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

Evaluating Recent 2D Human Pose Estimators for 2D-3D Pose Lifting

Soroush Mehraban[1,4], Yiqian Qin[2], Babak Taati[1,3,4]
[1] KITE Research Institute, University Health Network
[2] Department of Electrical and Computer Engineering, University of Toronto
[3] Department of Computer Science, University of Toronto
[4] Institute of Biomedical Engineering, University of Toronto

***Abstract*— Monocular 3D human pose estimation involves predicting the 3D pixel coordinates of key body joints from a 2D image or video. Typically, a 2D estimation model is employed to initially determine joint locations in an image, followed by training a separate model to lift these positions to 3D coordinates. In this paper, we evaluate the performance of recently proposed 2D human pose estimation models as different inputs for training and evaluation of 2D-3D lifting models. In addition, we propose four simple merging strategies to combine the outputs of these 2D human pose estimators and generate less noisy 2D inputs. To evaluate, four recent 2D pose estimators—ViTPose, PCT, MogaNet, and TransPose—are selected, and their corresponding 2D outputs are generated on the Human3.6M dataset. Subsequently, MotionAGFormer and PoseFormerV2 are trained and evaluated using each created 2D input and its corresponding 3D motion-capture ground truth. ViTPose stands out as the top-performing 2D estimator, and employing all merging strategies proves beneficial in generating a less noisy 2D input. Code and data are available at https://github.com/TaatiTeam/2DEstimatorEval.**

I. INTRODUCTION

Monocular 3D human pose estimation entails predicting 3D pixel coordinates of key body joints, such as knees, hips, and elbows, based on a 2D image or video. The method is used in a variety of applications, ranging from augmented [11] and virtual reality [15], to autonomous vehicles [2], clinical monitoring [3] and human-computer interaction [16]. Nevertheless, the inherent challenge in this process lies in its ill-posed nature, mainly due to depth ambiguities present in the 2D input data.

Since the majority of datasets in this domain are collected in controlled laboratory settings, directly estimating 3D poses from images lacks generalizability. On the other hand, 2D pose estimation models are trained with a wide range of data reflecting different environments. Consequently, a common approach for 3D human pose estimation involves a two-step process: (i) locating the 2D positions of key body joints in video frames, followed by (ii) lifting these 2D pixel coordinates to 3D.

While there have been numerous models suggested for each of these two stages, the connection between 2D human pose estimation and the subsequent 3D lifting process has not been thoroughly investigated. Typically, practitioners employ an off-the-shelf 2D pose detection model fine-tuned on 2D-3D lifting dataset video frames. However, this method comes with a drawback during evaluation. The 2D detection model has already been trained on a recording environment similar to the test dataset, resulting in less noisy 2D data that does not hold true for in-the-wild videos, making the evaluation less representative of real-world conditions. This work aims to experimentally find the best available 2D human pose estimation model(s) for the specific subsequent task of 3D lifting, without fine-tuning on the training dataset. Our main contributions are:

- We examine the utility of four 2D pose estimators, including from the top of the 2D leaderboards [13], [1], for the specific task of 3D lifting, and compare them with Detectron [7] (with a ResNet101-FPN backbone [12]), and CPN [4] that are fine-tuned on 2D-3D lifting dataset. The models are: TransPose [25], MogaNet [9], ViTPose [24], and PCT [6].
- We propose four simple merging strategies to combine 2D key joints coordinates of aforementioned 2D human pose estimation models and generate less noisy 2D inputs for 3D lifting, thereby improving 3D human pose lifting performance.

II. RELATED WORK

2D human pose estimation. These models receive a single RGB image as input and output locations of main joints in 2D pixel coordinate. Cascaded Pyramid Network (CPN) [4] introduces GlobalNet, a feature pyramid network aimed at localizing keypoints that are easily detectable, such as eyes and hands. Furthermore, the CPN incorporates an additional module called RefineNet, specifically devised to handle the localization of occluded keypoints. Stacked Hourglass [17] employs several stacked hourglass modules, enabling iterative bottom-up and top-down inference processes. TransPose [25] uses a CNN backbone to extract high-level image features and then uses a transformer encoder to process these extracted features. MogaNet [9] proposes a new family of pure ConvNet structure which shows competitive results in various computer vision tasks, including object detection, semantic segmentation, and 2D human pose estimation. ViTPose [24] employs a pure vision transformer for extracting image features and by using two deconvolution layers as the decoder, it generates heatmaps containing the 2D keypoints of different areas of the body. PCT [6] proposes a structured representation to constrain joint locations and prevent the model output to generate unrealistic pose estimates.

Monocular 3D human pose estimation. Earlier methods involved determining the 3D coordinates of joints directly from video frames, without the need for any intermediary

979-8-3503-9494-8/24/$31.00 ©2024 IEEE

processing to locate 2D pixel coordinates [18], [19], [22], [28]. Inspired by the rapid development and availability of accurate 2D pose estimation models, more recent models receive a sequence of 2D human pose as input and lift them to 3D coordinate system. VideoPose3D [20] uses dialated temporal convolutions over 2D keypoints to infer the 3D pose sequence. PoseFormer [27] is the first method that proposes spatial transformers to extract intra-frame information between joints and temporal transformers to extract inter-frame information. PoseFormerV2 [26] enhances its computational efficiency by using a frequency-domain representation, which also conferred robustness against abrupt movements in noisy data. STCFormer [23] proposes two parallel branches, one using spatial transformers and the other using temporal transformers. P-STMO [21] introduces masked pose modeling and achieves a lower final error through self-supervised pretraining. Enfalt *et al.* [5] reduce computational complexity by utilizing masked token modeling. In StridedFormer [10], the traditional fully-connected layers in the feed-forward network of the transformer encoder are substituted with strided convolutions. This modification aims to gradually reduce the sequence length and effectively enhance the central frame. MotionBERT [29] further improves the performance by using spatial-temporal stack of transformers in one branch and temporal-spatial transformers in another branch. MotionAGFormer [14] uses spatial-temporal transformers in one branch and Graph Convolutional Networks (GCNs) in another branch to capture a complementary information and output more accurate results.

In the experiments, we chose MotionAGFormer and PoseFormerV2 for the 2D-3D pose lifting task. For a fair comparison, MotionAGFormer is modified to accept 2D inputs without confidence scores and trained with the same data preprocessing as PoseFormerV2.

III. Method

We use recent 2D estimation models, trained on the MS COCO Keypoint dataset [13], to estimate 2D keypoints on the Human3.6M dataset [8]. Following that, we use the estimated 2D pose sequences as input to train the 2D-3D lifting models. Finally, we evaluate four simple merging strategies to combine different estimated 2D sequences and further improve the final 3D lifting accuracy.

A. 2D Human Pose Estimation

State-of-the-art models such as ViTPose [24], PCT [6], MogaNet [9], and TransPose [25], trained on the MS COCO dataset, are used to estimate 2D pose sequences for Human3.6M dataset. However, the 2D pose output format in MS COCO differed from that of the Human3.6M dataset. To align them, we converted the keypoints as illustrated in Figure 1. In addition to the abovementioned four models, we also used two 2D pose sequences used in VideoPose3D [20], i.e., CPN [4] fine-tuned on Human3.6M and Detectron [7] with and without fine-tuning. While the fine-tuned sequences were already in Human3.6M format, we converted the Detectron sequences without fine-tuning to match the required format.

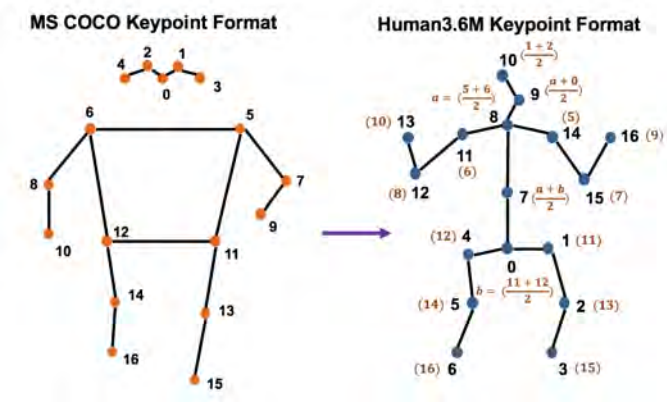

Fig. 1: MS COCO and Human3.6M keypoints format. For models trained on MS COCO dataset, we convert them to Human3.6M format.

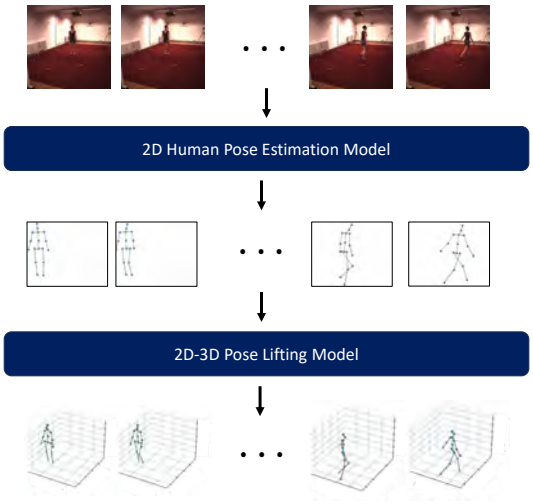

Fig. 2: **3D human pose estimation pipeline.** Initially, a 2D pose sequence is estimated from the RGB video using a 2D pose estimator. Subsequently, a 2D-3D lifting model is trained to lift 2D poses to 3D.

B. Merging Strategy

Four different simple merging strategies are proposed to improve the final 2D-3D lifting performance by introducing less noisy 2D data.

Winner-take-all (WTA) merging. For this merging strategy, for a single keypoint (e.g., right knee) estimated with different 2D estimators, the 2D estimate that has the least distance with the ground truth among all the training frames in Human3.6M is selected. For the ground truth, we project the motion capture 3D coordinates into 2D pixels by leveraging the camera intrinsic and extrinsic parameters. Specifically, given 3D coordinates P_W in world coordinate system, we use

$$P_C = R(P_W - T) \tag{1}$$

to convert it to camera coordinates system $P_C = (X_c, Y_c, Z_c)$ where R and T are rotation and translation parameters, respectively. Next, It is projected to 2D coordinates using

$$u = f\frac{X_c}{Z_c} + c_x, \tag{2}$$

$$v = f\frac{Y_c}{Z_c} + c_y, \tag{3}$$

where $P_p = (u, v)$ is 2D coordinates in pixel coordinates system. The intrinsic parameters, f, c_x, c_y, denote focal length and image center, respectively. Finally, for selecting the estimator d for a joint j, the 2D coordinates $P^t_{p',j}$ is represented as

$$\begin{aligned} P^t_{p',j} &= P^t_{d,j}, \\ \text{where } d &= \arg\min_{1\le i\le D} \sum_{t=1}^{T} ||P^t_{p,j} - P^t_{i,j}||. \end{aligned} \tag{4}$$

Here, $D = 4$ is the number of 2D estimators (trained on MS COCO dataset) and T is the total number of frames in Human3.6M used for training.

Average merging. In this merging approach, we compute the average of ViTPose, PCT, and MogaNet for each individual frame within the sequence. This averaging process aims to mitigate the impact of noise in the 2D input. Given that each estimator introduces varying levels of noise for a specific frame, combining their outputs through averaging is anticipated to yield a less noisy 2D input. Based on preliminary experiments, we decided not to use TransPose for average merging, because in general its output was significantly noisier than the other three model (see experimental results for more details).

Weighted average merging. In this approach, we incorporate the confidence scores of each 2D estimation as weights in a weighted average. We normalize the confidence scores of PCT (provided as logits) to probabilities. This strategy allows us to account for the confidence levels associated with each estimator's output, offering a more informed combination of results.

Concatenate merging. In this merging strategy, we concatenate results from ViTPose, PCT, and MogaNet, and train a model to lift $T \times J \times 2N$ data, with T as frames, J as joints, and N as number of 2D estimators.

C. 2D-3D Lifting

Following estimation of 2D pose sequences, PoseFormerV2 [26] and MotionAGFormer-B [14] are trained for the task of 2D-3D lifting (Figure 2). Among PoseFormerV2 variants, the model with the receptive field of 27 and $f = 3$ was selected for fast training and inference time. For the evaluation, Mean Per Joint Position Error (MPJPE) is used, defined as:

$$MPJPE = \Sigma_{t=1}^{T}\Sigma_{j=1}^{J}\|\hat{\mathbf{P}}_{t,j} - \mathbf{P}_{t,j}\|, \tag{5}$$

where J is number of joints, T is number of frames in batch of data, and $\hat{P}$ and P are the ground-truth 3D motion capture and estimated 3D pose, respectively.

IV. Experimental Results

A. Quantitative Comparison of 2D Sequences

The 2D sequences generated by different 2D estimators are initially transformed into the Human3.6M format, as illustrated in Figure 1. These converted sequences are then compared with the 2D ground truth, calculated through the 3D-2D camera projection process outlined in Equations 2 and 3. For the comparison, we incorporate all the training data from human participants 1, 5, 6, 7, and 8 in Human3.6M. Subsequently, we calculate the average for each joint by considering all frames across all the videos. The comparison for a subset of joints is illustrated in Figure 3. ViTPose generally surpasses other estimators in terms of mean per-joint position error, leading to more precise keypoint outputs. Nevertheless, for certain keypoints, PCT tends to yield more accurate keypoints on average compared to ViTPose. We hypothesize that the disparities in errors across various body regions can be attributed to the distinct biases inherent in each model, stemming from the use of different architectures. Building upon this concept, during WTA merging, ViTPose is predominantly employed, except for the following keypoints, where PCT exhibits lower average errors on training data: Left Knee, Upper Torso, Center Head, and Left Shoulder.

B. Quantitative Comparison of 3D Sequences

Table I compares the estimated 3D sequences with the motion capture 3D ground truth on the Human3.6M dataset after training 2D-3D lifting models using different 2D estimations as input. By comparison, ViTPose attains the lowest mean per-joint position error among the four recent models assessed for this task. The ultimate performance is 2.96 mm and 3.22 mm higher when compared to the scenario where PoseFormerV2 and MotionAGFormer are trained with the fine-tuned CPN model, respectively. It is important to note that CPN model used was fine-tuned on the Human3.6M dataset. We consider this approach unfair since the training and testing data in the Human3.6M dataset share identical environments and cameras, with subjects positioned at nearly the same distances. Consequently, CPN may acquire biases from the Human3.6M dataset that may not be applicable

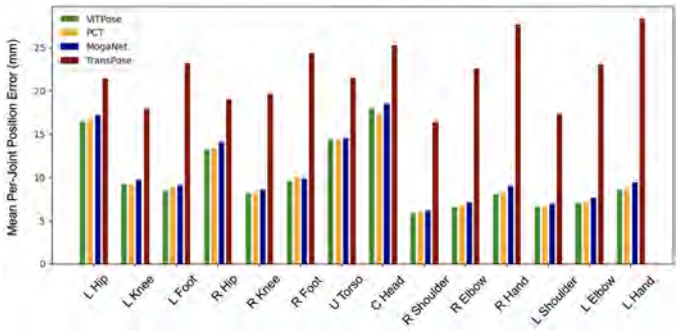

Fig. 3: Mean per-joint position error between each tested 2D estimator and the 2D ground truth. (L: left, R: right, U: upper, C: center)

to real-world scenarios where the subject is situated in a completely different environment. The performance of CPN without fine-tuning confirms our assertion, as the trained 2D-3D lifters exhibit inferior behavior compared to both PCT and ViTPose. Through the usage of the 2D sequences obtained via the merging strategies, we can enhance the performance of 2D-3D lifting models. Among the various merging strategies, WTA merging and concatenate merging were effective in reductions 3D error compared to ViTPose in PoseFormerV2 and MotionAGFormer, respectively.

TABLE I: The mean per-joint position error (mm) comparisons of estimated 3D keypoints on Human3.6M after training the PoseFormerV2 and MotionAGFormer using different 2D estimations.

2D Estimator	Finetuned	MPJPE (mm)	
		PoseFormerV2	MotionAGFormer
Detectron [7]	×	59.56	52.00
Detectron [7]	✓	55.91	47.48
CPN [4]	×	55.96	48.27
CPN [4]	✓	49.65	42.63
MogaNet [9]	×	54.77	48.52
TransPose [25]	×	66.20	52.87
PCT [6]	×	53.26	46.61
ViTPose [24]	×	52.61	45.85
Merge (WTA)	×	**51.96**	45.78
Merge (Average)	×	52.53	46.32
Merge (Weighted Average)	×	52.50	45.54
Merge (Concatenate)	×	52.13	**45.27**

C. Qualitative Comparison of 3D Sequences

Figure 4 visualizes the difference between sample estimated 3D sequences and the motion capture 3D ground truth on the Human3.6M dataset after training PoseFormerV2 and MotionAGFormer using different 2D estimations as input. CPN generally aligns better with the ground-truth. Among different merging strategies for PoseFormerV2 in Figure 4 (a), the WTA strategy has slightly fewer errors in some keypoints (e.g. right hand, left ankle), though it's a bit worse in others (e.g. left hand). Overall, WTA shows a bit better performance. Figure 4 (b) exhibits a similar trend, where concatenate merging demonstrates superior alignment compared to other merging strategies.

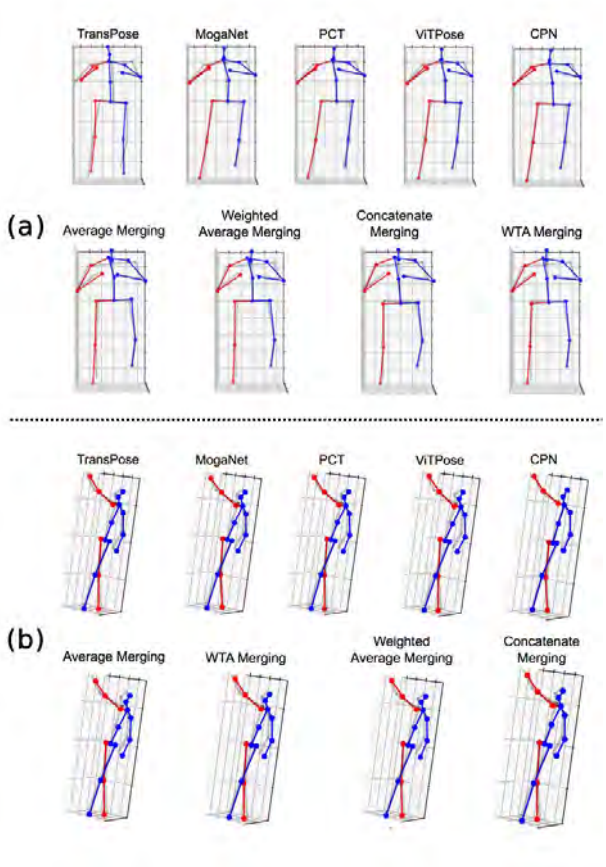

Fig. 4: Qualitative comparisons of estimated 3D keypoints on Human3.6M after training the (a) PoseFormerV2 and (b) MotionAGFormer using different 2D estimations. Transparent gray skeleton represents the ground-truth 3D pose. The right part is shown in red, and the left part and torso are shown in blue. WTA: Winner-take-all

V. Conclusion

Among the four recent 2D human pose estimators used in the 2D-3D pose lifting process, ViTPose exhibited the most promising results. Specifically, it generates the most precise 2D estimations for the majority of the keypoints, and achieves the lowest mean per-joint position error of estimated 3D sequences. Additionally, we investigated four merging strategies to combine the outputs of the 2D estimators to further reducing the final error in the estimated 3D sequences.

REFERENCES

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[2] P. Bauer, A. Bouazizi, U. Kressel, and F. B. Flohr. Weakly supervised multi-modal 3d human body pose estimation for autonomous driving. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, pages 1–7. IEEE, 2023.

[3] A. Bigalke, L. Hansen, J. Diesel, C. Hennigs, P. Rostalski, and M. P. Heinrich. Anatomy-guided domain adaptation for 3d in-bed human pose estimation. *Medical Image Analysis*, 89:102887, 2023.

[4] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018.

[5] M. Einfalt, K. Ludwig, and R. Lienhart. Uplift and upsample: Efficient 3d human pose estimation with uplifting transformers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2023.

[6] Z. Geng, C. Wang, Y. Wei, Z. Liu, H. Li, and H. Hu. Human pose as compositional tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 660–671, 2023.

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[8] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.

[9] S. Li, Z. Wang, Z. Liu, C. Tan, H. Lin, D. Wu, Z. Chen, J. Zheng, and S. Z. Li. Efficient multi-order gated aggregation network. *arXiv preprint arXiv:2211.03295*, 2022.

[10] W. Li, H. Liu, R. Ding, M. Liu, P. Wang, and W. Yang. Exploiting temporal contexts with strided transformer for 3d human pose estimation. *IEEE Transactions on Multimedia*, 25:1282–1293, 2022.

[11] H.-Y. Lin and T.-W. Chen. Augmented reality with human body interaction based on monocular 3d pose estimation. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 321–331. Springer, 2010.

[12] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[14] S. Mehraban, V. Adeli, and B. Taati. MotionAGFormer: Enhancing 3d human pose estimation with a transformer-gcnformer network. *arXiv preprint arXiv:2310.16288*, 2023.

[15] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. VNect: Real-time 3d human pose estimation with a single rgb camera. *Acm transactions on graphics (tog)*, 36(4):1–14, 2017.

[16] T. L. Munea, Y. Z. Jembre, H. T. Weldegebriel, L. Chen, C. Huang, and C. Yang. The progress of human pose estimation: A survey and taxonomy of models applied in 2d human pose estimation. *IEEE Access*, 8:133330–133348, 2020.

[17] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 483–499. Springer, 2016.

[18] G. Pavlakos, X. Zhou, and K. Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7307–7316, 2018.

[19] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7025–7034, 2017.

[20] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7753–7762, 2019.

[21] W. Shan, Z. Liu, X. Zhang, S. Wang, S. Ma, and W. Gao. P-STMO: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 461–478. Springer, 2022.

[22] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei. Integral human pose regression. In *Proceedings of the European conference on computer vision (ECCV)*, pages 529–545, 2018.

[23] Z. Tang, Z. Qiu, Y. Hao, R. Hong, and T. Yao. 3d human pose estimation with spatio-temporal criss-cross attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4790–4799, 2023.

[24] Y. Xu, J. Zhang, Q. Zhang, and D. Tao. ViTPose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022.

[25] S. Yang, Z. Quan, M. Nie, and W. Yang. TransPose: Keypoint localization via transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11802–11812, 2021.

[26] Q. Zhao, C. Zheng, M. Liu, P. Wang, and C. Chen. PoseFormerV2: Exploring frequency domain for efficient and robust 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8877–8886, June 2023.

[27] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11656–11665, 2021.

[28] K. Zhou, X. Han, N. Jiang, K. Jia, and J. Lu. HEMlets pose: Learning part-centric heatmap triplets for accurate 3d human pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2344–2353, 2019.

[29] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang. MotionBERT: Unified pretraining for human motion analysis. *arXiv preprint arXiv:2210.06551*, 2022.