Geometry-driven Self-supervised Method for 3D Human Pose Estimation

Yang Li[12] **Kan Li**[1*], **Shuai Jiang**[12], **Ziyue Zhang**[2]
Congzhentao Huang[2], **Richard Yi Da Xu**[2]
[1]School of Computer Science and Technology, Beijing Institute of Technology, China
[2]Faculty of Engineering and Information Technology, University of Technology Sydney, Australia
{yanglee, likan}@bit.edu.cn, {shuai.jiang-1, ziyue.zhang-2, congzhentao.huang}@student.uts.edu.au, yida.xu@uts.edu.au

Abstract

The neural network based approach for 3D human pose estimation from monocular images has attracted growing interest. However, annotating 3D poses is a labor-intensive and expensive process. In this paper, we propose a novel self-supervised approach to avoid the need of manual annotations. Different from existing weakly/self-supervised methods that require extra unpaired 3D ground-truth data to alleviate the depth ambiguity problem, our method trains the network only relying on geometric knowledge without any additional 3D pose annotations. The proposed method follows the two-stage pipeline: 2D pose estimation and 2D-to-3D pose lifting. We design the transform re-projection loss that is an effective way to explore multi-view consistency for training the 2D-to-3D lifting network. Besides, we adopt the confidences of 2D joints to integrate losses from different views to alleviate the influence of noises caused by the self-occlusion problem. Finally, we design a two-branch training architecture, which helps to preserve the scale information of re-projected 2D poses during training, resulting in accurate 3D pose predictions. We demonstrate the effectiveness of our method on two popular 3D human pose datasets, Human3.6M and MPI-INF-3DHP. The results show that our method significantly outperforms recent weakly/self-supervised approaches.

Introduction

3D human pose estimation has attracted substantial interest for its vast potential on various applications including human-computer interaction, virtual reality and action recognition. With the great success of deep learning, many researchers (Martinez et al. 2017; Pavllo et al. 2019) applied the neural network to predict 3D human poses from monocular images. Estimating 3D poses using neural networks mainly faces two main challenges. Firstly, a typical neural network model needs a large amount of training data. 3D pose annotations are collected through the marker-based Motion Capture (MoCap) system, which is an expensive process. Secondly, there are well-founded geometrical theories on how to project 2D images to 3D skeletons. Simply using a neural network to approximate this projection may lead to the network subject to overfitting training data.

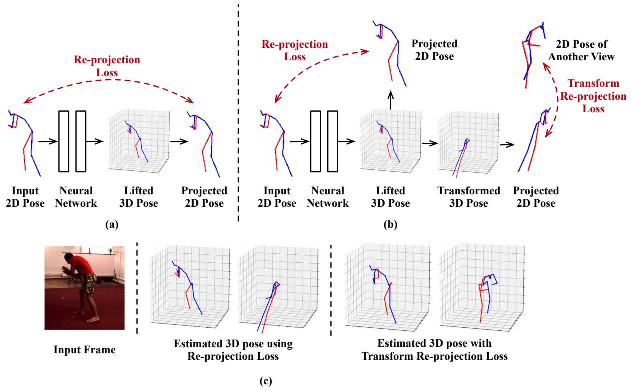

Figure 1: (a) The network trained with re-projection loss. (b) The proposed network trained with transform re-projection loss. (c) The comparisons of results estimated by the above two networks. The results are shown in two different views.

To alleviate the above challenges, weakly/self-supervised learning paradigms have been increasingly explored in recent works (Rhodin et al. 2018; Wu et al. 2016; Chen et al. 2019; Rhodin, Salzmann, and Fua 2018). Re-projection loss (Tung et al. 2017), which does not require explicit 3D ground-truth, has become a commonly used technique. It re-projects estimated 3D poses back to the 2D space and calculates the loss between the input and re-projected 2D poses as supervision. However, due to the depth ambiguity problem where multiple 3D body configurations can explain the same 2D projection, the re-projection loss cannot yield accurate and realistic 3D poses. For example, as shown in Figure 1, since re-projection loss only constrains the estimated 3D pose at a specific camera angle, it may result in an invalid human pose when observed from another angle. Although some techniques such as adversarial loss (Tung et al. 2017; Wandt and Rosenhahn 2019) and kinematic constraints (Habibie et al. 2019; Pavllo et al. 2019), have been proposed to constrain the estimated 3D poses into a semantic sub-space, they usually require some extra unpaired 3D pose

[*]Corresponding Author

annotations (without 2D-3D correspondence) to make the network memorize the distribution of real 3D skeletons.

3D pose datasets (Ionescu et al. 2013; Mehta et al. 2017) are usually collected under the configuration with multiple calibrated cameras. The consistency information between multiple camera views has not been fully explored in recent weakly/self-supervised methods. Although the recently proposed work (Kocabas, Karagoz, and Akbas 2019) has explored multi-view geometry to train a network, they utilized triangulation on detected multi-view 2D poses to generate 'ground-truth' 3D poses, which are subsequently used to train a 3D pose network. However, this naive application of 3D multi-view geometry is sub-optimal due to the noises introduced in the 2D pose detection at each individual camera. The detected 2D poses are combined to produce its 3D pose, which may further produce noisy supervision signals. Besides, the process of generating pseudo ground-truth is redundant.

In this paper, we propose a novel self-supervised approach to take advantage of the geometric prior for training a 3D pose estimation model. We formulate 3D pose estimation as 2D keypoint estimation followed by 2D-to-3D pose lifting. The first stage is compatible with any state-of-the-art 2D keypoint detector, and our work concentrates on training the 2D-to-3D lifting network without using any additional 3D ground-truth data. Specifically, in order to overcome the depth ambiguity problem, we design the transform re-projection loss. As shown in Figure 1(b), it transforms the lifted 3D poses from current view to another randomly selected view through rigid transformation, and then calculates the re-projection loss between the transformed 3D pose and the 2D pose of the target view. As a result, it can effectively constrain the estimated 3D poses by considering multi-view consistency. Due to the self-occlusion problem, some 2D joints may be invisible at the frame of a particular camera angle, which may lead to inaccurate 2D keypoint detections. However, they may be visible from other camera angles. Thus, the same human joint will obtain different 2D detection confidences on different camera views. We acquire the confidence weights from estimated 2D keypoint heatmaps and use them to integrate losses of different camera views, which makes our method more robust to noisy 2D detections. Finally, we introduce a root position regression branch to restore the global 3D poses during training. In this way, we can reserve the scale information of re-projected 2D poses, which can improve the accuracy of the predicted 3D poses. Moreover, in order to train the root position branch and lifting branch simultaneously from scratch, we propose a pre-training technique to help the network converge.

We perform extensive experiments on two popular 3D human pose datasets: Human3.6M (Ionescu et al. 2013) and MPI-INF-3DHP (Mehta et al. 2017). The results demonstrate that our method achieves state-of-the-art performance. The contributions of our work are summarized as follows:

- We propose a self-supervised approach to train the 2D-to-3D lifting network without any 3D pose annotations. It only relies on geometry knowledge to construct supervision signals, which leads to a better generalization ability.
- We design the transform re-projection loss, which is an effective technique to exploit multi-view consistency information and constrain the estimated 3D poses during training. Moreover, we integrate it with the 2D joint confidences of different camera views to alleviate the self-occlusion problem.
- The proposed method achieves state-of-the-art results on two popular 3D pose benchmarks compared with recent weakly/self-supervised methods.

Related Work

3D Human Pose Estimation

3D human pose estimation is a long-standing problem and has been considerably studied in the past few years. Recently, following the great success of deep learning, modern 3D human pose estimation techniques are usually formulated as learning-based frameworks. These works can be generally classified into two categories. The first class of methods (Tekin et al. 2017; Pavlakos et al. 2017a; Mehta et al. 2017; Habibie et al. 2019; Sun et al. 2018) directly predict the depth from monocular images through the deep convolutional neural networks (DCNNs). The second category (Iqbal et al. 2018; Fang et al. 2018; Tome, Russell, and Agapito 2017; Martinez et al. 2017; Fang et al. 2018; Chen and Ramanan 2017) is the two-stage pipeline, which first obtains 2D joint locations through the advanced 2D keypoint detector such as Stacked Hourglass (SH) network (Newell, Yang, and Deng 2016) and Cascaded Pyramid Network (CPN) (Chen et al. 2018), and then lifts them into 3D space. In order to learn the mapping between 2D and 3D joint positions, various 2D-to-3D lifting network backbones were designed. For example, (Martinez et al. 2017) proposed a simple baseline using a simple neural network with only two fully connected layers, while achieved surprising results. Since human skeletons are with the graph-like structure, several works (Zhao et al. 2019) also attempted to exploit the novel Graph Convolution Networks (GCNs) to capture the semantic relationships between human joints for accurate 3D human pose regression. Besides, there are also works (Pavllo et al. 2019; Arnab, Doersch, and Zisserman 2019; Zhou et al. 2016; Zhou and De la Torre 2016) that considers temporal information from frame sequence to produce more robust predictions. In our work, we follow the two-stage pipeline. Moreover the proposed approach is compatible with any recent 2D-to-3D lifting network backbone.

Weakly/Self-supervised Approaches

Recently, weakly/self-supervised approaches have received much attention due to the difficulty of gathering 3D pose annotations. In order to train the network without explicit 3D pose annotations, the prior of camera projection geometry was commonly explored, and some geometry-driven methods were proposed. Among them, re-projection loss is one of the most widely used technique (Kanazawa et al. 2018; Wu et al. 2016; Wandt and Rosenhahn 2019; Pavllo et al. 2019; Wang et al. 2019; Brau and Jiang 2016). However, using re-projection loss alone cannot accurately con-

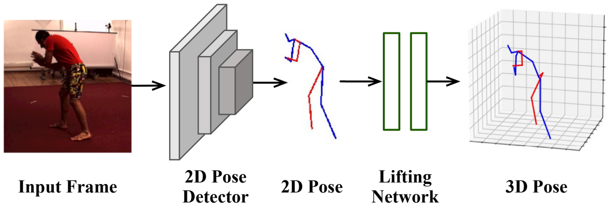

Figure 2: The overall architecture of our method that follows a two-stage pipeline.

strain the depth of skeletons due to the depth ambiguity problem. (Habibie et al. 2019; Pavllo et al. 2019; Rhodin et al. 2018) alleviated this problem by enforcing the bone length similarity between predicted and ground-truth skeletons. Adversarial loss (Yang et al. 2018; Tung et al. 2017; Wandt and Rosenhahn 2019; Kanazawa et al. 2018) is another popular technique to regularize the predicted 3D poses. It encourages the output 3D poses on the real human manifold by introducing a real/fake 3D skeleton discriminator. For example, (Tung et al. 2017) proposed the Adversarial Inverse Graphical Network (AIGN), which uses the adversarial prior to match the distribution between the predictions and a collection ground-truth for the task such as 2D-to-3D lifting and image-to-image translation. (Wandt and Rosenhahn 2019) proposed a weakly-supervised method with the adversarial supervision for 3D human pose estimation. (Chen et al. 2019) exploits the geometric self-consistency of the lift-reproject-lift process with the adversarial prior of 2D poses. As we analyzed above, the bone length constraint and adversarial loss still require unpaired 3D pose annotations for counting the bone length or training the real/fake 3D skeleton discriminator. Differently, we proposed a self-supervised approach that solely relies on the camera geometry prior, which can result in better generalization ability.

Multi-view Approaches

Early methods (Belagiannis et al. 2014) reconstructed 3D poses from multi-view inputs through the triangulation or the 3D pictorial structures model. Recent works (Pavlakos et al. 2017b; Tome et al. 2018; Dong et al. 2019) combined the traditional techniques with the novel CNNs to improve the robustness of the framework. Different from these methods requiring multi-view inputs at both training and testing stages, our method only requires multi-view inputs during training and requires monocular images during testing.

There are two latest methods that also exploit multi-view information for 3D pose estimation in a self-supervised way. (Rhodin, Salzmann, and Fua 2018) pre-trained an encoder-decoder network that predicts an image from one view to another to learn a geometry-aware body representation, and then use a small amount of supervision to learn a mapping from 2D to 3D poses. (Kocabas, Karagoz, and Akbas 2019) applied triangulation on the detected 2D joint positions to generate 'ground-truth' 3D poses for training the 3D pose estimation network. In comparisons, our method directly adopts the multi-view consistency information for training rather than generating the pseudo 3D ground-truth, which is a simpler way and more robust to the noisy 2D detections. Moreover, our method does not require any additional 3D pose annotations.

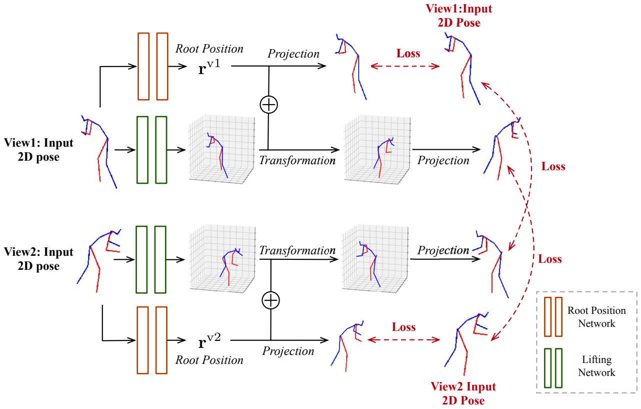

Figure 3: The architecture of the proposed self-supervised training approach.

Method

Overall, the proposed method follows a two-stage pipeline, as shown in Figure 2. First, we use the state-of-the-art 2D pose estimation network to predict 2D poses from input frames. Here, we denote $\mathbf{X} \in \mathbb{R}^{N\times 2}$ as N detected 2D joint locations. Meanwhile, we obtain their corresponding confidence scores $\mathbf{w} \in \mathbb{R}^{N}$ from estimated keypoint heatmaps through the max operation. With the detected 2D poses, we learn a neural network $\mathcal{N}$ to project them into 3D space. Similarly, we define $\mathbf{Y} \in \mathbb{R}^{N\times 3}$ as the output 3D joint locations. Following the protocol with previous works, we estimate zero-centered 3d poses where the values of $\mathbf{Y}$ are the 3D positions relative to the fixed root joint (pelvis).

The architecture of the lifting network $\mathcal{N}$ is designed inspired by (Martinez et al. 2017). The input layer takes the concatenated coordinates of N human joints and applies a fully connected layer with 1024 output channels. Then it is followed by four blocks that are surrounded by residual connections. For each block, several fully connected layers (1024 channels) followed by Batch Normalization, rectified linear units, and dropout, are stacked for efficiently mapping the 2D pose features to high-level features. Finally, the features extracted by the last residual block are fed into an extra linear layer ($N \times 3$ channels) to output 3D poses.

Self-supervised Approach

In this section, we introduce the proposed self-supervised approach for training the lifting network. The training process takes as input the detected 2D poses of a pair of frames that are captured from two different views at the same time. With the paired frames, we first detect their 2D poses $\mathbf{X}^{v1}$ and $\mathbf{X}^{v2}$ and their corresponding confidence weights of each joint $\mathbf{w}^{v1}$ and $\mathbf{w}^{v2}$. Then, we feed the 2D poses into the lifting network and obtain their estimated 3D poses $\mathbf{Y}^{v1}$ and $\mathbf{Y}^{v2}$.

Two-branch Training Architecture For training the lifting network without 3D ground-truth annotations, we design the transform re-projection loss. It involves the perspective projection and view transformation operations, which require global 3D joint positions. Without global 3D joint positions, we can not obtain the absolute depth of the person in the camera coordinate, which results in unknown scale when re-projecting 3D poses back to 2D space. Existing methods commonly normalize the scale of 2D skeletons to overcome the scale ambiguity problem. However, it must be used in conjunction with the kinematic constraint or adversarial loss to output realistic 3D poses.

In our work, we design another branch, named root position branch, to help train the lifting network. It predicts root joint positions, $\mathbf{r}^{\mathrm{v1}}$ and $\mathbf{r}^{\mathrm{v2}}$, which are added to relative 3D poses predicted by the lifting network to restore global 3D poses, $\tilde{\mathbf{Y}}^{\mathrm{v1}}$ and $\tilde{\mathbf{Y}}^{\mathrm{v2}}$. The root position network has the same architecture with the lifting network, and they do not share any weights. The two branches can be optimized simultaneously using multi-view consistency information, and the loss function and detailed training procedure will be discussed in the following sections.

Loss Function With the global 3D poses, we first re-project them back to the 2D space following the perspective projection ρ.

$$\begin{aligned} \rho(\tilde{\mathbf{Y}}_i^{\mathrm{v1}}) &= \begin{bmatrix} f_x^{\mathrm{v1}}\tilde{\mathbf{Y}}_i^{\mathrm{v1}}(x)/\tilde{\mathbf{Y}}_i^{\mathrm{v1}}(z) + c_x^{\mathrm{v1}} \\ f_y^{\mathrm{v1}}\tilde{\mathbf{Y}}_i^{\mathrm{v1}}(y)/\tilde{\mathbf{Y}}_i^{\mathrm{v1}}(z) + c_y^{\mathrm{v1}} \end{bmatrix}, \\ \rho(\tilde{\mathbf{Y}}_i^{\mathrm{v2}}) &= \begin{bmatrix} f_x^{\mathrm{v2}}\tilde{\mathbf{Y}}_i^{\mathrm{v2}}(x)/\tilde{\mathbf{Y}}_i^{\mathrm{v2}}(z) + c_x^{\mathrm{v2}} \\ f_y^{\mathrm{v2}}\tilde{\mathbf{Y}}_i^{\mathrm{v2}}(y)/\tilde{\mathbf{Y}}_i^{\mathrm{v2}}(z) + c_y^{\mathrm{v2}} \end{bmatrix}, \end{aligned} \tag{1}$$

where f_x and f_y refer to the focal lengths, c_x and c_y define the principal points, $\tilde{\mathbf{Y}}_i^{\mathrm{v1}}(x)$ indicates the value of x coordinate of i^{th} joint position of $\tilde{\mathbf{Y}}^{\mathrm{v1}}$. And then, we calculate the l_2 loss between the input and re-projected 2D poses as supervisions,

$$\begin{aligned} \mathcal{L}_{\text{reproj}} = \sum_i^N & \mathbf{w}_i^{\mathrm{v1}} \|\mathbf{X}_i^{\mathrm{v1}} - \rho(\tilde{\mathbf{Y}}_i^{\mathrm{v1}})\|^2 \\ & + \mathbf{w}_i^{\mathrm{v2}} \|\mathbf{X}_i^{\mathrm{v2}} - \rho(\tilde{\mathbf{Y}}_i^{\mathrm{v2}})\|^2, \end{aligned} \tag{2}$$

where $\mathbf{w}_i^{\mathrm{v1}}$ and $\mathbf{w}_i^{\mathrm{v2}}$ are the confidence scores of the i^{th} joints of two views. Here, we use the confidence scores of detected 2D poses to integrate the re-projection loss of different views. The view with smaller 2D confidence value makes less contribution to the loss value, which reduces the impact of the noisy 2D detections for the lifting network training.

However, simply using the re-projection consistency will encounter the depth ambiguity problem. To overcome the problem, we design the transform re-projection loss, which constrains the predicted 3D skeletons from multiple perspectives. Specifically, we transform the estimated 3D pose from one view to another through the rigid transformation τ as follows:

$$\begin{aligned} \tau(\tilde{\mathbf{Y}}_i^{\mathrm{v1}}) &= \mathbf{R}_{\mathrm{1to2}} \left(\tilde{\mathbf{Y}}_i^{\mathrm{v1}} - \mathbf{t}_{\mathrm{1to2}}\right), \\ \tau(\tilde{\mathbf{Y}}_i^{\mathrm{v2}}) &= \mathbf{R}_{\mathrm{2to1}} \left(\tilde{\mathbf{Y}}_i^{\mathrm{v2}} - \mathbf{t}_{\mathrm{2to1}}\right), \end{aligned} \tag{3}$$

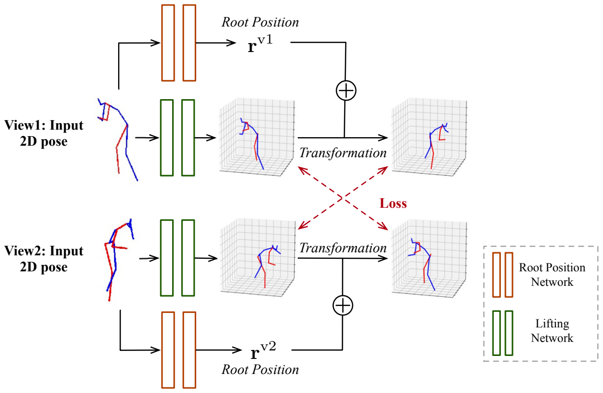

Figure 4: Illustration of the model pre-training.

where $\mathbf{R}_{\mathrm{1to2}}, \mathbf{R}_{\mathrm{2to1}} \in \mathbb{R}^{3\times3}$ are the rotation matrixes, and $\mathbf{t}_{\mathrm{1to2}}, \mathbf{t}_{\mathrm{2to1}} \in \mathbb{R}^3$ are the transformation vectors. With the extrinsic parameters of two cameras $\mathbf{R}_1, \mathbf{t}_1$ and $\mathbf{R}_2, \mathbf{t}_2$, we can directly obtain the rigid transformation parameters,

$$\begin{aligned} \mathbf{R}_{\mathrm{1to2}} &= \mathbf{R}_2\mathbf{R}_1^{\mathrm{T}}; \quad \mathbf{t}_{\mathrm{1to2}} = \mathbf{R}_1\left(\mathbf{t}_2 - \mathbf{t}_1\right), \\ \mathbf{R}_{\mathrm{2to1}} &= \mathbf{R}_1\mathbf{R}_2^{\mathrm{T}}; \quad \mathbf{t}_{\mathrm{2to1}} = \mathbf{R}_2\left(\mathbf{t}_1 - \mathbf{t}_2\right). \end{aligned} \tag{4}$$

If extrinsic parameters of cameras do not exist, we can use the positions of 2D joins of two views as calibration targets (Kocabas, Karagoz, and Akbas 2019). We assume the first camera as the center of the coordinate system, which means $\mathbf{R}_1$ is an identity matrix and $\mathbf{t}_1$ is a zero vector. For corresponding joints in $\mathbf{X}^{\mathrm{v1}}$ and $\mathbf{X}^{\mathrm{v2}}$, we find the fundamental matrix $\mathbf{F}$ satisfying $\mathbf{X}_i^{\mathrm{v1}}\mathbf{F}\mathbf{X}_i^{\mathrm{v2}} = 0, i = 1 \ldots N$, using RANSAC algorithm. From $\mathbf{F}$, we calculate the essential matrix $\mathbf{E}$ by $\mathbf{E} = \mathbf{P}_{\mathrm{v2}}^{\mathrm{T}}\mathbf{F}\mathbf{P}_{\mathrm{v1}}$, where $\mathbf{P}_{\mathrm{v1}}$ and $\mathbf{P}_{\mathrm{v2}}$ are the projection matrixes of cameras. By decomposing $\mathbf{E}$ with SVD, we obtain four possible solutions to $\mathbf{R}_{\mathrm{1to2}}$ and $\mathbf{t}_{\mathrm{1to2}}$. We decide on the correct one by verifying possible pose hypotheses using cheirality check. In the similar way, we can get $\mathbf{R}_{\mathrm{2to1}}$ and $\mathbf{t}_{\mathrm{2to1}}$. Since the calibrated $\mathbf{t}_{\mathrm{1to2}}$ and $\mathbf{t}_{\mathrm{2to1}}$ are unit vectors, we need to multiply them by the distance between two camera centers.

Next, according to multi-view consistency that the 2D projection of the transformed 3D skeleton should be the same with the 2D input of the target view, we design the transform re-projection loss as follows:

$$\begin{aligned} \mathcal{L}_{\text{t-reproj}} = \sum_i^N & \mathbf{w}_i^{\mathrm{v1}} \|\mathbf{X}_i^{\mathrm{v2}} - \rho(\tau(\tilde{\mathbf{Y}}_i^{\mathrm{v1}}))\|^2 \\ & + \mathbf{w}_i^{\mathrm{v2}} \|\mathbf{X}_i^{\mathrm{v1}} - \rho(\tau(\tilde{\mathbf{Y}}_i^{\mathrm{v2}}))\|^2. \end{aligned} \tag{5}$$

In this way, we construct supervision signals entirely relying on camera geometric prior. Compared with existing techniques that require unpaired 3D pose annotations or kinematic constraints, the proposed approach is simple and effective.

Training

It is challenging to train two inter-dependent branches from scratch without ground-truth annotations. We find that the

Table 1: Detailed results on H36M dataset under Protocol #1 and Protocol #2.

Protocol #1	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SittingD	Smoke	Wait	WalkD	Walk	WalkT	Avg
Reproj	390.2	441.6	479.3	422.8	503.4	479.0	400.6	471.5	568.5	662.2	483.6	423.8	473.2	414.5	413.5	468.5
Reproj+ADV	81.7	93.0	99.3	97.3	106.8	134.7	81.8	101.0	113.2	151.2	100.7	97.0	121.3	111.6	108.3	106.6
Trans_Reproj	49.7	54.5	58.0	56.8	63.4	80.0	52.4	52.7	71.4	78.3	58.9	55.2	60.0	43.8	49.6	59.0
Trans_Reproj+DA	48.7	53.6	54.7	55.1	61.3	76.1	51.5	50.3	68.0	75.9	56.7	53.8	58.8	42.6	47.9	57.0
Protocol #2	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SittingD	Smoke	Wait	WalkD	Walk	WalkT	Avg
Reproj	147.1	148.0	174.7	153.1	165.5	162.7	176.1	136.2	156.5	192.5	230.1	147.6	150.5	160.0	154.6	163.7
Reproj+ADV	64.8	69.4	77.4	74.4	78.4	94.2	60.4	68.9	81.5	113.1	74.3	70.4	84.8	82.4	81.6	78.4
Trans_Reproj	39.6	42.6	45.7	46.0	47.6	57.1	41.0	39.2	55.4	59.9	46.4	42.5	47.1	34.4	41.0	45.7
Trans_Reproj+DA	38.2	41.3	43.5	44.4	45.4	54.7	39.3	38.0	53.2	59.2	45.0	40.7	46.2	33.0	39.4	44.1

[1] ADV refers to the adversarial loss; DA means that the network is trained with data augmentation.

network cannot converge if we train it with random initialization. Thus, we design a pre-training technique to warm-up the network. As shown in Figure 4, the pre-training loss can be formulated as:

$$\mathcal{L}_{\text{pre-train}} = \sum_{i}^{N} \|\tau(\tilde{\mathbf{Y}}_i^{\text{v1}}) - \tilde{\mathbf{Y}}_i^{\text{v2}}\|^2 + \|\tau(\tilde{\mathbf{Y}}_i^{\text{v2}}) - \tilde{\mathbf{Y}}_i^{\text{v1}}\|^2. \quad (6)$$

It is designed according to multi-view consistency that the transformed 3D pose and the estimated 3D pose of the target view should be the same. Although this loss is not able to guide the lifting network to produce valid 3D poses, it effectively regularizes the output space of the root position branch. It can be regarded as an advanced initialization of the root position branch, which greatly reduces the difficulty of network convergence.

After pre-training, the network is fine-tuned using the re-projection loss and transform re-projection loss,

$$\mathcal{L}_T = \mathcal{L}_{\text{reproj}} + \lambda \mathcal{L}_{\text{t-reproj}}, \quad (7)$$

where λ is a hyper-parameter that is adapted to set under different datasets.

Experiments

Datasets

We perform extensive evaluations on two publicly available benchmarks. **Human3.6M (H36M)** (Ionescu et al. 2013) is one of the largest datasets for 3D human pose estimation, which is captured by MoCap system. It consists of 3.6 million images with 11 actors performing 15 actions such as eating, sitting and walking. They are captured from 4 calibrated cameras with known intrinsic and extrinsic parameters. In our experiments, we follow the standard protocol with 17-joint subset, use subjects S1, S5, S6, S7, S8 for training and S9, S11 for testing. **MPI-INF-3DHP (3DHP)** (Mehta et al. 2017) is a recently proposed 3D pose dataset constructed with both constrained indoor scenes and complex outdoor scenes. We use the five chest-height cameras and the provided 17 joints (compatible with H36M) for training, and we use the official test set, which contains 2929 frames from six subjects performing seven actions, for evaluation.

Table 2: Comparisons of different backbones on the H36M datasets.

Backbone	Protocol 1	Protocol 2
ResLinear	59.7	45.0
Ours	57.0	44.1
TemporalDilated	56.1	43.2

Evaluation Metrics For the H36M dataset, we consider two popular evaluation protocols. **Protocol 1** is the Mean Per Joint Position Error (MPJPE) in millimeters (mm). MPJPE is the mean euclidean distance between the ground-truth and predicted positions of the joints. **Protocol 2** is the Procrustes MPJPE (P-MPJPE), which aligns the estimated 3D pose to the ground-truth by a rigid transformation called Procrustes Analysis before computing the MPJPE.

The evaluation metrics for the 3DHP dataset include the adapted Percentage of Correct Keypoints (PCK) and corresponding Area Under Curve (AUC) (Mehta et al. 2017). The PCK indicates the percentage of joints whose estimated position is within 15cm of the ground-truth.

Data Augmentation The H36M dataset has only four calibrated camera views. Training with more camera views can improve model performance and generalization ability. We follow the technique proposed by (Fang et al. 2018) to simulate a series of virtual camera views. We extend the H36M dataset from 4 views to 12 views containing 8 virtual camera views, and we obtain the corresponding 2D pose of each sample through perspective projection to augment the training set. The detailed analysis will be shown in the following sections.

Implementation Details

In order to enable the proposed two-branch network to converge without any explicit 3D pose supervision, the training procedure contains two stages. First, we pre-train the network using the $\mathcal{L}_{\text{pre-train}}$ loss. We use the Adam as the optimizer and train the network for 20 epoches with learning rate 0.001. Next, the network is trained using the $\mathcal{L}_T$ loss for 300 epoches. The learning rate starts from 0.001 and drops by 0.1 each 100 epoches. During evaluation, for consistency

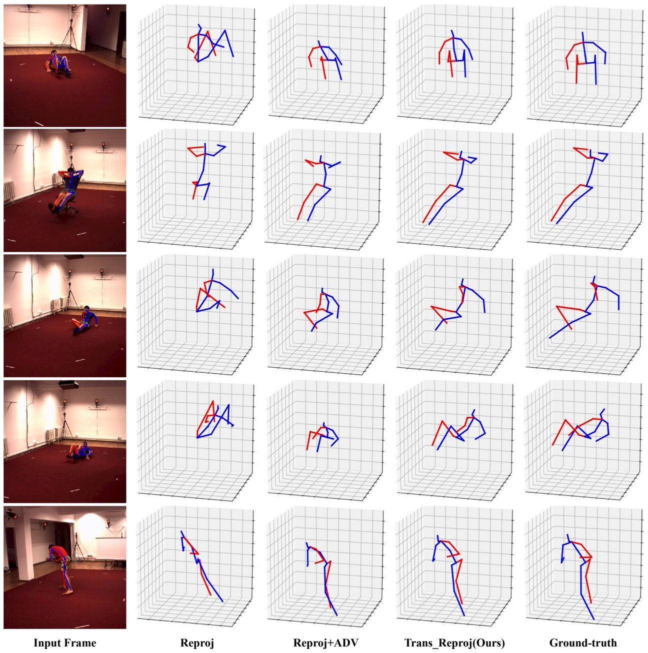

Figure 5: Results of different variants in some hard examples.

with other works, we only use the 2D-to-3D lifting branch to predict the relative 3D poses in the camera space, and not use the root position branch. We implement our method using the deep learning toolbox Pytorch.

Ablation Study

Analysis of Transform Re-projection Loss In order to evaluate the effectiveness of the proposed transform re-projection loss, we compare it with the existing popular technique, adversarial loss. We design several variants and compare the results under Protocol #1 (MPJPE) and Protocol #2 (P-MPJPE) on the H36M dataset. All variants use 2D poses extracted by the CPN network as inputs. Table 1 presents the quantitative results, and Figure 5 shows the results of different variants on several hard samples, i.e., with serious self-occlusion or far from the camera. It is obvious that only using the re-projection loss will obtain strange 3D skeletons that do not conform the human kinematics. Although adversarial loss can constrain the 3D poses using unpaired 3D pose annotations, it still can not produce precise 3D poses, especially when encountering samples with serious self-occlusion. Compared with adversarial loss, our method achieve significant performance improvements, and the MPJPE and P-MPJPE decrease by 47.6 and 32.7 (mm). This shows that the transform re-projection loss can effectively help the network learn geometric knowledge, which further constrains the estimated 3D poses to get more accurate results. Moreover, the MPJPE and P-MPJPE will decrease by extra 2.0 and 1.6 (mm) when using data augmentation, which verifies that training with more camera views can effectively facilitate the model performance.

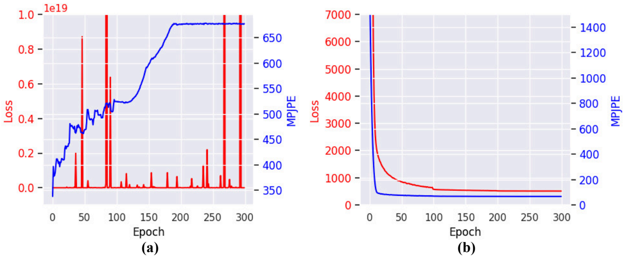

Figure 6: (a) and (b) are the loss and MPJPE curves of the network trained without and with pre-training respectively.

Analysis of Backbones Our method does not depend on any particular backbone. In this part, we investigate the performance of our method with different 2D-to-3D network backbones. ResLinear (Martinez et al. 2017) is the earliest and most commonly used backbone, which consists of fully connected layers and residual connections. TemporalDilated (Pavllo et al. 2019) is the latest proposed backbone that can explore the temporal information using dilated temporal convolutions. We feed it with 243 neighboring frames as inputs during training and testing. As shown in Table 2, our approach can achieve competitive results when using the simple ResLinear backbone. Therefore, the improvements of our method are not merely due to the better backbone. When using the TemporalDilated, our method gains obvious improvements, which benefits from the exploration of temporal information. These results illustrate that the proposed self-supervised training technique has strong versatility and is suitable for any novel 2D-to-3D network architecture.

Analysis of Network Pre-training In this section, we show the effectiveness of the network pre-training. Since the proposed network is trained without any 3D pose annotation, the pre-training is very important for our two-branch networks. As shown in Figure 6, the loss and MPJPE curves without pre-training violently oscillate. The network fails to converge despite our best efforts at tuning the hyperparameters. In contrast, the loss curve rapidly decreases, and we achieve low MPJPE value when using the proposed pre-training technique. It illustrates that pre-training technique is vital and effective in our work.

Analysis of Generalization Ability To demonstrate the generalization ability of our model, we train the network on the H36M dataset and evaluate it on the test split of the 3DHP dataset, which includes challenging outdoor scenes. We present some examples in Figure 7. It shows that our approach can successfully recover 3D poses on the datasets without being trained on them.

Comparisons with State-of-the-art Methods

In this section, we compare our method with recent weakly/self-supervised methods. First, we compare with them on the H36M dataset using protocol #1 and protocol #2 in Table 3. (Tung et al. 2017; Wandt and Rosenhahn 2019; Zhou et al. 2017) are based on re-projection loss and require additional unpaired 3D pose annotations. Compared

Table 3: Comparisons with recent weakly/self-supervised methods on the H36M dataset under evaluation Protocol #1 and Protocol #2.

Protocol #1	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SittingD	Smoke	Wait	WalkD	Walk	WalkT	**Avg**
Pavlakos *et al.* CVPR'17	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	118.4
Tung *et al.* ICCV'17 (†)	77.6	91.4	89.9	88	107.3	110.1	75.9	107.5	124.2	137.8	102.2	90.3	-	78.6	-	97.2
Wandt *et al.* CVPR'19 (†)	77.5	85.2	82.7	93.8	93.9	101.0	82.9	102.6	100.5	125.8	88.0	84.8	72.6	78.8	79.0	89.9
Wang *et al.* PAMI'19 (⋆;†)	50.0	60.0	54.7	56.6	65.7	52.7	54.8	85.9	118.0	62.5	79.6	59.6	41.5	65.2	48.5	63.7
Kocabas *et al.* CVPR'19	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	76.6
Ours	49.7	54.5	58.0	56.8	63.4	80.0	52.4	52.7	71.4	78.3	58.9	55.2	60.0	43.8	49.6	59.0
Protocol #2	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SittingD	Smoke	Wait	WalkD	Walk	WalkT	Avg
Zhou *et al.* ICCV'17 (†)	54.8	60.7	58.2	71.4	62.0	65.5	53.8	55.6	75.2	111.6	64.1	66.0	5.4	63.2	55.3	64.9
Drover *et al.* ECCV'18	60.2	60.7	59.2	65.1	65.5	63.8	59.4	59.4	69.1	88.0	64.8	60.8	64.9	63.9	65.2	64.6
Rhodin *et al.* ECCV'18 (†)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	98.2
Wandt *et al.* CVPR'19 (†)	53.0	58.3	59.6	66.5	72.8	71.0	56.7	69.6	78.3	95.2	66.6	58.5	63.2	57.5	49.9	65.1
Kocabas *et al.* CVPR'19	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	67.5
Chen *et al.* CVPR'19 (⋆)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	68.0
Ours	39.6	42.6	45.7	46.0	47.6	57.1	41.0	39.2	55.4	59.9	46.4	42.5	47.1	34.4	41.0	45.7

[1] (⋆) denotes that it takes advantage of the temporal information; (†) denotes that it requires unpaired or part of 3D pose annotations.

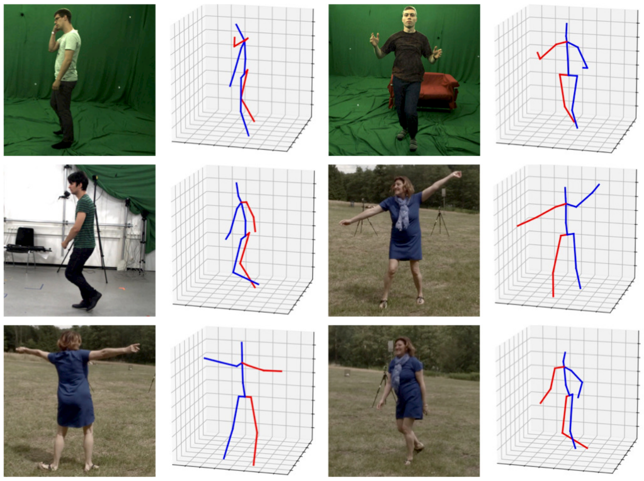

Figure 7: Quantitative results of our method (trained on the H36M dataset) on the 3DHP dataset.

with them, our method has an explicit improvement and obtains average errors of 59.0mm and 45.7mm under two evaluation protocols. Our method also outperforms (Kocabas, Karagoz, and Akbas 2019) that adopts multi-view information. We present its result obtained in the case of using ground-truth extrinsic parameters for fair comparisons. It illustrates that the proposed approach is a more effective way to exploit multi-view information. Table 4 shows the comparisons with state-of-the-art methods on the 3DHP dataset. In this setting, we train the network on the train set of the 3DHP dataset, and evaluate it on the test set following the PCK and AUC metrics. As seen, the PCK and AUC of our method reach 74.1 and 41.4 respectively, which outperform previous methods.

Table 4: Comparisons with recent weakly/self-supervised methods on the 3DHP dataset.

Method	PCK	AUC
Zhou *et al.* ICCV'17	69.2	32.5
Kocabas *et al.* CVPR'19	64.7	-
Chen *et al.* CVPR'19	71.1	36.3
Ours	74.1	41.4

Conclusion

In this work, we proposed a new self-supervised approach for 3D human pose estimation. The approach explored multi-view consistency to construct supervision signals for training a 2D-to-3D lifting network, which can effectively overcome the depth ambiguity problem. Note that our method simply applied multi-view information during training, and required only single view inputs during inference. Meanwhile, we designed a two-branch training architecture and pre-training technique to ensure the network can successfully converge and achieve excellent performance. Extensive ablation studies on the H36M and 3DHP datasets illustrated the effectiveness and generalization ability of our approach. The experiment results showed that our method obtained a superior performance over recent weakly/self-supervised methods.

Acknowledgments

This research was supported by National Key R & D Program of China (No. 2016YFB0801100), Beijing Natural Science Foundation (No. 4172054, L181010), National Basic Research Program of China (No. 2013CB329605), and Graduate Technological Innovation Project of Beijing Institute of Technology (No.2018CX20026).

References

Arnab, A.; Doersch, C.; and Zisserman, A. 2019. Exploiting temporal context for 3d human pose estimation in the wild. In *CVPR*, 3395–3404.

Belagiannis, V.; Amin, S.; Andriluka, M.; Schiele, B.; Navab, N.; and Ilic, S. 2014. 3d pictorial structures for multiple human pose estimation. In *CVPR*, 1669–1676.

Brau, E., and Jiang, H. 2016. 3d human pose estimation via deep learning from 2d annotations. In *3DV*, 582–591.

Chen, C.-H., and Ramanan, D. 2017. 3d human pose estimation= 2d pose estimation + matching. In *CVPR*, 7035–7043.

Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; and Sun, J. 2018. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, 7103–7112.

Chen, C.-H.; Tyagi, A.; Agrawal, A.; Drover, D.; MV, R.; Stojanov, S.; and Rehg, J. M. 2019. Unsupervised 3d pose estimation with geometric self-supervision. In *CVPR*, 5714–5724.

Dong, J.; Jiang, W.; Huang, Q.; Bao, H.; and Zhou, X. 2019. Fast and robust multi-person 3d pose estimation from multiple views. In *CVPR*, 7792–7801.

Fang, H.; Xu, Y.; Wang, W.; Liu, X.; and Zhu, S. 2018. Learning pose grammar to encode human body configuration for 3d pose estimation. In *AAAI*, 6821–6828.

Habibie, I.; Xu, W.; Mehta, D.; Pons-Moll, G.; and Theobalt, C. 2019. In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In *CVPR*, 10905–10914.

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. *PAMI* 36(7):1325–1339.

Iqbal, U.; Doering, A.; Yasin, H.; Krüger, B.; Weber, A.; and Gall, J. 2018. A dual-source approach for 3d human pose estimation from single images. *CVIU* 172:37–49.

Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-end recovery of human shape and pose. In *CVPR*, 7122–7131.

Kocabas, M.; Karagoz, S.; and Akbas, E. 2019. Self-supervised learning of 3d human pose using multi-view geometry. In *CVPR*, 1077–1086.

Martinez, J.; Hossain, R.; Romero, J.; and Little, J. J. 2017. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2640–2649.

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

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked hourglass networks for human pose estimation. In *ECCV*, 483–499.

Pavlakos, G.; Zhou, X.; Derpanis, K. G.; and Daniilidis, K. 2017a. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *CVPR*, 7025–7034.

Pavlakos, G.; Zhou, X.; Derpanis, K. G.; and Daniilidis, K. 2017b. Harvesting multiple views for marker-less 3d human pose annotations. In *CVPR*, 6988–6997.

Pavllo, D.; Feichtenhofer, C.; Grangier, D.; and Auli, M. 2019. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, 7753–7762.

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 *CVPR*, 8437–8446.

Rhodin, H.; Salzmann, M.; and Fua, P. 2018. Unsupervised geometry-aware representation for 3d human pose estimation. In *ECCV*, 750–767.

Sun, X.; Xiao, B.; Wei, F.; Liang, S.; and Wei, Y. 2018. Integral human pose regression. In *ECCV*, 529–545.

Tekin, B.; Márquez-Neila, P.; Salzmann, M.; and Fua, P. 2017. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *ICCV*, 3941–3950.

Tome, D.; Toso, M.; Agapito, L.; and Russell, C. 2018. Rethinking pose in 3d: Multi-stage refinement and recovery for markerless motion capture. In *3DV*, 474–483.

Tome, D.; Russell, C.; and Agapito, L. 2017. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *CVPR*, 2500–2509.

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

Wandt, B., and Rosenhahn, B. 2019. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *CVPR*, 7782–7791.

Wang, K.; Lin, L.; Jiang, C.; Qian, C.; and Wei, P. 2019. 3d human pose machines with self-supervised learning. *arXiv preprint arXiv:1901.03798*.

Wu, J.; Xue, T.; Lim, J. J.; Tian, Y.; Tenenbaum, J. B.; Torralba, A.; and Freeman, W. T. 2016. Single image 3d interpreter network. In *ECCV*, 365–382.

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 *CVPR*, 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 *CVPR*, 3425–3435.

Zhou, F., and De la Torre, F. 2016. Spatio-temporal matching for human pose estimation in video. *PAMI* 38(8):1492–1504.

Zhou, X.; Zhu, M.; Leonardos, S.; Derpanis, K. G.; and Daniilidis, K. 2016. Sparseness meets deepness: 3d human pose estimation from monocular video. In *CVPR*, 4966–4975.

Zhou, X.; Huang, Q.; Sun, X.; Xue, X.; and Wei, Y. 2017. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *ICCV*, 398–407.