Deformable Transformer with Feature Upsampling for Human Mesh Recovery

Jeffrey Heo
Department of Computer Science
Stanford University
jeffheo@stanford.edu

George Hu
Department of Computer Science
Stanford Unviersity
gehu@stanford.edu

Abstract

Human mesh recovery (HMR), is a challenging task for reconstructing a 3D mesh of the human body across a variety of environments. Various previous works have built relatively mature paradigms for parametrization of mesh coordinates with respect to joints and poses, building upon the body of work from pose estimation, but the actual modules involved in encoding and decoding embeddings do not yet result in optimal mesh recovery. The use of more contemporary vision transformers has improved the localization and pose angles in HMR, and we continue in this line of work, exploring how a form of sparse attention, deformable attention, can be improved by superresolution of low-dimensional image features. We find that using well-posed feature upsampling visibly improves feature quality and downstream performance, achieving near equal results to dense attention while training quicker. Our code can be found at: https://github.com/jeffheo/DeformableHMR.

1. Introduction

Motion capture (MoCap) technology has applications in numerous fields such as film, gaming, AR/VR, as well as sports medicine by providing a tool to capture and analyze human movement in 3D. Traditional marker-based MoCap systems utilizing multi-view cameras and marker suits recover highly accurate human motion but suffer from poor accessibility due to high cost and rigorous set-up. In contrast, a single camera with the correct algorithm can perform 3D Monocular Human Mesh Recovery (HMR), which computes a mesh of a human body in 3D given an input image or video, as a more accessible alternative using deep learning [8].

One common approach to 3D HMR leverages the Skinned Multi-Person Linear (SMPL) [12] representation model, consisting of joint articulations (also called pose parameter s) and body shape vectors (body shape parameters), to generate realistic human body meshes. Current challenges in HMR include occlusion situations and consistency between video frames [5], but underlying these issues is simply a lack of correct positional identification to help models output correct joint angles [9].

More recently, advances in vision transformers have demonstrated versatility and overall impressive performance across a wide range of vision tasks and domains, particularly in determining complex spatial relations [5]. In the field of object detection deformable attention has emerged as one promising solution for accurate, space-aware localization, and extending such an approach to HMR requires even greater focus on priors for extracting precise positional semantics [18].

In parallel, issues of data generalization across diverse real world applications for vision models have been diminished by the release of large vision transformer models pretrained on self-supervision tasks on web-scale datasets [14] [2]. The ability of these foundation models to generate meaningful features across all data spectra for downstream application has created a new effective learning paradigm, and more recently, works [4] have begun on improving the spatial resolution of these vision foundation model features for ever better results.

By integrating the information derived from large, pretrained vision transformer features with novel upsampling methods and deformable attention decoding, we aim to develop a novel transformer-based HMR framework that significantly improves upon current methods in both precision and computation efficiency.

2. Related Work

2.1. Transformer-Based HMR

End-to-end 3D human mesh recovery, not relying upon intermediate 2D keypoints or joints, was first proposed by Kanazawa et al. [8]. This was achieved by levraging novel deep learning advancements at the time and regressing the SMPL parameters along with a camera model to derive the 3D meshes. In [5], Goel et al. would extend on this framework and incorporate a vision transformer architecture, us-

ing a single query token fed into the decoder for SMPL and camera parameter predictions. HMR-2.0 established a new competitive baseline on single human mesh recovery, and in particular, they show how their transformer network can decode complex pose positions that have traditionally been hard to model due to complex spatial relations. Thus the authors integrate HMR-2.0 into a video system Humans4D to effectively support realistic human tracking and mesh reconstruction. For the purposes of this paper, we will call the vision transformer encoder pretrained on 2D pose estimation they use as **ViT-Pose**.

2.2. Deformable Attention

The Deformable Transformer [18] architecture, first proposed in end-to-end object detection, has demonstrated comparable performance to other SOTA methods without needing any hand-designed components commonly used in object detection. The deformable attention module is designed for efficiency and complex relational parameterization, having the keys and values be sparsely sampled learned offsets from a reference location determined by a given query. Zhu et al. show that this increases model training and inference speed while also incorporating inductive biases for precise spatial modeling beneficial for object detection.

Yoshiyasu (2023) [16] extends this notion of deformable attention to 3D HMR with the DeFormer architecture, using the joint and shape query tokens at each layer to generate reference points and offsets on multi-scale maps to be used in the attention computation. DeFormer works directly with positional information without the SMPL parameterization for dense mesh reconstruction, and it improves upon previous baselines of similar model size.

2.3. Feature Upsampling

Machine learning architectures can compress information from images into embedding spaces very well, but it often comes with the cost of lowering spatial resolution. The ResNet [6] and ConvNext [11] CNN variants compress the height and width by a factor of 32 at the last embedding layer, and ViT architectures cby a factor of 16 commonly. Fine-grained details useful for downstream applications can be lost with the low resolution features, so feature upsampling aims to increase the spatial resolution of feature maps while retaining useful semantics.

In FeatUp [4] Fu et al. introduce a lightweight, model-agnostic method to upsample visual feature maps that perform well in downstream applications. They introduce two modes: IMPLICIT, which can be used to overfit on one image for the highest quality upsampling, and JBU, which uses stacked joint bilateral upsamplers trained on multi-view consistency for generalizable feature upsampling.

3. Methodology

3.1. Pretraining FeatUp

We train JBU FeatUp for the features produced by ViT-Pose (used in HMR-2.0 [5]) in a similar configuration to the standard method in Fu et al. [4] that upsamples by a factor of 16, using an attention-based downsampler to go from high resolution to low resolution features. Multi-view consistency loss is applied between the original low resolution feature encodings and the downsampled features. We experiment with upsampling by smaller factors than 16, and for these experiments, we keep the downsampler and multi-view transformations consistent, always downsampling by a factor of 16. To ensure matching dimensions, we add additional bilinear interpolation when learning upsamplers that only upsample by factors of 4 and 8 to recover the original image resolution. We use a 20% subset of COCO data for training due to time constraints—ideally we could have pretrained on all of our training data.

From this stage we procure upsamplers $u_{16}^{(r)}$ and $u_{32}^{(r)}$ based on the outputs of the 16th and 32nd layers of ViT-Pose for upsample factors of $r = 4, 8, 16$.

3.2. Generating Feature Maps

We use a the frozen ViT-Pose from Goel et al. [5] as our initial feature encoder. Given an input image $x \in \mathbb{R}^{H\times W\times 3}$ and a patch size of 16, we represent the spatial output tokens at layer ℓ of the encoder f as $f_\ell(x) \in \mathbb{R}^{H/16\times W/16\times C}$. To get comprehensive feature maps at multiple stages of the encoder, we thus extract the outputs at layers 16 and 32, representing the middle and end of the transformer encoder.

To generate good quality higher resolution features that retain semantic and positional information, we use the JBU FeatUp upsamplers $u_{16}^{(r)}$ and $u_{32}^{(r)}$ for each of the spatial maps $f_16(x), f_{32}(x)$ to recover higher resoultion features with spatial dimension $r(H/16) \times r(W/16)$. At the end of this stage we have stacked feature maps

$$\{M_\ell := u_\ell^{(r)}\left(f_\ell(x)\right) \in \mathbb{R}^{r(H/16)\times r(W/16)\times C}\}_{\ell=[16,32]}$$

to be used in the deformable attention decoder.

3.3. Deformable Attention Decoder

We emulate the multi-scale deformable attention decoder in DeformableDETR, replacing the object queries with SMPL parameter queries for pose and shape instead and using 4 scales as provided from the encoder feature maps. Here is where our method provides significant theoretical benefit; in DeformableDETR, the reference + offset locations are floating point values in the feature map coordinate space, and bilinear interpolation is used to extract the relevant key and value information. However, we hypothesize this results in significant information loss due to

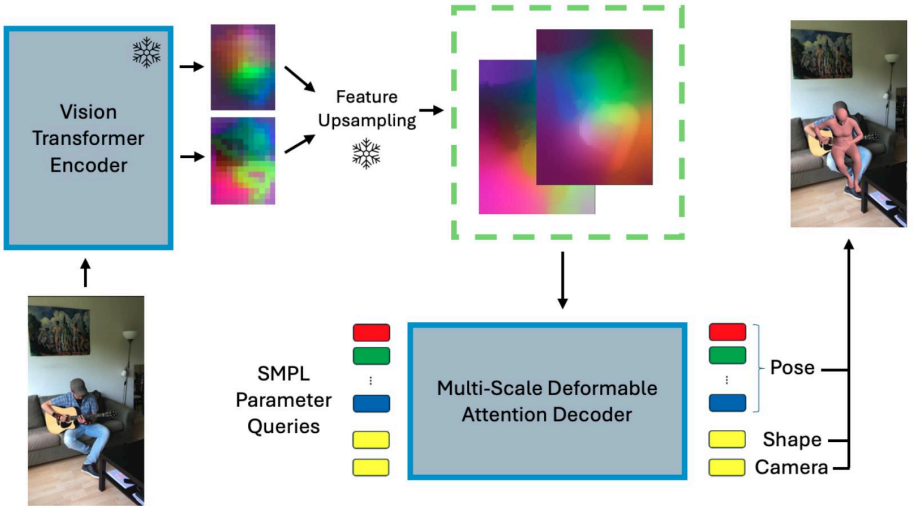

Figure 1. Model Architecture. Images are encoded, features upsampled, and the decoder queries reconstruct the mesh.

low spatial resolution, so we believe our deformable attention decoder can learn significantly better relations with the higher resolution image features.

The input SMPL query tokens are learnable $t \in \mathbb{R}^{26\times1280}$, representing 24 pose tokens, 1 shape token, and 1 camera token. After passing the queries through the decoder, we learn linear projections W_{pose}, W_{shape}, and W_{camera} to get the desired outputs of pose parameters $\theta \in \mathbb{R}^{24\times6}$, shape parameters $\beta \in \mathbb{R}^{10}$, and camera parameters $\pi \in \mathbb{R}^{3\times3}$. For the pose rotation angles in the SMPL parameters, we use the common 6D representation pioneered by Zhou et al. (2020) [17] for a more continuous loss-landscape, converting to the actual pitch/roll/yaw and rotation matrix afterwards. Moreover, we use one round of iterative error feedback [3], starting with the mean SMPL values from Humans 3.6M [7] to condition our predictions better. These are passed into the SMPL model to generate our 3D meshes.

3.4. Training Details

Following [9], we train with reconstruction loss on the SMPL parameters, the 3D joint positions, the 3D mesh vertices, and the projected 2D joint positions, all using mean square error. The relative loss weight for SMPL parameters is $\lambda_{SMPL} = 1$, 2D and 3D joint positions is $\lambda_{joint} = 5$, and mesh vertices $\lambda_{mesh} = 60$. For all training runs, we freeze the ViT-Pose to explore efficient decoding methods.

We train all models on four real world datasets, two with 3D SMPL ground truth derived from from motion capture—3DPW [15] and MPI-INF-3DHP [13]—and two psuedo-labeled from 2D pose ground truth using the CLIFF-annotator [9], COCO [10] and MPII [1]. To save on training time and still do a fair comparison, we use the same 20% subsample of the training data and train for 50 epochs. The evaluation is performed on the test split of 3DPW, and we use mean joint position error (MPJPE), procrustes analysis MPJPE (PA-MPJPE), and per vertex error (PVE) to determine how well the predicted mesh matches the ground truth.

4. Results

We compare various model architectures using our evaluation metrics, parameter count, and training time in table 1. BEDLAM-CLIFF is a non-transformer benchmark that improves upon HMR-1.0. The Deformable DETR adaptation to human mesh recovery here replaces the object queries with SMPL queries and the box prediction head with the SMPL prediction head—we retain the smaller model size and usage of ResNet50, just testing how deformable attention directly works with the older architecture they use. We re-implement HMR-2.0 and modify the transformer decoder to accept a 26 query input. The FeatUp module we report results for here is with 4 times upsampling, as we found that performed best in terms of PA-MPJPE.

We can observe that although our method fails to match the performance of modified HMR-2.0, the efficient deformable attention still achieves very competitive metrics, and this form of attention is improved through FeatUp. Despite having slightly worse performance, the speed of the deformable cross-attention between queries and the image, sampling a fixed $p = 8$ points per attention head as opposed to linear with respect to the image size, really shines in the

Table 1. Metrics comparison between model architectures

Method	**MPJPE ↓**	**PA-MPJPE ↓**	**PVE ↓**	***Params***	***GPU Time (Hours)***
BEDLAM-CLIFF	76.61	48.56	90.70	80M	4.3
Deformable DETR	102.16	65.06	120.48	89M	5.0
HMR-2.0	66.67	41.35	79.49	725M	24.4
ViT-Pose + Deformable	69.78	44.42	83.51	713M	13.9
Ours (ViT-Pose + FeatUp + Deformable)	68.25	43.18	82.02	714M	14.5

training speed. Moreover, the methods that do not use ViT-Pose, BEDLAM-CLIFF and Deformable DETR, both perform significantly worse, showing that well-conditioned initial image features are likely essential for any human mesh recovery model.

4.1. Analysis

Our HMR model exhibits a robust capability in capturing the general body pose and proportions of individuals across various scenarios, as seen in the visualizations on Figure 2. Upon rendering the recovered meshes on each of four distinct images from the validation set of the 3DPW [15] dataset, we confirm our model comprised of the Vit-Pose transformer encoder, FeatUp model for feature upsampling, and the transformer decoder using Deformable attention generalizes well across various in-the-wild image datapoints. Our model demonstrates plausible meshes for humans dancing sideways, sitting on the staircase, running, and conversing sideways, and in particular, we show strength in accuracy of upper body articulation and orientation.

Despite these strengths, there are several areas where the

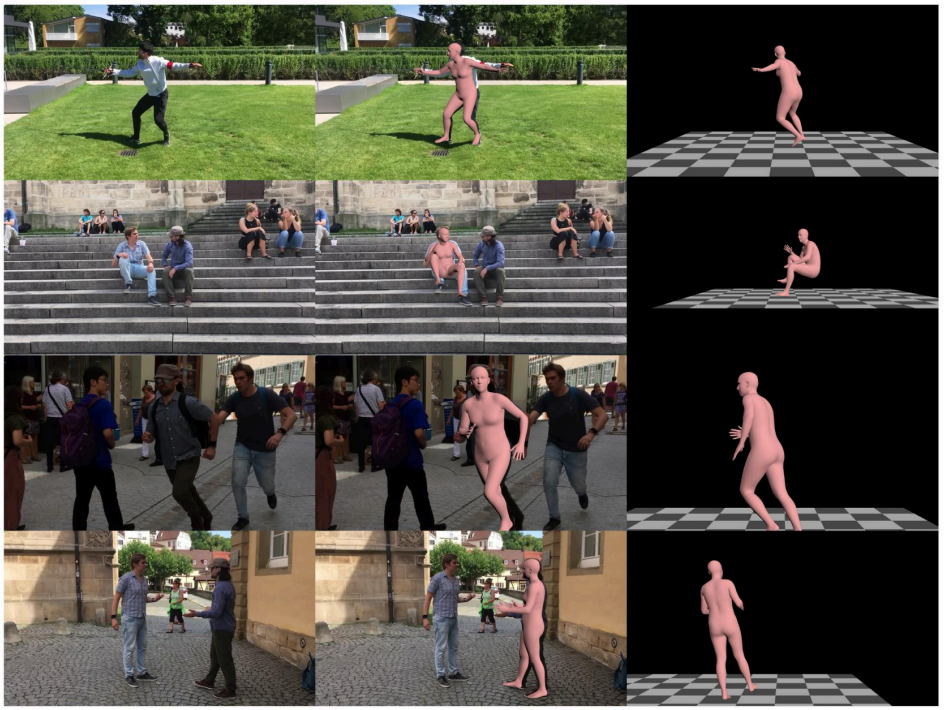

Figure 2. Qualitative results on validation set. We visualize the original image, the predicted mesh projected onto the original image, and the mesh alone on the validation set of the 3DPW dataset [15]. Upon visualizing our model's recovered meshes on four distinct scenarios, we deduce that our framework demonstrates competitive 3D HMR performance compared to existing SOTA models, as the model recovers decent meshes that demonstrate high accuracies particularly in the upper body region. However, the model struggles to accurately recover feet locations and suffers from high localization errors under difficult scenarios such as self-occlusions.

model could be improved. Notably, the model frequently exhibits inaccuracies in foot positioning and orientation. Such errors can detract significantly from the realism of the rendered meshes, as correct foot alignment is essential for the overall stability and appearance of the pose.

Self-occlusion presents another challenge for the model, particularly noticeable in scenarios where one limb occludes another, such as arms or legs during walking motions. These situations often result in unrealistic and inaccurate limb positioning. Such errors are noticeable in the third and fourth row of Figure 2; on the third row, the occluded right leg suffers from positional inaccuracy, and the same holds for the occluded right arm of the fourth row.

We also compare the effect of FeatUp JBU upsampler scaling factor on performance.

Upsample Factor	**MPJPE** ↓	**PA-MPJPE** ↓	**PVE** ↓
4x	68.25	43.18	82.02
8x	67.73	43.71	82.10
16x	68.89	44.14	83.14

Table 2. Results for using different FeatUp JBU upsample ratios

In table 2, we observe that taking all metrics in aggregate, 4x does slightly better than 8x, which both surpass 16x. This was a bit surprising to us as we expected better spatial resolution to result in better results. However, taking a look at the features themselves, as in figure 3, we can see how all the FeatUp JBU results are much crisper and seemingly more useful than the bilinear interpolation in the absence of FeatUp, but there begin to be slight artifacts at 8x and there are significant artifacts at 16x. Fu et al. [4] did mention the possibility of FeatUp overfitting and generalizing poorly to unseen images, and we believe these artifacts on the 16x results are indicative of this behavior. Although the 16x features look the sharpest, the 4x actually has better leg resolution as it seems like the artifacts in the 16x cover up the space between the legs.

In Figure 4, we visualize what the deformable attention decoder is attending to. The black squares are reference offsets with low attention values and the more red, the larger the attention value. In general the deformable attention seems to be focusing on a good amount of relevant areas, such as the individual's limbs and other sort of body key points, but there is quite a few harder to explain larger attention values to points such as the ground or in the zero-padding location. A significant general trend is that the points all are clustered towards the middle of the image when important context does exist at the extremities, so the model may just be missing out on that. We suspect that since the locations of the offsets and reference points are calculated as the sigmoid output, mapping $(0, 1)$ range to the image's height and width. For these offsets pre-sigmoid to approach large positive or negative values during training to allow for image extremities to be reached is quite difficult, so we suspect doing some temperature scaling or other method could improve the deformable attention.

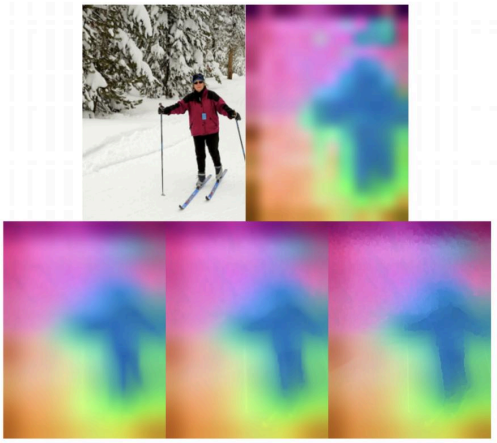

Figure 3. PCA Visualization of FeatUp. Top Row: Original Image and Bilinear Interpolation of features (no FeatUp). Bottom Row (Left to Right): 4x, 8x, 16x FeatUp.

Figure 4. Deformable Attention visualization for last layer of decoder. All splotched squares are offset locations, and red pigment corresponds to attention value.

5. Conclusion

With the proliferation of large, pretrained vision transformers across various vision tasks, human mesh recovery is no exception. The method in which researchers choose to use the resulting features of these models still has significant variability, as concerns regrading parameter count and training time for lightweight decoding exist. We show that a deformable attention decoder does result in some performance decline from full-attention; however the efficient attention module allows for significantly faster training, up to 40%. By adding feature upsampling, we improve the deofmrable attention and show that spatial quality of feature maps for deformable attention applications does matter—we observe that feature map quality does correlate to downstream performance.

The task of human mesh recovery still remains open to future methods of exploration. Aside from aspects like multiple humans or occlusions not specifically covered in this work, HMR errors are still quite easy to discern, and there certainly exists a small gap between the best HMR models and true motion capture detection. Deformable attention with feature upsampling improves upon other methods of similar training time, and we believe this line of work, focusing on the spatial quality of features, holds further potential for improvement. Some aspects we considered exploring in this area of informative spatial feature maps but rejected given resources and time include incorporating depth estimation and other backbones, and we call for future researchers to try these inquiries.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 3

[2] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022. 1

[3] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback, 2016. 3

[4] Stephanie Fu, Mark Hamilton, Laura Brandt, Axel Feldman, Zhoutong Zhang, and William T. Freeman. Featup: A model-agnostic framework for features at any resolution, 2024. 1, 2, 5

[5] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers, 2023. 1, 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 2

[7] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 3

[8] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose, 2018. 1

[9] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation, 2022. 1, 3

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 3

[11] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022. 2

[12] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6), oct 2015. 1

[13] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. 3

[14] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez,

Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 1

[15] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018. 3, 4

[16] Yusuke Yoshiyasu. Deformable mesh transformer for 3d human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17006–17015, 2023. 2

[17] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks, 2020. 3

[18] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection, 2021. 1, 2