2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

Uncalibrated Multi-view 3D Human Pose Estimation with Geometry Driven Attention

Victor Galizzi and Bertrand Luvison
University Paris-Saclay, CEA, List, F-91120, Palaiseau, France

***Abstract*— To make up for the inherent challenging nature of 3D pose estimation, most multi-view frameworks rely on camera calibration, often leading to impractical or constrained architectures. Accurate human pose estimation is key to enhancing human-computer interaction, gaming, health, sport and surveillance systems. By capturing precise and reliable body positions, our approach enables efficient and innovative downstream tasks. We leverage monocular 3D pose estimations and a novel geometry driven attention mechanism inside of a transformer lightweight architecture to produce high precision, occlusion aware refined 3D poses, with varying number of uncalibrated cameras. Our method shows competitive results on the in-lab dataset Human3.6M and in the in-the-wild environment of SkiPose PTZ-Camera, both in camera frames or in a disentangled person centric referential allowing practical downstream uses. Our approach matches state-of-the-art performance on Human3.6M, while being at least 3 times lighter. On the SkiPose base acquired under particularly difficult conditions, our results exceed those of the state of the art by being at least 3 times faster.**

I. Introduction

Human Pose Estimation (HPE) is a highly studied research area in the field of Computer Vision that has received significant attention in recent years. The goal of HPE is to accurately determine the position of the human body in a given image or video. This position can be represented in a variety of ways, including 2D keypoints, which define the locations of main body parts in an image, or 3D points, which provide information about the body's location in a 3D space. Such information can be used as inputs for numerous purposes, making HPE a starting step of a wide range of practical applications, including human activity recognition, augmented reality and motion analysis. An example of important field of application is sport. In elite sport, precise movements are performed at often very high speed and intensity, making it very challenging to monitor, analyze and correct. Relying on pose estimation allows for instance to compute 3D spatial information during elite sprinting, as shown in Fig. 1. without having to use motion capture system that would impair the athlete performance or hard-to-use calibrated systems. In any cases, for the method to be reliable, high precision is needed and must generalize well between different view angles, camera setup and not be too sensitive to occlusion.

The simplest approach is the monocular framework, where a single RGB camera is used to compute the 3D position of the body. Although being flexible and simple to use, a single image is often not accurate enough to capture full information about the space arrangement of the human body. Indeed, HPE is a ill-posed problem by nature, given the depth ambiguities. Occlusions, i.e. when a part of the body is not visible from the camera, because of an element in the environment or hidden by the body itself, are also an important source of error. To overcome these issues, multi-view approaches is an answer.

The use of synchronized views is often a viable solution, as it can greatly improve the accuracy of the pose estimation process (cf. Fig. 2). Seeing the same subject from multiple angles allows to correct occlusion errors and reduce the uncertainty caused by 2D projection. However, these approaches often require camera calibration to fully benefit from the multi camera setup by triangulating the views. Calibration limits the ease of deployment and use, therefore practical applications of such systems. The intuition presented in this paper is that a calibration step is not necessary when dealing with multi-view human posture estimation. Indeed, given two synchronized views, the observed person is at that moment in the same pose. Depending on the point of view of each camera on the person, some components of the 3D positioning of the joints in space can be determined in a very reliable way whereas other ones are particularly difficult to determine, especially along the line of sight to the camera.

It is by relying on this complementarity according to the incidence of view on the joints, that we propose a transformer architecture for merging N estimates of imperfect 3D skeletons into a consolidated unified one. The proposed model is designed to be flexible and easy to use. There are no constraints on the required number of views, and no calibration between cameras needs to be performed. Additionally, the network operates directly on keypoints in relatively low dimension. As a consequence, the network is lightweight, easy and fast to train but also to infer with. Finally, it is highly modular since any on the shelf monocular 2D and 3D HPE models can be used upstream.

II. Related work

By analogy with 2D HPE [17], state of the art for multi-person detection can be distinguished in two different ways, either the person concept is inferred lately by a reconstruction according to its articulations (bottom-up approach) as in [3], [16], or it is inferred upstream by a prior detection of the person in its entirety (top-down approach) as in [2].

However, when dealing with 3D estimation from 2D information, the main problem is the 3D estimation itself. Thus, many methods put aside the multi-person aspect, either by

979-8-3503-9494-8/24/$31.00

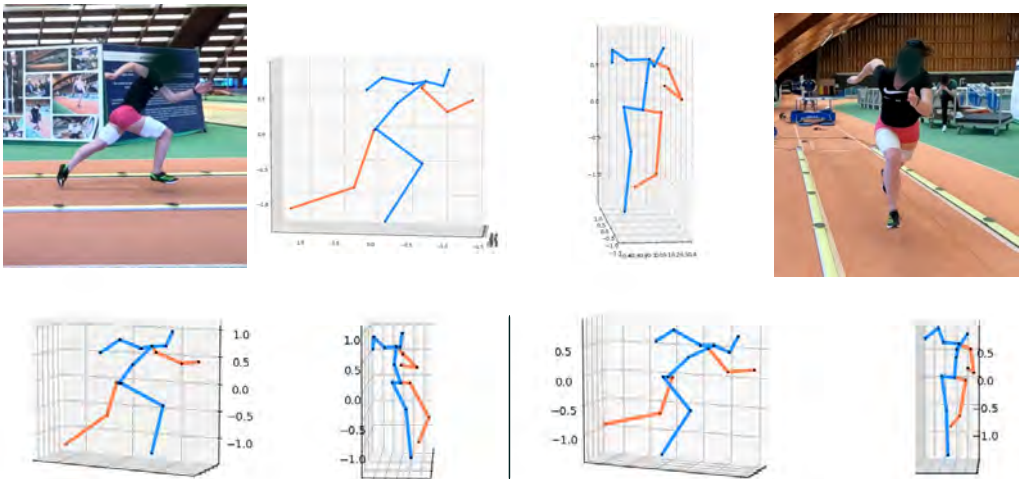

Fig. 1. Qualitative example of an In-the-wild 2 cameras scenario. On the bottom, two different angles of the refinement module estimation. On top, the aggregation module output from two different angles.

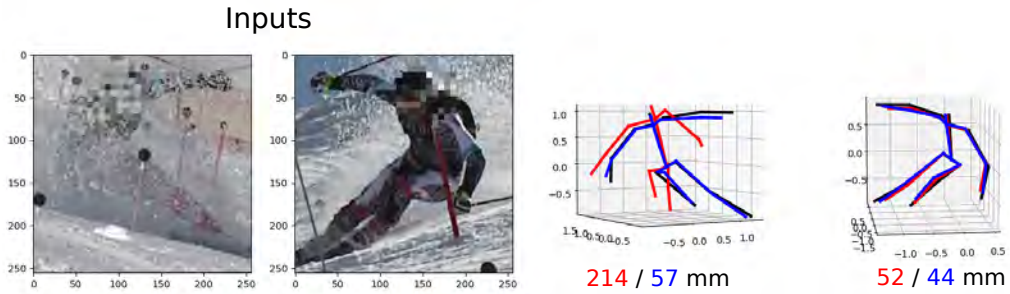

Fig. 2. Qualitative example on SkiPose. Estimation given by the monocular lifter and associated MPJPE (mm) are given in red, in both camera frames. In blue, estimation and scores after the 2-view refinement. Ground truth in black.

making the hypothesis that the processed image is adjusted and centred on the individual of interest [15], or by exploiting a top-down approach whose vocation is precisely to centre the analysis on a person after having detected it [2].

When focusing on the 3D estimate, many other works can be highlighted. Indeed, recovering 3D information using a single image is an ill-posed problem. However, deep learning methods are capable of producing satisfying results even from a single image. Most methods take advantage of excellent 2D pose estimation to infer depth information from a set of 2D keypoints [19], [21]. Other works [5], [18] tried to mitigate the problem of generalization on "in the wild" context and rare pose that are not present in public datasets generally made in lab environments. Ultimately, to reliably regress 3D position from images, more information can be added. It can be either temporal [11], [20], using multiple adjacent frames from a single camera, or spatial, using multiple views from several synchronized cameras [4], [14]. Some methods use both to obtain better performance such as [12], [6].

Concerning multi-view approaches which are at stake in this paper, to fully integrate this geometric knowledge into architectures, some work directly use calibrated cameras [16], [9], [4]. The cameras extrinsic information are used in [16] to augment the image features, using convolutions, with geometric information. In the attention formulation, camera parameters are used to sample features of the different views corresponding to the same local 3D neighborhood. Given a 2D feature map of each view, [4] create a cuboid around the person pelvis, project it into each view 2D image plane thanks to camera parameters, and use bilinear interpolation to extract 2D features at the cuboid coordinates. The view-dependent features of the cuboid are then aggregated using a softmax operation. In a similar way, [9] build a voxel feature space in which detection is made, followed by person tracking and pose estimation.

While having the best performance, the need for a precise camera calibration make theses approaches very restrictive. To alleviate these constraints, work has been made on 3D HPE without camera calibration [6], [7], [12], [14]. Geometric clues are used in different ways to replace the direct information granted by camera parameters. The fact that bone lengths and joint angles do not depend on camera positions is used by [6] to regress a single, camera independent 3D skeleton using the aggregated 2D poses of each view. The invariant bone length is also used in [7], in which

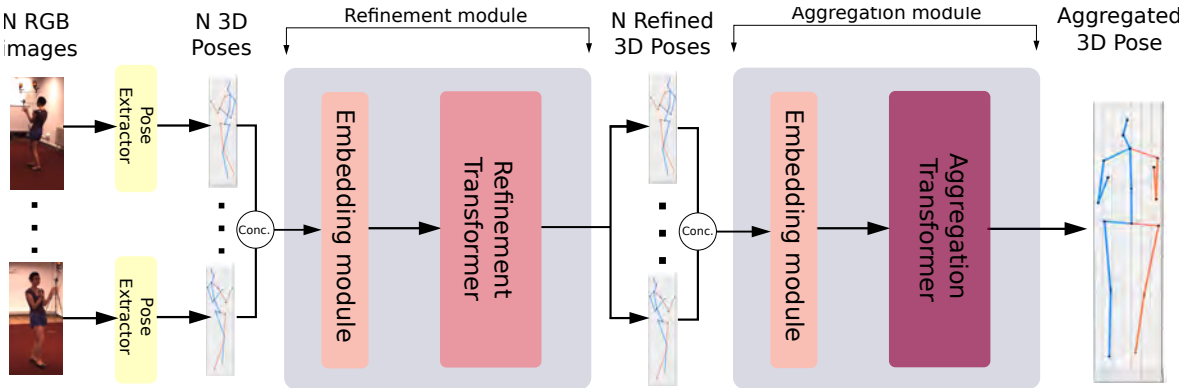

Fig. 3. The architecture overview. The refinement module shares the information between all views to refine them in their own camera frame. A second module, the aggregation module, produce a fused 3D pose in person-centric frame, freed from any camera representation.

two camera calibration matrices are estimated, then used to perform a stereoscopic triangulation on two 2D poses. Similarly, an unrestricted number of cameras relative position are estimated in [14] using rough 3D poses obtained by a monocular 3D pose estimator. The different 3D poses are then rigidly aligned using the estimated calibration, and a refined 3D pose is obtained by computing the average pose. This pose is then refined using the 2D joint location heatmaps in each view. Finally, [12] proposes a relative-attention module based on the attention mechanism adapted for the search for relative relationships in different views of the same pose.

III. Method

We present our architecture, capable of producing a high precision, occlusion aware single person human 3D pose estimation from several synchronised image without calibration. An overview of the approach is given in Fig. 3. Triangulating from two viewpoints is easily provided when the calibration between the cameras is available. Avoiding calibration implies understanding the relative position between the cameras based on the observed person's posture, which can be challenging. Our proposal is to make use of this information explicitly through an attention mechanism, leveraging a 3D estimation of the posture in each camera. Additionally, when dealing with multiple viewpoint, the choice of the final working frame is often ignored or avoided by taking an empirical convention (ex. first camera frame. To overcome this problem, we propose an extra module to merge the estimation into a single skeleton expressed in a person-centric frame.

A. *Input preparation*

*1) **3D pose extraction**:* We first employ a top-down pipeline to extract 3D skeletons from images. Given N synchronised RGB images, we use a pre-trained human detector followed by a 2D pose estimator to independently extract 2D keypoints from each images. We get $P_{2D} \in \mathbb{R}^{N\times K\times 2}$ the position on the image of the K keypoints, and $C \in \mathbb{R}^{N\times K}$ the corresponding keypoint confidence, $0 \leq c_k \leq 1$

Next, we use an off-the-shelf 3D pose lifter L on each 2D pose

$$P_{3D} = L(P_{2D}) \tag{1}$$

with $P_{3D} \in \mathbb{R}^{N\times K\times 3}$ the 3D coordinates of each of the K keypoints. The 3D position is root-relative, and is expressed in the camera frame of each of the N views.

*2) **View specific geometric confidence**:* It is intuitive that, given a specific camera, a lot of ambiguities will lie in the depth axis. We want to quantify this factor, in order to integrate this knowledge into our architecture .

Let $\mathscr{F} = (X_p, Y_p, Z_p)$ be a person-centric frame, with its origin lying on the person pelvis. The first vector X_p is pointing toward the left hip, Y_p is aligned with the spinal column. The last vector Z_p is defined so that (X_p, Y_p, Z_p) is a direct frame. In most cases Z_p will point in the direction the person is facing. This representation is showed in Fig. 4. Defining such frame disentangled from camera configuration will allow us to quantify uncertainty in the same frame for all the different cameras.

Let us now consider the two following examples, in Fig. 5, on the left example, the person has the camera on its left side. Intuitively, estimating X_p coordinates of all joints, in the person-centric frame, from this view will be very

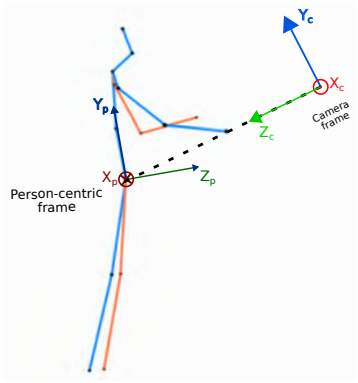

Fig. 4. Representation of the person-centric frame and the camera frame.

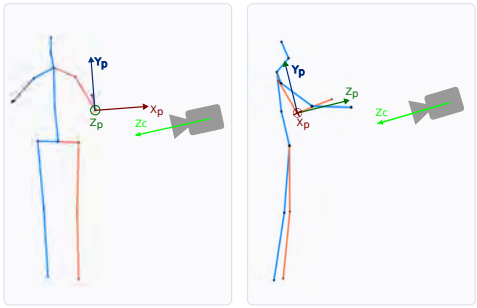

Fig. 5. On left image X_p component of the elbow cannot be well estimated from a camera on the left of the person whereas Z_p can. On the right, it is the contrary.

challenging. Notice how the camera optical axis Z_c and the X_p vector are collinear. On the right image, the person is facing the camera. This view will now allow for a good estimation of the X_p coordinates. Now, the camera optical axis and X_p are normal.

Thus, we argue the following : given a camera optical axis d, the collinearity between d and a given person-centric axis allows for a measure of the estimation capacity along this axis. This capacity being minimal when the two vectors are collinear, and maximum when orthogonal. The score quantifies the reconstruction ability for the whole pose within a particular camera angle.

Formally, let Z_c be the camera optical axis and X_p, Y_p and Z_p the person-centric frame axis in the camera frame, as shown in Fig. 4.

We define the view-specific geometric confidence c_X, c_Y, c_Z as

$$c_X = 1-|X_p.Z_c|,\ \ c_Y = 1-|Y_p.Z_c|,\ \ c_Z = 1-|Z_p.Z_c| \quad (2)$$

with $0 \leq c_X, c_Y, c_Z \leq 1$. Adding these 3 coefficients to the estimator confidence c, we obtain a score vector $S \in \mathbb{R}^{N\times K\times 4}$

B. *Embedding module*

Given the two vectors $P_{3D} \in \mathbb{R}^{N\times K\times 3}$ and $S \in \mathbb{R}^{N\times K\times 4}$, containing respectively the 3D coordinates in each of the camera frames, and the confidences scores c, c_x, c_y, c_z for each keypoints in all N views. Previous work [6], [12] showed the benefits of integrating confidences into the embedding modules. Inspired by these previous works, we design the following embedding module.

As we operate equally along all views and keypoints, the first two dimensions are omitted for the sake of clarity. Let us then consider a keypoint $p_k \in \mathbb{R}^3$ and the associated score vector $s_k \in \mathbb{R}^4$. We define two linear layers f_P, f_Q, such that

$$e_{P_k} = f_P(p_k), \quad e_{S_k} = f_S(s_k) \quad (3)$$

with $e_{P_k} \in \mathbb{R}^{D_{emb}}$, $e_{S_k} \in \mathbb{R}^{(3*D_{emb})}$ and D_{emb} the embedding dimension. The correctness can vary greatly between the different 3D input keypoints, as it depends on the image quality, camera point of view and 2D detections. We want the model to operate with this uncertainty in the embedding module.

The position embedding e_{S_k} is thus used to modulate the position p_k, added to e_{P_k} and finally combined with a final linear layer h. To add body structural information, and help the networks relate the associated keypoints in the different views, the combination is summed with a learned vector $\mathcal{L} = \{l_k \in \mathbb{R}^{\times D_{emb}} | k \in [\![1, K]\!]\}$, that acts as positional encoding. This vector is unique for each joint type.

$$x_k = h(e_{P_k} + p_k.e_{S_k}) + l_k \quad (4)$$

Operating the same way for every keypoints in every view, we get a high dimension representation $X \in \mathbb{R}^{N\times K\times D_{emb}}$ of our poses.

C. *Refinement transformer*

The goal of this module is to refine each 3D pose in its own camera frame, by sharing information between all of the different representations.

Let $M = K * N$ be the total number of keypoints on all views, we first reshape the input X into $\mathbb{R}^{M\times D_{emb}}$. Let $i \in [\![1, M]\!]$ and $x_i \in \mathbb{R}^{D_{emb}}$ a particular keypoint we want to refine using the set of keypoints $\mathcal{X} = \{x_j \mid j \in [\![1, M]\!]\}$ and the associated scores $\mathcal{S} = \{s_j \mid j \in [\![1, M]\!]\}$. Let f_q, f_k and f_v be three fully connected layers used to compute query, keys and values from the desired vectors :

$$\begin{aligned} q_i &= f_q(x_i) \\ \mathcal{K} &= \{k_j = f_k(x_j) \mid x_j \in \mathcal{X}\} \\ \mathcal{V} &= \{v_j = f_v(x_j) \mid x_j \in \mathcal{X}\} \end{aligned} \quad (5)$$

We then use two MLPs f_{qk}, f_s to get attention vectors a_{ij}

$$a_{ij} = f_{qk}(q_i - k_j) + f_s(s_i - s_j) \quad (6)$$

Attention vectors are now combined with values v_j

$$x_i^{'} = \sum_{j=1}^{M} \sigma(a_{ij}) \odot v_j \quad (7)$$

with $\odot$ the element-wise product and $\sigma(.)$ the softmax function defined by

$$\sigma(a_{ij}) = \frac{e^{a_{ij}}}{\sum_{j=1}^{M} e^{a_{ij}}} \quad (8)$$

Lastly, the embed keypoint is projected back into a three dimensional space using a final MLP , $p_i = f_{proj}(x_i^{refined})$. Proceeding the same for all keypoints, we obtain $P^{refined} \in \mathbb{R}^{N\times K\times 3}$, holding N refined 3D pose estimation, each in its own specific camera frame.

The loss for training this module is a standard MSE loss computed between $P^{refined}$ and the ground truth $\tilde{P}$ expressed in each corresponding camera frame.

TABLE I

QUANTITATIVE COMPARISON ON HUMAN3.6M OF MULTI-VIEW APPROACHES THAT DO NOT REQUIRE CAMERA CALIBRATION. MPJPE IN MM IS GIVEN FOR ALL THE DIFFERENT ACTIONS FEATURED IN THE DATASET, USING ALL 4 CAMERAS. REF. DENOTES THE REFINEMENT MODULE, AGG. THE AGGREGATION MODULE. ⋆,† INDICATE THAT 27 FRAMES WERE USED DURING TRAINING AND INFERENCE RESPECTIVELY. BEST IN **BOLD**, SECOND BEST UNDERLINED

	Dir.	Disc.	Eat.	Greet.	Phone	Photo	Pose	Purch.	Sit.	SitD.	Smoke	Wait	WalkD.	Walk.	WalkT.	Avg	Parameters
Flex ⋆,† [6]	**23.1**	28.8	26.8	28.1	31.6	37.1	25.7	31.4	36.5	39.6	35.0	29.5	35.6	26.8	26.4	30.9	70.6 M
MFT † [12]	24.2	**26.4**	**26.1**	**25.6**	**29.4**	**29.7**	**25.1**	**25.4**	**32.4**	37.4	**27.1**	**25.4**	**29.5**	**23.8**	**24.4**	**27.5**	10.1 M
MetaPose[14]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	49	3 M
Ours (Ref.)	28.7	28.7	28.8	29.0	31.0	34.7	27.4	26.0	33.1	**36.1**	30.1	27.4	31.3	26.1	27.7	29.8	**1.7 M**
Ours (Ref. + Agg.)	31.1	30.1	28.0	30.2	30.7	34.7	30.2	27.3	32.4	35.7	31.7	28.7	33.3	28.6	31.6	30.9	3.7 M

D. *Aggregation transformer*

This module answers the question of which frame should be used after the refinement module. Without any prior, no camera frame is assumed to be better than another. Of all the possible frames, those defined by the skeleton itself stand apart and allow to work with a normalized representation. We choose the skeleton frame, $\mathscr{F}$, defined in part III-A.2

The aggregation transformer operates the exact same way as the previous module, but instead of refining each views given the others, we use a learnable set of embedding $X^l \in \mathbb{R}^{K \times D_{emb}}$ that is used as queries in our transformer architecture. Hence, we will obtain a unique refined estimation instead of N previously.

In a similar fashion, let us consider the set of embed refined keypoints $\mathcal{X}^r = \{x_j^r \mid j \in [\![1, M]\!]\}$, the associated scores $\mathcal{S} = \{s_j | j \in [\![1, M]\!]\}$, (h_q, h_k, h_v) three linear layers, (h_{qk}, h_s, h_{proj}) three MLPs. We get queries q_i, keys k_j and values v_j $\forall i \in [\![1, K]\!]$, $\forall j \in [\![1, M]\!]$

$$q_i^{agg} = h_q(x_i^l), \;\; k_j^{agg} = h_k(x_j^r), \;\; v_j^{agg} = h_v(x_j^r) \tag{9}$$

$$a_{ij}^{agg} = h_{qk}(q_i^{agg} - k_j^{agg}) + h_s(s_j) \tag{10}$$

$$x_i^{l'} = \sum_{j=1}^{M} \sigma(a_j^{agg}) \odot v_j^{agg}, \quad p_i^{agg} = h_{proj}(x_i^{l'}) \tag{11}$$

We get $P^{agg} \in \mathbb{R}^{K \times 3}$ the aggregated 3D pose estimation, expressed in the skeleton frame $\mathscr{F}$.

IV. EXPERIMENTS

A. *Datasets*

Human3.6M [8] is a substantial dataset comprising a total of 3.6 million images featuring seven different individuals engaged in 15 distinct tasks, all captured by four calibrated cameras. This dataset provides both 3D and 2D pose annotations for each frame. Following established practices in prior research, we employ subjects 1, 3, 5, 6, and 7 for training purposes, reserving subjects 9 and 11 for evaluation.

Ski-Pose PTZ-Camera [13] is a rare in-the-wild multi-view dataset, capturing alpine skiers during slalom runs using six calibrated cameras. This dataset offers 3D pose annotations along with corresponding 2D projections across a collection of 20,000 images. Ski-Pose PTZ-Camera adheres to the MPII [1] joint convention. It's worth noting that the SkiPose joint convention differs from that of H3.6M, and the performance of the monocular 3D estimator used is significantly affected. Consequently, we train our monocular 3D lifter on the standard train-test split of the dataset, following the approach outlined in [14].

For both datasets, we employ the Mean Per Joint Position Error (MPJPE) metric, measured in millimeters. This metric quantifies the mean Euclidean distance between the estimated 3D joint positions and the corresponding ground truth points, with the two pelvis joints aligned to the origin for consistency.

TABLE II

EVALUATION ON SKIPOSE PTZ-CAMERA. MPJPE AFTER OPTIMAL RIGID ALIGMENT IS REPORTED FOR 2 AND 6 CAMERAS. INFERENCE TIME FOR 6 CAMERAS ON A V100 IS ALSO REPORTED. BEST IN **BOLD**.

	P-MPJPE		Δt
Camera number	6	2	
MetaPose [14]	42	50	0.4
Ours (Ref.)	41.0	53.9	**0.066**
Ours (Ref. + Agg.)	**39.4**	**48.5**	0.133

B. *Implementation details*

a) ***Architecture*** *:* We employ [10] to do human detection, [17] for 2D pose estimation and [19] for 3D monocular pose estimation. [17] use several joints labelling convention, but not Human3.6. In order to adapt its convention, we finetune [17] so that it outputs joints in the same convention used by [19]. This first part of the architecture is then frozen.

b) ***Hyperparameters*** *:* D_{emb} is equal to 128 throughout the model. The two modules, refinement and aggregation, are trained in a sequential manner. We first train the refinement module for 100 epochs, with a learning rate of 10^{-4}, divided by 10 at epoch 50 and 75. ==We found experimentaly that this set of parameters lead to the best results.== This first module is frozen, and the aggregation module is trained with the same hyperparameters, using the refined 3D poses as an input.

c) ***Data augmentation*** *:* We follow three different masking strategies, by setting selected values to $-\infty$ in attention matrices. This amounts to removing connections between certain keypoints inside the transformers. First, no masking is applied during training : all keypoints from all cameras can be used to compute all other keypoints values.

TABLE III

ARCHITECURE ABLATIONS ON H3.6M. WE REPORT THE MPJPE FOR BOTH REFINEMENT AND AGGREGATION MODULES AND VARYING NUMBER OF CAMERAS DURING INFERENCE. BEST IN **BOLD**.

	Refinement				Aggregation			
Camera number	1	2	3	4	1	2	3	4
2D base	56.9	43.1	38.3	35.9	59.2	45.8	40.5	38.5
3D base	53.6	39.6	34.4	32.1	57.1	41.7	36.2	34.0
Full model /wo positional encoding $\mathcal{L}$	53.4	41.5	37.2	35.3	57.5	43.1	38.0	35.9
Full model	**53.1**	**37.1**	**31.9**	**29.8**	**52.0**	**38.0**	**32.9**	**30.9**

Secondly, we try random elements masking with a proportion $p = 0.4$, as suggested in [12]. Lastly, we randomly mask complete views during training, by masking all keypoints related to said views. The number of masked views is uniformly drawn in $[\![0, N-1]\!]$

C. *Quantitative results*

a) ***On Human3.6M*** *:* Results of our approach and state of the art multi-view approaches that do not use camera calibration are given in Table I. Our approach competes with state of the art architectures, ranging 2mm below [12], placing second on average and have 20mm improvement on [14]. Both [12] and [6] are trained with a 27 frames windows, and use respectively 1 and 27 frames at inference. While not using any temporal information nor modelling, our method is capable of producing competitive estimation with a lighter model. We believe that leveraging 3D poses and geometry confidences closes this gap. We also report results for the aggregation module.Using this module leads to a drop of 1.1 mm in performance, which is likely caused by the change of referential, from the camera frames to the skeleton frame.

b) ***On Ski-Pose PTZ-Camera*** *:* We also perform a quantitative evaluation on Ski-Pose PTZ-Camera. This dataset is interesting for two main reasons. First, it is an in-the-wild environment, featuring challenging poses allowing to exhibit capacity of approaches outside of a lab environment. Secondly, the camera are rotating during the runs, making it a perfect example of a scenario where precise calibration is hard to acquire. As shown in Table II, our method is sensibly better than [14], while being significantly faster. Our approach performs better on both 6 and 2 camera scenario. A relevant 2-cameras refinement example on SkiPose is showed in Fig. 2. Left estimation, originally disturbed by snow occlusion, is remarkably refined thanks to the second camera. More interestingly, even the right estimation benefits from the process with a 8mm MPJPE upgrade.

D. *Qualitative results*

We also showcase three distinct qualitative examples. In Fig. 6, we feature a particularly challenging example from the H3.6M test set. Furthermore, we present in Fig. 7 an example from SkiPose captured using six cameras, effectively demonstrating the advantages of leveraging multiple viewpoints. Notably, the monocular lifter yields suboptimal estimations for the 4th and 6th views, but these are successfully refined through the fusion process. We finally show an example in-the-wild of an athlete sprinter recorded by 2 cameras, in Fig. 1. We observe that estimations are still better in axes favored by the camera angle, even after refinements. On the bottom left, elbow and knee angles are more precise, while on the other view, legs and arms placement are better. Ideally, the two estimations would be the same, except for a rotation between the two camera frames. In pratice, this is not always desirable, as some views can be occluded, noisy or more challenging. The aggregation module effectively addresses this issue by successfully capturing geometric cues, resulting in a reduction of the impact of camera angles on the estimation. In the middle, the output of the aggregation module exhibits improved articulation angles and correct body placement.

E. *Ablation study*

a) ***Architecture study*** *:* We now want to study the impact of two key components of our architecture : the direct use of 3D keypoints inside of our transformer architecture and the geometric confidence. We thus compare three different models : the 2D base, where the estimation is now done with 2D keypoints as an input. The use of geometry confidence is also removed, as it is impossible to get without a first 3D regression. Secondly, we study a 3D architecture, where the geometric confidence is removed, and the scoring vector is only made of the 2D detector confidence. Lastly, the last architecture is the full model. Results are reported in Table III.

Using 3D as an input instead of 2D leads to a gain of at least 3mm in all cameras situations, for both modules. Using the geometric confidence paired with 3D input also leads to performance gains, that increase as the number of cameras goes up. No benefits are observed when using only one camera, which is expected . When using 2 or more cameras, we observe a performance upgrade of 1mm in both modules. This validates our assumptions that geometric confidence is useful for extracting relevant information in the different views.

b) ***Encoding*** *:* We also report the benefits of using the position encoding $\mathcal{L}$, used in (4), in Table III. Compared to the full model, the gap in performance increases with the number of cameras, hence confirming that prior information

TABLE IV

MASKING ABLATIONS ON H3.6M. WE REPORT MPJPE ON THE TESTING SET FOR ALL THE DIFFERENT STRATEGIES, USING VARIOUS NUMBER OF CAMERAS DURING INFERENCE. BEST IN **BOLD**.

	Refinement				Aggregation			
Camera number	1	2	3	4	1	2	3	4
No Masking	202.0	81.6	43.9	**29.2**	172.3	62.7	38.9	30.7
Rand. ind. Masking	170.1	81.9	44.8	29.4	145.8	64.1	39.3	**30.2**
View masking	**53.1**	**37.1**	**31.9**	29.8	**52.0**	**38.0**	**32.9**	30.9

about body structure help the network relates relevant keypoints from different camera together to share information between different viewpoints.

c) ***Masking*** *:* Results for different masking strategies during training are shown in Table IV. For all strategies, the training is done with the maximum amount of cameras available, which is 4 in H3.6M. The number of camera showed is the one used during inference on the test set. View masking allows our model to generalise well on different number of cameras whereas the other masking strategies always have 4 cameras as inputs and overfit on this configuration. View masking results are overall far better in all configuration except for the overfitting case of 4 cameras where results are slightly lower.

d) ***Aggregation study*** *:* We now want to study the benefits of the aggregation mechanism and module. We thus compare the output of the aggregation module with a mean aggregation mechanism : we rotate refined skeletons into their respective skeleton frame before computing the per-joint average. This mean-aggregated skeleton is then compared to the ground truth of the test set of both datasets. We compare this score to the one obtained with the aggregation module in Table V. On H3.6M, the aggregation module does sensibly worse compared to the mean aggregation. On the contrary, it does better on SkiPose. This outcome is anticipated due to the geometric and angle similarity among all cameras, resulting in per-camera refined skeletons of comparable quality within the H3.6M dataset. On SkiPose, the camera configuration is not as uniform. Because of occlusion or challenging camera angles, the per-camera estimation is not of consistent quality throughout the views. A good example of such situation is showed in Fig. 2, where computing an aggregated pose from the average would badly hurt the final estimation, and the fusion has to be weighted toward the right skeleton. To support this intuition, we compute the per-camera MPJPE standard deviation on the test set of both dataset : for each time frame, we compute the MPJPE for each camera. We then compute the standard deviation of this error for the different views, and report the mean standard deviation of all time frames in in Table V. As expected, it is much lower on H3.6M where the camera distribution is much more regular.

TABLE V

QUANTATIVE ANALYSIS OF THE AGGREGATION MODULE. MEAN ESTIMATION DENOTES THE AVERAGE OF REFINEMENT MODULE OUPUTS (4 FOR H3.6M AND 6 FOR SKIPOSE) AFTER ALIGNEMENT IN THE PERSON-CENTRIC FRAME. MPJPE IN MILIMETERS IS GIVEN ON THE TEST SPLIT OF BOTH DATASETS

	MPJPE	
Dataset	H3.6M	SkiPose
Mean estimation	30.73	54.1
Agg. module	30.93	52.1
Camera standard deviation	3.2	17.3

V. CONCLUSION

In this article, we present a multi-view 3D pose estimation network capable of competing with state-of-the-art approaches. Our method achieves this by refining multiple monocular estimations together, all without relying on temporal information or modeling.

Instead of designing a system that requires calibrated cameras, we leverage 3D inputs, departing from the typical 2D approach. This allows us to establish a geometry-driven attention system that guides the network to establish connections between different poses and share relevant information across various viewpoints. The result is a lightweight approach that achieves state-of-the-art performance, all without the need for temporal information or complex modeling.

However, it's important to note that our approach remains sensitive to the performance of 2D and 3D monocular estimators and relies on acceptable initial estimations. Addressing this initial weakness is a priority for further research. Additionally, acquiring multi-view 3D data can be challenging. Exploring self-supervised approaches based on geometry constraints to overcome this issue presents an intriguing research direction.

Finally, we believe that the attention mechanism we propose could benefit a wide range of attention-based architectures. This is a hypothesis that we intend to verify through evaluation with the latest state-of-the-art approaches in the field.

ACKNOWLEDGEMENTS

This work was supported by the project FULGUR. The research program named FULGUR (Team: 30 researchers, Grant:1.9 MC) benefits from a French Research Agency aid (reference ANR-19-STPH-0003). This program is part of the perspective of the Paris 2024 Olympic and Paralympic Games. This publication was also made possible by the use

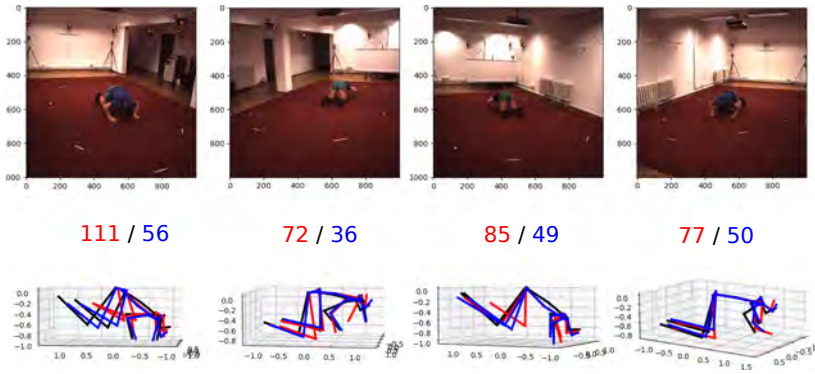

Fig. 6. 4 cameras example on H3.6M. Estimation given by the monocular lifter and associated MPJPE (mm) are given in red, for each camera frames. In blue, estimation and scores after the 4-view refinement. Ground truth in black.

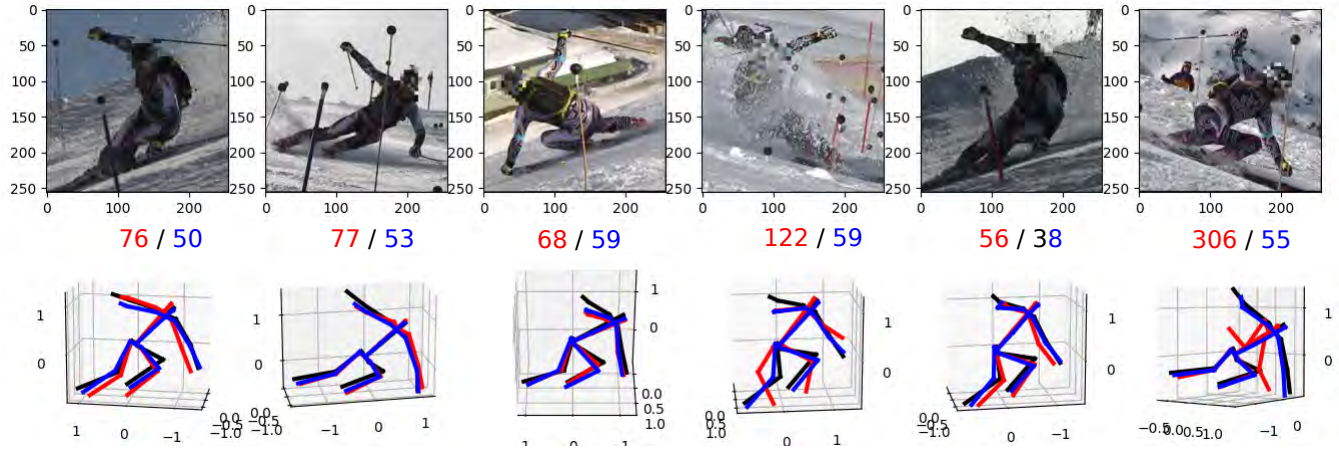

Fig. 7. 6 cameras example on SkiPose. Estimation given by the monocular lifter and associated MPJPE (mm) are given in red, for each camera frames. In blue, estimation and scores after the 6-view refinement. Ground truth in black.

of the FactoryIA supercomputer, financially supported by the Ile-de-France Regional Council.

REFERENCES

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, June 2014.

[2] A. Benzine, F. Chabot, B. Luvison, Q. C. Pham, and C. Achard. PandaNet: Anchor-Based Single-Shot Multi-Person 3D Pose Estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6855–6864, June 2020.

[3] A. Benzine, B. Luvison, Q. C. Pham, and C. Achard. Single-shot 3D multi-person pose estimation in complex images. *Pattern Recognition*, 112:107534, Apr. 2021.

[4] S. Chun, S. Park, and J. Y. Chang. Learnable Human Mesh Triangulation for 3D Human Pose and Shape Estimation. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2849–2858, Jan. 2023.

[5] M. Gholami, B. Wandt, H. Rhodin, R. Ward, and Z. J. Wang. AdaptPose: Cross-Dataset Adaptation for 3D Human Pose Estimation by Learnable Motion Generation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13065–13075, June 2022.

[6] B. Gordon, S. Raab, G. Azov, R. Giryes, and D. Cohen-Or. FLEX: Extrinsic Parameters-free Multi-view 3D Human Motion Reconstruction. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 176–196, Berlin, Heidelberg, Oct. 2022. Springer-Verlag.

[7] C. Grund, J. Tanke, and J. Gall. ElliPose: Stereoscopic 3D Human Pose Estimation by Fitting Ellipsoids. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2870–2880, Jan. 2023.

[8] C. Ionescu, D. Papava, V. Olaru, and C. 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, July 2014.

[9] N. D. Reddy, L. Guigues, L. Pishchulin, J. Eledath, and S. G. Narasimhan. TesseTrack: End-to-End Learnable Multi-Person Articulated 3D Pose Tracking. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15185–15195, June 2021.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, June 2016.

[11] W. Shan, Z. Liu, X. Zhang, S. Wang, S. Ma, and W. Gao. P-STMO: Pre-trained Spatial Temporal Many-to-One Model for 3D Human Pose Estimation. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision – ECCV 2022*, volume 13665, pages 461–478. Springer Nature Switzerland, Cham, 2022.

[12] H. Shuai, L. Wu, and Q. Liu. Adaptive Multi-view and Temporal Fusing Transformer for 3D Human Pose Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–14, 2022.

[13] J. Spörri. *Reasearch Dedicated to Sports Injury Prevention - the 'Sequence of Prevention' on the Example of Alpine Ski Racing*. PhD thesis, Oct. 2017.

[14] B. Usman, A. Tagliasacchi, K. Saenko, and A. Sud. MetaPose: Fast 3D Pose from Multiple Views without 3D Supervision. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*

(CVPR), pages 6749–6760, June 2022.

[15] B. Wandt, J. J. Little, and H. Rhodin. ElePose: Unsupervised 3D Human Pose Estimation by Predicting Camera Elevation and Learning Normalizing Flows on 2D Poses. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6625–6635, June 2022.

[16] T. Wang, J. Zhang, Y. Cai, S. Yan, and J. Feng. Direct Multi-view Multi-person 3D Pose Estimation. In *Advances in Neural Information Processing Systems*, Jan. 2022.

[17] Y. Xu, J. Zhang, Q. Zhang, and D. Tao. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. In *Advances in Neural Information Processing Systems*, Oct. 2022.

[18] C.-Y. Yang, J. Luo, L. Xia, Y. Sun, N. Qiao, K. Zhang, Z. Jiang, J.-N. Hwang, and C.-H. Kuo. CameraPose: Weakly-Supervised Monocular 3D Human Pose Estimation by Leveraging In-the-wild 2D Annotations. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2923–2932, Waikoloa, HI, USA, Jan. 2023. IEEE.

[19] A. Zeng, X. Sun, F. Huang, M. Liu, Q. Xu, and S. Lin. SRNet: Improving Generalization in 3D Human Pose Estimation with a Split-and-Recombine Approach. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, *Computer Vision – ECCV 2020*, volume 12359, pages 507–523. Springer International Publishing, Cham, 2020.

[20] J. Zhang, Z. Tu, J. Yang, Y. Chen, and J. Yuan. MixSTE: Seq2seq Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in Video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13222–13232, New Orleans, LA, USA, June 2022. IEEE.

[21] W. Zhao, W. Wang, and Y. Tian. GraFormer: Graph-oriented Transformer for 3D Pose Estimation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20406–20415, June 2022.