.8	60.4	50.9

There is only a minor improvement of the 3DPCK and AUC and even a minor deterioration of the MPJPE compared to the network trained on Human3.6M. This suggests that the critic network converges to a similar distribution of feasible human poses for both training sets.

4.3. Plausibility of the Reconstructions

The metrics used for evaluation in Sec. 4.1 and 4.2 compare the estimated 3D pose to the ground truth. However, a

low error in this metrics is not necessarily an indication for a plausible human pose since the reconstructed pose can still violate joint angle limits or symmetries of the human body. For this purpose we introduce a new metric based on bone length symmetry. We calculate bone lengths of the lower and upper arms and legs since there is the largest error per joint. By summing the absolute differences of all matching bones on the right and left side of the body we can calculate a *symmetry error*. The mean symmetry error of the ground

Table 2. Results for the reconstruction of the Human3.6M dataset compared to other state-of-the-art methods following *Protocol-II* (rigid alignment). All numbers are taken from the referenced papers, except rows marked with * that are taken from [35]. Although we do not improve over supervised methods on this specific dataset our method clearly outperforms all other weakly supervised approaches (column WS). The best results for the weakly supervised methods are marked in bold. The second best approach that is not ours is underlined. For comparison the last row *RepNet+2DGT* shows the error when using the ground truth 2D labels.

Protocol-II	WS	Direct.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD	Smoke	Wait	Walk	WalkD	WalkT	Avg.
Akther and Black [1]		199.2	177.6	161.8	197.8	176.2	186.5	195.4	167.3	160.7	173.7	177.8	181.9	198.6	176.2	192.7	181.1
Ramakrishna et al. [30]		37.4	149.3	141.6	154.3	157.7	158.9	141.8	158.1	168.6	175.6	160.4	161.7	174.8	150.0	150.2	157.3
Zhou et al. [49]		99.7	95.8	87.9	116.8	108.3	107.3	93.5	95.3	109.1	137.5	106.0	102.2	110.4	106.5	115.2	106.7
Bogo et al. [4]		62.0	60.2	67.8	76.5	92.1	77.0	73.0	75.3	100.3	137.3	83.4	77.3	79.7	86.8	87.7	82.3
Moreno-Noguer [24]		66.1	61.7	84.5	73.7	65.2	67.2	60.9	67.3	103.5	74.6	92.6	69.6	78.0	71.5	73.2	74.0
Martinez et al. [20]		44.8	52.0	44.4	50.5	61.7	59.4	45.1	41.9	66.3	77.6	54.0	58.8	35.9	49.0	40.7	52.1
Luo et al. [19]		40.8	44.6	42.1	45.1	48.3	54.6	41.2	42.9	55.5	69.9	46.7	42.5	36.0	48.0	41.4	46.6
3DInterpreter* [44]	✓	78.6	90.8	92.5	89.4	108.9	112.4	77.1	106.7	127.4	139.0	103.4	91.4	79.1	-	-	98.4
AIGN [35]	✓	<u>77.6</u>	91.4	<u>89.9</u>	<u>88.0</u>	<u>107.3</u>	<u>110.1</u>	<u>75.9</u>	107.5	<u>124.2</u>	<u>137.8</u>	<u>102.2</u>	<u>90.3</u>	78.6	-	-	97.2
RepNet (Ours)	✓	53.0	58.3	59.6	66.5	72.8	71.0	56.7	69.6	78.3	95.2	66.6	58.5	63.2	57.5	49.9	65.1
RepNet-noKCS (Ours)	✓	63.1	67.4	71.5	78.5	85.9	82.6	70.8	82.7	92.2	116.6	77.6	72.2	65.3	73.2	69.6	77.9
RepNet+2DGT (Ours)	✓	33.6	38.8	32.6	37.5	36.0	44.1	37.8	34.9	39.2	52.0	37.5	39.8	34.1	40.3	34.9	38.2

Table 3. Performance of our method regarding the median and PCK3D errors for the Human3.6M dataset.

	mean	median	PCK3D
RepNet	65.1	60.0	93.0
RepNet+2DGT	38.2	36.0	98.6

Table 4. Results for the MPI-INF-3DHP dataset. All numbers are taken from the referenced papers, except the row marked with * which is taken from [22]. Without training on this dataset the proposed method outperforms every other method. The row *RepNet 3DHP* shows the result when using the training set of MPI-INF-3DHP. The column WS denotes weakly supervised approaches. A higher value is better for 3DPCK and AUC while a lower value is better for MPJPE. The best results are marked in bold and the second best approach is underlined.

Method	WS	3DPCK	AUC	MPJPE
Mehta et al. [21]		76.5	40.8	117.6
VNet [23]		76.6	40.4	124.7
LCR-Net[32]*		59.6	27.6	158.4
Zhou et al. [47]		69.2	32.5	137.1
Multi Person [22]		75.2	37.8	122.2
OriNet [19]		<u>81.8</u>	45.2	89.4
Kanazawa [17]	✓	77.1	40.7	113.2
Yang et al. [45]	✓	69.0	32.0	-
RepNet H36M (Ours)	✓	81.8	54.8	92.5
RepNet 3DHP (Ours)	✓	82.5	58.5	97.8

Table 5. Symmetry error in *mm* of the reconstructed 3D poses on the different datasets with and without the KCS. Adding the KCS layer to the critic networks results in significantly more plausible poses.

Method	mean	std	max
H36M noKCS	31.9	9.3	61.3
H36M KCS	8.2	3.8	20.5
3DHP noKCS	32.9	21.9	143.9
3DHP KCS	11.2	8.0	54.7

truth poses from the test set of Human3.6M and MPI-INF-3DHP for all subjects is $0.7mm \pm 0.8mm$ (max. $2.6mm$) and $2.1mm \pm 1.3mm$ (max. $7.6mm$), respectively. This leads us to the conclusion that an equality between the left and right side and therefore a low symmetry error is one reasonable metric for the plausibility of a human pose. Table 5

compares several implementations of our network in terms of the symmetry error. It can be clearly seen that the KCS layer has a significant impact on this metric. The higher values for the MPI-INF-3DHP dataset can be explained by the larger differences in symmetry of the ground truth data.

4.4. Noisy observations

Since the performance of our network appears to depend a lot on the detections of the 2D pose detector we evaluate our network on different levels of noise. Following [24] we add gaussian noise $\mathcal{N}(0, \sigma)$ to the ground truth 2D joint positions, where σ is the standard deviation in pixel. The results for Human3.6M under protocol-II are shown in Table 6. The error scales linearly with the standard deviation. This indicates that the noise of the 2D joint detector has a major impact on the results. Considering Tables 1 and 2 an improved detector will enhance the results to a level where they outperform current state-of-the-art supervised approaches.

Please note that the maximum person size from head to toe is approximately 200px in the input data. Therefore, gaussian noise with a standard deviation of $\sigma = 20px$ can be considered as extremely large. However, due to the critic network using the KCS layer the output of the pose estimation network is still a plausible human pose. To demonstrate this we additionally investigated the average, standard deviation and maximal symmetry error for the different noise levels which is also shown in Table 6. As expected the error increases only slightly since the critic network enforces plausible human poses. Even for noise levels as high as $\mathcal{N}(0, 20)$ we achieve an average symmetry error of only $22.7mm \pm 4.5mm$ which can be considered as very low.

4.5. Qualitative Evaluation

For a subjective evaluation we use the Leeds Sports Pose dataset (LSP) [16]. This dataset contains 2000 images of different people doing sports. There is a large variety in poses including stretched poses close to the limits of pos-

Table 6. Evaluation results for protocol-II (rigid alignment) with different levels of gaussian noise $\mathcal{N}(0, \sigma)$ (σ is the standard deviation) added to the ground truth 2D positions (*GT*). The 2D detector noise has large impact on the 3D reconstruction. The right three columns show the mean, standard deviation, and maximal symmetry error in millimeter.

Protocol-II	Direct.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD	Smoke	Wait	Walk	WalkD	WalkT	Avg.	symmetry		
																	mean	std	max
GT	33.6	38.8	32.6	37.5	36.0	44.1	37.8	34.9	39.2	52.0	37.5	39.8	34.1	40.3	34.9	38.2	6.2	3.7	20.8
GT + $\mathcal{N}(0, 5)$	54.0	56.8	52.7	56.5	54.4	59.7	55.7	54.1	56.3	68.5	56.1	58.7	57.6	56.7	55.3	56.9	9.6	4.0	25.0
GT + $\mathcal{N}(0, 10)$	70.4	72.2	72.8	75.1	70.2	84.1	68.4	89.3	74.0	94.1	68.3	74.3	67.7	73.5	70.0	74.9	13.0	3.8	24.2
GT + $\mathcal{N}(0, 15)$	86.3	88.0	87.5	89.9	84.0	98.1	84.0	104.2	87.4	107.7	82.3	89.3	85.1	89.0	86.0	89.9	17.6	4.2	32.1
GT + $\mathcal{N}(0, 20)$	101.6	103.0	101.6	104.5	97.5	112.2	99.3	118.1	100.9	121.5	95.9	104.0	101.6	104.7	102.3	104.6	22.7	4.5	37.5

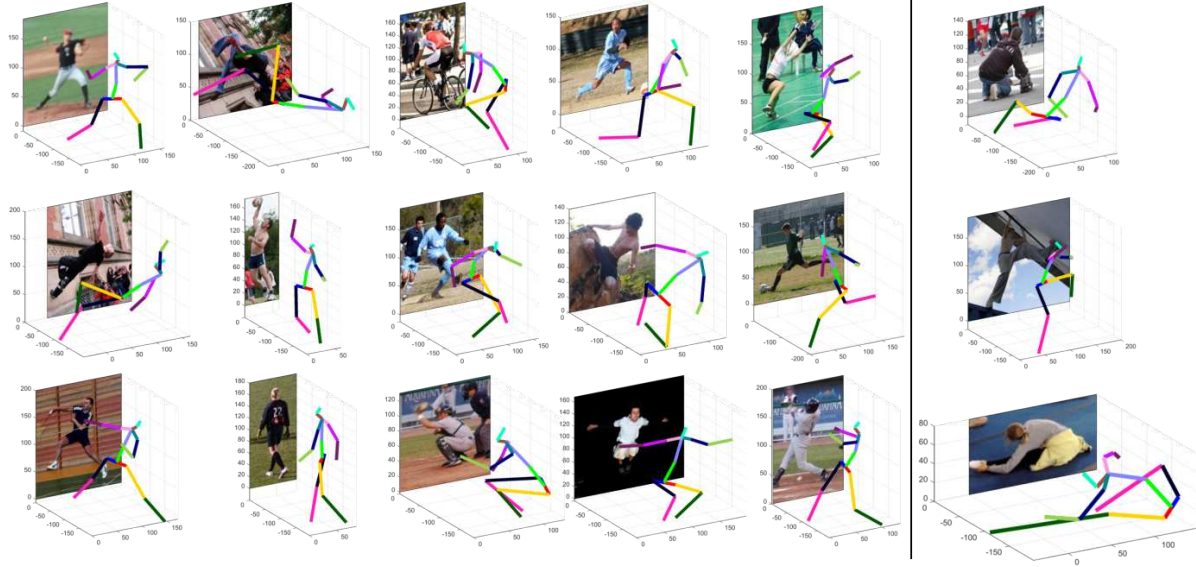


Figure 5. Example 3D pose estimations from the LSP dataset. Good reconstructions are in the left columns. The right column shows some failure cases with very unusual poses or camera angles. Although not perfect the poses are still plausible and close to the correct poses.

sible joint angles. Some of these poses and camera angles were never seen before by our network. Nevertheless, it is able to predict plausible 3D poses for most of the images. Fig. 5 shows some of the reconstructions achieved by our method. There are many subjectively well reconstructed poses, even if these are extremely stretched and captured from uncommon camera angles. Note that our network was only trained on the camera angles of Human3.6M. This underlines that an understanding of plausible poses and 2D projections is learned. The right column in Fig. 5 shows some failure cases and emphasizes a limitation of this approach: poses or camera angles that are too different from the training data cannot be reconstructed well. However, the reconstructions are still plausible human poses and in most cases at least near to the correct pose.

4.6. Computational Time

We see our method as a building block in a larger image-to-3D points system. Current state-of-the-art 2D keypoint detectors such as [5] achieve real-time performance (approximately 100ms per frame) on standard hardware. Our network adds another 0.05ms per frame and therefore has nearly no impact on the runtime. Assuming the 2D keypoint detection takes no time we achieve a frame rate of 20000fps

on an Nvidia TITAN X.

5. Conclusion

This paper presented RepNet: a weakly supervised training method for a 3D human pose estimation neural network that infers 3D poses from 2D joint detections in single images. We proposed to use an additional camera estimation network and our novel reprojection layer that projects the estimated 3D pose back to 2D. By exploiting state-of-the-art techniques in neural network research, such as improved Wasserstein GANs [10] and kinematic chain spaces [41], we were able to develop a weakly supervised training procedure that does not need 2D to 3D correspondences. This not only outperforms previous weakly supervised methods but also avoids overfitting of the network to a limited amount of training data. We achieved state-of-the-art performance on the benchmark dataset Human3.6M, even compared to most supervised approaches. Using the network trained on Human3.6M to predict 3D poses from the unseen data of the MPI-INF-3DHP dataset showed an improvement over all other methods. We also performed a subjective evaluation on the LSP dataset where we achieved good reconstructions even on images with uncommon poses and perspectives.

References

- [1] I. Akhter and M. J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2015)*, pages 1446–1455, June 2015. 2, 7
- [2] T. Alldieck, M. Kassubeck, B. Wandt, B. Rosenhahn, and M. Magnor. Optical flow-based 3d human motion estimation from monocular video. In *German Conference on Pattern Recognition (GCPR)*, Sept. 2017. 2
- [3] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. 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 *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science. Springer International Publishing, Oct. 2016. 7
- [5] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 2, 8
- [6] C. Chen and D. Ramanan. 3d human pose estimation = 2d pose estimation + matching. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pages 5759–5767, 2017. 2
- [7] Y.-L. Chen and J. Chai. 3d reconstruction of human motion and skeleton from uncalibrated monocular video. In H. Zha, R. I. Taniguchi, and S. J. Maybank, editors, *Asian Conference on Computer Vision (ACCV)*, volume 5994 of *Lecture Notes in Computer Science*, pages 71–82. Springer, 2009. 2
- [8] Y. Du, Y. Wong, Y. Liu, F. Han, Y. Gui, Z. Wang, M. Kankanhalli, and W. Geng. Marker-less 3D human motion capture with monocular image sequence and height-maps. In *European Conference on Computer Vision*, pages 20–36. Springer, 2016. 1, 2, 6
- [9] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS’14, pages 2672–2680, Cambridge, MA, USA, 2014. MIT Press. 1, 2, 5
- [10] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5767–5777. Curran Associates, Inc., 2017. 5, 8
- [11] A. Gupta, J. Martinez, J. J. Little, and R. J. Woodham. 3d pose from motion for cross-view action recognition via non-linear circulant temporal encoding. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2601–2608, 2014. 2
- [12] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), ICCV ’15*, pages 1026–1034, Washington, DC, USA, 2015. IEEE Computer Society. 3
- [13] 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, jul 2014. 2, 5
- [14] 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. 6
- [15] H. Jiang. 3d human pose reconstruction using millions of exemplars. *2010 20th International Conference on Pattern Recognition*, pages 1674–1677, 2010. 2
- [16] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, 2010. doi:10.5244/C.24.12. 2, 5, 7
- [17] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2018. 2, 7
- [18] H.-J. Lee and Z. Chen. Determination of 3d human body postures from a single view. *Computer Vision, Graphics, and Image Processing*, 30(2):148 – 168, 1985. 2
- [19] C. Luo, X. Chu, and A. L. Yuille. Orinet: A fully convolutional network for 3d human pose estimation. In *British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018*, page 92, 2018. 1, 2, 5, 6, 7
- [20] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 1, 2, 5, 6, 7
- [21] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. 1, 2, 5, 7
- [22] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll, and C. Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *3D Vision (3DV), 2018 Sixth International Conference on*. IEEE, sep 2018. 7
- [23] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. volume 36, 7 2017. 1, 2, 7
- [24] F. Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 4, 7
- [25] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV (8)*, volume 9912 of *Lecture Notes in Computer Science*, pages 483–499. Springer, 2016. 1, 3, 5

- [26] M. Omran, C. Lassner, G. Pons-Moll, P. V. Gehler, and B. Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. In *3DV*, Sept. 2018. 2
- [27] 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 - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III*, pages 156–169, 2016. 1, 2, 6
- [28] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1263–1272, 2017. 1, 2, 6
- [29] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *CVPR*, 2018. 2
- [30] V. Ramakrishna, T. Kanade, and Y. A. Sheikh. Reconstructing 3d human pose from 2d image landmarks. In *European Conference on Computer Vision (ECCV)*, October 2012. 2, 7
- [31] M. Rayat Imtiaz Hossain and J. J. Little. Exploiting temporal information for 3d human pose estimation. In *The European Conference on Computer Vision (ECCV)*, September 2018. 1, 2
- [32] G. Rogez, P. Weinzaepfel, and C. Schmid. LCR-Net: Localization-Classification-Regression for Human Pose. In *CVPR 2017 - IEEE Conference on Computer Vision & Pattern Recognition*, pages 1216–1224, Honolulu, United States, July 2017. IEEE. 1, 2, 7
- [33] E. Simo-Serra, A. Ramisa, G. Aleny, C. Torras, and F. Moreno-Noguer. Single image 3d human pose estimation from noisy observations. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2673–2680. IEEE, 2012. 2
- [34] B. Tekin, A. Rozantsev, V. Lepetit, and P. Fua. Direct prediction of 3d body poses from motion compensated sequences. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 991–1000, 2016. 6
- [35] H. F. Tung, A. W. Harley, W. Seto, and K. Fragkiadaki. Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4364–4372, Oct 2017. 5, 7
- [36] H.-Y. Tung, H.-W. Tung, E. Yumer, and K. Fragkiadaki. Self-supervised learning of motion capture. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5236–5246. Curran Associates, Inc., 2017. 2
- [37] T. v. Marcard, G. Pons-Moll, and B. Rosenhahn. Human pose estimation from video and imus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1533–1547, Aug 2016. 2
- [38] T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European Conference on Computer Vision (ECCV)*, volume Lecture Notes in Computer Science, vol 11214, pages 614–631. Springer, Cham, Sept. 2018. 2
- [39] T. von Marcard, B. Rosenhahn, M. J. Black, and G. Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. *Computer Graphics Forum*, 36(2):349–360, 2017. 2
- [40] B. Wandt, H. Ackermann, and B. Rosenhahn. 3d reconstruction of human motion from monocular image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1505–1516, 2016. 2
- [41] B. Wandt, H. Ackermann, and B. Rosenhahn. A kinematic chain space for monocular motion capture. In *ECCV Workshops*, Sept. 2018. 1, 2, 4, 8
- [42] C. Wang, Y. Wang, Z. Lin, A. Yuille, and W. Gao. Robust estimation of 3d human poses from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2
- [43] X. K. Wei and J. Chai. Modeling 3d human poses from uncalibrated monocular images. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1873–1880, Sept 2009. 2
- [44] 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 (ECCV)*, 2016. 5, 7
- [45] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang. 3d human pose estimation in the wild by adversarial learning. In *CVPR*, 2018. 2, 7
- [46] P. Zell, B. Wandt, and B. Rosenhahn. Joint 3d human motion capture and physical analysis from monocular videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. 2
- [47] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Towards 3d human pose estimation in the wild: A weakly-supervised approach. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 6, 7
- [48] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei. Deep kinematic pose regression. pages 186–201, 2016. 6
- [49] X. Zhou, M. Zhu, S. Leonardos, and K. Daniilidis. Sparse representation for 3d shape estimation: A convex relaxation approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(8):1648–1661, Aug 2017. 2, 7
- [50] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2, 6