Towards Alleviating the Modeling Ambiguity of Unsupervised Monocular 3D Human Pose Estimation

Zhenbo Yu[1,2], Bingbing Ni[1,2]*, Jingwei Xu[1,2], Junjie Wang[1,2], Chenglong Zhao[1,2], Wenjun Zhang[1,2]
[1]Shanghai Jiao Tong University, [2]Shanghai Key Lab of Digital Media Processing & Transmission
{yuzhenbo, nibingbing, xjwxjw, dreamboy.gns, cl-zhao, zhangwenjun}@sjtu.edu.cn

Abstract

In this work, we study the ambiguity problem in the task of unsupervised 3D human pose estimation from 2D counterpart. On one hand, without explicit annotation, the scale of 3D pose is difficult to be accurately captured (scale ambiguity). On the other hand, one 2D pose might correspond to multiple 3D gestures, where the lifting procedure is inherently ambiguous (pose ambiguity). Previous methods generally use temporal constraints (e.g., constant bone length and motion smoothness) to alleviate the above issues. However, these methods commonly enforce the outputs to fulfill multiple training objectives simultaneously, which often lead to sub-optimal results. In contrast to the majority of previous works, we propose to split the whole problem into two sub-tasks, i.e., optimizing 2D input poses via a scale estimation module and then mapping optimized 2D pose to 3D counterpart via a pose lifting module. Furthermore, two temporal constraints are proposed to alleviate the scale and pose ambiguity respectively. These two modules are optimized via a iterative training scheme with corresponding temporal constraints, which effectively reduce the learning difficulty and lead to better performance. Results on the Human3.6M dataset demonstrate that our approach improves upon the prior art by 23.1% *and also outperforms several weakly supervised approaches that rely on 3D annotations. Our project is available at https://sites.google.com/view/ambiguity-aware-hpe.*

1. Introduction

Human pose estimation has received considerable attention in computer vision community [2, 6, 28, 37]. As a fundamental module, it is widely used in many downstream applications, such as body reconstruction [14], robotics manipulation [26], and augmented reality [10]. In this paper, we are interested in unsupervised monocular 3D pose estimation. Due to the high-cost and time-consuming annotation procedure of 3D skeleton, unsupervised / weakly supervised 3D pose estimation [40, 3] has turned into an emerging trend in this field.

Recent unsupervised approaches [29, 12, 19, 17], i.e., without access to 3D annotations in any form, mainly use 2D annotations [3], unlabelled multi-view imagery [12] or learned 3D priors [17] to bypass the need of 3D annotation. Compared to easily accessible 2D annotations, the manual 3D priors are tedious and employing the multi-view images requires specific multi-camera equipment. Recently, Chen et.al. [3] propose a geometric constraint for single-frame unsupervised 3D pose estimation, which get rid of the need for the multi-view cameras.

However, there still exist two challenges remained to be solved: (a) **Scale ambiguity**. The scale of 3D pose is hard to be accurately captured if without the supervision of 3D annotations. We experimentally find that the scale of estimated 3D skeletons under unsupervised setting is prone to drift far away from ground truth. Simply enforcing the scale consistency of the predicted 3D skeleton (i.e., bone length consistency loss [20]) only leads to marginal improvements, which is supported by our ablation study in Sec 4.4. (b) **Pose ambiguity**. Lifting 2D pose to 3D counterpart is inherently ambiguous [19], where a single 2D pose possibly corresponds to multiple 3D poses. Multi-view data [3] is able to effectively address such kind of ambiguity. Chen et.al. [3] firstly propose to generate pseudo view to alleviate the ambiguity, which, however, ignores the temporal constraints between frames. 3D constraints (e.g., cycle loss [3], bone length consistency loss [20], camera projection loss [17]) have been previously proposed to address the above two challenges. However, they are commonly considered as auxiliary losses, i.e., enforcing the outputs to fulfill multiple training objectives simultaneously. Such kind of training scheme often leads to sub-optimal results [36].

To solve the above challenges, we propose to split the whole problem into two sub-tasks, i.e., optimizing 2D pose via scale estimation (short termed as scale estimation) and lifting optimized 2D pose to 3D counterpart (short termed as pose lifting). Furthermore, two temporal consistency con-

*corresponding author

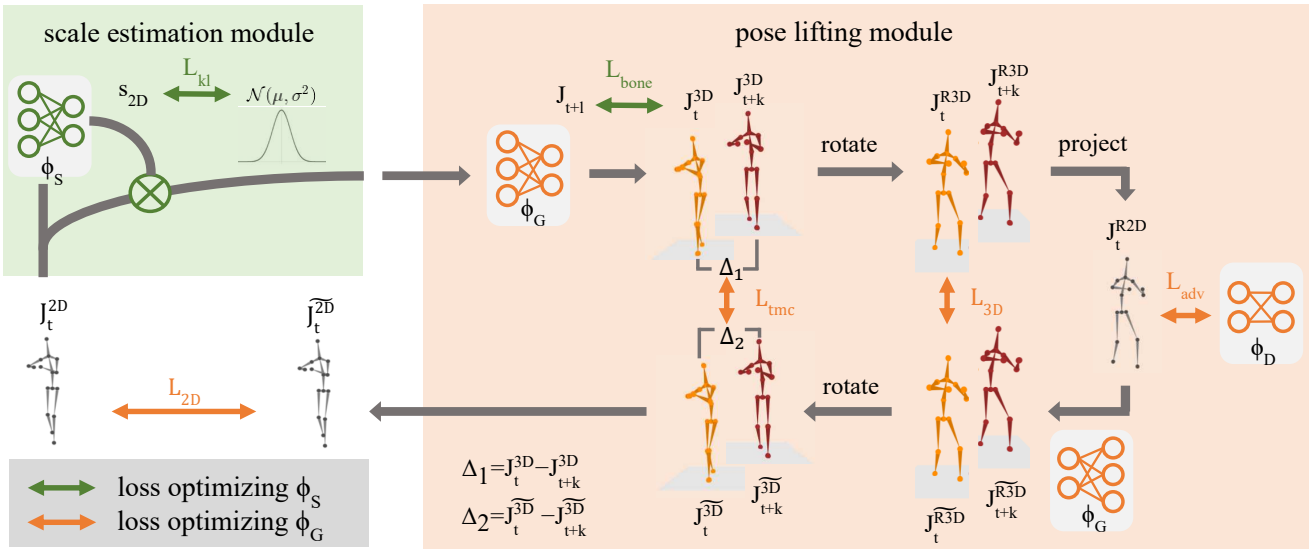

Figure 1: The detailed architecture of proposed framework. 3D pose estimation task is split into two parts, i.e., scale estimation module (green box) and pose lifting module (orange box). Pose lifting module contains two lifting networks and one discriminator. The lifting network takes the scaled 2D pose as input and outputs estimated 3D skeleton. Random projection of generated 3D skeleton goes through the lifting network, inverse transformation, and re-projection process again, allowing the network to self-supervise the training process by exploiting temporal geometric consistency. It should be noted that the scale estimation module and the pose lifting modules are trained iteratively in an end-to-end manner.

straints are incorporated into these two sub-tasks respectively: **(a) Temporal scale consistency**. Scale estimation module is used to optimize the scale of 3D pose. According to perspective projection [3], optimizing the scale of 3D skeletons equals to changing the scale of 2D pose and we experimentally find that constraining 2D scale is slightly better than 3D counterpart. Firstly, we propose a distribution constraint to coarsely adjust the 2D scale at the video level. Secondly, a bone constraint is proposed to optimize the scale of 2D pose at the frame level. Both aspects of constraints are seamlessly integrated into the scale estimation module, which effectively overcomes the scale ambiguity problem along the temporal direction. Pose lifting module then takes optimized 2D pose as input, where the scale estimation module helps reduce the learning difficulty of pose lifting and improve the estimation accuracy on in-the-wild data [24] by a large margin. **(b) Multi-view motion consistency**. Multi-view data [5], even synthesised from imagery [5], has shown its efficiency on single-frame 3D pose estimation. Inspired by the above operation, we propose a simple yet effective temporal constraint, which naturally generalizes the single-frame multi-view constraint to the video data (i.e., the motion trajectories across different views are encouraged to match with each other). To enhance the training stability, we propose an iterative training strategy, which achieves promising performance.

Extensive experiments demonstrate that our model achieves the state-of-the-art performance on two widely used 3D human motion datasets. Results on the Human3.6M [11] dataset for 3D human pose estimation exhibit that our approach improves upon the previous unsupervised methods by 23.1% and also outperforms several weakly supervised approaches that explicitly use 3D annotations. We also conduct detailed ablation studies to demonstrate the contribution of each component of the proposed framework.

2. Related Work

Fully Supervised pose estimation. Catalin et.al. [11] firstly provides a large-scale indoor dataset, i.e., Human3.6M [11], for 3D poses estimation with densely annotated joints. Facilitated by strong supervision signal, Julieta et.al. [22] proposes to utilize a simple yet effective model to predict the location of 3D joints with 2D key-points as inputs. This leads to a promising way for 3D pose estimation, i.e, based on the accurately estimated 2D joints. With access to the skeleton topology of human subject, structured constraint [18] is incorporated into the training procedure for better estimation of 3D key-points. Following the routine of the above work, Sun et.al. [30] decomposes the 3D coordinates of skeleton joints in a parameterized way. Meanwhile, a compositional loss function leveraging the temporal relationship in the pose is proposed to better modeling spatial-temporal structure of the human subjects. Noguer et.al. [25] considers the monocular 3D pose estimation task as a regression problem between matrices represented with

3D and 2D key-point distances. Also facilitated by structural property of the skeleton and correlations among key points, such representations facilitate reducing the inherent ambiguity of this task. Wang et.al. [34] propose a two-stage framework to facilitate the 3D pose estimation problem. A depth ranking algorithm is involved to fully utilize 2D and 3D key-point information within images. Pavllo et.al. [27] extends single-frame pose estimation to a video sequence, which effectively utilizes the temporal information and also could be trained in a semi-supervised way.

Weakly Supervised pose estimation. The acquisition of 3D key-point annotations is generally high cost, which motivates researches towards weakly supervised 3D pose estimation. Brau et.al. [1] proposes a deep model involving an additional output layer that maps predicted 3D key-points onto 2D image plane and body part lengths in 3D space is constrained to match with a prior distribution from dataset. Geometry-aware representation is proposed by Chen et.al. [5] to learn the 3D key-point relationship in a weakly-supervised manner. Li et.al. [20] utilizes the statistic property of 3D poses, which should be low rank and temporally smooth, to optimize the 3D trajectory of video sequences. Predicted results are treated as pseudo-annotations to achieve a more robust 3D pose training procedure. Fang et.al. [9] propose a so-called pose grammar framework, which explicitly leverages a set of domain information in terms of human body structure (i.e., kinematics, symmetry, motor coordination) As important prior knowledge, the geometric configuration of the human subject is further utilized in the work of Zhou et.al. [40] to combine spatial-temporal information to model the 3D geometry and account for the ambiguities resulting from the pose estimation model. Yang et.al. [38] shows that adversarial training is an effective tool to facilitate realistic 3D pose estimation. Boosted by the discriminator, predicted 3D poses should be visually natural and valid, i.e., following the human skeleton topology.

Unsupervised pose estimation. In contrast to fully supervised and weakly supervised pose estimation, unsupervised setting does not allow usage of any ground truth 3D pose information or relevant projection, which is much more challenging. Rhodin et.al. [29] firstly propose to learn a geometry-aware skeleton structure constructed via multi-view images and without any 3D labels. The image viewpoint along with 3D geometry information is predicted to facilitate the unsupervised location of 3D key-points. The multi-view information is an effective guidance signal for learning geometry-aware representation. Another work [3] bypasses the reliance of multi-view inputs and supervision signal is provided by rotation consistency in the 3D space. Further boosted by the generative adversarial training incorporated on the recovered 2D joints, the predicted 3D key-points are indirectly optimized to be more realistic and accurate. Kundu et.al. [17] estimate 3D skeletons relying on a minimal set of prior knowledge that defines the underlying kinematic 3D structure, such as skeletal joint connectivity information with bone-length ratios in a fixed canonical scale. Kim et.al. [15] propose to learn the 2D key-points in an unsupervised way via background/foreground disentangling, which is theoretically extensible to an arbitrary object. Similarly, based on background / foreground disentangling, [17] utilize encoder and decoder module by paired 2D images in a self-supervised manner.

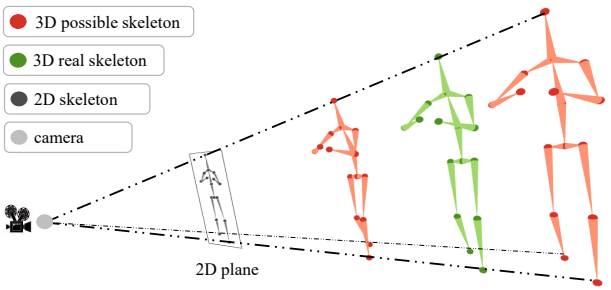

Figure 2: Examples of predicted results caused by scale ambiguity. Red and green 3D skeletons represent the estimated 3D skeletons with different scales.

3. Methods

In this section, we propose a unified temporal framework to effectively lift a 2D pose to the 3D skeleton, where the temporal scale consistency and multi-view motion consistency are combined into a pose estimation model. The overall framework is illustrated in Fig. 1, where the proposed model consists of two main parts, i.e., scale estimation module and pose lifting module. Given a monocular video sequence with a length of T time stamps, we first apply a pre-trained 2D pose detector (e.g., CPN [6]) to obtain crude 2D joints as initial inputs. Then the scale estimation module is used to optimize the scale of crude 2D pose. After that, we estimate the 3D skeletons with the refined 2D poses through the pose lifting module. More detailed descriptions are given as follows.

3.1. Optimizing 2D Pose via Scale Estimation

We denote initial 2D pose as $\mathbf{J}^{2D} = \{(x_i, y_i)\}_{i=1}^{N}$, and the corresponding 3D pose for each 2D joint can be represented as $\mathbf{J}^{3D} = \{(X_i, Y_i, Z_i)\}_{i=1}^{N}$. Here N indicates the number of joints for single person, and i refers to the ith joint. For a video sequence, the 2D and 3D coordinates of the ith joint at time stamp t are denoted as $\mathbf{J}_{i,t}^{2D}$ and $\mathbf{J}_{i,t}^{3D}$ respectively. For each time stamp, projected 2D joints and 3D joints should obey perspective projection [36] as follows:

$$x_i = \frac{X_i}{Z_i} f_x + c_x, y_i = \frac{Y_i}{Z_i} f_y + c_y, \quad (1)$$

where $\mathbf{f} = [f_x, f_y]$ and $\mathbf{c} = [c_x, c_y]$ are focal length and point respectively. Note that $Z_i = D + d_i$, where D indicates the absolute depth of the root joint of the person and d_i is the depth offset of ith joint relative to the root joint.

Limitation: Scale ambiguity. Previous unsupervised 3D pose estimation [29, 3, 12, 17] can not effectively estimate the size of the output 3D skeletons $\mathbf{J}^{3D}$ due to the lack of 3D annotations. As illustrated in Fig. 2, without 3D supervision, we can not obtain the absolute depth of the person in the camera coordinate, which results in ambiguous scale when lifting 2D pose to 3D pose.

Solution: Scale estimation via temporal constraint. Following [3], we assume a camera with unit focal length centered at the origin $(0, 0, 0)$ and fix the distance of the skeleton to the camera to a constant D units and normalize the 2D joints such that the mean distance from the head joint to the root joint is $\frac{1}{D}$ units in 2D. This ensures that 3D skeleton will be generated with a scale of ≈ 1 unit (head to root joint distance). Since $D \gg d_i$ and $Z_i = D + d_i$, we have $Z_i \approx D$, where the prospective projection can be approximated as follows:

$$x_i = \frac{X_i}{D} \cdot f_x. \tag{2}$$

We can see that x_i is proportional to X_i (i.e., $x_i \sim X_i$). Scale estimation module is firstly utilized to infer the scale of 3D pose S_{3D}, i.e., $x_i = S_{3D} \cdot X_i$. It can be alternatively written as $S_{2D} \cdot x_i = X_i$, where S_{2D} is the scale of 2D pose. We can see that limiting the scale of 3D pose or optimizing the scale of 2D pose S_{2D} has the similar effect on this task, and we experimentally find better results if estimating the scale of 2D pose.

If 3D poses are given, we are able to calculate the scale of the projected 2D joints (S_{2D}^{ref}) according to Eqn. 1 as follows:

$$S_{2D}^{ref} = \frac{1}{2}[\frac{\mathcal{H}(\{X_i/Z_i\}_{i=1}^{N})}{\mathcal{H}(\{x_i\}_{i=1}^{N})} + \frac{\mathcal{H}(\{Y_i/Z_i\}_{i=1}^{N})}{\mathcal{H}(\{y_i\}_{i=1}^{N})}], \tag{3}$$

where $\mathcal{H}(x) = \max(x) - \min(x)$. And $\max(\cdot)$ and $\min(\cdot)$ operation mean the maximum and minimum value of corresponding 2D joints. We experimentally find considerable improvement if feeding in S_{2D}^{ref} as input, which drives us to capture the underlying value of S_{2D}^{ref} (modelled with S_{2D}) with the temporal cues in a supervised manner.

S_{2D}^{ref} of different video sequences has been shown in Fig. 3(A / B). However, We can see that the value of S_{2D}^{ref} fluctuates across the specific value in the whole video irregularly. And two curves of corresponding video sequences in Fig. 3(A / B) have a different pattern. Thus it is highly difficult to learn S_{2D} only using monocular information without the supervision of S_{2D}^{ref}. We are thus motivated to utilize the temporal information to learn the variation of S_{2D}^{ref}.

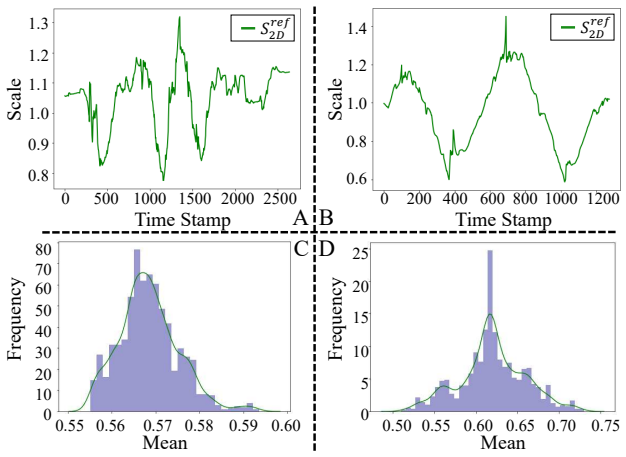

Figure 3: Analysis on the scale of the projected 2D joints (S_{2D}^{ref}). A/B corresponds to S_{2D}^{ref} from two different video sequences. C/D indicates two corresponding frequency histogram of S_{2D}^{ref}.

Temporal Scale Consistency. Fig. 3(C and D) indicate two distributions of frequency histogram of S_{2D}^{ref} in different video sequences. It can be seen that the distribution of frequency histogram of S_{2D}^{ref} is uni-modal, i.e., following $\mathcal{N}(\mu, \sigma)$. We thus model the distribution of S_{2D} with a parameterized gaussian distribution, i.e., $\mathcal{N}(\hat{\mu}, \hat{\sigma})$, where $\hat{\mu}$ and $\hat{\sigma}$ are learnable parameters. Our goal is to encourage the distribution of estimated S_{2D} to approximate that of S_{2D}^{ref}. We use the Kullback-Leibler divergence to optimize the learnable parameters $\hat{\mu}$ and $\hat{\sigma}$ as follows:

$$\begin{aligned} \mathcal{L}_{kl} &= D_{KL}(\mathcal{N}(\hat{\mu}, \hat{\sigma}) || \mathcal{N}(\mu, \sigma)) \\ &= \log \frac{\sigma}{\hat{\sigma}} + \frac{\hat{\sigma}^2 + (\mu - \hat{\mu})^2}{2\sigma^2} - \frac{1}{2}, \end{aligned} \tag{4}$$

$\mathcal{L}_{kl}$ is able to constrain the distribution (i.e., mean and variance) of all sequences, but can not constrain the distribution of each sequence, not to mention more detailed information of single frame. To achieve more accurate results, we introduce bone consistency loss $\mathcal{B}$, which is defined as follows:

$$\mathcal{L}_{bone} = \|\mathcal{B}(\mathbf{J}_t^{3D}) - \mathcal{B}(\mathbf{J}_{t+l}^{3D})\|^2, \tag{5}$$

where $\mathcal{B}$ denotes the bone length of the 3D skeleton. $\mathbf{J}_t^{3D}$ and $\mathbf{J}_{t+l}^{3D}$ represent estimated 3D poses at time stamp t and $t + l$.

In summary, the first part loss function (Temporal Scale Consistency loss, $\mathcal{L}_{tsc}$) can be represented as:

$$\mathcal{L}_{tsc} = w_{kl}\mathcal{L}_{kl} + w_{bone}\mathcal{L}_{bone}, \tag{6}$$

where w_{kl} and w_{bone} are hyper-parameters. We would like to emphasize that $\mathcal{L}_{kl}$ constrains the general range (i.e., distribution) of S_{2D}, more accurate value at each time stamp will be achieved with the second regularization term.

3.2. Lifting Optimized 2D Pose to 3D Counterpart

2D-to-3D Pose Lifting Model. We adopt the work of Chen et.al. [3] as the baseline model, which outputs 3D poses as follows:

$$\mathbf{J}_i^{3D} = (x_i Z_i, y_i Z_i, Z_i), Z_i = max(1, D + d_i). \quad (7)$$

Specifically, we estimate the depth d_i for the ith joint and obtain the final 3D poses through Eqn. 7. As shown in Fig. 1, the 2D-to-3D pose lifting model contains two parts, i.e., lifting network Φ_G and discriminator Φ_D. Compared with [3], we further optimize the baseline model in terms of adversarial loss, network architecture and training strategy. Please refer to the supplementary materials for more details.

Limitation: Pose ambiguity. Lifting 2D poses to 3D counterparts is inherently ambiguous [17]. Given a specific 2D pose, there might exist multiple reasonable 3D poses matching with the 2D input. Li et.al. [17] indicate that multi-view data is able to effectively alleviate such kind of ambiguity. However, multi-camera equipment is not accessible for monocular 3D pose estimation [12].

Solution: Multi-view Motion Constraint. To alleviate pose ambiguity, we produce another (pseudo) view of 3D trajectories via a geometric random rotation [5] scheme to construct multi-view information. This process denoted by $\mathcal{V}$ can be shown as follows:

$$\mathbf{J}_t^{R3D} = \mathcal{V}(\mathbf{J}_t^{3D}) = R * \left(\mathbf{J}_t^{3D} - \mathbf{J}_{0,t}^{3D}\right) + T, \quad (8)$$

where $T = [0, 0, D]$ is the translation vector and R is the rotation matrix. As is illustrated in Fig. 1, $\mathbf{J}_t^{R2D}$ can be obtained by camera projection from $\mathbf{J}_t^{R3D}$. Then $\mathbf{J}_t^{R2D}$ is sent to lifting network Φ_G to get $\tilde{\mathbf{J}}_t^{R3D}$. $\tilde{\mathbf{J}}_t^{R3D}$ is transformed to $\tilde{\mathbf{J}}_t^{3D}$ by applying the inverse of rigid transformation $\mathcal{V}$. The 3D skeleton $\tilde{\mathbf{J}}_t^{3D}$ is finally projected to the 2D pose $\tilde{\mathbf{J}}_t^{2D}$. More details will be shown in supplementary materials. Generated multi-view motion is utilized to pursue a unique 3D structure. Our goal is to keep the pose difference between two random frames from two different views as close as possible.

Multi-view Motion Consistency. As shown in Fig. 1, our approach takes two frames from one sequence as inputs for training. We enforce the temporal consistency between different views via the following loss $\mathcal{L}_{tmc}$ to refine the lifting network,

$$\mathcal{L}_{tmc} = \|(\mathbf{J}_t^{3D} - \mathbf{J}_{t+k}^{3D}) - (\tilde{\mathbf{J}}_t^{3D} - \tilde{\mathbf{J}}_{t+k}^{3D})\|^2, \quad (9)$$

where $\mathbf{J}_t^{3D}$ and $\mathbf{J}_{t+k}^{3D}$ represent estimated 3D skeletons in frame t and $t + k$. $\tilde{\mathbf{J}}_t^{3D}$ and $\tilde{\mathbf{J}}_{t+k}^{3D}$ mean transformed 3D skeletons in frame t and $t + k$. Transformed 3D skeletons can be seen via a camera from another view. Therefore, the proposed temporal motion consistency loss $\mathcal{L}_{tmc}$ is formed by cross-view motion constraint, which is to tackle the lifting ambiguity to pursue a more reasonable 3D structure. In summary, the second part loss function (Pose lifting loss, $\mathcal{L}_{lifter}$) can be represented as :

$$\mathcal{L}_{lifter} = w_{2D}\mathcal{L}_{2D} + w_{3D}\mathcal{L}_{3D} + w_{adv}\mathcal{L}_{adv} + w_{tmc}\mathcal{L}_{tmc}, \quad (10)$$

where w_{2D}, w_{3D}, w_{adv} and w_{tmc} are hyper-parameters. Formally, the loss function $\mathcal{L}_{2D}$, $\mathcal{L}_{3D}$ and $\mathcal{L}_{adv}$ are defined as:

$$\mathcal{L}_{2D} = \|\mathbf{J}_t^{2D} - \tilde{\mathbf{J}}_t^{2D}\|^2, \mathcal{L}_{3D} = \|\mathbf{J}_t^{3D} - \tilde{\mathbf{J}}_t^{3D}\|^2, \quad (11)$$

$$\min_{\theta_D} \mathcal{L}_{adv}(\Phi_D) = \mathbb{E}[\Phi_D(\mathbf{J}_t^{2D}) - 1]^2 + \mathbb{E}[\Phi_D(\mathbf{J}_t^{R2D})]^2, \quad (12)$$

$$\min_{\theta_G, \theta_S} \mathcal{L}_{adv}(\Phi_G) = \mathbb{E}[\Phi_D(\mathbf{J}_t^{R2D}) - 1]^2, \quad (13)$$

where θ_S and θ_G are the parameters of the scale estimation module Φ_S and the pose lifting module Φ_G (as illustrated in Fig. 1), and θ_D is the parameters of discriminator.

It is worth mentioning that $\mathcal{L}_{tsc}$ and $\mathcal{L}_{lifter}$ constrain two parts (i.e., scale estimation module and pose lifting module) separately, which is split by 3D pose estimation. Different from traditional loss constraints, these two losses constrain above two parts respectively, which shows excellent performance compared to traditional training strategy.

Note that [3] uses an extra temporal discriminator to learn the temporal consistency, which is computational inefficiency with marginal improvement. By contrast, the proposed constraint only utilizes multi-view motion information to learn temporal consistency instead of relying on discriminator, which still achieves considerably better results.

3.3. Iterative Training Strategy

In this work, we split 3d pose estimation into two subtasks by the scale estimation module and the pose lifting module. In such serial design, the input of the pose lifting module depends on the output of the scale estimation module, which may suffer from large variance in input distribution. Therefore, to stabilize the optimization process, we train the scale estimation module and the pose lifting module iteratively (e.g. we train the pose lifting module 4 times and then train the scale estimation module once.) When training the pose lifting module, we freeze the weights of the scale estimation module, and vice versa. We empirically find that such iterative practice is effective and the ablation study can be found in Sec. 4.4.

3.4. Implementation Details

Our networks are generally shallow and can be trained in an end-to-end manner efficiently. Following [23], we use the residual block as the building block. Specifically, we

use 5 blocks for the pose lifting module, 2 blocks for the discriminator and 1 block for the scale estimation module. Please refer to the supplementary materials for more details.

As in earlier discussion, the loss functions for optimizing the scale estimation module and the pose lifting module are given by $\mathcal{L}_{tsc}$ and $\mathcal{L}_{lifter}$. where $w_{kl} = 0.001$, $w_{bone} = 1.0$ $w_{2D} = 0.5$, $w_{3D} = 5.0$, $w_{adv} = 1.0$ and $w_{tmc} = 1.0$ respectively. For the KL prior, the parameters of target Gaussian distribution(μ, σ) need to be predefined. We experiment with different pairs selected from ranges $(0.55, 0.75)$ and $(0.05, 0.15)$ and choose a best pair($\mu \approx 0.71$, $\sigma \approx 0.06$). As for other hyper-parameters, we set constant depth $D = 10$, set the batch size to 1024 and the learning rate of both lifter and discriminator to 0.0002 with decay rate of 0.95 per epoch. The dropout rate is set to 0.25. We adopt Adam optimizer with default parameters and train the whole network for 200 epochs.

4. Experiments

4.1. Datasets and Metrics

Human3.6M [11]. Human3.6M is one of the largest indoor datasets with Mosh [21] available. We report mean per-joint position error (MPJPE) and PMPJPE (MPJPE after rigid alignment).

MPI-INF-3DHP [24]. MPI-INF-3DHP is collected both indoors and outdoors. In addition to PMPJPE, we report the Percentage of Correct Keypoints (PCK) thresholded at 150mm and the Area Under the Curve (AUC).

Surreal [32]. Surreal contains many video clips with human characters of various shapes and poses.

LSP [13]. LSP consists of 2000 in-the-wild images without ground-truth 3D annotation. We perform qualitative evaluation to illustrate the generalization ability.

4.2. Quantitative Evaluation

Results on Human3.6M Dataset [11]. As illustrated in Tab. 1, we report the unsupervised pose estimation results in terms of MPJPE and PMPJPE. We show results from fully supervised (Full), weakly supervised (Weak) and unsupervised (Unsup) methods. Our method outperforms the state-of-the-art unsupervised method (Kundu et.al. [16]) by a significant margin (52.3 vs.62.4) in terms of PMPJPE. This is mainly facilitated by multi-view motion consistency, which provide more accuracy and reasonable pose. Moreover, our method surpasses Rhodin et.al. [29] by 29.8 % in MPJPE, which is possibly boosted by temporal scale consistency. Notably, our method is comparable with several weakly supervised approaches that explicitly use 3D data.

Results on MPI-INF-3DHP [24]. As shown in Tab. 2, we present the pose estimation results in terms of PCK and AUC. For more comprehensive comparison, we also report the performance of several recent fully-supervised and weakly-supervised methods trained on various datasets. Among these models trained on 3DHP [24] dataset, our approach achieves higher accuracy than both unsupervised and weakly-supervised methods. Similarly, among the models trained on Human3.6M [11] dataset, our approach outperforms all the unsupervised / weakly-supervised methods and is comparable with the performance of fully-supervised models, which demonstrates the generalization ability of our model. By explicitly incorporating temporal scale consistency and multi-view motion consistency into deep models, Ambiguity problems have been well alleviated and output structure is more reasonable, which finally leads to higher estimation accuracy.

Mode	Algorithm	GT		PRE	
		MPJPE	PMPJPE	MPJPE	PMPJPE
Full	Martinez et al. [22]	45.5	37.1	62.9	47.7
	Pavllo et al. [27]	37.2	27.2	46.8	36.5
	Wang et al. [33]	-	-	42.6	32.7
Weak	3DInterpreter [35]	-	88.6	-	98.4
	AIGN [31]	-	79.0	-	97.4
	Drover et al. [8]	-	38.2	-	64.6
	Li et al. [20]	-	-	88.8	66.5
Unsup	Rhodin et.al.[29]	-	-	131.7	98.2
	Chen et.al.[3]	-	51.0	-	68.0
	Kundu et al.[16]	-	-	-	62.4
	Kundu et al.[17]	-	-	-	63.8
	Ours	**85.3**	**42.0**	**92.4**	**52.3**

Table 1: Experimental results on the test set of Human3.6M [11]. For a more comprehensive comparison, we list results from several kinds of supervision. GT and PRE denote results using ground truth 2D pose and estimated 2D pose by 2D detector, respectively.

4.3. Qualitative Evaluation

We present qualitative results on Human3.6M [11], 3DHP [24], Surreal [32] and LSP [13] datasets. Note that to demonstrate the generalization ability of the proposed model, the pose on LSP [13] dataset is estimated with the model trained on Human3.6M [11] dataset. As illustrated in Fig. 4, we visualize predicted skeletons (green) and the ground truth (red) in the same coordinate system. Note that scaling and rigid alignment are NOT performed. We can see that the scale of 3D poses are well estimated (Fig. 4A,B,C) on Human3.6M [11], 3DHP [24], and Surreal [32] datasets, which is visually reasonable and mainly facilitated with the temporal scale constraint. Moreover, the estimation results on the unseen appearance (Fig. 4D) are still visually satisfying. The qualitative results have shown that scale ambiguity and pose ambiguity are properly handled by our proposed method. For comparison, we also provide more visual re-

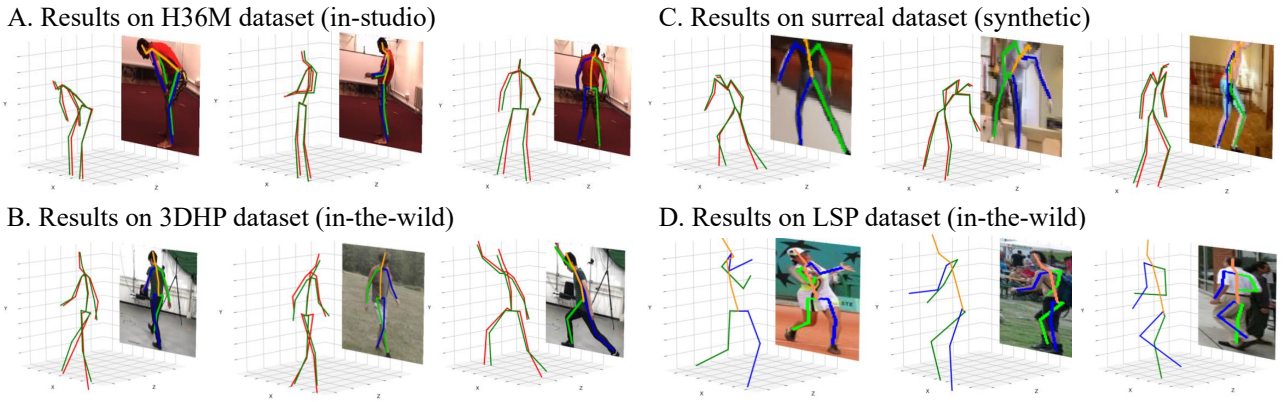

Figure 4: Qualitative results on 4 different datasets. **Top Left:** predictions(red) along with ground truth(green) in Human3.6M [11]. **Top Right:** predicted predictions(red) along with ground truth(green) in 3DHP. **Bottom Left:** predicted predictions(red) along with ground truth(green) in Surreal. **Bottom Right:** results without training on the corresponding train-set.

Supervision	Algorithm	Trainset	PCK	AUC
Full	Mehta et al.[24]	H36M	64.7	31.7
	Mehta et al.[24]	3DHP	72.5	36.9
	Yu et al.[7]	H36M	84.1	-
	Zeng et al.[39]	H36M	77.6	43.8
	Chen et al.[4]	3DHP	87.9	54.0
	Wang et al.[33]	3DHP	86.9	62.1
Weak	Zhou et al.[24]	H36M	69.2	32.5
	Kanazawa et al.[14]	3DHP	77.1	40.7
Unsup	Chen et al.[3]	H36M	64.3	31.6
	Chen et al.[3]	3DHP	71.1	36.3
	Kundu et al.[16]	YTube+H36M	84.6	**60.8**
	Kundu et al.[17]	H36M+3DHP	80.2	44.8
	Ours	H36M	82.2	46.6
	Ours	3DHP	**86.2**	51.7

Table 2: Experimental results on the test set of MPI-INF-3DHP [24]. When **Trainset** is h36m, we simply use pre-trained model on Human3.6M [11] to evaluate on MPI-INF-3DHP [24] without any fine-tuning.

sults of estimated poses with and without scale module respectively. Please refer to the supplementary material.

4.4. Ablation Study

Analysis on loss configuration and supervision signals. In Tab. 3 we report the performance under different loss configurations on the Human3.6M [11] data. Without access to the source code, we re-implement the baseline model [3]. Experimental results show that our implementation is much better than the original one in [3]. We present the differences in detail in supplementary material and please also refer to the released code. Specifically, by simply exploiting geometric information and utilizing discriminator to prevent irrational poses, we achieved satisfying results. Furthermore, based on the baseline, we can

Loss Function	MPJPE	PMPJPE
$\mathcal{L}_{2D}$+$\mathcal{L}_{3D}$+$\mathcal{L}_{adv}$ in [3]	-	58.0
$\mathcal{L}_{2D}$+$\mathcal{L}_{3D}$+$\mathcal{L}_{adv}$	105.0	46.0
$\mathcal{L}_{2D}$+$\mathcal{L}_{3D}$+$\mathcal{L}_{adv}$+$\mathcal{L}_{tmc}$	101.7	43.5
$\mathcal{L}_{2D}$+$\mathcal{L}_{3D}$+$\mathcal{L}_{adv}$+$\mathcal{L}_{tmc}$+$\mathcal{L}_{kl}$	96.0	42.9
$\mathcal{L}_{2D}$+$\mathcal{L}_{3D}$+$\mathcal{L}_{adv}$+$\mathcal{L}_{tmc}$+$\mathcal{L}_{kl}$ +$\mathcal{L}_{bone}$	**85.3**	**42.0**

Table 3: The analysis on different loss configurations. Performance is evaluated on test set of Human3.6M [11] and we use **GT** 2d as the input. We re-implement the baseline according to [3] and list the differences in supplementary material in detail.

Ablation study on scale estimation	MPJPE	PMPJPE
*Reference scale as inputs	72.3	39.7

2D Normalization	Scale Module	MPJPE	PMPJPE
None	None	260.7	57.8
None	On 2D	233.4	55.0
Universal scale	None	105.0	46.0
Step-wise scale	None	94.5	48.5
Universal scale	On 3D	97.5	43.6
Universal scale	On 2D	**85.3**	**42.0**

Table 4: The analysis on methods to address scale ambiguity. * can be viewed as the ceiling performance of our design. Universal scale normalization indicates we normalise all 2D skeletons with a same constant. Step-wise scale indicates we normalise each 2D skeleton independently. We evaluate the performance on test set of Human3.6M [11].

observe that adding temporal motion consistency $\mathcal{L}_{tmc}$ can boost the performance by about 6%, proving the effectiveness of the temporal constraint in our model. Then we per-

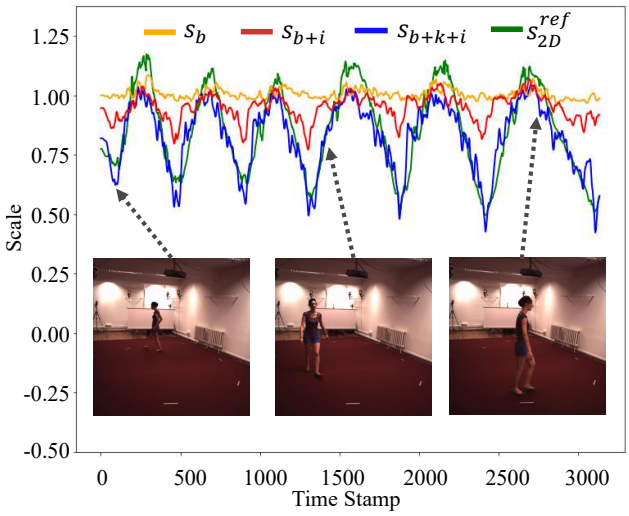

Figure 5: Illustration of scales and scale predictions. s_{2D}^{ref} (green line) denotes target scales computed manually. s_b (yellow line) denotes using only $\mathcal{L}_{bone}$ during training. s_{b+c} (red line) denotes $\mathcal{L}_{bone}$ with iterative training. s_{b+k+c} (blue line) denotes $\mathcal{L}_{bone} + \mathcal{L}_{kl}$ with iterative training.

form ablation studies on the scale estimation module. We observe that applying $\mathcal{L}_{kl}$ prior alone leads to a slight drop $\approx$ 1.3% in PMPJPE. This is because $\mathcal{L}_{kl}$ is a relatively weak constraint, which only constrains the range of estimated scales instead of pose accuracy. Besides, in terms of MPJPE, the scale estimation module brings about 16.1% improvement compared with the baseline.

Finally, we visualize the scale information predicted by different components of the proposed method in Fig. 6. It can be easily seen that s_b (yellow line) is closed to a constant value 1, which means little scale information can be learned only relying on $\mathcal{L}_{bone}$. s_{b+i} (red line) has obvious fluctuations, proving the effectiveness of the iterative training strategy. The predicted scale s_{b+k+i} based on $\mathcal{L}_{bone} + \mathcal{L}_{kl}$ with iterative training is highly close to s_{2D}^{ref}, which is computed by 3D information. Note that we learn the scale information without the supervision of s_{2D}^{ref} during the training procedure. The excellent results show that our method is able to learn the scale information well.

Analysis on methods to address scale ambiguity. As shown in Tab. 4, we first feed reference scales (calculated with ground truth) as inputs and evaluate on the test set of Human3.6M [11], which can be seen as the upper-bound performance of our framework (the first row). Then, without any pre-processing procedure, we report the performance with and without scale module. We can observe that scale module can boost the performance still, proving the effectiveness of our design. Then we experiment with some other pre-processing techniques. A simple alternative solution is to use step-size scale in normalization(i.e. normalise the head-root distance of all skeletons to $1/D$). Step-wise scale leads to a drop in PMPJPE and we speculate the reason is that information about the real inputs distribution can not be preserved. In contrast, using scale module to address this problem is much more effective. Finally, we try to multiply scales directly on 3d outputs of the pose lifting module, which does not work as well as on 2d counterparts.

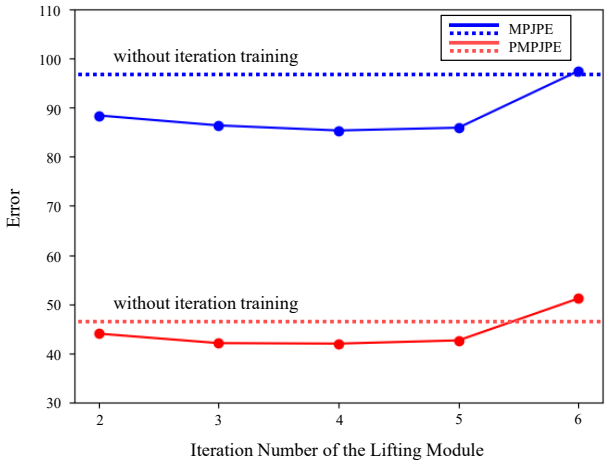

Figure 6: Illustration of the effects of training strategy on model performance. Iteration numbers of the lifting module indicate how many times we train the lifting module when scale module is trained once. We show MPJPE, PMPJPE w.r.t. different ratios, as well as in setting without iterative training, which is shown in dash lines.

Analysis on training strategy. Intuitively, scales will dramatically affect the distribution of the input 2d joints of the lifting module, so it is important to employ training methods properly, otherwise, the training will be quite unstable. As is illustrated in Fig. 6, we compare different training methods and report their performance. When we remove iterative training, we simply combine Eqn. 10 and Eqn. 6, optimizing scale estimation module and pose lifting module together. We experiment with several ratios in iterative training and empirically choose the best one.

5. Conclusions

In this paper, our method splits unsupervised monocular 3D pose estimation into two sub-tasks, scale estimation module and pose lifting module. Both two modules are optimized via an iterative training scheme with the corresponding temporal constraints. Extensive experiments show that our model achieves state-of-the-art performance on related human pose estimation datasets.

Acknowledgment This work was supported by National Science Foundation of China (U20B2072,61976137). The authors would like to give a personal thanks to the Student Innovation Center of SJTU for providing GPUs.

References

[1] Ernesto Brau and Hao Jiang. 3d human pose estimation via deep learning from 2d annotations. In *3DV*, pages 582–591, 2016. 3

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, pages 1302–1310. 1

[3] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Rohith MV, Stefan Stojanov, and James M. Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *CVPR*, pages 5714–5724, 2019. 1, 2, 3, 4, 5, 6, 7

[4] Tianlang Chen, Chen Fang, Xiaohui Shen, Yiheng Zhu, Zhili Chen, and Jiebo Luo. Anatomy-aware 3d human pose estimation in videos. *CoRR*, abs/2002.10322, 2020. 7

[5] Xipeng Chen, Kwan-Yee Lin, Wentao Liu, Chen Qian, and Liang Lin. Weakly-supervised discovery of geometry-aware representation for 3d human pose estimation. In *CVPR*, pages 10895–10904, 2019. 2, 3, 5

[6] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, pages 7103–7112, 2018. 1, 3

[7] Yu Cheng, Bo Yang, Bo Wang, and Robby T. Tan. 3d human pose estimation using spatio-temporal networks with explicit occlusion training. In *AAAI*, pages 10631–10638. 7

[8] Dylan Drover, M. V. Rohith, Ching-Hang Chen, Amit Agrawal, Ambrish Tyagi, and Cong Phuoc Huynh. Can 3d pose be learned from 2d projections alone? In *ECCV*, pages 78–94, 2018. 6

[9] Haoshu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. Learning pose grammar to encode human body configuration for 3d pose estimation. In *AAAI*, pages 6821–6828, 2018. 3

[10] David C. Hogg. Model-based vision: a program to see a walking person. *IVC*, 1(1):5–20, 1983. 1

[11] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, 2014. 2, 6, 7, 8

[12] Umar Iqbal, Pavlo Molchanov, and Jan Kautz. Weakly-supervised 3d human pose learning via multi-view images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 4, 5

[13] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, 2010. doi:10.5244/C.24.12. 6

[14] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2018. 1, 7

[15] Yunji Kim, Seonghyeon Nam, In Cho, and Seon Joo Kim. Unsupervised keypoint learning for guiding class-conditional video prediction. In *NeurIPS*, pages 3809–3819, 2019. 3

[16] Jogendra Nath Kundu, Siddharth Seth, Varun Jampani, Mugalodi Rakesh, R. Venkatesh Babu, and Anirban Chakraborty. Self-supervised 3d human pose estimation via part guided novel image synthesis. In *CVPR*, pages 6151–6161, 2020. 6, 7

[17] Jogendra Nath Kundu, Siddharth Seth, Rahul M. V., Mugalodi Rakesh, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Kinematic-structure-preserved representation for unsupervised 3d human pose estimation. In *AAAI*, pages 11312–11319, 2020. 1, 3, 4, 5, 6, 7

[18] Sijin Li, Weichen Zhang, and Antoni B. Chan. Maximum-margin structured learning with deep networks for 3d human pose estimation. In *ICCV*, pages 2848–2856, 2015. 2

[19] Yang Li, Kan Li, Shuai Jiang, Ziyue Zhang, Congzhentao Huang, and Richard Yi Da Xu. Geometry-driven self-supervised method for 3d human pose estimation. In *AAAI2020*. 1

[20] Zhi Li, Xuan Wang, Fei Wang, and Peilin Jiang. On boosting single-frame 3d human pose estimation via monocular videos. In *ICCV*, pages 2192–2201, 2019. 1, 3, 6

[21] Matthew Loper, Naureen Mahmood, and Michael J. Black. Mosh: motion and shape capture from sparse markers. *ACM Trans. Graph.*, 33(6):220:1–220:13, 2014. 6

[22] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, pages 2659–2668, 2017. 2, 6

[23] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 5

[24] 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 *3DV*, pages 506–516, 2017. 2, 6, 7

[25] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *CVPR*, pages 1561–1570, 2017. 2

[26] Itay Mosafi, Eli (Omid) David, and Nathan S. Netanyahu. Deepmimic: Mentor-student unlabeled data based training. In *Artificial Neural Networks and Machine Learning - ICANN, pages = 440–455, year = 2019,*. 1

[27] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, pages 7753–7762, 2019. 3, 6

[28] Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. In *ICCV*, pages 1913–1921, 2015. 1

[29] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *ECCV*, pages 765–782, 2018. 1, 3, 4, 6

[30] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2621–2630, 2017. 2

[31] Hsiao-Yu Fish Tung, Adam W. Harley, William Seto, and Katerina Fragkiadaki. Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision. In *ICCV*, pages 4364–4372, 2017. 6

[32] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 6

[33] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion guided 3d pose estimation from videos. *CoRR*, abs/2004.13985, 2020. 6, 7

[34] Min Wang, Xipeng Chen, Wentao Liu, Chen Qian, Liang Lin, and Lizhuang Ma. Drpose3d: Depth ranking in 3d human pose estimation. In *IJCAI*, pages 978–984, 2018. 3

[35] Jiajun Wu, Tianfan Xue, Joseph J. Lim, Yuandong Tian, Joshua B. Tenenbaum, Antonio Torralba, and William T. Freeman. Single image 3d interpreter network. In *ECCV*, pages 365–382, 2016. 6

[36] Jingwei Xu, Zhenbo Yu, Bingbing Ni, Jiancheng Yang, Xiaokang Yang, and Wenjun Zhang. Deep kinematics analysis for monocular 3d human pose estimation. In *CVPR2020*, June 2020. 1, 3

[37] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Learning feature pyramids for human pose estimation. In *ICCV*, pages 1290–1299, 2017. 1

[38] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy S. J. Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *CVPR*, pages 5255–5264, 2018. 3

[39] Ailing Zeng, Xiao Sun, Fuyang Huang, Minhao Liu, Qiang Xu, and Stephen Lin. Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. *CoRR*, abs/2007.09389, 2020. 7

[40] Xiaowei Zhou, Menglong Zhu, Georgios Pavlakos, Spyridon Leonardos, Konstantinos G. Derpanis, and Kostas Daniilidis. Monocap: Monocular human motion capture using a CNN coupled with a geometric prior. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 901–914, 2019. 1, 3