Two Views Are Better than One: Monocular 3D Pose Estimation with Multiview Consistency

Christian Keilstrup Ingwersen[1,2] Rasmus Tirsgaard[1] Rasmus Nylander[2]
Janus Nørtoft Jensen[1] Anders Bjorholm Dahl[1]
Morten Rieger Hannemose[1]
[1] Visual Computing, Technical University of Denmark
[2] TrackMan A/S, Denmark
{cin, rany}@trackman.com, {rhti, abda, jnje, mohan}@dtu.dk

Abstract

Deducing a 3D human pose from a single 2D image is inherently challenging because multiple 3D poses can correspond to the same 2D representation. 3D data can resolve this pose ambiguity, but it is expensive to record and requires an intricate setup that is often restricted to controlled lab environments. We propose a method that improves the performance of deep learning-based monocular 3D human pose estimation models by using multiview data only during training, but not during inference. We introduce a novel loss function, consistency loss*, which operates on two synchronized views. This approach is simpler than previous models that require 3D ground truth or intrinsic and extrinsic camera parameters. Our consistency loss penalizes differences in two pose* sequences *after rigid alignment. We also demonstrate that our consistency loss substantially improves performance for fine-tuning without requiring 3D data. Furthermore, we show that using our consistency loss can yield state-of-the-art performance when training models from scratch in a semi-supervised manner. Our findings provide a simple way to capture new data,* e.g. *in a new domain. This data can be added using off-the-shelf cameras with no calibration requirements. We make all our code and data publicly available.*

1. Introduction

Inferring a 3D human pose from a single 2D image or 2D keypoints is an ill-posed problem, where the 3D pose is not uniquely defined. The consequence is that most existing models inaccurately predict the depth of the keypoints [11]. It is, however, possible to learn a consistent mapping from a single 2D image to a 3D pose [14, 55]. Previous methods train on 3D data [55] or on two or more views with a calibrated camera setup [14]. Such data is expensive and restricts the settings where data can be recorded. Instead, we use a much simpler setup with two synchronized stationary or moving cameras. Here, the 3D pose in each frame is identical up to a transformation, and in the stationary case, this transformation is the same for all the frames recorded by the two cameras. We estimate this transformation using the Procrustes algorithm and optimize for consistent 3D pose predictions during training. At inference, we use a single 2D image, and our approach leads to state-of-the-art performance. Fig. 1 sketches the concept of our consistency loss.

With this inherent ambiguity, models inferring 3D poses need a representation of how the body can move. It has become popular to rely on large foundation models [55], which have learned how the body moves in general based on supervised learning. Even so, foundation models do not always generalize to new movements in specific applications such as sports. This introduces the need for fine-tuning the foundation models to accommodate the less frequently seen movements specific to that domain [55], for which our model is particularly useful since collecting the data is inexpensive.

Methods for adapting 3D human pose models to different domains and movements have traditionally relied on the availability of new 3D data [18, 52, 53]. However, it can be costly and may not be feasible to set up systems for capturing 3D data in the new domain. Alternative methods have been explored to address these challenges. These methods have demonstrated that 3D pose models can be fine-tuned using 2D data, as suggested by previous work [2]. This fine-tuning process ensures that the inferred 3D joint positions align with 2D keypoints in an image, which can be obtained accurately using readily available methods [16, 49]. Unfortunately, state-of-the-art models based solely on 2D supervision from a single view have too inaccurate depth predic-

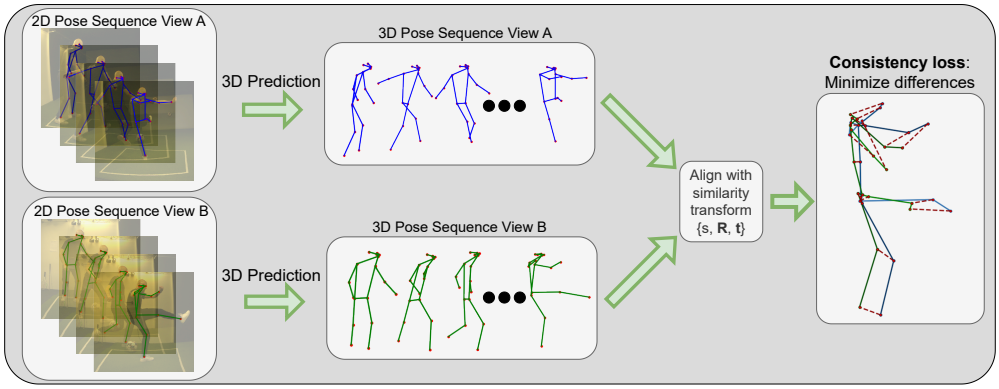


Figure 1. We improve monocular performance by applying our consistency loss during training to predicted 3D pose sequences from two different views. The consistency loss penalizes variations between the two predicted pose sequences of the same activity. Note that we only use multiple views during training. For every predicted 3D pose sequence obtained from View A and View B, we compute a similarity transform with Procrustes Analysis. This transformation aligns the predicted poses in Sequence A with Sequence B. The consistency loss is the average 3D distance between the two pose sequences post-alignment, shown as dashed red lines. Using Procrustes analysis for this transformation enables us to use cameras with unknown intrinsics and extrinsics.

tions to be used in actual applications such as sports [11]. Thus, there is a need to advance 3D pose prediction from 2D images.

While we focus on increasing the performance of monocular models and not models utilizing multiple views, many datasets used for training monocular models have multiple views of the scene available [12, 17, 31, 33]. We use the multiple views in the SportsPose [12], Human3.6M [13], and SkiPose [44] datasets to evaluate our approach. We use multiple views during training while only using a single view at inference. Additionally, we investigate how many views are necessary to obtain improvements in 3D predictions. Specifically, we use the SportsPose dataset [12] to investigate the performance using varying numbers of views. We have gained access to all seven views of this dataset, and the original authors have allowed us to release the additional data as part of this paper. The SportsPose dataset features complex and challenging sports scenarios, enabling us to test poses that are rare in other datasets. The new views are available on our website[1]. Additionally, we test our consistency loss on the Human3.6M dataset trained semi-supervised, resulting in superior performance. By utilizing our consistency loss, we obtain state-of-the-art among semi-supervised methods.

Our contributions extend beyond introducing the view-consistency loss for domain-adaptive 3D pose estimation. We also present the first set of baseline results on the SportsPose dataset, demonstrating the effectiveness of our approach. We illustrate how our method enhances 3D pose estimation accuracy in dynamic and complex environments by showcasing a model fine-tuned on the SportsPose dataset. This research opens new possibilities for domain adaptation in 3D pose estimation, providing a practical and cost-effective solution to customize models for specific applications, and unlocks the possibility of increasing the state-of-the-art accuracy in monocular pose estimation by training on large amounts of two-view in-the-wild data.

[1] http://christianingwersen.github.io/SportsPose

2. Related work

2.1. Monocular 3D human pose models

There are two primary approaches for monocular 3D human pose estimation. One solely predicts the 3D locations of the human skeleton [36, 42, 52], while the other includes estimating the body shape [2, 18, 24, 28, 50]. Most models that include the body shape employ parametric body models such as SMPL [30], which describes the body through shape parameters and pose parameters. Notably, our proposed loss remains applicable to both approaches, as 3D joint positions, as well as shape and pose parameters can be penalized based on the variation between multiple views.

Irrespective of whether the goal is to estimate pose alone or to include shape parameters, monocular 3D human pose estimation commonly adopts either a one-stage or a two-stage approach. In the one-stage approaches, the estimation is directly derived from an image or video input, while the two-stage approaches involve lifting estimated 2D poses to 3D space. State-of-the-art monocular models that employ the two-stage approach, lifting 2D poses to 3D, achieve mean-per-joint precision errors (MPJPE) as low as 17mm [55] when lifting ground truth 2D poses on the Human3.6M dataset [13], and 37mm when lifting estimated 2D poses [55]. However, when the same methods are evaluated on other datasets they have much higher MPJPE, it is clear that further work is still required [11].

Models that adopt the alternative approach of inferring the 3D pose by estimating the parametric SMPL model directly from image input, have achieved MPJPE scores of 60mm [43] on the in-the-wild 3DPW dataset [46].

2.2. Multiview 3D human pose models and datasets

Multiple synchronized and calibrated cameras have been extensively used to generate data to develop human pose estimation models [4, 15, 38]. Utilizing calibrated camera setups in such approaches has yielded impressive results, and has also been the basis for generating state-of-the-art 3D human pose datasets [12, 17, 31, 33]. These datasets have been essential for developing monocular pose models. However, the practical implementation of multi-camera setups involves a calibrated camera setup. Thus, most data has been collected in controlled laboratory environments, which does not reflect data variability in many scenarios.

Approaches that require limited or no 3D supervision have also been explored. Some of these are unsupervised and train on single images by lifting a 2D pose to 3D followed by a random rotation and re-projection to 2D [3, 5, 10, 48]. Others use multiple views of the same person [6, 14, 23, 29, 32], which is close to our approach. However, these approaches require camera calibration [29]. Some methods only rely on intrinsic calibration and estimate relative camera poses by decomposing the essential matrix estimated from 2D poses predicted in multiple views. Then, the 3D poses are triangulated using the estimated relative poses and used as training data [6, 23]. Mitra et al. [32] add additional training data from multiview images and use metric learning to enforce that images of the same pose have similar embeddings. Similar to our approach, Iqbal et al. [14] rigidly align 3D poses predicted from multiple views but opposed to our method they require known camera intrinsic and can only predict scale normalized poses. During training, they penalize the model for differences in the predicted poses. Both Iqbal et al. [14] and Mitra et al. [32] apply multiview consistency on single image pairs and not sequences, thus risk reducing accuracy and potential use by ignoring temporal information.

3. Multiview consistency loss

In our approach, we propose a loss function that optimizes for consistency between a sequence of 3D poses predicted from multiple views. The consistency is measured as the distance between Procrustes aligned pose sequences. Therefore, we do not need to know the cameras' intrinsic or extrinsic calibration or other prior information. Instead, the consistency loss applies a similarity transformation and penalizes differences in the poses of two sequences of the same activity, see Figure 1. Avoiding camera calibration simplifies the training pipeline and gives an efficient alternative for handling data from multiple views.

Specifically, the loss is based on the difference between poses computed from two or more views after alignment with a similarity transformation, τ. We compute the mean difference over every pair of two cameras, which results in the loss

$$\mathcal{L}_{\text{con}} = \sum_{s=1}^{S} \frac{1}{|V_s|} \sum_{(a,b)\in V_s} \mathcal{L}_{\text{c}} \left(\hat{J}_a, \hat{J}_b \right). \tag{1}$$

Here, S is the total number of sequences, and V_s is the set of possible pairs of views of the sequence s. Therefore, with N different cameras available in a sequence, $|V_s| = \binom{N}{2}$. The consistency loss $\mathcal{L}_{\text{c}}$ is calculated between $\hat{J}_a$ and $\hat{J}_b$, which are the predicted 3D body joints for all frames of the sequence from view a and view b, respectively. The term $\mathcal{L}_{\text{c}}$ is

$$\mathcal{L}_{\text{c}}(\hat{J}_a, \hat{J}_b) = \frac{1}{n} \sum_{i=1}^{n} \left\| \tau \left(\hat{J}_{a,i}; \hat{\theta}_{ab} \right) - \hat{J}_{b,i} \right\|_2, \tag{2}$$

where $\hat{J}_{a,i}$ is element i, from the sequence of predicted 3D poses from view a, which has length n. Similarly, $\hat{J}_{b,i}$ is element i from the sequence of predicted 3D poses from view b. τ is a similarity transform with parameters $\hat{\theta}_{ab}$ that are estimated such that τ transforms $\hat{J}_{a,i}$ to be as close as possible to $\hat{J}_{b,i}$ by scaling, rotating, and translating the 3D joints from $\hat{J}_{a,i}$.

To compute the scaling, rotation, and translation used to transform $\hat{J}_{a,i}$, we estimate the optimal parameters $\hat{\theta}_{ab}$, as in Equation (3). Here, it should be noted that contrary to how similarity transformations are traditionally computed in 3D human pose estimation to calculate the Procrustes aligned MPJPE, we only compute one transformation, $\hat{\theta}_{ab}$, for the entire sequence and not one per frame as in the PA-MPJPE metric [11]

$$\hat{\theta}_{ab} = \arg\min_{\theta} \sum_{i=1}^{n} \left\| \tau \left(\hat{J}_{a,i}; \theta \right) - \hat{J}_{b,i} \right\|_2^2. \tag{3}$$

The optimal solution to Equation (3) is found using Procrustes analysis [9], such that we obtain the optimal scaling, rotation, and translation to transform $\hat{J}_{a,i}$ as follows

$$\tau \left(\hat{J}_{a,i}; \hat{\theta}_{ab} \right) = s \hat{J}_{a,i} R + t. \tag{4}$$

By transforming $\hat{J}_a$, the idea is to directly estimate the similarity transformation that transforms from the camera coordinate system of camera a to the coordinate system of camera b instead of relying on knowing the camera extrinsics to perform the transformation.

4. Experiments

We conduct experiments with fine-tuning a pretrained pose estimator on a new dataset with ground truth 3D (Sec. 4.2)

Soccer Tennis Baseball Volley Jump

Figure 2. The five activities from SportsPose [12]. The top row displays the publicly available view "right". The bottom row features a view rotated 90 degrees relative to "right", which we refer to as "View 1".

and without any 3D data (Sec. 4.3). In Sec. 4.4 we show the applicability of our loss when training from scratch in a weakly supervised setting. Finally, we investigate the number of views our consistency loss needs (Secs. 4.5 and 4.6).

Datasets To evaluate our method we have chosen several datasets that all contain multiple views and ground truth 3D poses.

Human3.6M [13] is the most widely used dataset in human pose estimation. It is recorded with four fixed cameras and a total of 3.6 million frames across seven subjects.

SkiPose [40, 44] is a domain-specific dataset of 12 sequences of alpine skiers recorded with six PTZ cameras with a total of 10,197 frames. This lets us test the performance of our loss in the challenging setting of non-fixed cameras.

SportsPose contains several movements common in sports that pretrained models struggle with [12]. It is recorded with seven fixed cameras and has a total of 1.5 million frames, which lets us test the impact of the number of views on our loss.

SportsPose test protocol As Ingwersen et al. [12] do not provide a specified test protocol, we employ a test protocol inspired by Human3.6M [13], wherein subjects are distributed across sets to ensure that no subject appears in the same set.

We use subjects S04, S07, S09, S14, and S22 for validation. Subsequently, we employ subjects S06, S12, and S19 for testing. To focus on monocular performance, we opt to use only the currently available view, "right", during the testing and validation of the model. This decision streamlines the evaluation process, as we are interested in assessing the proficiency of the model when exposed to a single front-facing view. Examples of this view are in the first row of Figure 2.

4.1. Implementation

While our consistency loss is versatile and applicable to any monocular 3D human pose method, we choose to adapt the MotionBERT model by Zhu et al. [55] due to its impressive performance on multiple datasets. For fine-tuning the SportsPose [12] dataset we used 2D poses predicted by RTMPose [16].

Details of the preprocessing of the detected 2D poses are in the supplementary material.

For the fine-tuning of MotionBERT [55], we employ the weights provided for the DSTformer with a depth of five and eight heads. The sequence length is 243, and the feature and embedding sizes are 512 as in the original paper. Adhering to the training protocol suggested by Zhu et al. [55], we fine-tune the models for 30 epochs, using a learning rate of 0.0002 and utilizing the Adam optimizer [22].

4.2. Fine-tuning with 3D data

To compare to the situation when 3D data is available, we also experiment with fine-tuning with 3D data. We implement the proposed fine-tuning configuration from MotionBERT. This involves using a positional loss, $\mathcal{L}_{\text{pos}}$, directly on the 3D poses, coupled with losses on joint velocities, $\mathcal{L}_{\text{vel}}$, and scale only loss, $\mathcal{L}_{\text{scale}}$, as suggested by Rhodin et al. [39]. This combination results in the combined loss for 3D data,

$$\mathcal{L}_{\text{3D}} = \lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{scale}}\mathcal{L}_{\text{scale}}, \tag{5}$$

where λ_{pos}, λ_{vel}, and λ_{scale} are weights for the respective losses. Our proposed consistency loss is added as a regularization term, $\lambda_{\text{con}}\mathcal{L}_{\text{con}}$, to the total loss, resulting in Equation (6),

$$\mathcal{L}_{\text{3D}_{\text{con}}} = \lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{scale}}\mathcal{L}_{\text{scale}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}. \tag{6}$$

After an extensive parameter search, aligning with suggestions from Zhu et al. [55], we identify the optimal configuration for Equation (6) as $\lambda_{\text{pos}} = 1$, $\lambda_{\text{vel}} = 20$, $\lambda_{\text{scale}} = 0.5$, and $\lambda_{\text{con}} = 0.2$. These parameters are employed to obtain the results presented in Table 1, utilizing two camera views from SportsPose [12], one from the right side, as illustrated in the first row of Figure 2, and another 90 degrees to the view facing the back of the subject as in the second row of Figure 2. The second view behind the subject is based on the assumption that this view contains the most information when joints are occluding each other in the original "right" view from SportsPose [12].

The results in Table 1 show the impact of the consistency loss on model performance. Even when ground truth 3D data is available, the consistency loss yields marginal improvements, with a 0.8mm decrease in MPJPE and 0.2mm in PA-MPJPE. This slight enhancement suggests that our

Table 1. Results on SportsPose [12]. Baseline is MotionBERT [55], which is then fine-tuned with either 2D ($\mathcal{L}_{2D}$) or 3D ($\mathcal{L}_{3D}$) supervision with and without our proposed multiview consistency loss $\mathcal{L}_{con}$. All results are in mm where lower is better. MPJPE is mean per joint precision error, and PA is Procrustes aligned MPJPE. All results use predicted 2D poses [16]. **Bold** is the best performance with only 2D data and **bold gray** is best performance with 3D ground truth. The two views are shown in Figure 2. The consistency loss improves performance for both 2D and 3D but substantially more when 3D supervision is not used.

	Soccer kick		Tennis serve		Baseball pitch		Volley		Jumping		All	
	MPJPE	PA	MPJPE	PA	MPJPE	PA	MPJPE	PA	MPJPE	PA	MPJPE	PA
Baseline												
MotionBERT [55]	64.2	39.5	70.7	39.7	85.0	42.2	86.8	50.0	78.0	48.9	77.1	44.1
Iqbal et al. [14][2]	42.8	28.5	39.9	26.2	47.0	30.1	40.3	27.9	44.9	30.1	42.9	28.5
Fine-tuning with 3D data (2 views)												
$\mathcal{L}_{3D}$ (5)	26.7	**20.2**	27.3	20.1	30.1	22.5	31.3	24.2	27.9	21.4	28.7	21.7
$\mathcal{L}_{3D_{con}}$(6), Ours	**26.1**	20.5	**25.4**	**18.8**	**29.4**	**22.4**	**30.8**	**23.8**	**27.9**	**20.9**	**28.0**	**21.3**
Only 2D fine-tuning (2 views)												
$\mathcal{L}_{2D}$(7)	59.0	44.1	59.1	42.0	73.8	45.1	64.7	47.8	65.0	45.6	64.4	45.0
$\mathcal{L}_{2D_{con}^{frame}}$, Ours	**33.9**	21.9	**31.0**	**20.1**	**36.1**	**23.0**	36.5	23.9	**34.1**	23.5	**34.4**	**22.5**
$\mathcal{L}_{2D_{con}}$(8), Ours	35.4	**20.9**	36.2	22.7	40.9	26.1	**33.5**	**22.6**	35.1	**21.5**	36.2	22.8

Table 2. Results on SkiPose [44] Methods in gray use all six views during inference, the rest are monocular. "F" indicates fine-tuning a pretrained model, "S" is self supervised using only 2D on SkiPose. Best performance without ground truth 2D annotations is marked in **bold**.

Method	MPJPE	PMPJPE	Uses
Metapose[45]	-	42	S
Zhou et al. [54]	42.2	29.4	F+3D
MotionBERT [55]	259	132	
CanonPose: [47]	128.1	89.6	S
MHCanonNet [21]	122	50.7	S
Yang et al. [51]	-	68.4	S
Kim et al. [20]	115.2	78.8	S
Pavllo et al. [35][3]	106	88.1	F+S
PoseAug [8][3]	105.4	83.5	F+S
AdaptPose[7]	99.4	83.0	F+S
$\mathcal{L}_{2D_{con}^{frame}}$, Ours	**62.0**	**40.5**	F+S
$\mathcal{L}_{2D_{con}^{frame}}$, Ours, $2D^{GT}$	49.1	32.3	F+S

regularization term can be seamlessly integrated even when 3D data are available without compromising performance. It is crucial to emphasize that our consistency loss is intentionally crafted for scenarios lacking 3D data. But this underscores its utility as a valuable regularization technique for monocular pose estimation, acknowledging that the use of 3D data remains superior to achieve a well-performing model.

4.3. Fine-tuning without 3D data

Our consistency loss is especially useful when fine-tuning a model on a new dataset where multiple views are available but 3D poses are not.

The common way of utilizing additional data without 3D poses during training is to penalize deviations between the 2D poses and the reprojection of the predicted 3D poses [19]. When this is done with ground truth 2D it can work, but it can result in poor results due to the 2D-3D ambiguity [11]. Obtaining ground-truth 2D poses further requires manual annotations. Instead, predicted 2D poses are commonly used. We denote this 2D reprojection loss as

$$\mathcal{L}_{2D} = \lambda_{2D_{reproj}} \mathcal{L}_{2D_{reproj}}. \tag{7}$$

We also add our consistency loss as regularization

$$\mathcal{L}_{2D_{con}} = \lambda_{2D_{reproj}} \mathcal{L}_{2D_{reproj}} + \lambda_{con} \mathcal{L}_{con}. \tag{8}$$

Through experimentation, we found that we achieve the best performance, with $\lambda_{2D_{reproj}} = 1$ and $\lambda_{con} = 0.3$, and use these values unless otherwise specified.

Using the same two views as described in Section 4.2 we fine-tune on the SportsPose dataset with the loss from Equation (7) or Equation (8) and present the results in Table 1. Here we also present results with a variant of our consistency loss $\mathcal{L}_{2D_{con}^{frame}}$, where the optimal similarity transform is computed per frame, to allow for moving cameras. In Tab. 2 we present results on fine-tuning on the SkiPose dataset only with $\mathcal{L}_{2D_{con}^{frame}}$, as the cameras are not fixed.

Tables 1 and 2 highlights the impact of our consistency loss, particularly when 3D supervision is unavailable. The addition of the consistency loss on SportsPose leads to a substantial improvement in MPJPE, demonstrating a reduction of 39.2mm compared to relying solely on the reprojection loss. Figure 3 shows predictions from models with and without the consistency loss and we see the same substantial increase in accuracy when using the consistency loss.

The improvement for 2D data is this big because the consistency loss has improved the network's ability to resolve ambiguities during the process of lifting 2D to 3D from a single view. In addition, our method proves beneficial in situations where joints might be occluded in one of the views, enhancing the overall robustness of the model.

However, a closer examination of the PA-MPJPE in Table 1 reveals an interesting observation. Fine-tuning the model solely on 2D body keypoints based on Eq. (7) results in an increase in error. A likely cause is that the 3D

[2] As they have not released their code, we re-implement an improved version of their method by incorporating MotionBERT as the backbone with sequence length 1.

[3] Value for [8] and [35] reported by [7].

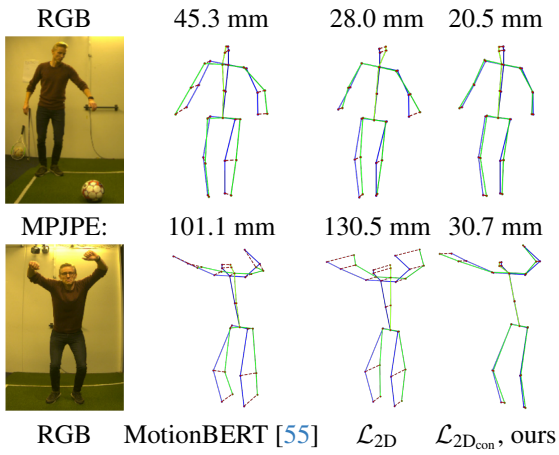


Figure 3. Visual comparison of predictions in green and the ground truth pose in blue. The magnitude of errors, measured in millimeters and indicated at the top, highlights the superiority of our consistency loss $\mathcal{L}_{2D_{con}}$ in achieving more accurate results. The notable improvement is especially evident in the bottom row, where the method employing our consistency loss successfully captures the complex movement.

Table 3. MPJPE and PA-MPJPE for different combinations of two views. The view "right" is included in all combinations. All experiments have been conducted with the loss $\mathcal{L}_{2D_{con}}$ and $\lambda_{con} = 1$. It is clear that the two-view combination matters with right + view 1 and right + view 2 achieve substantially lower errors.

Right + view x	MPJPE	PA-MPJPE
View 1	21.8	**22.4**
View 2	**21.6**	24.3
View 3	27.3	31.8
View 4	25.6	26.7
View 5	31.9	35.6
View 6	25.8	27.2

points that reproject to the same 2D points are not unique. Consequently, the model may struggle to provide accurate depth estimates of joint locations, as highlighted by Ingwersen et al. [11]. This underscores the importance of the consistency loss in mitigating such challenges and emphasizes its role in refining the performance of the model without ground truth 3D data.

4.4. Comparison with state-of-the-art on Human3.6M data

Table 4 shows results from our method with consistency loss compared with other semi-supervised methods trained on the Human3.6M [13] dataset. For the evaluation, we adopt the same protocol as Iqbal et al. [14] using 3D supervision from S1 and use S5, S6, S7, S8 with our consistency loss, $\mathcal{L}_{2D_{con}}$, from Eq. (8). For evaluation, we use the standard protocol testing on subjects S9 and S11 [13, 14].

Table 4. Comparison with reported state-of-the-art Semi-Supervised methods trained from scratch on the Human3.6M [13] dataset using only 3D data from S1 and weak supervision on S5, S6, S7, S8 during training. Unreported values are marked with "-". All 2D poses are predicted.

Methods	**MPJPE** ↓	**PA-MPJPE** ↓
Rhodin et al. [39] (ECCV'18)	131.7	98.2
Pavlakos et al. [34] (ICCV'19)	110.7	74.5
Li et al. [26] (ICCV'19)	88.8	66.5
Rhodin et al. [40] (CVPR'18)	-	65.1
Kocabas et al. [23] (CVPR'19)	-	60.2
Iqbal et al. [14] (CVPR'20)	62.8	51.4
Roy et al. [41] (3DV'22)	60.8	48.4
Li and Pun [25] (AAAI'23)	51.9	-
Ours, $\mathcal{L}_{2D_{con}^{frame}}$	52.1	41.0
Ours, $\mathcal{L}_{2D_{con}}$ (8)	**50.5**	**40.4**

Table 4 demonstrates that our novel consistency loss significantly improves performance in scenarios where 3D data is limited and with unlabeled multiview data. A key innovation of our approach, compared to Iqbal et al. [14], lies in incorporating the temporal aspect by computing a similarity transform per sequence instead of per frame, and not requiring known camera intrinsics. This advancement establishes a new state-of-the-art in semi-supervised performance on the Human3.6M dataset [13]. Recent work [7, 37] obtain very similar performance under the same weakly supervised protocol, but they utilize ground truth 2D poses, making the results not directly comparable.

4.5. How many views do we need?

Examining the experiments carried out in Sections 4.2 and 4.3, a logical inquiry arises regarding the scalability of the results when more than two views are incorporated into the experiments. To investigate the correlation between the number of views and performance, we have calculated the results for scenarios where one to seven views are available, encompassing the total number of views in the SportsPose dataset [12].

Without available 3D data. In the absence of ground-truth 3D data, the influence of including multiple views on accuracy is evident as shown in Section 4.3 and the results in Table 1. To compute the results that involve more than two views without access to 3D data, we utilize the loss function from Equation (8) with a consistent configuration, specifically setting $\lambda_{2D_{reproj}} = 1$ and $\lambda_{con} = 1$ for all experiments. It is essential to note that this configuration is not fine-tuned for a specific number of views, which may result in variations compared to the results presented in Table 1. The outcome of this ablation study is detailed in Figure 4a.

Examining the results for 2D supervision in Figure 4a reveals a substantial increase in accuracy as we progress from one to two views. However, the accuracy curve for both MPJPE and PA-MPJPE appears to plateau beyond two views, with marginal gains observed when incorporating more than two views.

This observed plateau could be attributed to diminishing returns in information gain beyond the second view. While additional views contribute valuable perspectives, they may not necessarily introduce new information that significantly refines the precision of the predicted joints. Interestingly, this property of the loss underscores its utility, particularly in scenarios where capturing new data becomes significantly more manageable requiring only two views of the activity from an uncalibrated camera setup.

With available 3D data. Even when 3D data is available, incorporating our consistency loss with two views results in a modest performance gain in MPJPE or PA-MPJPE, as illustrated in Section 4.2 and detailed in Table 1. This raises the question of whether this incremental gain will persist with an increasing number of views or reach a plateau, similar to the findings with only 2D. In these experiments, we employ the loss function from Equation (6) with $\lambda_{\text{pos}} = 1$, $\lambda_{\text{vel}} = 20$, $\lambda_{\text{scale}} = 0.5$, and $\lambda_{\text{con}} = 1$. Notably, these values are not fine-tuned for any specific number of views and may thus differ from the results presented in Table 1. The outcomes of this experiment are illustrated in Figure 4b. We observe a slight increase in performance when additional views are added, along with the inclusion of our consistency loss. However, the variation in performance is generally small, and the overarching conclusion remains unchanged: when 3D data is available, there is no need to adapt the consistency loss.

4.6. More views or more data?

Examining Figures 4a and 4b, one may question if both the marginal accuracy improvements for 3D supervision and the substantial gains with 2D supervision with our consistency loss are only due to the increased amount of training data. To explore this we have conducted the same experiments but without including $\mathcal{L}_{\text{con}}$ in the loss.

Figure 4c shows the experiment analogous to 2D supervision illustrated in Figure 4a, but without the consistency loss. It reveals that neither MPJPE nor PA-MPJPE exhibit improvement with the addition of more training data when incrasing the number of views. The observed plateau after two views contradicts the substantial accuracy increases depicted in Figure 4a, suggesting that these improvements are attributed to the introduction of our consistency loss.

However, examining the experiments adding data to the 3D supervision in Figure 4d, we observe a trend similar to that depicted in Figure 4b with the error decreasing marginally when we add more data to the training. This suggests that the marginal improvements in accuracy when employing our consistency loss with 3D supervision can be attributed to the increased volume of data rather than solely to the presence of the consistency loss. This finding supports the overarching conclusion that 3D data is superior, and supports that the main advantage of our consistency loss lies in enhancing accuracy in scenarios where obtaining 3D data is impractical.

4.7. Which views to choose

In the experiments of Figures 4a and 4b, the selection of views followed a deterministic process. Specifically, the first view was consistently chosen as the "right" view from SportsPose [12], and the second view was facing the back of the subject *i.e*. the one positioned closest to a 90-degree angle relative to the initial view. For scenarios involving three or more views, the remaining views were selected arbitrarily but maintained the same order across all experiments.

As the ambiguities of a 3D pose can theoretically be solved when observing from any two different views, a question arises if there is a practical advantage of certain viewpoints. This was investigated in Table 3 where the model was trained using the "right" view in combination with all other available views, where View 1 corresponds to the perspective positioned 90 degrees relative to the view facing the back of the subject.

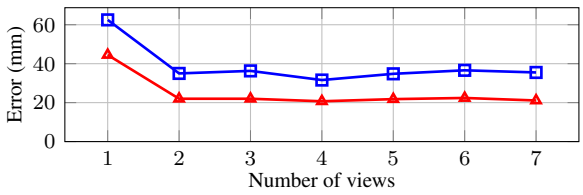
Analyzing the errors depicted in Table 3, it is evident that the choice of the view for multiview supervision significantly influences the outcomes of using our consistency loss. This aligns with intuition, as certain views are more effective in resolving ambiguities and identifying occluded joints, while others may not contribute new information. The results indicate that an optimal configuration involves using views from two sides that are 90 degrees apart when only two cameras are available.

5. Discussion and conclusion

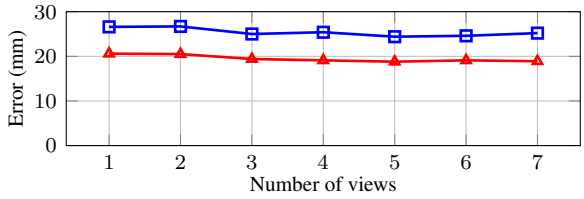
Limitations. While our results underscore a notable improvement in accuracy achieved through the implementation of our consistency loss, it is important to use the method with care. As Table 3 demonstrates, the effect of the consistency loss depends on which views are used during training, with the least favorable combination resulting in performance comparable to using a single camera view. However, accuracy increases substantially in four out of five combinations.

Furthermore, emphasizing the consistency loss too much, indicated by a large λ_{con} value, can lead to a degenerate solution. Specifically, an optimal solution to Equation (2) may predict the same position of all joints.

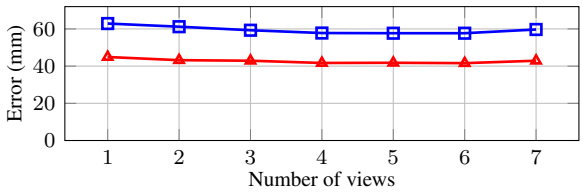
Our proposed consistency loss estimates a single transformation for a pose sequence for aligning the 3D poses from different views. While this requires the cameras to



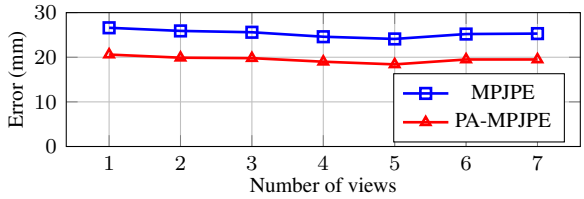
(a) Loss: $\mathcal{L}_{2D_{con}}$. With just two available views, the error decreases significantly, but adding more views only decreases the error marginally.



(b) Loss: $\mathcal{L}_{3D_{con}}$. There is a slightly lower error for more views, but the increase is far from as significant as when 3D data is not available.



(c) Loss: $\mathcal{L}_{2D}$. The aim is to discern whether the increase in accuracy observed in Figure 4a is influenced by the consistency loss or the augmented data availability. The nearly flat lines for both errors indicate that the accuracy boost associated with two or more views primarily stems from incorporating the consistency loss.



(d) Loss: $\mathcal{L}_{3D}$. The purpose is to explore whether the marginal improvements in accuracy in Figure 4b are attributable to the consistency loss or the increased availability of data. Notably, we observe a slight decreasing trend in error as the number of views is increased, even without the consistency loss.

Figure 4. Investigation of how MPJPE and PA-MPJPE are affected when varying the number of views available with different losses. Top: With consistency loss, $\lambda_{con} = 1$, bottom without consistency loss. Left: 2D, right: 3D.

be stationary throughout the sequence, our method can be extended to movable cameras by finding a similarity transformation for each time step.

It is worth noting that incorporating the proposed consistency loss necessitates temporal synchronization of pose sequences from different views. Desynchronisation will appear as correlated noise in the 2D predictions that in the worst case can confuse the model, especially for fast movements. This requirement imposes constraints on the camera system used for data capture. Using two cameras capturing frames with logged timestamps, it is possible to manually identify the same point in time in both sequences, or to use the audio to time-synchronize the views after acquisition [27]. However, if using Android smartphones as cameras, the frame-synchronization can be obtained using an app and WiFi hotspots as described by Akhmetyanov et al. [1].

While we demonstrate our loss on the SportsPose dataset [12], which contains ground truth 3D data as well as a full multi-camera calibration, we only use the 3D data for evaluation, relying purely on predicted 2D keypoints for training. Because of the similarity transformation, our view-consistency loss eliminates the need for knowing the camera intrinsic or extrinsics, and only needs synchronized cameras. Foregoing camera calibration unlocks new opportunities for scaling multiview data acquisition as the footage can be captured on smartphones outside of the lab in diverse settings. The problem synchronization can be done with existing apps over WiFi [1] or solved as a post-processing step using audio cues [27].

Conclusion. We present a novel method to enhance monocular 3D human pose estimation performance. By incorporating our multiview consistency loss during training in scenarios where 3D data is unavailable, we achieve notable performance improvements when compared to relying solely on 2D reprojection loss or no fine-tuning without requiring knowledge of the camera's extrinsic or intrinsic parameters.

We demonstrate the efficacy of the proposed consistency loss by evaluating it on the Human3.6M [13], SkiPose [44], and SportsPose [12] datasets. Following the semi-supervised protocol for Human3.6M, we advance the weakly supervised state-of-the-art precision.

A thorough analysis exploring various configurations involving the number of views and camera placement reveals that an effective enhancement is achieved with just two appropriately positioned views. We observe that positioning the cameras at a 90-degree angle yields consistently good performance compared to other combinations of views. This demonstrates that, through the use of our multiview consistency loss, it is feasible to capture new domain data for fine-tuning a 3D model with a simple setup needing only two appropriately positioned and time-synchronized cameras.

With this paper, we also release six new views of sports activities to the SportsPose [12] dataset. Together with the new data we propose a new test protocol for the dataset and provide a simple baseline relying on MotionBERT [55] and our proposed consistency loss.

References

[1] Azat Akhmetyanov, Anastasiia Kornilova, Marsel Faizullin, David Pozo, and Gonzalo Ferrer. Sub-millisecond video synchronization of multiple android smartphones. *2021 IEEE Sensors*, pages 1–4, 2021. 8

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*. Springer International Publishing, 2016. 1, 2

[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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5714–5724, 2019. 3

[4] Sungho Chun, Sungbum Park, and Ju Yong Chang. Representation learning of vertex heatmaps for 3d human mesh reconstruction from multi-view images, 2023. 3

[5] Dylan Drover, Rohith MV, Ching-Hang Chen, Amit Agrawal, Ambrish Tyagi, and Cong Phuoc Huynh. Can 3d pose be learned from 2d projections alone? In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 3

[6] Mohsen Gholami, Ahmad Rezaei, Helge Rhodin, Rabab Ward, and Z Jane Wang. Tripose: A weakly-supervised 3d human pose estimation via triangulation from video. *arXiv preprint arXiv:2105.06599*, 2021. 3

[7] Mohsen Gholami, Bastian Wandt, Helge Rhodin, Rabab Ward, and Z Jane Wang. Adaptpose: Cross-dataset adaptation for 3d human pose estimation by learnable motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13075–13085, 2022. 5, 6

[8] Kehong Gong, Jianfeng Zhang, and Jiashi Feng. Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8575–8584, 2021. 5

[9] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. 3

[10] Peter Hardy and Hansung Kim. Links" lifting independent keypoints"-partial pose lifting for occlusion handling with improved accuracy in 2d-3d human pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3426–3435, 2024. 3

[11] Christian Keilstrup Ingwersen, Janus Nørtoft Jensen, Morten Rieger Hannemose, and Anders B. Dahl. Evaluating current state of monocular 3d pose models for golf. In *Proceedings of the Northern Lights Deep Learning Workshop*, 2023. 1, 2, 3, 5, 6

[12] Christian Keilstrup Ingwersen, Christian Mikkelstrup, Janus Nørtoft Jensen, Morten Rieger Hannemose, and Anders Bjorholm Dahl. Sportspose: A dynamic 3d sports pose dataset. In *Proceedings of the IEEE/CVF International Workshop on Computer Vision in Sports*, 2023. 2, 3, 4, 5, 6, 7, 8

[13] 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 Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 2, 4, 6, 8

[14] 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*, pages 5243–5252, 2020. 1, 3, 5, 6

[15] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *International Conference on Computer Vision (ICCV)*, 2019. 3

[16] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose. *arXiv e-prints*, art. arXiv:2303.07399, 2023. 1, 4, 5

[17] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015. 2, 3

[18] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 1, 2

[19] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5614–5623, 2019. 5

[20] Hyun-Woo Kim, Gun-Hee Lee, Myeong-Seok Oh, and Seong-Whan Lee. Cross-view self-fusion for self-supervised 3d human pose estimation in the wild. In *Proceedings of the Asian Conference on Computer Vision*, pages 1385–1402, 2022. 5

[21] Hyun-Woo Kim, Gun-Hee Lee, Woo-Jeoung Nam, Kyung-Min Jin, Tae-Kyung Kang, Geon-Jun Yang, and Seong-Whan Lee. Mhcanonnet: Multi-hypothesis canonical lifting network for self-supervised 3d human pose estimation in the wild video. *Pattern Recognition*, 145:109908, 2024. 5

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[23] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-supervised learning of 3d human pose using multi-view geometry. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1077–1086, 2019. 3, 6

[24] Nikos Kolotouros, Georgios Pavlakos, Michael Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2252–2261, 2019. 2

[25] Haolun Li and Chi-Man Pun. Cee-net: Complementary end-to-end network for 3d human pose generation and estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1):1305–1313, 2023. 6

[26] Zhi Li, Xuan Wang, Fei Wang, and Peilin Jiang. On boosting single-frame 3d human pose estimation via monocular videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2192–2201, 2019. 6

[27] Junwei Liang, Poyao Huang, Jia Chen, and Alexander Hauptmann. Synchronization for multi-perspective videos in the wild. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1592–1596. IEEE, 2017. 8

[28] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1954–1963, 2021. 2

[29] Yanchao Liu, Xina Cheng, and Takeshi Ikenaga. Motion-aware and data-independent model based multi-view 3d pose refinement for volleyball spike analysis. *Multimedia Tools and Applications*, pages 1–24, 2023. 3

[30] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 2

[31] 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, 2017. 2, 3

[32] Rahul Mitra, Nitesh B Gundavarapu, Abhishek Sharma, and Arjun Jain. Multiview-consistent semi-supervised learning for 3d human pose estimation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 6907–6916, 2020. 3

[33] Aiden Nibali, Joshua Millward, Zhen He, and Stuart Morgan. ASPset: An outdoor sports pose video dataset with 3d keypoint annotations. *Image and Vision Computing*, 111: 104196, 2021. 2, 3

[34] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. Texturepose: Supervising human mesh estimation with texture consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 803–812, 2019. 6

[35] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7753–7762, 2019. 5

[36] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[37] Qucheng Peng, Ce Zheng, and Chen Chen. A dual-augmentor framework for domain generalization in 3d human pose estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2240–2249, 2024. 6

[38] N Dinesh Reddy, Laurent Guigues, Leonid Pishchulin, Jayan Eledath, and Srinivasa G. Narasimhan. Tessetrack: End-to-end learnable multi-person articulated 3d pose tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15190–15200, 2021. 3

[39] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation learning for 3d human pose estimation. In *ECCV*, 2018. 4, 6

[40] 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. 4, 6

[41] S. Roy, L. Citraro, S. Honari, and P. Fua. On triangulation as a form of self-supervision for 3d human pose estimation. In *2022 International Conference on 3D Vision (3DV)*, pages 1–10, Los Alamitos, CA, USA, 2022. IEEE Computer Society. 6

[42] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 461–478. Springer, 2022. 2

[43] Karthik Shetty, Annette Birkhold, Srikrishna Jaganathan, Norbert Strobel, Markus Kowarschik, Andreas Maier, and Bernhard Egger. Pliks: A pseudo-linear inverse kinematic solver for 3d human body estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 574–584, 2023. 3

[44] Jörg Spörri. Research dedicated to sports injury prevention-the'sequence of prevention'on the example of alpine ski racing. *Habilitation with Venia Docendi in Biomechanics*, 1(2): 7, 2016. 2, 4, 5, 8

[45] Ben Usman, Andrea Tagliasacchi, Kate Saenko, and Avneesh Sud. Metapose: Fast 3d pose from multiple views without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6759–6770, 2022. 5

[46] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[47] Bastian Wandt, Marco Rudolph, Petrissa Zell, Helge Rhodin, and Bodo Rosenhahn. Canonpose: Self-supervised monocular 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13294–13304, 2021. 5

[48] Bastian Wandt, James J Little, and Helge Rhodin. Elepose: Unsupervised 3d human pose estimation by predicting camera elevation and learning normalizing flows on 2d poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6635–6645, 2022. 3

[49] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3349–3364, 2021. 1

[50] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6184–6193, 2020. 2

[51] Geon-Jun Yang, Jun-Hee Kim, and Seong-Whan Lee. Geometry-driven self-supervision for 3d human pose estimation. *Neural Networks*, 174:106237, 2024. 5

[52] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13232–13242, 2022. 1, 2

[53] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11656–11665, 2021. 1

[54] Kangkang Zhou, Lijun Zhang, Feng Lu, Xiang-Dong Zhou, and Yu Shi. Efficient hierarchical multi-view fusion transformer for 3d human pose estimation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7512–7520, 2023. 5

[55] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1, 2, 4, 5, 6, 8