Deductive Learning for Weakly-Supervised 3D Human Pose Estimation via Uncalibrated Cameras

Xipeng Chen[1], **Pengxu Wei**[1*], **Liang Lin**[1, 2]
[1] Sun Yat-Sen University
[2] DarkMatter AI Research
chenxp37@mail2.sysu.edu.cn, weipx3@mail.sysu.edu.cn, linliang@ieee.org

Abstract

Without prohibitive and laborious 3D annotations, weakly-supervised 3D human pose methods mainly employ the model regularization with geometric projection consistency or geometry estimation from multi-view images. Nevertheless, those approaches explicitly need known parameters of calibrated cameras, exhibiting a limited model generalization in various realistic scenarios. To mitigate this issue, in this paper, we propose a Deductive Weakly-Supervised Learning (DWSL) for 3D human pose machine. Our DWSL firstly learns latent representations on depth and camera pose for 3D pose reconstruction. Since weak supervision usually causes ill-conditioned learning or inferior estimation, our DWSL introduces deductive reasoning to make an inference for human pose from a view to another and develops a reconstruction loss to demonstrate what the model learns and infers is reliable. This learning by deduction strategy employs the view-transform demonstration and structural rules derived from depth, geometry and angle constraints, which improves the reliability of the model training with weak supervision. On three 3D human pose benchmarks, we conduct extensive experiments to evaluate our proposed method, which achieves superior performance in comparison with state-of-the-art weak-supervised methods. Particularly, our model shows an appealing potential for learning from 2D data captured in dynamic outdoor scenes, which demonstrates promising robustness and generalization in realistic scenarios. Our code is publicly available at https://github.com/Xipeng-Chen/DWSL-3D-pose.

Introduction

3D human pose estimation is a fundamental problem in computer vision for many applications, such as human-robot interaction, virtual reality, and action recognition, etc. However, it is greatly bottlenecked by the availability of abundant 3D annotated data, since 3D images are usually subject to specific conditions with constrained laboratory environments and thus have limited pose variations and simple backgrounds, and particularly, accurate 3D annotation demands prohibitively expensive cost. Accordingly, they cause the poor generalization of 3D pose models to the cases in the wild.

Without any 3D pose annotation, many researchers resort to Weakly-Supervised Learning (WSL) methods (Kocabas, Karagoz, and Akbas 2019; Rhodin et al. 2018; Rhodin, Salzmann, and Fua 2018; Chen et al. 2019a), which inherit the benefits of rich annotation and diversity of 2D pose datasets. They usually utilize annotated 2D pose images by lifting 2D poses to the 3D space together with geometric consistency constraints and train models without 3D pose labels for 3D human pose estimation. (Chen et al. 2019a) proposes a method to learn from single-view self-supervision, but requires a very large amount of diverse 2D human poses. (Kocabas, Karagoz, and Akbas 2019; Rhodin et al. 2018; Rhodin, Salzmann, and Fua 2018) propose a multi-view consistency from images which are taken for the same person from different viewpoints. Nevertheless, these methods have to obtain well-defined rigid transformation from annotations (Rhodin, Salzmann, and Fua 2018) or predictions from off-the-shelf methods (Kocabas, Karagoz, and Akbas 2019; Rhodin et al. 2018). Meanwhile, they employ the view synthesis strategy to produce 3D poses which supervise the training of 3D pose detectors. This casts the weakly-supervised learning problem of 3D pose estimation with only 2D annotation into a conventional fully-supervised learning task with synchronized information from multi-view images (Chen et al. 2019a). Essentially, fully supervised models are trained inductively in a data-driven manner, which greatly depends on abundant observations or samples with labels. Nevertheless, following the same spirit, with weak supervision or without annotation, the training of models suffers from a large knowledge of uncertainty or controversial ambiguity, which would cause ill-conditioned learning or inferior estimation.

To mitigate this problem, we propose Deductive Weakly-Supervised Learning (DWSL) for 3D human pose estimation. Rather than following the spirit of data-driven inductive learning in most existing methods, the proposed paradigm of learning by deduction utilizes deduction with view-transform demonstration and structural rules to infer the plausible 2D pose from another view and develop a reconstruction loss for training. This is regarded as a self-demonstration with deductive reasoning from one view to another view, namely, deduction with view-transform demonstration, and the derived reconstruction loss provides a checkpoint for the current weakly-supervised learning. At the same time, we also introduce structural rules to further promote the learning by deduction, which would ease the model training and reduce the searching space of parameters. We conduct experiments

*Pengxu Wei is the corresponding author.

on three public 3D human estimation benchmarks, where our superior performance demonstrates an appealing reliability and robustness of our method.

Overall, our main contributions are summarized as follows:

1) We propose a deductive weakly-supervised learning method for 3D human pose machines with multi-view images and only 2D pose annotations. Instead of view synthesis which involves given camera parameters and complex view alignment, it employs deductive reasoning for human pose inference and develops a mechanism of self-demonstration to guide the model learning.
2) We propose learning by deduction with view-transform demonstration and structural rules, aiming to make an inference reasonably for human pose from a view to another and improve the reliability of the model training with weak supervision.
3) Quantitative and qualitative experimental results on challenging 3D human pose datasets show a superior performance of our proposed method, demonstrating the effectiveness of our learning by deduction for weakly-supervised 3D human pose estimation, even in unconstrained scenes.

Related Work

Fully-supervised learning methods Recent advances in 3D human pose estimation attribute to the availability of large-scale datasets and sophisticated deep networks. Some methods (Sun et al. 2018, 2017; Pavlakos et al. 2017; Mehta et al. 2017a) directly predict 3D human poses from images. Despite the great advance on standard 3D pose datasets, these methods do not generalize well to outdoor scenes, since the 3D datasets are collected in constrained laboratory environments. While two-stage methods (Martinez et al. 2017; Wang et al. 2018; Chen and Ramanan 2017; Fang et al. 2018) first predict 2D human poses from images and then lift them to 3D. These methods are robust to the diverse appearance and background, due to the advance in 2D human estimation (Cao et al. 2018). However, they are still constrained to limited motion and viewpoints for directly modeling 3D mapping from a given dataset.

Weakly-supervised learning methods WSL methods provide a promising way for 3D human pose estimation, which do not heavily rely on the 3D annotations. Some methods use adversarial training for learning from single-view 2D data (Kudo et al. 2018; Drover et al. 2018; Chen et al. 2019a). These methods project the predicted 3D human poses onto the image plane from random viewpoints and a discriminator distinguishes whether these projections are realistic. Even these methods do not need extra supervision, the training process needs a very large amount of diverse 2D human poses. (Tung et al. 2017a; Kanazawa et al. 2018) utilize the 3D human model SMPL (Loper et al. 2015) to assist the learning of 3D human pose. They predict parameters of SMPL from images and the rendered 3D human pose is projected to match the 2D human poses in images. Although SMPL provides a strong structure prior of the human body, it is generated from the statistics of 3D human body, which may still be limited to some unseen human motions. Besides, some methods seek help from multi-view consistency (Kocabas, Karagoz, and Akbas 2019; Rhodin et al. 2018; Suwajanakorn et al. 2018), learning from multi-view 2D images. (Rhodin et al. 2018) trains a deep neural network to predict 3D pose from an image, with the supervision that the predicted 3D pose from different viewpoints shares the same shape. These methods still need a small part of 3D annotations for warming up, since lacking learnable view-transform. (Rhodin, Salzmann, and Fua 2018) proposes to get geometry-aware representations via novel viewpoint image generation, which still rely on annotated view-transform. (Kocabas, Karagoz, and Akbas 2019) uses epipolar geometry to obtain 3D pose annotations from multi-view 2D human pose. Although its images is unnecessarily captured from calibrated cameras, the reconstructed 3D poses may be not very stable with the limited number of viewpoints. In our work, rather than separate the learning of view-transform and 3D human pose into two steps, we propose a novel deductive learning framework to simultaneously learn the view-transform and 3D human pose in our framework.

Deductive Weakly-Supervised Learning Model

Our goal is to train a model for monocular 3D human pose estimation with only 2D human poses. To this end, We propose a novel deductive weakly-supervised learning model to automatically learn from multi-view 2D human poses, shown in Fig.1. Given a pair of images observed from two views, our proposed model formulates the depth for the input view image and the camera pose between two images as latent variables, and proposes learning by deduction with a view-transform demonstration and structural rules in a deductive module.

Formally, given a pair of images I_1 and I_2, which are taken from different viewpoints for the same human body, their 2D human poses are estimated from images by a trained 2D pose estimator (Duan et al. 2019). Considering our deductive learning strategy with view-transform demonstration from one view to another, for convenience, one of two images is regarded as the input view and the other one is as the target view. Let $[\boldsymbol{u}, \boldsymbol{v}] \in \mathbb{R}^{N\times 2}$ and $[\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}] \in \mathbb{R}^{N\times 2}$ be the pixel coordinate of the input and target respectively, where N is the number of human pose keypoints.

Latent Representations for Depth and Camera Pose

We aim to train a model for monocular 3D human pose estimation. In contrast to previous methods (Chen et al. 2019b; Rhodin, Salzmann, and Fua 2018), which first learn an implicit 3D geometry-aware representation and then need fine-tuning on the 3D annotations, our framework is capable of directly predicting the depth $\boldsymbol{z}$ for every 2D keypoints, from which the accurate 3D human pose is reconstructed. The neural network P aims to predict the depth of each keypoint and neural network Q predicts the view transformation, i.e., camera pose. We use the simple backbone in (Martinez et al. 2017) for neural networks P and Q. The architecture of back-

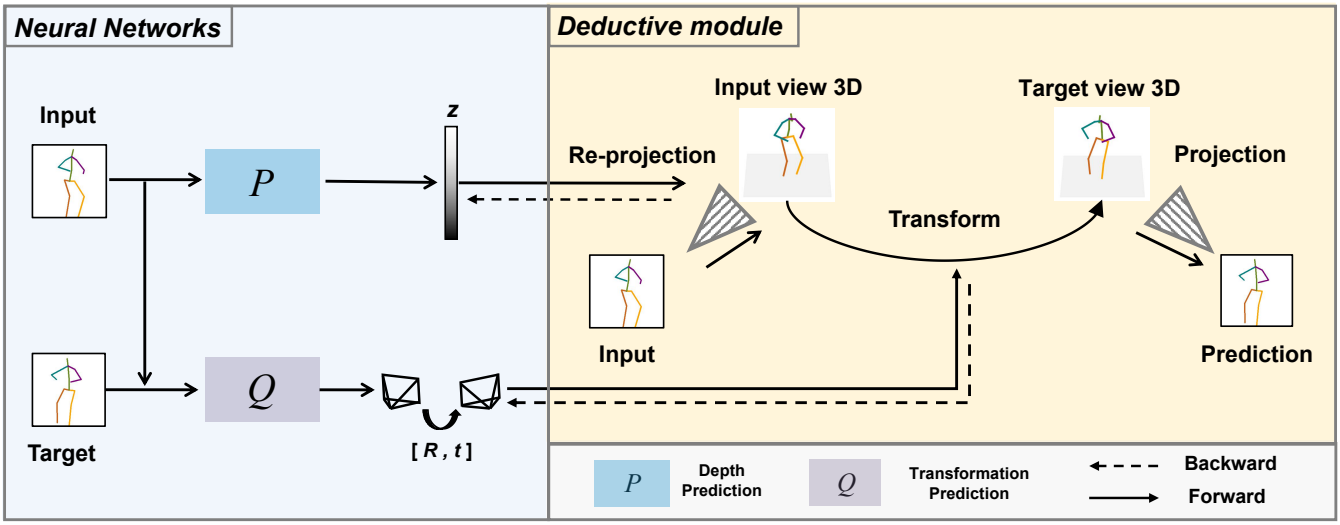

Figure 1: Illustration of our framework. Given a pair of 2D human poses predicted by a 2D pose estimator, our framework learns to predict 3D human pose via learning by deduction with view-transform demonstration and structural rules. The neural networks P and Q aim to predict the depth and camera pose; the deduction module mainly consists of three parts: re-projection, transform and projection.

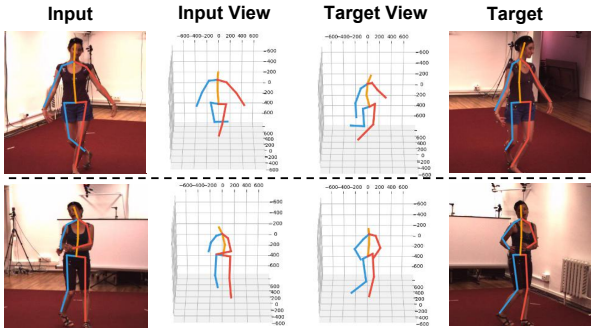

Figure 2: Visualization of the learned 3D human pose with view-transform. The input view 3D pose and the target view 3D pose are shown in the second and third column.

bone is a multilayer neural network with residual blocks. Specifically, the neural network P is parameterized by $\boldsymbol{\theta}_P$ and generate implicitly the depth $\boldsymbol{z}$ from the 2D coordinate $[\boldsymbol{u}, \boldsymbol{v}]$ of the input view.

$$\boldsymbol{z} = P([\boldsymbol{u}, \boldsymbol{v}]; \boldsymbol{\theta}_P). \quad (1)$$

We utilize the neural network Q parameterized by $\boldsymbol{\theta}_Q$ to estimate the transformation $\boldsymbol{R} \in \mathbb{R}^{3\times3}$ and $\boldsymbol{t} \in \mathbb{R}^{3\times1}$ from the input 2D pose $[\boldsymbol{u}, \boldsymbol{v}]$ and target 2D pose $[\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}]$.

$$[\boldsymbol{R}, \boldsymbol{t}] = Q([\boldsymbol{u}, \boldsymbol{v}, \hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}]; \boldsymbol{\theta}_Q). \quad (2)$$

Learning by Deduction

Our DWSL leverages the learning by deduction strategy, which is regarded as a self-demonstration with deductive reasoning from one view to another, namely, deduction with view-transform demonstration, and the derived reconstruction loss provides a checkpoint for the current weakly-supervised learning. At the same time, we also introduce structural rules to further promote the learning by deduction, which would ease the model training and reduce the searching space of parameters.

Deduction with view-transform demonstration. In the deductive module, our deduction with view-transform demonstration firstly employs re-projection to guarantees the input data is lifted from 2D to 3D based on the learned depth $\boldsymbol{z}$; then it transforms the 3D pose prediction from one view to another based on the learned camera pose between these two views; finally, the derived 3D pose for another view is projected to its 2D counterpart.

We assume that the re-projection and the projection both use an ideal pinhole camera with fixed intrinsic parameters, the camera center $[c_x, c_y] = [0, 0]$ and focal length $[f_x, f_y] = [1, 1]$. This constrains the input view 3D pose within a fixed camera coordinate system, reducing the learning difficulty of the transformation. Specifically, the neural network P is employed to generate implicitly the depth $\boldsymbol{z}$ from 2D coordinate $[\boldsymbol{u}, \boldsymbol{v}]$, from which the 3D pose of the input view $[\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}]$ is obtained.

$$[x_i, y_i, z_i] = [u_i \cdot z_i,\ v_i \cdot z_i,\ z_i] \quad (i = 1, 2, ..., N). \quad (3)$$

Different from previous work (Chen et al. 2019b; Kocabas, Karagoz, and Akbas 2019; Rhodin et al. 2018; Rhodin, Salzmann, and Fua 2018), the relative transformation is acquired from annotation or estimated by off-the-shelf algorithms, which limits the generalization of these methods to realistic dynamic scenes with moving cameras. In our framework, we propose to predict the relative transformation end-to-end together with the 3D pose. If the input view 3D pose is correct, there will exist a certain viewpoint, from which the 3D

pose is projected to match the target 2D pose. Specifically, we utilize the neural network Q to estimate the transformation $\boldsymbol{R} \in \mathbb{R}^{3\times3}$ and $\boldsymbol{t} \in \mathbb{R}^{3\times1}$. The target view 3D pose $[\boldsymbol{x}',\ \boldsymbol{y}',\ \boldsymbol{z}']$ is obtained via applying the transform.

$$[x_i',\ y_i',\ z_i']^\top \ = \ \boldsymbol{R}\cdot[x_i,\ y_i,\ z_i]^\top + \boldsymbol{t} \quad (i = 1, 2, ..., N). \tag{4}$$

With view-transform, the target view 3D pose $[\boldsymbol{x}',\ \boldsymbol{y}',\ \boldsymbol{z}']$ is first projected onto the image plane to obtain the predicted 2D pose $[\boldsymbol{u}',\ \boldsymbol{v}']$, and then the reconstruction loss is defined with the target 2D pose $[\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}]$ as supervision.

$$[u_i',\ v_i',\ z_i'] \ = \ [x_i' \ / \ z_i',\ y_i' \ / \ z_i',\ z_i'] \quad (i = 1, 2, ..., N), \tag{5}$$

$$\mathcal{L}_{\text{rec}} \ = \ \frac{1}{N}\sum_{i=1}^{N} \left\| [u_i', v_i']^\top - [\hat{u}_i, \hat{v}_i]^\top \right\|^2. \tag{6}$$

Deduction with structural rules. With the limited number of viewpoints and sparse 2D keypoints correspondence, even trivial errors of the predicted locations of 2D pose can result in a remarkable position offset or even the failure of 3D pose recovery. 3D pose estimation suffers from an intractable challenge of the view ambiguity; to address this issue, apart from deduction with view-transform demonstration, we explore common sense as structural rules to deductively constrain the model learning, which includes rules of positive depth, symmetry length and valid angle.

Positive depth rule requests each element of the learnt depth $\boldsymbol{z}$ to be non-negative. This reduces the searching parameter space of model optimization. Positive depth loss is defined as follows,

$$\mathcal{L}_{\text{pos}} \ = \ \frac{1}{N}\sum_{i=1}^{N}(|z_i| - z_i). \tag{7}$$

Symmetry length rule requests the length of the same body part in the left half and the right half to be equal, since in 3D space, each limb has a nearly equal length no matter from which viewpoint human poses are observed. Thus, we define the length of the left body parts is $\{l_{l_i}\}_{i=1}^{M}$, and its counterpart in the right body is $\{l_{r_i}\}_{i=1}^{M}$, where M is the number of half body parts. This loss function can be defined as follows,

$$\mathcal{L}_{\text{sym}} \ = \ \frac{1}{M}\sum_{i=1}^{M} \left\| l_{l_i} - l_{r_i} \right\|^2. \tag{8}$$

Valid angle rule helps to remove invalid 3D human poses, which guides the model to distinguish the left from the right. This rule penalizes the 3D poses that violate the knee or elbow joint-angle limits. The examples are shown in Figure 3. This constrains 3D poses generated in a feasible manner. Similar to (Dabral et al. 2018), the valid angle rule is formulated as follows,

$$\mathcal{L}_{\text{ang}} \ = \ \mathcal{L}_{\text{left-arm}} + \mathcal{L}_{\text{right-arm}} + \mathcal{L}_{\text{left-leg}} + \mathcal{L}_{\text{right-leg}}. \tag{9}$$

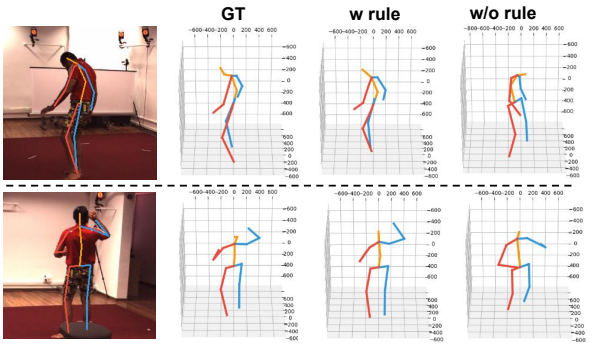

Figure 3: Illustration of the effect of the valid angle rule. The predicted 3D pose is from our model trained with/without the valid angle rule. As we can see, removing this rule will result in the reversed left and right, while the correct human shape is kept. The left part of the human pose is denoted in red color and the right part is in blue color.

Training and Inference

In the training stage, we train our model in an end-to-end manner by minimizing the following total loss function, where $\{\alpha_{\text{rec}}, \alpha_{\text{pos}}, \alpha_{\text{sym}}, \alpha_{\text{ang}}\}$ are the weights for losses.

$$\mathcal{L} \ = \ \alpha_{\text{rec}}\cdot\mathcal{L}_{\text{rec}} + \alpha_{\text{pos}}\cdot\mathcal{L}_{\text{pos}} + \alpha_{\text{sym}}\cdot\mathcal{L}_{\text{sym}} + \alpha_{\text{ang}}\cdot\mathcal{L}_{\text{ang}}. \tag{10}$$

In the inference stage, 2D human pose is first estimated from single image I by a 2D pose estimator, and then 2D coordinate $[\boldsymbol{u}, \boldsymbol{v}]$ is fed into P to generate the depth $\boldsymbol{z}$. The 3D human pose is efficiently recovered based on the depth $\boldsymbol{z}$ using Equation 3, without the aforementioned deduction.

Experiments

Experimental Settings

Datasets. Our experiments have been conducted on three 3D human pose datasets, i.e., Human3.6M (Ionescu et al. 2013), MPI-INF-3DHP dataset (Mehta et al. 2017a), and Ski-PosePTZ (Rhodin et al. 2018). **Human3.6M** (Ionescu et al. 2013) consists of 3.6 million images and has 11 subjects with 15 daily actions from 4 different viewpoints. Following the standard protocol in (Martinez et al. 2017), subject 1, 5, 6, 7, 8 are for training and subject 9, 11 are for evaluation. **MPI-INF-3DHP** (Mehta et al. 2017a) contains about 1.3 million frames taken from different viewpoints. With the dataset split similar to (Yang et al. 2018), its training set covers five chest-height cameras and 17 joints (compatible with H36M) are for training; its test set has 2929 frames taken from both indoor and outdoor scenes. **Ski-PosePTZ** (Rhodin et al. 2018) is more challenging since its frames are captured in dynamic outdoor scenes with 6 pan-tilt-zoom cameras; namely, it is a ski dataset with competitive racers going down alpine slalom courses. It contains images of 6 subjects. The subject 1 to 5 are for training (8481 frames) and the subject 6 is for testing (1716 frames).

Metrics. We report our results on Human3.6M in terms of MPJPE (Mean Per joint Position Error) and PMPJPE (Procrustes aligned Mean Per Joint Position Error), similar to

Supervision	Method	PMPJPE	MPJPE
Full	Our baseline (Martinez et al. 2017)	52.1	62.9
Weak	Tung et al. (2017a)	98.4	-
	Tung et al. (2017b)	97.2	-
	Kocabas, Karagoz, and Akbas (2019)	70.7	N/A
	Wandt and Rosenhahn (2019)	65.1	-
	Drover et al. (2018)	64.6	-
	Wang, Kong, and Lucey (2019)	**57.5**	83.0
	Ours (SH detections)	60.6	80.2
	Ours	58.6	**76.7**

Table 1: Comparison results on Human3.6M. 'Full' refers to fully-supervised methods; 'Weak' refers to weakly-supervised methods; N/A means the result is not available and '-' is not provided by authors. 'SH detections' indicates that we trained and tested our model with Stacked Hourglass (Newell, Yang, and Deng 2016) 2D detections.

Supervision	Method	Training Data	PCK	AUC
Full	Kanazawa et al. (2018)	H36M+MPI	86.3	47.8
	Mehta et al. (2017b)	H36M+MPI	83.9	47.3
Weak	Kanazawa et al. (2018)	H36M+MPI	77.1	40.7
	Ours	H36M	86.3	49.8
	Ours	H36M + MPI	**88.1**	**50.5**

Table 2: Results on MPI-INF-3DHP. H36M refers to the Human3.6M dataset and MPI refers to the MPI-INF-3DHP dataset. All the methods are evaluated on the test set of MPI-INF-3DHP dataset.

(Wang, Kong, and Lucey 2019). Besides, in order to evaluate on MPI-INF-3DHP, following (Chen et al. 2019a), we also report PCK (Percentage of Correct Keypoints) and AUC (Area Under the Curve) calculated based on PMPJPE.

Implementation details. 2D poses used in our model are normalized to unit size together with their pelvis points at origin. For the input 2D poses, they are augmented by a random rotation within $\pm 30^{\circ}$, and are re-scaled by a factor within 1 ± 0.1. For the target 2D poses, we introduce the flip operation to generate data as if it was taken from a virtual camera, via multiplying the x coordinate by -1. In every training epoch, we randomly choose a pair of viewpoints for every frame when there exist multiply ones. The weights in the losses, $(\alpha_{\text{rec}},\ \alpha_{\text{pos}},\ \alpha_{\text{sym}},\ \alpha_{\text{ang}})$, are set to $(1,\ 1,\ 10,\ 10^{-3})$. We train the network on an RTX 2080 GPU with a batchsize of 64 for 100 epochs.

Comparison Results with State-of-the-art Methods

We compare our proposed method with existing state-of-the-art weakly-supervised learning methods on Human3.6M dataset, shown in Table 1. It is observed that for PMPJPE, our method achieves a performance of 58.5mm and outperforms most of the state-of-the-art methods. It also obtains comparable performance with (Wang, Kong, and Lucey 2019), in comparison of which our method just utilize a more light-weighted backbone. When compared with the method (Kocabas, Karagoz, and Akbas 2019) that generates 3D supervisions via conventional 3D reconstruction algorithm, our method outperforms it by a large margin (58.6mm vs. 70.7mm in PMPJPE). Besides, our model also achieves the best performance 76.7mm on the MPJPE, which has a significant improvement by 8.3mm in comparison with (Wang, Kong, and Lucey 2019), demonstrating that our method can capture the 3D orientations of the human body more accurately. We also compare our weakly-supervised framework with our fully-supervised baseline (Martinez et al. 2017), which is trained with 3D annotations. As shown in Table 1, our method has a comparable performance with the fully-supervised baseline by only 6.5mm gaps under PMPJPE on Human3.6M dataset. This indicates that our method can effectively learn 3D information from multi-view images.

Method	Training Data	PMPJPE	MPJPE
Martinez et al. (2017)	H36M-3D	111.3	141.3
Zhao et al. (2019)	H36M-3D	108.8	125.1
Ours	H36M-MV	108.7	130.2
Ours	H36M-MV+Ski-MV	**74.7**	**99.4**

Table 3: Results on Ski-PosePTZ. -3D refers to using the 3D ground truth in the training set of the dataset. -MV refers to using multi-view images in the training set of the dataset. All the methods are evaluated on the test set of Ski-PosePTZ dataset.

Method	PMPJPE	Δ
Ours	58.6	-
w/o reconstruction	N/A	N/A
w/o positive depth rule	73.2	14.6
w/o symmetry length rule	N/A	N/A
w/o valid angle rule	104.0	45.4
w/o data augmentation	62.1	3.5

Table 4: Ablation study on Human3.6M dataset. Δ indicates the performance decrease in comparison with our DWSL. N/A means the result is not available.

To further validate the generalization ability, we further conduct experiments on MPI-INF-3DHP and Ski-PosePTZ, which contain challenging outdoor scenes. Following the similar experimental setting (Chen et al. 2019b), we first train our model on Human3.6M, and then test on MPI-INF-3DHP. Our method generalizes well and achieves a performance of 86.3 in PCK and 49.8 in AUC. When it is further trained on the multi-view images of MPI-INF-3DHP, we have a better performance of 88.1 in PCK and 50.5 in AUC. As shown in Table 3, when it is trained only on the multi-view images of Human3.6M and Ski-PosePTZ, our model outperforms the fully-supervised method (Zhao et al. 2019) by reducing the error by 31.4% (74.7mm vs. 108.8mm) in PMPJPE.

Ablation Studies

We conduct ablation experiments to verify the effectiveness of each component in our framework on Human3.6M, shown in Table 4. The absence of reconstruction loss leads to model

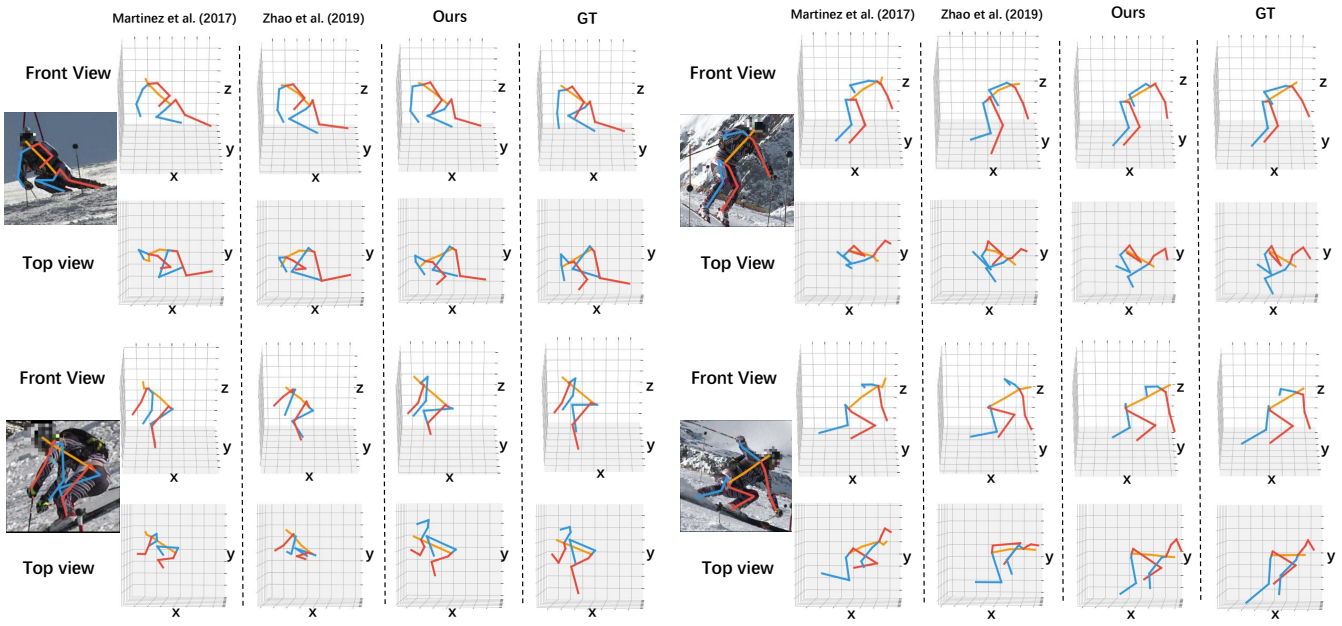

Figure 4: 3D human pose predictions observed in the front view and top view on Ski-PosePTZ dataset.

collapse, demonstrating that multi-view images provide the information for the learning of depth. Without positive depth rule, the error increases by 14.6mm. When the predicted depth is negative, the re-projection is applied along the negative z axis, which increases the difficulty of the model learning. Symmetry length rule ensures reasonable 3D human poses to be predicted from noisy multi-view data. In Figure 5, without symmetry length rule, the model fails to predict reasonable 3D human pose, even if 2D human pose can be predicted. Due to noisy 2D human pose predictions, the limited number of viewpoints and unknown camera parameters, it is hard to reconstruct the correct 3D human pose under such conditions. The valid angle rule teaches the method to distinguish the left from right, as shown in Figure 3. Removing this rule will result in the reversed left and right, while the correct human shape is kept. When this component is removed, the error increases by 45.4mm. When the data augmentation is not applied, the small increase in the error is 3.5mm.

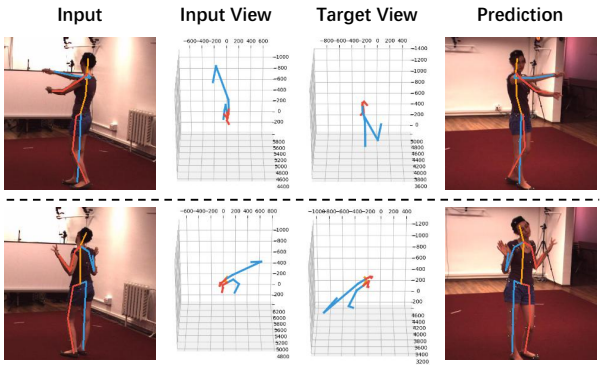

Figure 5: Illustration of the effect of the symmetry length rule. 'Input view' and 'Target view' are reconstructed 3D human poses in input and target viewpoints, respectively. Without the symmetry length rule, the model fails to predict reasonable 3D pose, even if 2D human poses can be reconstructed.

Model Analysis

Learning from multi-view images of unconstrained scenes. To validate the ability of our model to learn from multi-view images of unconstrained scenes, we conduct experiments on Ski-PosePTZ. Specifically, this dataset is challenging due to diverse distributed camera poses and complex human poses. The result is shown in Table 3. It is shown that the model trained only on Human3.6M is difficult to be generalized to the Ski-PosePTZ dataset. The fully-supervised method (Zhao et al. 2019) get the performance of 108.8mm in PMPJPE, and our model trained only on the multi-view images of Human3.6M also has difficulty to be adopted on Ski-PosePTZ. However, only using extra multi-view images from the training set of Ski-PosePTZ, without any 3D supervisions (including camera parameters), we achieve the performance of 74.7mm in PMPJPE, which reduces the error by 31.2 compared to our baseline (74.7mm vs. 108.7mm). The qualitative results are shown in Figure 4 and we present the predicted 3D human pose in front viewpoints and top viewpoints. Even if all the methods show correct projections from the front viewpoints, the difference of depth can be viewed from the top viewpoints.

Conclusion

In this paper, we propose a deductive weakly-supervised learning method for 3D human pose machines with multi-view images from uncalibrated cameras and only 2D pose annotations. To mitigate the issue of ill-conditioned learning

and inferior estimation due to weak supervision, our method employs deductive reasoning for human pose inference and develops a mechanism of self-demonstration to guide the model learning. Our learning by deduction is performed with view-transform demonstration and structural rules to make an inference reasonably for human pose from a view to another. This ensures the reliability of the model training with weak supervision. Extensive experiments on 3D human pose datasets show that our method has achieved a remarkable performance improvement. Especially, our method demonstrates appealing effectiveness and generalization for more challenging scenes in the wild. Our work provides a fresh insight with learning by deduction for weakly-supervised 3D human pose estimation.

Acknowledgments

This work was supported in part by NSFC (No.62006253, U1811463, 61836012, 61976233), China Postdoctoral Science Foundation (No.2020M672968), State Key Development Program (No.2018YFC0830103), Fundamental Research Funds for the Central Universities (No.19lgpy228), and Major Project of Guangzhou Science and Technology of Collaborative Innovation and Industry under Grant 201605122151511.

References

Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2018. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *arXiv preprint arXiv:1812.08008* .

Chen, C.-H.; and Ramanan, D. 2017. 3d human pose estimation= 2d pose estimation+ matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7035–7043.

Chen, C.-H.; Tyagi, A.; Agrawal, A.; Drover, D.; Stojanov, S.; and Rehg, J. M. 2019a. Unsupervised 3d pose estimation with geometric self-supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5714–5724.

Chen, X.; Lin, K.-Y.; Liu, W.; Qian, C.; and Lin, L. 2019b. Weakly-supervised discovery of geometry-aware representation for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10895–10904.

Dabral, R.; Mundhada, A.; Kusupati, U.; Afaque, S.; Sharma, A.; and Jain, A. 2018. Learning 3d human pose from structure and motion. In *Proceedings of the European Conference on Computer Vision*, 668–683.

Drover, D.; Chen, C.-H.; Agrawal, A.; Tyagi, A.; and Phuoc Huynh, C. 2018. Can 3d pose be learned from 2d projections alone? In *Proceedings of the European Conference on Computer Vision*.

Duan, H.; Lin, K.-Y.; Jin, S.; Liu, W.; Qian, C.; and Ouyang, W. 2019. TRB: A Novel Triplet Representation for Understanding 2D Human Body. In *Proceedings of the IEEE International Conference on Computer Vision*, 9479–9488.

Fang, H.-S.; Xu, Y.; Wang, W.; Liu, X.; and Zhu, S.-C. 2018. Learning pose grammar to encode human body configuration for 3d pose estimation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2013. 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.

Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7122–7131.

Kocabas, M.; Karagoz, S.; and Akbas, E. 2019. Self-supervised learning of 3d human pose using multi-view geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1077–1086.

Kudo, Y.; Ogaki, K.; Matsui, Y.; and Odagiri, Y. 2018. Unsupervised adversarial learning of 3d human pose from 2d joint locations. *arXiv preprint arXiv:1803.08244* .

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A skinned multi-person linear model. *ACM transactions on graphics* 34(6): 1–16.

Martinez, J.; Hossain, R.; Romero, J.; and Little, J. J. 2017. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2640–2649.

Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; and Theobalt, C. 2017a. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *International Conference on 3D Vision*, 506–516. IEEE.

Mehta, D.; Sridhar, S.; Sotnychenko, O.; Rhodin, H.; Shafiei, M.; Seidel, H.-P.; Xu, W.; Casas, D.; and Theobalt, C. 2017b. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics* 36(4): 1–14.

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, 483–499. Springer.

Pavlakos, G.; Zhou, X.; Derpanis, K. G.; and Daniilidis, K. 2017. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7025–7034.

Rhodin, H.; Salzmann, M.; and Fua, P. 2018. Unsupervised geometry-aware representation for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision*, 750–767.

Rhodin, H.; Spörri, J.; Katircioglu, I.; Constantin, V.; Meyer, F.; Müller, E.; Salzmann, M.; and Fua, P. 2018. Learning monocular 3d human pose estimation from multi-view images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8437–8446.

Sun, X.; Shang, J.; Liang, S.; and Wei, Y. 2017. Compositional human pose regression. In *Proceedings of the IEEE International Conference on Computer Vision*, 2602–2611.

Sun, X.; Xiao, B.; Wei, F.; Liang, S.; and Wei, Y. 2018. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 529–545.

Suwajanakorn, S.; Snavely, N.; Tompson, J. J.; and Norouzi, M. 2018. Discovery of latent 3d keypoints via end-to-end geometric reasoning. In *Advances in Neural Information Processing Systems*, 2059–2070.

Tung, H.-Y.; Tung, H.-W.; Yumer, E.; and Fragkiadaki, K. 2017a. Self-supervised learning of motion capture. In *Advances in Neural Information Processing Systems*, 5236–5246.

Tung, H.-Y. F.; Harley, A. W.; Seto, W.; and Fragkiadaki, K. 2017b. Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision. In *IEEE International Conference on Computer Vision*, 4364–4372. IEEE.

Wandt, B.; and Rosenhahn, B. 2019. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7782–7791.

Wang, C.; Kong, C.; and Lucey, S. 2019. Distill Knowledge from NRSfM for Weakly Supervised 3D Pose Learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 743–752.

Wang, M.; Chen, X.; Liu, W.; Qian, C.; Lin, L.; and Ma, L. 2018. Drpose3d: Depth ranking in 3d human pose estimation. *arXiv preprint arXiv:1805.08973* .

Yang, W.; Ouyang, W.; Wang, X.; Ren, J.; Li, H.; and Wang, X. 2018. 3d human pose estimation in the wild by adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5255–5264.

Zhao, L.; Peng, X.; Tian, Y.; Kapadia, M.; and Metaxas, D. N. 2019. Semantic graph convolutional networks for 3D human pose regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3425–3435.