Cross-View Self-Fusion for Self-Supervised 3D Human Pose Estimation in the Wild

Hyun-Woo Kim[1], Gun-Hee Lee[2], Myeong-Seok Oh[2], and Seong-Whan Lee[1,2]

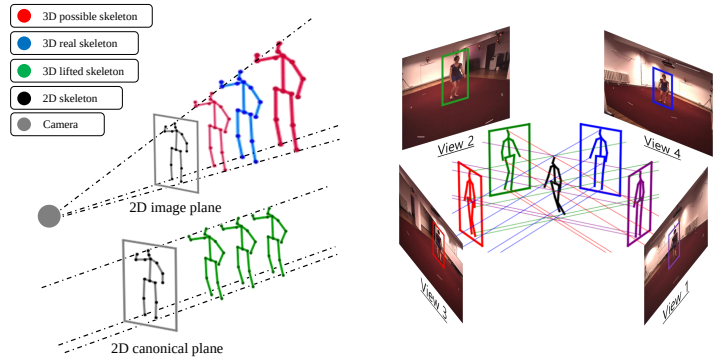
[1] Department of Artificial Intelligence, Korea University, Seoul, Korea
[2] Department of Computer Science and Engineering, Korea University, Seoul, Korea
{kim_hyun_woo,gunhlee,ms_oh,sw.lee}@korea.ac.kr

Abstract. Human pose estimation methods have recently shown remarkable results with supervised learning that requires large amounts of labeled training data. However, such training data for various human activities does not exist since 3D annotations are acquired with traditional motion capture systems that usually require a controlled indoor environment. To address this issue, we propose a self-supervised approach that learns a monocular 3D human pose estimator from unlabeled multi-view images by using multi-view consistency constraints. Furthermore, we refine inaccurate 2D poses, which adversely affect 3D pose predictions, using the property of canonical space without relying on camera calibration. Since we do not require camera calibrations to leverage the multi-view information, we can train a network from in-the-wild environments. The key idea is to fuse the 2D observations across views and combine predictions from the observations to satisfy the multi-view consistency during training. We outperform state-of-the-art methods in self-supervised learning on the two benchmark datasets Human3.6M and MPI-INF-3DHP as well as on the in-the-wild dataset SkiPose. Code and models are available at https://github.com/anonyAcc/CVSF_for_3DHPE

1 Introduction

Human Pose Estimation (HPE) is widely used in various AI applications such as video analysis, AR/VR, human action recognition, and 3D human reconstruction [1–8]. Owing to the variety of applicability, HPE has received considerable attention in computer vision. Recent methods for 3D HPE have achieved remarkable results in a supervised setting, but they require large amounts of labeled training data. Collecting such datasets is expensive, time-consuming, and mostly limited to fully controlled indoor settings that require a multi-camera motion capture system. Therefore, self-supervised 3D HPE, which does not require 3D annotation, has become an emerging trend in this field.

In this study, we propose a novel self-supervised training procedure that does not require camera calibrations (including camera intrinsic and extrinsic parameters) and any annotations in the multi-view training dataset. Specifically, our model requires at least two temporally synchronized cameras to observe a person of interest from different orientations, but no further knowledge regarding



(a) Examples of results lifted from general camera (top) and canonical camera (bottom)

(b) Visual understanding of canonical space

Fig. 1. (a) Examples of results predicted from general and canonical camera. Depth-scale ambiguity in general camera results with numerous 3D candidates of varying scales. In canonical camera, 2D pose is lifted to 3D poses of the same scale irrespective of depth. Red and blue skeletons represent estimated 3D skeletons at varying scales. Scale of green skeletons is the same as that of canonical cameras. (b) Visual understanding of a canonical space. In the space, the scales of all observed human poses are equal, and the relationship between cameras is relative.

the scene and camera parameters is required. Note that multi-view images are used to train a network, but only a single image is used at inference time.

There are only a few comparable methods [9–12] that apply to our self-supervised setting. They require additional knowledge about the observed person such as bone length constraints [11] and 3D human structures [10], or traditional computer vision algorithms to obtain a pseudo ground truth pose [9]. On the other hand, CanonPose [12] learns a monocular 3D human pose estimator using the multi-view images without any prior information. However, the research does not address the 2D pose errors caused by a pretrained 2D pose estimator, which remains fixed during the training. The 2D pose errors not only propagate to the 3D prediction, but also may affect the multi-view consistency requirement during training, which can yield an inaccurate camera rotation estimation.

To address this issue, we refine the 2D pose errors influencing the 3D prediction and then lift the refined 2D pose to a 3D pose. In principle, we train a neural network by satisfying the multi-view consistency between the 2D poses through refining the incorrect 2D pose, as well as the multi-view consistency between the 3D outputs and 2D inputs. However, it is necessary to know the multi-view relationship to refine an incorrect 2D pose in a multi-view setting. The multi-view relationship can be represented using the parameters of each camera. We assume a training setting in which the camera parameters are not given. Also, estimating the camera parameters of each camera is complex and computationally intensive. Therefore, we deploy a canonical form [13] that fixes one camera and represents the remainder with the relative camera parameters

based on the fixed camera. According to this form, the relationship between the cameras can be represented as a relative rotation and translation.

On the other hand, every camera position is different for a particular 3D target, so all the scales of the observed 2D poses are different in each view due to perspective. Therefore, there exists an infinite number of 3D poses with multiple scales corresponding to a given 2D pose due to the depth-scale ambiguity, which is illustrated in Fig. 1 (a). To address the ambiguity, we transform the 2D poses into a canonical space by normalizing the position and scale of the 2D poses observed with different scales in all views. The transformation allows the distances between the 3D target and each camera to be the same, so we don't need to consider the relative translation. In other words, the relationship between the cameras can be represented only by relative rotation, and all lifted 3D poses bear the same scale. Fig. 1 (b) illustrates the canonical space in which it has the same scale for all transformed 2D poses and a lifted 3D pose satisfies the multi-view consistency.

The flow of our approach is as follows: First, we transform the estimated 2D pose in an image plane coordinate system into a canonical plane coordinate system. Second, we input the transformed 2D pose into a lifting network. Then, the network predicts a 3D pose in the canonical coordinate system and a camera rotation to rotate the pose to the canonical camera coordinate system. Third, the proposed cross-view self-fusion module takes the 2D poses along with the camera rotations predicted by the lifting network as input. Subsequently, it refines incorrect 2D poses by fusing all the 2D poses with the predicted rotations. Lastly, the refined 2D poses are lifted to 3D poses by the lifting network.

We evaluate our approach on two multi-view 3D human pose estimation datasets, namely Human3.6M [14] and MPI-INF-3DHP [15], and achieve the new state-of-the-art in several metrics for self-supervised manner. Additionally, we present the results for the SkiPose dataset that represents all the challenges arising from outdoor human activities, which can be hard to perform in the limited setting of traditional motion capture systems.

The contributions of our research can be summarized as follows:

- We propose a Cross-view Self-fusion module that refines an incorrect 2D pose using multi-view data without camera calibration. This can be performed in any in-the-wild setting as it does not require camera calibration.
- We improve a self-supervised algorithm to lift a 2D pose to a 3D pose by refining poses across views. Refinement enhances multi-view consistency, and the enhanced consistency enables more accurate refinement.
- We achieve state-of-the-art performance on 3D human pose estimation benchmarks in a self-supervised setting.

2 Related Work

Full Supervision. Recent supervised approaches depend primarily on large datasets with 3D annotations. These approaches can be classified into two categories: image-based and lifting-based 3D human pose estimations. The image-

based approaches [16–25] directly estimate the 3D joint locations from images or video frames. Although these approaches generally deliver exceptional performance on similar images, their ability to generalize to other scenes is restricted. In this regard, certain studies [30–33] have attempted to resolve this problem using data augmentation. The lifting-based approaches [34–43] leverage 2D poses from input images or video frames to lift them to the corresponding 3D poses, which is more popular among the state-of-the-art methods in this domain. Martinez et al. [34] showed the prospect of using only 2D joint information for 3D human pose estimation by proposing "a simple and effective baseline for 3D human pose estimation", which uses only 2D information but achieves highly accurate results. Owing to its simplicity, it serves as a baseline for several future studies. However, the main disadvantage of all full-supervised approaches is that they are not appropriately generalized for the unseen poses. Therefore, their application is substantially limited to new environments or in-the-wild scenes.

Self-Supervision with Multi-view. Recently, the research interest in self-supervised 3D pose estimation using unlabeled multi-view images has increased, and our research pertains to this category as well. The self-supervised approaches use 2D poses estimated from unlabeled multi-view images. These approaches usually follow a lifting-based pipeline and therefore, they extract the 2D poses from the images using 2D pose estimators [44–47]. In our case, to get 2D joints from the images, we exploit AlphaPose [46] that is pretrained on a MPII dataset [48]. In constraints to the calibrated multi-view supervised approaches [49–52], the self-supervised approaches do not require the camera parameters to use multi-view data and thus, do not use traditional computer algorithms such as triangulation to recover the 3D poses. Kocabas et al. [9] leveraged epipolar geometry to acquire a 3D pseudo ground-truth from multi-view 2D predictions and then used them to train the 3D CNN network. Although this effective and intuitive approach shows promising results, the errors caused by incorrectly estimated joints in 2D estimation lead to an incorrect pseudo ground-truth. Iqbal et al. [11] trained a weakly-supervised network that refines the pretrained 2D pose estimator which predict pixel coordinates of joints and their depth in each view during training. Unfavorably, this method is not robust in environments other than the datasets employed for training, which is a limitation of the method of estimating the 3D pose from the image unit. More recently, Wandt et al. [12] reconstructed the 3D poses in a canonical pose space that was consistent across all views. We take advantage of the canonical space to fuse poses between the multiple views without using any camera parameters to refine the 2D pose incorrectly estimated by the 2D pose estimator.

3 Methods

Our goal is to train a neural network to accurately predict a 3D pose from an estimated 2D pose. At training time, we use 2D poses observed in multi-view images to train the network. At inference time, the network estimates a 3D pose from a 2D pose observed in a single image. The overall process of our framework

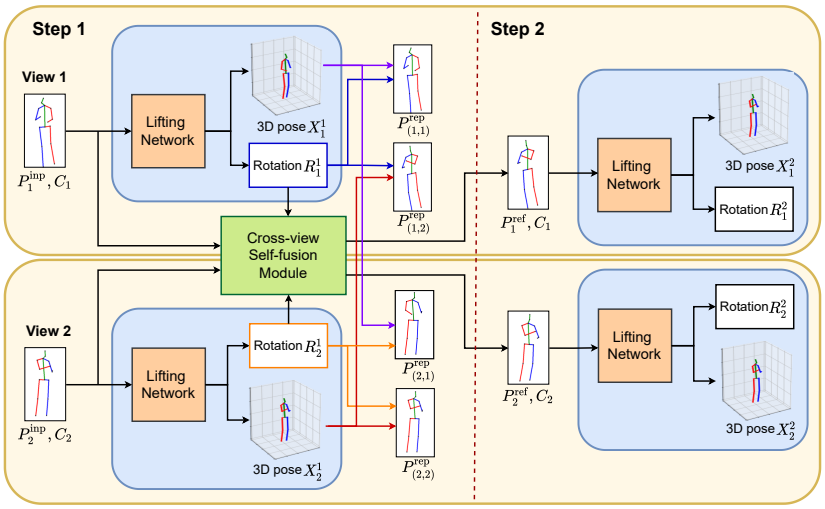


Fig. 2. A framework for learning a single 3D pose estimation from multi-view self-supervision while refining a 2D pose that adversely affects 3D pose estimation. At inference time, only a single view (blue box) is used for estimating a 3D human pose.

is as follows. For each view, a neural network takes a 2D pose as an input, and subsequently predict a camera rotation $R^1 \in \mathbb{R}^{3\times3}$ and 3D pose $X^1 \in \mathbb{R}^{3\times J}$ with J joint positions in the first lifting step. Then, the proposed Cross-view Self-fusion module fuses the input 2D poses with the predicted rotations from all views to refine the 2D poses as outputs. In the second lifting step, the refined 2D pose of each view is input into the lifting network to output a second 3D pose X^2 and camera rotation R^2 for each view. We define losses to each step and a total loss using the outputs of multiple weight-sharing neural networks and describe them in Section 3.4. The proposed framework with two cameras is illustrated in Fig. 2, which can be conveniently expanded with the availability of more cameras.

3.1 Lifting Network

Before inputting a 2D pose for each view to a lifting network, we normalize the 2D pose by centering it on the root joint and dividing it with its Euclidean norm. As we do not have any 3D annotations, we can train the lifting network by satisfying the multi-view consistency. Although the 3D poses lifted in each view must be identical to satisfy the multi-view consistency, the scales of the lifted 3D poses are different since the scales of the 2D poses are different in each view. Therefore, we transform the estimated 2D pose in each view into a canonical space, where the distances between the 3D target and each camera are the same, by normalizing it. $P^{\text{inp}} \in \mathbb{R}^{2\times J}$ is a transformed 2D pose that is input to the lifting network. We concatenate the confidences $C \in \mathbb{R}^{1\times J}$, provided by the 2D

pose estimator [46] for each predicted 2D joint, to the 2D input vector P^{inp} for input in the lifting network. The lifting network predicts a 3D pose $X \in \mathbb{R}^{3\times J}$ and rotation $R \in \mathbb{R}^{3\times 3}$ to rotate the pose to the canonical camera coordinate system. This division of the output into a 3D pose and camera rotation enables cross-view self-fusion and self-supervised learning to be possible.

3.2 Reprojection

As the canonical camera neglects the perspective in the canonical space, projecting the 3D prediction into the camera plane is accomplished by discarding the three dimensions, which is expressed as:

$$P^{\text{rep}} \;=\; \mathbf{I}_{[0:1]} \cdot R \cdot X, \tag{1}$$

where $P^{\text{rep}} \in \mathbb{R}^{2\times J}$ is the reprojected 2D pose and $\mathbf{I}_{[0:1]}$ is a truncated identity matrix that projects the 3D pose to 2D. The 3D pose X in the canonical space is rotated by the predicted rotation R to a canonical camera coordinate system. We rotate the m canonical 3D poses into the camera coordinate system of each camera through m rotations, in which the combining provides m^2 combinations. For instance, there are four possible combinations of rotations and poses for two cameras. During training, all possible combinations are reprojected onto the respective cameras. For example, $P^{\text{rep}}_{(2,1)}$ can be obtained by reprojecting a 3D pose X_1 predicted in view-1 to view-2 using the camera rotation R_2 predicted in view-2 as visualized in Fig. 2.

3.3 Cross-view Self-fusion

An inaccurate 2D pose results in incorrect 3D pose estimation. Typically, existing cross-view fusion methods utilize camera parameters to fuse cross-view data for accurate 2D poses. In the case of the canonical space, the relationship between multi-views can be represented only by relative rotation. Therefore, we propose a Cross-view Self-fusion Module (CSM) that refines an incorrect 2D pose using the predicted rotations and other input 2D poses. Fig. 3 illustrates our proposed module and a refinement process for a wrist joint that is incorrectly estimated by occlusion. The module takes a set of 2D poses, corresponding confidences and predicted rotations of all views as inputs, and outputs a set of the refined 2D poses: $\mathbf{P}^{\text{ref}} = \text{CSM}(\mathbf{P}^{\text{inp}}, \mathbf{C}, \mathbf{R})$.
Formally, our proposed module is defined as:

$$\begin{aligned}&\text{CSM}\left(\mathbf{P}^{\text{inp}}, \mathbf{C}, \mathbf{R}\right) = \\ &\bigcup_{m\in V}\left\{\left\{ argmax\Big(\mathbf{H}\left(p^i_m, c^i_m\right) + \sum_{\substack{n\in V,\\ n\neq m}} \mathbf{H}\left(\mathbf{E}\left(p^i_n, R_{(n,m)}\right), c^i_n\right)\Big) \;\Big|\; \forall\, i \in J \right\}\right\},\end{aligned} \tag{2}$$

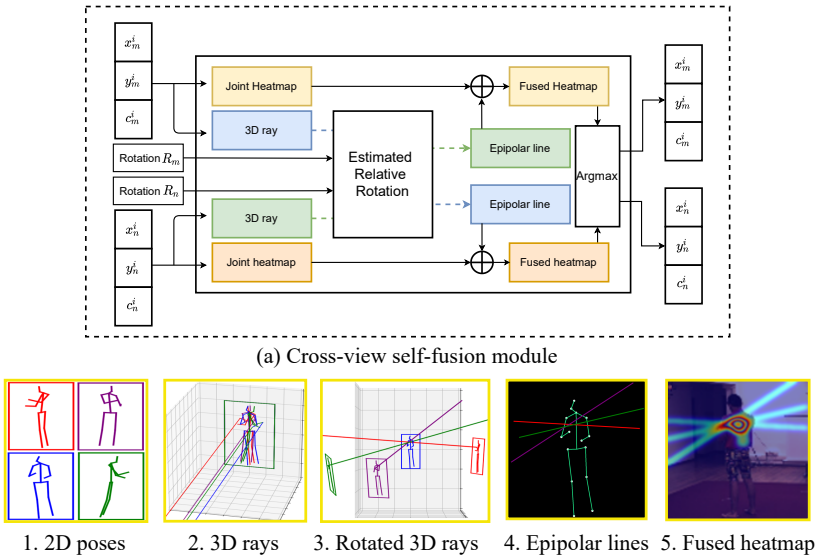


Fig. 3. Cross-view Self-fusion module. (a) shows our cross-view self-fusion module and (b) provides a process of fusing multi-view information about the wrist joint to refine the joint incorrectly estimated by occlusion.

where $V = \{1, 2, \ldots, v\}$ is a set of views and $J = \{1, 2, \ldots, j\}$ is a set of joints. p_m^i is the i-th joint of 2D pose P_m^{inp} from view-m. It has canonical plane coordinates $\{x_m^i, y_m^i\}$. A set of confidences corresponding to the 2D pose is $C_m = \{c_m^1, \ldots, c_m^j\}$. $R_{(n,m)}$ is the relative rotation between view-n and view-m. The relative rotation $R_{(n,m)}$ by rotating from view-n to view-m using rotation matrices R_n and R_m is defined as $R_{(n,m)} = R_n R_m^T$.

First, a heatmap generator $\mathbf{H}(\cdot)$ takes a joint p_m^i of 2D pose P_m^{inp} from view-m and a confidence c_m^i of the joint as inputs and generates a gaussian heatmap for the joint, in which the maximum value of the heatmap is its confidence. Second, an epipolar line generator $\mathbf{E}(\cdot)$ takes a joint p_n^i of 2D pose P_n^{inp} from view-n and a relative rotation $R_{(n,m)}$ as input and outputs an epipolar line for the joint on the view-m, as illustrated in Fig. 3 (b). The input joint is lifted to a 3D ray by simply adding the third dimension since the input is on canonical space where perspective is neglected. It is rotated to view-m by a relative rotation $R_{(n,m)}$ and projected to view-m similar to Sec. 3.2. A rotated and projected 3D ray is represented as a line on a view, which is called an epipolar line. Next, an epipolar line heatmap is generated with the maximum value of the heatmap as the confidence value of the joint. Finally, the position with the maximum value of a heatmap fused with a joint heatmap from view-m and epipolar line heatmaps from the other views becomes the coordinate value of the newly refined 2D joint position. It is repeated for all views.

3.4 Self-supervised Training

Due to the absence of supervision, we train the lifting network using the observed 2D information as well as the properties of multi-view consistency constraints. Our training procedures can be distinguished in two stages. In the first stage, we train the lifting network so that predictions satisfy the multi-view consistency for the input 2D poses. We proceed to the second stage when the total loss converges in the first stage because we determine that the lifting network sufficiently satisfies the consistency for the input 2D poses. In the second stage, we train the lifting network so that predictions satisfy the multi-view consistency for the refined 2D poses until the total loss second converges. This enables more accurate camera rotation estimation and enhances multi-view consistency. The prediction of accurate camera rotation results in a more precise refined 2D pose, which in turn leads to a more accurate camera rotation prediction. The key idea is to enhance the multi-view consistency through various losses defined by combining the rotations and 3D poses predicted from different views and fusing the 2D poses observed from different views with the rotations.

Reprojection Loss. Upon comparing the input 2D poses and the 2D reprojections of the combined 3D poses, the loss can be defined as:

$$\mathcal{L}_{\text{rep}} = \left\| \left(P_m^{\text{inp}} - \frac{P_n^{\text{rep}}}{\|P_n^{\text{rep}}\|_E} \right) \odot C \right\|_1, \tag{3}$$

where $\|\cdot\|_1$ denote the L_1 norm and $\odot$ indicates the Hadamard product. In particular, each deviation between the input and reprojected 2D pose is linearly weighted along with its confidence in order to a strong weight to the predicted joint in certainty and less weight to the predicted joints in uncertainty. Since the global scale of the 3D human pose is not given, the reprojection P^{rep} is scaled by the Euclidean norm. m and n indicate the camera indices.

Refinement Loss. Upon comparing the refined 2D poses and input 2D poses, the loss can be defined as:

$$\mathcal{L}_{\text{ref}} = \left\| \left(P_m^{\text{inp}} - P_n^{\text{ref}} \right) \odot C \right\|_1. \tag{4}$$

According to multi-view geometric consistency, if the 2D joints of all views are accurate and the camera relationship is known correctly, the intersection of epipolar lines on one view from the other views should be on a joint observed in one view. In our initial training, the predicted rotation is not accurate, so the 2D pose is refined to the wrong position. Therefore, we learn the initial refined 2D pose to be equal to the input 2D pose. This loss makes the camera rotation estimation more accurate and ensures that the refined 2D poses are plausible.

Refinement-Reprojection Loss. The loss between the refined 2D poses and 2D reprojections is defined as:

$$\mathcal{L}_{\text{ref-rep}} = \left\| \left(P_m^{\text{ref}} - \frac{P_n^{\text{rep}}}{\|P_n^{\text{rep}}\|_E} \right) \odot C \right\|_1. \tag{5}$$

We learn that the 2D reprojection of the lifted 3D pose is the same as the refined 2D pose, which is the result of the cross-view self-fusion module. This enables

the lifting network to learn a 3D pose similar to the 3D poses estimated by other views so that it does not violate multi-view consistency even if it takes an incorrectly estimated 2D pose as an input.

Multi-view 3D consistency Loss. To ensure multi-view consistency, previous work [12] has attempted to introduce a loss between the 3D poses predicted by multi-view to train a lifting network. As reported, the lifting network learned the 3D poses that were invariant toward the view but were no longer in close correspondence to the input 2D pose, thereby preventing the convergence of the network to plausible solutions. For each view, we lift a refined 2D pose to a 3D pose, as depicted in Fig. 2. More specifically, we enhance the consistency by adding a loss in 3D units using the 3D pose predicted from the refined 2D pose for each view. This loss is defined as the deviation between the 3D poses generated by lifting the refined 2D pose for each view.

$$\mathcal{L}_{\text{3D}} = \left\| X_m^2 - X_n^2 \right\|_1, \tag{6}$$

where X_m^2 is a second 3D pose lifted from a refined 2D pose of view-m and X_n^2 is a second 3D pose lifted from a refined 2D pose of view-n.

Total Loss. We sum up the losses described above for all views. To this end, we can define total loss as follows.

$$\mathcal{L} = \sum_{m=1}^{V} \sum_{n=1}^{V} \left(w_1 \mathcal{L}_{\text{ref}}^{m,n} \; + \; w_1 \mathcal{L}_{\text{rep}}^{m,n} \; + \; w_2 \mathcal{L}_{\text{ref-rep}}^{m,n} + w_2 \mathcal{L}_{\text{3D}}^{m,n} \right). \tag{7}$$

Until total loss $\mathcal{L}$ first converges, the weight w_1 is set to 1, and the weight w_2 is set to 0.01 in the first stage. Then, until the end of the training, we set the weight w_1 to 0.01, and the weight w_2 to 1 in the second stage.

4 Experiments

4.1 Datasets and Metrics

Dataset. We perform experiments on two standard benchmark datasets: Human3.6M [14] (H36M) and MPI-INF-3DHP [15] (3DHP). We also evaluate our method on the SkiPose dataset [53, 54] with six moving cameras to demonstrate the generality of our method to various in-the-wild scenarios. To conform with a setting of self-supervised training for a particular set of activities, we train one network for each dataset without using additional datasets. For each dataset, we follow the self-supervision protocols for training and evaluation [12, 9].

Metrics. For quantitative evaluation, we adopt the common protocol, *Normalized Mean Per Joint Position Error* (NMPJPE) and *Procrustes aligned Mean Per Joint Position Error* (PMPJPE) that measure the mean euclidean distance between the ground-truth and the predicted 3D joint positions after applying the optimal rigid alignment and scale (for NMPJPE), or optimal shift and scale (for PMPJPE) to poses. For 3DHP and SkiPose, we report the N-PCK, which is *Percentage of Correct Keypoints* (PCK) normalized by scale. The N-PCK indicates the percentage of joints whose estimated position is within 150mm of

Table 1. Per-action PMPJPE of different variants on the Human3.6M dataset.

MethodAct	Direct	Discuss	Eating	Greet	Phone	Photo	Pose	Purch
Baseline	50.5	49.0	46.1	54.0	51.2	53.2	59.5	49.1
Ours (S1)	41.9	42.3	41.1	45.3	45.3	45.0	48.8	43.4
Ours (S1+S2)	40.1	40.4	39.9	44.1	44.5	44.2	48.0	42.3
MethodAct	Sitting	SittingD	Smoke	Wait	Walk	WalkD	WalkT	**Avg**
Baseline	65.7	83.7	53.7	57.7	48.9	60.1	50.7	55.5
Ours (S1)	59.7	73.0	46.2	46.2	40.6	47.9	40.7	47.1
Ours (S1+S2)	59.0	72.1	45.3	44.3	39.4	46.1	39.9	**45.9**

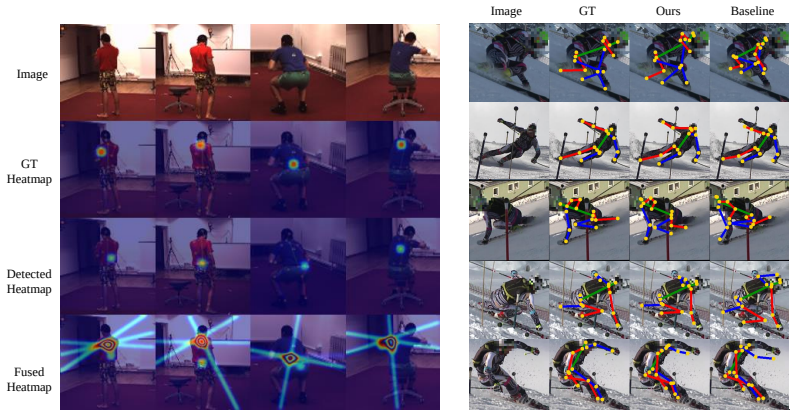
Table 2. Evaluation of 2D pose refinement accuracy for each dataset. We show JDR (%) for six important joints about each dataset.

Method	Dataset	Hip	Knee	Ankle	Shoulder	Elbow	Wrist
Single	H36M	97.1	97.5	97.5	98.5	96.7	98.2
Ours	H36M	**98.2**	**98.5**	**97.8**	**98.9**	**98.5**	**99.6**
Single	3DHP	97.4	97.8	99.8	96.9	97.0	96.9
Ours	3DHP	**98.8**	**97.8**	**99.9**	**98.4**	**98.4**	**98.3**
Single	Ski	97.0	73.7	**81.2**	90.0	70.0	**60.9**
Ours	Ski	**98.7**	**77.0**	75.1	**91.9**	**71.7**	56.4

the ground-truth. Lastly, for evaluating our proposed CSM, we measure the refined 2D pose accuracy by *Joint Detection Rate* (JDR), which is the percentage of the successfully detected joints. If the distance between the estimated and ground-truth locations is smaller than a threshold, this joint can be deemed to be successfully detected. The threshold is set to half of the head size.

4.2 Ablation Studies

We analyze the effectiveness of the proposed losses. Specifically, we design several variants of our method, and the details of these variants are shown as follows.
Baseline: The baseline does not consider the Cross-view Self-fusion module. The baseline is trained simply using the reprojection loss.
Step1 (S1): This variant adopts the refinement loss and refinement-reprojection loss by the CSM. It does not consider the lifting of Step 2.
Step2 (S2): It lifts a refined 2D pose to a 3D pose. This variant considers a multi-view 3D consistency loss between second 3D poses of all views.
We train all variants (Baseline, S1, S1+S2) on the H36M, and Table 1 demonstrates the per-action PMPJPE of all variants on the H36M.
Compared with the baseline, our CSM is helpful to obtain better results. We experiment with whether the CSM has the effect of intuitively refining incorrectly



(a) Samples of heatmaps (b) Examples of refined 2D poses for SkiPose

Fig. 4. (a) "Detected heatmap" indicated that it is extracted from a image of the target view. "Fused heatmap" is obtained by summing the "Detected heatmap" and the "Epipolar heatmaps" fused from the three remaining views. **(b)** Qualitative results of 2D refined pose for the challenging SkiPose dataset. We compare the visual results of ground-truth, ours and baseline.

estimated 2D poses due to occlusion. Table 2 shows the 2D pose estimation accuracy with JDR (%) for six important joints on the H36M, 3DHP and Skipose datasets. It compares our approach with the 2D pose estimator [46], termed *Single*, which estimates a 2D pose from a single image without performing cross view self-fusion. It can be seen that using CSM improves overall accuracy except for some joints. Fig. 4 (a) shows examples of the fused heatmap during the cross-view self-fusion process for the H36M. It shows examples of incorrectly estimated wrist joints by occlusions. It can be seen that the 2D pose is correctly refined by fusing the estimated epipolar line heatmaps. If the correct 2D pose is lifted to 3D, a more accurate 3D pose will be estimated. Fig. 4 (b) shows examples of the refined 2D poses for the SkiPose.

4.3 Comparison with State-of-the-Art Methods

We compare the results of the proposed method with other state-of-the-art approaches. For a fair comparison with [12], we follow the implementation and evaluation performed on [12]. We employ an off-the-shelf detector [46] to extract a 2D human pose required as an input to the proposed method.

Most methods using a lifting network without knowledge of scenes show the large gap between the NMPJPE and PMPJPE as a small error of the 2D pose incorrectly estimated in a particular view among all views leads to a large 3D NMPJPE error. They train their network without addressing the incorrectly estimated 2D pose, which further impacts to estimate the camera rotation because the network is trained with the violation of the multi-view consistency

Table 3. Evaluation results on the Human3.6M and comparison of the 3D pose estimation errors NMPJPE and PMPJPE (*mm*) of previous approaches. The best results are marked in bold. Our model outperforms all self-supervised methods.

Supervision	Method	**NMPJPE ↓**	**PMPJPE ↓**
Full	Martinez [34]	67.5	52.5
Weak	Rhodin [55]	122.6	98.2
	Rhodin [53]	80.1	65.1
	Wandt [56]	89.9	65.1
	Kolotouros [57]	-	62.0
	Kundu [58]	85.8	-
Self	Kocabas [9]	76.6	67.5
	Jenni [10]	89.6	76.9
	Iqbal [11]	69.1	55.9
	Wandt [12]	74.3	53.0
	Ours (S1)	63.6	46.1
	Ours (S1+S2)	**61.4**	**45.9**

Table 4. Evaluation results on the MPI-INF-3DHP. NMPJPE and PMPJPE are reported in millimeters, and N-PCK is in %. The best results are marked in bold.

Supervision	Method	**NMPJPE ↓**	**PMPJPE ↓**	**N-PCK ↑**
Weak	Rhodin [53]	121.8	-	72.7
	Kolotouros [57]	124.8	-	66.8
	Li [59]	-	-	74.1
	Kundu [58]	103.8	-	**82.1**
Self	Kocabas [9]	125.7	-	64.7
	Iqbal [11]	110.1	68.7	76.5
	Wandt [12]	104.0	70.3	77.0
	Ours (S1)	95.2	57.3	79.3
	Ours (S1+S2)	**94.6**	**56.5**	81.9

Table 5. Evaluation results on the SkiPose. NMPJPE and PMPJPE are given in *mm*, N-PCK is in %. The best results are marked in bold.

Supervision	Method	**NMPJPE ↓**	**PMPJPE ↓**	**N-PCK ↑**
Weak	Rhodin [53]	**85.0**	-	**72.7**
Self	Wandt [12]	128.1	89.6	67.1
	Ours (S1)	118.2	79.3	70.1
	Ours (S1+S2)	115.2	**78.8**	72.4

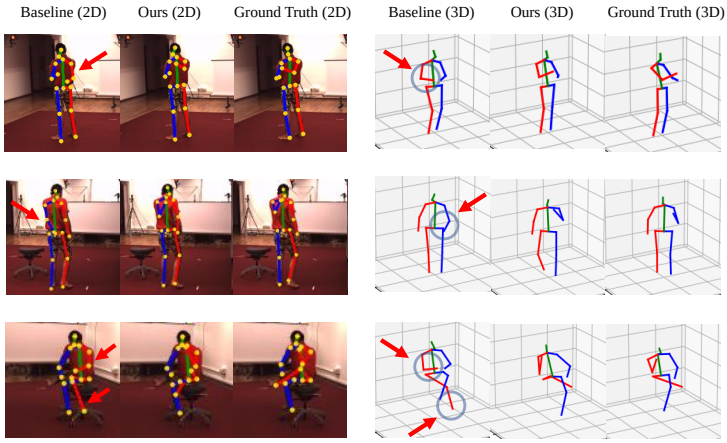


Fig. 5. Qualitative results of our approach on Human3.6M. We present both skeletons of 2D pose on the image and 3D pose in the space by comparing the visual results of baseline, ours, and ground-truth.

constraints. This results in the incorrect 3D pose and rotation estimation. The MPJPE can be considerably enhanced by refining the incorrect 2D pose of any multi-views in training to meet the multi-view consistency constraints.

Results on Human3.6M. A 2D skeleton morphing network introduced by Wandt et al. [12] is used to circumvent the offset between the 2D pose from [46] and the ground-truth 2D pose in the H36M dataset. As illustrated in Table 3, we report the self-supervised pose estimation results in terms of the NMPJPE and PMPJPE. As can be seen, the proposed model outperforms every other comparable approach in terms of the aforementioned metrics. Notably, the achieved performance surpassed our baseline, CanonPose [12], that outperforms the fully supervised method of Martinez et al. [34], which has a lifting network. In Fig. 5, we present some challenging examples on the H36M dataset and qualitatively compare the visualization results. These pictures include some occlusions and show the results of our baseline, ours, and ground-truth.

Our baseline model is already able to output plausible results. However, it does not solve occlusion, so we can visually confirm that an incorrect 2D pose is lifted to an incorrect 3D pose. We demonstrate that our approach solves the problem of occlusion that was not solved in the baseline approach.

Results on MPI-INF-3DHP. We evaluate the proposed approach on the 3DHP dataset [15] following the self-supervised protocols and metrics. The results are presented in Table 4. For a more comprehensive comparison, we report the performance of several recently fully and weakly-supervised methods. The proposed model outperforms all other self-supervised methods. In addition, the visualization results for the test dataset are presented in Fig. 6. Our model yields satisfactory results even for some dynamic action and unseen outdoor scenes.

Results on SkiPose. Our primary motivation is to train 3D human pose es-

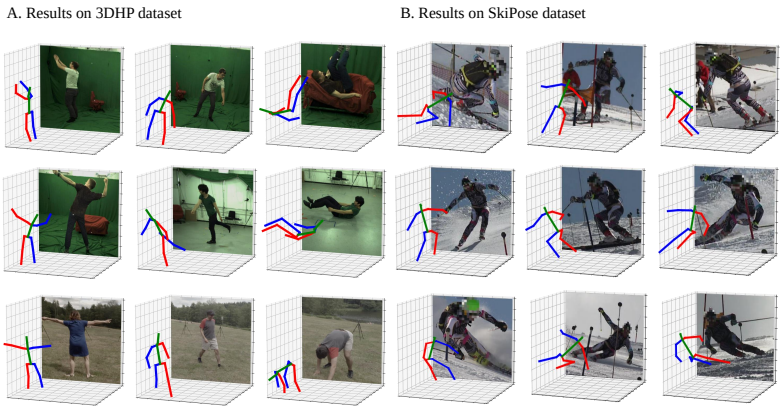


Fig. 6. Qualitative results on the MPI-INF-3DHP dataset and the SkiPose dataset.

timation in-the-wild with multiple uncalibrated cameras. Moreover, we intend to experiment on in-the-wild human activity data that cannot measure 3D annotation with traditional motion capture systems. The dataset, which is best suited to these conditions, is the SkiPose [53, 54]. Our approach can handle all these challenges since it operates without relying on static or calibrated cameras. Table 5 shows our results in comparison to Rhodin et al. [53] and Wandt et al. [12]. Rhodin et al. [53] considers a weakly supervised setting and known camera locations, so direct comparison with ours is impossible. We outperform the baseline approach [12] on the SkiPose and the qualitative results for the dataset are presented in Fig. 6.

5 Conclusion

In this paper, we introduced a novel self-supervised learning method for monocular 3D human pose estimation from unlabeled multi-view images without camera calibration. We exploited multi-view consistency to disentangle 2D estimations into canonical predictions (a 3D pose and camera rotation) that were used to refine the errors of the 2D estimations and reproject the 3D pose on the 2D for self-supervised learning. We conducted quantitative and qualitative experiments on three 3D benchmark datasets and achieve state-of-the-art results. The results demonstrated that our method could be applied to real-world scenarios, including dynamic outdoor human activities like sports.

Acknowledgements. This work was partially supported by the Institute of Information & communications Technology Planning Evaluation (IITP) funded by the Korea government(MSIT) (No. 2019-0-00079, Artificial Intelligence Graduate School Program(Korea University)) and the Technology Innovation Program (No. 20017012, Business Model Development for Golf Putting Simulator using AI Video Analysis and Coaching Service at Home) funded by the Ministry of Trade, Industry & Energy(MOTIE, Korea)

References

1. Liu, W., Mei, T.: Recent advances of monocular 2d and 3d human pose estimation: A deep learning perspective. ACM Computing Surveys (CSUR) (2022)
2. Lim, Y.K., Choi, S.H., Lee, S.W.: Text extraction in mpeg compressed video for content-based indexing. In: Proceedings 15th International Conference on Pattern Recognition. ICPR-2000. Volume 4., IEEE (2000) 409–412
3. Lee, G.H., Lee, S.W.: Uncertainty-aware human mesh recovery from video by learning part-based 3d dynamics. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 12375–12384
4. Yang, H.D., Lee, S.W.: Reconstruction of 3d human body pose from stereo image sequences based on top-down learning. Pattern Recognition **40** (2007) 3120–3131
5. Ahmad, M., Lee, S.W.: Human action recognition using multi-view image sequences. In: 7th International Conference on Automatic Face and Gesture Recognition (FGR06), IEEE (2006) 523–528
6. Roh, M.C., Shin, H.K., Lee, S.W.: View-independent human action recognition with volume motion template on single stereo camera. Pattern Recognition Letters **31** (2010) 639–647
7. Ji, X., Fang, Q., Dong, J., Shuai, Q., Jiang, W., Zhou, X.: A survey on monocular 3d human pose estimation. Virtual Reality Intelligent Hardware **2** (2020) 471– 500
8. Roh, M.C., Kim, T.Y., Park, J., Lee, S.W.: Accurate object contour tracking based on boundary edge selection. Pattern Recognition **40** (2007) 931–943
9. Kocabas, M., Karagoz, S., Akbas, E.: Self-supervised learning of 3d human pose using multi-view geometry. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 1077–1086
10. Jenni, S., Favaro, P.: Self-supervised multi-view synchronization learning for 3d pose estimation. In: Proceedings of the Asian Conference on Computer Vision. (2020)
11. Iqbal, U., Molchanov, P., Kautz, J.: Weakly-supervised 3d human pose learning via multi-view images in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2020) 5243–5252
12. Wandt, B., Rudolph, M., Zell, P., Rhodin, H., Rosenhahn, B.: Canonpose: Self-supervised monocular 3d human pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2021) 13294–13304
13. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003)
14. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. (2013) 1325–1339
15. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 2017 International Conference on 3D Vision. (2017) 506–516
16. Tekin, B., Katircioglu, I., Salzmann, M., Lepetit, V., Fua, P.: Structured prediction of 3d human pose with deep neural networks. arXiv preprint arXiv:1605.05180 (2016)
17. Tekin, B., M´arquez-Neila, P., Salzmann, M., Fua, P.: Learning to fuse 2d and 3d image cues for monocular body pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 3941–3950

18. Sun, X., Shang, J., Liang, S., Wei, Y.: Compositional human pose regression. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 2602–2611
19. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3d human pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 7025–7034
20. Rogez, G., Weinzaepfel, P., Schmid, C.: Lcr-net: Localization-classificationregression for human pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 3433–3441
21. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., Theobalt, C.: Single-shot multi-person 3d pose estimation from monocular rgb. In: 2018 International Conference on 3D Vision. (2018) 120–130
22. Yang, W., Ouyang, W., Wang, X., Ren, J., Li, H., Wang, X.: 3d human pose estimation in the wild by adversarial learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 5255–5264
23. Fang, H.S., Xu, Y., Wang, W., Liu, X., Zhu, S.C.: Learning pose grammar to encode human body configuration for 3d pose estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence. (2018)
24. Moon, G., Chang, J.Y., Lee, K.M.: Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 10133–10142
25. Wang, C., Li, J., Liu, W., Qian, C., Lu, C.: Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation. In: Proceedings of the European Conference on Computer Vision. (2020) 242–259
26. Xi, D., Podolak, I.T., Lee, S.W.: Facial component extraction and face recognition with support vector machines. In: Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition, IEEE (2002) 83–88
27. Lee, S.W., Verri, A.: Pattern Recognition with Support Vector Machines: First International Workshop, SVM 2002, Niagara Falls, Canada, August 10, 2002. Proceedings. Volume 2388. Springer (2003)
28. Lee, S.W., Kim, S.Y.: Integrated segmentation and recognition of handwritten numerals with cascade neural network. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **29** (1999) 285–290
29. Lee, S.W., Kim, J.H., Groen, F.C.: Translation-, rotation-and scale-invariant recognition of hand-drawn symbols in schematic diagrams. International Journal of Pattern Recognition and Artificial Intelligence **4** (1990) 1–25
30. Rogez, G., Schmid, C.: Mocap-guided data augmentation for 3d pose estimation in the wild. Advances in neural information processing systems (2016)
31. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 109–117
32. Cheng, Y., Yang, B., Wang, B., Yan, W., Tan, R.T.: Occlusion-aware networks for 3d human pose estimation in video. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 723–732
33. Gong, K., Zhang, J., Feng, J.: Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2021) 8575–8584
34. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 2640–2649

35. Xu, T., Takano, W.: Graph stacked hourglass networks for 3d human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2021) 16105–16114
36. Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T.J., Yuan, J., Thalmann, N.M.: Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 2272–2281
37. Ci, H., Wang, C., Ma, X., Wang, Y.: Optimizing network structure for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 2262–2271
38. Chen, T., Fang, C., Shen, X., Zhu, Y., Chen, Z., Luo, J.: Anatomy-aware 3d human pose estimation with bone-based pose decomposition. IEEE Transactions on Circuits and Systems for Video Technology **32** (2021) 198–209
39. Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z.: 3d human pose estimation with spatial and temporal transformers. In: Proceedings of the IEEE International Conference on Computer Vision. (2021) 11656–11665
40. Hossain, M.R.I., Little, J.J.: Exploiting temporal information for 3d human pose estimation. In: Proceedings of the European Conference on Computer Vision. (2018) 68–84
41. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Elgharib, M., Fua, P., Seidel, H.P., Rhodin, H., Pons-Moll, G., Theobalt, C.: Xnect: Real-time multi-person 3d motion capture with a single rgb camera. Acm Transactions On Graphics (TOG) **39** (2020) 82–1
42. Cao, X., Zhao, X.: Anatomy and geometry constrained one-stage framework for 3d human pose estimation. In: Proceedings of the Asian Conference on Computer Vision. (2020)
43. Liu, K., Zou, Z., Tang, W.: Learning global pose features in graph convolutional networks for 3d human pose estimation. In: Proceedings of the Asian Conference on Computer Vision. (2020)
44. Cao, Z., Simon, T.,Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 7291–7299
45. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 7103–7112
46. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: Rmpe: Regional multi-person pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 2334–2343
47. Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H.S., Lu, C.: Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2019) 10863–10872
48. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. (2014) 3686–3693
49. Iskakov, K., Burkov, E., Lempitsky, V., Malkov, Y.: Learnable triangulation of human pose. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 7718–7727
50. Qiu, H., Wang, C., Wang, J., Wang, N., Zeng, W.: Cross view fusion for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 4342–4351

51. He, Y., Yan, R., Fragkiadaki, K., Yu, S.I.: Epipolar transformers. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2020) 7779–7788
52. Ma, H., Chen, L., Kong, D., Wang, Z., Liu, X., Tang, H., Yan, X., Xie, Y., Lin, S.Y., Xie, X.: Transfusion: Cross-view fusion with transformer for 3d human pose estimation. Proceedings of the British Machine Vision Conference (2021)
53. Rhodin, H., Spörri, J., Katircioglu, I., Constantin, V., Meyer, F., Müller, E., Salzmann, M., Fua, P.: Learning monocular 3d human pose estimation from multi-view images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 8437–8446
54. Spörri, J.: Reasearch dedicated to sports injury prevention-the'sequence of prevention'on the example of alpine ski racing. Habilitation with Venia Docendi in Biomechanics **1** (2016) 7
55. Rhodin, H., Salzmann, M., Fua, P.: Unsupervised geometry-aware representation for 3d human pose estimation. In: Proceedings of the European Conference on Computer Vision. (2018) 750–767
56. Wandt, B., Rosenhahn, B.: Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 7782–7791
57. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 2252–2261
58. Kundu, J.N., Seth, S., Jampani, V., Rakesh, M., Babu, R.V., Chakraborty, A.: Self-supervised 3d human pose estimation via part guided novel image synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2020) 6152–6162
59. Li, Y., Li, K., Jiang, S., Zhang, Z., Huang, C., Da Xu, R.Y.: Geometry-driven self-supervised method for 3d human pose estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence. (2020) 11442–11449