On Boosting Single-Frame 3D Human Pose Estimation via Monocular Videos

*Zhi Li[1], *Xuan Wang[2], Fei Wang[2], and Peilin Jiang[1]

[1]School of Software Engineering, Xi'an Jiaotong University, China
[2]School of Electronic and Information Engineering, Xi'an Jiaotong University, China
{maleficentlee, xwang.cv}@gmail.com, {wfx, pljiang}@xjtu.edu.cn

Abstract

The premise of training an accurate 3D human pose estimation network is the possession of huge amount of richly annotated training data. Nonetheless, manually obtaining rich and accurate annotations is, even not impossible, tedious and slow. In this paper, we propose to exploit monocular videos to complement the training dataset for the single-image 3D human pose estimation tasks. At the beginning, a baseline model is trained with a small set of annotations. By fixing some reliable estimations produced by the resulting model, our method automatically collects the annotations across the entire video as solving the 3D trajectory completion problem. Then, the baseline model is further trained with the collected annotations to learn the new poses. We evaluate our method on the broadly-adopted Human3.6M and MPI-INF-3DHP datasets. As illustrated in experiments, given only a small set of annotations, our method successfully makes the model to learn new poses from unlabelled monocular videos, promoting the accuracies of the baseline model by about 10%. By contrast with previous approaches, our method does not rely on either multi-view imagery or any explicit 2D keypoint annotations.

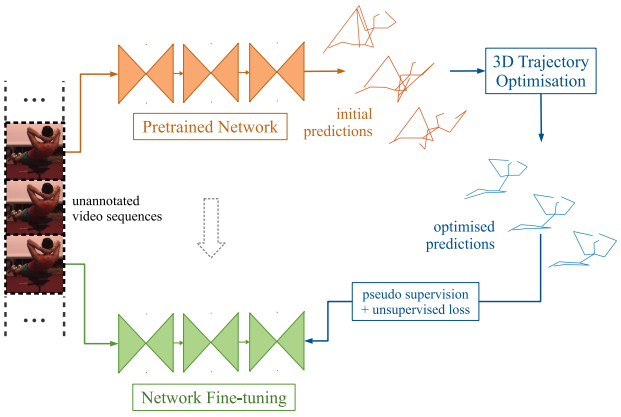

Figure 1. **Two-stage 3D human pose estimation framework.** Unannotated image frames from video sequences are fed into the initial 3D human pose estimation network which is trained by only a few annotated data to get initial predictions. Next, a 3D trajectory optimisation operation utilising the low rank property and temporal smoothness of video sequences is applied to these predictions to generate pseudo-annotations. These optimised predictions are then applied to the initial network as pseudo supervision, plus a geometry loss, to boost the performance of the network.

1. Introduction

Estimating the accurate 3D human pose from single image is a key foundation to massive applications in computer vision and graphics, which attracts lots of attentions from both societies. Recently, great progress in 3D human pose estimation have been made by leveraging the deep learning techniques. Despite the success, most of these approaches heavily rely on the availability of extensive training datasets. Nevertheless, using motion capture systems to capture the 3D annotations is usually constrained in studio environments. Furthermore, the manual annotation of 3D human pose is quite time-consuming and error-prone.

In the past few years, several approaches that employ unlabelled multi-view imagery or 2D-annotated images have been proposed to solve the problem of training data availability. The manual 2D keypoint annotation could be accurate enough, but is still tedious. In addition, employing the multi-view images requires specific multi-camera equipment. Nevertheless, a question still remains: How to exploit merely the unlabelled monocular videos to complement the training datasets for single-image 3D human pose estimation tasks? To this end, we propose an automatic method that collects accurate annotations of human motions

*These authors contributed equally to this work.

from monocular videos. As is shown in Figure 1, a baseline model is pre-trained with a small set of annotations. Then the outputs of this base model is optimised and exploited as annotations for further training. Compared with the previous methods, capturing monocular videos does not require any specific equipment, e.g. multi-camera system, and is not constrained in controlled environments. Furthermore, there is no manual intervention involved when complementing the datasets with videos.

Specifically, the automatic collection of annotations can be regarded as a problem of completing the consecutive 3D human motions. Relying on low-rank representation and temporal smoothness priors, we optimise the consecutive poses across the entire video by fixing the reliable estimations of the part of joints produced by the baseline model. As a by-product, the accurate geometry information, e.g. the individual limb length ratio of a specific subject, can be easily estimated from the optimised poses. Incorporating this geometry information encoded in loss function, we use the optimised poses as annotations to fine-tune the baseline model. Experiments show that our method significantly increases the accuracy of the 3D human pose estimations.

In summary, we focus on the single-image 3D human pose estimation tasks and propose a method to solve the data scarcity problem. Different from the previous approaches that learn the poses from consecutive image sequences, for our method, the video data is only exploited during training. Given a baseline model pre-trained with only a small set of annotations, a matrix completion based method is employed to automatically collect the 3D annotations from monocular videos. In this process, except for the small set of 3D annotations, our method does not require any human intervention such as manual annotation of 2D poses or calibration of multi-camera system which are usually employed by existing weakly-supervised approaches. As illustrated in the experiments on both Human3.6M and MPI-INF-3DHP datasets, our method successfully fine-tunes the pre-trained model to novel actions and subjects depicted in unlabelled monocular videos. Relying on this fine-tuning process, the accuracy of the estimated 3D human poses is promoted by about 10%.

2. Related Works

Nowadays, 3D human pose estimation [4, 6, 20, 11, 17, 16, 15, 19, 21, 22, 24, 25, 26, 28, 29, 30] has grown to the point where it can yield accurate poses even in realtime. Nonetheless, estimating the motions without sufficiently labelled datasets still remains an open problem. In this section, we briefly review the previous algorithms that focus on the data scarcity problem.

Data augmentation. Learning with synthetic data [5, 23, 27] is an alternative to tackle the data scarcity problem. With the advances in computer graphics, a number of methods synthetically generate the training images by replacing the background or subject appearance. In [23], a collage approach is introduced. They synthesise the training images with known 3D poses by composing human parts from different images to generate more realistic results. However, the diversity of the appearance and motion is not adequate, resulting in that the accuracy of the model trained using such datasets is limited.

Multi-view approaches. There are several methods that focus on learning the poses from synchronised multi-view images via view consistency regardless of the availability of the 3D ground truth. Given the generic 2D pose detector, the method in [17] automatically collects the annotations relying on the constraints from the calibrated multi-camera system. In [12], a marker-less motion capture system is employed to record the multi-view images and the 3D ground truth is estimated to each view. The weakly-supervised approach in [22] removes the requirement of calibration of the multi-view system and achieves a good level of performance. In [21], the amount of annotated training data is further reduced by learning a latent representation that can be integrated into semi-supervised learning. The main drawback of such methods is their requirements of specific devices to establish the multi-camera system. Moreover, for most multi-view approaches, the multi-camera system is assumed to be calibrated and synchronised.

Single-view approaches. There are also other authors who explore methods of complementing the training datasets with single-view images. In [30], a weakly-supervised approach is introduced, which aims to augment the fully-supervised training data with 2D annotations only. Furthermore, an adversarial learning method is presented in [29] which adopts the estimator from [30] as the generator. Additionally, incorporating 2D annotations with depth ordinal relation [15] is proved to be effective for learning 3D human poses. Though being more convenient than capturing the accurate 3D ground truth, manually annotating the 2D keypoints or depth ordinal relations is still labour-consuming.

In this paper, we aim to solve the problem of the scarcity of training data with neither multi-view imagery nor 2D poses. In such a challenging scenario, the proposed method effectively complements the training dataset with only unannotated monocular videos. We empirically show that the performance of the networks fine-tuned with the produced datasets improves significantly.

3. Technical Approach

In this section, we present our approach for augmenting the training datasets of single-image 3D human pose estimation with monocular videos.

We design a framework to fine-tune the network initialised with a small set of annotated training data with un-

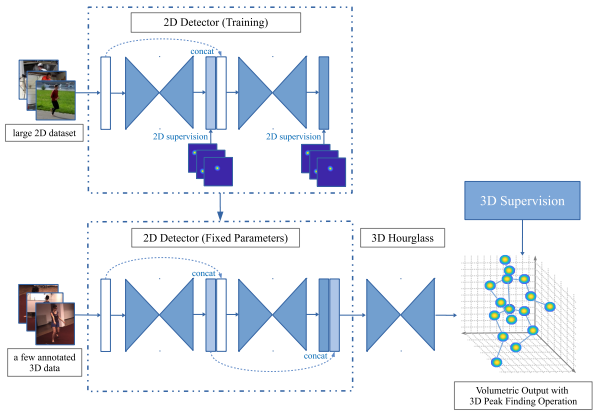

Figure 2. **Network architecture and the initialisation process.** Top: the 2D detection network (traditional stacked-hourglass network), trained by existing 2D human pose estimation datasets. Down: transfers feature extraction layers of the 2D detector to a 3D reconstruction network. A volumetric hourglass network is concatenated with the previously trained 2D detector (whose parameters are fixed) to lift the 2D features to 3D poses. A few annotated 3D data are used to perform the fully-supervised training.

annotated monocular video sequences. We firstly train a deep network with a small number of 3D annotated data to produce plausible 3D human poses. Then we use matrix completion methods to optimise the 3D poses of unannotated video sequences predicted by the insufficiently trained network, utilising the low-rank representation and temporal smoothness property of 3D human pose sequences. In the meantime, relatively accurate bone lengths per video can be collected from the predicted 3D human poses. The optimised predictions are then used as pseudo annotations to further train (fine-tune) the initial network. To better employ the accurate predictions and pass over the inaccurate ones, we introduce the weighted fidelity loss to be the pseudo-supervised term. In addition, we introduce the bone-length consistency loss as the unsupervised term.

3.1. Baseline Models

To reconstruct 3D human poses from monocular images, accurate 2D representations are usually required. The network architecture in [13], namely Stacked Hourglass network, is effective for extracting 2D features from images for the use of predicting 3D poses. Furthermore, inspired by [12], existing 2D human pose estimation datasets can be used to train a 2D human pose detector whose feature extraction layers can then be transferred to a 3D human pose estimation network.

To get accurate 3D poses directly from monocular images, we refer to the work of [16] which introduces a volumetric version of the stacked hourglass network. The 2D features extracted by the previously trained 2D detector can then be fed into a 3D hourglass network that requires only a small amount of 3D annotated data to get plausible 3D predictions on unannotated video sequences. Different from the network settings in [18], our 3D network directly outputs the 3D poses from single-frame images, without requiring the 2D poses to be the intermediate results, thus no 2D keypoint data is needed either to fine-tune the 2D detector or to train the 3D network. Figure 2 shows our network architecture and the whole process of initialising the network.

3.2. Auto-collecting the 3D Annotations

The network trained by only a few annotated 3D data gives initial predictions on unannotated video sequences. These predictions can be saved and augmented to perform as pseudo annotations for further (unsupervised) training of the network. However, due to the insufficient training of the 3D hourglass, the initial predictions cannot be precise enough. As human poses in video sequences are non-rigid and have properties such as being low-rank and temporally smooth, we can optimise the initial predictions by methods of matrix completion applied to 3D trajectories.

Formulations. The optimisation can be regarded as a matrix completion problem in which we utilise some of the poses with high confidence in one video and fill in the ones with low confidence. We combine the low-rank constraint of human poses in videos and their temporal smoothness property to get the following problem formulation:

$$\begin{aligned} \min \quad & \left\|\mathbf{X}^{\sharp}\right\|_{*}+\lambda_d\left\|\mathbf{X}\mathbf{D}\right\|_F^2+\lambda_c\left\|\mathbf{C}\right\|_{*}+\lambda_e\left\|\mathbf{E}\right\|_1 \\ s.t. \quad & \mathbf{X}=\mathbf{X}\mathbf{B}+\mathbf{E},\ P_{\mathbf{\Omega}}(\mathbf{X})=P_{\mathbf{\Omega}}(\mathbf{S}), \\ & \mathbf{B}=\mathbf{C},\ \mathbf{X}^{\sharp}=\mathbf{X}. \end{aligned} \tag{1}$$

In Eq. 1, λ_e, λ_c and λ_d are the weights for $\|\mathbf{E}\|_1$, $\|\mathbf{C}\|_*$ and $\|\mathbf{X}\mathbf{D}\|_F^2$, respectively. $\mathbf{X} \in \mathbb{R}^{3P\times F}$ (P for number of joint points per person and F for number of frames per video sequence) contains the 3D poses (one matrix for one subject per video). $\mathbf{S} \in \mathbb{R}^{3P\times F}$ is a constant matrix that equals to the pre-processed version of the initial pose predictions, in which the low-confidence elements are set to 0, with $P_{\mathbf{\Omega}}(\cdot)$ representing the operation of picking out the elements with high confidence. The confidence here can be expressed by the scores extracted from the heatmap outputs of the network, and we have the threshold score value $\tau \in [0, 1]$. $\mathbf{B} \in \mathbb{R}^{F\times F}$ is the self-representation coefficient matrix inspired by [10] and [31], and $\mathbf{E} \in \mathbb{R}^{3P\times F}$ is the error matrix. $\mathbf{D} \in \mathbb{R}^{F\times(F-1)}$ is a block-diagonal sparse matrix composed of 1 and -1 for computing the temporal smoothness of pose sequences. The matrices $\mathbf{X}^{\sharp} \in \mathbb{R}^{3P\times F}$ and $\mathbf{C} \in \mathbb{R}^{F\times F}$ equal to $\mathbf{X}$ and $\mathbf{B}$ respectively, which per-

form as the auxiliaries for minimising the nuclear norm in the process of the optimisation.

Optimisation. Eq.1 can be solved by the Augmented Lagrangian Methods (ALMs) [2]. By converting Eq.1 in to the augmented Lagrangian form, the converted cost function can be split into five subproblems, the variables inside which can be efficiently solved by the ALM, minimising the cost function $\mathcal{L}$ by iteratively updating individual variable while the other variables are fixed. Each of the subproblems is solvable individually following some developed techniques such as the Singular Value Thresholding (SVT) operation[3], the Shrinkage Operator[9] or simply solving the linear equations. In our settings, $\mathbf{X}$ and $\mathbf{X}^\sharp$ are initialised by the predicted poses. $\mathbf{B}$ and $\mathbf{C}$ are initialised by identity matrices. The elements in $\mathbf{E}$ are all set to zero at the beginning. After the convergence of the ALM, we take $\mathbf{X}$ as the optimised results who will further be used as pseudo supervision in the network fine-tuning step.

3.3. Fine-tuning the Initial Network

In the previous steps, image frames from unannotated video sequences are fed into the initial network to get 3D predictions, then these predictions are optimised and saved. In the fine-tuning step, these saved predictions are used with augmentation to "supervise" the further training of the initial network. These operations are like automatically collecting "annotations" for unannotated videos.

However, in the optimised predictions there are still many errors, and these errors are very likely to mislead the further training of the network. To alleviate the erroneous impact caused by these errors, we propose to weight the pseudo-supervision term in the loss function by the confidence score of each prediction, and add a constraint on bone-length consistency for the same person in the same video. The complete loss function for fine-tuning the initial network is as follows:

$$L(\theta) = S_w(\theta; \mathbf{p}, \mathbf{s}) + \gamma U(\theta; \mathbf{b}), \tag{2}$$

where $S_w(\cdot)$ is the weighted fidelity pseudo-supervision term of the loss function and $U(\cdot)$ is the unsupervised one (bone-length consistency loss). θ denotes the set of network parameters, $\mathbf{p}$ represents the matrix of human poses, $\mathbf{s}$ are the confidence scores for the corresponding predictions and $\mathbf{b}$ denotes the bone lengths. γ is a weight which can be empirically set to balance the two terms.

Weighted Fidelity Loss. The optimised predictions of the unannotated images are utilised in the network fine-tuning process as pseudo-supervision. As we regard the initial predictions with high confidence as precise ones, we do not want the corresponding outputs of the fine-tuned network to drift too much from their initially predicted values. Since our network outputs volumetric heatmaps, we can directly extract heatmap confidence scores for the predicted joints. Therefore, we make use of these confidence scores and propose a weighted fidelity pseudo-supervised loss function, written as:

$$S_w(\theta; \mathbf{p}, \mathbf{s}) = \sum_{i=1}^{F} \sum_{j=1}^{P} W(s_{ij}) \cdot \left\| \hat{p}_{ij} - p_{ij} \right\|^2, \tag{3}$$

where F denotes the number of frames and P denotes the number of joints per skeleton. $W(\cdot)$ is the weight function which is defined as:

$$W(s) = \begin{cases} 1, & s > \tau \\ s, & s \leq \tau \end{cases}, \tag{4}$$

where $\tau \in [0, 1]$ denotes the threshold value defined to judge whether the predicted joint is reliable enough to supervise the network fine-tuning.

Bone-length Consistency Loss. The unsupervised geometry term, namely bone consistency loss, is defined as follows:

$$U(\theta; \mathbf{b}) = \sum_{i=1}^{F} \sum_{k=1}^{B} \left\| \hat{b}_{ik} - b_{ik} \right\|^2, \tag{5}$$

where B is the number of bones of one skeleton. The bone length b can be calculated and collected through the initial predictions. In this paper we adopt 11 bones (right and left lower and upper limbs, pelvis to left and right hips, chin to headtop) whose lengths are basically invariant while the subjects are moving.

As our initial network outputs volumetric heatmaps instead of coordinates of each joint, during the training in the network initialisation step (which has only the supervised loss), the L2 loss is applied to the heatmaps. To perform network inference and validation, we can extract joint coordinates from the predicted heatmaps by argmax operations. However, during network fine-tuning, as the argmax operation is not differentiable, our bone-length consistency loss whose calculation requires coordinates cannot be directly applied. To back-propagate the network outputs through direct coordinates, we replace the argmax operation with a 3D version of the peak finding operation in [7], in which the weighted sum of heatmap confidence scores within the cube centred around the coarse location of the maximum score serves as the predicted joint coordinate.

4. Experiments

In this section, we conduct experiments in various settings to analyse the performance of our trajectory optimisation and its ability on boosting the insufficiently trained 3D pose estimation network.

4.1. Experimental Configurations

Datasets. Our methods are firstly tested on the well-known Human3.6M (H36M)[8] dataset. It contains video sequences of totally 3.6 million frames of 11 human subjects performing different actions, in which subjects 1, 5, 6, 7, 8, 9 and 11 are annotated with 3D poses. The original videos are recorded at a frame rate of 50fps, and we downsample them to 10fps. Due to the fact that our 2D feature extraction layers are trained on MPII 2D human pose dataset[1], we adopt the 16-joint skeleton of this dataset for H36M as well. Subjects 1, 5, 6, 7 and 8 are treated as training subjects (fully-supervised or semi-supervised), and subjects 9 and 11 are for validation. We also test our methods on the more recent MPI-INF-3DHP (3DHP) dataset[12]. The training set of 3DHP is similar to H36M but with more challenging actions, containing 8 subjects acting in the same indoor scenes with green screen. The original videos are recorded at 25 or 50fps, and we downsample them to 12.5fps. The test set of 3DHP is much smaller, containing roughly 3k image frames in total of 6 subjects. The two of the 6 sequences are in the same green screen indoor scene as the training set, two are in different indoor scenes (without green screen), and the other two are outdoors. Our network only requires the 3D pose annotations in camera coordinates, and we train a single model for all actions.

Training protocols. We split the training datasets into two parts: the labelled one and the unlabelled one. The baseline models are trained with the labelled data, and the annotations of the unlabelled data are automatically collected and exploited for the further fine-tuning stage. To illustrate the effectiveness of our approach, we design different data split patterns on the H36M and 3DHP datasets. **a) Subject-wise protocols.** We split the training datasets according to different subjects. For H36M, we mainly use the annotations of subject 1 as the labelled set and S5,6,7,8 as the unlabelled set; for 3DHP, we use labelled S1 to train initially and unlabelled S2-8 to fine-tune (*subject-wise protocol S1*). While testing the influence of varying number of labelled subjects on network performance with H36M, we progressively increase the number of subjects in the labelled set (*subject-wise protocol S1, S15, S156 and S1567*). **b) Action-wise protocols.** We evaluate the framework's ability of learning new actions by splitting the H36M dataset according to different actions. Firstly, regardless of the different degrees of difficulties of the actions, we use the first half of the actions (Directions, Discussion, Eating, Greeting, Phoning, Photo, Posing and Purchases) as the labelled set and the rest as unlabelled set (*action-wise protocol half*). Then, to illustrate that our auto-collected annotations can enhance the network's performance on hard poses, we use the 11 easier poses to train fully-supervised and the rest 4 unlabelled, harder poses (Photo, Purchases, Sitting, SittingDown) to fine-tune (*action-wise protocol hard*).

Evaluation Metrics. The accuracy of the predicted 3D poses are commonly evaluated in terms of the mean per-joint position error (MPJPE), namely the average per-joint distance from the predictions centred around the root joints (pelvis) to the ground-truth annotations (in mm). In addition, inspired by [22], we adopt another two metrics – normalised mean per-joint error (NMPJPE) and mean per-joint position error after Procrustes alignment (PMPJPE). NMPJPE is set to avoid the dependence of subjects height, in which a scale factor is applied to the predictions so as to minimise the squared distance between annotations and predictions. When Procrustes alignment is applied to the predictions before calculating the mean per-joint error, the metric becomes PMPJPE, which is independent to both scale and orientation. In this paper, all of the three evaluation metrics are demonstrated in mm.

Implementation Details. As is described in section 3.1, our initial 3D human pose estimation network is a 3 stack volumetric hourglass network, with the first 2 stacks initialised by transfer learning of a 2D detector trained on the MPII 2D human pose dataset. The parameters of these two stacks are then fixed (without fine-tuning with the 2D annotations from the 3D datasets), therefore no explicit 2D keypoint annotations of the 3D datasets are required in the training of the 3D network. For the fully-supervised training on subsets of the datasets, we train with batch size 2 at a learning rate of 2.5×10^{-4}. For the network fine-tuning, we train for 2 epochs on the whole unannotated image sequences with batch size 4 at learning rate 2.5×10^{-5}. Both the annotations for fully-supervised training and the pseudo-annotations (optimised predictions) for network fine-tuning are augmented with 50% chance flipping, $-30^{\circ} \sim +30^{\circ}$ rotation and $0.75 \sim 1.25$ times scaling.

4.2. 3D Trajectory Optimisation Results

The baseline network trained on the labelled part of the datasets yields plausible 3D pose predictions on the unlabelled part of the datasets, and the predictions are then optimised using our trajectory completion method. Table 1 and Table 2 gives the 3D trajectory optimisation results of H36M S5-8 and 3DHP S2-8 under subject-wise protocol S1. Table 1 and 2 show obvious improvements in all of the evaluation metrics after optimisation. In addition, it can be observed that reliable predictions and optimisations can be selected using the heatmap score values. As the optimisation imposes temporal smoothness property to the video sequences, some of the erroneous joints are corrected by the preceding or subsequent poses. Figure 3 visualises some 3D pose examples before and after optimisation, showing that the optimised predictions (the blue skeletons) are closer to the ground truth, especially for those occluded joints.

	MPJPE	NMPJPE	PMPJPE
Predictions (all)	91.20	86.83	75.52
Optimisations (all)	**81.67**	**77.36**	**66.88**
Preds (scores $> \tau$)	71.12	68.77	62.92
Optis (scores $> \tau$)	**64.16**	**60.88**	**57.79**

Table 1. **H36M optimisation results under subject-wise protocol S1.** The predictions (all) are the direct outputs of S5-8 of H36M by the network trained on S1, and the optimisations (all) are the optimised results. There are 86.38% predictions whose scores are greater than the threshold $\tau = 0.7$.

	MPJPE	NMPJPE	PMPJPE
Predictions (all)	132.25	128.87	109.15
Optimisations (all)	**125.97**	**123.53**	**98.41**
Preds (scores $> \tau$)	100.24	101.57	86.65
Optis (scores $> \tau$)	**98.63**	**98.32**	**81.08**

Table 2. **3DHP optimisation results under subject-wise protocol S1.** The predictions (all) are the direct outputs of S2-8 of 3DHP by the network trained on S1, and the optimisations (all) are the optimised results. There are 80.96% predictions whose scores are greater than the threshold $\tau = 0.5$.

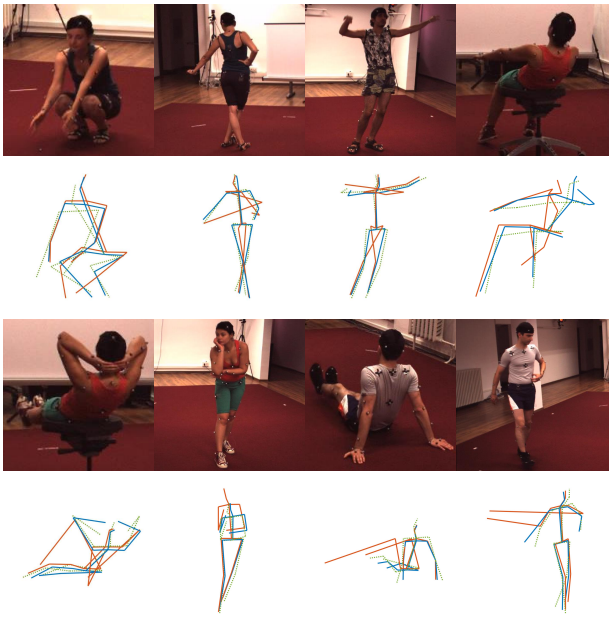

Figure 3. **Qualitative H36M optimisation results under subject-wise protocol S1.** Visualisations of some 3D poses on H36M S5-8 before and after 3D trajectory optimisation. The green dashed skeletons are the ground truth poses, the red ones are the initial network predictions, and the blue ones are the optimised predictions.

4.3. Network Fine-tuning Results

Subject-wise results. Table 3 shows the prediction errors of the network trained on H36M under subject-wise protocol S1, fine-tuned with different loss settings and their comparison with some state-of-the-arts methods. The data splitting pattern of our experiment, i.e. S1 for fully-supervised training (baseline) and S5-8 for network fine-tuning (fine-tune), without 2D fine-tuning, is the same as that of Rhodin et al. [22], who utilise multi-view information to perform the semi-supervised training. Our baseline is comparative to theirs, but our network fine-tuning results significantly outperforms theirs with the usage of the same amount of data. Pavllo et al. [18] design a deep network which accepts 2D annotations (or detections) as input to reconstruct 3D poses, and their semi-supervised training is performed via reprojecting the predicted 3D poses to 2D annotations. Their results in Table 3 are generated with the provided stacked hourglass detections without fine-tuning (SH PT), in single frame setting. The baselines are also comparative to ours, but their semi-supervised training fails when there are no ground truth 2D annotations available to fine-tune the 2D detector (i.e. the 2D detections are not highly accurate). Under the same configurations, our two-stage training framework works well, and the network works in single frame, which is much easier and faster to train and does not require consecutive image frames during training or in test time.

	MPJPE	NMPJPE	PMPJPE
Rhodin[22] baseline	99.6	91.5	-
Rhodin[22] semi-sup	98.5	88.8	-
Pavllo[18] baseline	98.6	93.8	70.3
Pavllo[18] semi-sup	119.3	113.7	92.8
Ours baseline	97.7	93.4	75.6
original supervision	94.5	89.7	69.3
+ weighted fidelity	93.0	88.6	69.0
+ bone consistency	92.6	87.7	68.8
optimised supervision	91.0	82.1	67.6
+ weighted fidelity	90.5	81.4	67.1
+ bone consistency	**88.8**	**80.1**	**66.5**

Table 3. **H36M fine-tuning results under subject-wise protocol S1.** The network (initialised by H36M, S1) is fine-tuned with unannotated S5-8 images under different loss configurations. Results are generated on the H36M validation set.

It can also be observed from Table 3 how our trajectory optimisation, weighted fidelity loss and bone length consistency loss incrementally improve the baseline network. Each component improves the network a little further, and finally the initial network gains about 10% performance boost in all of the three evaluation metrics. Table 4 provides a more detailed, action-wise results of this ablative experiments. For all of the actions, results supervised by optimised predictions outperform the baseline, demonstrating the effectiveness of our annotation auto-collecting scheme. For most of the actions, optimised supervision with weighted fidelity plus bone length consistency loss obtain the best results. Figure 4 visualises the predictions of the

	Direct	Discuss	Eat	Greet	Phone	Photo	Pose	Purchase
Baseline	78.90	92.80	82.09	86.34	94.10	113.21	83.75	110.55
+ optimised supervision	71.47	87.45	77.36	79.94	87.96	107.69	74.47	108.40
+ weighted fidelity	70.86	85.55	77.40	78.92	87.93	108.48	73.53	107.03
+ bone consistency	**70.44**	**83.61**	**76.59**	**77.91**	**85.43**	**106.14**	**72.26**	**102.93**
	Sit	SitDown	Smoke	Wait	WalkDog	Walk	WalkPair	**Average**
Baseline	125.45	185.76	90.57	82.24	99.83	67.04	79.86	97.72
+ optimised supervision	118.34	168.85	84.25	77.28	94.71	60.97	72.71	91.05
+ weighted fidelity	117.91	171.93	83.52	75.82	**93.55**	60.52	71.15	90.50
+ bone consistency	**115.79**	**164.99**	**82.43**	**74.34**	94.61	**60.15**	**70.65**	**88.77**

Table 4. **Ablative H36M fine-tuning results under subject-wise protocol S1.** Network fine-tuning results shown in MPJPE, evaluated by actions on H36M validation set under different loss configurations. Best in bold.

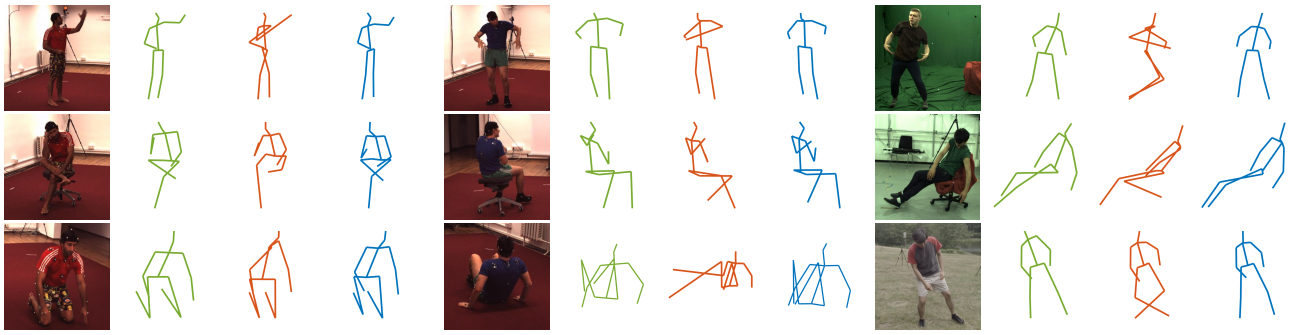

Figure 4. **Qualitative fine-tuning results under subject-wise protocol S1.** Visualisations of some of the final network fine-tuning results on the validation set, under subject-wise protocol S1. The first 8 columns are H36M results, and the last 4 are 3DHP results (of green screen studio, studio without green screen, and outdoors). The green skeletons are the ground truth poses, the red ones are the baseline network predictions, and the blue ones are the finial results of the fine-tuned network.

initial network and the network fine-tuned with weighted fidelity optimised supervision with bone length consistency loss, in which obvious improvements can be observed. The joints which are undetected or wrongly detected (interchanged with other joints) are effectively corrected through network fine-tuning.

We further test our method with varying number of labelled subjects. With the varying of the labelled and unlabelled subjects, our method still works well in boosting the performance of the baseline network. Figure 5 shows the network fine-tuning results under subject-wise protocol S15, S156 and S1567. It is demonstrated that our framework consistently improves the network with varying labelled subjects.

Action-wise results. To test our framework's ability of learning new actions, we perform action-wise training on H36M under action-wise protocol half and action-wise protocol hard. Table 5 compares the baseline and the fine-tuned results of the half of unlabelled actions on the H36M validation set under action-wise protocol half. Table 6 gives the results on the H36M validation set of the 4 unlabelled hard actions under action-wise protocol hard. It can be observed

	Sit	SitDown	Smoke	Wait
Baseline	111.88	254.42	69.25	66.80
Fine-tune	**101.08**	**215.02**	**64.71**	**61.57**
	WalkDog	Walk	WalkPair	**Average**
Baseline	75.95	57.65	61.36	98.92
Fine-tune	**71.95**	**53.63**	**56.75**	**88.75**

Table 5. **H36M fine-tuning results under action-wise protocol half.** The difficulties of labelled and unlabelled actions are similar. Results are generated on the part of unlabelled training actions in H36M validation set, shown in MPJPE.

	Photo	Purch.	Sit	SitDown	**Avg.**
Baseline	85.00	85.01	107.85	252.06	138.09
Fine-tune	**83.22**	**82.76**	**97.44**	**200.32**	**119.76**

Table 6. **H36M fine-tuning results under action-wise protocol hard.** The 4 unlabelled actions are more challenging. Results are generated on this 4 hard actions in H36M validation set, shown in MPJPE.

that the performance of the networks under both protocols are boosted on the new actions, showing that our two-stage framework has the ability to help with learning new actions from the unsupervised data, especially for challenging ac-

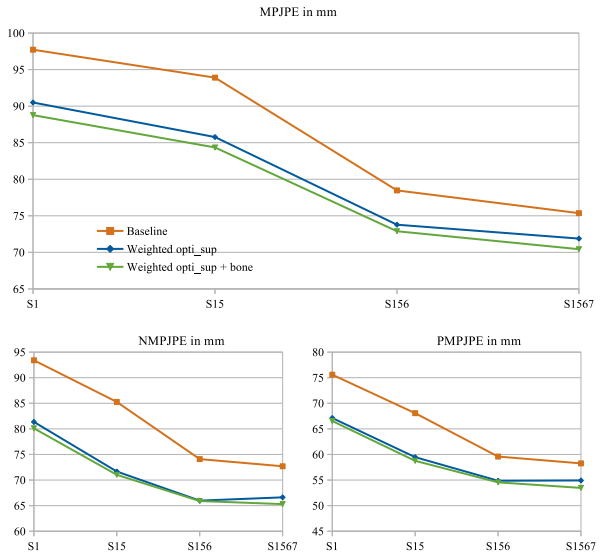

Figure 5. **Training with varying number of labelled H36M subjects.** MPJPEs, NMPJPEs and PMPJPEs under subject-wise protocol S1, S15, S156 and S1567.

		MPJPE	NMPJPE	PMPJPE
Studio GS	Baseline	124.52	121.33	98.41
	Fine-tune	**113.49**	**111.65**	**90.85**
Studio no GS	Baseline	151.45	149.15	122.46
	Fine-tune	**138.25**	**136.43**	**103.52**
Outdoors	Baseline	187.06	180.46	158.67
	Fine-tune	**171.16**	**167.18**	**148.89**
All	Baseline	149.79	146.07	122.07
	Fine-tune	**136.76**	**134.40**	**109.84**

Table 7. **Results of different scenes on 3DHP.** Network fine-tuning results on 3DHP validation set under subject-wise protocol S1. All of the videos in the training set are recorded in an indoor studio with green screen (Studio GS). Results are generated on the validation set, in which all of the three scenes (Studio GS, Studio no GS, and Outdoors) are presented.

tions.

In addition, recent studies on semi-supervised learning [14] addresses the problem of "class distribution mismatch", which refers to whether labelled and unlabelled data coming from different classes limits the performance of semi-supervised learning of classification tasks. To validate our framework in this setting, we choose actions contained in labelled and unlabelled part as dissimilar as possible. We train an initial model with annotated data of H36M "discussion" (mainly standing), and fine-tune with unannotated "phoning" (mainly sitting). After fine-tuning, both the actions get MPJPE drop: Discussion from 82.2mm to 79.2mm, Phoning from 144.5mm to 134.2mm, showing that our framework is valid despite class distribution mismatch.

Generalising to new scenes and outdoor captures. As both the training and the validation set of H36M are indoors, we also test our two-stage training scheme on a more recent and more challenging dataset – MPI-INF-3DHP, whose validation set contains images of not only the same indoor scenes as of the training set, but also new indoor scenes and outdoor captures. The experiments on 3DHP are conducted under subject-wise protocol S1. Table 7 shows the results of different scenes on 3DHP validation set. It can be observed that our two-stage training framework improves the results in all of the scenes on all the three evaluation metrics, showing that the framework generalises to different scenes and outdoor captures. Some results are visualised in Figure 4. Obvious improvements can be observed from the reconstructed 3D poses for all of the three types of scenes.

Cross-dataset validation. To validate our framework in a more real-world like scenario, we conduct experiments of cross-dataset transfer learning, transferring the model trained on one dataset to a new dataset using only the unlabelled data of the new dataset. We use unlabelled data from 3DHP to fine-tune the model trained fully supervised on H36M. We randomly select 5 videos from 3DHP s1-5 as unannotated training data, and 3 videos from s6-8 for testing. The optimisation makes MPJPE on s1-5 drops from 192.6mm to 176.6mm; after using the optimised predictions to fine-tune the network, results on s6-8 drop from 206.8mm to 153.8mm, showing that the initial network is promoted by 25.6% on the 3DHP dataset without introducing any 3DHP annotation.

5. Conclusion

In this paper, we introduce a two-stage framework for single-image 3D human pose estimation to boost the neural network's performance by auto-collecting annotations for unannotated monocular videos. Extensive experiments demonstrate the effectiveness of our framework, showing that it can successfully assist the learning of new 3D human poses from unannotated monocular videos. This framework can be applied when there are not enough annotated data to train a 3D human pose estimation network while a lot of unannotated monocular videos are available, which is often the case in real world scenarios.

6. Acknowledgements

This work is supported by National Science and Technology Major Project 2018ZX01008103, the Fundamental Research Funds for the Central Universities and Scientific and Technological Innovation Project 201809162CX3JC4.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark

and state of the art analysis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3686–3693, 2014.

[2] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[3] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[4] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation = 2d pose estimation + matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7035–7043, 2017.

[5] Wenzheng Chen, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Synthesizing training images for boosting human 3d pose estimation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 479–488. IEEE, 2016.

[6] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3d human pose from structure and motion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 668–683, 2018.

[7] Xuanyi Dong, Shoou-I Yu, Xinshuo Weng, Shih-En Wei, Yi Yang, and Yaser Sheikh. Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. In *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 360–368, 2018.

[8] Catalin Ionescu, Joao Carreira, and Cristian Sminchisescu. Iterated second-order label sensitive pooling for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1661–1668, 2014.

[9] Zhouchen Lin, Minming Chen, and Yi Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

[10] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(1):171–184, 2013.

[11] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017.

[12] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 International Conference on 3D Vision (3DV)*, pages 506–516. IEEE, 2017.

[13] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.

[14] Avital Oliver, Augustus Odena, Colin Raffel, Ekin D Cubuk, and Ian J Goodfellow. Realistic evaluation of semi-supervised learning algorithms. 2018.

[15] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7307–7316, 2018.

[16] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7025–7034, 2017.

[17] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Harvesting multiple views for marker-less 3d human pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6988–6997, 2017.

[18] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. pages 7753–7762, 2019.

[19] Alin-Ionut Popa, Mihai Zanfir, and Cristian Sminchisescu. Deep multitask architecture for integrated 2d and 3d human sensing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6289–6298, 2017.

[20] Mir Rayat Imtiaz Hossain and James J Little. Exploiting temporal information for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 68–84, 2018.

[21] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 750–767, 2018.

[22] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. Learning monocular 3d human pose estimation from multi-view images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8437–8446, 2018.

[23] Grégory Rogez and Cordelia Schmid. Mocap-guided data augmentation for 3d pose estimation in the wild. In *Advances in neural information processing systems*, pages 3108–3116, 2016.

[24] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3433–3441, 2017.

[25] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2602–2611, 2017.

[26] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2500–2509, 2017.

[27] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 109–117, 2017.

[28] Min Wang, Xipeng Chen, Wentao Liu, Chen Qian, Liang Lin, and Lizhuang Ma. Drpose3d: Depth ranking in 3d human pose estimation. 2018.

[29] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5255–5264, 2018.

[30] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 398–407, 2017.

[31] Yingying Zhu, Dong Huang, Fernando De La Torre, and Simon Lucey. Complex non-rigid motion 3d reconstruction by union of subspaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1542–1549, 2014.