Generalizing Monocular 3D Human Pose Estimation in the Wild

Luyang Wang[1,2*] Yan Chen[1,2*] Zhenhua Guo[2] Keyuan Qian[2]
Mude Lin[1] Hongsheng Li[3] Jimmy S. Ren[1]
[1]SenseTime Research [2]Tsinghua University [3]CUHK-SenseTime Joint Lab
{wangluyang, chenyan, linmude, rensijie}@sensetime.com

Abstract

The availability of the large-scale labeled 3D poses in the Human3.6M dataset plays an important role in advancing the algorithms for 3D human pose estimation from a still image. We observe that recent innovation in this area mainly focuses on new techniques that explicitly address the generalization issue when using this dataset, because this database is constructed in a highly controlled environment with limited human subjects and background variations. Despite such efforts, we can show that the results of the current methods are still error-prone especially when tested against the images taken in-the-wild. In this paper, we aim to tackle this problem from a different perspective. We propose a principled approach to generate high quality 3D pose ground truth given any in-the-wild image with a person inside. We achieve this by first devising a novel stereo inspired neural network to directly map any 2D pose to high quality 3D counterpart. We then perform a carefully designed geometric searching scheme to further refine the joints. Based on this scheme, we build a large-scale dataset with 400,000 in-the-wild images and their corresponding 3D pose ground truth. This enables the training of a high quality neural network model, without specialized training scheme and auxiliary loss function, which performs favorably against the state-of-the-art 3D pose estimation methods. We also evaluate the generalization ability of our model both quantitatively and qualitatively. Results show that our approach convincingly outperforms the previous methods. We make our dataset and code publicly available.[1]

1. Introduction

3D human pose estimation is one of the fundamental problems in computer vision. It is widely used in a large number of areas such as action recognition, virtual reality, human-computer interaction, and video surveillance. Recently, significant advances have been achieved in 2D human pose estimation due to the powerful deep Convolutional Neural Networks (CNNs) and the availability of large-scale in-the-wild 2D human pose datasets with manual annotations. However, advances in 3D human pose estimation remain limited.

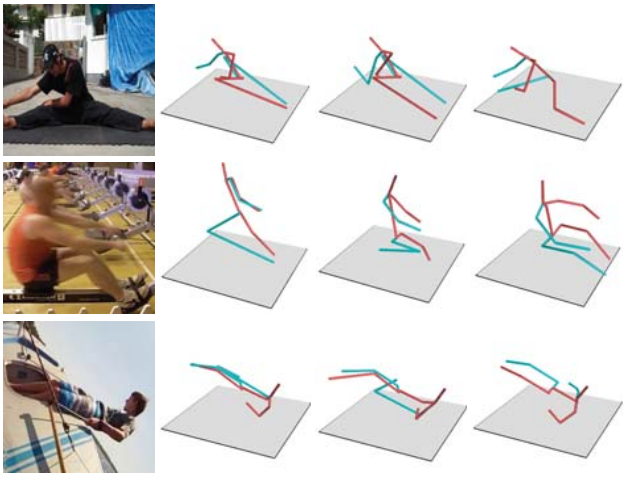


Figure 1. 3D pose estimation on challenging images. The proposed method performs favorably against the state-of-the-art 3D pose estimation algorithms due to our generated in-the-wild 3D pose dataset. The red lines denote the skeletons of the left and torso part of the human body while blue lines represent the right part. All the predicted poses are demonstrated at the same viewpoint.

This problem is widely studied in the literature and is mainly tackled with the following types of technical methodologies namely 2D-to-3D pose estimation [6, 19], monocular image based 3D pose estimation [47, 44], and multi-view images based 3D pose estimation [31]. Human3.6M dataset [12] plays an important role in the passive 3D human pose estimation methods. It is collected in a highly constrained environment with limited subjects, and background variations. The innovation of these methods mainly focuses on new techniques that explicitly address

*Indicates equal contribution.
[1]https://github.com/llcshappy/Monocular-3D-Human-Pose

the generalization issues when using this dataset. As shown in Figure 1, the current methods are still problematic when tested against the in-the-wild images.

To solve the problem, we can improve the generalization ability with well-annotated in-the-wild 3D pose data. Rogez et al. [32] propose a method to solve the limitations of the laboratory 3D datasets by artificially composing different images to generate a synthetic one based on the 3D Motion Capture (MoCap) data. However, the details and variety level of these synthetic images are limited compared with the in-the-wild images.

In this paper, we introduce a principled method to generate high quality 3D labels of the in-the-wild images. Inspired by [31], to solve the depth ambiguity problem in 3D human pose estimation, we devise a stereo inspired 3D label generator utilizing the 2D poses from multi-view to generate a high quality 3D human pose. We also propose a geometric search scheme to further refine the predicted 3D human pose. Given any image with the 2D ground truth, the proposed 3D label generator can produce its high quality counterpart.

To this end, based on the 3D label generator, we collect more than 400,000 in-the-wild images with high quality 3D labels from the wildly used 2D pose datasets [3, 13, 43]. With the proposed in-the-wild 3D pose dataset, we train a high performance baseline network which achieves favorable results against the state-of-the-art methods , both quantitatively and qualitatively. Furthermore, we introduce a method that utilizes the predicted 3D human pose on the task of action classification to evaluate the generalization ability quantitatively.

Our contributions can be summarized as follows:

- We propose a novel stereo inspired neural network to generate high quality 3D pose labels for in-the-wild images. We also devise a geometric searching scheme to further refine the 3D joints.
- We build a large-scale dataset with 400,000 in-the-wild images and the corresponding high quality 3D pose labels.
- We train a baseline network with the proposed dataset that performs favorably against the state-of-the-art approaches, both quantitatively and qualitatively. Experimental results demonstrate that the proposed dataset can significantly boost the generalization performance on the realistic scenes.

2. Related Work

Synthetic Images and Additional Annotations. Most of the existing 3D pose datasets such as Human3.6M [12], and HumanEva [34] are collected in indoor scenes and cannot cover various activities. To solve it, several methods attempt to use the graphics methods [4, 20, 29, 41] to enrich the training samples. Rogez and Schmid [32] introduce a computer graphics engine that artificially composes different images to generate synthetic poses based on the 3D Motion Capture (MoCap) data. Recently, to alleviate the need for the accurate 3D labels of in-the-wild images, Pavlakos et al. [27] and Shi et al. [33] provide in-the-wild images with additional annotations which contain the forward or backward information of each bone. However, the aforementioned methods either have limited details and variety level of the synthetic images or require a large number of manual annotations. Different from them, we propose a network to automatically generate a large-scale in-the-wild 3D human pose dataset.

2D-to-3D Pose Estimation. Several methods tackle 3D pose understanding from 2D pose [1, 5, 15, 30, 37, 42, 45, 48]. Martinez et al. [19] propose a simple multi-layer perceptron to regress the locations of the 3D joints. Fang et al. [6] introduce a model to encode the mapping function of human pose from 2D to 3D by explicitly encoding human body configuration with pose grammar. Despite the consideration of the domain knowledge of the human body, models trained with 2D/3D key-points from Human3.6M containing only fifteen activities cannot perform well to various actions. Training with 2D/3D pairs including more than 2,500 activities generated by the unity toolbox [2], we devise a network can map an in-the-wild 2D pose to its high quality 3D counterpart.

Monocular Image Based 3D Pose Estimation. Recently, several methods have been proposed to estimate the 3D pose on the monocular image [11, 16, 20, 28, 36, 38, 40, 49]. To improve the generalization on realistic scenes, some attempt to estimate 3D human pose in a semi-supervised way. Zhou et al. [47] employ a weakly-supervised transfer learning method with a 3D geometric loss.Yang et al. [44] propose an adversarial learning framework, which distills the 3D human pose structures learned from the fully annotated dataset to in-the-wild images with only 2D pose annotations. However, their mechanisms are to transfer the domain knowledge of the constrained dataset to the in-the-wild images without adding new subjects or more activities. Trained with our proposed in-the-wild 3D dataset, the network performs better on the in-the-wild images.

Multi-View Images Based 3D Pose Estimation. Some methods attempt to estimate the 3D human pose from multiple views of different cameras [2, 8, 9, 31]. Amin et al. [2] propose the evidence across multiple viewpoints to allow for robust 3D pose estimation. Rhodin et al. [31] propose to predict the same 3D pose in all views with only a small number of labeled images. However, compared with generating 2D joints from different cameras, obtaining the images

[2] https://unity3d.com

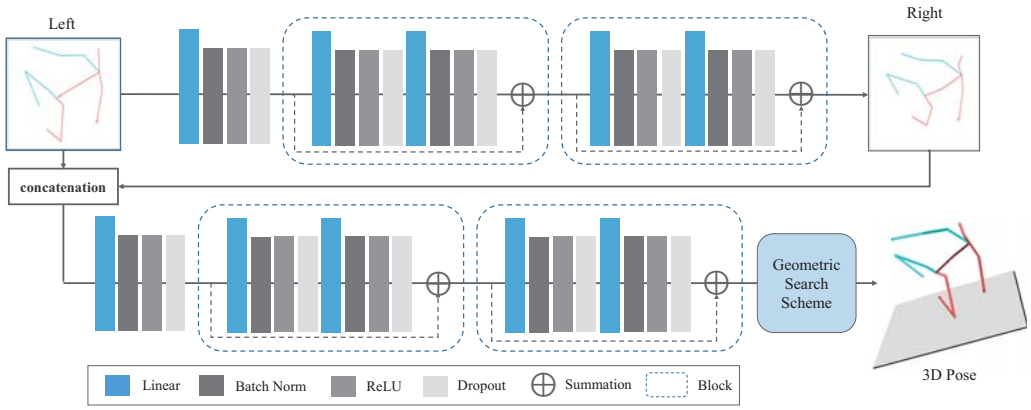


Figure 2. Architecture of the 3D label generator. The generator consists of stereoscopic view synthesis subnetwork, 3D pose reconstruction subnetwork, and a geometric search scheme. Given the 2D pose from the left viewpoint, stereoscopic view synthetic subnetwork aims to generate the 2D pose from the right viewpoint. 3D pose reconstruction subnetwork utilizes the multi-view 2D poses to estimate a coarse 3D human pose. Geometric search scheme is applied to further refine the predicted 3D human pose.

from multi-views are more difficult. Based on these methods, we devise a network with the simple 2D joints from the multi-view to generate the high quality 3D human poses.

3. Methodology

In this section, we first introduce the principles of the network design. As shown in Figure 2, we present a novel 3D label generator, which consists of stereoscopic view synthesis subnetwork, 3D pose reconstruction subnetwork, and a geometric search scheme. In addition, we propose a large-scale in-the-wild 3D pose dataset, and its 3D pose labels are provided by the 3D pose generator. Furthermore, we adopt a baseline network to evaluate the proposed in-the-wild 3D pose dataset.

3.1. Principles of Network Design

2D-to-3D human pose estimation inherently accompanies with the depth ambiguity since the mapping function from 2D to 3D is not unique. Amin et al. [2] propose the evidence across multiple viewpoints to allow for robust 3D pose estimation. Recently, Luo et al. [18] confirm that utilizing images from two different cameras can achieve excellent performance in the stereo matching area. Inspired by [18], we devise the stereo inspired neural network utilizing the 2D key-points from two different viewpoints to alleviate the depth ambiguity of predicting 3D poses. Different from the previous methods [31, 2] using multi-view images as inputs to estimate the 3D pose, our method is relatively easier to obtain the training data, since the unity toolbox can generate a large number of 2D/3D pairs automatically.

In addition, most existing 2D-to-3D methods [10, 6] mainly focus on the domain-knowledge of the human body or the architecture of the network while ignoring that a reasonable predicted 3D human pose can be re-projected to its 2D input with zero-pixel error. Based on this principle, we devise a geometric search scheme to further refine the predicted coarse 3D human pose.

3.2. Stereoscopic View Synthesis Subnetwork

Stereoscopic view synthesis subnetwork is proposed to synthesize the 2D pose from the right viewpoint . Given an image with 2D key-points from the left viewpoint, we generate the 2D key-points from the right viewpoint. As shown in Figure 2, we input the left-view 2D pose (u_L, v_L) to regress the location of the right-view 2D pose (u_R, v_R). However, the challenge is how to obtain the ground truth of the 2D pose from the right viewpoint .

We employ a large bunch of 3D key-points and their corresponding camera intrinsic matrix from the realistic 3D pose dataset (i.e., Human3.6M [12]) and the synthetic data generated by the unity toolbox to train the subnetwork. To obtain the ground truth of the right-view 2D pose, we move 3D joints to the right direction slightly along the X axis in the camera coordinate system while keep Y and Z unchanged. We then re-map them into the 2D key-points based on the camera calibration by the following equation:

$$s\begin{bmatrix} u_R \\ v_R \\ 1 \end{bmatrix} = \begin{bmatrix} \alpha_x & 0 & u_0 \\ 0 & \alpha_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_c + \Delta x \\ y_c \\ z_c \end{bmatrix} = M_c P_c, \quad (1)$$

where s denotes the scale factor, (u_R, v_R) represents the right-view 2D pose, α_x and α_y are the scale factors, (u_0, v_0) denotes the origin coordinate of the RGB image. $P_c = (x_c, y_c, z_c)$ is the location of 3D key-points in the camera coordinate system. M_c represents the camera intrinsic matrix. $\Delta_x = 500\text{mm}$ denotes the shift distance.

The subnetwork contains the linear-ReLU layers, residual connections, batch normalization layers, and dropout layers with max-norm constraint.

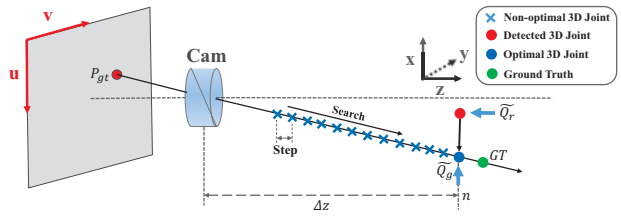


Figure 3. Geometric search scheme. $\widetilde{Q_r} = (\widetilde{x_r}, \widetilde{y_r}, \widetilde{z_r})$ denotes the predicted 3D human pose by the 3D pose reconstruction subnetwork with hypothetical depth to the camera. $\widetilde{Q_g} = (\widetilde{x_g}, \widetilde{y_g}, \widetilde{z_g})$ represents the 3D human pose with the absolute depth to the camera.

3.3. 3D Pose Reconstruction Subnetwork

As shown in Figure 2, 3D pose reconstruction subnetwork directly regresses the location of 3D key-points based on the input left-view 2D pose and synthesized right-view 2D pose. It shares the same architecture as the stereoscopic view synthesis subnetwork and takes the left-view and synthetic right-view 2D poses as the inputs. More precisely, when inputting the multi-view 2D poses, we combine them by the concatenation operation. After the operation of a fully connected network structure, we obtain a coarse 3D human pose. The 3D pose reconstruction can be represented by the following equation:

$$Q_r = f_r((u_L, v_L), (u_R, v_R)), \tag{2}$$

where $Q_r = (x_r, y_r, z_r) \in \mathbb{R}^{3\times N}$ denotes the predicted 3D human pose by the 3D pose reconstruction subnetwork, , N denotes the joint number.

3.4. Geometric Search Scheme

The geometric search scheme aims to further refine the coarse 3D human pose. It ensures the refined 3D human pose can be projected to the input 2D joints via the camera intrinsic matrix with zero-pixel error. Mathematically, our geometric search scheme can be represented by the following equation:

$$Q_g = f_{geo}(P_{gt}, Q_r), \tag{3}$$

where $P_{gt} = (u_{gt}, v_{gt}) \in \mathbb{R}^{2\times N}$ represents ground truth of the real 2D pose, $Q_g = (x_g, y_g, z_g) \in \mathbb{R}^{3\times N}$ is the final output of the 3D label generator.

Actually, the 3D pose reconstruction subnetwork outputs the 3D human pose aligned to the root joint (pelvis). What our model predicts in our case is 3D key-points with relative depth (relative to hip). Therefore, the projection is not possible because it requires absolution depth. According to the camera calibration principle, we propose the heuristic projection to constrain the consistence between the input 3D pose and projected 2D pose. Figure 3 shows the procedure of the geometric search scheme. Based on the z_r and P_{gt}, we initialize $\Delta z = 0$ (Δz represents the hypothetical depth to the camera). Then We can infer the $\widetilde{x}_r$ and $\widetilde{y}_r$ according to the camera intrinsic matrix and Δz, and the process of searching the optimal 3D joint can be described as follows:

Algorithm 1 Geometric Search Scheme

Require: $Q_r = (x_r, y_r, z_r)$, $P_{gt} = (u_{gt}, v_{gt})$, and camera intrinsic parameters (c_x, c_y, f_x, f_y)
Ensure: Q_g
$thres = +\infty$, $dist = 20000\text{mm}$
for $\Delta z = 0; \Delta z \leq dist$ **do**
  $\widetilde{x}_r = (u_{gt} - c_x)(z_r + \Delta z)/f_x$
  $\widetilde{y}_r = (v_{gt} - c_y)(z_r + \Delta z)/f_y$ (Eqn.4)
  $L_x = (\widetilde{x}_r - x_r)^2$, $L_y = (\widetilde{y}_r - y_r)^2$
  $L_{geo} = ||L_x + L_y||_2^2$ (Eqn.5)
  if $L\_geo \leq thres$ **then**
    $thres = L\_geo$
    $x_g = \widetilde{x}_r$, $y_g = \widetilde{y}_r$, $z_g = z_r$
return $Q_g = (x_g, y_g, z_g)$

$$\begin{aligned} \widetilde{x}_r &= (u_{gt} - c_x)(z_r + \Delta z)/f_x \\ \widetilde{y}_r &= (v_{gt} - c_y)(z_r + \Delta z)/f_y \end{aligned} \tag{4}$$

, where f_x and f_y denote focal length of the camera and c_x and c_y represent the optical axis points of the camera. By increasing Δz with $step = 1\text{mm}$, until we obtain the optimal value that can satisfy the following loss function,

$$L_{geo} = \arg\min_{\Delta z} \left\| ((\widetilde{x}_r - x_r)^2 + (\widetilde{y}_r - y_r)^2) \right\|_2^2 \tag{5}$$

As shown in Algorithm 1, we empirically set the maximum search distance to $dist = 20000\text{mm}$. 3D pose reconstruction subnetwork outputs the 3D pose ($Q_r = (x_r, y_r, z_r)$) aligned to the root joint. Therefore, we need to estimate the Δz (the same Δz for the entire pose) to make re-projection possible. The 3D pose with absolute depth can be formulate as $(x_r, y_r, z_r + \Delta z)$, and then we put it into the Eq.4 and Eq.5 to obtain the final $\widetilde{Q}_g = (\widetilde{x}_r, \widetilde{y}_r, z_r + \Delta z)$. Finally, the 3D pose aligned to the root joint is $Q_g = (x_g, y_g, z_g)$. We take the step $\Delta z = 1\text{mm}$ because the accuracy of Human3.6M is also in the millimeter level.

In this way, we can obtain a reasonable Δz, the location of the 3D pose $\widetilde{Q}_g$ in the 3D space, and the final output of the 3D label generator Q_g.

3.5. A Large-Scale in-the-Wild 3D Pose Dataset

The proposed 3D label generator can map a 2D human pose to its high quality 3D counterpart. The existing datasets for 2D human pose estimation such as Leeds Sports Pose dataset (LSP) [13], MPII human pose dataset (MPII) [3] and Ai Challenger dataset for 2D human pose estimation (Ai-Challenger) [43] can be used to extract the high quality 3D labels by the 3D label generator. Given the

well-annotated 2D key-points, the 3D label generator can output the high quality 3D labels of the in-the-wild images. Finally, we collect a large-scale in-the-wild dataset containing more than 400,000 images (320,000 training images and the rest for testing) with high quality 3D labels.

3.6. Baseline Network

We adopt the backbone of Zhou et al. [47] as the baseline network to evaluate the in-the-wild 3D pose dataset quantitatively and qualitatively. This network can be viewed as a two-stage pose estimator. The first stage is to use the stacked hourglass network [23] for 2D human pose estimation. Each stack is in an encoder-decoder structure. The second stage is a depth regression module. Given the 2D body joints heat-maps and intermediate features generated from stacked hourglass network, it can predict the depth of each joint. Since we have a large-scale in-the-wild 3D pose dataset, we discard the weakly-supervised designs employed by Zhou et al. [47] and Yang et al. [44]. We train the network only using the first two stages of the method with Human3.6M and in-the-wild 3D pose dataset.

4. Experiments

In this section, we present the experiments and results of the 3D label generator. Trained with 2D ground truth, our 3D label generator achieves state-of-the-art results on the Human3.6M dataset [12]. Experimental results denote that the generator can provide high quality labels for the in-the-wild images. Meanwhile, we investigate the efficacy of the stereoscopic view synthesis subnetwork, 3D pose reconstruction subnetwork, and geometric search scheme respectively. In addition, we compare with the methods of [47, 44] to verify the effectiveness of in-the-wild 3D pose dataset. To further verify the generalization ability of the model, we attempt to use our predicted 3D poses to the task of classification on the Penn Action dataset [46].

4.1. Datasets and Evaluation Metrics

We numerically evaluate the publicly available 3D human pose estimation dataset: Human3.6M [12]. We also conduct qualitative experiments on in-the-wild images.

3D Pose Datasets. Human3.6M is a large-scale dataset with 2D joint locations and 3D ground truth collected by the MoCap system in the laboratory environment. It consists of 3.6 million RGB images of 11 different professional actors performing 15 everyday activities. Following our baseline method [47], we employ data from subjects S1, S5, S6, S7, S8 for training and evaluate on the data from subjects S9 and S11. We refer the MPJPE that evaluated on the predicted 3D pose after alignment of the root without any rigid alignment transformation as protocol#1.

MPI-INF-3DPH [20] is a recent dataset that includes both indoor and outdoor scenes, which contains 2929 frames from six subjects performing seven actions, to evaluate the generalization ability quantitatively. We only use the test split of this dataset to demonstrate the generalization ability of the trained model.

2D Pose Datasets. MPII and LSP are the most widely used dataset for 2D human pose estimation. Ai-Challenger is proposed recently for multi-person 2D pose estimation, which consists of 210,000 images for training, 30,000 images for validation and 60,000 images for testing. We qualitatively compare the generalization ability on these 2D datasets.

Penn Action Dataset. Penn Action Dataset [46] contains 2,326 video sequences of 15 different actions, e.g., pull-up, squat and push-up, with 1,258 clips for training and 1,068 clips for testing. The performance is measured by the mean classification accuracy across the splits [35].

4.2. Implementation Details

We introduce the implementation details of the 3D label generator and the baseline network based on the RGB images for 3D human pose estimation.

3D Label Generator. We train the proposed 3D label generator using Pytorch [26] toolbox and Adam [14] solver to optimize the parameters. We set momentum, momentum2, and weight decay as 0.9, 0.99, and 10^{-4}, respectively. Kaiming initialization [7] is used to initialize the weights of our linear layers. The network is trained for a total of 200 epoch. The learning rate is set to be 10^{-3} and exponential decay. We train the stereoscopic view synthesis subnetwork with 4.8 million 2D/3D key-points pairs from the Unity toolbox and Human3.6M. We set the batch size as 64 and normalize the dataset to $[-1, 1]$. Training on a Nvidia TITAN X GPU, the network converges within one day. When training the 3D pose reconstruction subnetwork, we fix the parameters of the stereoscopic view synthesis subnetwork. We adopt the same training scheme when training the 3D pose reconstruction subnetwork.

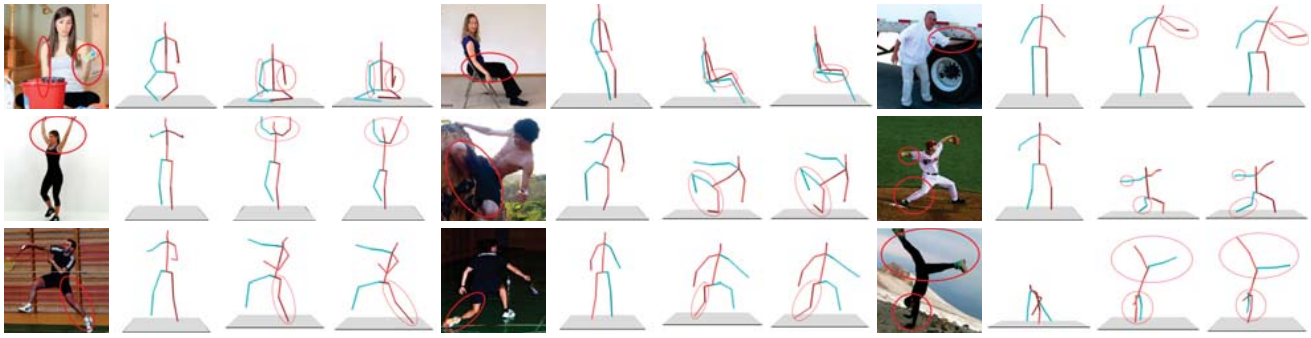
Baseline Network. Stochastic Gradient Descent (SGD) optimization is used for training. Each training batch contains both the Human3.6M and in-the-wild images in the ratio of 1:1. We fine-tune the 2D module based on the checkpoint of Zhou et al. [47] with the 2D annotated Human3.6M dataset and the in-the-wild dataset containing MPII, LSP, and Ai-Challenger. Both data from Human3.6M and the proposed in-the-wild 3D pose dataset are employed for training the two stage baseline network. We train the network with the loss following Zhou et al. [47].

4.3. Evaluations on 3D Label Generator

Quantitative Results. Table 2 denotes the comparisons with Martinez et al. [19] on the Human3.6M [12]. All the

Table 1. Quantitative evaluations on the Human3.6M [12] under Protocol#1 (no rigid alignment or similarity transform is applied in post-processing). GT indicates that the network was trained on ground truth 2D pose. GS denotes the geometric search scheme. Unity denotes the model trained with the additional 2D/3D key-points generated by the unity toolbox. The bold-faced numbers represent the best results.

Protocol#1 (↓)	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SittingD.	Smoke	Wait	WalkD	Walk	WalkT.	Average
Martinez et al. [19] (GT) w/o GS	37.7	44.4	40.3	42.1	48.2	54.9	44.4	42.1	54.6	58.0	45.1	46.4	47.6	36.4	40.4	45.5
Martinez et al. [19] (GT) w/ GS	33.1	39.8	34.5	37.5	39.5	**45.7**	40.4	31.7	44.9	49.2	37.8	39.2	39.8	**30.3**	33.8	38.5
Ours (GT) w/o GS	35.6	41.3	39.4	40.0	44.2	51.7	39.8	40.2	50.9	55.4	43.1	42.9	45.1	33.1	37.8	42.0
Ours (GT) w/ GS	**32.1**	**39.2**	**33.4**	**36.4**	**38.9**	45.9	**38.4**	**31.7**	**42.5**	**48.1**	**37.8**	**37.9**	**38.7**	30.6	**32.6**	**37.6**
Ours (GT) w/ GS + unity	36.5	42.7	38.2	39.6	45.3	50.8	40.2	34.8	45.0	50.3	39.4	39.9	42.5	32.2	33.8	40.8



(a) (b) (c) (d) (a) (b) (c) (d) (a) (b) (c) (d)

Figure 4. Qualitative evaluations on the in-the-wild images. (a) Original in-the-wild images, (b) Results of Martinez et al. [19], (c) Our results w/o geometric search scheme, (d) Our results w/ geometric search scheme. The proposed 3D label generator outperforms the method of Martinez et al. [19]. The proposed geometric search scheme can refine the coarse 3D human pose. (All the predicted 3D poses are demonstrated from the front viewpoint.)

methods are trained with 2D key-points ground truth. The experimental results show that 3D label generator boosts the performance on the Human3.6M dataset. For protocol#1, the generator trained with 2D/3D ground truth from Human3.6M has 17% (37.6mm vs. 45.5mm) improvements compared with the method of Martinez et al. [19]. To improve the generalization ability, we also train the network with synthetic 2D/3D pairs generated by the unity toolbox. There is 10% improvement compared with the method of Martinez et al. [19]. Since the domain gap between the synthetic data and the real data, the model trained with both dataset performs slight worse than the model trained with only the Human3.6M dataset. However, the qualitative performance on the in-the-wild images is significantly improved, as we will show in the following and our supplementary materials.

Qualitative Results. Compared with the method of Martinez et al. [19], we demonstrate the generalization ability qualitatively on the images from MPII and LSP. Both of the networks to estimate the 3D human pose are based on the 2D ground truth of these datasets, As shown in Figure 4 (b), (d), the generalization ability of our algorithm outperforms the generalization results of Martinez et al. [19]. Because of the 2D/3D key-points pairs generated by the unity toolbox, the generalization ability of the network is highly improved.

Validation of Stereoscopic View Synthesis Subnetwork. To evaluate the quality of synthetic right-view 2D pose, we apply the PCKh metrics following 2D human pose estimation method [23]. The subnetwork trained with merely Human3.6M achieves 98.2%, and the subnetwork trained with the Human3.6M and synthesized dataset by unity toolbox obtains 95.3% in PCKh-0.5 scores. It demonstrates that the stereoscopic view synthesis is able to generate high quality right-view 2D pose based on the left-view 2D pose.

Validation of 3D Pose Reconstruction Subnetwork. In Section 3.1, we note that the stereoscopic architecture can alleviate the depth ambiguity in 3D human pose estimation. As shown in Table 2, compared with the monocular structure of Martinez et al. [19], our network without the geometric search scheme has 7.7% (42.0mm vs. 45.5mm) improvements. Both of the networks have the same architecture and parameters, the only difference is that we take the 2D poses from multi-view as inputs. Results illustrate that the rationality of the designed stereoscopic network structure which can boost the performance in 3D human pose estimation.

Validation of the Geometric Search Scheme. We indicate that a high quality 3D human pose can be projected to its 2D counterpart with zero-pixel error. Based on this premise, we devise a geometric search scheme to further refine the coarse 3D human pose. As shown in Table 2, we further analyze the effectiveness of the geometric search scheme by comparing the performance between the method w/o or w/ using it. The experimental results validate the effectiveness of the geometric search scheme. As one can see in the figure 4(c) and (d), after the geometric search scheme,

Table 2. Quantitative evaluations on the Human3.6M [12] under Protocol#1 (no rigid alignment or similarity transform is applied in post-processing). $*$ denotes the methods using the same backbone. The bold-faced numbers represent the best results.

Protocol#1 (↓)	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SittingD.	Smoke	Wait	WalkD	Walk	WalkT.	Average
LinKDE [12]	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. [39]	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
Park et al. [25]	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
Nie et al. [24]	90.1	88.2	85.7	95.6	103.9	103.0	92.4	90.4	117.9	136.4	98.5	94.4	90.6	86.0	89.5	97.5
Metha et al. [21]	57.5	68.6	59.6	67.3	78.1	82.4	56.9	69.1	100.0	117.5	69.4	68.0	76.5	55.2	61.4	72.9
Metha et al. [22]	62.6	78.1	63.4	72.5	88.3	93.8	63.1	74.8	106.6	138.7	78.8	73.9	82.0	55.8	59.6	80.5
Lin et al. [17]	58.0	68.2	63.3	65.8	75.3	93.1	61.2	65.7	98.7	127.7	70.4	68.2	72.9	50.6	57.7	73.1
Tome et al. [40]	65.0	73.5	76.8	86.4	86.3	110.7	68.9	74.8	110.2	173.9	84.9	85.8	86.3	71.4	73.1	88.4
Tekin et al. [38]	54.2	61.4	60.2	61.2	79.4	78.3	63.1	81.6	70.1	107.3	69.3	70.3	74.3	51.8	63.2	69.7
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
Martinez et al. [19]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Fang et al. [6]	50.1	54.3	57.0	57.1	66.6	73.3	53.4	55.7	72.8	88.6	60.3	57.7	62.7	47.5	50.6	60.4
Sun et al. [36]	52.8	54.8	54.2	54.3	61.8	67.2	53.1	53.6	71.7	86.7	61.5	53.4	61.6	47.1	53.4	59.1
Yang et al. [44]	51.5	58.9	50.4	57.0	62.1	65.4	49.8	52.7	69.2	85.2	57.4	58.4	43.6	60.1	47.7	58.6
Zhou et al. [47]*	54.8	60.7	58.2	71.4	62.0	65.5	53.8	55.6	75.2	111.6	64.1	66.0	51.4	63.2	55.3	64.9
Yang et al. [44]*	53.0	60.8	**47.9**	57.1	61.5	**65.5**	50.8	49.9	73.3	**98.6**	58.8	58.1	42.0	62.3	43.6	59.7
Ours*	**47.4**	**56.4**	49.4	**55.7**	**58.0**	67.3	**46.0**	**46.0**	**67.7**	102.4	**57.0**	**57.3**	**41.1**	**61.4**	**40.7**	**58.0**

(a) (b) (c) (d) (a) (b) (c) (d) (a) (b) (c) (d)

Figure 5. Visual results on the in-the-wild images. The proposed dataset helps to generate more reasonable results in terms of 3D human skeletons. (a) Original in-the-wild images, (b) Results with geometric loss [47], (c) Results with GANs [44], (d) Results with the proposed dataset. 3D human pose presents in a novel viewpoint.

Table 3. Quantitative evaluations on the Human3.6M [12] under Protocol#1 without using the geometric search scheme and the dataset from the unity toolbox.

Δx/mm	Martinez et al. [19]	250	500	750
Ave./mm	45.5	42.2	**42.0**	42.3

the predicted 3D human poses become more reasonable.

Ablation Study We discuss the truth of Δx in (1). As shown in Table 3, we conduct experiments on 3D label generator with different Δx. The results show that our generator is not sensitive to Δx. The value of Δx around 500mm can meet our requirement for generating precise 3D labels.

4.4. Evaluations on Baseline Network

In this section, we mainly compare the two different methods [47, 44] with the baseline network trained with the proposed in-the-wild 3D pose dataset. The baseline network is an upgraded version of the Stacked Hourglass Network [23] with an additional depth regression module. To focus on the analysis of the proposed 3D pose dataset, we set all the backbone with the same components consisting of 2 stacked hourglass modules, 2 residual blocks, and 2 depth regression modules.

Quantitative Results. As shown in Table 2, trained with in-the-wild 3D pose dataset without the additional geometric loss [47] or the adversarial learning method [44],

Table 4. Quantitative evaluations on the MPI-INF-3DPH [20]. No training data from this dataset have been used for training by any method.

	Zhou et al. [47]	Yang et al. [44]	Ours
PCK	69.2	69.0	**71.2**
AUC	32.5	32.0	**33.8**

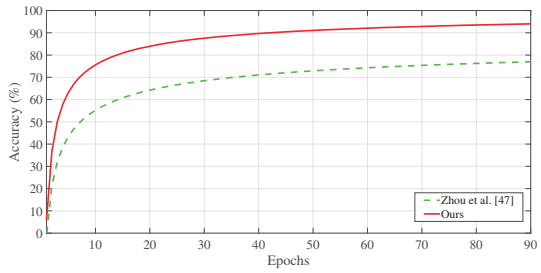


Figure 6. Evaluation results of action classification on Penn action dataset [46]. The detected 3D human pose by our method is more conducive to the action classification task.

our baseline network can outperform them. Compared with Zhou et al. [47], there is about 10.7% improvement (58.0mm vs. 64.9mm) on the Human3.6M dataset. It proves that our dataset promotes the accuracy of 3D human estimation on the images taken laboratory scenarios.

Qualitative Results. By visualizing the 3D skeleton of the predicted human body, we show that the model trained with the in-the-wild 3D pose data is robust in the realistic scenes. Figure 5 shows the visualization results by different methods. As one can see in the figure, our baseline network trained with the proposed in-the-wild 3D pose dataset can estimate more reasonable results, which in term proves the quality of our proposed dataset. In addition, we discover that our model can handle challenging samples such as leaning over, sitting cross-legged, and jumping.

Cross-Domain Generalization. We further verify the generalization introduced by our proposed 3D pose dataset on the MPI-INF-3DHP [21]. Without any retraining the model on this dataset, we compare the results by Zhou et al. [47], Yang et al. [44] and our baseline network. As reported in Table 4, one can observe that the method trained with in-the-wild 3D data significantly improves the generalization ability.

Generalization Evaluations by Action Classification. To further evaluate the generalization ability of different methods, we propose an approach to utilize detected 3D joints for action classification on the Penn Action dataset [46]. Using only the coordinate location of predicted 3D joints, we devise a simple network consist of multiple fully connected layers, the detailed network structure can be found in our supplementary file. When training this network, we use the location of predicted 3D joints of 25 consecutive frames in Penn Action dataset [46] as the inputs.

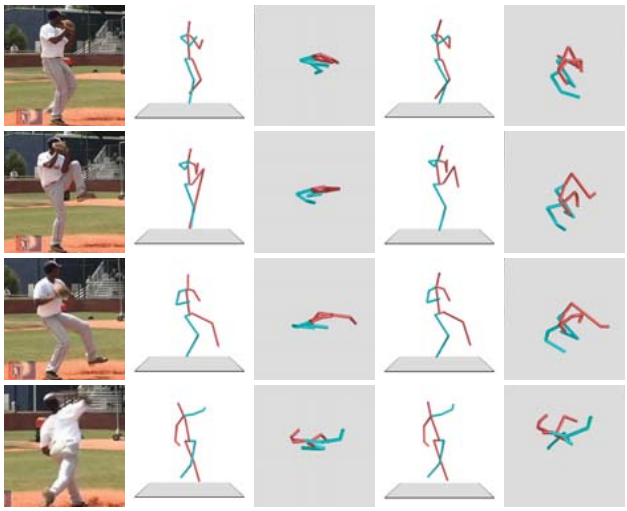


Figure 7. Visualization of the 3D skeletons of the video sequence extracted in the Penn [46] dataset. (a) Video sequences, (b) 3D human skeletons from the front viewpoint predicted by Zhou et al. [47], (c) 3D human skeletons from the top viewpoint predicted by Zhou et al. [47], (d) 3D human skeletons from the front viewpoint predicted by our method, (e) 3D human skeletons from the top viewpoint predicted by our method.

As shown in Figure 6, we can find that the model trained with our predicted 3D joints are more precise than the model of Zhou et al. [47] (93% vs. 80%). In addition, we analyze the difference between the detected 3D joints of the two models. Figure 7 shows an example that the method of Zhou et al. [47] predicts 3D pose at almost the same depth level, while the depth of the predicted 3D poses by our baseline network varies with the baseball player movements. The results demonstrate that the inaccurate predicted depth leads to a lower accuracy of action classification. The superior performance in action recognition task can further prove the generalization of our baseline network.

5. Conclusions

In this paper, we try to solve the generalization problem of 3D human pose estimation from a novel perspective. We propose a principled approach to generate high quality 3D labels given an in-the-wild image. Based on the stereo inspired structure, the proposed network with a carefully designed geometric search scheme significantly outperforms other methods quantitatively and qualitatively. We proposed an in-the-wild 3D pose dataset containing more than 400,000 images by employing this network as a 3D label generator. Experiments show that the baseline network trained with the proposed dataset can significantly improve the performance on the public Human3.6M and boost the generalization ability on the in-the-wild images.

References

[1] I. Akhter and M. J. Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[2] S. Amin, M. Andriluka, M. Rohrbach, and B. Schiele. Multi-view pictorial structures for 3d human pose estimation. In *British Machine Vision Conference*, 2013. 2, 3

[3] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 2, 4

[4] W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen. Synthesizing training images for boosting human 3d pose estimation. In *International Conference on 3D Vision*, 2016. 2

[5] X. Fan, K. Zheng, Y. Zhou, and S. Wang. Pose locality constrained representation for 3d human pose reconstruction. In *European Conference on Computer Vision*, 2014. 2

[6] H. Fang, Y. Xu, W. Wang, X. Liu, and S.-C. Zhu. Learning knowledge-guided pose grammar machine for 3d human pose estimation. *arXiv*, 2017. 1, 2, 3, 7

[7] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision*, 2015. 5

[8] M. Hofmann and D. M. Gavrila. Multi-view 3d human pose estimation in complex environment. *International Journal of Computer Vision*, 2012. 2

[9] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund. Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments. *IEEE Journal of selected topics in signal processing*, 2012. 2

[10] M. R. I. Hossain and J. J. Little. Exploiting temporal information for 3d human pose estimation. In *European Conference on Computer Vision*, 2018. 3

[11] C. Ionescu, J. Carreira, and C. Sminchisescu. Iterated second-order label sensitive pooling for 3d human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 2

[12] 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*, 2014. 1, 2, 3, 5, 6, 7

[13] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *British Machine Vision Conference*, 2010. 2, 4

[14] D. Kinga and J. B. Adam. A method for stochastic optimization. In *IEEE International Conference on Representation Learning*, 2015. 5

[15] H.-J. Lee, C. Zen, et al. Determination of 3d human-body postures from a single view. *Computer Vision Graphics and Image Processing*, 1985. 2

[16] S. Li and A. B. Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *Asian Conference on Computer Vision*, 2014. 2

[17] M. Lin, L. Lin, X. Liang, K. Wang, and H. Cheng. Recurrent 3d pose sequence machines. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 7

[18] Y. Luo, J. Ren, M. Lin, J. Pang, W. Sun, H. Li, and L. Lin. Single view stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

[19] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *IEEE International Conference on Computer Vision*, 2017. 1, 2, 5, 6, 7

[20] 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 *International Conference on 3D Vision*, 2017. 2, 5, 8

[21] 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*, 2016. 7, 8

[22] 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. *ACM Transactions on Graphics*, 2017. 7

[23] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, 2016. 5, 6, 7

[24] B. X. Nie, P. Wei, and S.-C. Zhu. Monocular 3d human pose estimation by predicting depth on joints. In *IEEE International Conference on Computer Vision*, 2017. 7

[25] S. Park, J. Hwang, and N. Kwak. 3d human pose estimation using convolutional neural networks with 2d pose information. In *European Conference on Computer Vision*, 2016. 7

[26] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *Neural Information Processing Systems Workshops*, 2017. 5

[27] G. Pavlakos, X. Zhou, and K. Daniilidis. Ordinal depth supervision for 3D human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[28] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 7

[29] L. Pishchulin, A. Jain, M. Andriluka, T. Thormählen, and B. Schiele. Articulated people detection and pose estimation: Reshaping the future. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 2

[30] V. Ramakrishna, T. Kanade, and Y. Sheikh. Reconstructing 3d human pose from 2d image landmarks. In *European Conference on Computer Vision*, 2012. 2

[31] H. Rhodin, J. Spörri, I. Katircioglu, V. Constantin, F. Meyer, E. Müller, M. Salzmann, and P. Fua. Learning monocular 3d human pose estimation from multi-view images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 3

[32] G. Rogez and C. Schmid. Mocap-guided data augmentation for 3d pose estimation in the wild. In *Neural Information Processing Systems*, 2016. 2

[33] Y. Shi, X. Han, N. Jiang, K. Zhou, K. Jia, and J. Lu. Fbipose: Towards bridging the gap between 2d images and 3d human poses using forward-or-backward information. *arXiv*, 2018. 2

[34] L. Sigal, A. O. Balan, and M. J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 2010. 2

[35] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Neural Information Processing Systems*, 2014. 5

[36] X. Sun, J. Shang, S. Liang, and Y. Wei. Compositional human pose regression. In *IEEE International Conference on Computer Vision*, 2017. 2, 7

[37] C. J. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *Computer Vision and Image Understanding*, 2000. 2

[38] B. Tekin, P. Marquez Neila, M. Salzmann, and P. Fua. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *IEEE International Conference on Computer Vision*, 2017. 2, 7

[39] B. Tekin, A. Rozantsev, V. Lepetit, and P. Fua. Direct prediction of 3d body poses from motion compensated sequences. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 7

[40] D. Tome, C. Russell, and L. Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 7

[41] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[42] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao. Robust estimation of 3d human poses from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 2

[43] J. Wu, H. Zheng, B. Zhao, Y. Li, B. Yan, R. Liang, W. Wang, S. Zhou, G. Lin, Y. Fu, et al. Ai challenger: A large-scale dataset for going deeper in image understanding. *arXiv*, 2017. 2, 4

[44] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang. 3d human pose estimation in the wild by adversarial learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 5, 7, 8

[45] H. Yasin, U. Iqbal, B. Kruger, A. Weber, and J. Gall. A dual-source approach for 3d pose estimation from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[46] W. Zhang, M. Zhu, and K. G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *IEEE International Conference on Computer Vision*, 2013. 5, 8

[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 *IEEE International Conference on Computer Vision*, 2017. 1, 2, 5, 7, 8

[48] X. Zhou, S. Leonardos, X. Hu, and K. Daniilidis. 3d shape estimation from 2d landmarks: A convex relaxation approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[49] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei. Deep kinematic pose regression. In *European Conference on Computer Vision*, 2016. 2

[50] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 7