ProxyCap: Real-time Monocular Full-body Capture in World Space via Human-Centric Proxy-to-Motion Learning

Yuxiang Zhang[1], Hongwen Zhang[2]*, Liangxiao Hu[3], Jiajun Zhang[4], Hongwei Yi[5], Shengping Zhang[3], Yebin Liu[1]*
[1] Tsinghua University [2] Beijing Normal University [3] Harbin Institute of Technology
[4] Beijing University of Posts and Telecommunications [5] Max Planck Institute for Intelligent Systems

Figure 1. ProxyCap is a real-time full-body capture solution to produce accurate motions with plausible ground contact in world space.

Abstract

Learning-based approaches to monocular motion capture have recently shown promising results by learning to regress in a data-driven manner. However, due to the challenges in data collection and network designs, it remains challenging to achieve real-time full-body capture while being accurate in world space. In this work, we introduce ProxyCap, a human-centric proxy-to-motion learning scheme to learn world-space motions from a proxy dataset of 2D skeleton sequences and 3D rotational motions. Such proxy data enables us to build a learning-based network with accurate world-space supervision while also mitigating the generalization issues. For more accurate and physically plausible predictions in world space, our network is designed to learn human motions from a human-centric perspective, which enables the understanding of the same motion captured with different camera trajectories. Moreover, a contact-aware neural motion descent module is proposed to improve foot-ground contact and motion misalignment with the proxy observations. With the proposed learning-based solution, we demonstrate the first real-time monocular full-body capture system with plausible foot-ground contact in world space even using hand-held cameras.

*Corresponding authors.

1. Introduction

Motion capture from monocular videos is an essential technology for various applications such as gaming, VR/AR, sports analysis, *etc*. One ultimate goal is to achieve real-time capture while being accurate and physically plausible in world space. Despite the recent advancements, this task is still far from being solved, especially under the settings of in-the-wild captures with hand-held moving cameras.

Compared to optimization-based methods [4, 10, 12, 20, 38, 48, 62], learning-based approaches [15, 19, 68, 70] can directly regress human poses from images, significantly enhancing computational efficiency while addressing the inherent issues in optimization-based methods of initialization sensitivity and local optima entrapment. As data-driven solutions, the performance and generalization capabilities of learning-based methodologies are heavily bounded by the accuracy and diversity of the training data. Unfortunately, existing datasets are unable to simultaneously meet these requirements. On the one hand, datasets with sequential ground truth 3D pose annotations [11, 13, 28, 49] are mostly captured by marker-based or multi-view systems, which makes it hard to scale up to a satisfactory level of diversity in human appearances and backgrounds. On the other hand, numerous in-the-wild datasets [1, 23] excel in the richness of human and scenario diversity but they lack

real-world 3D motions and most of them only provide individual images instead of videos. Recently, researchers tried to create synthetic data [2, 3, 36] by rendering human avatars with controllable cameras, but it remains difficult to bridge domain gaps between the real-world images and the rendered ones, and is too expensive to scale up.

In this paper, we follow the spirit of creating synthetic data, but turn to render 2D proxy representations instead of person images. By using proxy representations (*i.e.*, silhouettes [37, 58], segmentations [16, 16, 35, 45], IUV [60, 69] and 2D skeletons [21, 26, 29, 30, 37, 50, 52, 56]), the whole motion capture pipeline can be divided into two steps: image-to-proxy extraction and proxy-to-motion lifting. In the divided pipeline, the image-to-proxy extraction is pretty accurate and robust as there are plenty of annotated 2D ground truths in real-world datasets, while the proxy-to-motion step can leverage more diverse training data to mitigate the generalization issue and reduce the domain gap. Here we adopt the 2D skeleton sequences as the proxy representation for its simplicity and high correlation with 3D motion. Meanwhile, we combine random virtual camera trajectories upon the existing large-scale motion sequence database, *i.e.*, AMASS [27], to synthesize nearly infinite data for learning proxy-to-motion lifting.

Though the proposed proxy dataset can be generated at large scales, there remain two challenges for regression-based networks to learn physically plausible motions from proxy data: how to recover i) world-space human motion under moving cameras, and ii) physically plausible motion with steady body-ground contact. For hand-held captured videos, the trajectories/rotations of humans and cameras are coupled with each other, making the recovery of world-space human motion extremely difficult. To address this issue, the latest solutions [18, 61] estimate the camera poses from the background using SfM [46, 54], and then estimate the human motion from a camera-centric perspective. However, the SfM requires texture-rich backgrounds and may fail when the foreground moving character dominates the image. Their post-processing optimization pipelines are also not suitable for real-time applications due to expensive computational costs. Besides, these previous solutions learn human motion in a camera-centric perspective like [17], which is actually ambiguous for the regression network.

In this paper, we would point out that one of the main challenges arises from the camera-centric settings in previous solutions. In such a setting, the same motion sequence captured under different cameras will be represented as multiple human motion trajectories, making it difficult for the network to understand the inherent motion prior. In contrast, we propose to learn human-centric motions to ensure consistent human motion outputs under different camera trajectories in synthetic data. Specifically, our network learns the local translations and poses in a human coordinate system, together with the relative camera extrinsic parameters in this space. After that, we accumulate the local translations of humans in each frame to obtain the global camera trajectories. Benefiting from the proposed proxy-to-motion dataset, we are able to synthesize different camera trajectories upon the same motion sequence to learn the human motion consistency. In this way, our network can disentangle the human poses from the moving camera more easily via the strong motion prior without SfM.

On top of human-centric motion regression, we further enhance the physical plausibility by introducing a contact-aware neural motion descent module. Specifically, our network first predicts coarse motions and then refines them iteratively based on foot-ground contact and motion misalignment with the proxy observations. Compared with the global post-processing optimization used in previous work [18, 61, 65], our method learns the descent direction and step instead of explicit gradient back-propagation. We demonstrate that our method, termed ProxyCap, is more robust and significantly faster to support real-time applications. To sum up, our contributions can be listed as follows:

- To tackle the data scarcity issue, we adopt 2D skeleton sequences as proxy representations and generate proxy data in world space with random virtual camera trajectories.
- We design a network to learn motions from a human-centric perceptive, which enables our regressor to understand the consistency of human motions under different camera trajectories.
- We further propose a contact-aware neural descent module for more accurate and physically plausible predictions. Our network can be aware of foot-ground contact and motion misalignment with the proxy observations.
- We contribute a real-time mocap system with plausible ground contact in world space under moving cameras.

2. Related Work

Monocular motion capture has been an active research field recently. We give a brief review of the works related to ours and refer readers to [55] for a more comprehensive survey.

Motion Capture Datasets. Existing motion capture datasets are either captured with marker-based [13, 49] or marker-less [40, 63, 73, 74] systems. Due to the requirement of markers or multi-view settings, the diversity of these datasets is limited in comparison with in-the-wild datasets. To enrich the motion datasets, numerous efforts [14, 19, 32, 34, 47] have been dedicated to generating pseudo-ground truth labels with better alignment in the image space but do not consider the motion in world space. On the other hand, researchers have also resorted to using synthetic data [36, 51, 57] by rendering human models with controllable viewpoints and backgrounds. However, such synthetic datasets are either too expensive to create or have large domain gaps with real-world images.

Proxy Representations for Human Mesh Recovery. Due to the lack of annotated data and the diversity of human appearances and backgrounds, learning accurate 3D motions from raw RGB images is still challenging. To alleviate this issue, previous approaches have exploited the different proxy representations, including silhouettes [37, 58], 2D/3D landmarks [21, 26, 29, 30, 37, 50, 52, 56], segmentation [16, 16, 35, 45], and IUV [60, 69]. These proxy representations can provide guidance for the neural network and hence make the learning process easier. However, the proxy representations simplify the observation and introduce additional ambiguity in depth and scale, especially when using proxy representations in a single frame [50, 60, 69]. In this work, we alleviate this issue by adopting 2D skeleton sequences as proxy representations and generate synthetic proxy motion data with GT world space annotations.

Full-body Motion Capture. Recent state-of-the-art approaches [16, 68] have achieved promising results for the estimation of body-only [16, 68], hand-only [22], and face-only [8] models. By combining the efforts together, these regression-based approaches have been exploited for monocular full-body motion capture. These approaches [5, 7, 33, 44, 70, 76] typically regress the body, hands, and face models by three expert networks and integrate them together with different strategies. For instance, PIXIE [7] learns the integration by collaborative regression, while PyMAF-X [70] adopts an adaptive integration strategy with elbow-twist compensation to avoid unnatural wrist poses. Despite the progress, it remains difficult for existing solutions to run at real-time while being accurate in world space. In this work, we achieve real-time full-body capture with plausible foot-ground contact by introducing new data generation strategies and novel network architectures.

Neural Decent for Motion Capture. Traditional optimization-based approaches [4] typically fit 3D parametric models to the 2D evidence but suffer from initialization sensitivity and the failure to handle challenging poses. To achieve more efficient and robust motion prediction, there are several attempts to leverage the learning power of neural networks for iterative refinement. HUND [66] proposes a learning-to-learn approach based on recurrent networks to regress the updates of the model parameters. Song *et al*. [50] propose the learned gradient descent to refine the poses of the predicted body model. Similar refinement strategies are also exploited in PyMAF [68] and LVD [6] by leveraging image features as inputs. In our work, we propose a contact-aware neural decent module and exploit the usage for more effective motion updates.

Plausible Motion Capture in World Space. Though existing monocular motion capture methods can produce well-aligned results, they may still suffer from artifacts such as ground penetration and foot skating in world space. For more physically plausible reconstruction, previous works [17, 65] have made attempts to leverage more accurate camera models during the learning process. To encourage proper contact of human meshes, Rempe *et al*. [42] propose a physics-based trajectory optimization to learn the body contact information explicitly. HuMoR [43] introduces a conditional VAE to learn a distribution of pose changes in the motion sequence, providing a motion prior for more plausible human pose prediction. LEMO [71] learns the proposed motion smoothness prior and optimizes with the physics-inspired contact friction term. Despite their plausible results, these methods typically require high computation costs and are unsuitable for real-time applications. For more effective learning of the physical constraints, there are several attempts [25, 64] to incorporate the physics simulation in the learning process via reinforcement learning. However, these methods [25, 64] typically depend on 3D scene modeling due to the physics-based formulation. Recently, there are also attempts to recover camera motion via SLAM techniques [18, 61] or regress the human motion trajectory [41, 53]. Despite the progress, it remains challenging for these methods to run in real-time or produce physically plausible in world space. In our work, we achieve real-time capture with plausible foot-ground contact in world space by designing novel networks to learn human-centric motion.

3. Proxy Data Generation

To tackle the data issue, we synthesize sequential proxy-to-motion data based on 2D skeletons and their corresponding 3D rotational motions in world space. In the following, we describe the synthesis and integration of different types of labels in our proxy data, including body motions, hand gestures, and contact labels.

Body proxy data. We adopt the motion sequences from the AMASS dataset [27] to generate proxy-to-motion pairs for the body part. The AMASS dataset is a large-scale body motion sequence dataset that comprises 3,772 minutes of motion sequence data, featuring diverse and complex body poses. We downsample the motion data to 60 frames per second, resulting in 9407K frames.

Integration with hand gestures. Since the hand poses in the AMASS dataset are predominantly static, we augment the proxy data with hand poses from the InterHand [31] dataset, which contains 1361K frames of gesture data captured at 30 frames per second. We employ Spherical Linear Interpolation (Slerp) to upsample the hand pose data to 40, 50, and 60 fps and randomly integrate them with the body poses in the AMASS motion sequences.

Integration with contact labels. We calculate the continuous contact indicators ind for 3D skeletons as follows:

$$\widehat{ind}_i = Sigmoid(\frac{v_{max} - v_i}{k_v}) \cdot Sigmoid(\frac{z_{max} - z_i}{k_z}) \quad (1)$$

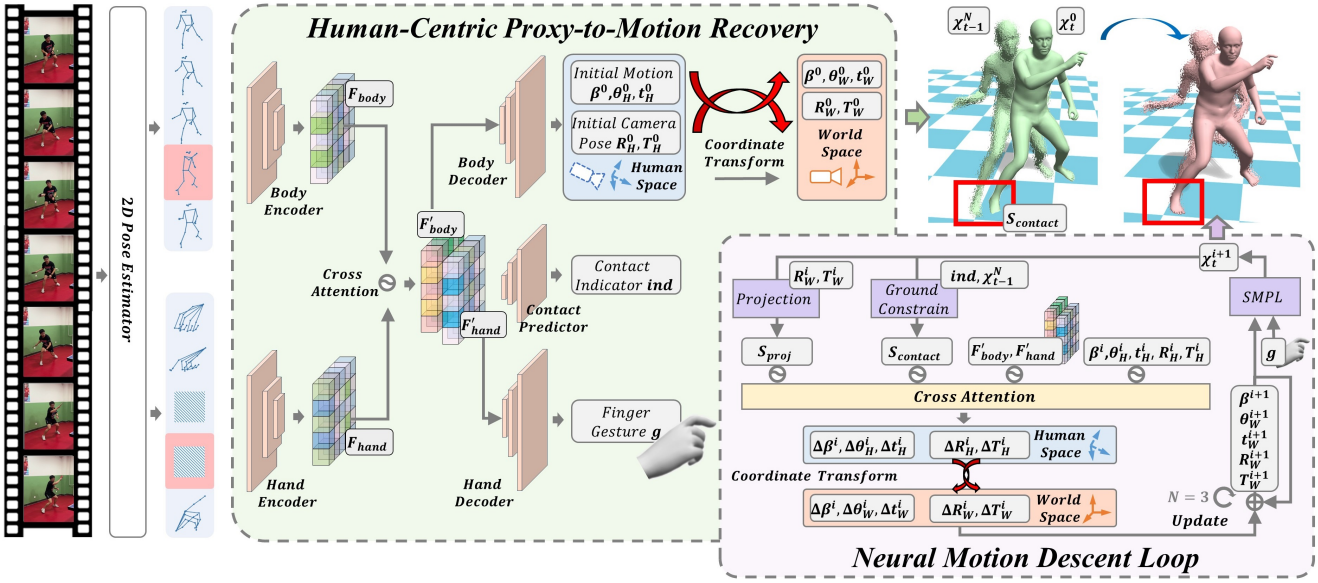

Figure 2. Illustration of the proposed method ProxyCap. Our method takes the estimated 2D skeletons from a sliding window as inputs and estimates the relative 3D motions in the human coordinate space. These local movements are accumulated frame by frame to recover the global 3D motions. For more accurate and physically plausible results, a contact-aware neural motion descent module is proposed to refine the initial motion predictions.

where v_i and z_i denote the velocity and the height to the xz-plane of the given joint. v_{max} and z_{max} is the threshold of foot sliding as $0.2m/s$ and $0.08m$, and k_v and k_z is the normalization factor as 0.04 and 0.008.

Camera setting. For each 3D motion sequence, we generate 2D proxy data under different virtual camera trajectories (four cameras in practice). Such proxy data enhances the learning of the inherent relationship between 2D proxy and 3D motions and the consistency across different viewpoints.

Specifically, we uniformly sample the field of view from $30°$ to $90°$ for the intrinsic parameters of cameras. When setting the extrinsic parameters for camera trajectories, we position the cameras at distances ranging from 1 meter to 5 meters around the human, and at heights varying from 0.5 meters to 2 meters to the ground. Finally, we generate pseudo 2D skeleton annotations by projecting 3D joints into the 2D pixel-plane using these virtual cameras. Moreover, to simulate the jitter of 2D detectors, we add Gaussian noise $\Delta X \sim \mathcal{N}(0, 0.01)$ to 3D joints before projections.

4. Method

As illustrated in Fig. 2, we first detect the 2D skeletons from images and input them into our proxy-to-motion network to recover 3D local motions in human space. These relative local motions are transformed into a world coordinate system and accumulated along the sliding window. Additionally, we leverage a neural descent module to refine the accuracy and physical plausibility. In this section, we start with introducing the human-centric setting of our motion prediction.

4.1. Human-Centric Motion Modeling

For more accurate camera modeling, we adopt the classical pinhole camera model instead of using the simplified orthogonal or weak perspective projection [9, 15, 19, 39, 68, 75]. As shown in Fig. 3, we transform the global human motion and the camera trajectories into local human space from two adjacent frames, where we adopt $\{\beta \in \mathbf{R^{10}}, \theta \in \mathbf{R^{22\times 3}}, t \in \mathbf{R^3}, g \in \{0,1\}\}_t$ to denote the parameters of shape, pose, translation and gender at frame t, and $\{R \in \mathbf{SO(3)}, T \in \mathbf{R^3}\}_t$ to denote the camera extrinsic parameters. Given the pose and shape parameters, the joints and vertices can be obtained within the SMPL Layer: $\{J, V\} = \mathcal{X}(\beta, t, \theta, g)$. In the following, we use the subscript $\boldsymbol{H}$ and $\boldsymbol{W}$ to distinguish between the human coordinate system and the world coordinate system.

During the learning of human-centric motions, we adopt a similar setting used in previous works on temporal human motion priors [24, 43, 65]. Specifically, each motion sequence will be normalized by a rigid transformation to remove the initial translation in x-z plane and rotate itself with the root orientation heading to the z-axis direction. With this setting, our network can learn the relative movements of humans, which are independent of observation viewpoints. The detailed implementation of the human-to-world coordinate transformation and the global motion accumulation in world space can be found in our supplementary material.

4.2. Sequential Full-body Motion Initialization

As shown in Fig. 2, at the first stage of our method, the skeleton sequences are processed by temporal encoders and

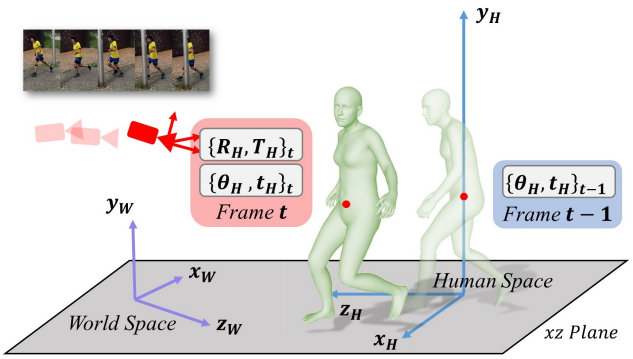

Figure 3. Illustration of decoupling the world-space motion into the human-centric coordinates and relative camera poses.

then fed into decoders for the predictions of the initial motion $\{\beta^0, \theta^0, \mathbf{t}^0\}$, initial camera $\{\mathbf{R}_\mathbf{H}^0, \mathbf{T}_\mathbf{H}^0\}$, the contact indicator $\mathbf{ind}$, and the hand poses $\mathbf{g}$. Following previous baseline [39], we build these encoders upon temporal dilated convolution networks.

For better body-hand compatibility, we exploit the cross-attention mechanism to facilitate the motion context sharing during the body and hand recovery. Specifically, we first obtain the initial body features F_{body} and hand features F_{hand} from the temporal encoders and map them as Query, Key, Value matrices in forms of $Q_{body/hand}$, $K_{body/hand}$, and $V_{body/hand}$, respectively. Then we update the body features F'_{body} and hand features F'_{hand} as follows:

$$\begin{aligned} F'_{body} &= V_{body} + Softmax(\frac{Q_{hand}K_{body}^\top}{\sqrt{d_k}})V_{body}, \\ F'_{hand} &= V_{hand} + Softmax(\frac{Q_{body}K_{hand}^\top}{\sqrt{d_k}})V_{hand}. \end{aligned} \tag{2}$$

The updated features $\{F'_{body}, F'_{hand}\}$ can be further utilized in the contact indicators ind and serve as the temporal context in the Neural Descent module, as will be described shortly. In our experiments, we demonstrate that the feature fusion in Eq. 2 can make two tasks benefit each other to produce more comparable wrist poses in the full-body model.

4.3. Contact-aware Neural Motion Descent

At the second stage of our method, the initial motion predictions will be refined to be more accurate and physically plausible with the proposed contact-aware neural motion descent module. As shown in Fig. 2, this module takes the 2D misalignment and body-ground contact status as input and updates motion parameters during iterations.

Misalignment and Contact Calculation. At the iteration of $i \in \{0, 1, ..., N\}$, we calculate the 2D misalignment status by projecting the 3D joints on the image plane and calculate the differences between the re-projected 2D joints and the proxy observations: $\mathcal{S}_{proj} = \Pi(J_i, \{K, R_H^i, T_H^i\}) - \widehat{J}_{2D}$. Here, $\Pi(\cdot)$ denotes the perspective projection function, and K denotes the intrinsic parameter.

For the contact status, we calculate the velocity of 3D joints v_{xz}^i in xz-plane and the distance to the ground as d_y^i. Moreover, we also leverage the temporal features from inputs 2D skeletons to predict the contact labels ind, which will be used as an indicator to mask the body-ground contact. Then, the contact status of the current predictions can be obtained as $\mathcal{S}_{contact} = ind \odot (v_{xz}^i, d^i)$, where $\odot$ denotes the Hadamard product operation.

Motion Update. After obtaining the contact and misalignment status, we feed them into the neural motion descent module for motion updates. As shown in Fig. 4, the descent module takes the two groups of tensors as input: i) the state group includes the current SMPL parameters in the human coordinate system $\beta^i, t_H^i, \theta_H^i$, camera pose R_H^i, T_H^i and the sequential motion context $F_{seq} = \{F'_{body}, F'_{hand}\}$; ii) the deviation group includes the current misalignment status $\mathcal{S}_{proj}$ and contact status $\mathcal{S}_{contact}$.

A straightforward solution would be using an MLP to process these two groups of tensors. However, the values of these two groups exhibit significant differences. For instance, the values of the state tensors change smoothly while the values of the deviation tensors may change rapidly along with the misalignment and contact status. Simply concatenating them as inputs introduces difficulty in the learning process. Note that the magnitude of the deviation tensors is highly correlated with the parameter updates. When the body model is well-aligned without foot skating or ground penetration, the values of the deviation tensors are almost zero, indicating that the refiner should output zeros to prevent further changes in the pose parameters. Otherwise, the refiner should output larger update values for motion adjustments. To leverage such a correlation property, we exploit a cross-attention module to build a more effective architecture.

As shown in Fig. 4, two fully connected layers are leveraged to process the tensors of the state and deviation groups and produce the Query, Key, and Value for the cross-attention module. In this way, our contact-aware neural motion descent module can effectively learn the relationship between the state and deviation groups and hence produce more accurate motion updates. Moreover, the sequential motion context F_{seq} is also leveraged in our neural descent module to mitigate the depth uncertainty and improve the motion predictions.

Compared with previous work [43, 50, 66, 71], the proposed contact-aware neural motion descent module offers the advantage of freeing us from the heavy cost of explicit gradient calculations or the manual tuning of hyper-parameters during testing. Furthermore, the module is capable of learning human motion priors with contact information from our synthetic dataset, which provides a more suitable descent direction and steps to escape the local minima and achieve faster convergence.

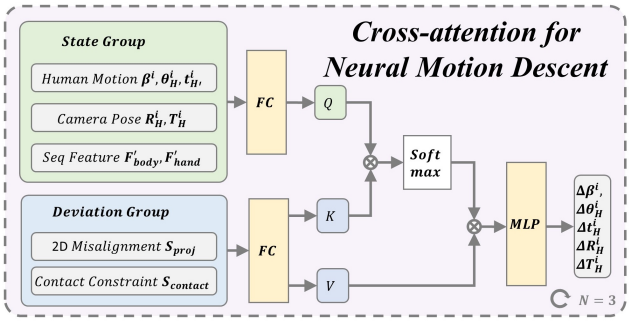

Figure 4. Implementations of the neural descent module.

4.4. Loss Function

In our solution, the full-body motion recovery module and the contact-aware neural motion descent module are trained sequentially. Benefiting from the proxy-to-motion learning, the ground-truth full-body pose θ, g, and human body shape β_b can be obtained for supervision from our synthetic dataset. Overall, the objective function of motion recovery can be written as follows:

$$\mathcal{L}_{rec} = \mathcal{L}_{3D} + \mathcal{L}_{2D} + \mathcal{L}_{\theta} + \mathcal{L}_{\beta} + \mathcal{L}_{cam} + \mathcal{L}_{consist} + \mathcal{L}_{smooth} \tag{3}$$

Specifically, $\mathcal{L}_{3D}$ involves 3D MPJPE loss and 3D trajectory L1 loss while $\mathcal{L}_{2D}$ is the projected 2D MPJPE loss. $\mathcal{L}_{\theta}, \mathcal{L}_{\beta}, \mathcal{L}_{cam}$ represents L1 loss between the estimated human pose, shape and camera pose to our synthetic ground truth. $\mathcal{L}_{consist}$ is a L1 loss to restrict the consistency of the local motion outputs θ_H, t_H of the same 3D motion sequence via different observations from virtual cameras. $\mathcal{L}_{smooth}$ is adopted from [67] by penalizing the velocity and acceleration between the estimation and the ground truth. For the neural descent module, the objective loss can be written as:

$$\begin{cases} \mathcal{L}_{desc} = \sum_k u^{N-k}(\mathcal{L}_{rec} + \mathcal{L}_{contact}) \\ \mathcal{L}_{contact} = \sum_i ind^{gt} \odot (||v_{xz}||_2 + ||d_y||_2) \\ \mathcal{L}_{ind} = \sum_i Entropy(ind^{gt}, ind^{est}) \end{cases} \tag{4}$$

where $k = 1, 2, ..., N$ is the iteration time and u is the decay ratio to emphasize the last iteration. We set iteration counts $K = 3$, decay rate $u = 0.8$ in practice. $\mathcal{L}_{contact}$ involves the error of trajectory drifting, foot floating, or ground penetration. $\mathcal{L}_{ind}$ refers to the loss between the predicted contact label to the ground truth.

5. Experiments

In this Section, we validate the efficacy of our method and demonstrate accurate human motion capture results with physically plausible foot-ground contact in world space.

Dataset. The RICH dataset [11] is collected with a multi-view static camera system and one moving camera that the ground truth 3D human motions can be recovered using spatial stereo geometry. The EgoBody [72] is captured by a multi-camera rig and a head-mounted device, focusing on the social interactions in 3D scenes. Dynamic Human3.6M is a benchmark proposed in [65] to simulate the moving cameras on Human3.6M [13] by randomly cropping with a small view window around the person.

Table 1. Quantitative comparison on EgoBody [72] and RICH [11]. The symbol † denotes the methods relying on SLAM.

Methods	W-MPJPE ↓	WA-MPJPE ↓	PA-MPJPE ↓	ACCEL ↓
EgoBody dataset				
† SLAHMR [61]	**141.1**	101.2	79.13	25.78
† PACE [18]	147.9	**101.0**	**66.5**	**6.7**
GLAMR [65]	416.1	239.0	114.3	173.5
Ours	**385.5**	**131.3**	**73.5**	**49.6**
RICH dataset				
†SLAHMR [61]	571.6	323.7	52.5	9.4
†PACE [18]	**380.0**	**197.2**	**49.3**	**8.8**
GLAMR [65]	653.7	365.1	79.9	107.7
Ours	**629.8**	**343.6**	**56.0**	**25.3**

Metrics. In our experiments, we follow previous work [65] to report various metrics, primarily focusing on the evaluation of the world coordinate system. The WA-MPJPE metric reports the MPJPE after aligning the entire trajectory of both the predicted and GT through Procrustes Alignment. The W-MPJPE metric reports the MPJPE after aligning the first frame of sequences. The PA-MPJPE metric reports the MPJPE error after applying the GT trajectories to each frame. The ACCEL metric evaluates the joint acceleration.

5.1. Comparison with the State of the Art

We compare our approach with the state-of-the-art approaches to human motion recovery under dynamic cameras, including GLAMR [65], SLAMHR [61] and PACE [18]. Both the SLAMHR and PACE require a pre-processing SLAM to reconstruct the scene to solve the camera trajectories (refer to Tab. 3). Such a process is time consuming and requires texture-rich backgrounds, which narrows their applications. To validate the effectiveness of the proposed solution, we primarily compare our method with GLAMR, as it also runs without SLAM. We also conduct comparison experiments on the RICH and EgoBody datasets, as shown in Tab. 1. As shown in the table, our method achieves significant improvements over previous solutions in all metrics. Visual comparisons with previous different solutions are also depicted in Fig. 6 and the video in our supplementary materials, where our method again shows superior results in terms of model-image alignment and foot-ground contact in world space.

5.2. Ablation Studies

We conduct ablation studies to validate the effectiveness of the proposed neural descent method on the Dynamic Human3.6M dataset following the setup of [65]. As shown in

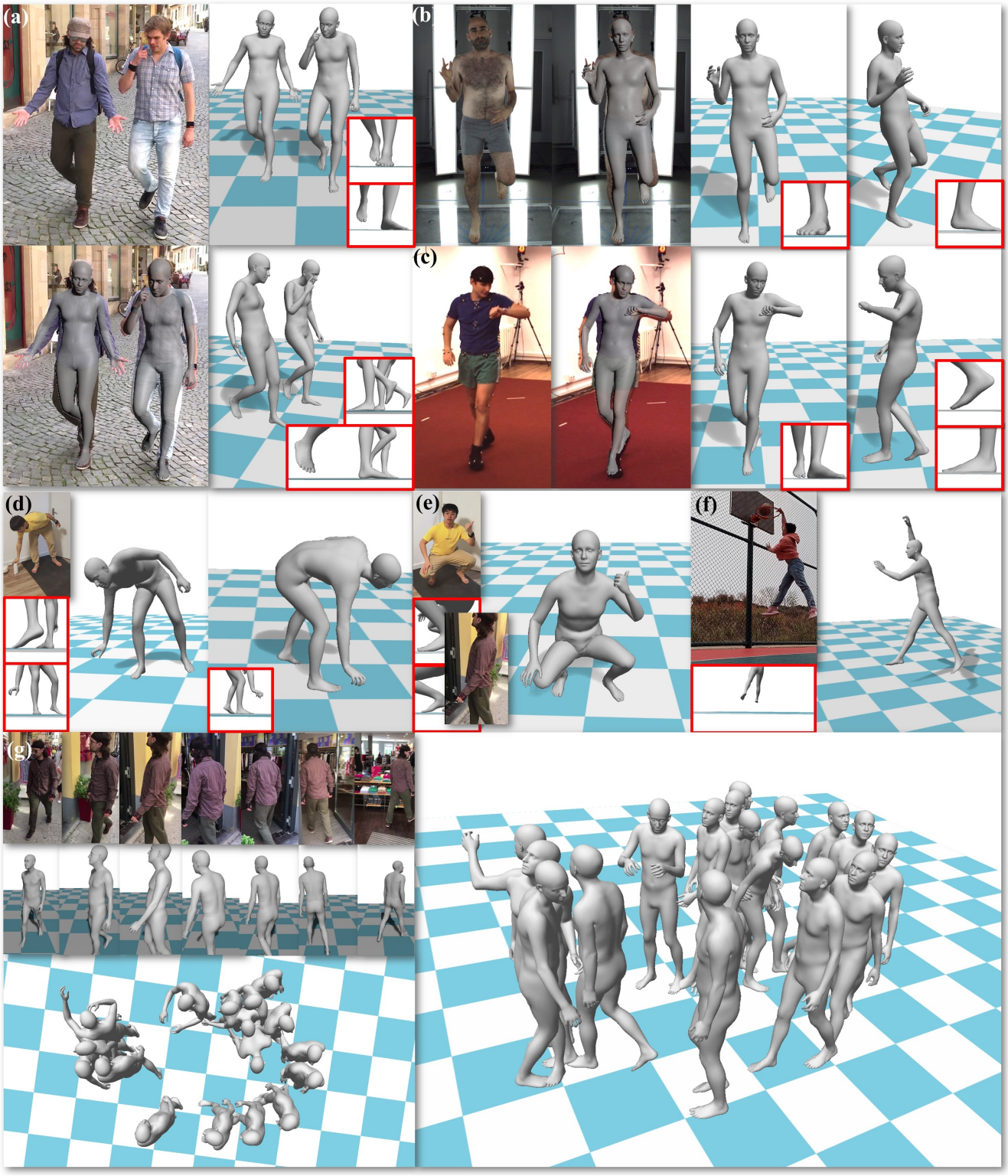

Figure 5. Results across different cases in the (a,g) 3DPW [59], (b) EHF [38], and (c) Human3.6M [13] datasets and (d,e,f) internet videos. We demonstrate that our method can recover the accurate and plausible human motions in moving cameras at a real-time performance. Specifically, (g) demonstrates the robustness and the temporal coherence of our method even under the occlusion inputs.

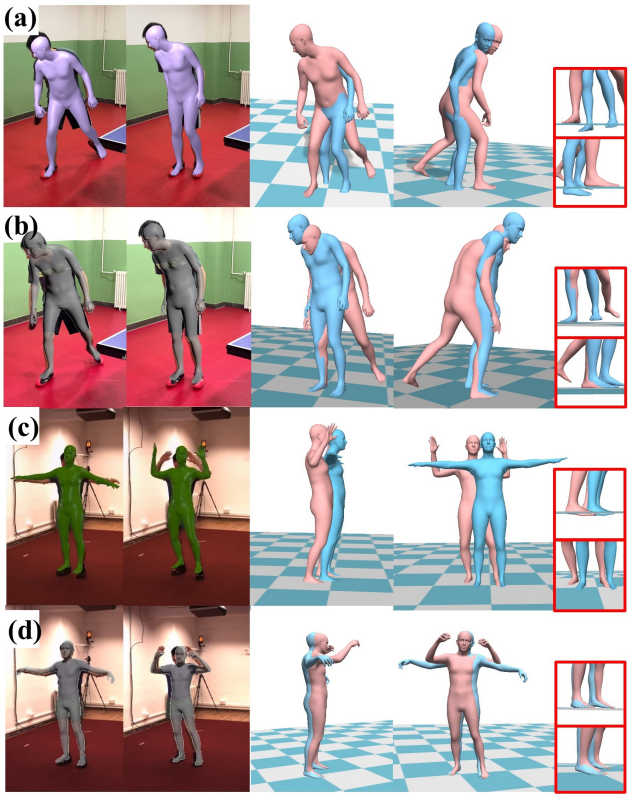

Figure 6. Qualitative comparison with previous state-of-the-art methods: (a) PyMAF-X [70], (c) GLAMR [65], (b)(d) Ours.

Table 2. Ablation study of the Neural Descent on Dynamic H36M.

Neural Descent	W-MPJPE ↓	PA-MPJPE ↓
w/o	644.8	48.4
w/	**605.4**	**45.9**

Tab 2, the Neural Descent module can significantly reduce the motion errors in world space.

We also report the metric of ground penetration [64] (GP) and foot floating (FF) in the Human3.6M [13] dataset. The GP is defined as the percentage of the frames that penetrate to the ground. The FF is defined as the percentage of frames with foot-ground distances far from the given threshold. We report the curves of GP and FF with respect to the distance to ground in Fig. 7 with a logarithmic scale, where we can conclude that the neural descent algorithm can significantly improve the ground contact plausibility.

5.3. Runtime

It is also worth noting that the proposed method has improved the speed by an order of magnitude compared to the previous methods. The comparisons of speed and complexity are reported in Tab. 3 and Tab. 4. Our method can reach real-time performance at 30 FPS in a laptop with RTX4060, which is very promising to enable various applications related to virtual humans.

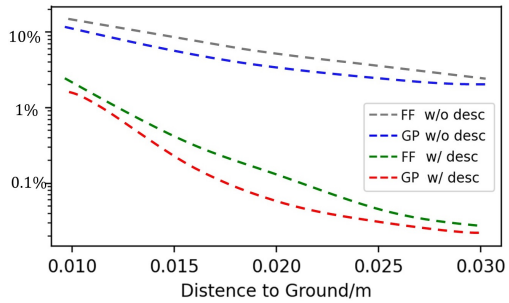

Figure 7. The ablation study on the percentage of foot floating (FF) and ground penetration (GP). We vary the threshold from 1cm to 3cm to calculate the corresponding FF and GP metrics.

Table 3. Runtime Comparison with other methods

Method	Initial Estimation	Optimization
GLAMR	2.2fps:FasterRCNN+HyberIK	5.4fps:Stage II,III,IV
SLAMHR	1.9fps:ViTDet+HMR2.0+SLAM	0.07fps:HuMoR
Ours	**60fps**:PoseDet+RTMPose	**73fps**:Neural descent

Table 4. Complexity Comparison with other methods

	GLAMR		SLAMHR		Ours		
Method	FasterRCNN	HyberIK	ViTDet	HMR2	PoseDet	RTMPose	NeuDescent
FLOPs	134G	22.7G	3600G	245G	38.9G	4.16G	33.2M
Params	41.8M	76.1M	692M	670M	24.7M	27.7M	7.3M

6. Conclusion

In this paper, we present ProxyCap, a real-time monocular full-body motion capture approach with physically plausible foot-ground contact in world space. We leverage a proxy dataset based on 2D skeleton sequences with accurate 3D rotational motions in world space. Based on the proxy data, our network learns motions from a human-centric perceptive, which enhances the consistency of human motion predictions under different camera trajectories. For more accurate and physically plausible motion capture, we further propose a contact-aware neural motion descent module so that our network can be aware of foot-ground contact and motion misalignment. Based on the proposed solution, we demonstrate a real-time monocular full-body capture system under hand-held cameras.

Limitations. As our method recovers 3D motion from 2D joint observations, the depth ambiguity issue remains especially when the person is captured in nearly static poses, and we assume that people move in a planar ground with foot contact, so we can not handle motions such as going upstairs, boardskating, etc.

Acknowledgement. This paper is supported by National Key R&D Program of China (2022YFF0902200), the NSFC project No.62125107, the Beijing Municipal Science&Technology Z231100005923030, and the Fundamental Research Funds for the Central Universities.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, pages 3686–3693, 2014. 1

[2] Eduard Gabriel Bazavan, Andrei Zanfir, Mihai Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Hspace: Synthetic parametric humans animated in complex environments. *arXiv preprint arXiv:2112.12867*, 2021. 2

[3] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *CVPR*, pages 8726–8737, 2023. 2

[4] 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 *ECCV*, pages 561–578. Springer, 2016. 1, 3

[5] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *ECCV*, pages 20–40. Springer, 2020. 3

[6] Enric Corona, Gerard Pons-Moll, Guillem Alenyà, and Francesc Moreno-Noguer. Learned vertex descent: A new direction for 3D human model fitting. In *ECCV*, pages 146–165, 2022. 3

[7] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Collaborative regression of expressive bodies using moderation. In *I3DV*, pages 792–804, 2021. 3

[8] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *TOG*, 40(4):88:1–88:13, 2021. 3

[9] Kehong Gong, Bingbing Li, Jianfeng Zhang, Tao Wang, Jing Huang, Michael Bi Mi, Jiashi Feng, and Xinchao Wang. Posetriplet: co-evolving 3d human pose estimation, imitation, and hallucination under self-supervision. In *CVPR*, pages 11017–11027, 2022. 4

[10] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *ICCV*, pages 1381–1388. IEEE, 2009. 1

[11] Chun-Hao P Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J Black. Capturing and inferring dense full-body human-scene contact. In *CVPR*, pages 13274–13285, 2022. 1, 6

[12] 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 *I3DV*, pages 421–430. IEEE, 2017. 1

[13] 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*, 36(7):1325–1339, 2014. 1, 2, 6, 7, 8

[14] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. In *I3DV*, pages 42–52, 2021. 2

[15] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2018. 1, 4

[16] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. PARE: Part attention regressor for 3D human body estimation. In *ICCV*, pages 11127–11137, 2021. 2, 3

[17] Muhammed Kocabas, Chun-Hao P Huang, Joachim Tesch, Lea Muller, Otmar Hilliges, and Michael J Black. SPEC: Seeing people in the wild with an estimated camera. In *ICCV*, pages 11035–11045, 2021. 2, 3

[18] Muhammed Kocabas, Ye Yuan, Pavlo Molchanov, Yunrong Guo, Michael J. Black, Otmar Hilliges, Jan Kautz, and Umar Iqbal. PACE: Human and motion estimation from in-the-wild videos. In *3DV*, 2024. 2, 3, 6

[19] 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*, pages 2252–2261, 2019. 1, 2, 4

[20] 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 *CVPR*, pages 6050–6059, 2017. 1

[21] 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 *CVPR*, pages 3383–3393, 2021. 2, 3

[22] Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. Interacting attention graph for single image two-hand reconstruction. In *CVPR*, pages 2761–2770, 2022. 3

[23] 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 *ECCV*, pages 740–755. Springer, 2014. 1

[24] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel van de Panne. Character controllers using motion vaes. *ACM Trans. Graph.*, 39(4), 2020. 4

[25] Zhengyi Luo, Shun Iwase, Ye Yuan, and Kris Kitani. Embodied scene-aware human pose estimation. In *NeurIPS*, 2022. 3

[26] Xiaoxuan Ma, Jiajun Su, Chunyu Wang, Wentao Zhu, and Yizhou Wang. 3d human mesh estimation from virtual markers. In *CVPR*, pages 534–543, 2023. 2, 3

[27] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, pages 5442–5451, 2019. 2, 3

[28] 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 *I3DV*, pages 506–516, 2017. 1

[29] 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 *ECCV*, pages 752–768. Springer, 2020. 2, 3

[30] Gyeongsik Moon and Kyoung Mu Lee. Pose2Pose: 3D positional pose-guided 3D rotational pose prediction for expressive 3D human pose and mesh estimation. *arXiv preprint arXiv:2011.11534*, 2020. 2, 3

[31] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In *ECCV*, pages 548–564. Springer, 2020. 3

[32] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. NeuralAnnot: Neural annotator for 3D human mesh training sets. In *CVPRW*, pages 2299–2307, 2022. 2

[33] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3D hand pose estimation for whole-body 3D human mesh estimation. In *CVPRW*, 2022. 3

[34] Lea Müller, Ahmed AA Osman, Siyu Tang, Chun-Hao P Huang, and Michael J Black. On self-contact and human pose. In *CVPR*, pages 9990–9999, 2021. 2

[35] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. In *I3DV*, pages 484–494. IEEE, 2018. 2, 3

[36] 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*, pages 13468–13478, 2021. 2

[37] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *CVPR*, pages 459–468, 2018. 2, 3

[38] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 1, 7

[39] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3D human pose estimation in video with temporal convolutions and semi-supervised training. pages 7753–7762, 2019. 4, 5

[40] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, pages 9054–9063, 2021. 2

[41] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people by predicting 3D appearance, location and pose. In *CVPR*, pages 2740–2749, 2022. 3

[42] Davis Rempe, Leonidas J Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *ECCV*, pages 71–87. Springer, 2020. 3

[43] 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. 3, 4, 5

[44] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. FrankMocap: A monocular 3D whole-body pose estimation system via regression and integration. In *ICCV*, 2021. 3

[45] Nadine Rueegg, Christoph Lassner, Michael Black, and Konrad Schindler. Chained representation cycling: Learning to estimate 3D human pose and shape by cycling between representations. In *AAAI*, pages 5561–5569, 2020. 2, 3

[46] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[47] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3D human pose and shape estimation in the wild. In *BMVC*, 2020. 2

[48] Leonid Sigal, Alexandru Balan, and Michael J Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *NeurIPS*, pages 1337–1344, 2008. 1

[49] Leonid Sigal, Alexandru O Balan, and Michael J Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87(1-2):4, 2010. 1, 2

[50] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In *ECCV*, pages 744–760. Springer, 2020. 2, 3, 5

[51] Jiajun Su, Chunyu Wang, Xiaoxuan Ma, Wenjun Zeng, and Yizhou Wang. Virtualpose: Learning generalizable 3d human pose models from virtual data. In *ECCV*, pages 55–71. Springer, 2022. 2

[52] 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*, pages 5349–5358, 2019. 2, 3

[53] Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J Black. TRACE: 5D temporal regression of avatars with dynamic cameras in 3D environments. In *CVPR*, pages 8856–8866, 2023. 3

[54] Zachary Teed and Jia Deng. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. *Advances in neural information processing systems*, 2021. 2

[55] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. Recovering 3D human mesh from monocular images: A survey. *TPAMI*, 2023. 2

[56] Hsiao-Yu Fish Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. *NeurIPS*, pages 5236–5246, 2017. 2, 3

[57] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, pages 109–117, 2017. 2

[58] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *ECCV*, pages 20–36, 2018. 2, 3

[59] 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)*, 2018. 7

[60] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. DenseRaC: Joint 3D pose and shape estimation by dense render-and-compare. In *ICCV*, pages 7760–7770, 2019. 2, 3

[61] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 6

[62] Hongwei Yi, Chun-Hao P. Huang, Dimitrios Tzionas, Muhammed Kocabas, Mohamed Hassan, Siyu Tang, Justus Thies, and Michael J. Black. Human-aware object placement for visual environment reconstruction. In *CVPR*, pages 3959–3970, 2022. 1

[63] 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*, pages 2990–3000, 2020. 2

[64] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. SimPoE: Simulated character control for 3D human pose estimation. In *CVPR*, pages 7159–7169, 2021. 3, 8

[65] 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. 2, 3, 4, 6, 8

[66] Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Neural descent for visual 3D human pose and shape. In *CVPR*, pages 14484–14493, 2021. 3, 5

[67] Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. SmoothNet: a plug-and-play network for refining human poses in videos. In *ECCV*, pages 625–642. Springer, 2022. 6

[68] 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 *ICCV*, pages 11446–11456, 2021. 1, 3, 4

[69] Hongwen Zhang, Jie Cao, Guo Lu, Wanli Ouyang, and Zhenan Sun. Learning 3D human shape and pose from dense body parts. *TPAMI*, 2022. 2, 3

[70] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. PyMAF-X: Towards well-aligned full-body model regression from monocular images. *TPAMI*, 2023. 1, 3, 8

[71] 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. 3, 5

[72] 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 *European conference on computer vision (ECCV)*, 2022. 6

[73] Yuxiang Zhang, Liang An, Tao Yu, Xiu Li, Kun Li, and Yebin Liu. 4d association graph for realtime multi-person motion capture using multiple video cameras. In *CVPR*, pages 1324–1333, 2020. 2

[74] Yuxiang Zhang, Zhe Li, Liang An, Mengcheng Li, Tao Yu, and Yebin Liu. Lightweight multi-person total motion capture using sparse multi-view cameras. In *ICCV*, pages 5560–5569, 2021. 2

[75] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3D human pose estimation with spatial and temporal transformers. In *ICCV*, pages 11656–11665, 2021. 4

[76] Yuxiao Zhou, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, and Feng Xu. Monocular real-time full body capture with inter-part correlations. In *CVPR*, pages 4811–4822, 2021. 3