Human Pose Estimation in Trampoline Gymnastics: How to Improve Performance on Extreme Poses

Léa Drolet-Roy*[1,2], Victor Nogues*[1], Bérenger Chedal-Anglay[1], Sylvain Gaudet[2], Eve Charbonneau[3], Mickaël Begon[4], and Lama Séoud[1]

[1] Polytechnique Montreal, Canada
{lea.drolet-roy,victor.nogues}@etud.polymtl.ca
[2] Institut National du Sport du Québec, Canada
[3] University of Sherbrooke, Canada
[4] University of Montreal, Canada

Abstract. Trampoline gymnastics involves extreme human poses and uncommon viewpoints, on which state-of-the art pose estimation models tend to under-perform. We demonstrate that this problem can be addressed by fine-tuning a pose estimation model on a combination of real extreme poses and domain-specific synthetic poses (STP). We generate STP from motion capture recordings of trampoline routines. We propose a pipeline to fit noisy motion capture data to a parametric human model, then generate multi-view realistic images with high-fidelity keypoint labels. The fine-tuned ViTPose model tested on real multi-view images exhibits accuracy improvements in 2D which translate to improved 3D triangulation. In 2D, we obtain a performance similar to state-of-the-art models on the MS COCO validation set while evaluating on significantly more challenging data, bridging the performance gap between common and extreme poses. In 3D, we reduce the MPJPE by 46.1 mm with our best model, which represents an improvement of 42.7% compared to the pretrained ViTPose model. Our code and data are available at https://github.com/VisionICLab/trampoline_syn_data.

Keywords: Human Pose Estimation · Motion Capture · Synthetic Data · Trampoline gymnastics

1 Introduction

Understanding human motion is fundamental to applications ranging from sport analytics to animation. Among these scenarios, acrobatic motions such as those performed in trampoline gymnastics represent some of the most challenging forms of human movement. During trampoline routines, athletes execute combinations of somersaults and twists, performed in tucked, piked and straight

* Equal contribution.

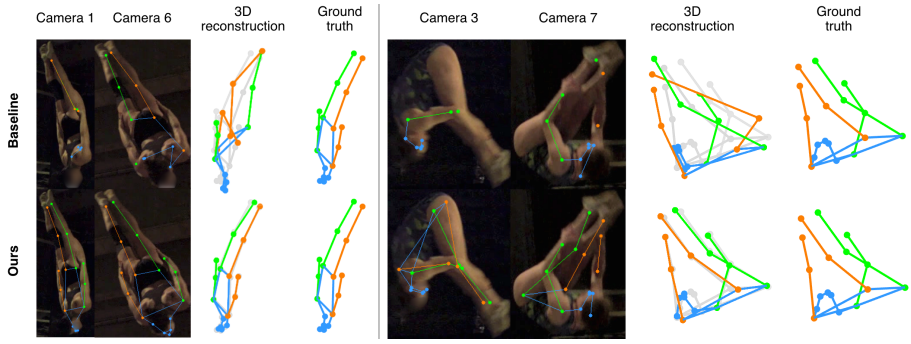

Fig. 1: Qualitative comparison of the ViTPose-S without *(Baseline)* and with fine-tuning *(Ours)* in extreme cases, 2 views out of 8 are selected. Illustrated triangulations are obtained from 8 views, using Pose2Sim [21] weighted triangulation function. Ground-truth poses are shown in gray behind the reconstructed poses.

positions. This involves atypical poses with extreme joint angles, strong self-occlusions and rapid body orientation changes leading to motion blur. These challenges are even tangible in manual pose annotation, emphasizing the complexity of the task. Accurate pose estimation in such scenarios can enable quantitative analysis of sport performance, new forms of markerless motion capture for sports and automatic scoring, like demonstrated in a pilot project by Fujitsu in artistic gymnastics [6]. However, despite recent progress in human pose estimation (HPE), current solutions remain unreliable for such extreme motions, as depicted in Fig. 1, largely because they are trained on datasets dominated by upright everyday activities [9, 13, 18]. Creating datasets that capture these challenging cases is difficult. Marker-based motion capture systems can record complex movements with high-precision, but trampoline sequences suffer from marker detachment and occlusions, making it hard to directly reconstruct consistent body motion.

In this work, we address these challenges by introducing a pipeline for improving HPE in trampoline gymnastics. The pipeline begins with marker-based motion capture recordings of trampoline routines. Because these sequences frequently suffer from markers detachment or occlusions, we propose a method for fitting a parametric human body model (from the SMPL [16] family) to noisy motion capture sequences. We then leverage the recovered motions to generate a synthetic multi-view dataset with diverse clothing, backgrounds and camera viewpoints, with accurate 2D and 3D pose annotations. This synthetic dataset is used to fine-tune a ViTPose [29] model. Finally, we evaluate 2D pose estimation and multi-view 3D pose reconstruction via triangulation on real multi-view trampoline routines recordings.

Our contributions are:

1. A new method to fit a SMPL body-model to noisy and complex motion capture data.

2. A collection of SMPL trampoline motion sequences featuring real-world extreme human poses under-represented in existing datasets such as AMASS.
3. A new end-to-end open-source tool for HPE synthetic data generation.
4. A ViTPose model fine-tuned on trampoline motion sequences and real sports data [11] that achieves SOTA HPE performance on real-world complex acrobatic data, both on 2D and triangulated 3D joints positions.

More broadly, the challenges addressed in this paper, namely markerless motion capture under fast and atypical movement, robustness to self-occlusion, and the domain gap between everyday motion training data and rare movement patterns, are shared by other real-world human motion settings. Even for clinical gait and movement analysis, standard pose estimators are similarly trained on datasets dominated by typical and unoccluded motion.

2 Related works

2.1 Human pose estimation

Most HPE pipelines decompose the problem into detecting 2D keypoints in images, optionally followed by 3D reconstruction using either monocular inference or multi-view geometry. A first category of HPE approaches relies on CNN for detecting 2D keypoints like OpenPose [4] and RTMpose [10]. The second one includes transformer-based architectures. In particular, ViTPose [29] leverages Vision Transformer backbones to achieve state-of-the-art performance on large-scale benchmarks such as COCO, demonstrating the effectiveness of large scale pretraining combined with fine-tuning strategies.

Recovering 3D human pose from multiple synchronized cameras provides a reliable alternative to monocular 3D HPE by leveraging geometric consistency across views. Independently detected 2D keypoints in each camera view are used to reconstruct 3D keypoints through geometric triangulation. Several works have demonstrated that accurate multi-view reconstruction can be achieved when high-quality 2D detections are available and cameras are properly calibrated [1, 5, 23]. However, the performance of these systems strongly depends on the robustness of underlying 2D pose estimators. In challenging scenarios such as acrobatic movements, self-occlusions and uncommon viewpoints can significantly degrade the quality of 2D keypoint predictions, thereby reducing the accuracy of the reconstructed 3D pose. These situations are rarely represented in existing datasets, motivating the development of specialized datasets and synthetic data generation strategies to improve robustness in such settings.

2.2 Human motion capture datasets

Large-scale human motion capture datasets have played a pivotal role in advancing research in HPE and motion analysis. A widely used resource is the AMASS collection [18], which aggregates motion capture recordings from multiple datasets into a unified representation using the SMPL body model. While

AMASS contains a broad range of activities, the large majority of recorded sequences correspond to everyday movements. Among the datasets included in AMASS, MoYo [26] introduces motions with larger joint angles ranges and more dynamic movements. However, even this dataset remains limited in its coverage of highly dynamic acrobatic movements involving frequent body inversion and rapid rotations. Such motions are difficult to capture reliably with traditional marker-based motion capture systems due to occlusions, marker detachment and the high-speed nature of the motion. Camera positioning is also a challenge, as we need to cover jumps up to eight meter high.

In this work, we focus specifically on acrobatic trampoline motion, a class of movements largely absent from existing motion capture collections. To address the challenges associated with capturing such motion, we propose a method for fitting a SMPL model to noisy motion capture data.

2.3 SMPL body-model fitting

Transferring human motion to animatable avatars is a widely explored field as several works have focused on recovering SMPL parameters from marker-based motion capture data. Early approaches such as MoSh [15] and MoSh++ [18] enabled large-scale unification of heterogeneous motion capture datasets resulting in the AMASS collection. More recent works, like SOMA [7] or DAMO [12], have explored learning-based approaches to improve robustness to noisy or incomplete marker data. While these approaches significantly improve automation of motion capture processing, they are typically designed for conventional scenarios and standard human motions. In contrast, acrobatic movements introduce additional challenges, including rapid motion and frequent marker detachment. Methods like MoSh [15] or SOMA [7] do not handle these cases well, produce unrealistic, distorted outputs, and are extremely slow to run on large batches of data. In this work, we address these challenges by proposing a robust GPU-accelerated SMPL fitting method, tailored to noisy motion capture data of trampoline routines.

2.4 Synthetic human pose datasets

Synthetic data generation has become common practice in HPE, with large datasets like SURREAL [27], AGORA [22] and BEDLAM [2, 25]. These works demonstrate that combining synthetic and real data improves HPE as synthetic datasets enable greater diversity through variations in body shape, appearance, motion and environments, as well as the inclusion of atypical poses and uncommon camera configurations that are difficult to capture or annotate in real-world.

However, specific contexts remain under-represented in existing datasets. For example, scenarios involving people in wheelchairs are not adequately covered in standard HPE datasets. To address such gaps, recent works such as WheelPose [8] and RePoGen [24] propose specialized synthetic datasets designed to improve HPE in specific settings. In contrast, our work focuses on generating synthetic

training data derived from real-world trampoline motion capture, enabling HPE under extreme acrobatic poses.

3 Methods

3.1 SMPL body-model fitting to motion-capture data

With athletes moving fast and experiencing large acceleration forces, the markers used for traditional motion capture tend to detach from the athletes, leading to incomplete marker data, or diverging trajectories. Our primary goal is not to match exactly a set of markers, but to target smooth, plausible, complex motion that accounts for missing or diverging markers. To retrieve realistic motion from a set of markers, we make the following prior assumptions:

- markers order remains constant over time, indicating that the acquisition system maintains consistent marker identities throughout the sequence;
- a user-defined marker layout specifying the association between markers and body segments is available a priori;
- markers attached to the same rigid body segment should maintain fixed pairwise distances over time.

Assuming these hypotheses are valid, and before starting the fitting, we automatically filter the raw marker sequences to remove unreliable trajectories. The procedure discards markers with invalid labels or high residual errors (from the motion capture system), non-moving markers, and those whose vertical dynamics deviate from the dominant jump pattern observed across markers. This last part is specific to trampoline data, and can be deactivated. It is also important to note that the user-defined marker layout is only required once per marker set, allowing us to set these correspondences once to fit ten sequences executed by three different athletes. As motion capture data collections usually employ well-defined and unchanging marker sets, our method is suitable for large-scale dataset generation.

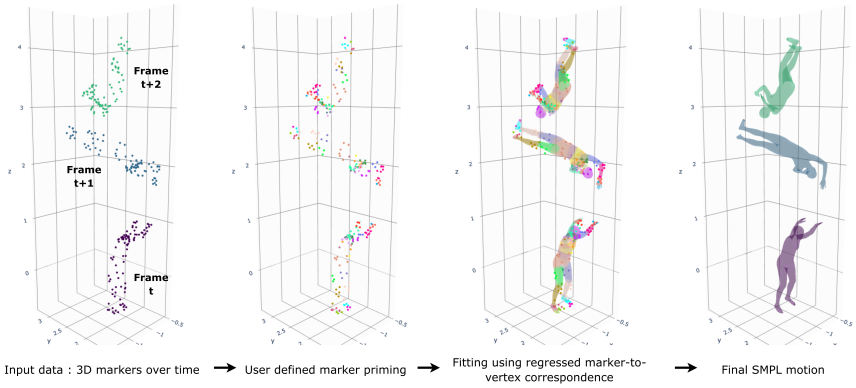

Fig. 2: Markers-to-SMPL fitting pipeline. Illustrated frames are not consecutive.

We then proceed to a rough marker-to-vertex fitting process on a few poses sampled uniformly in the target sequence. The synthetic markers, positioned 9.5 mm above the skin surface, as the real markers were, are expected to match the observed marker locations. We estimate the marker–vertex correspondence by accumulating, over a downsampled sequence, the number of frames in which each marker projects onto each vertex, and assigning the marker to the vertex with the highest support. A visual evaluation tool is provided to qualitatively determine if the markers placement is plausible. Fig. 2 illustrates the fitting process, from filtered marker coordinates to final SMPL fittings.

Once a satisfactory vertex-to-marker correspondence set is found, a marker-to-synthetic-marker distance loss is applied with temporal smoothing and Gaussian mixture model regularization, used also in SMPLify [3], to avoid converging toward unrealistic poses. The fitting phase is fully customizable, although we used a succession of four steps during our experiments:

- LBFGS [14] optimization only focused on translation and orientation,
- Adam optimization only focused on translation and orientation refinement,
- short Adam optimization phase to regress pose, shape, translation and orientation.
- LBFGS optimization to refine pose, shape, translation and orientation.

The output is a smooth, realistic SMPL sequence that closely follows the original motion, while ignoring detached markers. We compare our approach to SOMA on four sequences from three athletes. SOMA produced unrealistic results, including physically implausible joint angles, whereas our method yields smooth realistic motion and reduces computation time by 57–70%, depending on the marker layout and sequence length. Tab. 1 provides quantitative metrics.

Table 1: Marker-to-closest-vertex after fitting. SOMA uses SMPLX, whereas our method uses SMPL. Standard deviation is provided for the error. Both methods were run on the same machine equipped with an Intel Core i9-13900 CPU and an NVIDIA GeForce RTX 4090 GPU.

		Error (mm) ↓		Processing time ↓	
Subject ID	# frames	SOMA	Ours	SOMA	Ours
1	3347	41.1 ± 54.5	$\mathbf{28.1 \pm 30.2}$	42m 27s	**15m 10s**
2	2412	36.4 ± 25.9	$\mathbf{17.3 \pm 12.0}$	37m 01s	**10m 56s**
3a	5079	41.3 ± 44.7	$\mathbf{22.7 \pm 23.5}$	49m 54s	**21m 27s**
3b	4537	30.9 ± 25.3	$\mathbf{16.2 \pm 9.3}$	45m 23s	**19m 07s**

Our SMPL fitting pipeline is not specific to sport and could equally be applied to generate domain-specific fine-tuning data for other under-represented, real-world motion distributions.

3.2 TramPoseFit motion dataset

Our SMPL fitting method is applied to motion capture data of ten trampoline sequences executed by three elite athletes [28]. Each sequence contains one to five acrobatic jumps. For this data collection, a total of 95 reflective markers were placed on trampolinists. Markers positions were captured using 18 Vicon optoelectronic cameras at 200 FPS. The filtering process ended up only using a mean number of 59 ± 19 markers (about 62% of the whole 95 markers set). We refer to the resulting motion collection as TramPoseFit.

3.3 SynTramPose multi-view synthetic dataset (STP)

With TramPoseFit, we proceed to generating realistic multi-view synthetic images using Blender (Stitching Blender Foundation, Amsterdam). First, a camera combination is chosen: here, we define an eight-camera setup similar to the one used during our real-world acquisitions, illustrated on Fig. 4, and another setup with modified camera viewpoints. The cameras were arranged to ensure a wide range of viewpoints while maintaining sufficient coverage of the trampoline volume. The TramPoseFit avatar performing the motion is then placed in the center of the world reference frame.

An HDRI background is then added to the scene to ensure realistic surroundings around the avatar. Three different HDRI backgrounds (under CC0 license) are used. Their combination represents common environment challenges such as low contrast and luminosity, multiple background objects, or vibrant colors. A body texture from BEDLAM [2] is applied to the avatar. Avatars skin textures are randomly sampled and modified for each camera view. A simple clothing set is added by selecting a submesh of the SMPL avatar, and expanding this mesh to place its surface a few millimeters away from the body. Colors are then applied to the clothing submesh to distinguish it from the skin. We do not add any physics-based deformation due to the tightness of actual athletes clothing.

Fig. 3: Examples of synthetic images in STP (images are cropped for clarity)

The resulting SynTramPose (STP) dataset contains a total of 3,624 synthetic images and their corresponding 2D joint automatic annotations. Sample images are provided in Fig. 3. We cover a diversity of camera views, backgrounds, avatar skin textures and clothes, as explained in the pipeline. To avoid any form of data leakage athletes visible in SRT (validation set) or in MRT (test set) and also present in TramPoseFit are excluded from STP (ID 3). STP is then generated from six sequences executed by two athletes.

Joint annotations. We generate automatic joint annotations in the format [x, y, v] where v stands for visibility. Following COCO format guidelines [13], v has three possible values: v=0 if not visible, v=1 if labeled but occluded and v=2 if labeled and visible. We define the joint visibility first by verifying if the joint position falls within the image bounds. If not, v=0. Subsequently, we verify if the joint is occluded or visible, using the method detailed by [24].

4 Experimental setup

4.1 Implementation details

We considered the ViTPose-Small model [29] to allow comparison with prior work [24] while considering the trade-off between computational cost and performance. We use a batch size of 256 and a base learning rate of 1e-4 with the AdamW optimizer [17]. We use the OpenMMLab MMPose toolbox for conducting our experiments [19]. We consider the models pre-trained on the COCO dataset as our baseline models. Then, we fine-tune a model on extreme poses, including our STP dataset and a publicly available dataset of real images of extreme sports poses (LSPe [11]).

4.2 Datasets

Experiments are conducted using the following datasets. In all datasets, we crop the images to ground truth bounding boxes, since our goal is to evaluate only the pose estimator, excluding detection errors. Manual annotation is too time-consuming to consider annotating a dataset for model training, although it is reasonable to thoroughly annotate subsets for model validation and testing. The manual annotation method is described at the end of this section.

Leeds Sports Pose extended (LSPe) [11]. LSPe contains 10,000 real-world sports images including gymnastics and parkour, where poses resemble trampoline ones. 2D HPE annotations are gathered through Amazon Mechanical Turk. LSPe is considered in our experiments to mitigate the domain gap between synthetic and real images, without harming performance on extreme poses. It is optionally used for fine-tuning.

Singleview Real TramPose (SRT). SRT is a private manually annotated dataset of real trampoline recordings. It consists of a combination of two subsets: a) eight jumps executed by an elite trampolinist and captured with a webcam at 30 FPS (103 images) and b) one additional jump executed by the same trampolinist and captured with eight OptiTrack Prime Color cameras at 120 FPS and

downsampled at 60 FPS for annotation (746 images). The OptiTrack cameras were synchronized and calibrated. The eight-camera setup is shown on Fig. 4. The camera positions were chosen so that every point of the acquisition volume was seen by at least three cameras to facilitate the triangulation process. Constraints on the security distance around the trampoline were taken into account, which explains the limited size of athletes in the images. Cameras were calibrated using a traditional calibration method [31] and a ChArUco board, using OpenCV library [20]. After calibration, triangulating a point from an image to the world reference frame and projecting it back on the image gives a RMSE of 1.1 pixels. Data collection was approved by the Research Ethics Committee of our university and all participants signed written consents. This dataset is used only for model validation.

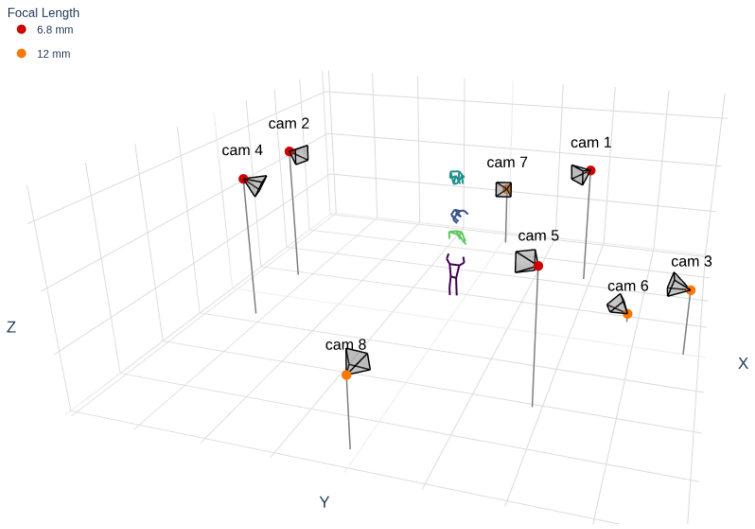

Fig. 4: Multi-view acquisition setup. A trampolinist in motion is represented as a skeleton at different positions.

Multiview Real TramPose (MRT). MRT is a private manually annotated dataset of real multi-view trampoline recordings. The RGB images were acquired using eight OptiTrack Prime Color cameras running at 120 FPS, during trampolinists training sessions, using the setup presented on Fig. 4. MRT is a combination of two subsets: MRTa corresponds to the recordings of two jumps executed by two different athletes (292 images) and MRTb corresponds to two consecutive jumps executed by the same athlete (264 images). All sequences are downsampled to 12 FPS to ensure enough pose variation between images. MRT is only used for model testing. It allows a 2D error quantification in different views as well as an extensive 3D error quantification through multi-view triangulation.

4.3 Manual annotations

Two different GUIs are used for annotating: Label Studio and MV-PoseAnnotator[1] (MVPA), a home-made tool that allows to assess multi-view consistency through triangulation and reprojection. SRT and the first part of MRT are annotated with Label Studio, while the second part of MRT is annotated using MVPA. When using MVPA, the keypoints were positioned on three selected views, triangulated, reprojected on all views, and then rectified to ensure both image-level accuracy and multi-view consistency. MVPA is helpful in ambiguous cases because the other views help the annotator understand where a given joint is. The triangulation-and-reprojection functionality makes the annotation process faster and more consistent between views.

Table 2: Summary of manually annotated datasets details

Dataset	Athlete(s) ID	# images	FPS	Annotation tool
SRTa	3	103	30	Label Studio
SRTb	3	746	60	Label Studio
MRTa	4, 5	292	12	Label Studio
MRTb	5	264	12	MVPA

4.4 2D HPE performance against baseline

We first compare the fine-tuned models with the baseline models in terms of 2D HPE performance on MRT test set, following COCO guidelines [13]. Reported average precision (AP) and average recall (AR) are OKS-based. Results are provided in Tab. 3. Baseline corresponds to COCO pre-trained ViTPose and RTMPose models, with available checkpoints. ViTPose is chosen for fine-tuning as it is the best baseline. The fine-tuning is conducted using 5,000 common images from COCO2017-val in addition to real images of uncommon sports poses (LSPe) images and/or synthetic trampoline images (STP). Individually, both new datasets demonstrate an accuracy gain compared to the baseline model, but the best performance is reached when combining both datasets. Our model achieves comparable performance to ViTPose-S on the MS COCO validation set (AP=73.8, AR=79.2 as reported in [30]) while being evaluated on significantly more challenging data. Assuming ViTPose can be considered as an established solution for standard 2D HPE, these results suggest that the fine-tuned model provides a practical solution for pose estimation in trampoline scenarios.

RePoGen [24] is included for comparison, as their model is also a ViTPose specifically trained on synthetic data of extreme poses and uncommon viewpoints. Since trampoline athletes are always in rotation and extreme poses, and

[1] Available at https://github.com/VisionICLab/MV-PoseAnnotator

Table 3: Performance of pose estimators pre-trained on COCO and fine-tuned on COCO (5,000 images) and additional datasets - testing on MRT. The results are the mean performance with 5 different seeds. (*) means that we did not train on RePoGen, but only use the provided weights [24]

Model	Fine-tuning data	N images	AP	AP 50	AR	AR 50
RTMPose-S	None (*Baseline*)	-	40.1	69.8	47.9	77.4
ViTPose-S	None (*Baseline*)	-	50.5	72.9	55.6	77.9
	STP	3,624	59.3	84.6	64.0	88.0
	LSPe	10,000	64.1	87.4	70.2	90.8
	LSPe + STP *(Ours)*	13,624	**68.4**	**88.0**	**74.5**	**91.7**
	RePoGen *	-	59.1	80.9	70.0	88.7

thus can be seen from top, bottom and side views from a single camera viewpoint, as shown on Fig. 5, we chose to train only one generic model rather than one specialized model per viewpoint, like RePoGen [24].

Fig. 5: Examples of images acquired from a same camera viewpoint showing the trampolinist from (a) bottom, (b) top and (c) side orientations. Images are cropped for clarity.

4.5 Multi-view 3D reconstruction accuracy against baseline

Real-world 3D reconstruction accuracy is evaluated on the MRT dataset. Once 2D poses are estimated in each view, the 3D poses are computed by triangulation, using the Pose2Sim pipeline [21]. The triangulation process is an iterative optimization, trying all possible combinations of given cameras until reaching a valid triangulation. The number of used cameras must stay above the minimal number of cameras, set to three for all experiments. The reprojection error

threshold to consider that a joint triangulation is valid is set to 15px, unless otherwise specified, as in the default Pose2Sim configuration file. For invalid joints, the best triangulation is kept even if the error is above the threshold to allow a fair comparison between the models.

Since the triangulation step relies primarily on camera calibration and 2D keypoints predictions, and assuming the camera calibration is accurate, this evaluation provides a meaningful robustness test for assessing the view-invariance properties of pose estimation models. In addition, it also evaluates the model's ability to accurately predict a sufficient number of joints across views to ensure complete and accurate triangulated 3D poses. This metric is more informative than COCO AP for real-world applications like sports, where the goal is to accurately retrieve a 3D pose from a minimal number of camera views.

We compare the triangulations obtained from the 2D poses estimated by the baseline ViTPose-S and by our fine-tuned version (LSPe+STP). Both predictions are compared to the ground-truth (GT) which corresponds to the triangulation of the manually annotated MRT 2D poses, using all eight views. Fig. 1 provides qualitative results both in 2D and after 3D reconstruction. The example on the left shows the importance of side disambiguation for 3D reconstruction, as the skeleton ends up twisted with the baseline model, due to inconsistencies across views. The example on the right shows missing detections in 2D leading to inaccurate 3D reconstruction.

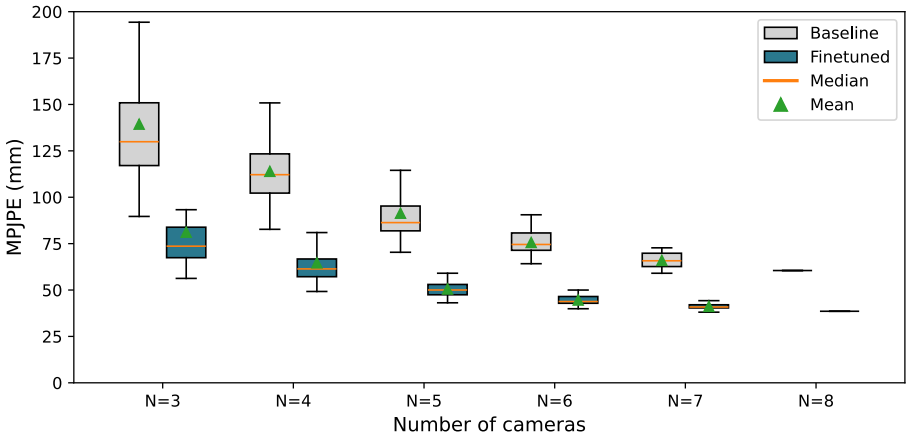

Fig. 6: MPJPE distribution among camera combinations of N cameras. MPJPE is computed on 3D triangulated poses using baseline 2D predictions *(gray)* or fine-tuned model 2D predictions *(blue)*. Outliers are represented as circles.

Fig. 6 shows the distribution of the MPJPE for all combinations of N=3 to N=8 cameras. The baseline model consistently produces larger reconstruction errors. Both mean and maximal MPJPE values are smaller with our fine-tuned

model. Overall, we observe an average reduction of 42.7% (46.1 mm) in MPJPE. The lowest improvement is 21.9 mm when using all eight cameras and the highest is 58.1 mm when using only three cameras. The improvement observed when using seven or eight cameras shows that, even with multiple predictions, the fine-tuned model leads to a more accurate 3D pose reconstruction. Although one might expect that errors would be mitigated when multiple views are combined, our results suggest that triangulation tends to propagate 2D pose errors instead of mitigating them. Using only three well-selected camera views with our fine-tuned model (MPJPE=56.3 mm) is enough to reach a better accuracy than using eight camera views and the baseline model (MPJPE=61.4 mm).

Fig. 7 presents the percentage of camera combinations leading to triangulation below MPJPE thresholds. The MPJPE of a camera combination is the average MPJPE of all its triangulated joints. It shows that, for all numbers of cameras, the triangulations from the fine-tuned model reach a lower MPJPE with more camera combinations than triangulations from the pre-trained model. It also means that, for a given MPJPE threshold, we need less cameras using the fine-tuned model and we have more liberty in the positioning of these cameras. This is particularly interesting in sport contexts, where we might be restricted in both camera number and positioning.

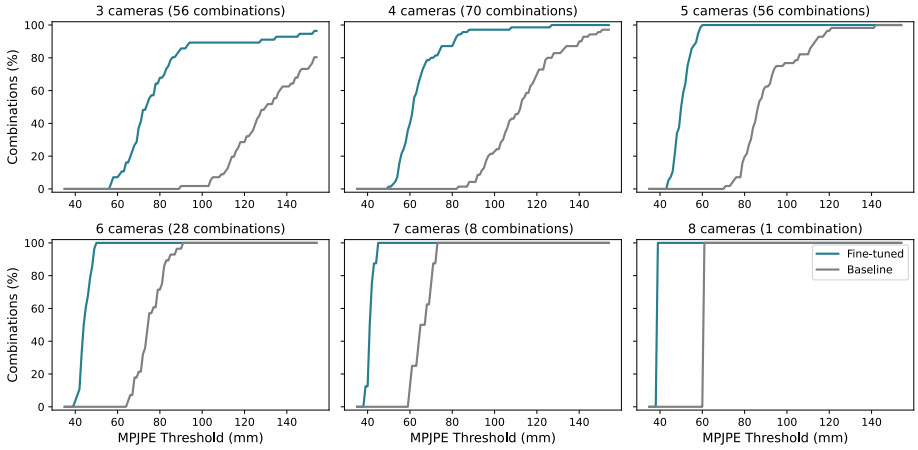

Fig. 7: Percentage of camera combinations leading to triangulation error below MPJPE. Triangulations are done using Baseline predictions *(gray)* and Fine-tuned model predictions *(blue)*. The percentage of joints is the average for all combinations of N cameras.

4.6 Limitations

The ground-truth we use in both SRT and MRT comes from manual annotations. As previously mentioned, the images are challenging: even human annotators are

faced with uncertainty for some joints. When the location of a joint could not be determined with sufficient confidence, it was left unannotated, and hence not considered in the metrics. Moreover, direct comparison with ViTPose's COCO AP is limited due to differences in the annotation protocol. In 3D, triangulated ground-truth joints are more reliable since the Pose2Sim triangulation process and the large number of annotated views act as a consistency check across camera views.

STP is a small set of synthetic images, limited by the size of TramPoseFit and the exclusion of a subject to avoid any form of data leakage. Our results (Tab. 3) reveal that fine-tuning exclusively on STP underperforms LSPe alone, which suggests either that STP's size is insufficient or that the domain gap between our synthetic and real data is too important. Our current evaluation cannot disentangle these two factors. All our fine-tuning experiments outperform RePoGen, showing that pose diversity and domain-specificity are essential.

The small scale of the MRT test set limits the generalization to other athletes, routines or sports. Moreover, as results are obtained with a ViTPose model, known for its heavy architecture, additional work is required to test whether these gains would transfer to lightweight real-time edge models suited for real-world deployment.

5 Conclusion

In this work, we address the challenge of HPE in trampoline gymnastics, an extreme benchmark where existing pipelines for motion capture and pose estimation break down, revealing gaps in current models and datasets. By introducing a realistic physics-consistent set of acrobatic motions and releasing the full toolchain used to generate and process them, we close part of this gap and enable broader research on human pose and shape estimation in challenging, under-represented regimes. Our work shows that combining synthetic and real data helps to improve HPE in a given context, but also that the fidelity of both motion and appearance is critical when operating in highly dynamic, non-standard scenarios. Our work is a proof-of-concept, showing strong in-domain performance with our fine-tuning method. We further demonstrate that these improvements in 2D predictions translate to more reliable multi-view 3D pose reconstruction in a real-world trampoline scenario.

Acknowledgements

This work was supported through Mitacs Accelerate program in partnership with Own the Podium. Léa Drolet-Roy gratefully acknowledges scholarship support from the Fonds de recherche du Québec – Nature et technologies (FRQNT). The authors thank the Institut national du sport du Québec, the trampoline athletes, as well as Youstina Daoud and Alexandre Naaim for their participation in data collection.

References

1. Bartol, K., Bojanić, D., Petković, T., Pribanić, T.: Generalizable human pose triangulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11028–11037 (2022)
2. Black, M.J., Patel, P., Tesch, J., Yang, J.: BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8726–8737. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.00843, https://ieeexplore.ieee.org/document/10205479/, 8 citations (Crossref) [2024-04-08]
3. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3d human pose and shape from a single image. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016, vol. 9909, pp. 561–578. Springer International Publishing (2016). https://doi.org/10.1007/978-3-319-46454-1_34, http://link.springer.com/10.1007/978-3-319-46454-1_34, series Title: Lecture Notes in Computer Science
4. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE transactions on pattern analysis and machine intelligence **43**(1), 172–186 (2019)
5. Dong, J., Jiang, W., Huang, Q., Bao, H., Zhou, X.: Fast and robust multi-person 3d pose estimation from multiple views. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7792–7801 (2019)
6. Fujiwara, H., Ito, K.: Ict-based judging support system for artistic gymnastics and intended new world created through 3d sensing technology. Fujitsu scientific & technical journal **54**(4), 66–72 (2018)
7. Ghorbani, N., Black, M.J.: SOMA: Solving optical marker-based MoCap automatically. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11117–11126 (2021), https://openaccess.thecvf.com/content/ICCV2021/html/Ghorbani_SOMA_Solving_Optical_Marker-Based_MoCap_Automatically_ICCV_2021_paper.html
8. Huang, W., Ghahremani, S., Pei, S., Zhang, Y.: WheelPose: Data Synthesis Techniques to Improve Pose Estimation Performance on Wheelchair Users (Apr 2024). https://doi.org/10.1145/3613904.364255, http://arxiv.org/abs/2404.17063, arXiv:2404.17063 [cs]
9. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: 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 (Jul 2014). https://doi.org/10.1109/TPAMI.2013.248, https://ieeexplore.ieee.org/document/6682899/
10. Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., Li, Y., Chen, K.: Rtmpose: Real-time multi-person pose estimation based on mmpose. arXiv preprint arXiv:2303.07399 (2023)
11. Johnson, S., Everingham, M.: Learning effective human pose estimation from inaccurate annotation. In: CVPR 2011. pp. 1465–1472 (Jun 2011). https://doi.org/10.1109/CVPR.2011.5995318, https://ieeexplore.ieee.org/document/5995318/, iSSN: 1063-6919
12. Kim, K., Seo, S., Han, D., Kang, H.: DAMO: A deep solver for arbitrary marker configuration in optical motion capture. Association for Computing Machinery **44**(1), 1–14 (2025). https://doi.org/10.1145/3695865, https://dl.acm.org/doi/10.1145/3695865

13. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft COCO: Common Objects in Context (Feb 2015). https://doi.org/10.48550/arXiv.1405.0312, http://arxiv.org/abs/1405.0312, arXiv:1405.0312 [cs]
14. Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large scale optimization. Mathematical Programming **45**(1), 503–528 (1989). https://doi.org/10.1007/BF01589116, https://doi.org/10.1007/BF01589116
15. Loper, M., Mahmood, N., Black, M.J.: MoSh: motion and shape capture from sparse markers. ACM Transactions on Graphics **33**(6), 1–13 (2024). https://doi.org/10.1145/2661229.2661273, https://dl.acm.org/doi/10.1145/2661229.2661273
16. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: a skinned multi-person linear model. ACM Transactions on Graphics **34**(6), 1–16 (2015). https://doi.org/10.1145/2816795.2818013, https://dl.acm.org/doi/10.1145/2816795.2818013, 1737 citations (Crossref) [2023-11-20]
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net (2019), https://openreview.net/forum?id=Bkg6RiCqY7
18. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.: AMASS: Archive of motion capture as surface shapes. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5441–5450. IEEE (2019). https://doi.org/10.1109/ICCV.2019.00554, https://ieeexplore.ieee.org/document/9009460/
19. MMPose Contributors: OpenMMLab Pose Estimation Toolbox and Benchmark (Aug 2020), https://github.com/open-mmlab/mmpose, original-date: 2022-04-27T01:09:19Z
20. OpenCV: OpenCV: Detection of ChArUco Boards, https://docs.opencv.org/3.4/df/d4a/tutorial_charuco_detection.html
21. Pagnon, D., Domalain, M., Reveret, L.: Pose2Sim: An open-source Python package for multiview markerless kinematics. Journal of Open Source Software **7**(77), 4362 (Sep 2022). https://doi.org/10.21105/joss.04362, https://joss.theoj.org/papers/10.21105/joss.04362
22. Patel, P., Huang, C.H.P., Tesch, J., Hoffmann, D.T., Tripathi, S., Black, M.J.: AGORA: Avatars in geography optimized for regression analysis. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13463–13473. IEEE (2021). https://doi.org/10.1109/CVPR46437.2021.01326, https://ieeexplore.ieee.org/document/9577889/
23. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Harvesting multiple views for marker-less 3d human pose annotations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6988–6997 (2017)
24. Purkrabek, M., Matas, J.: Improving 2D Human Pose Estimation in Rare Camera Views with Synthetic Data. In: 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–9 (May 2024). https://doi.org/10.1109/FG59268.2024.10582011, https://ieeexplore.ieee.org/abstract/document/10582011, iSSN: 2770-8330
25. Tesch, J., Becherini, G., Achar, P., Yiannakidis, A., Kocabas, M., Patel, P., Black, M.J.: BEDLAM2.0: Synthetic humans and cameras in motion. In: The thirty-ninth annual conference on neural information processing systems datasets and benchmarks track (2025)

26. Tripathi, S., Muller, L., Huang, C.H.P., Taheri, O., Black, M.J., Tzionas, D.: 3D Human Pose Estimation via Intuitive Physics . In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4713–4725. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2023). https://doi.org/10.1109/CVPR52729.2023.00457, https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.00457
27. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4627–4635 (2017). https://doi.org/10.1109/CVPR.2017.492, http://arxiv.org/abs/1701.01370, 486 citations (Crossref) [2023-11-19]
28. Venne, A., Bailly, F., Charbonneau, E., Dowling-Medley, J., Begon, M.: Optimal estimation of complex aerial movements using dynamic optimisation. Sports Biomechanics **22**(2), 300–315 (Feb 2023). https://doi.org/10.1080/14763141.2022.2066015, https://doi.org/10.1080/14763141.2022.2066015, _eprint: https://doi.org/10.1080/14763141.2022.2066015
29. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: ViTPose: simple vision transformer baselines for human pose estimation. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. pp. 38571–38584. NIPS '22, Curran Associates Inc. (2022)
30. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: ViTPose++: Vision Transformer for Generic Body Pose Estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(2), 1212–1230 (Feb 2024). https://doi.org/10.1109/TPAMI.2023.3330016, https://ieeexplore.ieee.org/abstract/document/10308645, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence
31. Zhang, Z.: A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(11), 1330–1334 (Nov 2000). https://doi.org/10.1109/34.888718, https://ieeexplore.ieee.org/document/888718/