Occluded Human Body Capture with Frequency Domain Denoising Prior

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

Abstract

Monocular human motion capture in occlusion scenarios presents significant challenges. Although a few works have explicitly considered the occlusion problem, image-based methods are unreliable due to the lack of temporal constraints while video-based approaches cannot gain sufficient knowledge from time domain motion priors to address long-term occlusions. However, occluded human motion typically exhibits periodic patterns and consistent momentum. Inspired by this observation, we exploit reliable image observations in frequency domain and formulate the motion capture task as a wavelet coefficients selection process. Specifically, we first construct probabilistic distributions for the occluded 2D keypoints, and then introduce a frequency domain diffusion model to refine the distributions by learning long-term periodic information and physical momentum with Discrete Wavelet Transform (DWT). Consequently, the learned denoising prior can select valid wavelet components to facilitate the 3D motion capture with a 3D decoder. By employing a joint reprojection strategy, we can also use the same diffusion process to train the 3D decoder. To further promote human occlusion-related tasks, we also present the first 3D occluded motion dataset, ***OcMotion****, which serves as a new benchmark for both training and evaluation. Experimental results demonstrate that our method can produce accurate and coherent human motions from occluded videos. More information is available at* https://github.com/boycehbz/FreqMotion.

1. Introduction

Recovering 3D human motion from monocular images is a long-standing problem, which has wide applications such as computer animation, human behavior understanding, and human well-being. Recently, researches in this area have gained significant progress [9, 23, 41], but most of them do not consider the occlusion scenarios that are very common in the real world.

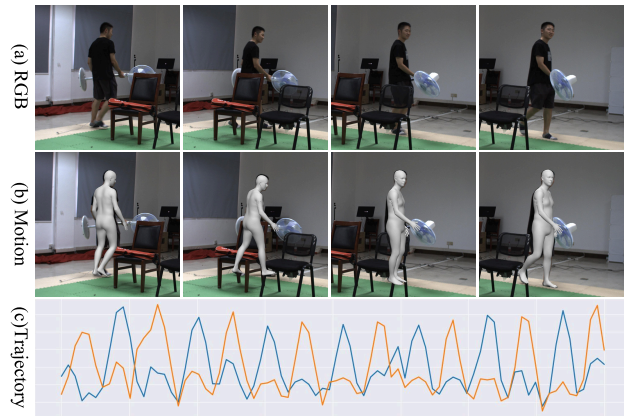


Figure 1. We visualize the trajectories along the X-axis of the left and right knee pose parameters (c) from the SMPL model in an occluded motion sequence. Despite partial occlusion, the knee joints maintain periodic and consistent patterns, which help alleviate the effects of long-term occlusion.

Only a few works explicitly focus on occluded 3D human pose estimation. For example, techniques for human representation [13, 14, 42], data augmentation [24, 37] and training strategy [26, 57] have been proposed to improve performance under occlusion scenarios. Nonetheless, the results from these image-based methods are often unreliable due to the lack of temporal constraints. Recently, some works [15, 50, 52, 53] directly train motion priors to compensate for the occluded parts with temporal consistency, but the time domain priors cannot provide sufficient knowledge for long-term occlusions and thus often lead to over-smoothed motions. Therefore, human motion capture in severely occluded scenarios remains an open issue in the community.

However, as shown in Fig. 1, we found that the occluded parts of human motion often follow periodic patterns and maintain consistent momentum. Based on this observation, we formulate the occluded human motion capture as a wavelet coefficient selection process, and learn

[1]This work was supported in part by the National Natural Science Foundation of China (No. 62076061), in part by the Natural Science Foundation of Jiangsu Province (No. BK20220127), in part by the Postgraduate Research&Practice Innovation Program of Jiangsu Province (No. SJCX25_0080). Corresponding author: Yangang Wang, yangangwang@seu.edu.cn.

complex spatio-temporal dependencies of the occluded motion in the frequency domain with a denoising prior. In contrast to phase-based representation [39, 40] and Discrete Cosine Transform (DCT) [17, 58], we adopt Discrete Wavelet Transform (DWT) to capture the correlations among different frequency components, which can address non-stationary signals (*e.g.*, sudden motions) and is more robust to both low- and high-frequency noises in occluded motions.

Specifically, given an occluded human video, we first use an off-the-shelf 2D pose estimation framework [47] to detect human keypoints. Since occlusions induce severe pixel-level ambiguities, only the visible keypoints are accurate, while the occluded parts are not reliable. We therefore construct distributions to model the uncertainty associated with the occluded keypoints and then introduce a diffusion model to refine the distributions using the visible observations. To fully capture the complex spatio-temporal dependencies, we design the diffusion model in the frequency domain, which incorporates periodic information to mitigate the impact of long-term occlusions. As shown in Fig. 2, the noisy occluded keypoints drawn from the distributions are combined with the visible keypoints as input to the model, which are then decomposed into multiple wavelet subbands using DWT. Next, the diffusion model with a 2D decoder is enforced to select the valid wavelet coefficients, and the noisy components caused by the occlusions are removed during the selection process. The filtered coefficients are then reconstructed into clean keypoints via inverse Discrete Wavelet Transform (iDWT). Subsequently, the predicted keypoints are used to generate the occluded keypoints in the next diffusion timestep. Once the diffusion model is trained, we freeze the network parameters of the encoder, and use a 3D decoder to predict the 3D wavelet subbands under the guidance of image features. The output subbands can also be mapped to 3D motions using iDWT. We also project the 3D joints onto the 2D image plane, and thus the same diffusion process can be used to train the 3D decoder. Finally, we can use the trained encoder and 3D decoder to regress 3D human motions from occluded videos over several diffusion timesteps. In addition, to further promote the network training and evaluation, we build **OcMotion**, the first video-based 3D occluded motion dataset, which contains 43 sequences with 6 views and more than 300K frames. The proposed dataset may serves as a new benchmark for human occlusion-related tasks. To sum up, the contributions of this paper are as follows:

- We formulate occluded human motion capture as a wavelet coefficient selection process that incorporates local periodic information to alleviate the impact of occlusions.
- We propose a two-branch frequency domain denoising framework to exploit reliable image observations to learn complex spatio-temporal dependencies for occluded human motion capture.
- We build OcMotion, the first 3D occluded motion dataset, which contains 300K images captured in real occlusion scenarios. The dataset is suitable for both training and testing.

2. Related Work

Occluded 3D human pose estimation. Although the 3D human pose estimation has progressively developed in recent years, it still cannot achieve satisfactory performance in occlusion scenarios. Historically, there are few works [14, 24] that explicitly focus on occluded human pose estimation. Recently, some works [13, 37, 42] regresses the occluded human from a single image, but they are not flexible enough to exploit temporal information. In addition, the lack of sufficient real occluded training data has been a bottleneck for regression-based methods over an extended period. To improve occlusion-robustness, [37] use synthetic occlusion data during training. To reduce the gap between synthetic and real occlusions, [14] represents the occluded human in the UV map and builds the first image-based object-occluded human dataset. DPMesh [59] employs pre-trained diffusion models to compensate for missing information. Other works [26, 57] alleviate the ambiguous occluded features with contrastive learning. Nonetheless, previous works that do not employ temporal information cannot obtain reliable results. Without the motion data for training, existing methods [15, 36, 39] can only utilize the temporal information based on motion priors via a time-consuming optimization. In this work, we extend the dataset proposed by [14] to have 43 real occluded 3D motions with complete and accurate annotations, thus we can train a temporal model for real-world occlusion problems. We also leverage frequency domain denoising prior to recover human motion from reliable observations, which is more robust to long-term occlusions.

Video-based monocular human mesh recovery. With the development of deep learning, temporal neural networks [20] are applied to learn the dependencies among frames. However, due to the limited training data, they use pseudo ground-truth labels for training, which are unreliable for modeling accurate 3D human motion. [4, 23, 30, 50] follow [20] to use the static features generated by image-based methods [25] for modeling temporal relations. These methods depend heavily on the static features and ignore the kinematic information, which results in severe temporal inconsistency [23] and motion oversmoothness [4]. [35, 45] models the kinematic and temporal relations by attention mechanism, which has been proved useful for human mesh recovery from occluded inputs [45]. Diffusion model [41] is also used to refine temporal consistency for

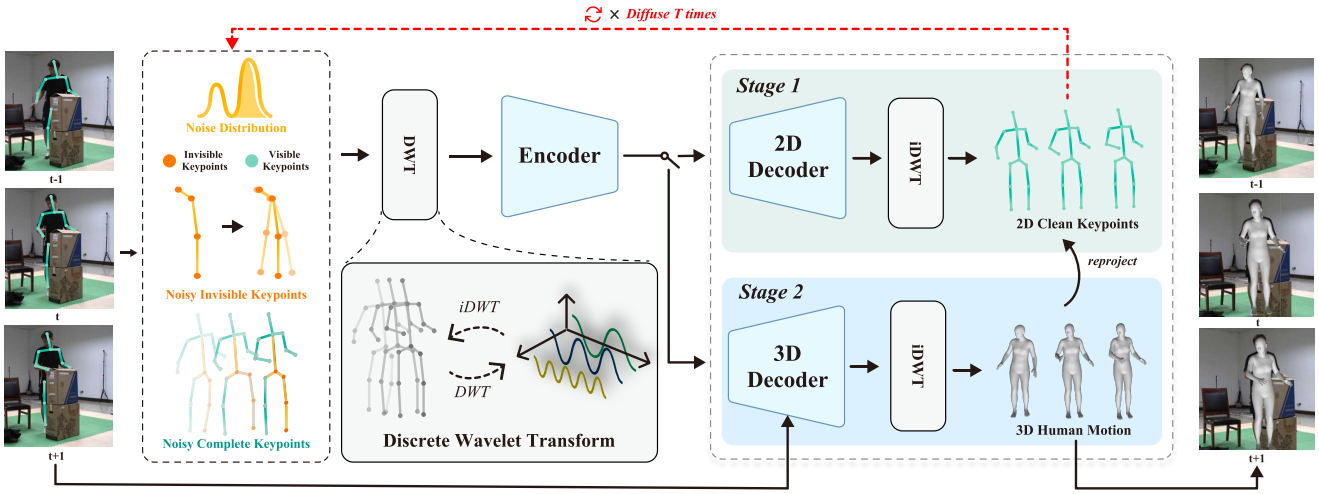


Figure 2. The pipeline of our method. Given an occluded video, we first detect 2D keypoints and model the noisy keypoints in the occluded regions using Gaussian distributions. We then combine the visible and noisy invisible keypoints, and decompose them into multiple wavelet subbands using DWT. Subsequently, we design a diffusion model with a 2D decoder to select valid frequency components for reconstructing the clean data. Once the prior model is trained, we employ the encoder with a 3D decoder to facilitate 3D motion capture within the same diffusion process. Finally, the reconstructed motion can be regressed from the input keypoints and images after several diffusion time steps.

human motion. Although previous methods achieve competitive results on specific datasets, they cannot achieve satisfactory results on long-term occlusion scenarios since the temporal consistency in time domain cannot provide sufficient information. In contrast, our method conducts denoising process in frequency domain, which incorporates periodic information to alleviate the occlusions.

Human motion analysis in frequency domain. The spectral decomposition in frequency domain simplifies the periodic characteristics analysis of human motion and thus promotes a lot of motion generation tasks, such as motion editing [3], motion prediction [12, 40, 56], and style transferring [51]. Since the frequency domain can better reveal the natures of human motion, a few recent works [17, 39, 43, 58] employ frequency domain representations in human motion capture to improve the reconstruction performance [6, 11]. Specifically, FTCM [43] adopt Fast Fourier transform (FFT) to capture global correlations for different pose frequencies. Other works [17, 58] leverage Discrete Cosine Transform (DCT) to prevent high-frequency motion jittering in the reconstructed results by directly discarding high-frequency DCT coefficients, which cannot consider local periodicity and may lose a lot of detailed motion information. In addition, these methods cannot distinguish valid signals from image observations and thus fail to alleviate long-term noises. PhaseMP [39] learns motion phase manifolds to constrain the occluded motions with test-time optimization, which confront severe depth ambiguity. Discrete Wavelet Transform (DWT) has also shown its strengths in motion generation tasks [2, 8]. However, since the wavelet coefficient selection process in pure 3D space cannot effectively distinguish reliable signals from visual observation, directly applying DWT to 3D motions like MotionWavelet [8] may result in inconsistencies with image observations in motion capture scenarios. In contrast to these works, our method denoises clean motion information from reliable partial image observations with a 2D frequency domain prior, which is more robust to long-term occlusions and high-frequency noises.

3. Method

3.1. Preliminaries

We adopt SMPL model [29] with 6D representation to describe 3D motions, and thus the parameters for an N-frame motion sequence are denoted as $\mathbf{x}^{1:N} = \{\theta^{1:N}, \tau^{1:N}\}$, which consist of pose $\theta \in \mathbb{R}^{144}$ and translation $\tau \in \mathbb{R}^3$. For each subject, the shape parameters $\beta \in \mathbb{R}^{10}$ are also included. We use $\hat{\mathbf{x}}$ to denote ground-truth data. To leverage the reliable image observations, we employ an off-the-shelf 2D detector [47] to estimate 2D keypoints $p \in \mathbb{R}^{J\times 2}$ and corresponding confidence scores $c \in \mathbb{R}^J$ for each input image.

3.2. Occluded Keypoints Uncertainty Modeling

Occluded human videos exhibit severe pixel-level ambiguities, and even the state-of-the-art human parsing models [22, 47] fail to produce reliable human semantics. In addition, the occluded regions in 2D images can correspond to multiple plausible 3D poses, which further introduces uncertainty into the reconstruction process. However, previous motion capture methods [24, 58] directly regress 3D human motions from image features or 2D keypoints, which

ignore the ambiguities and uncertainties induced by occlusions. As a result, they often fail in such situations.

To explicitly address the above two problems, we design a diffusion model that accounts for the uncertainty of occluded parts while fully exploiting the reliable image observations. Given the detected human keypoints $\{p^i\}_{i=0}^{N\times J}$ in a motion sequence, we treat the visible keypoints with high confidence as reliable predictions and directly adopt their coordinates. For the occluded joints, which often have low confidence scores and are less accurate, we model the uncertainty by constructing Gaussian distributions for keypoints with confidence below a predefined threshold. Specifically, for an occluded keypoint, we use its detected pixel coordinates p^i as the mean of the distribution $\mathcal{N}\left(p^i, \mathrm{I}\right)$, and I is an identity matrix. Thus, for a 2D pose, we represent reliable and unreliable keypoints as a set of coordinates and a set of distributions, respectively. We then leverage the reliable keypoints to refine the distributions by training a frequency-domain diffusion model.

3.3. Frequency Domain Denoising Prior

With the initial distributions, we then use the visible keypoints to reduce the uncertainty of occluded parts using a diffusion model.

Forward diffusion process. In contrast to previous diffusion-based pose estimation works [7, 10] that inject time-dependent noises towards standard Gaussian distributions, we diffuse the ground-truth 2D keypoints to the constructed initial distributions, which can leverage the prior knowledge from the 2D detector for the estimation. We thus modify the forward diffusion process as follows:

$$q(p_t \mid \hat{p}_0) = p + \sqrt{\hat{\alpha}_t}(\hat{p}_0 - p) + \sqrt{1-\hat{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}\left(0, \mathrm{I}\right), \tag{1}$$

where t is the time step. α_t is a constant hyper-parameter, and $\hat{\alpha}_t = \prod_{i=0}^{t} \alpha_i$. With this formulation, the noisy keypoint p_t follows the initial distribution $\mathcal{N}\left(p, \mathrm{I}\right)$. It should be noted that we only diffuse the unreliable keypoints, and finally combine the diffused keypoints with the reliable ones to construct the 2D pose sequence $P_t \in \mathbb{R}^{N\times 2J}$ for input to the diffusion model.

Reverse diffusion process. We then learn a diffusion model to reverse the diffusion process, starting by sampling from the initial distribution $\mathcal{N}\left(p, \mathrm{I}\right)$, which is defined as:

$$q(P_{t-1} \mid P_t) = \mathcal{N}(P_{t-1}; \mu_\alpha(P_t), \tilde{\gamma}_t \mathrm{I}), \tag{2}$$

where $\mu_\alpha(P_t)$ is the estimated mean predicted by the diffusion model at timestep $t-1$. $\tilde{\gamma}_t$ is the variance calculated using the hyper-parameters γ_t, $\hat{\alpha}_t$ and $\hat{\alpha}_{t-1}$.

Specifically, our diffusion model predicts P_0 at each timestep and then diffuse the unreliable keypoints in P_0 to p_{t-1} with Eq. (1) to construct P_{t-1}. That is, the reliable keypoints retain their original values and remain unchanged during both the forward and reverse diffusion processes. Consequently, the model can learn relationships from the reliable keypoints to refine the occluded parts.

Frequency domain diffusion model. However, predicting occluded keypoints from partially visible observations requires complex spatio-temporal dependencies, and conventional time domain priors [36, 52] tend to produce over-smoothing motions in long-term occlusion scenarios. For occluded joints, the most reasonable assumption is to preserve their original motion momentum and local periodicity. Consequently, we focus on addressing this challenging problem in the frequency domain, which reveals the periodic characteristics. In this field, a few pioneering works [12, 39] in motion analysis have adopted phase-based representations to model temporal dynamics, but they struggle to capture non-stationary signals (*e.g.*, sudden and irregular motions), which are very common in real motion data. Other works [17, 58] utilize DCT and directly discard global high-frequency components to alleviate the noises. However, since occlusion only occurs in the local area of the motion, DCT-based methods often result in suboptimal reconstructions.

To this end, we adopt DWT [2] as our frequency domain representation since it is capable of capturing local periodicity due to its multi-scale analysis. Based on this representation, we formulate the occluded human motion capture as a wavelet coefficient selection process. Specifically, given the noisy 2D keypoints P, we first apply the DWT operation along both spatial and temporal dimensions.

$$y_{h,v}\left[k_1, k_2\right] = \sum_m \sum_n P[m,n] f_h\left[m-2k_1\right] f_v\left[n-2k_2\right], \tag{3}$$

where $h, v \in \{L, H\}$. (m, n) and (k_1, k_2) are indices on the input and output map. $f_L\left[i\right] = h[i]$ and $f_H\left[i\right] = g[i]$ are low-pass and high-pass filters, respectively. Therefore, we obtain four subbands $y = \mathrm{cat}(y_{L,L}, y_{H,L}, y_{L,H}, y_{H,H})$, each representing different frequency components and spatial information of the original signal. We then design a transformer-based network to regress coefficient maps and refined wavelet from the input subbands:

$$\hat{y}_{h,v}, m_{h,v} = \mathcal{F}_{h,v}(y), \tag{4}$$

where $\mathcal{F}$ denotes the network and m is the coefficient map. The elements of m represent the scaling factors of the corresponding wavelet coefficients, which can be regarded as a selection of wavelet coefficients by applying Hadamard Product on the subband $\bar{y}_{h,v} = m_{h,v} \cdot \hat{y}_{h,v}$. Unlike conventional DWT-based image filtering, which only processes high-frequency bands, we learn a coefficient map for each subband since the occluded human motion contains both low- and high-frequency noises. Finally, we reconstruct the

2D keypoints from the filtered subbands via iDWT:

$$P[m,n] = \sum_{k_1}\sum_{k_2}\Big(\sum_{v\in\{L,H\}}\sum_{h\in\{L,H\}} \bar{y}_{h,v}[k_1,k_2]f_h[m-2k_1]f_v[n-2k_2]\Big). \quad (5)$$

The reconstructed P is then used in the next diffusion time step. We train the frequency domain prior with L1 loss on 2D keypoints:

$$\mathcal{L}_{keyp} = |P - \hat{P}|. \quad (6)$$

Once the diffusion model is trained, the prior can learn dependencies in frequency domain from the reliable keypoints to alleviate the occlusions and we use it to achieve 2D-to-3D lifting. This design ensures that only reliable signals contribute to the 3D frequency prediction. Without the 2D prior, the wavelet coefficient selection process lacks access to 2D observations and therefore cannot differentiate clean signals from noise.

3.4. Occluded Human Motion Capture

The trained prior is then used to enhance 3D occluded human motion capture. We use the same input as in Sec. 3.3 for the network, and the parameters of the encoder are frozen. As shown in Fig. 2, the encoded latent embeddings and decomposed subbands are fed into a 3D decoder in this stage, which is also a transformer network. To estimate the body shape, the RGB images are also regressed via a ViT backbone [5], and the extracted features are used as a condition for the decoder. We enforce the 3D decoder to predict shape parameters and wavelet subbands for 3D motion.

$$Y_{h,v}, \beta = \mathcal{D}(y, z, I), \quad (7)$$

where Y represents the subbands for 3D motion, and z and I are latent embedding and image features, respectively. The pose and translation parameters are then mapped from the subbands via iDWT.

$$\mathbf{x} = \text{iDWT}(Y). \quad (8)$$

We can also obtain 3D joints using the SMPL model and then reproject the joints onto the image plane to produce 2D keypoints. Consequently, the diffusion process in Sec. 3.3 can still be applied.

We train the 3D decoder following the diffusion process in Sec. 3.3 with additional loss functions, which is given by:

$$\mathcal{L} = \mathcal{L}_{smpl} + \mathcal{L}_{joint} + \mathcal{L}_{verts} + \mathcal{L}_{keyp}, \quad (9)$$

which is the sum of the supervisions from the SMPL parameters: $\mathcal{L}_{smpl} = \|[\beta,\theta] - [\hat{\beta},\hat{\theta}]\|_2^2$, 3D joint and vertex positions: $\mathcal{L}_{joint} = \|J_{3D} - \hat{J}_{3D}\|_2^2$ and $\mathcal{L}_{vert} = \|V_{3D} - \hat{V}_{3D}\|_2^2$.

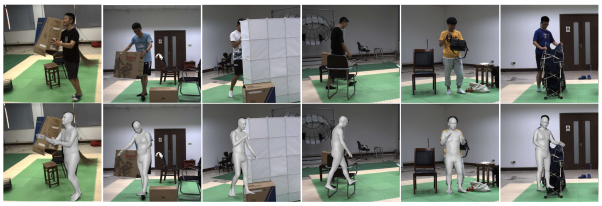


Figure 3. Samples of the proposed OcMotion dataset, which contains human videos under occlusion scenarios with accurate 3D annotations.

$\mathcal{L}_{keyp}$ is the same as in Eq. (6) but with the reprojected keypoints.

Given the input images and predicted keypoints from an occluded video, the final 3D human motion can be obtained after several diffusion time steps in the frequency domain.

4. OcMotion Dataset

Although 3D human datasets have been exponentially increasing in recent years, only a few are specifically designed for occlusion problems. AGORA [34] contains frequent occlusions, but it is a synthetic dataset. 3DOH50K [54] is the first 3D human dataset that explicitly considers object occlusion. However, it is an image dataset and cannot be used for evaluating video-based methods. In this work, we extend the 3DOH50K dataset to include complete motion annotations. We obtain human motions using [15], and the samples in 3DOH50K are also used to constrain the optimization. For severely occluded cases, we manually adjust the 3D motion. To evaluate the accuracy of the dataset, we randomly select 5K images and manually annotate 2D poses. The re-projection error on these images is 7.3 pixels, which is sufficient for motion capture tasks. Finally, the dataset contains 300K images captured at 10 FPS, 43 sequences with 6 viewpoints, 3D motion annotations represented by SMPL, 2D poses, and camera parameters.

We show the comparison with commonly used 3D human datasets in Sec. 3.3. Despite the large number of samples and the wide variety of actions, most existing datasets give little consideration to the occlusion problem. MPI-INF-3DHP [31], MuPoTs-3D [32], and Panoptic Studio [19] include only a few occluded cases. GPA [46] has a limited variety of occluders. In addition, AGORA [34] and 3DOH50K [54] are image-based datasets and cannot be applied to video-based methods. In contrast, our dataset contains complete motion annotations and explicitly considers object-occluded scenarios, which may promote future research on human-object occlusion tasks.

Table 1. Comparison with commonly used 3D human datasets. OcMotion is the first motion dataset that contains diverse real object occlusions with complete and accurate annotations.

Dataset	Occlusion Data	Sequence	Real Data	3D Pose	Mesh	Frames	Views
Human3.6M [18]	–	✓	✓	✓	–	3.6M	4
AIST++ [27]	–	✓	✓	✓	✓	10.1M	9
HUMBI [49]	–	✓	✓	✓	✓	17.3M	107
MPI-INF-3DHP [31]	+	✓	✓	✓	–	1.3M	14
3DPW [44]	++	✓	✓	✓	✓	50K	1
MuPoTs-3D [32]	++	✓	✓	✓	–	8K	1
Panoptic Studio [19]	++	✓	✓	✓	–	1.5M	480
GPA [46]	++	✓	✓	✓	–	700K	5
3DOH50K [54]	++++	–	✓	✓	✓	50K	6
AGORA [34]	++++	–	–	✓	✓	17K	–
OcMotion	++++	✓	✓	✓	✓	300K	6

5. Experiments

5.1. Dataset

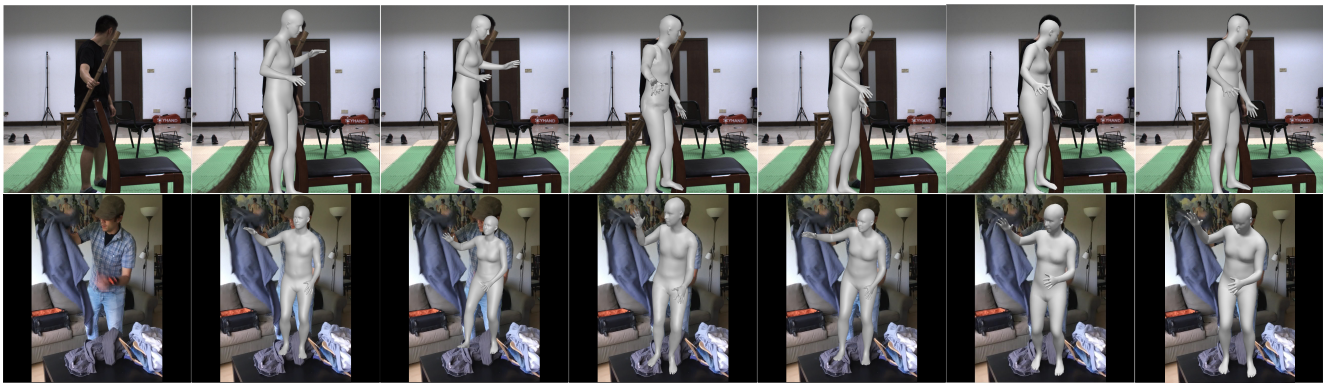
Common single-person setting. We use the same training data as previous works [4, 23] for fair comparisons. Specifically, 2D (PoseTrack [1], InstaVariety [20] and PennAction [55]) and 3D (Human3.6M [18] and MPI-INF-3DHP [31]) human motion datasets are used for training. Only the 2D keypoints from 2D datasets are used to train our 2D and 3D diffusion models. We then evaluate different methods on the standard test splits of Human3.6M, 3DPW and 3DPW-OC [14]. **Object occlusion setting.** We then conduct experiments on OcMotion dataset to investigate the performance of different methods on object-occluded scenarios, in which sequences from 3 subjects are used for testing, and the rest are used for training. **Inter-person occlusion setting.** We also use Hi4D [48], a dataset containing severe inter-person occlusions, to demonstrate the strength of our method.

Metrics. We adopt the metrics in previous works [23] to evaluate our method. The Mean Per Joint Position Error (MPJPE) and the MPJPE after rigid alignment of the prediction with ground truth using Procrustes Analysis (PA-MPJPE) are used for measuring joint positions. The Per Vertex Error (PVE) and acceleration error (Accel.) are applied to evaluate mesh quality and motion smoothness.

5.2. Comparison to state-of-the-art results

We first compare our method to state-of-the-art approaches on OcMotion dataset. To the best of our knowledge, OcMotion is the first video dataset designed for the object-occluded human mesh recovery task. We conducted experiments on this dataset to demonstrate the superiority of our method in occluded cases. As shown in Tab. 2, since previous methods do not explicitly consider the occlusion problem, our method significantly outperforms previous video-based methods on all metrics. In addition, JOTR [26], DPMesh [59], and OOH [54] are image-based methods designed for the occluded human reconstruction task. Benefiting from the spatial-temporal information, our method achieves more robust results. Moreover, while OOH [54] uses a UV map representation, which can explicitly describe an occluded human, the resampled meshes show many artifacts Fig. 4 (d). Furthermore, we found that video-based methods exhibit more motion jitters and higher acceleration errors in occluded cases. In contrast, our method incorporates periodic information from the frequency domain and alleviates the ambiguities induced by occlusions. With the denoising prior, our method achieves the best performance in terms of acceleration error and produces more temporally coherent results on the occlusion dataset. Furthermore, by leveraging the frequency representation, our method is capable of accurately recovering periodic motions even under long-term occlusions, as demonstrated in the Supplementary Video.

We further conduct experiments on the 3DPW dataset to demonstrate the strengths of our method. The results in Tab. 2 show that our method produces similar results to state-of-the-art methods on the non-occluded dataset. To further evaluate performance in more challenging occluded environments, we follow [54] to test our method on the 3DPW-OC dataset, a subset of 3DPW that contains occlusions. Our training data is the same as that used by VIBE, MEVA, and TCMR. Specifically, TCMR [4] and MEVA [30] develop time-domain motion priors to exploit temporal cues, but these priors do not provide sufficient knowledge for long-term occlusions and may lead to oversmoothed motions. Our supplementary video also clearly demonstrate that our method consistently outperforms ScoreHMR and other methods under occlusion. As a result, these methods remain sensitive to occlusions Fig. 4 (b,c,f), whereas our method is more robust. In addition, PhaseMP [39] uses a phase-based representation to incorporate frequency knowledge to address occlusions. However, this phase-based representation struggles to capture non-stationary signals, such as highly dynamic human



(a) RGB (b) VIBE (c) TCMR (d) OOH (e) PARE (f) ScoreHMR (g) Ours

Figure 4. Qualitative comparison among the methods that utilize temporal information (b, c, f) and explicitly consider the occlusion problem (d, e). Our method is more robust to occlusions than other methods.

Table 2. Quantitative comparison with state-of-the-art methods. Our method produces good results and achieves the best performance in some metrics on occluded datasets. * means the image-based method. † denotes the method that explicitly considers the occlusion problem.

Method	OcMotion			3DPW			3DPW-OC		
	MPJPE	PA-MPJPE	Accel.	MPJPE	PA-MPJPE	PVE	MPJPE	PA-MPJPE	Accel.
*†OCHMR [21]	–	–	–	89.7	58.3	107.1	–	–	–
VIBE [23]	106.3	69.1	51.6	93.5	56.5	113.4	98.3	69.7	39.0
TCMR [4]	112.9	72.7	23.7	95.0	55.8	111.5	90.3	63.0	**8.0**
*†OOH [14]	103.9	73.8	42.2	86.7	55.2	105.2	90.4	57.0	45.3
MEVA [30]	108.0	70.4	34.4	86.9	54.7	–	91.4	63.5	17.8
PSVT [35]	–	–	–	79.1	45.7	92.6	–	–	–
*†JOTR [26]	–	–	–	76.4	48.7	92.6	75.7	52.2	–
*†DPMesh [59]	–	–	–	73.6	47.4	90.7	70.9	48.0	–
†PhaseMP [39]	97.8	65.4	28.8	83.5	51.6	100.4	86.4	54.4	13.4
*HMR2.0 [9]	89.2	54.1	31.9	70.0	**44.5**	78.7	71.0	50.5	15.8
ScoreHMR [41]	81.1	53.1	24.6	68.7	47.9	76.3	65.9	46.3	10.3
GVHMR [38]	80.6	**51.6**	20.5	–	–	–	–	–	–
†**Ours** (w/o OcMotion training)	**79.2**	51.7	**20.1**	**67.3**	44.8	**75.3**	**63.1**	**45.2**	9.1

motions. Furthermore, PhaseMP adopts an optimization framework for occluded human motion capture, which also faces severe depth ambiguity. In contrast, our method, with the frequency domain prior, can obtain coherent and highly dynamic motions.

Table 3. **Comparisons on Hi4D.** Our method can achieve state-of-the-art performance on inter-person occluded cases.

Method	MPJPE	PA-MPJPE	PVE	Accel
Human4D [9]	72.1	52.4	88.6	21.1
CLIFF [28]	91.3	53.6	109.6	28.8
BEV [42]	91.8	52.2	101.2	43.5
BUDDI [33]	96.8	70.6	116.0	46.1
CloseInt [16]	63.1	47.5	76.4	19.9
Ours	**61.5**	**45.7**	**75.5**	**15.6**

We also compare our method with state-of-the-art methods on Hi4D, which contains complex two-person close interactions. The dataset exhibits severe inter-person occlusions, which induce ambiguous image features for reconstruction. Previous works [16, 33] exploit social interactions to improve the reconstruction, which requires high quality multi-person data for the training. In contrast, we fully leverage the reliable image observations and learn spatio-temporal dependencies in the frequent domain, which helps decompose the occluded motion from the complex interactions. As shown in Tab. 3, our method demonstrates competitive performance compared to multi-person methods in terms of joint accuracy and achieves better motion coherence.

5.3. Ablation study

Frequency domain denoising. We ablate the frequency domain denoising to reveal the properties of this module in occluded human motion capture. As shown in Tab. 4 (Temporal Regression), we use temporal regression as the baseline method, which directly regresses 3D human motions from extracted image features. First, we combine the ground-truth 2D keypoints with the image features to estimate 3D motions, which demonstrate that 2D keypoints can

provide valid information for reconstruction. Next, we replace the ground-truth data with predicted keypoints from ViTPose [47] (Tab. 4 (+ Predicted Keypoints)). Since the occluded videos are ambiguous, the 2D keypoints contain a lot of noise, which even affects the original performance. We also follow previous time-domain methods to enforce the model to refine the noisy keypoints with a temporal transformer (Tab. 4 (+ Denoised Keypoints)). We found that some high-frequency noise can be removed using temporal consistency. However, it cannot provide sufficient information for long-term occlusions. In contrast to time domain priors, our DWT-based transformer model learns the correlations among different frequency components, and selects valid wavelet coefficients to reconstruct the clean data, which is more robust to both low- and high-frequency noise induced by the occlusions.

DWT vs. DCT vs. Phase-based representation. There are several frequency domain representations that can be used for human motion analysis. The phase-based representation [40] takes the Fourier phase components to model temporal dynamics, which has been demonstrated to be effective in generating periodic motions like locomotion. However, real-world motions are non-stationary signals and do not persist with constant frequencies. As shown in Tab. 2, although PhaseMP gained improvements from the frequency domain knowledge, its performance is still inferior to SOTA methods, which reveals that this representation may not be appropriate for motion capture. We also compare DWT to DCT in Tab. 4. DCT decomposes the signals into different global frequency components, and previous DCT-based methods directly eliminate high-frequency components to achieve denoising. However, occlusion often occurs in a local region of the entire motion. Consequently, this method cannot effectively address occlusions and may diminish some delicate motion details. In contrast, DWT focuses on the local region and can preserve spatial and temporal information, making it more appropriate for occluded motion capture.

Frequency domain lifting. We further investigate the effectiveness of the frequency domain prior in occluded human motion capture. We first directly train a diffusion model to lift the predicted keypoints to 3D motion under the guidance of image features, and the results are shown in Tab. 4 (+ 3D Diffusion). We found its performance is similar to Tab. 4 (+ Predicted Keypoints), which reveals that time domain diffusion cannot remove the noise in the 2D detections. We then transform the predicted keypoints into the frequency domain via DWT and train the diffusion model without the pretraining from stage 1. This strategy is similar to MotionWavelet [8]. Although this method constrains the inference with frequency decomposition, it is difficult to regress 3D wavelet subbands from the noisy 2D components. Therefore, we incorporate the pretrained prior to assist in the lifting. In contrast to denoised keypoints with DWT, latent embeddings from the prior encoder contain more information and can significantly improve the 3D reconstruction. Since the 3D branch can also produce 2D keypoints through reprojection, we can also use the diffusion process to train the 3D decoder, which can further improve the accuracy.

Table 4. Ablation studies on different key components. All models are trained and evaluated on the proposed OcMotion dataset. "Frame Regression" and "Temporal Regression" refer to directly regressing SMPL parameters from image features using MLP and a temporal transformer, respectively. "+" denotes the addition of a specific module to the Temporal Regression model.

Method	MPJPE	PA-MPJPE	PVE	Accel.
Frame Regression	54.0	39.4	60.1	31.4
Temporal Regression	51.7	37.7	57.3	28.2
+ Ground-Truth Keypoints	32.5	22.2	36.1	12.5
+ Predicted Keypoints	52.3	37.7	57.8	28.2
+ Denoised Keypoints	51.4	37.7	57.2	23.3
+ Denoised Keypoints + DWT	49.6	36.4	56.9	18.8
+ Denoised Keypoints + DCT	51.6	37.5	57.3	17.6
+ 3D Diffusion	51.3	37.4	57.0	20.0
+ 3D Diffusion + DWT	51.0	36.9	56.9	19.8
+ Prior	49.1	36.3	56.8	18.5
+ Prior + 3D Diffusion	48.5	35.4	55.9	15.5

6. Limitations and future works

Although the current implementation achieves satisfactory results on single-person occluded videos, the 2D detection still affects the accuracy of 3D motion capture. For instance, in Sup. Mat. Fig. 5, the 2D joint positions of the left and right legs in the top-left case are incorrectly predicted but assigned high confidence scores. Consequently, the model regresses an erroneous 3D motion. Additionally, the training data include few challenging poses, making it difficult for the model to generalize to these cases. Future work could incorporate more low-level visual features into the estimation process and utilize enhanced behavioral priors to filter noise and mitigate such issues.

7. Conclusion

In this paper, we propose a novel frequency domain denoising prior that learns complex spatio-temporal dependencies from reliable image observations to alleviate both low- and high-frequency noises induced by occlusions. To train the prior and the lifting decoder, we introduce a diffusion process that can be applied in both the 2D and 3D stages. With the proposed prior, we formulate occluded human motion capture as a wavelet coefficient selection process, which improves the robustness and accuracy of single-view reconstruction. To promote research related to human occlusion, we further build the first 3D occluded motion dataset (OcMotion), which can be used for training and evaluation under occlusion scenarios.

References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *CVPR*, 2018. 6

[2] Philippe Beaudoin, Pierre Poulin, and Michiel van de Panne. Adapting wavelet compression to human motion capture clips. In *GI*, 2007. 3, 4

[3] Armin Bruderlin and Lance Williams. Motion signal processing. In *PACMCGIT*, 1995. 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. 2, 6, 7

[5] 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. 5

[6] Qiao Feng, Yebin Liu, Yu-Kun Lai, Jingyu Yang, and Kun Li. Fof: Learning fourier occupancy field for monocular real-time human reconstruction. *NeurIPS*, 35:7397–7409, 2022. 3

[7] Runyang Feng, Yixing Gao, Tze Ho Elden Tse, Xueqing Ma, and Hyung Jin Chang. Diffpose: Spatiotemporal diffusion model for video-based human pose estimation. In *ICCV*, pages 14861–14872, 2023. 4

[8] Yuming Feng, Zhiyang Dou, Ling-Hao Chen, Yuan Liu, Tianyu Li, Jingbo Wang, Zeyu Cao, Wenping Wang, Taku Komura, and Lingjie Liu. Motionwavelet: Human motion prediction via wavelet manifold learning. *arXiv preprint arXiv:2411.16964*, 2024. 3, 8

[9] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 1, 7

[10] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. In *CVPR*, 2023. 4

[11] Chengan He, Xin Sun, Zhixin Shu, Fujun Luan, Sören Pirk, Jorge Alejandro Amador Herrera, Dominik L Michels, Tuanfeng Y Wang, Meng Zhang, Holly Rushmeier, et al. Perm: A parametric representation for multi-style 3d hair modeling. In *ICLR*, 2025. 3

[12] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *TOG*, 36 (4):1–13, 2017. 3, 4

[13] Buzhen Huang, Tianshu Zhang, and Yangang Wang. Pose2uv: Single-shot multi-person mesh recovery with deep uv prior. *TIP*, 2022. 1, 2

[14] Buzhen Huang, Tianshu Zhang, and Yangang Wang. Object-occluded human shape and pose estimation with probabilistic latent consistency. *TPAMI*, 45(4):5010–5026, 2022. 1, 2, 6, 7

[15] Buzhen Huang, Jingyi Ju, Yuan Shu, and Yangang Wang. Simultaneously recovering multi-person meshes and multi-view cameras with human semantics. *TCSVT*, 2023. 1, 2, 5

[16] Buzhen Huang, Chen Li, Chongyang Xu, Liang Pan, Yangang Wang, and Gim Hee Lee. Closely interactive human reconstruction with proxemics and physics-guided adaption. In *CVPR*, 2024. 7

[17] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V Gehler, Javier Romero, Ijaz Akhter, and Michael J Black. Towards accurate marker-less human shape and pose estimation over time. In *3DV*, 2017. 2, 3, 4

[18] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 2013. 6

[19] H Joo, T Simon, X Li, H Liu, L Tan, L Gui, S Banerjee, T Godisart, B Nabbe, I Matthews, et al. Panoptic studio: A massively multiview system for social interaction capture. *TPAMI*, 41(1):190–204, 2017. 5, 6

[20] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *CVPR*, 2019. 2, 6

[21] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery. In *CVPR*, 2022. 7

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 3

[23] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 1, 2, 6, 7

[24] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *ICCV*, 2021. 1, 2, 3

[25] 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. 2

[26] 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 *ICCV*, pages 9110–9121, 2023. 1, 2, 6, 7

[27] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*, 2021. 6

[28] 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. 7

[29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *TOG*, 34(6):1–16, 2015. 3

[30] Zhengyi Luo, S Alireza Golestaneh, and Kris M Kitani. 3d human motion estimation via motion compression and refinement. In *ACCV*, 2020. 2, 6, 7

[31] 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 *3DV*, 2017. 5, 6

[32] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *3DV*, 2018. 5, 6

[33] Lea Müller, Vickie Ye, Georgios Pavlakos, Michael Black, and Angjoo Kanazawa. Generative proxemics: A prior for 3d social interaction from images. In *CVPR*, pages 9687–9697, 2024. 7

[34] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. Agora: Avatars in geography optimized for regression analysis. In *CVPR*, 2021. 5, 6

[35] Zhongwei Qiu, Qiansheng Yang, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Chang Xu, Dongmei Fu, and Jingdong Wang. Psvt: End-to-end multi-person 3d pose and shape estimation with progressive video transformers. In *CVPR*, 2023. 2, 7

[36] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *ICCV*, 2021. 2, 4

[37] István Sárándi, Timm Linder, Kai O Arras, and Bastian Leibe. How robust is 3d human pose estimation to occlusion? *arXiv*, 2018. 1, 2

[38] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 7

[39] Mingyi Shi, Sebastian Starke, Yuting Ye, Taku Komura, and Jungdam Won. Phasemp: Robust 3d pose estimation via phase-conditioned human motion prior. In *ICCV*, pages 14725–14737, 2023. 2, 3, 4, 6, 7

[40] Sebastian Starke, Ian Mason, and Taku Komura. Deepphase: Periodic autoencoders for learning motion phase manifolds. *TOG*, 41(4):1–13, 2022. 2, 3, 8

[41] Anastasis Stathopoulos, Ligong Han, and Dimitris Metaxas. Score-guided diffusion for 3d human recovery. In *CVPR*, pages 906–915, 2024. 1, 2, 7

[42] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *CVPR*, 2022. 1, 2, 7

[43] Zhenhua Tang, Yanbin Hao, Jia Li, and Richang Hong. Ftcm: Frequency-temporal collaborative module for efficient 3d human pose estimation in video. *TCSVT*, 34(2):911–923, 2023. 3

[44] 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 *ECCV*, 2018. 6

[45] 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. 2

[46] Zhe Wang, Liyan Chen, Shaurya Rathore, Daeyun Shin, and Charless Fowlkes. Geometric pose affordance: 3d human pose with scene constraints. *arXiv preprint arXiv:1905.07718*, 2019. 5, 6

[47] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *NeurIPS*, 2022. 2, 3, 8

[48] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *CVPR*, pages 17016–17027, 2023. 6

[49] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. Humbi: A large multiview dataset of human body expressions. In *CVPR*, 2020. 6

[50] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *CVPR*, pages 11038–11049, 2022. 1, 2

[51] M Ersin Yumer and Niloy J Mitra. Spectral style transfer for human motion between independent actions. *TOG*, 35(4): 1–8, 2016. 3

[52] Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, and Siyu Tang. Learning motion priors for 4d human body capture in 3d scenes. In *ICCV*, pages 11343–11353, 2021. 1, 4

[53] 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. 1

[54] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *CVPR*, 2020. 5, 6

[55] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *ICCV*, 2013. 6

[56] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *CVPR*, pages 3372–3382, 2021. 3

[57] Yi Zhang, Pengliang Ji, Angtian Wang, Jieru Mei, Adam Kortylewski, and Alan Yuille. 3d-aware neural body fitting for occlusion robust 3d human pose estimation. In *ICCV*, pages 9399–9410, 2023. 1, 2

[58] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In *CVPR*, pages 8877–8886, 2023. 2, 3, 4

[59] Yixuan Zhu, Ao Li, Yansong Tang, Wenliang Zhao, Jie Zhou, and Jiwen Lu. Dpmesh: Exploiting diffusion prior for occluded human mesh recovery. In *CVPR*, pages 1101–1110, 2024. 2, 6, 7