Evaluating 3d human pose estimation in occluded multi-sensor scenarios : dataset and annotation approach

Kévin Riou, Kaiwen Dong, Yujie Huang, Kévin Subrin, Patrick Le Callet, Yanjing Sun

EVALUATING 3D HUMAN POSE ESTIMATION IN OCCLUDED MULTI-SENSOR SCENARIOS: DATASET AND ANNOTATION APPROACH.

Kévin Riou[1,*] *, Kaiwen Dong*[2,*]*, Yujie Huang*[1]*, Kévin Subrin*[1]*, Patrick Le Callet*[1,3]*, Yanjing Sun*[2]

1. Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, Nantes, France
2. School of Information and Control Engineering, China University of Mining and Technology
3. Institut universitaire de France (IUF)

ABSTRACT

Obtaining ground truth annotations for 3D pose estimation (3D HPE) typically depends on motion capture equipment (Mocap), which is not only expensive but impractical for widespread deployment. In contrast, triangulation can reconstruct 3D poses solely from multi-view 2D poses with known camera parameters, eliminating the need for Mocap. However, inherent noise in 2D pose predictions introduces uncertainties, compromising the reliability of the results. To obtain more reliable annotations with noisy input, we introduce an annotation approach for the 3D HPE task, driven by prior knowledge of the skeletal configuration. We split our approach into two steps: first a parametric model is designed to enhance confidence predictions. Then, a differentiable weighted triangulation is employed to estimate the 3D pose in world space, leveraging the predicted confidence scores as weights. The pipeline is trained using a bone length loss. Moreover, we collect a multi-view dataset for 3D HPE and annotate it using our proposed annotation tool. This dataset is characterized by more construction scenarios, including heavier occlusion cases, diverse viewing directions, and the integration of various optical sensors, setting it apart from existing datasets. Experiments on both our dataset and Human3.6M demonstrate the effectiveness of our method.

***Index Terms*—** 3D Human Pose Estimation, Multi-view fusion, Dataset

1. INTRODUCTION

The significant progress in 3D Human Pose Estimation (3D HPE) owes much to the wealth of ground truth annotations in training data. Nevertheless, the simultaneous capture of videos and corresponding ground truth necessitates costly specialized motion capture equipment [1]. Alternatively, with calibrated camera parameters, the 3D pose can be directly triangulated from synchronized 2D poses in the image space of multiple cameras [2], albeit with a compromise in accuracy. Typical triangulation in 3D pose estimation can be divided into two stages: first estimating 2D poses in multi-view images, and then applying triangulation to derive the 3D human pose [3]. While substantial progress has been made within this pipeline, traditional triangulation methods [4] still face two constrains which hinder its application for 3D HPE: **(1) Information loss in the 2D pose estimation process; (2) Information loss of human skeleton prior in triangulation.** To alleviate **(1)**, Iskakov *et al.* [2] and Tu *et al.* [5] opt to bypass the 2D estimation step entirely, integrating heatmaps directly into a discrete volume before regressing the 3D pose. However, these works necessitate 3D ground truth for training, which is the factor we're searching for in annotation process. Pavlakos [6] introduced a method for training 3D human pose estimation without ground truth, while eliminating the need for calibration parameters during inference. However, their performance in annotation, when calibration parameters are available, still falls behind basic triangulation methods. However, these approaches are designed for individual joints, frequently overlooking the relationship among these joints. The human body, on the other hand, inherently exhibits a structured arrangement with interrelated joints, offering robust priors for 3D HPE and introducing the constraint **(2)** to the stage. Several earlier study [7] incorporate

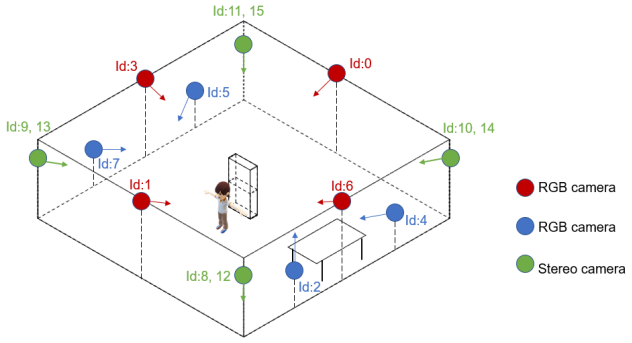


Fig. 1: **Overview of data acquisition setup.** 3D human pose estimation under : occlusions, various sensors, various optics, various viewpoints. How to balance views? How to generalize across views?

∗: Equal Contribution

prior knowledge of the human skeleton, surpassing the precision of earlier triangulation-based methods. [7] introduce a weighted triangulation-based approach, enhanced by incorporating bone length as a supervisory signal. In contrast to the prototype triangulation method [4], [7] employs confidence scores from the 2D pose estimator as weights for the Direct Linear Transform (DLT) process. While achieving impressive performance, the accuracy highly depends on the confidence estimation associated to the 2D pose prediction. However, bunch of factors, including, e.g., occlusions, distance from the cameras, variations in camera configurations (sensor and optics), may affect the reliability of the 2D pose prediction and associated confidence. To this end, we propose leveraging a parametric model to enhance the robustness of confidence scores by incorporating prior information derived from human skeletal structures. Consequently, a more dependable confidence score can be employed as the weighting factor in the triangulation process, thereby enhancing the overall performance of 3D pose estimation.

To validate our annotation approach, we present a novel multi-view 3D pose dataset called 3D-Labor. Numerous multi-view datasets have been proposed with joint position labels, such as Human3.6M[8], CMU Panoptic [9], and HumanEva [10]. However, the majority of these datasets have limitations, with few cameras and fixed viewing angles directed towards a central stage, limited occlusions, and monotonous sensor diversity. Additionally, these datasets primarily capture social interactions and basic movements such as jogging and walking. Therefore, there is a notable absence of datasets emphasizing specific scenarios, such as construction operations involving activities like shifting, lifting, hammering, etc. Accordingly, we design a unique dataset collection setup tailored specifically for the task of multi-view 3D pose estimation in construction scenarios.

Moeover, we engineered our dataset collection process to exhibit versatility across varying motion scales and optical configurations. This involved introducing a substantial number of cameras (16 cameras), with diverse optical sensors, and incorporating challenging viewing directions, as well as occluding objects in the scene.

The contributions of our method are threefold:

- We present a triangulation-based parametric model designed for annotating 3D Human Pose in multi-view datasets. This model enhances triangulation by leveraging more reliable confidence scores derived from the 2D pose input, estimated by an off-the-shelf estimator. It capitalizes on constraints provided by the human skeletal prior to improve accuracy.
- We introduce a novel 3D HPE dataset focused on construction scenarios, employing our annotation approach for labeling. Distinguished by heavy occlusions, a substantial number of views, challenging viewing directions, and the integration of diverse optical sensors, our dataset stands out from others.
- Extensive experiments on our dataset and Human3.6M demonstrate that our proposed annotation method significantly enhances traditional triangulation in multi-view 3D HPE.

Code for our annotation tools and link to data can be found at: https://github.com/KevinRiou22/supervised_hall6_pose_estimation

2. WEIGHTED TRIANGULATION AND DYNAMIC WEIGHTS PREDICTON

Triangulation for 3D Human Pose Estimation, as introduced by [7], is a 2 step process that allows to recover the 3D positions of a set of J joints defining the human skeleton. The first step consists in the simultaneous detection of the 2D joints in multiple views $\{\mathbf{u}_i^j\}_{i=1}^N$. The 2D detection of a joint j in the view $i \in \{1..N\}$ corresponds to its pixel position $\mathbf{u}_i = [u_i, v_i]$ in the corresponding image. Such detection can be done using off the shelf 2D pose estimation models, such as HRNet [15]. A confidence in the detection C_i^j is also associated to each joint j, in each view i by the 2D detector.

The second step allows to lift the 2D joint detections to 3D, using Direct Linear Transform. The relationships between the 2D detections u_i, $i \in [1, N]$, the corresponding homogeneous 3D joint $\tilde{x} \in \mathbb{R}^4$, and cameras projection matrices $P_i \in \mathbb{R}^{3\times4}$, can be written as $A\tilde{x} = 0$, where

$$A = \begin{bmatrix} u_1 p_{1,3}^T - p_{1,1}^T \\ v_1 p_{1,3}^T - p_{1,2}^T \\ \vdots \\ u_N p_{N,3}^T - p_{N,1}^T \\ v_N p_{N,3}^T - p_{N,2}^T \end{bmatrix} \in \mathbb{R}^{2N\times4}$$

$\tilde{x}$ can be recovered as the unit singular vector corresponding to the smallest singular value of the Singular Value Decomposition of A. The final 3D position of the joint can be recovered by dividing the 3 first values of $\tilde{x}$ by its fourth value: $x = \frac{\tilde{x}}{(\tilde{x})_4}$

Triangulation with confidence weights The reliability of 2D detections can vary across different views. Factors such as occlusions, joints positioned outside the field of view, distance from the cameras, and variations in camera configurations (including sensor and optics) can all contribute to this variability. To address this variability, a straightforward approach involves assuming that the confidence levels associated with the 2D detections accurately represent their reliability. Subsequently, joints in each view are weighted based on their respective confidence levels, by multiplying each row of A by a weight $w_j = \left(\frac{C_1}{\|a_1^T\|}, \frac{C_1}{\|a_2^T\|}, \ldots, \frac{C_N}{\|a_{2N-1}^T\|}, \frac{C_N}{\|a_{2N}^T\|}\right)$.

Triangulation with dynamic weights Confidence levels associated with the 2D detections may not represent the optimal contribution of the corresponding joints to the triangulation, due to the high complexity of the factors that can affect

Table 1: Comparison for existing 3D Pose Estimation Datasets. Heavy occlusions, a sufficient number of views, challenging viewing directions, and the integration of diverse optical sensors distinguish our dataset from others.

Dataset	#Frames	#Cameras	#Subj	Tricky Viewing Direction	Optic sensor Variation	Environment	Characteristics/Issues
HumanEva [10]	80K	7	4	No	**Yes**	Indoor	Inconsistent body structure bettwen 3D annotation and 2D prediction
Human3.6M [8]	3.6M	4	11(5 female + 6 male)	No	No	Indoor	-
Shelf	-	3	4	No	No	Indoor	Heavy occlusion
Campus	-	2	3	No	No	Outdoor	-
HuMMan[11]	60M	11	1000	No	No	Indoor	Involving 500 actions
CMU Panoptic[9]	154M	480	Up to 8 subjects	No	No	Indoor	Various social interactions
MPI-INF-3DHP[12]	1.3M	14	8(4 female + 4 male)	No	No	Indoor& Outdoor	-
Total Capture[13]	1.892M	4	5(4 male + 1 female)	No	No	Indoor	Shift when reprojecting 3D GT to 2D
3D-Labor (Ours)	1.98M	16	5(2 male+3 female)	**Yes**	**Yes**	Indoor	Construction operations; Heavy occlusion

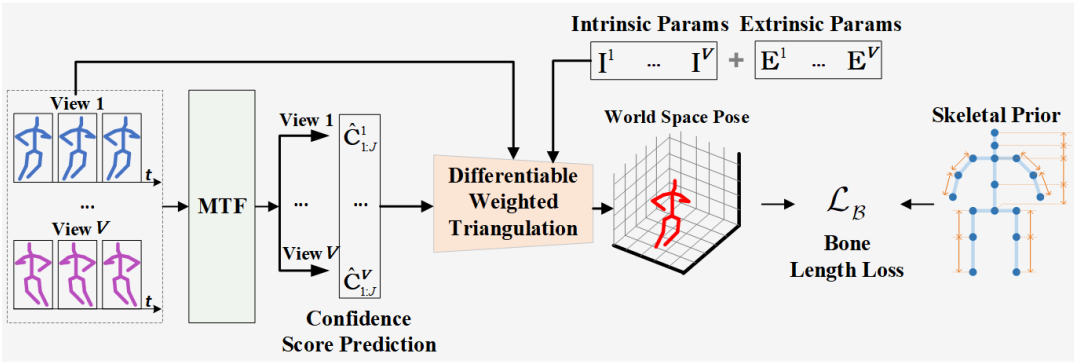


Fig. 2: **Overview of our method.** Rather than directly predicting 3D pose in world space, we employ a canonical weighted triangulation pipeline[7]. However, instead of using the weights provided by the off-the-shelf 2D pose estimator, we substitute them with predictions generated by the Adaptive Multi-view and Temporal Fusing Transformer (MTF) [14] model. Let $\hat{\mathbf{C}}^{v}_{1:J}$ denote the confidence scores for the j-th joint in the v-th view, where j ranges from 1 to J, and v ranges from 1 to V. Here, J represents the total number of joints, and V represents the total number of views. Camera parameters including Intrinsic Params ($\mathbf{I}^{1:V}$) and Extrinsic Params ($\mathbf{E}^{1:V}$) are provided by multi-view camera calibration.

the 2D detections. In this work, we propose to train a deep learning model to predict refined confidences , from the set of multi-view 2D detections and associated inital confidences :

$$\{\widehat{C}^{1:V}_{1:J}\} = MTF(\{\mathbf{u}^{1:V}_{1:J}\}, \{C^{1:V}_{1:J}\}),$$

where $\{\widehat{C}^{1:V}_{1:J}\}$, $\{\mathbf{u}^{1:V}_{1:J}\}$ and $\{C^{1:V}_{1:J}\}$ represent respectively the sets of refined confidences, 2D detections, and initial confidences; for each joints and in each view. MTF refers to the Adaptive Multi-view and Temporal Fusing Transformer (MTF) [14], originally proposed for 3D human pose prediction from multi-view 2D skeleton data. We adapted the final convolutional layer of MTF to predict a refined set of JV confidences instead of the JV 3D poses originally output by MTF. The overall pipeline is illustrated in Fig. 2. We utilize the refined confidences $\widehat{C}^{1:V}_{1:J}$ to triangulate the 3D joints and supervise the entire process using skeletal priors for subjects in the dataset. Specifically, we incorporate prior knowledge of participant bone lengths, supervising the pipeline to ensure that the lengths of the bones in the triangulated skeleton $\widehat{b}_k$ match the corresponding participant bone lengths b^{prior}_k:

$$L_B = \sum_{k=1}^{B} (\widehat{b}_k - b^{prior}_k)^2$$

3. DATA ANALYSIS

3.1. Datasets

3.1.1. 3D-Labor

We collected a new dataset for multi-view 3D human pose estimation, incorporating several factors that complicate the assessment of view reliability for the 3D reconstruction task. The setup consists of a room equipped with 16 calibrated and synchronized cameras. Occlusions in the camera views are caused by a shelf and a table placed within the scene. The setup includes a set of 4 RGB cameras (Flir GS3-U3-41C6C-C) with 12mm focal length optics positioned above the scene. These cameras, labeled as cameras 0, 1, 3, and 6 in Fig. 1, are collectively referred to as "Flir Top." Additionally, there is a set of 4 Flir cameras positioned lower and closer to the scene, equipped with optics causing vignetting that narrows the field of view. These cameras, labeled as cameras 2, 4, 5, and 7 in Fig. 1, are denoted as "Flir Bot." Furthermore, there are 4 Stereo Cameras (ZED Mini) placed around the corners of the scene, aligned with the shelf's height. Each ZED cam-

era provides two individual yet closely aligned views. With 8mm focal length optics, the ZED cameras offer wide fields of view. The cameras are labeled (12,13,14,15) and (8, 9, 10, 11) for the right and left views of the stereo cameras, respectively, and referred to as "Zed_r" and "Zed_l." We defined various multi-view scenarios comprising different combinations of these cameras, resulting in varying levels of occlusion and scene coverage. We recorded footage of 5 participants using the multi-view setup. Each participant completed 27 tasks, comprising 15 box-shifting tasks between the shelf and the table, along with 6 screwing tasks and 6 hammering tasks on the table. The cameras recorded at a rate of 30 frames per second.

We provide 2D joints annotations obtained using HRNet [15] pose estimation model trained on Coco Keypoint Dataset Format [16]. In section 3.4, we discuss triangulation approaches used to recover 3D pose ground truth from the 16 views.

3.1.2. Human3.6M

It's a widely used benchmark dataset for 3D human pose estimation [8], comprising 3.6 million 3D human poses using four synchronized cameras at 50Hz, which are organized by 11 subjects. All the 3D pose annotations for these frames are obtained using a professional motion capture system.

3.2. Metrics

Percentage of Failed Parts (PFP) A bone is deemed failed if its error exceeds half the ground truth length.

3D Bone Error on data cleaned from failed bones. Although the Percentage of Failed Parts (PFP) effectively identifies instances of substantial error in 3D reconstructions, it overlooks the accuracy of successful cases. To provide a more comprehensive assessment, we introduce the "3D Bone Error" metric. This metric computes the error between ground truth bones and reconstructed bones, excluding the extreme "failed bones." By focusing on more typical scenarios, we obtain a finer estimation of 3D error.

Ground truth bone lengths are determined by triangulating from a manually selected frame, combined with manually selected views corresponding to that frame.

Reprojection error in the best view (Reproj. Error) While metrics based on 3D Bone Error provide insight into the accuracy of the reconstructed skeleton, they may not suffice to evaluate the precise positioning of 3D joints. This is particularly relevant in scenarios where a deep learning model guides triangulation to minimize bone error, yet does not guarantee correct joint positioning. To address this limitation, a complementary metric is necessary. The reprojection loss assesses how well the 3D skeleton aligns with 2D projections. It assumes the presence of at least one view with accurate 2D detections and compares the 3D reconstructed pose, when projected into 2D space, with the input 2D poses in the best view. Concretely, the reprojection loss computes the Euclidean distance between the input 2D poses and the corresponding reprojected poses from the view that yields the lowest distance.

P-MPJPE The Procrustes-aligned Mean Per Joint Position Error (P-MPJPE) is the L2-error of the 3D estimation, calculated after applying optimal rigid alignment (shift and scale) to both the predicted 3D pose and the ground truth 3D pose.

3.3. Benchmark and Implementation Details

We train our model following a configuration similar to MTF [14] configurations (batch size, learning rate, learning decay, and dropout rate are set to 64, 1e3, 0.95, 0.1, respectively). The models are implemented with pytorch and trained using Adam Optimizer for 60 epochs on a single NVIDIA V100 GPU. Since our pipeline is designed for annotation purpose, we aim to overfit the trained model on the data that need to be annotated. Therefore, we train and test on all subjects for our data, and focus only on the training set subjects (S1, S5, S6, S7, S8) for the H36M dataset.

3.4. Comparing Annotation approaches

The Table 2 reports 3D Bone Error, Percentage of Failed Parts and Reprojection Error metrics while triangulating without weights (basic), with the 2D detector weights (w/2DConf) and with the weights predicted by our trained model (w/DynConf). The results are reported for all the camera setups defined in section 3.1.1 , resulting in various levels of occlusions.

Analyzing the Impact of Occlusions on Triangulation

The presence of occlusions poses a significant challenge to triangulation, especially in scenarios with strong occlusions, resulting in substantial errors in 3D reconstruction and subsequently higher Percentage of Failed Parts (PFP).

Using PFP as a metric allows us to explore the setups' sensitivity to strong occlusions and understand how the introduction of 2D confidences in triangulation addresses this issue.

Initially sorting setups by PFP on basic triangulation results, we observe the following ranking from best to worst: "16 views," "Zed r, Flir Top" (8 views), "Flirs Top" (4 views), Flirs Bot. (4 views), "Zed_r, Flir Bot." (8 views), "All Zeds" (8 views), "Zed_l" (4 views), "Zed_r" (4 views). Notably, the use of multi-view information is not optimized, evident in instances where triangulating solely from "Flirs Bot." yields lower PFP than combining "Flirs Bot." and "Zed_r," which inherently provides more information.

Subsequently, sorting setups by the percentage improvement in PFP when using triangulation with 2DConf compared to basic triangulation, we find the following order of improvement from best to worst: "16 views" (95.1%), "Zed_r, Flir Top" (92.9%), "Zed_r, Flir Bot." (92.7%), "All zed" (85.4%),

Table 2: Evaluation of 3D reconstruction approaches on our dataset.

	3D Bone Error (mm)							Reproj. Error (best view) (px)							PFP
	S1	S2	S3	S4	S5	Avg	Max	S1	S2	S3	S4	S5	Avg	Max	
16 views (basic)	30.3	32.2	44.5	33.6	33.2	25.9	2.14e+02	39.5	54.3	1.14e+02	54.1	68.0	68.2	1.06e+03	12.2
16 views (w/2DConf)	15.3	14.2	17.7	14.6	16.8	15.4	1.89e+02	4.7	4.6	6.3	4.5	4.6	5.0	6.83e+02	0.6
16 views (w/DynConf)	2.5	2.6	3.0	2.6	2.6	2.7	59.9	4.9	4.5	5.8	4.7	4.7	5.0	96.4	0.0
All Zeds (basic)	20.6	23.7	31.9	25.7	27.1	20.0	2.15e+02	28.7	42.8	1.06e+02	55.0	68.1	61.9	1.16e+03	17.8
All Zeds (w/2DConf)	16.3	15.1	20.2	16.1	19.0	16.3	1.92e+02	5.0	4.9	9.6	5.0	5.8	6.2	1.06e+03	2.6
All Zeds (w/DynConf)	5.1	5.1	7.1	5.2	5.8	5.6	1.92e+02	52.6	53.7	69.1	54.3	57.5	58.0	8.40e+02	0.2
Zed_r (basic)	20.5	23.6	30.5	24.9	26.4	20.3	2.15e+02	31.0	45.9	1.10e+02	57.4	72.8	65.3	1.07e+03	18.6
Zed_r (w/2DConf)	17.8	16.0	20.8	16.6	19.3	17.0	1.92e+02	7.1	6.6	10.8	6.6	7.1	7.8	1.01e+03	3.1
Zed_r (w/DynConf)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.0
Zed_l (basic)	20.9	23.7	30.3	25.2	26.4	20.3	2.15e+02	31.1	47.3	1.14e+02	58.8	72.4	66.7	9.68e+02	18.0
Zed_l (w/2DConf)	16.5	15.9	21.7	17.3	20.2	17.0	1.92e+02	6.5	6.3	11.2	7.1	7.5	7.9	1.31e+03	3.2
Zed_l (w/DynConf)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.0
Flirs Top (basic)	42.5	46.0	46.2	41.1	41.5	33.0	2.15e+02	1.35e+02	1.66e+02	2.17e+02	1.28e+02	1.64e+02	1.65e+02	1.02e+03	15.9
Flirs Top (w/2DConf)	20.9	21.9	27.5	22.5	22.4	21.1	1.92e+02	31.1	47.3	1.14e+02	58.8	72.4	66.7	9.68e+02	3.5
Flirs Top (w/DynConf)	5.9	6.2	6.8	6.7	6.6	6.4	1.75e+02	11.8	13.9	25.1	12.6	15.2	16.2	7.00e+02	0.2
Flirs Bot. (basic)	29.7	38.4	44.2	32.5	34.0	27.2	2.15e+02	76.0	1.04e+02	1.80e+02	1.09e+02	1.29e+02	1.22e+02	1.51e+03	16.3
Flirs Bot. (w/2DConf)	20.8	23.2	24.5	19.8	22.2	21.0	2.02e+02	13.1	13.8	14.8	12.3	13.1	13.5	1.26e+03	6.5
Flirs Bot. (w/DynConf)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.0
Zed_r, Flir Top (basic)	29.7	34.5	45.2	36.0	33.7	28.1	2.15e+02	54.1	66.3	1.19e+02	63.7	76.9	78.2	9.34e+02	14.1
Zed_r, Flir Top (w/2DConf)	16.7	15.1	19.7	16.2	17.8	16.5	1.91e+02	6.5	6.0	8.2	5.9	5.7	6.6	6.62e+02	1.0
Zed_r, Flir Top (w/DynConf)	3.5	3.7	4.3	3.7	3.8	3.8	1.06e+02	12.8	12.6	14.7	12.4	13.6	13.3	2.28e+02	0.0
Zed_r, Flir Bot. (basic)	26.3	30.9	45.8	30.9	32.8	21.8	2.15e+02	37.8	58.8	1.51e+02	64.6	83.5	82.5	1.25e+03	16.4
Zed_r, Flir Bot. (w/2DConf)	16.9	16.6	19.4	15.4	17.9	16.7	1.92e+02	6.8	6.5	9.0	6.3	6.5	7.1	8.19e+02	1.2
Zed_r, Flir Bot. (w/DynConf)	4.2	4.5	5.8	4.5	4.9	4.8	1.87e+02	15.1	13.6	16.2	13.7	14.0	14.6	3.51e+02	0.0

Table 3: Validation of our DynConf weight prediction approach on H36M dataset train subjects.

	P-MPJPE	3D Bone Error (mm)	PFP (%)
basic	20.78	12.10	0.0
weighted 2DConf	27.41	16.27	0.0
weighted DynConf	17.99	5.53	0.0

"Zed_r" (83.3%), "Zed_l" (82.2%), "Flir Top" (78.0%), and "Flir Bot." (60.1%). Sorting setups by triangulation with 2DConf PFP mirrors the order of PFP improvements.

It becomes evident that setups with more views and diverse perspectives benefit the most from weighted triangulation. For instance, "Zed_r, Flir Top" with diverse views outperforms "Zed_r, Flir Bot.," showcasing the advantage of incorporating diverse perspectives.

Shifting the focus to smaller occlusions, we employ 3D Bone Error on data cleaned from Failed parts as a proxy to assess sensitivity and understand how weighted triangulation influences these scenarios.

Sorting setups by 3D Bone Error on basic triangulation results, the order is as follows: "All zed," "Zed_r," "Zed_l," "Zed_r, Flir Bot.," "16 views," "Flir Bot.," "Zed_r, Flir Top," "Flir Top."

Subsequently, sorting setups by the percentage improvement in 3D Bone Error when using triangulation with 2DConf, the order from best to worst improvement is: "16 views" (40.5%), Zed r, Flir Top (41.3%), flir top (36.0%), Zed r, Flir Bot. (23.4%), flir bot (22.8%), all zeds (18.5%), zed_r (16.3%), and zed_l (16.3%).

When using triangulation with 2DConf, the setups can be sorted by 3D Bone Error, from best to worst as "16 views," "All Zeds," "Zed_r, Flir Top," "Zed_r, Flir Bot.," "Zed_r," "Zed_l, Flirs Bot.," "Flirs Top."

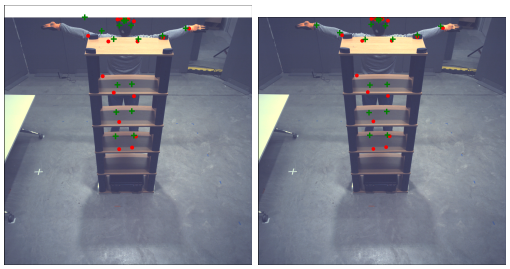
We observe a consistent pattern in the relationship between 3D Bone Error, derived from data cleaned of Failed parts, and the PFP metric. Notably, setups with a higher number of views and greater view diversity exhibit a more pronounced benefit from incorporating confidence weighting into the triangulation process.

However, integrating 2DConf in the triangulation process appears to be less significant than in the more challenging scenarios highlighted by the PFP metric. This observation suggests that the influence of 2DConf varies based on the complexity of the reconstruction task.

An intriguing insight emerges from the results presented in Table 3, which showcases the outcomes of both basic triangulation and triangulation with 2DConf on the H36M dataset—a dataset exclusively featuring self-occlusions. Despite both approaches yielding a PFP of 0.0%, utilizing triangulation with 2DConf in such scenarios leads to an increase in both 3D Bone Error and P-MPJPE. This unexpected outcome suggests that in specific instances, incorporating 2DConf in triangulation may detrimentally impact the overall triangulation performance.

Enhancing Occlusion Handling through Predicted Triangulation Weights

Because of the high amount of PFP when triangulating from few views, the models trained to predict DynConf weights from 4 views failed to converge. However, except for these 4-view setups, DynConf weights allowed to reduce the PFP to almost 0% for all the studied setups. Moreover, using



(a) 2DConf (b) DynConf

Fig. 3: Visualization of joints re-projected to 2D (green dots) following triangulation with 2DConf and DynConf weights (16 views). The red dots denote the initial 2D joint detections.

DynConf significantly reduces the 3D Bone Error (cleaned from Failed parts) compared to using 2DConf weights. The "16 views", "Zed r, Flir Top", "Zed r, Flir Bot." and "All zeds" setups reduced their 3D Bone Error respectively by 82.5%, 77.0%, 71.3% and 65.6%. We can notice that our approach also improves more the setups that showcase more views, and more diverse perspectives.

Overall, triangulating with 16 views and with DynConf weights allows 0.0% PFP and 2.7mm average 3D Bone Error on the whole dataset.

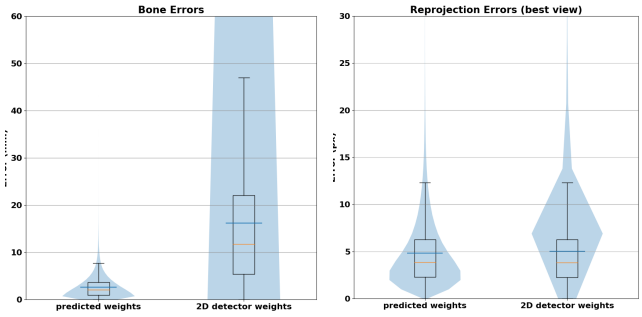
As a qualitative result, Fig. 3 highlights a triangulation that fails for the human's left wrist with 2DConf weights but succeds with DynConf weights.

On H36M dataset, as illustrated in Table 3, while using 2DConf weights decreased performances compared to basic triangulation, using the DynConf weights imporves both 3D Bone Error and P-MPJPE metrics.

Does our approach improve bone len at the expense of correct joint positioning?

As illustrated in Table 2, regarding the "16 views" scenario, using DynConf weights drastically improved the 3D bone error, while maintaining a stable re-projection error. Fig. 3 highlights the distribution of 3D bone errors and reprojection errors in the best view. Using DynConf weights squeezes most of the 3D Bone error distribution bellow 10mm, while showcasing a similar, and even slightly better Reprojection error distribution.

Interestingly, for setups that include Zed_r cameras, our model found a way to predict DynConf weights that reduce the 3D Bone error, while increasing the reprojection error. Fig. 5 highlights the re-projected joints compared to the predicted 2D joints for this case.



(a) 3D Bone Error (b) Reprojection Error

Fig. 4: Violin plot and box plot for 3D Bones and reprojection errors, when traingulating with 2DConf weights vs DynConf weights on our dataset.

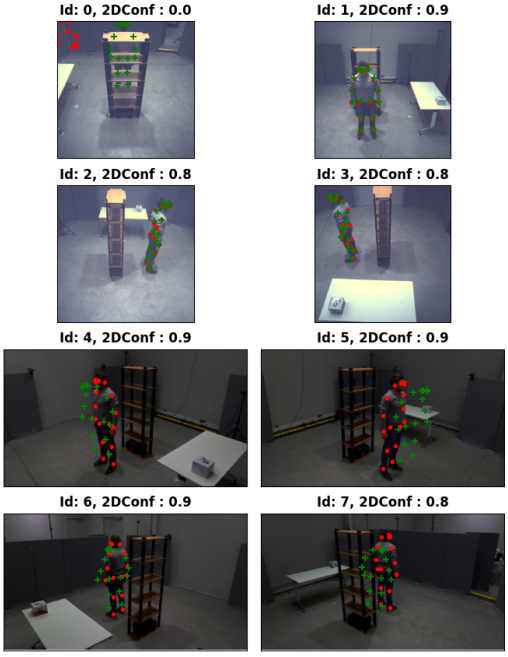


Fig. 5: Failure case with DynConf ("Zed r, Flir Top" setup on our dataset.)

4. CONCLUSION

This paper proposed a triangulation-based approach for the automatic annotation of 3D labels in the context of 3D HPE tasks, with the goal of eliminating the use of impractical Mocap equipment. Utilizing multi-view 2D poses and their corresponding confidence scores obtained from an off-the-shelf 2D pose estimator as input, our approach endeavors to learn a set of confidence scores aimed at enhancing the performance of triangulation. We introduced a novel 3D HPE dataset with a construction theme, featuring significant occlusions, a substantial number of cameras, various optical sensors, and diverse activity scales. Our annotation approach was trained using both the proposed dataset and Human3.6M, leading to significant improvements over existing triangulation-based methods.

5. ACKNOWLEDGMENTS

This work was performed using HPC resources from GENCI-IDRIS (Grant 2023-AD011014326).

6. REFERENCES

[1] Michael Gleicher, "Animation from observation: Motion capture and motion editing," *ACM SIGGRAPH Computer Graphics*, vol. 33, no. 4, pp. 51–54, 1999.

[2] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov, "Learnable triangulation of human pose," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7718–7727.

[3] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic, "3d pictorial structures revisited: Multiple human pose estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 1929–1942, 2016.

[4] Richard I Hartley and Peter Sturm, "Triangulation," *Computer vision and image understanding*, vol. 68, no. 2, pp. 146–157, 1997.

[5] Hanyue Tu, Chunyu Wang, and Wenjun Zeng, "Voxelpose: Towards multi-camera 3d human pose estimation in wild environment," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 197–212.

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

[7] Simon Bultmann and Sven Behnke, "Real-time multi-view 3d human pose estimation using semantic feedback to smart edge sensors," in *Proceedings of the Robotics: Science and Systems (RSS)*, 2021.

[8] 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*, vol. 36, no. 7, pp. 1325–1339, 2013.

[9] 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 *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3334–3342.

[10] Leonid Sigal, Alexandru O Balan, and Michael J Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *International journal of computer vision*, vol. 87, no. 1-2, pp. 4–27, 2010.

[11] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al., "Humman: Multi-modal 4d human dataset for versatile sensing and modeling," in *European Conference on Computer Vision*. Springer, 2022, pp. 557–577.

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

[13] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse, "Total capture: 3d human pose estimation fusing video and inertial sensors," in *Proceedings of 28th British Machine Vision Conference*, 2017, pp. 1–13.

[14] Hui Shuai, Lele Wu, and Qingshan Liu, "Adaptive multi-view and temporal fusing transformer for 3d human pose estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4122–4135, 2023.

[15] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.

[16] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo, "Whole-body human pose estimation in the wild," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer, 2020, pp. 196–214.