ScoreHypo: Probabilistic Human Mesh Estimation with Hypothesis Scoring

Supplementary Material

We elaborate more details about our network architecture designs in Section 7. Section 8 provides additional implementation details and experimental results. Finally, limitations and ethics are discussed in Section 9.

7. Architecture Details

We use HRNet-W48 [60] as the CNN backbone. The architectures of LatentNet in HypoNet and ScoreNet share the same design, where the encoder consists of a linear layer, GroupNorm [69], LeakyReLU, and a dropout, and encodes 3D joint position $\mathbf{J}$ and twists Φ into latent feature $\mathbf{F}^p$. Subsequently, LatentNet concatenates $\mathbf{F}^p$ with the pixel-aligned sampled local feature $\hat{\mathbf{F}}^l$ and outputs the concatenated feature $\mathbf{F}$, which is then fed into HypoFormer or ScoreFormer. The HypoFormer also shares the same architecture design as ScoreFormer.

The HypoFormer/ScoreFormer is based on the Transformer-encoder architecture [65], containing $B = 6$ basic blocks. Each basic block consists of three units: a Multi-Head Self Attention (MHSA) unit, a Multi-Head Cross-Attention (CA) unit, and a Feed-Forward-Network (FFN) [65] unit. Both the MHSA and CA units are equipped with 8 attention heads. In the MHSA units, the query, key, and value inputs for the multi-head attention layer are the concatenated feature $\mathbf{F}$ from LatenNet. For CA units, the query input is the output of the MHSA units, while the key and value inputs are the global image feature $\mathbf{F}^g$. The hidden layers in the FFN units are configured with the channel number of 2048.

Finally, the decoder in HypoNet consists of one linear layer. The scorer in ScoreNet is designed as an MLP including a dropout layer, a hidden layer with 1024 channels, GroupNorm, and LeakyReLU.

8. Experiments

8.1. Datasets

H3.6M [21] Following previous works [22, 29, 30, 33], we use the SMPL parameters generated from MoSh [41]. Following standard practice [22], we evaluate the quality of 3D pose of 14 joints derived from the estimated mesh. We report Mean Per Joint Position Error (MPJPE) and PA-MPJPE in millimeters (mm). PA-MPJPE uses Procrustes algorithm [18] to align the estimates to GT poses before computing MPJPE. To evaluate the quality of 3D mesh, we also report Mean Per Vertex Error (MPVE) which can be interpreted as MPJPE computed over the whole mesh.

3DPW [66] We use the SMPL parameters obtained by using IMUs officially. Following [38, 39], we use the train set of 3DPW for model training and evaluate on the test set.

MPI-INF-3DHP [47] is a 3D pose dataset with 3D GT pose annotations. Following [22, 29, 33], we only use the GT pose annotation for 2d pose supervision (Eq. (8)) and diffusion noise supervision (Eq. (6)), since this dataset does not provide 3D mesh annotations.

COCO [40] is a large wild 2D pose dataset. We use the pseudo SMPL mesh annotations provided by [49]. Since the pseudo 3D mesh annotations are not accurate, while joints regressed from meshes align well with the image if we project them to 2D images, we only use the regressed joints for supervision.

UP-3D [32] is a wild 2D pose dataset. The 3D poses and meshes are obtained by SMPLify [6]. As the fitted meshes are not accurate, we only use the joint annotations for training.

MPII [2] is a wild 2D pose dataset. We use the pseudo SMPL mesh annotations generated by CLIFF [37]. Due to the inaccuracy of these pseudo mesh annotations, we only use the projected joints for supervision.

8.2. Implementation Details

We implement the proposed method with PyTorch [51]. All the experiments are conducted on a Linux machine with 2 NVIDIA A800-PCIE-80GB GPUs. We follow the definition of joint and twist in HybrIK [33], which involves extending the original 24 SMPL joints with 5 additional vertices for the head, feet, and hands, and defining $\varphi = 23$ twists for 23 limbs. We perform common data augmentation including random rotations, scaling, horizontal flipping, random occlusion, and color jitter, following HybrIK [33]. We train HypoNet for 50 epochs with a batch size of 160, and ScoreNet for 10 epochs with a batch size of 80. For the training of HypoNet, the loss weights were set as follows: $\lambda_{noise} = 1$, $\lambda_{\boldsymbol{\beta}} = 10$, and $\lambda_{2d} = 40$. To train ScoreNet, the loss weights are $\lambda_j = 1$, $\lambda_v = 1$, $\lambda_{rank} = 1$, and $\lambda_{2d} = 1$.

For the diffusion process in HypoNet, we set the training time range to $T = 1000$ and employ a linear schedule for β:

$$\beta_t = \beta_0 + t \cdot (\beta_T - \beta_0). \tag{15}$$

Data amount	MPVPE↓	MPJPE↓	PA-MPJPE↓
full	84.6	72.4	44.5
1/16	**89.0**	**76.3**	**45.8**
1/64	92.1	78.9	48.7
1/256	93.9	80.6	48.9

Table 5. Ablation study on different amounts of training data.

We set $\beta_0 = 0.0001$ and $\beta_T = 0.02$. During inference, we adopt the DDIM acceleration technique [58] and take 4 steps for the whole reverse diffusion process.

8.3. Additional Quantitative Results

Amount of training data We further investigate the robustness of our framework to the amount of training data. By systematically sampling the complete dataset at ratios of $1/16$, $1/64$, and $1/256$, we evaluate the performance of our framework in Table 5 obtained using ScoreNet with $M = 100$ hypotheses. Remarkably, our framework showcases its resilience to data reduction, as it maintains impressive performance even with a significantly reduced amount of training data. Notably, when utilizing only $1/256$ of the data, which corresponds to approximately 2,000 data samples, our framework continues to deliver exceptional results.

Same training data as ProHMR [28] Note that the datasets used in existing works vary, we present our results in Table 1 as a proof of concept. Additionally, we report the minMPJPE and minPA-MPJPE of M hypotheses on the 3DPW test set [66] by using the same training datasets as ProHMR [31] in Table 6. The baseline results are obtained using their official codes and instructions.

M	ProHMR [31]		HuManiFlow [56]		**Ours**	
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
5	91.4	57.0	79.5	50.7	75.3	47.4
10	88.9	55.0	75.6	47.9	71.7	45.2
25	85.1	52.1	71.9	44.5	67.8	42.5
100	80.1	48.1	65.1	39.9	62.9	39.1
200	77.9	46.5	64.5	38.8	**60.7**	**37.1**

Table 6. Comparison to the state-of-the-art probabilistic methods on 3DPW datatset [66] by using the same training datasets as ProHMR [31]. The baseline results are obtained using their official codes and instructions.

Same evaluation method as Biggs *et al.* [5] We report the results by following a common practice proposed in Biggs *et al.* [5] in Table 7, where we first generate $M = 4,096$ hypotheses and then "quantize" them to n hypotheses using K-Means. This practice reduces the variance in performance when generating the hypotheses. Our method achieves superior performance.

Quantization	Biggs *et al.* [5]		ProHMR [28]		**Ours**	
n	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
1	93.8	59.9	96.5	59.8	74.9	45.0
5	82.2	57.1	90.8	56.5	72.3	42.8
10	79.4	56.6	88.4	54.6	70.8	41.8
25	75.8	55.6	85.2	52.4	**68.8**	**39.9**

Table 7. Comparison to the state-of-the-art probabilistic methods on 3DPW datatset [66] by using the same evaluation method as in [5].

8.4. Additional Qualitative Results

Variability in hypotheses We visualize 10 hypotheses of two cases in Figure 6. For each case, we display the input image, 2D projected meshes, normalized variance for all vertices, the 3D view, and a zoom-in, from left to right. The rightmost shows the variance bar. While our estimates projected back to 2D for the visible parts are well-aligned, noticeable diversity persists in 3D (blue box). In occluded regions (*e.g.*the right hand in the left case), a wider range of reasonable estimates is observed.

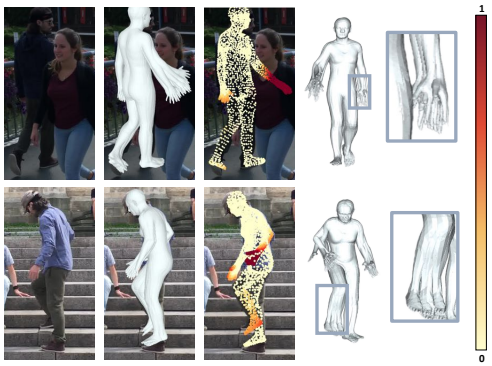

Figure 6. Visualization of variability in hypotheses with $M = 10$. Best viewed zoomed-in on a color screen.

We show additional qualitative results on challenging in-the-wild images in Figure 8. It is noticeable that the multi-hypotheses generated by HypoNet align well with the 2D observations, and the final results selected by ScoreNet are more reasonable, highlighting the robust generalization capabilities of our method. Please refer to https://xy02-05.github.io/ScoreHypo for more qualitative results.

9. Discussion

9.1. Limitations

Our method is model-based by using the SMPL model [42], which has limitations when tackling extreme body types due to the lack of training data. We show a typical failure case in Figure 7, where our estimate has a slimmer body shape.

Figure 7. A typical failure case of our method. Due to the lack of training data with diverse body shapes, our method struggles to handle extreme body shapes.

More diverse training data may alleviate this issue.

9.2. Ethics

In our experiments, we use public datasets that have IRB approval, adhering strictly to their licensing requirements. The method proposed in this paper does not violate ethical principles and has no harm to society. It is in strict compliance with relevant standards and regulations.

Figure 8. Qualitative results on challenging in-the-wild images. The yellow and blue-colored meshes are the generated results of HypoNet, while the green ones are the final results selected by ScoreNet. The last column overlaps the multiple estimates to unveil their differences.

References

[1] Jake K Aggarwal and Quin Cai. Human motion analysis: A review. pages 428–440. Elsevier, 1999. 1

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3686–3693, 2014. 6, 9

[3] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. 34:17981–17993, 2021. 3

[4] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021. 3

[5] Benjamin Biggs, David Novotny, Sebastien Ehrhardt, Hanbyul Joo, Ben Graham, and Andrea Vedaldi. 3d multi-bodies: Fitting sets of plausible 3d human models to ambiguous image data. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 20496–20507, 2020. 2, 5, 6, 10

[6] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*, pages 561–578, 2016. 2, 6, 9

[7] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, 2005. 2, 5

[8] Hanbyel Cho, Yooshin Cho, Jaesung Ahn, and Junmo Kim. Implicit 3d human mesh recovery using consistency with pose and shape from unseen-view. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21148–21158, 2023. 1, 6

[9] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *European Conference on Computer Vision (ECCV)*, pages 342–359, 2022. 1, 2, 6

[10] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *European Conference on Computer Vision (ECCV)*, pages 769–787, 2020. 5, 6

[11] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3d human mesh from in-the-wild crowded scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1475–1484, 2022. 8

[12] Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. Optimizing network structure for 3d human pose estimation. In *International Conference on Computer Vision (ICCV)*, pages 2262–2271, 2019. 4

[13] Hai Ci, Mingdong Wu, Wentao Zhu, Xiaoxuan Ma, Hao Dong, Fangwei Zhong, and Yizhou Wang. Gfpose: Learning 3d human pose prior with gradient fields. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4800–4810, 2023. 2

[14] Steffi L Colyer, Murray Evans, Darren P Cosker, and Aki IT Salo. A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system. *Sports medicine-open*, 4(1):1–15, 2018. 1

[15] Qi Fang, Kang Chen, Yinghui Fan, Qing Shuai, Jiefeng Li, and Weidong Zhang. Learning analytical posterior probability for human mesh recovery. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8781–8791, 2023. 2

[16] Lin Geng Foo, Jia Gong, Hossein Rahmani, and Jun Liu. Distribution-aligned diffusion for human mesh recovery. In *International Conference on Computer Vision (ICCV)*, pages 9221–9232, 2023. 2, 6

[17] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Košecká, and Ziyan Wu. Hierarchical kinematic human mesh recovery. In *European Conference on Computer Vision (ECCV)*, pages 768–784. Springer, 2020. 6

[18] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. 9

[19] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. 35:27953–27965, 2022. 3

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:6840–6851, 2020. 2, 3

[21] 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, 2013. 5, 6, 9

[22] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131, 2018. 1, 2, 5, 6, 9

[23] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1715–1725, 2022. 8

[24] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 8255–8263, 2023. 3

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 6

[26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[27] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. 29, 2016. 2

[28] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. Pare: Part attention regressor for 3d human body estimation. In *International Conference on Computer Vision (ICCV)*, pages 11127–11137, 2021. 6, 8, 10

[29] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *International Conference on Computer Vision (ICCV)*, pages 2252–2261, 2019. 2, 6, 8, 9

[30] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4501–4510, 2019. 2, 6, 9

[31] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *International Conference on Computer Vision (ICCV)*, pages 11605–11614, 2021. 1, 2, 5, 6, 7, 8, 10

[32] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6050–6059, 2017. 2, 6, 9

[33] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3383–3393, 2021. 2, 4, 5, 6, 9

[34] Jiefeng Li, Siyuan Bian, Qi Liu, Jiasheng Tang, Fan Wang, and Cewu Lu. Niki: Neural inverse kinematics with invertible neural networks for 3d human pose and shape estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12933–12942, 2023. 2, 6

[35] Jiahao Li, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, and Yi Yang. Jotr: 3d joint contrastive learning with transformers for occluded human mesh recovery. In *International Conference on Computer Vision (ICCV)*, pages 9110–9121, 2023. 8

[36] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13147–13156, 2022. 6

[37] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision (ECCV)*, pages 590–606, 2022. 2, 6, 9

[38] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *International Conference on Computer Vision (ICCV)*, pages 12939–12948, 2021. 2, 5, 6, 9

[39] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1954–1963, 2021. 1, 5, 6, 9

[40] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 6, 9

[41] Matthew Loper, Naureen Mahmood, and Michael J Black. Mosh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics (TOG)*, 33(6):1–13, 2014. 5, 9

[42] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):1–16, 2015. 2, 4, 5, 10

[43] Tianyu Luan, Yali Wang, Junhao Zhang, Zhe Wang, Zhipeng Zhou, and Yu Qiao. Pc-hmr: Pose calibration for 3d human mesh recovery from 2d images/videos. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 2269–2276, 2021. 2

[44] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11461–11471, 2022. 3

[45] Xiaoxuan Ma, Jiajun Su, Chunyu Wang, Wentao Zhu, and Yizhou Wang. 3d human mesh estimation from virtual markers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 534–543, 2023. 2

[46] Abed Malti. Robust monocular 3d human motion with lasso-based differential kinematics. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6607–6617, 2023. 1

[47] 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 *International Conference on 3D Vision (3DV)*, pages 506–516, 2017. 6, 9

[48] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *European Conference on Computer Vision (ECCV)*, pages 752–768, 2020. 1, 2, 6, 8

[49] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Neuralannot: Neural annotator for 3d human mesh training sets. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2299–2307, 2022. 9

[50] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*, pages 8162–8171. PMLR, 2021. 3

[51] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 9

[52] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 459–468, 2018. 5

[53] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2

[54] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning (ICML)*, pages 1530–1538. PMLR, 2015. 2

[55] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Hierarchical kinematic probability distributions for 3d human

shape and pose estimation from images in the wild. In *International Conference on Computer Vision (ICCV)*, pages 11219–11229, 2021. 2, 6

[56] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Humaniflow: Ancestor-conditioned normalising flows on so(3) manifolds for human pose and shape distribution estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4779–4789, 2023. 1, 2, 5, 6, 7, 8, 10

[57] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, pages 2256–2265. PMLR, 2015. 2

[58] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021. 3, 6, 10

[59] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021. 3

[60] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5693–5703, 2019. 6, 9

[61] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *European Conference on Computer Vision (ECCV)*, pages 529–545, 2018. 4

[62] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *International Conference on Computer Vision (ICCV)*, pages 11179–11188, 2021. 8

[63] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *International Conference on Learning Representations (ICLR)*, 2023. 3

[64] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2, 4, 7, 9

[66] Timo von Marcard, Roberto Henschel, Michael J 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)*, pages 601–617, 2018. 6, 7, 8, 9, 10

[67] Wenjia Wang, Yongtao Ge, Haiyi Mei, Zhongang Cai, Qingping Sun, Yanjun Wang, Chunhua Shen, Lei Yang, and Taku Komura. Zolly: Zoom focal length correctly for perspective-distorted human mesh reconstruction. In *International Conference on Computer Vision (ICCV)*, pages 3925–3935, 2023. 6

[68] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Photo wake-up: 3d character animation from a single photo. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5908–5917, 2019. 1

[69] Yuxin Wu and Kaiming He. Group normalization. In *European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 9

[70] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[71] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2955–2966, 2023. 3

[72] Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation. *Entropy*, 25 (10):1469, 2023. 3

[73] Yusuke Yoshiyasu. Deformable mesh transformer for 3d human mesh recovery. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17006–17015, 2023. 2, 6

[74] Mihai Zanfir, Andrei Zanfir, Eduard Gabriel Bazavan, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Thundr: Transformer-based 3d human reconstruction with markers. In *International Conference on Computer Vision (ICCV)*, pages 12971–12980, 2021. 6

[75] Wang Zeng, Wanli Ouyang, Ping Luo, Wentao Liu, and Xiaogang Wang. 3d human mesh regression with dense correspondence. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7054–7063, 2020. 2

[76] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *International Conference on Computer Vision (ICCV)*, pages 11446–11456, 2021. 1, 6, 8

[77] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7376–7385, 2020. 8

[78] Ce Zheng, Xianpeng Liu, Guo-Jun Qi, and Chen Chen. Potter: Pooling attention transformer for efficient human mesh recovery. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1611–1620, 2023. 6

[79] Wentao Zhu, Zhuoqian Yang, Ziang Di, Wayne Wu, Yizhou Wang, and Chen Change Loy. Mocanet: Motion retargeting in-the-wild via canonicalization networks. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 3617–3625, 2022. 1

[80] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *International Conference on Computer Vision (ICCV)*, pages 15085–15099, 2023. 1

[81] Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1