Supplementary material for STRIDE: Single-video based Temporally Continuous Occlusion Robust 3D Pose Estimation

1. Network Architecture

$\mathcal{M}$ contains two key components: 1) a *spatial block* to capture the orientation of joints, and 2) a *temporal block* to model the temporal dynamics of a joint. The spatial block refines poses in each frame, while the temporal block smooths the transitions between frames. We describe these components below:

Spatial block. This block utilizes *Spatial Multi-Head Self-Attention* (S-MHSA) to model relationships among joints within each pose in the input sequence. Mathematically, the S-MHSA operation is defined as:

$$\text{S-MHSA}(\mathbf{Q}_\text{S}, \mathbf{K}_\text{S}, \mathbf{V}_\text{S}) = [\texttt{head}_1; ...; \texttt{head}_h]\mathbf{W}_\text{S}^P;$$
$$\text{head}_i = \text{softmax}(\frac{\mathbf{Q}_\text{S}^i(\mathbf{K}_\text{S}^i)^\top}{\sqrt{d_K}})\mathbf{V}_\text{S}^i$$

Here, $\mathbf{Q}_\text{S}^i, \mathbf{K}_\text{S}^i, \mathbf{V}_\text{S}^i$ denote the query, key, and value projections for the i^{th} attention head, d_k is the key dimension, and $\mathbf{W}_\text{S}^P$ is the projection parameter matrix. We apply S-MHSA to features of different time steps in parallel. The output undergoes further processing, including residual connection and layer normalization (LayerNorm), followed by a multi-layer perceptron (MLP).

Temporal block. This block utilizes *Temporal Multi-Head Self-Attention* (T-MHSA) to model the relationships between poses across time steps, thereby enabling the smoothing of the pose trajectories over the sequence. It operates similarly to S-MHSA but is applied to per-joint temporal features parallelized over the spatial dimension:

$$\text{T-MHSA}(\mathbf{Q}_\text{T}, \mathbf{K}_\text{T}, \mathbf{V}_\text{T}) = [\texttt{head}_1; ...; \texttt{head}_h]\mathbf{W}_\text{T}^P;$$
$$\texttt{head}_i = \text{softmax}(\frac{\mathbf{Q}_\text{T}^i(\mathbf{K}_\text{T}^i)^\top}{\sqrt{d_K}})\mathbf{V}_\text{T}^i$$

By attending to temporal relationships, T-MHSA produces smooth pose transitions over time.

Dual-Stream Spatio-temporal Transformer. We then use the dual-stream architecture which employs spatial and temporal Multi-Head Self-Attention mechanisms. These mechanisms capture intra-frame and inter-frame body joint interactions, necessitating careful consideration of three key assumptions: both streams model comprehensive spatio-temporal contexts, each stream specializes in distinct spatio-temporal aspects, and the fusion dynamically balances weights based on input characteristics.

2. Implementation Details

We implement the proposed motion encoder DSTformer with depth N = 5, number of heads h = 8, feature size = 512, embedding size = 512. For pretraining, we use sequence length T = 243. The pretrained model could handle different input lengths thanks to the transformer-based backbone. During finetuning, we set the backbone learning rate to be 0.1× of the new layer learning rate.

Setup. We have implemented the proposed model using PyTorch. For our experiments, we utilized a CentOS machine equipped with 4 NVIDIA 3090 GPUs, specifically designed for accelerating pretraining tasks. It's worth noting that for finetuning and inference processes, a single GPU typically proves to be more than adequate.

Pretraining. We do large scale pertaining using AMASS and Training split of Human3.6M. For the implementation of AMASS [11], we initiate the process by rendering the parameterized human model SMPL+H. Subsequently, we extract 3D keypoints using a predefined regression matrix. The extraction of 3D keypoints from the Human3.6M dataset is accomplished through camera projection. Motion clips with a length of $T = 243$ are sampled for the 3D mocap data. The input channels are set to $C_\text{in} = 3$, representing the (x, y, z) coordinate. Data augmentation is applied through random horizontal flipping.

The entire network undergoes training for a total of 90 epochs, employing a learning rate of 0.0005 and a batch size of 64, facilitated by the Adam optimizer. The weights assigned to the loss terms are parameterized by $\lambda_\text{O} = 20$. Additionally, we set the 3D skeleton masking ratio to 15%, aligning with BERT's configuration. This involves using 10% frame-level masks and 5% joint-level masks. Despite variations in the proportion of these mask types, only marginal differences are observed.

To ensure the smoothness of the noise and prevent severe jittering, we initially sample noise $\boldsymbol{z} \in \mathbb{R}^{T_K \times J}$ for $T_K = 27$ keyframes. Subsequently, we upsample it to $\boldsymbol{z}' \in \mathbb{R}^{T \times J}$ and introduce a small Gaussian noise $\mathcal{N}(0, 0.002^2)$.

3D Pose Estimation. We conduct training during the

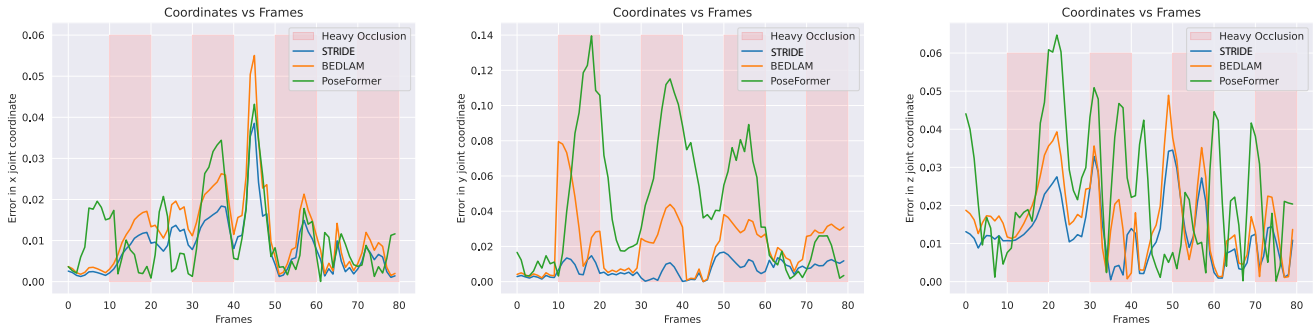


Figure 1. The figure above, from left to right, illustrates the variation in error values across the x, y, and z coordinates within a single video. Notably, STRIDE exhibits relatively lower error, particularly in scenarios involving occlusion. Furthermore, for y-coordinate, it is evident that the error demonstrates a remarkable level of smoothness.

inference stage for a duration of 30 epochs, employing the following hyperparameters:

- Batch size: 1
- Learning rate: 0.0002
- Weight decay: 0.01
- Learning rate decay: 0.99

The total loss, denoted as

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{mpjp}} + \lambda_2 \mathcal{L}_{\text{vel}} + \lambda_3 \mathcal{L}_{\text{lim}} + \lambda_4 \mathcal{L}_{\text{nmpjp}},$$

is comprised of multiple components, each weighted by specific coefficients. For this configuration, we set the weights as follows: $\lambda_1 = 1$, $\lambda_2 = 20$, $\lambda_3 = 200$, $\lambda_4 = 0.5$. These weightings contribute to the overall optimization objective, allowing for a fine-tuned balance during the training process.

3. Temporal Smoothness

The existing metric falls short in capturing temporal smoothness or assessing errors during occlusion. Additionally, there's a likelihood that a model excelling in occluded scenarios might not significantly impact overall performance if non-occluded cases dominate the results. This becomes particularly apparent in cases of sporadic temporal occlusion.

To address this issue and gain deeper insights into predictions during occlusions, we visualize various errors in Fig. 1. This plot illustrates how the error in the x, y, and z coordinates evolves in a video featuring occlusions. Notably, other methods demonstrate subpar performance during occlusions, with the error in the x and z coordinates being relatively minimal, exerting less influence on the final error. In contrast, the y-coordinate error predominantly contributes to the overall error, where STRIDE stands out

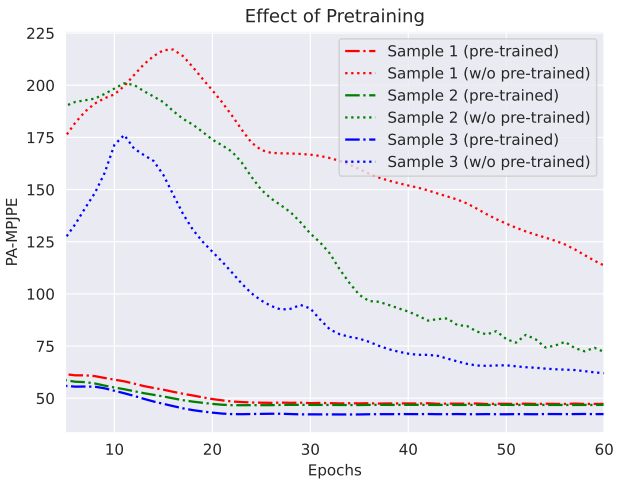


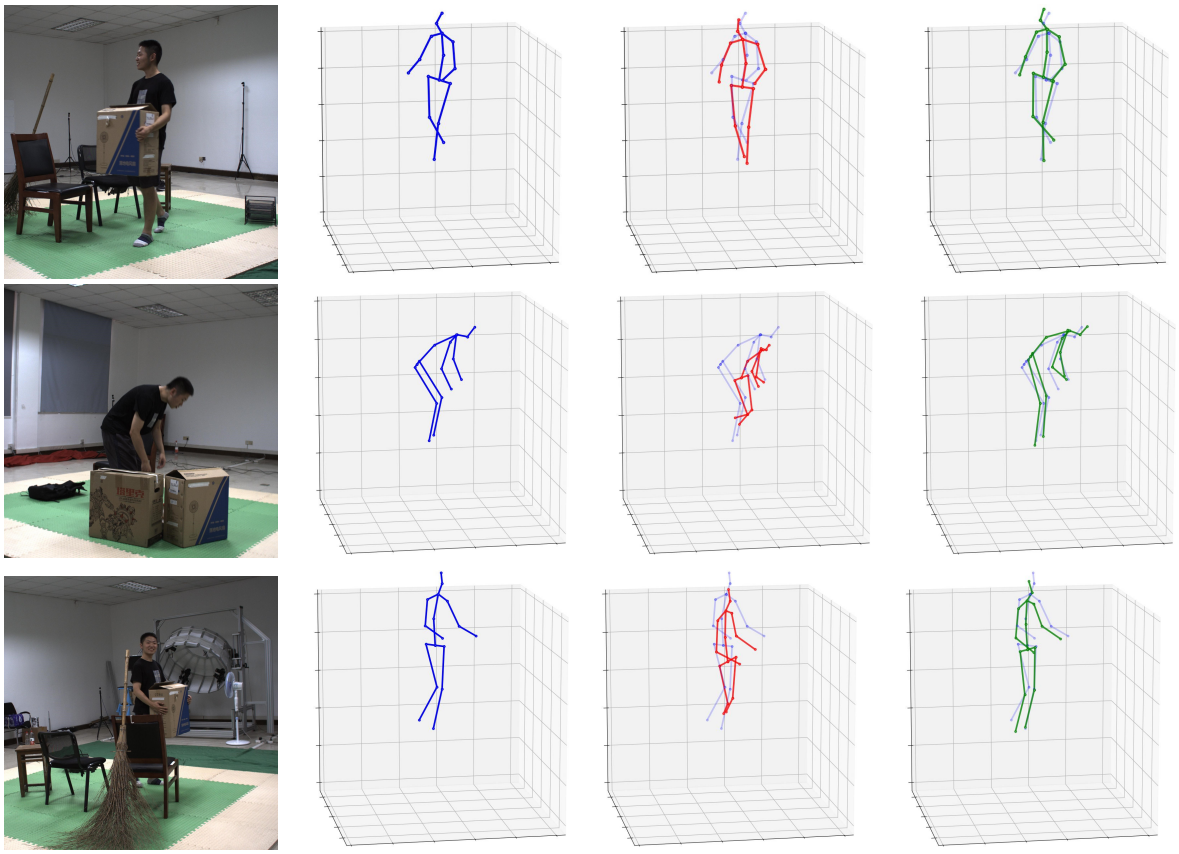
Figure 2. Effect of large-scale pre-training. We take 5 random samples from Occluded Human3.6M and try to align DSTFormer architecture. We find that when DSTFormner is initialised with motion-prior weights it converges faster.

by consistently having the least amount of error. The noteworthy aspect is the sustained and consistent performance throughout the occluded duration.

4. Additional Qualitative Results

In Fig. 3 we compare our method against a different state-of-the-art 3D pose estimation method named PoseFormerV2 [21]. We observe that STRIDE's skeleton is best aligned with the actual ground truth pose, even when there is significant occlusion.

One trivial way to improve the results of BEDLAM is by linear interpolation between frames. However, we qualitatively found that the interpolation was very smooth and misses to capture the intricate motion. Our loss optimization during inference helps to achieve the best results. In-



(a) OCMotion Image (b) Ground Truth (c) PoseFormerV2 (d) STRIDE (ours)

Figure 3. This figure shows how our method works when tested in natural occlusion cases. The translucent blue color in the *second column*, *third column*, and *fourth column* represents the ground truth. Blue, red, and green similarly represent Ground Truth, PoseformerV2 and STRIDE results, respectively.

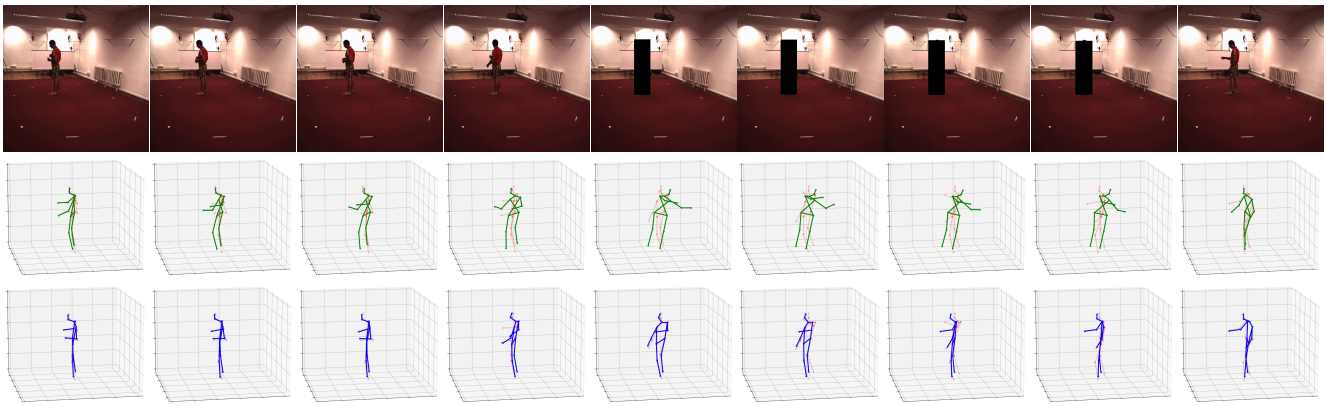


Figure 4. **3D pose estimation results on Occluded Human3.6M**. CycleAdapt (*second* row) fails to generalize in cases when there is complete occlusion. STRIDE (*third* row) produces temporally coherent pose infilling due to test time training. Note that the translucent red color represents the ground truth poses.

terpolation results are shown in Table 2.

5. Additional Quantitative Results

In our study, we conduct a comparison with the WHAM model [15], a method that has only recently been intro-

duced. WHAM employs a fully supervised training approach, benefiting from extensive datasets for its development. Notably, our method surpasses WHAM's performance on the Occluded Human3.6M dataset (refer to Table 2), demonstrating STRIDE's enhanced capability in managing significant occlusions. Additionally, we achieve results comparable to those of WHAM on the OCMotion dataset (refer to Table 1). It is important to highlight that the OCMotion dataset lacks substantial occlusions, which limits the opportunity to showcase STRIDE's strengths fully. STRIDE is particularly effective in scenarios with heavy occlusion, as evidenced by its performance on Occluded Human3.6M. Further, we also compare with 3DNBF [20], which was specially introduced to tackle occlusions while estimating the pose. It is evident from the results that STRIDE surpasses 3DNBF's performance (refer to Table 2) and also highlights the point that STRIDE performs the best when there are severe occlusions in the scene.

STRIDE is originally proposed to improve the temporal continuity of any existing pose estimation method. This makes STRIDE agnostic to any existing pose estimation method. Note that in Table 3 even if we use CLIFF, STRIDE outperforms existing SOTA methods on Occluded Human3.6.

	Method	PA-MPJPE	Accel	Avg
Image	OOH [19]	55.0	48.6	51.8
	PARE [7]	52.0	43.6	47.8
	BEDLAM [1]	47.1	49.0	48.0
Video	PoseFormerV2 [21]	126.3	28.5	77.4
	GLAMR [18]	89.9	51.3	70.6
	CycleAdapt [14]	74.6	57.5	66.0
	ROMP [16]	48.1	57.2	52.6
	SPIN† [8]	56.7	47.0	51.8
	VIBE† [6]	58.6	44.5	51.5
	WHAM [15]	**42.4**	**27.0**	**34.7**
	STRIDE (ours)	46.2	47.8	47.0

Table 1. **3D pose estimation results on OCMotion [4]**. WHAM performs better than other methods because it is a supervised method and has been trained on large amounts of data compared to STRIDE's backbone. Hence, it is able to generalize well on the OCMotion dataset.

6. Extending STRIDE for Mesh Generation and Recovery

STRIDE is originally proposed to extract the 3D pose estimation of the Human which is $(T, 17, 3)$ dimensional vector. where T is the number of frames. Extension of STRIDE to mesh recovery is trivial as we can simply train the DSTformer backbone to produce a sequence of temporally clean $(T, 24, 3)$ output from a sequence of noisy input SMPL parameters. Here $(24, 3)$ represents the θ SMPL [10] parameter shape. We can then extract human mesh using the SMPL parameters using an SMPL head. For the re-projection of the human mesh onto the image, we simply use the β and camera parameters predicted by BEDLAM-HMR/BEDLAM-CLIFF [1] and do linear interpolation/extrapolation to incorporate missing camera predictions.

	Method	PA-MPJPE	MPJPE	Accel
Image	CLIFF [9]	183.5	100.5	38.4
	BEDLAM [1]	179.5	98.9	39.1
	BEDLAM Interpolation [1]	64.1	83.3	-
	3DNBF [20]	204.3	260.4	39.3
Video	GLAMR [18]	213.9	380.3	42.3
	PoseFormerV2 [21]	193.9	260.2	38.7
	CycleAdapt [14]	77.6	132.6	48.7
	MotionBERT [23]	76.1	112.8	28.7
	WHAM [15]	119.5	237.7	46.8
	STRIDE (ours)	**59.0**	**80.7**	**26.6**

Table 2. 3D Pose estimation results on **Occluded Human3.6M**. This dataset is crucial as it is the only dataset that has significant occlusion. The results underscore that STRIDE surpasses all state-of-the-art including WHAM with substantial percentage improvements, affirming its robustness in handling occlusions.

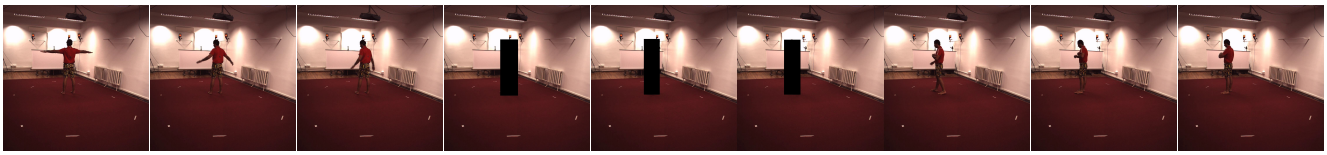
Method	Occluded H36M		Human3.6M		OCMotion	
	PA-MPJPE	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE	Accel
CLIFF	183.5	100.5	**39.4**	**62.9**	54.2	56.8
STRIDE (CLIFF)	**64.8**	**82.1**	40.1	63.2	**52.1**	**54.2**
BEDLAM	179.5	98.9	50.9	70.9	47.1	49.0
STRIDE (BEDLAM)	**59.0**	**80.7**	**50.4**	**69.7**	**46.2**	**47.8**

Table 3. Effect of various 3D Pose estimation method on STRIDE. We observe that we gain similar performance improvement if we have use any other backbone as well.

7. Additional Dataset Details

Human3.6M [5]. An indoor-scene dataset, Human3.6M is a pivotal benchmark for 3D human pose estimation from 2D images. Following [1], we retain every 1 in 5 frames in the test split comprising the S9 and S11 sequence. We perform experiments on the original publically available Human3.6M dataset to show that our method achieves comparable performance with other state-of-the-art methods.

OCMotion [4]. OCMotion is a video dataset that extends the 3DOH50K image dataset [19], incorporating natural occlusions. The dataset comprises 300K images captured at 10 FPS, featuring 43 sequences observed from 6 view-



Frame 1 Frame 6 Frame 9 Frame 10 Frame 16 Frame 19 Frame 20 Frame 26 Frame 29

Figure 5. **Samples of Occluded Humans3.6M dataset**. We add artificial occlusions on Human3.6M dataset that persist both spatially and temporally, covering the subject up to 100%. Here the person between the Frame 10 to Frame 19 remains occluded.

points. Its annotations for 3D motion include SMPL, 2D poses, and camera parameters. The sequences {0013, 0015, 0017, 0019} are designated for testing. Our method does not require supervised training, so we have only used the test split when performing all experiments.

Occluded Human3.6M. We curate the Occluded Human3.6M dataset to evaluate our method, specifically designed for assessing human pose estimation under significant occlusion, unlike existing datasets such as Human3.6M, MPI-INF-3DHP [13], and 3DPW [17]. To accomplish this, we use random erase occlusions on Human3.6M videos, completely covering a person up to 100%. These occlusions persist spatially and temporally for 1.6 seconds within 3.2 seconds of the video. Some samples are shown in Figure 5.

BRIAR [3]. BRIAR is a large-scale biometric dataset featuring videos of human subjects captured in extremely challenging conditions. These videos are recorded at varying distances $i.e$ close range, 100m, 200m, 400m, 500m, and unmanned aerial vehicles (UAV), with each video lasting around 90 seconds. Most of the pose estimation methods fail on this dataset due to the extreme amount of domain shifts. Additionally, BRIAR lacks ground truth data for poses, which means evaluations of pose estimation methods on this dataset can only be qualitative, relying on visual assessments rather than quantitative metrics.

8. Additional Related Works

2D-3D human pose lifting. Modern 3D human pose estimation encounters significant challenges in generalization due to limited labeled data for real-world applications. [12] addressed this issue by breaking down the problem into 2D pose estimation and 2D to 3D lifting. Subsequently, [2] improved on this by including self-supervised geometric regularization, by synthetic data usage [24], spatio-temporal transformers [22], and frequency domain analysis [21]. [23] achieved state-of-the-art results by modelling motion priors from a sequence of 2D poses. Although these works perform well up to a certain degree, they suffer from two problems: depth ambiguity of 2D human poses, inaccurate 3D human poses if the initial 2D human poses are noisy. In contrast, we focus on 3D pose estimation in a video-based setting and does not involve any 2D-3D pose lifting.

References

[1] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion, 2023. 4

[2] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Rohith Mv, Stefan Stojanov, and James M Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5714–5724, 2019. 5

[3] David Cornett, Joel Brogan, Nell Barber, Deniz Aykac, Seth Baird, Nicholas Burchfield, Carl Dukes, Andrew Duncan, Regina Ferrell, Jim Goddard, et al. Expanding accurate person recognition to new altitudes and ranges: The briar dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 593–602, 2023. 5

[4] Buzhen Huang, Tianshu Zhang, and Yangang Wang. Object-occluded human shape and pose estimation with probabilistic latent consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 4

[5] 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. 4

[6] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 4

[7] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *ICCV*, 2021. 4

[8] 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 *ICCV*, 2019. 4

[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. 4

[10] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. *SMPL: A Skinned Multi-Person Linear Model*. Association for Computing Machinery, New York, NY, USA, 1 edition, 2023. 4

[11] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes, 2019. 1

[12] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017. 5

[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. 5

[14] Hyeongjin Nam, Daniel Sungho Jung, Yeonguk Oh, and Kyoung Mu Lee. Cyclic test-time adaptation on monocular video for 3d human mesh reconstruction. In *International Conference on Computer Vision (ICCV)*, 2023. 4

[15] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. Wham: Reconstructing world-grounded humans with accurate 3d motion, 2023. 3, 4

[16] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, 2021. 4

[17] 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. 5

[18] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras, 2022. 4

[19] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *CVPR*, 2020. 4

[20] Yi Zhang, Pengliang Ji, Adam Kortylewski, Angtian Wang, Jieru Mei, and Alan L Yuille. 3D-Aware Neural Body Fitting for Occlusion Robust 3D Human Pose Estimation. In *The IEEE/CVF International Conference on Computer Vision*, 2023. 4

[21] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen 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*, pages 8877–8886, 2023. 2, 4, 5

[22] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming 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. 5

[23] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15085–15099, 2023. 4, 5

[24] Yue Zhu and David Picard. Decanus to legatus: Synthetic training for 2d-3d human pose lifting. In *Proceedings of the Asian Conference on Computer Vision*, pages 2848–2865, 2022. 5