Occluded Human Body Capture with Frequency Domain Denoising Prior —Supplementary Material—

Buzhen Huang[1,2] Chongyang Xu[3] Wentao Tang[4] Yuan Shu[1] Jingyi Ju[1]
Binghui Zuo[1] Yangang Wang[1]

[1]Southeast University [2]Tianjin University [3]Sichuan University [4]The University of Tokyo

In the supplementary material, we first provide additional implementation details to facilitate the reproduction of our experiments (Sec. 1). Subsequently, we present more ablation studies and results across different scenarios to further demonstrate the performance of our method (Sec. 2). Extensive qualitative results and comparisons are also included in the **supplementary video**. Finally, we discuss the broader impacts of our work (Sec. 3).

1. Implementation details

We adopt the PyTorch implementation of DWT [5] to construct the framework, with the decomposition level set to 1. We adopt the perspective camera used in CLIFF [11], where the principal point is assumed to lie at the image center, and the focal length is estimated as $\sqrt{w^2+h^2}$, with w and h representing the full image width and height, respectively. With this camera, the estimated poses are in the global camera coordinate system. For the prior encoder, we employ two separate transformers [6] to capture long-term spatial and temporal information in the frequency domain. The decoder of the 2D branch consists of only a LayerNorm [1] and a linear layer, enabling the features encoded by the prior to represent a complete motion. The 3D decoder is also implemented as a transformer. To train the diffusion model, the number of timesteps is set to 100. We employ VitPose++-B model [18] for 2D detection, which runs at 900 FPS. During inference, we adopt the DDIM sampling strategy [14] with 5 timesteps for regression, and thus the model runs at 32.3 FPS. The model is trained using the AdamW [12] optimizer with a learning rate of 1e-4. All experiments use a batch size of 32 and are conducted on a single NVIDIA RTX 4090 GPU with 24GB of memory over 45 epochs.

Data augmentation. Due to the limited amount of human motion data, we apply data augmentation techniques to train a more generalizable model. Specifically, we use the following three strategies: 1) Mirror flipping the motion: Leveraging the symmetry of the human body, we mirror flip the motion based on the kinematic tree of the human model. 2) Sampling at different rates: We resample the original sequences at various frame rates to generate new motion data. 3) Sampling in reverse order: By inverting the original sequences, we create additional motion sequences.

Occlusion synthesis. We further synthesize realistic occlusions on non-occluded motions to generate diverse occluded data. The synthetic occluded motions are used for training both the prior and the 3D decoder. Specifically, we calculate the bounding box of a motion clip and synthesize occlusions based on the intersection-over-union (IoU) between the bounding box and the occluder. The IoU values range from 0.3 to 0.7. By utilizing the synthetic data, the trained prior and decoders can effectively learn strong prior knowledge of occluded motions, making them robust to occlusions.

Dataset split. In the OcMotion dataset, we provide standard training and testing splits. The test set consists of 3 subjects, including the sequences *0013, 0015, 0017, 0019, 0038, 0039, 0040, 0041, 0044, 0045, 0046, 0047*. All remaining sequences are used for training.

2. Extended experiments

2.1. The performance on non-occluded cases

To quantitatively evaluate our method on non-occluded data, we conducted experiments on the Human3.6M dataset to further validate its effectiveness. Since the detected 2D keypoints are generally accurate, ScoreHMR demonstrates satisfactory performance in such scenarios. As shown in Sec. 1, our method achieves competitive results on Human3.6M in terms of MPJPE.

2.2. Diffusion timesteps

We also investigate the affect of diffusion timesteps. As shown in Tab. 6, 5 timesteps achieve the optimal trade-off between efficiency and accuracy.

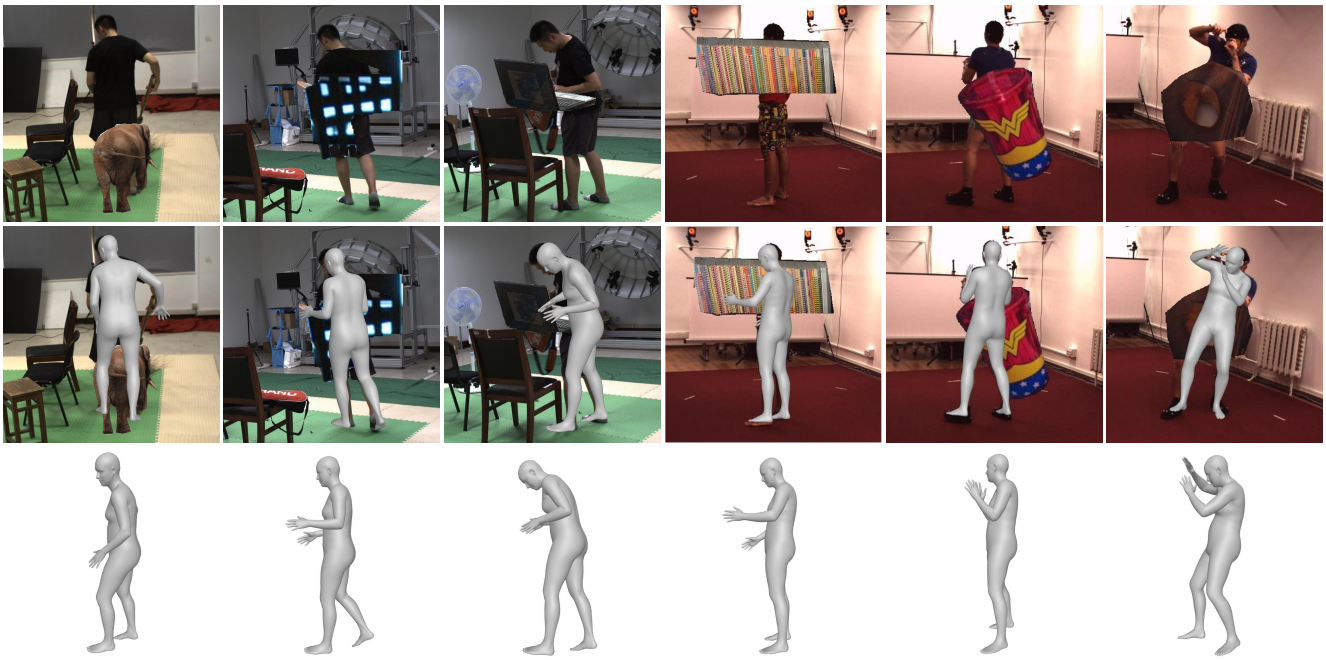

Figure 1. Results on synthetic occlusion data. Our method can obtain satisfactory results in synthetic cases.

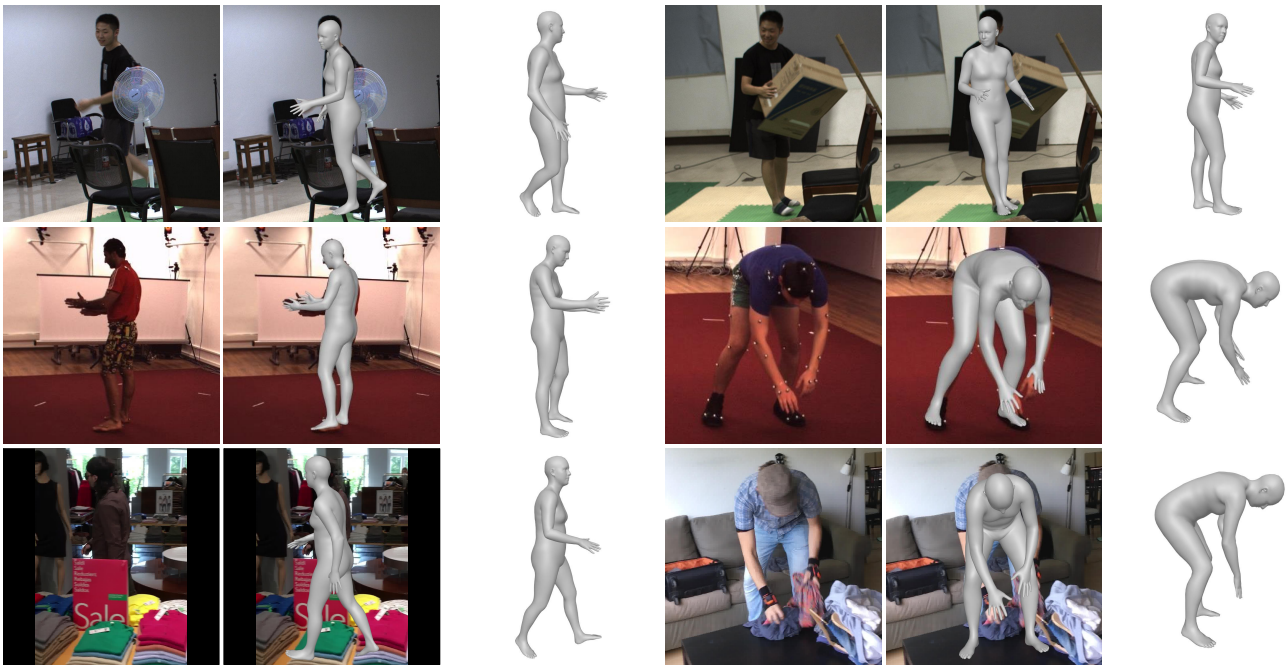

Figure 2. Our method achieves good performance in both occluded and non-occluded cases.

2.3. Window sizes and occlusion period

Temporal window size is important for frequency-based methods. A small window fails to capture sufficient motion cycles, which hinders effective frequency analysis. However, unlike DCT, we utilize DWT to capture local periodicity, which mitigates the impact of occlusions and reduces the need for a large window size. The influence of window size is summarized in Tab. 3. To balance efficiency and accuracy, we adopt a window size of 36 frames—equivalent to 3.6 seconds on OcMotion.

Occlusion often occurs during interactions. Since human interactions are dynamic, in most cases, specific body parts are not fully occluded for extended periods. In such cases, DWT can exploit local periodicity to alleviate the impact

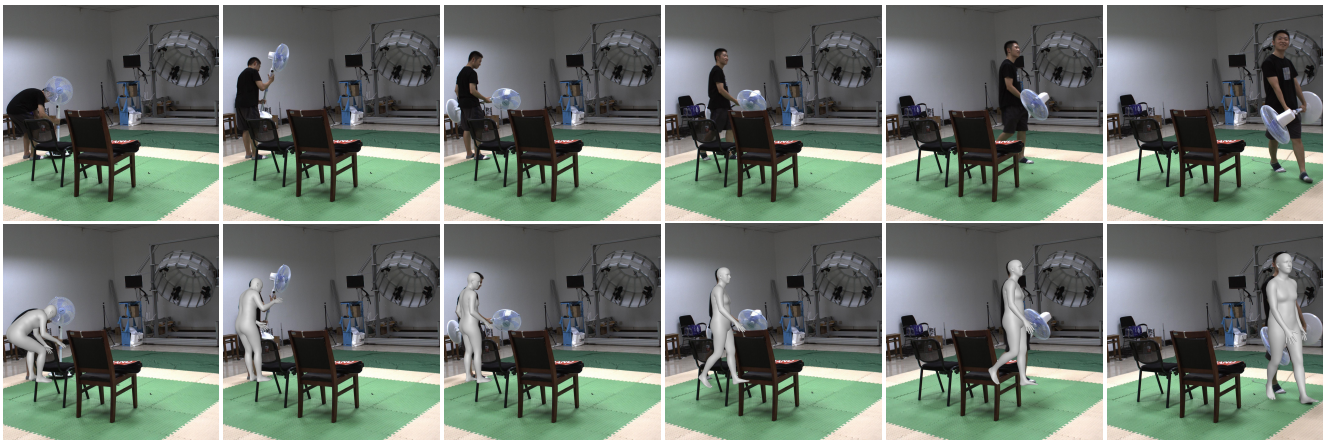

Figure 3. Qualitative results on consecutive frames in occlusion scenario. More results can refer to our supplemental video.

Figure 4. Our method can also predict satisfactory results with absolute positions in some multi-person cases.

Table 1. Quantitative comparisons on Human3.6M dataset. Our method achieves competitive performance as state-of-the-art methods.

Method	Protocol 2		
	MPJPE	PA-MPJPE	Accel.
HMMR [8]	–	56.9	–
MEVA [13]	76.0	53.2	15.3
PARE [10]	71.6	49.9	32.3
Pose2Mesh [3]	64.9	47.0	–
DSD-SATN [16]	59.1	42.4	–
VIBE [9]	65.9	41.5	27.3
OOH [21]	61.8	41.2	35.3
TCMR [4]	62.3	41.1	5.3
Chen *et al*. [2]	58.9	38.7	–
Wan *et al*. [17]	56.3	38.7	–
HMR2.0 [7]	52.8	35.6	–
ScoreHMR [15]	44.7	29.0	–
Ours	44.8	32.2	13.0

Table 2. The impact of window sizes.

Window size	8	36	81
MPJPE	52.1	48.5	46.7

of occlusions. We find that using 36 frames for both diffusion and DWT is sufficient to handle the majority of examples in the datasets. However, in our OcMotion dataset, there are indeed 2 sequences in which a certain body part may be completely occluded throughout the entire temporal window size. In such cases, it is also difficult for humans to accurately estimate the body pose from a single-view video. In these cases, our model can only estimate plausible poses for the occluded parts based on the observations of the visible regions and frequency-domain motion priors.

2.4. Occlusion ratio

We further analyzed the occlusion ratio of each joint on occluded frames in the OcMotion dataset and evaluated the

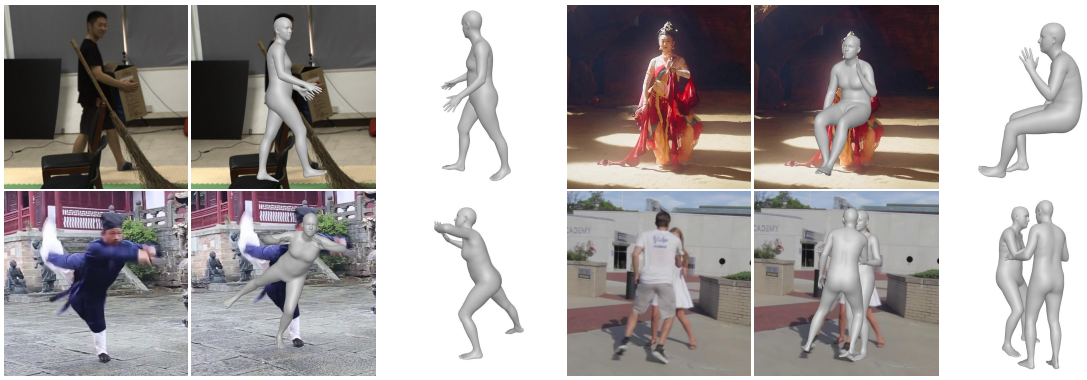

Figure 5. Failure cases. The detected incorrect 2D joints with high confidences may affect the reconstructed results.

MPJPE for each joint. In motion capture, end effectors typically exhibit larger errors, which is also consistent with our experimental results (*e.g.*, ankle and wrist). However, our evaluation shows that although the ankle has a higher occlusion ratio (34.3%) than the wrist (15.2%), their MPJPEs are 76.5 and 79.7 respectively, indicating that the ankle's accuracy is not more adversely affected. This demonstrates the effectiveness of our method in handling occlusions.

2.5. Keypoint confidence

We have conducted a performance analysis with different confidence thresholds using VitPose++-B. Setting the threshold too high may lead to the loss of valuable keypoint information, while a threshold that is too low may fail to effectively filter out invisible keypoints. The results in Tab. 4 indicate that a threshold of 0.7 achieves the best performance on the OcMotion dataset.

2.6. Qualitative reuslts

As shown in Fig. 2, our method produces satisfactory results on both occluded and non-occluded data. Leveraging the strengths of frequency modeling, the prior facilitates the recovery of accurate joint positions while preserving natural dynamics. Additional results on synthetic occlusion data are presented in Fig. 1. The first 3 columns display results from the OcMotion dataset, which features real occlusions along with realistic synthetic occluders that further increase the ambiguities. We observed that synthetic occlusions impact the accuracy of 2D pose detection; however, the frequency domain prior effectively suppresses incorrect joint coordinates. Consequently, our method can accurately recover 3D human motion by utilizing the learned prior knowledge. In the last 3 columns, we conducted experiments on the Human3.6M dataset. Since this is a non-occluded dataset, we introduced occlusions with a greater proportion. The results in Fig. 1 demonstrate that the model remains reliable even when approximately 40% of body parts are occluded.

We present additional results on in-the-wild occluded and non-occluded data in Fig. 6. The last 2 rows showcase results on internet images, demonstrating that our method performs effectively in these scenarios. Leveraging the prior, our method generates temporally coherent outputs on the occlusion dataset, as illustrated in Fig. 3. Further results on monocular videos are provided in the supplementary video.

2.7. More comparisons

RoHM [20] is an appropriate baseline method that recovers occluded motions using temporal priors through diffusion models. To quantitatively compare to RoHM, we evaluate our model on EgoBody dataset [19] using the same sequences as RoHM. The results show that our method outperforms RoHM in both occluded and non-occluded parts, which demonstrate the effectiveness of our frequency-based framework.

3. Broader Impacts

This work advances the field of human motion capture by addressing a critical and underexplored challenge: recovering 3D human motion under long-term occlusions. Our method leverages frequency-domain priors to extract periodic and physically plausible motion patterns, enabling more reliable motion reconstruction in scenarios where traditional image- or video-based approaches fail. This has promising implications for a wide range of real-world applications, including human-computer interaction, animation, sports analysis, and healthcare, where occlusions are common (e.g., crowded scenes, home environments, or clinical settings).

Furthermore, by releasing the OcMotion dataset—the first 3D human motion dataset specifically focused on occlusion—we provide a valuable resource to the community that can foster further research in robust motion estimation. While our approach enhances motion understanding under occlusion, care must be taken in its deployment to ensure ethical use, particularly in surveillance or biometric identification scenarios where privacy concerns are paramount. Overall, we believe this work contributes both technical in-

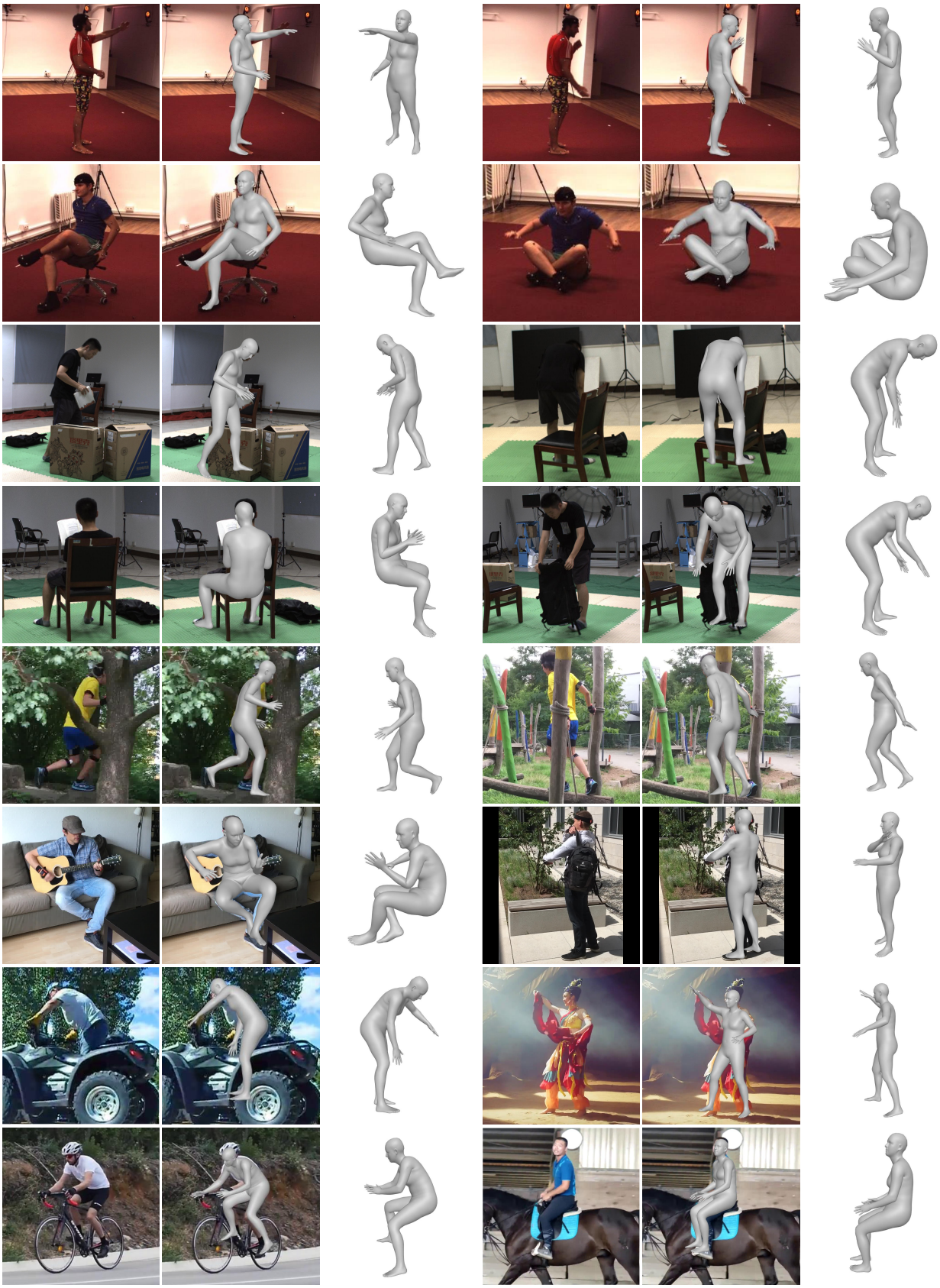

Figure 6. Results on the videos from Human3.6M, OcMotion, 3DPW and internet. Our method can achieve great performance in these cases. More qualitative results can be found in the **supplementary video**.

Table 3. The impact of occlusion ratio.

Part	R_{Ankle}	R_{Knee}	R_{Hip}	L_{Hip}	L_{Knee}	L_{Ankle}	R_{Wrist}	R_{Elbow}	$R_{Shoulder}$	$L_{Shoulder}$	L_{Elbow}	L_{Wrist}	Neck	Head
Occlusion rate (%)	33.1	20.3	20.0	20.8	19.3	35.4	14.5	10.0	6.0	6.3	11.1	16.0	5.7	42.9
MPJPE (mm)	74.6	47.1	9.7	9.7	43.5	78.3	79.6	54.6	34.2	34.7	56.7	79.8	32.0	48.2

Table 4. The impact of keypoint confidence.

Threshold	0.5	0.7	0.9
MPJPE	48.7	48.5	54.3

Table 5. Comparison to RoHM.

Method	MPJPE-vis	MPJPE-occ
RoHM	60.0	122.9
Ours	58.1	119.0

Table 6. Ablation on timestep on OcMotion.

Timestep	1	3	5	10
MPJPE	50.3	49.0	48.5	48.3

novation and infrastructure that can promote fair and responsible progress in motion capture research.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 1

[2] Yun-Chun Chen, Marco Piccirilli, Robinson Piramuthu, and Ming-Hsuan Yang. Self-attentive 3d human pose and shape estimation from videos. *CVIU*, 213:103305, 2021. 3

[3] 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 *ECCV*, 2020. 3

[4] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *CVPR*, 2021. 3

[5] Fraser Cotter. pytorch_wavelets: A pytorch implementation of the dwt and idwt, 2019. Accessed: 2025-01-22. 1

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[7] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 3

[8] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *CVPR*, 2019. 3

[9] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 3

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

[11] 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 *ECCV*, pages 590–606, 2022. 1

[12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[13] Zhengyi Luo, S Alireza Golestaneh, and Kris M Kitani. 3d human motion estimation via motion compression and refinement. In *ACCV*, 2020. 3

[14] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[15] Anastasis Stathopoulos, Ligong Han, and Dimitris Metaxas. Score-guided diffusion for 3d human recovery. In *CVPR*, pages 906–915, 2024. 3

[16] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *ICCV*, 2019. 3

[17] Ziniu Wan, Zhengjia Li, Maoqing Tian, Jianbo Liu, Shuai Yi, and Hongsheng Li. Encoder-decoder with multi-level attention for 3d human shape and pose estimation. In *ICCV*, 2021. 3

[18] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *NeurIPS*, 2022. 1

[19] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *ECCV*, pages 180–200, 2022. 4

[20] Siwei Zhang, Bharat Lal Bhatnagar, Yuanlu Xu, Alexander Winkler, Petr Kadlecek, Siyu Tang, and Federica Bogo. Rohm: Robust human motion reconstruction via diffusion. In *CVPR*, 2024. 4

[21] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *CVPR*, 2020. 3