Evaluation of Zhou et al.: “Towards 3D Human Pose Estimation in the Wild: a Weakly-supervised Approach”[28]

Marvin Alles
Technical University of Munich
marvin.alles@tum.de

Supervisor: Philip Häusser
Chair for Computer Vision & Artificial Intelligence
Technical University of Munich

Abstract

In this paper the proposed method by Zhou et al. on 3D human pose estimation in the wild [28] is evaluated. The task of 3D human pose estimation is challenging due to lack of training data and ambiguity of recovering 3D information from 2D images. Zhou. et al. provide a weakly-supervised transfer learning approach, which seamlessly integrates 2D and 3D labels with a geometric constraint. The network can be trained end-to-end and is based on an state-of-the-art 2D pose estimation network. Thereby 3D poses from controlled lab environments are combined with 2D in the wild images to fully exploit the correlation between 2D pose and depth estimation. Through their PyTorch implementation it was possible to overall reproduce the results, though their experimental evaluation leaves room for discussion: They do not use the MPI-INF-3DHP dataset[16] for training and furthermore they do not compare their method to Popa et al. [20]. Thereby the proposed method lacks important evaluation even though it might be superior to earlier approaches.

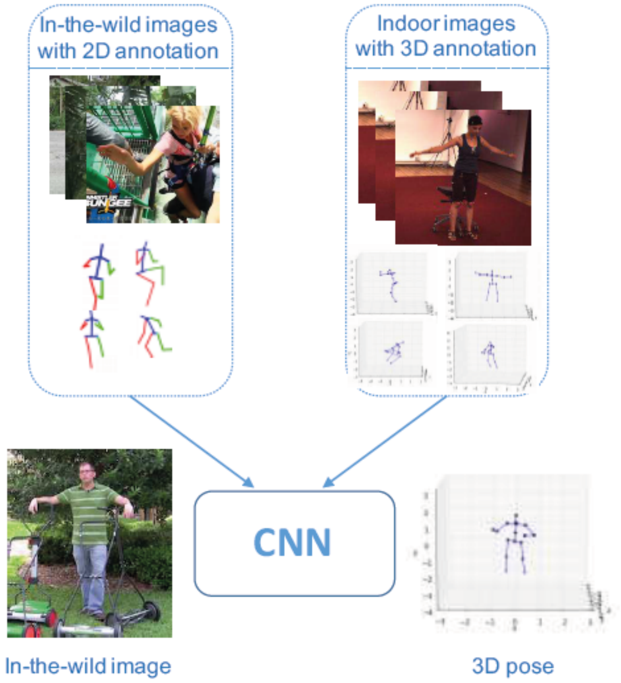

Figure 1. The proposed method combines in-the-wild images with 2D annotation and indoor images with 3D annotation to train a Convolutional Neural Network and later on predict the 3D pose given an in-the-wild image. [28]

1. Introduction

Human pose estimation has been studied heavily in the past as there are multiple interesting use-cases: Human-computer interaction, virtual reality applications, markerless motion capture, visual surveillance or ergonomic studies. Thereby the research can be divided into the field of 2D and 3D human pose estimation. The former recently improved a lot [17, 25, 12, 5, 9], though the more challenging task of 3D pose estimation is still limited. One of the main reasons are lack of training data and ambiguity of recovering 3D poses from 2D datasets. 3D datasets are captured just indoors with a high speed motion capturing system in a controlled lab environment [13, 22] and thus generalization on poses in the wild is challenging. For example do lab images not take changing backgrounds or clothes into account. On the other hand there are multiple datasets of 2D images in-the-wild available resulting in recent advances of 2D pose estimation.

The method proposed by Zhou et al. builds up on Metha

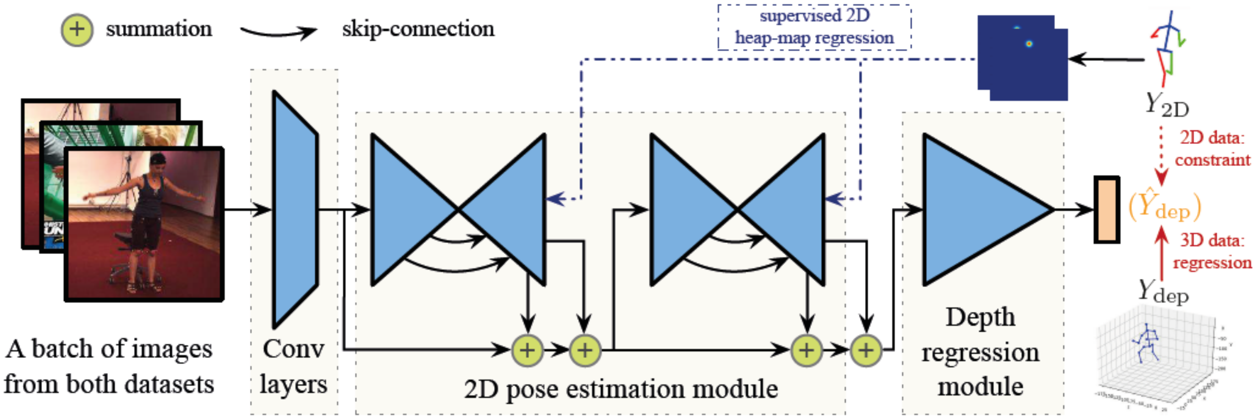

Figure 2. Illustration of the framework proposed by Zhou et al.[28]: It consists of a 2D Pose Estimation Module and a Depth Regression Module with shared features.

et al.[16] who have demonstrated the 2D-to-3D knowledge transfer and thereby show that 2D and 3D pose estimation tasks could share common representations. Thats why Zhou et al. constitute the knowledge transfer from 3D indoor to in-the-wild images and introduce a weakly supervised transfer learning approach with fully annotated images taken in a lab environment as the source domain and weakly-labeled images in-the-wild as target domain. Like previous work [30, 23, 7, 4, 26, 27] the framework is based on a 2D and a 3D module. Though unlike in those networks, which just feed the output of the 2D module as input to the 3D module, Zhou et al. connect intermediate layers of the 2D module to the 3D module for the purpose of shared representations. Furthermore end-to-end training with both 2D and 3D data differentiates their work from previous approaches. In addition the weakly supervised learning of the 3D module is regularized through a geometric constraint, which is based on the fact that relative bone lengths in a human skeleton stay more or less the same.

2. Related Work

Human pose estimation has been matter in multiple elaborations. For 3D human pose estimation with well-labeled 3D datasets, the problem can be formulated as a supervised learning problem. Though there are weakly-/unsupervised approaches, which will be introduced as well. Related work will be focused on the methods later on used for evaluation of the proposed method and can be divided into multiple categories:

No images in the wild. First of all there are multiple approaches which can not handle images in the wild. They use data from controlled lab environments for training and thereby do not generalize to in the wild settings. Zhou et al. [31] propose a method using a generative forward-kinematic layer to include a bone-length constraint. Pavlakos et al. [19] regress a volumetric representation of a 3D skeleton.

Synthetic data. The same goes for data generated using a human template model. [8, 21] Because of the challenge of modeling a 3D environment they do not generalize to in-the-wild images.

Estimation from 2D joints. Various different methods firstly generate 2D joint positions and afterwards estimate the 3D pose based on them and thereby lack an intermediate feature representation. Tome et al. [23] introduce a way to generate a 3D model from 2D heat-maps and then improves on them by combining 3D pose projection and image features. Chen et al. [7] apply nearest-neighbor search for matching estimated 2D poses to 3D poses.

Images in the wild. Mixed 2D and 3D pose estimation like the proposed method has been introduced before as well. Metha et al. [16] fine tuned a 2D pose estimation network with 3D data and Popa et al. [20] applied a framework for multi-task learning of 2D and depth regression with varying data. Though the proposed method is different by seamlessly integrating 2D and 3D data with a weakly supervised loss.

Weakly-/unsupervised constraints. In the absence of sufficient training data weakly-/unsupervised constraints among the prediction are a common approach. [18, 24, 11] Nevertheless it is new to use a geometric constraint for pose estimation in the wild.

3. Contribution

The proposed method aims to estimate the 3D joint coordinates of a human pose provided a 2D RGB image by training the network with both 2D and 3D images. Following the convention the first two 3D coordinates of the joint position are represented in pixel coordinates and the third one as joint depth. As illustrated in figure 2 the network is divided in a 2D Pose Estimation Module and Depth Regression Module

3.1. Architecture

Convolutional Layers. The first part of the Architecture, the Convolutional Layers, is implemented to preprocess the input images and reduce the overall size for the next steps.

2D Pose Estimation Module. The 2D Pose Estimation Module is based on stacked state-of-the-art hourglass networks as introduced by Newell et al.[17]. For reasons of speed a shallow version of the stacked hourglass, that consists of 2 stacks with 2 residual modules, is used.[19] The module is directly connected to the upstream convolutional layer. As output, heat-maps with the probability distribution of joint presence are generated. With one map for each joint, the predicted positions in the 2D pose $\hat{Y}_{2D}$ can be determined by the peak locations of the heat-maps.

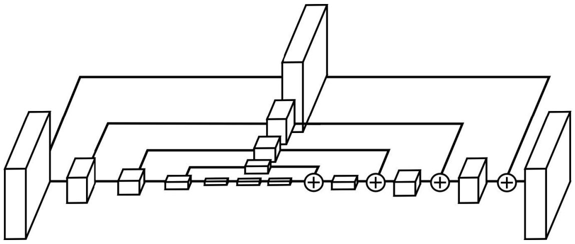

Figure 3. Illustration of the Hourglass Module by Moeslund et al.[17]: Each box is indicating a Residual Module (Figure 4)

The network is convolutional and aims to combine features across multiple resolutions by pooling and subsequent upsampling. As indicated in Figure 3 the network branches off and applies more convolutions at pre-pooled resolutions. Furthermore to later on bring together sets of adjacent resolutions nearest neighbor upsampling and element wise addition is used. Each box in figure 3 indicates a Residual Module as introduced by Newell et al. [17]. The Residual Module is displayed in figure 4 and consists of three convolutional steps and summation of the additional branch in the end. Using two stacked hourglasses allows to further evaluate high level features by processing them again and reassess higher order spatial relationships.[17]

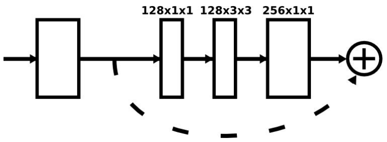

Figure 4. Residual Module by Zhang et al. [10]

Depth Regression Module. The architecture of the Depth Regression Module is based on the top-down part of one hourglass and consists of 4 sequential residual and pooling modules. The input is generated by summation of the 2D heat-maps and lower-layer image features of the 2D Pose Estimation Module for better shared representations.

3.2. Loss Functions

2D Pose Estimation Module. The 2D Pose Estimation Module is trained on the loss function (eq. 1). Therefore the L^2 distance between the predicted heat-maps $\hat{Y}_{HM}$ and heat-maps of the ground truth Y_{2D} processed by a Gaussian kernel is calculated.

$$L_{2D}(\hat{Y}_{HM}, Y_{2D}) = \sum_{h}^{H}\sum_{w}^{W}(\hat{Y}_{HM}^{(h,w)} - G(Y_{2D})^{(h,w)})^2 \quad (1)$$

Depth Regression Module. Since the Depth Regression Module needs to handle both 2D and 3D datasets, the loss function (eq. 2) depends on the dataset dimension.

$$L_{dep}(\hat{Y}_{dep}|I, Y_{2D}) = \begin{cases} \lambda_{reg}||Y_{dep} - \hat{Y}_{dep}||^2, if I \in I_{3D} \\ \lambda_{geo}L_{geo}(\hat{Y}_{dep}|Y_{2D}), if I \in I_{2D} \end{cases} \quad (2)$$

Euclidean loss between ground truth Y_{dep} and predicted $\hat{Y}_{dep}$ is used as measurement if the dataset is 3D. Furthermore, for calculating the loss of 2D datasets a geometric constraint is implemented as effective regularization for depth prediction. The concept behind this solution is, that ratios between bone lengths remain relative fixed in a human skeleton.

$$L_{geo}(\hat{Y}_{dep}|Y_{2D}) = \sum_{i} \frac{1}{|R_i|} \sum_{e \in R_i} (\frac{l_e}{\bar{l}_e} - \bar{r}_i)^2 \quad (3)$$

$$\bar{r}_i = \frac{1}{|R_i|} \sum_{e \in R_i} \frac{l_e}{\bar{l}_e} \quad (4)$$

R_i - set of involved bones

l_e - length of bone e

$\bar{l}_e$ - length of bone e in a canoncial skeleton

Therefore 4 groups of bones are considered: R_{arm} = (left/right lower/upper arms), R_{leg} = (left/right lower/upper

legs), $R_{shoulder}$ = (left/right shoulder bones), R_{hip} = (left/right hip bones). Thereby bones of different groups do not influence each other. Torso bones have been excluded, because of high variance.
The so called geometric loss L_{geo} (eq. 3) is the sum of variance among $\frac{l_e}{\bar{l}_e}$. In addition, continuous and differentiable with respect to $\hat{Y}_{dep}$. L_{geo} is defined on Y_{2D} instead of $\hat{Y}_{2D}$. Therefore back propagation into the 2D module is not necessary and thus leading to easier training.

Overall Loss. For combining loss functions of the Depth Regression Module and the 2D Pose Estimation module, the sum is calculated. This leads to the following overall loss for each training image $I \subset I_{2D} \cup I_{3D}$:

$$L(\hat{Y}_{HM}, \hat{Y}_{dep}|I) = L_{2D}(\hat{Y}_{HM}, Y_{2D}) + L_{dep}(\hat{Y}_{dep}|I, Y_{2D}) \tag{5}$$

3.3. Training

The framework is trained on the combined loss function (eq. 5) with stochastic gradient descent: Moreover, random sampled mini-batches consisting of half 2D and half 3D datasets.
Because of dependency between the Depth Regression Module and the 2D Pose Estimation Module and high nonlinearity of the geometric constraint, direct end-to-end training does not work. Therefore they make use of a 3 stage training scheme:

Stage 1: Initialize 2D pose module using 2D images like[17]

Stage 2: Initialize 3D pose module and fine-tuning 2D pose module by 2D and 3D images with inactive geometric constraint ($\lambda_{geo} = 0$)

Stage 3: Fine tuning of the whole network with activated geometric constraint

4. Experimental Evaluation

For experimental evaluation the framework has been trained by MPII(training) and Human3.6M(training) and evaluated using MPII(testing), Human3.6M(testing) and MPI-INF-3DHP(testing) datasets and the following metrics:

MPJPE Mean per joint position error in mm

PCK Percentage of correct key-points with a threshold of 150mm

AUC Area under the Receiver Operating Characteristic Curve

Symmetry Symmetric bone lengths' difference: Normalize 2D joints in 256x256 pixels. Depth normalized by same scale. Calculation of L1 distance between left and right symmetric bones e.g. $||Y^{(leftshoulder)} - Y^{(leftelbow)}|| - ||Y^{(rightshoulder)} - Y^{(rightelbow)}||$

Furthermore the experimental evaluation is separated in supervised 3D human pose estimation and transferred 3D human pose estimation in the wild.

4.1. Implementation

The framework was implemented in torch7 and trained with the 3 stage scheme explained in 3.3. Thereby 240k iterations with batch size of 6 were used for Stage 1. Overall the performance of the 2D pose estimation module is comparable to [17]. Training stage 2 took 200k and stage 3 40k iterations, resulting in a total of about 2 days training on one TitanX GPU with CUDA 8.0 and cudnn 5. A single forward pass for testing takes about 30ms. Furthermore, the loss function constants have been set as following: $\lambda_{reg} = 0.1$ and $\lambda_{geo} = 0.01$. The other hyper-parameters were set like in [17].

4.2. Datasets

The proposed method combines 2D images in-the-wild (MPII) with 3D images taken in an indoor environment (Human3.6M) for training. For evaluation the MPII and Human3.6M datasets are used again and furthermore the newly proposed MPI-INF-3DHP dataset.

MPII. (Max Planck Institute for Informatics Human Pose dataset) [3] The dataset contains around 25k training and 2958 validation images from human poses in-the-wild with annotated body joints and body part inclusions. All images were extracted from online videos. [3]

Human3.6M. [13] [6] The dataset provides 3.6 million 3D human poses and corresponding RGB images. They have been captured indoors with a high-speed motion capture system, so that accurate joint positions are included. The provided 50fps video has been down-sampled to 10fps for reducing redundancy. Like in [15] [29] [30] the subjects S1, S5, S6, S7 and S8 were utilized for training, and S9 and S11 for testing.

$$\hat{Y} = (Y_{out} - Y_{out}^{(root)}) * \frac{AvgSumLen}{SumLen_{out}} + Y_{GT}^{(root)} \tag{6}$$

Y_{out} - output of the network

$AvgSumLen$ - average sum-of-skeleton-length

$SumLen_{out}$ - sum-of-skeleton-length of output joints

	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases
Chen Ramanan[7]	89.87	97.57	89.98	107.87	107.31	139.17	93.56	136.09
Tome et al.[23]	64.98	73.47	76.82	86.43	86.28	110.67	68.93	74.79
Zhou et al.[31]	87.36	109.31	87.05	103.16	116.18	143.32	106.88	99.78
Metha et al.[16]	59.69	69.74	60.55	68.77	76.36	85.42	59.05	75.04
Pavlakos et al.[19]	58.55	64.56	63.66	**62.43**	66.93	70.74	57.72	62.51
3D/wo geo	73.25	79.17	72.35	83.90	80.25	81.86	69.77	72.74
3D/w geo	72.29	77.15	72.60	81.08	80.81	77.38	68.30	72.85
3D+2D/wo geo	55.17	61.16	**58.12**	71.75	62.54	67.29	54.81	56.38
3D+2D/w geo	**54.82**	**60.70**	58.22	71.41	**62.03**	**65.53**	**53.83**	**55.58**
	Sitting	SittingDown	Smoking	Waiting	WalkDog	Walking	WalkPair	Average
Chen Ramanan[7]	133.14	240.12	106.65	106.21	87.03	114.05	90.55	114.18
Tome et al.[23]	110.19	172.91	84.95	85.78	86.26	71.36	73.14	88.39
Zhou et al.[31]	124.52	199.23	107.42	118.09	114.23	79.39	97.70	79.9
Metha et al.[16]	96.19	122.92	70.82	68.45	54.41	82.03	59.79	74.14
Pavlakos et al.[19]	76.84	**103.48**	65.73	**61.56**	67.55	**56.38**	59.47	66.92
3D/wo geo	98.41	141.60	80.01	86.31	61.89	76.32	71.47	82.44
3D/w geo	93.52	131.75	79.61	85.10	67.49	76.95	71.99	80.98
3D+2D/wo geo	**74.79**	113.99	64.34	68.78	52.22	63.97	57.31	65.69
3D+2D/w geo	75.20	111.59	**64.15**	66.05	**51.43**	63.22	**55.33**	**64.90**

Table 1. Mean per joint position error (MPJPE) on Human3.6M Dataset [28] in mm

In training dataset 3D and 2D poses are aligned through the use of the ground truth 2D joint locations: Whereas in testing the calibration of 3D and 2D poses is not required, because of the method described in [19] [31]: The sum of all 3D bone lengths has to be equal to a predefined canonical skeleton.

MPI-INF-3DHP. (Max Planck Institute for Informatics 3D Human Pose dataset) [16] The dataset is newly proposed in 2017 and contains 3D human pose images captured by a motion capture system in outdoor and indoor environments. Over 1.3 million frames are included in the dataset, though only the test set of 2929 images was used. Because of different definitions compared to MPII and Human 3.6M, pelvis and hip position have been moved in the ratio of 0.2 towards neck.

4.3. Baselines for Ablation Study

For experimental evaluation of the proposed method three baseline models were created:

3D/wo geo: Just 3D data used for training of *Stage 2* and *Stage 3* (sec. 3.3). The 2D pose estimation module is still trained on 2D data like in *Stage 1*. No use of in-the-wild images.

3D/w geo: Same as first baseline, but including geometric constraint induced loss.

3D/wo geo	3D/w geo	3D+2D/wo geo	3D+2D/w geo
90.01%	90.57%	90.93%	91.62%

Table 2. 2D pose accuracy on Human3.6M dataset [28]

3D+2D/wo geo: Like the proposed method, but without the geometric constrain induced loss for 2D data during training of the Depth Regression Module.

3D+2D/w geo: Proposed method.

4.4. Supervised 3D Human Pose Estimation

Human3.6M. Based on the Human 3.6M testing dataset the results of the proposed and baseline methods are analyzed and compared to state-of-the-art frameworks. Table 1 displays the MPJPE for different settings. As indicated by bold numbers, the methods handle poses with varying degrees. Though overall the average MPJPE is the key indicator.

Compared by mean MPJPE the method by Chen & Ramanan [7] leads to the highest difference of $114.18mm$. The baselines which only uses 3D labeled data in training *Stage 2* and *Stage 3* (*3D/wo geo* and *3D/w geo*) can already compete with state-of-the-art methods by Tome et al. [23], Zhou et al. [31] and Metha et al. [16]. Thereby the network by Metha et.al [16] is similar to *3D/wo geo*. His approach is fine tuning a 2D pose network [12] with 3D labeled data, but differs to the proposed method by using a learning rate decay for transferred layers. The decay was also implemented, but resulted in worse performance. The

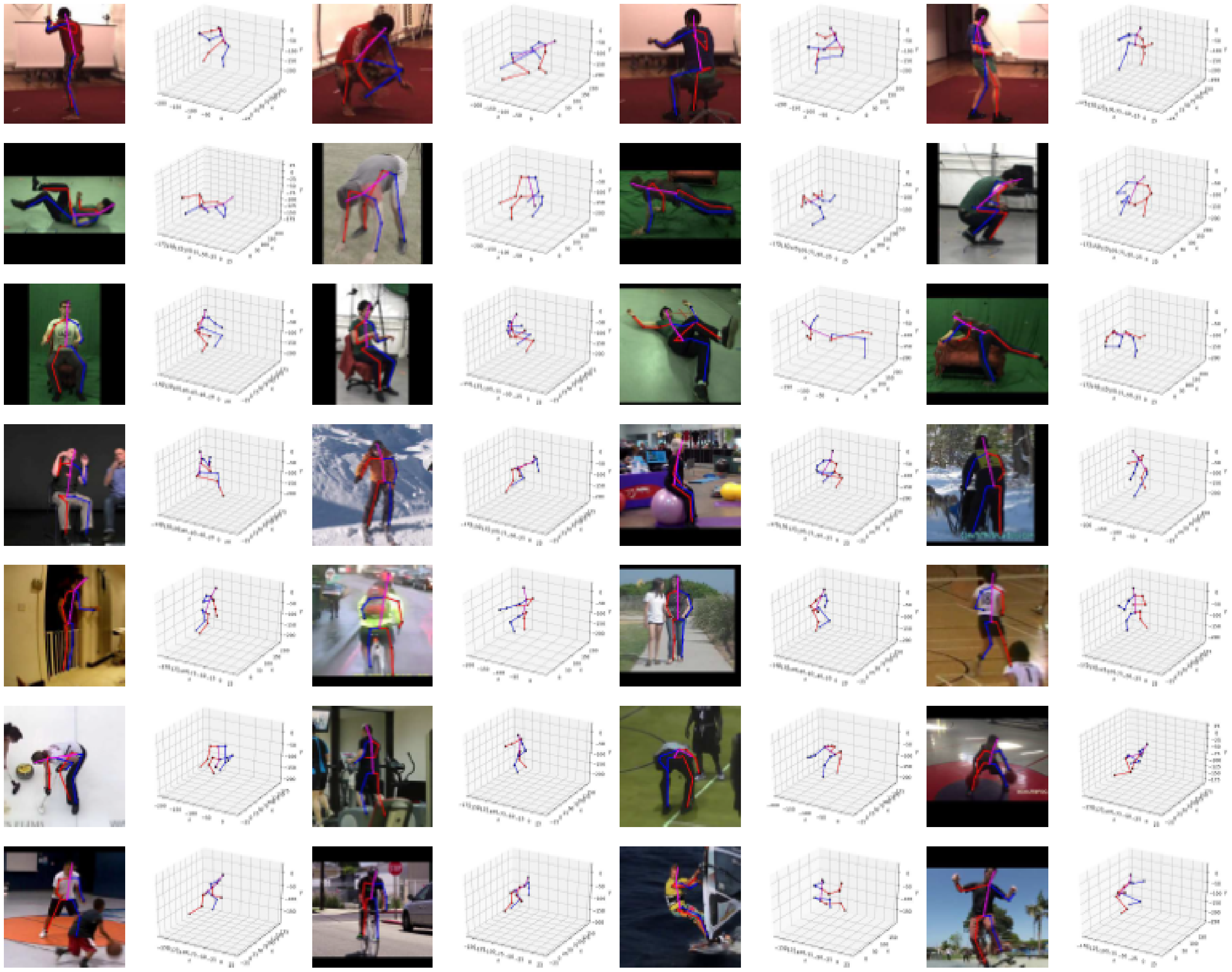

Figure 5. Qualitative results [28] - First line: Human 3.6M dataset; Second and third line: MPI-INF-3DHP dataset; Fourth to seventh line: MPII dataset

average of $80.98mm$ for *3D/w geo* compared to $82.44mm$ for *3D/wo geo* indicates that a geometric constraint increases the performance. Both baselines with 2D and 3D training datasets for *Stage 2* and *Stage 3* (*3D+2D/wo geo* and *3D+2D/w geo*) have a lower average MPJPE and thereby a better performance than all other state-of-the-art methods. With $64.90mm$ the proposed method yields to the best performance and demonstrates the benefits of the geometric constraint. Furthermore, since the constraints are applied on the disjoint 2D dataset, it shows that the prior knowledge is universal.

In table 2 the results of applying the standard metric PCKh0.5 [3] to the 2D Pose Estimation Module are displayed. The increase of accuracy between all baselines is negligible. Thus, it is indicating that adding 2D training data essentially helps the Depth Regression Module through shared future representation and does not increase the 2D accuracy.

4.5. Transferred Human Pose in the Wild

For evaluation of generalization performance of the proposed method testing datasets from MPI-INF-3DHP and MPII were used to actually analyze the results on pose estimation in the wild.

MPI-INF-3DHP In table 3 results of the baseline methods and Mehta et al. [16] are shown. The table is sorted by scene (studio with greenscreen, studio without greenscreen, outdoor) and includes the metrics PCK and AUC. Training only by 3D indoor images like *3D/wo geo* and *3D/w geo* yields to a low performance. Though even then the geometric constraint improves the results. *3D+2D/wo geo* already excels the network by Metha et al.

	Studio GS	Studio no GS	Outdoor	ALL PCK	AUC
Metha et al.(H36M+MPII) [16]	70.8	62.3	58.8	64.7	31.7
3D/wo geo	34.4	40.8	13.6	31.5	18.0
3D/w geo	45.6	45.1	14.4	37.7	20.9
3D+2D/wo geo	68.8	61.2	67.5	65.8	32.1
3D+2D/w geo	71.1	64.7	72.7	**69.2**	**32.5**
Metha et al.(MPI-INF-3DHP) [16]	84.1	68.9	59.6	**72.5**	**36.9**

Table 3. Evaluation of MPI-INF-3DHP [28]. GS is indicating a green screen background

	3D+2D/wo geo	3D+2D/w geo
Upper arm	42.4mm	**37.8mm**
Lower arm	60.4mm	**50.7mm**
Upper leg	43.5mm	**43.4mm**
Lower leg	59.4mm	**47.8mm**
Upper arm	6.27px	**4.80px**
Lower arm	10.11px	**6.64px**
Upper leg	6.89px	**4.93px**
Lower leg	8.03px	**6.22px**

Table 4. Left right symmetry [28] - MPI-INF-3DHP (Top) and MPII-Validation set (Bottom).

[16] trained on the Human 3.6M and MPII dataset with PCK of 69.2 compared to 72.5 (same goes for AUC). Only Metha et al.[16] trained on the MPI-INF-3DHP dataset is able to reach a higher performance. Still the results by the proposed method are close. Therefore the ability on in-the-wild images of the proposed method is confirmed. Furthermore table 4 (upper part) shows the evaluation of left-right symmetry (4.1) on *3D+2D/wo geo* and *3D+2D/w geo* for the MPI-INF-3DHP dataset. The geometric constraint yields to a better performance and so improves geometric validity.

MPII Figure 5 shows the qualitative results for all 3 datasets. Even for the challenging MPII dataset they seem valid. Further, table 4 (bottom) shows the results of the left-right symmetry evaluation, demonstrating again the improvement of geometric validity through the geometric constraint.

5. Personal Evaluation

To further evaluate and reproduce the results the following PyTorch implementation was used: `https://github.com/xingyizhou/Pytorch-pose-hg-3d`

In figure 6 two images from the datasets Human3.6M and MPI-INF-3DHO (figure 5 - first and third one in the first column) are used again to generate the 3D pose. When comparing to the actual results in the paper some differences for the right arm of the boxer and right leg of the guy sitting will become visible. The use of the PyTorch implementation instead of Torch7 or a different pre-trained model might be the reason for the differences. Still it is possible to reproduce the overall qualitative results.

For the next step of evaluation two new pictures were used. The pose of the tennis player is generated quite well (Figure 7 Top), whereby the one of the skier (Figure 7 Bottom) is wrong on the right arm. Comparing both it shows that a clear background makes a great difference. Multiple people with similar outfits are especially difficult to process.

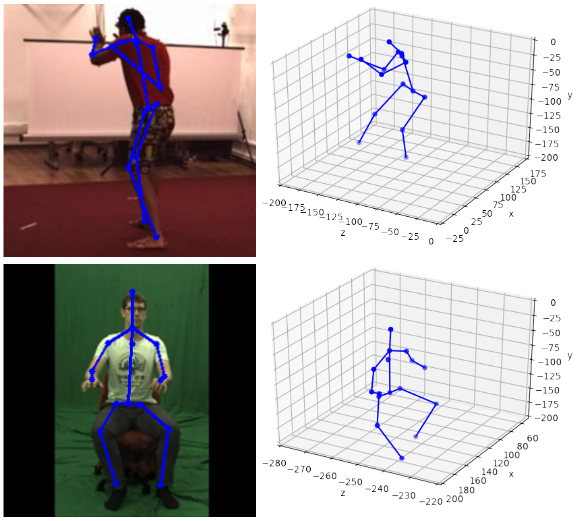

Figure 6. Results from Human3.6M [13] (Top) and MPI-INF-3DHP [16] (Bottom).

6. Conclusion

Overall the proposed method by Zhou et al. [28] generates reasonable results even for challenging situations like partly hidden body parts or unclear backgrounds. Furthermore, they use for the first time and end-to-end framework which seamlessly integrates 2D outdoor and 3D indoor labels through a weakly-supervised geometric constraint. Nonetheless the method leaves room for improvement: Especially use cases like ergonomic studies would benefit on a more accurate joint position since an average MPJPE

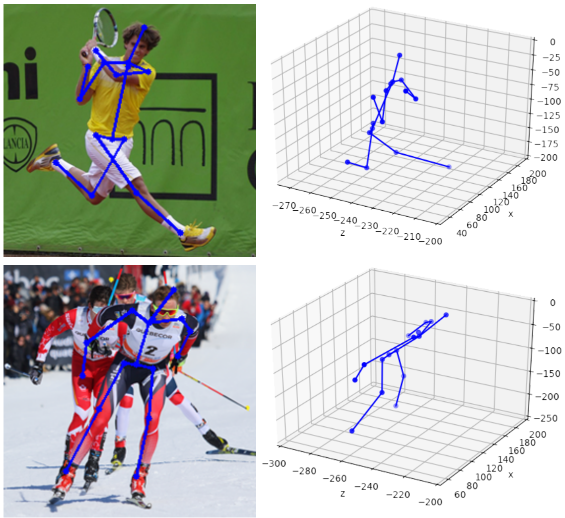

Figure 7. Results from additional images, not included in the datasets. (Bottom [1] and Top [2])

of $64.9mm$ (table 1) is good compared to other methods, but still relatively high. One way to improve could be to introduce more un-/weakly supervised constraints as indicated in the paper. The personal evaluation through the PyTorch implementation overall confirms the results by Zhou et al.

Nevertheless there are some important remarks which need to be discussed: First of all Zhou et al. mention that 3D pose estimation in the wild is challenging because of the lack of training data: "This task is challenging due to lack of training data, as existing datasets are either in the wild images with 2D pose or in the lab images with 3D pose."[28] But on the other hand they only make use of 2929 images from the test set of MPI-INF-3DHP neglecting $\sim$1.3M frames of 3D indoor and outdoor images, which are augmented but anyway better than just controlled lab images [16]. It would have been very interesting and necessary to see PCK and AUC of the proposed method if trained additionally on MPI-INF-3DHP as shown in table 3 for Metha. Furthermore they neglect the results of the method by Metha et al. trained on Human3.6M, MPII and MPI-INF-3DHP which is with an PCK of 76.5 and AUC of 40.8 [16] significantly better than the proposed method. By doing so Zhou et al. lack an important part of verification, even though their proposed method might be superior.
Lastly they mention the approach by Popa et al.[20] in section Related Work, but do not show any experimental evaluation compared to the proposed method. Popa et al. train the network on slightly different datasets: MPII and LSP (Leeds Sport Pose) [14] for 2D labels and Human3.6M and HumanEva[22] for 3D labels [20]. Validation on Human80K, a subset of Human3.6M, reaches an average MPJPE of $63.5mm$ which is compared to the proposed method with $64.90mm$ even lower (table 1). The slightly different training sets could be a reason that the Popa et al. method seems to be superior, but anyway Zhou et al. should evaluate on that to fully show the benefit of their proposed method.

References

[1] 2017 ski tour canada quebec city 17. `https://commons.wikimedia.org/wiki/File:2017_Ski_Tour_Canada_Quebec_city_17.jpg`.

[2] Jugend, spieler, tennis, sport, florenz, ballspiel, tennisspieler, schlaegersport. `https://pxhere.com/de/photo/1183116`.

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

[4] F. Bogo, A. Kanazawa, C. Lassner, P. V. Gehler, J. Romero, and M. J. Black. Keep it SMPL: automatic estimation of 3d human pose and shape from a single image. *CoRR*, abs/1607.08128, 2016.

[5] A. Bulat and G. Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. *CoRR*, abs/1609.01743, 2016.

[6] C. S. Catalin Ionescu, Fuxin Li. Latent structured models for human pose estimation. *International Conference on Computer Vision*, 2011.

[7] C. Chen and D. Ramanan. 3d human pose estimation = 2d pose estimation + matching. *CoRR*, abs/1612.06524, 2016.

[8] W. Chen, H. Wang, Y. Li, H. Su, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen. Synthesizing training images for boosting human 3d pose estimation. *CoRR*, abs/1604.02703, 2016.

[9] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. Multi-context attention for human pose estimation. *CoRR*, abs/1702.07432, 2017.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[11] J. Hoffman, D. Wang, F. Yu, and T. Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *CoRR*, abs/1612.02649, 2016.

[12] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. *CoRR*, abs/1605.03170, 2016.

[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.

[14] 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.

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

[16] 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. *3D Vision (3DV), 2017 Fifth International Conference on*, 2017.

[17] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. *CoRR*, abs/1603.06937, 2016.

[18] D. Pathak, P. Krähenbühl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. *CoRR*, abs/1506.03648, 2015.

[19] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. *CoRR*, abs/1611.07828, 2016.

[20] A. Popa, M. Zanfir, and C. Sminchisescu. Deep multitask architecture for integrated 2d and 3d human sensing. *CoRR*, abs/1701.08985, 2017.

[21] G. Rogez and C. Schmid. Mocap-guided data augmentation for 3d pose estimation in the wild. *CoRR*, abs/1607.02046, 2016.

[22] 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.

[23] D. Tomè, C. Russell, and L. Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. *CoRR*, abs/1701.00295, 2017.

[24] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. *CoRR*, abs/1510.02192, 2015.

[25] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. *CoRR*, abs/1602.00134, 2016.

[26] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman. Single image 3d interpreter network. *CoRR*, abs/1604.08685, 2016.

[27] H. Yasin, U. Iqbal, B. Krüger, A. Weber, and J. Gall. 3d pose estimation from a single monocular image. *CoRR*, abs/1509.06720, 2015.

[28] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Weakly-supervised transfer for 3d human pose estimation in the wild. *CoRR*, abs/1704.02447, 2017.

[29] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei. Deep kinematic pose regression. *CoRR*, abs/1609.05317, 2016.

[30] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. *CoRR*, abs/1511.09439, 2015.

[31] X. Zhou, M. Zhu, G. Pavlakos, S. Leonardos, K. G. Derpanis, and K. Daniilidis. Monocap: Monocular human motion capture using a CNN coupled with a geometric prior. *CoRR*, abs/1701.02354, 2017.