Evaluating current state of monocular 3D pose models for golf

Christian Keilstrup Ingwersen[1, 2], Janus Nørtoft Jensen[1], Morten Rieger Hannemose[1], and Anders Bjorholm Dahl[1]

[1]Visual Computing, Technical University of Denmark
[2]TrackMan A/S

Abstract

Monocular 3D human pose estimation has reached an impressive performance. State-of-the-art models predict joint locations that can be accurately reprojected back into the image, resulting in visually convincing detections. However, our aim is to use the predicted poses in a domain with high-frequency movements, that is, for video of athletes performing golf swings. Our investigation is based on accurate marker-based motion capture data. Also, for our data, the predicted 3D joint locations look convincing when we reproject them into the image. However, by quantitatively comparing the results with the motion capture data, we see significant model errors that are too erroneous to be used for any kinematic analysis of the movements. Thus we conclude that the current models cannot be used out of the box for advanced golf analytics.

1 Introduction

With recent years advancements in human pose estimation, we are approaching 2D human pose estimation with accuracy scores of more than 95% (probability of correct keypoint scores [3]). This is a state where methods can be used to generate new pseudo-ground truth data [44, 39, 43, 6, 38, 4]. The issue is that 2D pose estimation is often not sufficient for further analysis of movement due to the inherent parallax error from 2D analysis [29, 37]. This parallax error means that an observed joint angle can be different from two views of the same movement and lead to conflicting conclusions.

Instead, we need 3D. 3D pose methods are also improving with state-of-the-art multi-view methods achieving mean per joint precision errors (MJPE) as low as 2 cm in controlled lab setups [35, 17, 14]. But also in-the-wild methods achieve MPJPEs below 8 cm [10, 21, 23, 40], with only a single camera view of people performing various activities that results in convincing mesh reprojections [12]. With the introduction of the 3DPW dataset [42] and the accompanying challenge, people are now starting to standardize evaluation protocols and metrics to make it easier to quantitatively compare models in a consistent manner and not just judging a model based on visual inspection.

In this paper we will investigate how state-of-the-art methods for monocular 3D human pose estimation perform on data of golf movements, i.e. fast motions that differ from the motions in the 3DPW dataset [42]. Our aim is to determine if current models can be used for kinematic analysis of golf swings. In many ways one should think that golf would be an easy scenario for the 3D human pose models, as, in contrary to many other sports, the athlete in golf is performing the motion in one place and not moving around the scene. On the other hand the swing motion is rapid, and temporal consistency is key when using the data for kinematic analysis.

In our analysis, we investigate four state-of-the-art 3D monocular pose methods – two that are frame-based and two that are sequence-based – and evaluate their performance using standard quantitative metrics. Since we are interested in the performance of the golfer, we also investigate the predicted body rotations and kinematic metrics normally used for golf motion analysis.

https://doi.org/10.7557/18.6793

2 Related work

2.1 Monocular 3D pose estimation

Methods for monocular 3D pose estimation can be separated into two main categories. One category predicts the 3D joint locations directly or regresses it from a 2D pose [45, 27, 31, 2, 34], while methods in the other category fit a parametric body model to the image [5]. The most widely used body model is the SMPL model [25]. Currently, parametric methods based on SMPL achieve state-of-the-art results on monocular 3D pose benchmarks [18, 22, 24, 8, 19, 20] and are therefore the methods we will consider in this paper.

Human Mesh Recovery (HMR) by Kanazawa et al. [18], directly regresses a reconstruction of the SMPL model from an image. This is done by first encoding the image with a ResNet-50 model [13] and from this embedding regressing the pose and shape parameters of the SMPL model, as well as the camera parameters. The model is trained to predict the 3D representation of a human to minimize the 3D error as well as the 2D reprojection error. To achieve better results, they use a discriminator that validates if the predicted shape and pose parameters are valid. They argue that this will constrain the model to only predict valid poses. A potential issue with the discriminator could, however, be that the method will struggle to adapt to data containing motions not included in the dataset the prior is trained on [33].

The **SPIN** [22] method is the first method to combine regression and optimization based models into a unified model. They combine HMR with SMPLify [5], by first regressing the initial SMPL parameters with HMR and then iteratively using SMPLify updates the model parameters by fitting the model to detected 2D keypoints in the image.

Another recent extension to the HMR method is **METRO** by Lin et al. [24]. In METRO they replace the ResNet-50 backbone with a transformer architecture that directly regresses 3D vertex and joint positions instead of first regressing SMPL parameters followed by a mapping to vertices and joints.

None of the aforementioned methods utilizes the temporal aspect of human movement and are only interested in frame-wise accurate 3D poses. With **HMMR**, Kanazawa et al. [19] introduce a video-based human 3D pose estimation method that successfully incorporates temporal information, by predicting SMPL parameters from a sequence of frames. Kocabas et al. [20] extends this approach with their method **VIBE**, which in addition to utilizing temporal information adds a discriminator on the entire motion instead of a frame-wise discriminator as in HMR [18]. Both HMMR and VIBE includes a residual layer between their spatial feature extraction layer and the temporal enconding. Choi et al. [8] shows with their **TCMR** method that by removing this residual connection, they get smoother and more temporal consistent predictions. In this analysis we will use the HMR [18], SPIN [22], VIBE [20] and TCMR [8] methods, as they represent the most established single frame methods as well as the new state-of-art performing sequential methods.

2.2 3D human pose data

Obtaining 3D human pose data is a cumbersome and time consuming process usually involving a motion capture system with either reflective markers attached to the body [41, 1], or an IMU based system [7]. Most public available datasets are captured in controlled lab environments different from the physical domains where the 3D models are expected to be deployed in real use cases. In addition to this, the datasets are limited in the amount of subjects and motions available [16, 28, 36]. Mahmood et al. [26] have with their AMASS database, tried to address these issues by combining several motion capture dataset and unified their different marker protocols into the SMPL body model [25]. However, even with this large collections of datasets it still only contains 344 subjects doing a total of 11265 motions, in controlled environments. To solve this and move 3D human pose estimation out of the lab and into realistic environments, von Marcard et al. [42] introduced the dataset 3D Poses in the Wild (3DPW). The 3DPW dataset is based on smartphone video of subjects wearing an Xsens IMU system [7]. They the use a graph-based approach to align the video data with the recorded IMU data. Since this dataset truly is captured in the wild and outside the lab it is often what is used as a benchmark for new 3D pose estimation methods. It should however be noted that their method of aligning video with IMU data is not perfect and

they report a mean joint error at 26mm [42].

3 Evaluation metrics

To evaluate 3D pose methods, mean per joint precision error, or in short MPJPE, is often used. Given a set of predicted 3D joints $\hat{x}_1, \ldots, \hat{x}_n$ and the corresponding ground truth locations for the joints $x_1, \ldots, x_n$. The error is then usually given as,

$$\frac{1}{n}\sum_{i=1}^{n} ||\hat{x}_i - x_i||_2. \tag{1}$$

From this equation it is clear that the MPJPE metric should only be able to use the absolute pose and not take scale, translation, or orientation into account. It is, however, common practice to account for differences in scale by scaling the predictions such that the mean limb length is identical to the average value of all subjects in the training set [46]. In addition to adjusting the scale, translation is also accounted for by aligning the predicted and ground truth poses by their root coordinate, which in human pose estimation is set to be the center of the hips [46].

Another often used metric in 3D human pose estimation is the reconstruction error or Procrustes aligned MPJPE (PA-MPJPE). This is a variation of the MPJPE metric that takes scale, rotation and translation into account by aligning the predictions to the ground truth with Procrustes analysis [9], i.e.,

$$\min_{\tau} \frac{1}{n}\sum_{i=1}^{n} ||\tau(\hat{x}_i) - x_i||_2. \tag{2}$$

From Equation (2) it can be seen that the only difference between the MPJPE and PA-MPJPE metric is the similarity transformation τ. From how the MPJPE metric is used with hip alignment and fixing the scale, the only difference is how the transformation is found and that the rotation is the identity matrix for the MPJPE metric.

It should be noted that the MPJPE and PA-MPJPE metrics only look at a single frame at a time and thus have a transformation for each frame in the sequence. Depending on the experiment frame-based metrics with one rigid alignment per frame can be desirable, as the metric thus only focuses on evaluating the pose and nothing else. In some scenarios, it can, however, also hide a lot of the flaws in the prediction, as it does not evaluate if the predicted poses are inconsistent through a sequence of consecutive frames or if the estimated prediction has a wrong rotation.

For this reason, we introduce a variation of the PA-MPJPE metric operating on a sequence basis instead of a per-frame basis. This means that the transformation, τ, is found using all the poses in a sequence instead of just a single pose. In our setup with a stationary camera, we believe this metric results in a more fair evaluation of the estimated joints, as it penalizes the method for being temporally inconsistent. The sequence-based metric, however, still focuses on evaluating the quality of the methods' predicted poses and not the predicted position in world coordinates. This sequence-based metric will in the results be denoted Sequence PA-MPJPE.

4 Experiments

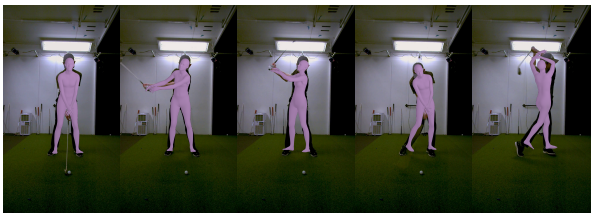
To validate the state-of-the-art methods performance on high frequency data, we have captured a small dataset of golf swings using the optical marker based motion capture system Qualisys [1], with a synchronized RGB camera. The dataset consists of four male golfers with skill levels from amateur to semi-professional and each subject has taken swings both with irons and drivers to include a diverse set of golf strokes. In total 124 swings are included in the dataset, each with a frontal camera view of the golfer. All of the videos are captured at 85 FPS totaling roughly 22 195 frames with ground truth 3D poses of the athletes.

4.1 Evaluation protocol

To evaluate the performance of the models, we use MPJPE, PA-MPJPE, and Sequence PA-MPJPE as described in Section 3. The Sequence PA-MPJPE is evaluated with one rigid alignment per swing. The PA-MPJPE and MPJPE are evaluated following the common protocols [45, 27, 31, 2, 34, 18, 22, 10]. This means the MPJPE is computed after transforming the estimated and ground-truth coordinates to the same reference coordinate system and then aligning them by their root joint, i.e. the mid-position between the left and right hip.

	MPJPE	Sequence PA-MPJPE	PA-MPJPE
VIBE^{s} [20]	144.95	123.74	106.13
HMR^{f} [18]	301.20	284.59	269.84
SPIN^{f} [22]	201.38	176.07	145.75
TCMR^{s} [8]	149.06	127.79	107.00

Table 1: Evaluation results on the golf motion capture dataset. All the metrics are presented in mm. The superscripted s indicates a sequence-based method and the superscripted f a frame-based method. Generally, it can be said that all models perform poorly compared to evaluation results on the 3DPW dataset [42]. It is however clear that the sequence-based methods, VIBE and TCMR, perform better than the frame-based methods.



(a) Address (b) Mid-backswing (c) Top-backswing (d) Impact (e) Follow through

Figure 1: Five frames from different key scenarios in the golf swing. The 3D mesh estimated by the VIBE method [20] has been overlaid on the images. The results look visually accurate with the only obvious issue being the "hand-crossings" in the follow-through.

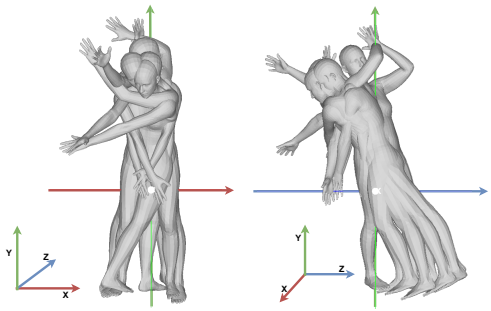
While PA-MPJPE is the frame-based rigid alignment where the alignment is found by Procrustes analysis [9].

4.2 Results

We compare four state-of-the-art methods for monocular 3D pose estimation, namely VIBE [20], TCMR [8], HMR [18] and SPIN [22] on the captured golf dataset. The evaluation results are presented in Table 1. From the results in Table 1 it is clear that all of the methods perform significantly worse than on the datasets they have been trained on, with more than double the errors they obtain on the 3DPW dataset [42]. These numbers indicate that the methods struggle to adapt to data from an unseen domain with high frequency movements. From the results we also see that the two sequence based methods VIBE [20] and TCMR [8] performs significantly better than the frame based methods.

In Figure 1 we have qualitatively evaluated the VIBE method [20] on a golf sequence. Visually, it seems like the method is performing well on a golf swing apart from the crossing hands at the end of the swing. This is a general trend, which we see in all of the four methods in all of the swings. These visual results do not align with the quantitative results in Table 1 as these numbers indicate that the methods perform poorly.

To investigate the results in depth we have rendered the predicted meshes from Figure 1 together from a frontal and a side view in Figure 2. From the renderings in Figure 2 it becomes apparent that the method struggles to correctly estimate the depth. It is no surprise that monocular models perform worse along the z-axis, but in this specific case it seems like the method simply rotates the predicted person as the athlete bends forward in the swing, which then results in a correct 2D projection but wrong 3D positions. We suspect that this behavior is encouraged by the design of optimization-based models, where they emphasize the 2D reprojection loss [20, 8, 22] which in the end is almost the same as prioritizing to minimize the loss along the x- and y-axes resulting in a visually pleasing result. Combined with the PA-MPJPE metric hiding errors related to the rotation of the predicted pose and generally a lack of ground truth 3D pose data, this makes it seems that the models are successful at adapting to unseen and different domain data as the results by visual inspection seem to be correct. This is, however, not the case when evaluated with the MPJPE metric on 3D pose data. To further investigate this for all the methods, we have computed the error along each axis. From the results in Table 2 we see that the error along the z -axis, that is, the depth component, is significantly larger than the errors along the x- and y-axes for all methods except HMR [18], which is also the only method not relying on optimization.



(a) Front view (b) Side view

Figure 2: The five meshes from Figure 1 visualized in 3D from a frontal and side view. From the frontal view, the poses seem visually plausible, but from the side view, it is clear that the estimated meshes exhibit large errors in the z-direction.

	x-error	y-error	z-error
VIBE [20]	55.0	55.6	101.1
SPIN [22]	89.9	77.8	128.9
HMR [18]	166.0	143.5	145.2
TCMR [8]	63.1	51.3	103.1

Table 2: MPJPE for each of the models divided into the error along each axis. Here it is seen that the error along the z-axis, i.e. the depth component is significantly larger than the errors along the x- and y-axis.

5 Kinematic analysis

The purpose of this paper is to analyze the state of current monocular 3D pose models and to see if they have reached a state in which they can be used for kinematic analysis of a golf swing. With the accuracy found in Table 1 and Figure 2, it has become clear that the models cannot be used for analyses where accurate world coordinates are required.

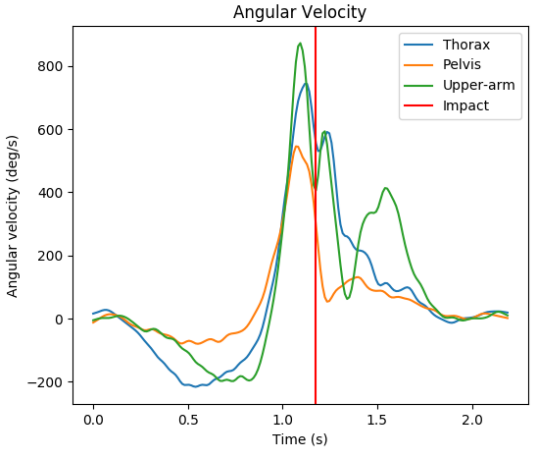
However, for analysis of the golf swing global coordinates often are not needed as the researchers and coaches are more interested in the kinematics of a swing. The kinematics of a swing is usually presented as relative rotations computed in local coordinate systems of the relevant body segments [32, 11, 30, 15]. This implies that the predicted 3D joint locations needed for a kinematic analysis do not need to be precise in the global position and orientation, but they do still need to be predicted with a consistent rotation.

To investigate whether or not the models can be used for analysis in the local coordinate systems of the body segment, we have taken the best-case model output, i.e., the swing with the lowest PA-MPJPE, and used the predicted 3D joint locations to compute the kinematic sequence of the swing. For the sake of this analysis, we will only focus on the kinematic sequence of the golf swing, which is an important indicator for the relative timings of the main body rotations in the golf swing. An ideal kinematic sequence will allow golfers to hit more consistent swings and achieve higher club speed. The relevant rotations in the kinematic sequence are the rotation of the pelvis, thorax, and lead arm, which for a right-handed golfer will be the left arm.

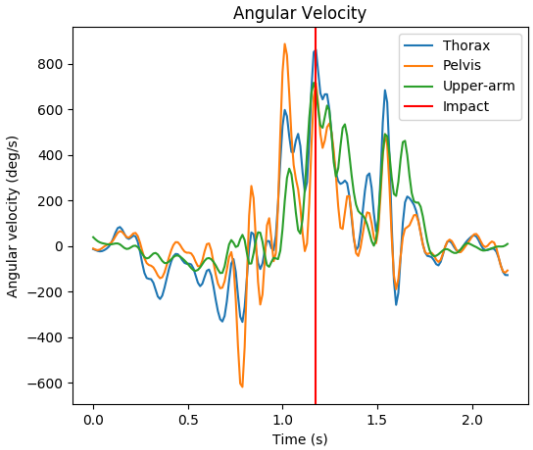
From Figure 3a we see a clear trend with the arm, thorax and pelvis rotations all peaking about 120 ms before impacting the golf ball. This is close to an ideal golf swing, which allows optimal transfer of the body rotational forces into the club head. If we instead consider the kinematic sequence computed from the predicted joints in Figure 3b it indicates that the pelvic rotation peaks close to 120 ms before impact with the golf ball, while the arm and thorax rotation have its peaks exactly at impact. In a coaching scenario, this would lead to a wrong conclusion, as it would suggest that the golfers timings are off and that he is not fully successful in transferring the body rotation velocities to the golf club. Based on this analysis of Figure 3a and Figure 3b it is clear that the predicted signal is too noisy to be used for swing analysis.

6 Recommendations

In this section we will present recommendations for future methods to address some of the issues found with existing state-of-the-art methods in this research, in the hope that future methods can reach a state where they can be used for sports analytics. In Figure 2, we saw a clear trend that the models struggle with accurate depth predictions while performing significantly better along the x- and y-axes. Especially the optimization-based methods show significantly better performance along the x- and y-axes compared to the depth.



(a) Kinematic sequence from ground truth 3D joint locations.



(b) Kinematic sequence from predicted 3D joint locations.

Figure 3: Kinematic sequence of a golf swing computed from (a) ground truth 3D joint locations, and (b) 3D estimates from the VIBE model [20]. The swing has been chosen as the swing with the lowest PA-MPJPE, i.e. the best prediction. Comparing (a) to (b) it is clear that the sequence based on predictions cant be used for analysis.

We believe that this is caused by models moving towards more 2D supervision instead of 3D supervision. The argument for moving towards 2D supervision is that accurate 3D data are scarce and expensive to obtain while pseudo-ground truth 2D data are easy to obtain. Several models also show that they achieve higher accuracy by including 2D supervision [18, 24, 22], while also showing better performance even on out-of-domain data. We believe that 2D supervision is beneficial to include but one should be careful not to overfit to the 2D reprojected data and thus loose depth information. Our recommendation is to continue to use 2D data, but to put more emphasis on 3D supervision and obtain 3D data for the intended domain.

We also see a trend that more methods include a discriminator to classify whether or not a predicted pose or pose sequence is likely to be a real human pose [33, 18, 22, 20]. This is an interesting approach, which especially for sports analysis will be able to constrain the method based on known information about the likely motions in the sport. The downside of the approach used in current methods is that the discriminator is trained on already available 3D datasets such as 3DPW or Human3.6M [42, 16]. This can be limiting when the objective is sports analysis as the poses seen in sports are not poses that are present in the datasets and could thus be classified as unlikely poses by the discriminator. For sports analysis, it is often the relative movements that are of interest, as also seen in the analysis of the kinematic sequence. To achieve accurate relative movements, temporally consistent methods are needed, but with current benchmarks [42] focusing on the PA-MPJPE metric, the rotational and temporal inconsistencies of the methods are not evaluated. We believe that rethinking the metrics to account for this, will encourage researchers to focus on this aspect of the methods performance. Some articles have started to report the acceleration error of the joint, which we believe is a step in the right direction toward more temporally consistent methods [8, 20], and thus methods useful for sports analytics.

7 Conclusion

In this paper, we presented a study of the current state-of-the-art monocular human 3D pose methods evaluated on a small 3D golf pose dataset obtained with a marker based motion capture setup. The purpose of the study was to investigate whether current methods are at a state where they can be used for accurate sports analytics. We have, through evaluation of the estimated joint positions,

shown that current models emphasizing the 2D reprojection loss fail to provide accurate depth estimates of the joint locations. For the golf swing, we see that the models are most likely to estimate the position of the hands incorrectly. Through a case study of the kinematic sequence of a golf swing, we have concluded that current methods are not accurate and temporally consistent enough to be used for sports analysis.

References

[1] Q. AB. Qualisys. `https://www.qualisys.com/`.

[2] I. Akhter and M. J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June-2015:1446–1455, 2015. ISSN 10636919. doi: 10.1109/CVPR.2015.7298751.

[3] 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, 2014. doi: 10.1109/CVPR.2014.471.

[4] B. Artacho and A. Savakis. Unipose: Unified human pose estimation in single images and videos. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7033–7042, 2020. doi: 10.1109/CVPR42600.2020.00706.

[5] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image BT - Computer Vision – ECCV 2016. pages 561–578, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46454-1. doi: 10.1007/978-3-319-46454-1_34. URL `https://doi.org/10.1007/978-3-319-46454-1_34`.

[6] A. Bulat, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. Toward fast and accurate human pose estimation via soft-gated skip connections. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 8–15, 2020. doi: 10.1109/FG47880.2020.00014.

[7] X. T. B.V. Xsens. `https://www.xsens.com/`.

[8] H. Choi, G. Moon, J. Y. Chang, and K. M. Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1964–1973, 2021. doi: 10.1109/CVPR46437.2021.00200.

[9] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. ISSN 00333123. doi: 10.1007/BF02291478.

[10] S. Guan, J. Xu, M. Z. He, Y. Wang, B. Ni, and X. Yang. Out-of-domain human mesh reconstruction via dynamic bilevel online adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–16, 2022. doi: 10.1109/TPAMI.2022.3194167.

[11] H. Gulgin, C. Armstrong, and P. Gribble. Hip rotational velocities during the full golf swing. *Journal of Sports Science and Medicine*, 8(2):296–299, 2009. ISSN 13032968. doi: 10.1249/00005768-200605001-02539.

[12] R. A. Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. doi: 10.1109/CVPR.2018.00762.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Identity Mappings in Deep Residual Networks. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 630–645, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46493-0. doi: 10.1007/978-3-319-46493-0_38. URL `https://doi.org/10.1007/978-3-319-46493-0_38`.

[14] Y. He, R. Yan, K. Fragkiadaki, and S.-I. Yu. Epipolar transformers. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7776–7785, 2020. doi: 10.1109/CVPR42600.2020.00780.

[15] S. A. Horan, K. Evans, N. R. Morris, and J. J. Kavanagh. Thorax and pelvis kinematics during the downswing of male and female skilled golfers. *Journal of Biomechanics*, 43(8):1456–1462, 2010. ISSN 00219290. doi: 10.1016/j.jbiomech.2010.02.005. URL http://dx.doi.org/10.1016/j.jbiomech.2010.02.005.

[16] 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, 2014. doi: 10.1109/TPAMI.2013.248.

[17] K. Iskakov, E. Burkov, V. Lempitsky, and Y. Malkov. Learnable triangulation of human pose. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7717–7726, 2019. doi: 10.1109/ICCV.2019.00781.

[18] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. doi: 10.1109/CVPR.2018.00744.

[19] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik. Learning 3d human dynamics from video. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5607–5616, 2019. doi: 10.1109/CVPR.2019.00576.

[20] M. Kocabas, N. Athanasiou, and M. J. Black. Vibe: Video inference for human body pose and shape estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5252–5262, 2020. doi: 10.1109/CVPR42600.2020.00530.

[21] M. Kocabas, C.-H. P. Huang, O. Hilliges, and M. J. Black. Pare: Part attention regressor for 3d human body estimation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11107–11117, 2021. doi: 10.1109/ICCV48922.2021.01094.

[22] N. Kolotouros, G. Pavlakos, M. Black, and K. Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2252–2261, 2019. doi: 10.1109/ICCV.2019.00234.

[23] K. Lin, L. Wang, and Z. Liu. Mesh graphormer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12919–12928, 2021. doi: 10.1109/ICCV48922.2021.01270.

[24] K. Lin, L. Wang, and Z. Liu. End-to-end human pose and mesh reconstruction with transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1954–1963, 2021. doi: 10.1109/CVPR46437.2021.00199.

[25] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34 (6), nov 2015. ISSN 0730-0301. doi: 10.1145/2816795.2818013. URL https://doi.org/10.1145/2816795.2818013.

[26] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. Black. Amass: Archive of motion capture as surface shapes. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5441–5450, 2019. doi: 10.1109/ICCV.2019.00554.

[27] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2659–2668, 2017. doi: 10.1109/ICCV.2017.288.

[28] C. M. A. H. Matthew Trumble, Andrew Gilbert and J. Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In G. B. Tae-Kyun Kim, Stefanos Zafeiriou and K. Mikolajczyk, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 14.1–14.13. BMVA Press, September 2017. ISBN 1-901725-60-X. doi: 10.5244/C.31.14. URL https://dx.doi.org/10.5244/C.31.14.

[29] S. Mehdizadeh, H. Nabavi, A. Sabo, T. Arora, A. Iaboni, and B. Taati. Concurrent validity of human pose tracking in video for measuring gait parameters in older adults: a preliminary analysis with multiple trackers, viewing angles, and walking directions. *Journal of NeuroEngineering and Rehabilitation*, 18(1): 1–16, 2021. ISSN 17430003. doi: 10.1186/s12984-021-00933-0.

[30] K. Mitchell, S. Banks, D. Morgan, and H. Sugaya. Shoulder Motions During the Golf Swing in Male Amateur Golfers. *Journal of Orthopaedic & Sports Physical Therapy*, 33 (4):196–203, 2003. doi: 10.2519/jospt.2003.33.4.196. URL `https://doi.org/10.2519/jospt.2003.33.4.196`.

[31] F. Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. *CoRR*, abs/1611.09010, 2016. URL `http://arxiv.org/abs/1611.09010`.

[32] S. M. Nesbit. A three dimensional kinematic and kinetic study of the golf swing. *Journal of Sports Science and Medicine*, 4(4):499–519, 2005. ISSN 13032968.

[33] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10967–10977, 2019. doi: 10.1109/CVPR.2019.01123.

[34] V. Ramakrishna, T. Kanade, and Y. Sheikh. Reconstructing 3D human pose from 2D image landmarks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7575 LNCS(PART 4):573–586, 2012. ISSN 03029743. doi: 10.1007/978-3-642-33765-9_41.

[35] 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, 2021. doi: 10.1109/CVPR46437.2021.01494.

[36] L. Sigal, A. O. Balan, and M. J. Black. HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion. *International Journal of Computer Vision*, 87 (1):4, 2009. ISSN 1573-1405. doi: 10.1007/s11263-009-0273-6. URL `https://doi.org/10.1007/s11263-009-0273-6`.

[37] J. Stenum, C. Rossi, and R. T. Roemmich. Two-dimensional video-based analysis of human gait using pose estimation. *PLoS Computational Biology*, 17(4), 2021. ISSN 15537358. doi: 10.1371/journal.pcbi.1008935. URL `http://dx.doi.org/10.1371/journal.pcbi.1008935`.

[38] Z. Su, M. Ye, G. Zhang, L. Dai, and J. Sheng. Cascade feature aggregation for human pose estimation, 2019.

[39] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5686–5696, 2019. doi: 10.1109/CVPR.2019.00584.

[40] Y. Sun, Q. Bao, W. Liu, Y. Fu, M. J. Black, and T. Mei. Monocular, one-stage, regression of multiple 3d people. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11159–11168, 2021. doi: 10.1109/ICCV48922.2021.01099.

[41] V. M. S. L. UK. Vicon. `https://www.vicon.com/`.

[42] T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll. Recovering Accurate 3D Human Pose in the Wild Using IMUs and a Moving Camera BT - Computer Vision – ECCV 2018. pages 614–631, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01249-6. doi: 10.1007/978-3-030-01249-6_37. URL `https://doi.org/10.1007/978-3-030-01249-6_37`.

[43] S. Yang, Z. Quan, M. Nie, and W. Yang. Transpose: Keypoint localization via transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*,

pages 11782–11792, 2021. doi: 10.1109/ICCV48922.2021.01159.

[44] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah. Deep Learning-Based Human Pose Estimation: A Survey. *arXiv e-prints*, art. arXiv:2012.13392, Dec. 2020.

[45] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4966–4975, 2016. doi: 10.1109/CVPR.2016.537.

[46] X. Zhou, M. Zhu, G. Pavlakos, S. Leonardos, K. G. Derpanis, and K. Daniilidis. Monocap: Monocular human motion capture using a cnn coupled with a geometric prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(4):901–914, 2019. doi: 10.1109/TPAMI.2018.2816031.