

Enhanced Spatio-Temporal Context for Temporally Consistent Robust 3D Human Motion Recovery from Monocular Videos

Sushovan Chanda
TCS Research
India

chanda.sushovan@tcs.com

Amogh Tiwari
CVIT, IIIT Hyderabad
India

amogh.tiwari@research.iiit.ac.in

Lokender Tiwari
TCS Research
India

lokender.tiwari@tcs.com

Brojeshwar Bhowmick
TCS Research
India

b.bhowmick@tcs.com

Avinash Sharma
CVIT, IIIT Hyderabad
India

asharma@iiit.ac.in

Hrishav Barua
TCS Research
India

hrishav.smit5@gmail.com

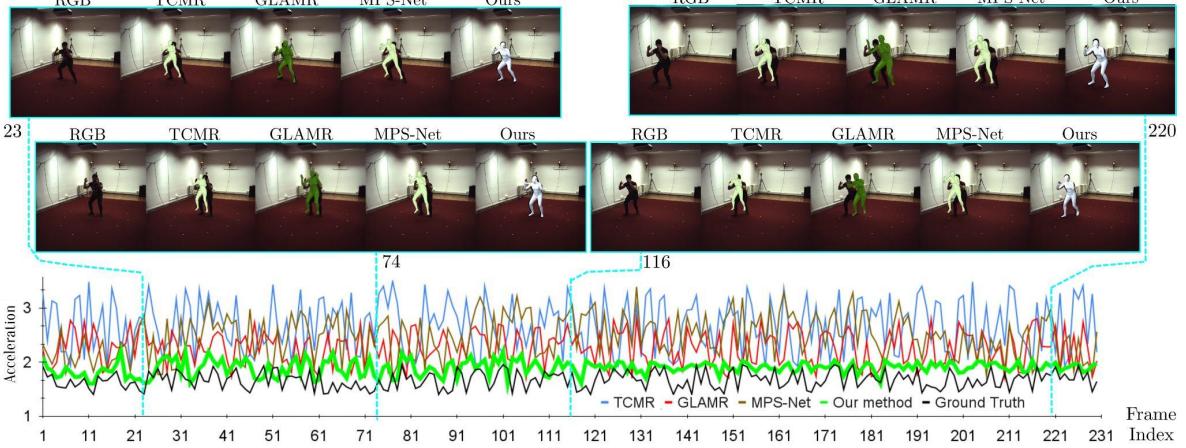


Figure 1. Our method yields superior and temporally consistent motion estimation. It can be observed that our method yield graph (green curve) significantly closer to the ground truth acceleration graph (black curve) compared to existing methods [2, 38, 39] (result inferred on unseen test video from Human3.6M dataset [11]).

Abstract

Recovering temporally consistent 3D human body pose, shape and motion from a monocular video is a challenging task due to (self-)occlusions, poor lighting conditions, complex articulated body poses, depth ambiguity, and limited availability of annotated data. Further, doing a simple per-frame estimation is insufficient as it leads to jittery and implausible results. In this paper, we propose a novel method for temporally consistent motion estimation from a monocular video. Instead of using generic ResNet-like features, our method uses a body-aware feature representation and an independent per-frame pose and camera initialization over a temporal window followed by a novel spatio-temporal fea-

ture aggregation by using a combination of self-similarity and self-attention over the body-aware features and the per-frame initialization. Together, they yield enhanced spatio-temporal context for every frame by considering remaining past and future frames. These features are used to predict the pose and shape parameters of the human body model, which are further refined using an LSTM. Experimental results on the publicly available benchmark data show that our method attains significantly lower acceleration error and outperforms the existing state-of-the-art methods over all key quantitative evaluation metrics, including complex scenarios like partial occlusion, complex poses and even relatively low illumination.

1. Introduction

Recovering 3D human body pose, shape and motion from a monocular video is an important task that has tremendous applications in augmented/virtual reality, healthcare, gaming, sports analysis, human-robot interaction in virtual environments, virtual try-on, etc. A lot of work has been done in estimating 3D body pose and shape from a single-image [3, 13, 18, 30, 32] by learning to regress the explicit 3D skeleton or parametric 3D body model (e.g., SMPL [23]). However, many applications such as human motion analysis, sports analytics, behavior analysis, etc., critically depend on the temporal consistency of human motion where single-image-based methods seem to fail frequently. Temporally consistent 3D human pose, shape and motion estimation from a monocular video is a challenging task due to (self-) occlusions, poor lighting conditions, complex articulated body poses, depth ambiguity, and limited availability of annotated data. Efforts on monocular video-based motion estimation [2, 16, 19, 24, 33, 35] typically introduce a CNN or RNN module to perform spatio-temporal feature aggregation from neighboring frames followed by SMPL [23] parameters regression, thus modeling relatively local temporal coherence. However, these methods tend to fail while capturing long-term temporal dynamics and show poor performance when the body is under partial occlusion. Some of the recent works [8, 26, 34, 39] also attempt to model the generative space of motion modeling using Conditional VAEs, often followed by a global, non-learning-based optimization at inference time using the entire video. Such global optimization is also used in a very recent work in [40] with a plug-and-play post-processing step for improving the existing methods by exploiting long-term temporal dependencies for human motion estimation. However, due to the post-processing over the entire sequence, such methods find limited applicability to real-world scenarios.

A highly relevant recent work, MPS-Net [38], attempts to attain a good balance between local to global temporal coherence using their MOtion Continuity Attention (MOCA) module. More specifically, their method explicitly models the visual feature similarity across RGB frames and uses it to guide the learning of the self-attention module for spatio-temporal feature learning. MOCA enables focusing on an adaptive neighborhood range for identifying the motion continuity dependencies. This is followed by a Hierarchical Attentive Feature Integration (HAFI) module to achieve local to global temporal feature aggregation through which they achieve SOTA performance. Nevertheless, similar to the majority of the existing methods, they use ResNet [9]-like generic deep features extracted from the RGB frames. However, such generic feature representations do not exploit the prior knowledge of human appearance(that the 3D human body has a fixed topology and can be represented by a parametric model). Additionally, [38]

do not exploit per-frame pose and shape initialization and uses a computationally heavy Hierarchical Attentive Feature Integration (HAFI) module. Finally, they only perform per-frame prediction using the aggregated spatio-temporal features, thereby completely neglecting the joint estimation performed by existing methods.

In this paper, we propose a holistic method that exploits enhanced spatio-temporal context and recovers temporally consistent 3D human pose/shape from monocular video. At first, we select a set of continuous frames in a temporal window and pass it to the *Initialization module* which extracts the body-aware deep features from individual frames and in-parallel predict initial per-frame estimates of body pose/shape and camera pose using an off-the-shelf method. Subsequently, we pass these initial estimates and features to novel *Spatio-Temporal feature Aggregation (STA) module* for recovering enhanced spatio-temporal features. Finally, we employ our novel *Motion estimation and Refinement module* to obtain temporally consistent pose/shape estimation using these enhanced features. [Figure 2](#) provides outline of our method.

In regard to functionality/relevance of these modules, the initialization module extracts a body-aware feature representation [29] for each frame of the local non-overlapping temporal frame window, instead of the generic ResNet feature used by existing methods and the independent per-frame pose and camera initialization estimated using [4]. This provides a strong spatial prior to our method. Further, our proposed novel STA module computes the self-similarity and the self-attention on initial spatial priors provided by the previous module. In particular, the self-similarity between the body-aware features in a temporal window helps us to correlate the body parts across frames even in the presence of occlusion. Similarly, the self-similarity among the pose parameters and the cameras reveals the continuity of the human motion along with the camera consistency. We also use self-attention on the camera parameters and the body-aware features. Together, they yield spatio-temporal aggregated features for every frame by considering the remaining past and future frames inside the window. Here, the joint characteristics of the self-similarity and the attention map find the more appropriate range in the input video to reveal the long-horizon context. Finally, our novel motion estimation and refinement module first predicts the per-frame coarse estimation of pose/shape using the spatio-temporally aggregated features from the STA module and subsequently passes it to an LSTM-based joint-temporal refinement network to recover the temporally consistent robust prediction of pose/shape estimates. In order to generate continuous predictions near the temporal window boundaries, we average the pose/shape parameters for consecutive border frames across neighboring windows. We empirically observed that applying LSTM-based joint

refinement on pose/shape yields superior performance instead of applying it on STA features and then predicting pose/shape parameters (see subsection 4.4).

As a cumulative effect, our method produces significantly lower acceleration errors in comparison to SOTA methods (see subsection 4.1). Figure 1 shows a plot of acceleration where our method yields the acceleration curve (in green) closest to the ground truth acceleration curve (in black). Moreover, owing to our enhanced spatio-temporal context and motion refinement, our method significantly outperforms the state-of-the-art (SOTA) methods even in relatively poor illumination and severe occlusion (please refer to subsection 4.2).

2. Related Works

Image based 3D human pose, shape and motion estimation: Existing methods either solve for the parameters of SMPL [23] from the images or directly regress the coordinates of a 3D human mesh [6]. [13, 18, 30] are some of the early successful works for human pose and shape estimation from monocular images.

HyBrik [21] and KAMA [12] leverage 3D key points for the 3D mesh reconstruction. In particular, HyBrik uses twist and swing decomposition for transforming the 3D joints to relative body-part rotations. Instead of full body, methods like HoloPose [5] and PARE [17] have introduced parts-based model. While HoloPose does part-based parameter regression, PARE uses a part-guided attention mechanism for exploiting the visibility of individual body parts and predicting the occluded parts using neighboring body-part information. While these methods are quite effective for estimating the 3D pose and shape from images, they are not capable of producing temporally consistent 3D human motion from video by frame-based processing.

Video based 3D human and pose estimation: Recently, a considerable amount of work has been carried out to address the challenge of temporally consistent 3D human pose and shape estimation from video. For instance, HMMR [14] trains a temporal encoder that learns a representation of 3D human dynamics from a temporal context of image features. Along with 3D human pose and shape, such representation is also used for capturing the changes in the pose in the nearby past and future frames. Similarly, VIBE [16] proposes a temporal encoder that encodes static features into a series of temporally correlated latent features and feeds them to a regressor to estimate the SMPL parameters. MEVA [24] uses a two-stage model that first captures the coarse overall 3D human motion followed by a residual estimation that adds back person-specific motion details. However, these methods fail to reconstruct the humans under partial occlusions. In a recent work by Choi *et al.* [2], GRU-based temporal encoders are used with different en-

coding strategies to learn better temporal features from images. Also, they propose a feature integration from the three encoders for the SMPL parameter regressor. GRU-based techniques can only deal with local neighborhoods which makes it difficult for them to learn long-range dependencies. Hence, [1, 42] uses a transformer to learn long-range temporal dependencies. However, such methods require a large number of consecutive frames (around 250), making them slower. Another class of methods like HuMoR [34] and GLAMR [39] use variational autoencoder which takes single frame-based human pose estimates to predict the human motion sequence in an auto-regressive way followed by a non-learning based global optimization on the human pose and trajectory obtained from the entire video for temporal refinement. Similarly, SmoothNet [40] also does a global optimization on the estimated trajectory of any human pose estimation method to improve their temporal continuity. The global optimization in the test time limits the applicability of such methods. In a recent attempt, MPS-Net [38] tries to produce locally global temporal coherence using a motion continuity attention module (see section 1 for more details on MPS-Net [38]).

3. Method

In this section, we provide a detailed overview of the key modules of our proposed method. As discussed in section 1 (and outlined in Figure 2), our method takes a set of consecutive frames as input and feeds it to the three key modules, namely, initialization, spatio-temporal feature aggregation and motion prediction & refinement, to predict temporally consistent body pose and shape parameters of SMPL [23], a statistical body model.

More specifically, given an input video $V = \{F_i\}_{i=1}^N$ composed of N frames, with F_i representing the i^{th} frame, we aim to recover SMPL-based human body pose and shape parameters for each frame, i.e., $\Theta_i^{pred} = \{T_i, R_i, \theta_i, \beta_i\}$. Here, $T_i \in \mathbb{R}^3$ and $R_i \in \mathbb{R}^3$ represents the translation and rotation (in axis-angle format) of the root joint, $\theta_i \in \mathbb{R}^{23 \times 3}$ represents the relative rotations of the remaining 23 joints while $\beta_i \in \mathbb{R}^{10}$ represents body shape parameters. Please note that we sample a temporal window (a subset of continuous frames) of size W (we choose $W = 16$) from the input video and learn/infer over it instead of doing inference on all frames in the video sequence.

3.1. Initialization

Per-frame Body Pose and Camera Estimation: We perform independent estimation of per-frame body pose (θ^{init}) and camera (ω^{init}) parameters using a SOTA method (HMR2.0 [4]) and feed it as initialization to our STA module.

Body-aware Spatial Feature Extraction: Recently,

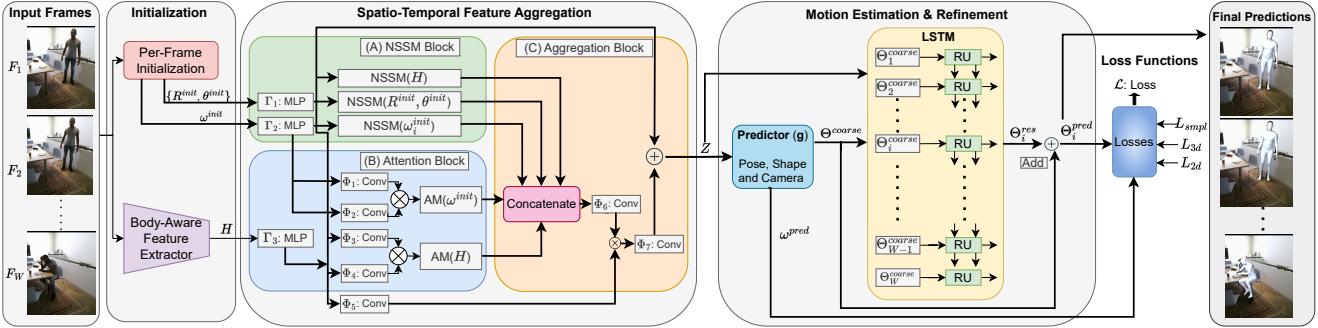


Figure 2. Overview of our proposed method.

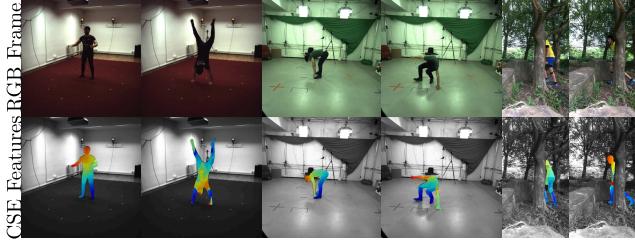


Figure 3. Sample three-channel visualization of CSE Embeddings.

Continuous Surface Embedding (CSE) [29] was proposed to learn body-aware feature representation for obtaining dense correspondences across images of humans. CSE predicts, per pixel 16-dimensional embedding vector (associated with the corresponding vertex in the parametric human mesh), thereby establishing dense correspondences between image pixels and 3D mesh surface, even in the presence of severe illumination conditions and (self-) occlusions. Figure 3 shows the color-coded visualization of CSE embeddings demonstrating its robustness to severe illumination and/or occlusion scenarios. Thus, we propose to extract and use the 16-dimensional body-aware spatial features $H = \{H_i\}_{i=1}^N$ using a pre-trained CSE encoder for each frame F_i , such that:

$$H_i = \Psi(F_i) \quad (1)$$

where, $H_i \in \mathbb{R}^{112 \times 112 \times 16}$.

3.2. Spatio-Temporal Feature Aggregation (STA)

The spatial features H_i extracted from each frame can be directly regressed to estimate per-frame motion and shape parameters. However, this typically leads to jittery and implausible motion estimates as the predictions are not temporally consistent. One possible remedy to this is to use self-attention across frames in a temporal window [37]. Interestingly, [38] showed that a regular attention network is unreliable and can give high attention scores between temporally distant frames which would lead to inaccurate results.

They address this problem by using a **Normalized Self-Similarity Matrix (NSSM)** in their MOCA module. Nevertheless, their method only exploited the spatial features for such self-attention guidance. Instead, as per the recent trend of exploiting per-frame pose initialization [20, 39], we propose to encode additional information to our temporal features in terms of initial estimates of body pose and camera parameters. More specifically, we obtain for each i^{th} frame the initial pose/shape and camera parameters using [4] as: $\Theta_i^{init} = \{T_i, R_i, \theta_i, \beta_i\}$ and camera parameters $\omega_i^{init} \in \mathbb{R}^3$ (assume a weak perspective camera model). It is important to note that we represent rotation using the 6-dimensional vector representation [41] and then flatten them into a single 144-dimensional vector to recover body pose as: $[R_i, \theta_i] \in \mathbb{R}^{144}$

Our STA module has three key blocks: (1) Frame-wise Similarity Computation, (2) Frame-wise self-attention, and (3) Feature Aggregation.

The first block deals with the computation of the three $\{W \times W\}$ self-similarity matrices, namely, NSSM (H) for the Body-aware spatial features, NSSM ($[R, \theta]$) for initial body pose estimates and NSSM (ω^{init}) for initial camera estimates. More specifically, we uplift $[R_i, \theta_i]$ and $\omega_i^{init} \in R^3$ to 512 dimensions using linear layers Γ_1 and Γ_2 and similarly transform the spatial feature H_i to 2048 dimensions using Γ_3 . These multiple NSSMs help us to correlate the frames based upon body parts appearance, body pose, and cameras thereby giving robustness to occlusions as well as revealing the continuity of the human motion along with the camera consistency.

The second block obtains a self-attention map on our spatial features i.e., $AM(H)$ and initial camera estimates i.e., $AM(\omega^{init})$, respectively. When applying self-attention on our spatial features $H_i \in \mathbb{R}^{112 \times 112 \times 16}$, first we transform them to 2048 dimension using a linear layer Γ_3 , and later downsample them to $\mathbb{R}^{N \times 1024}$ by learning two different 1×1 convolutional layers Φ_3 and Φ_4 . Similarly, when applying self-attention on the initial camera estimates, we first uplift this vector to 512 dimension vector using an MLP Γ_2 and

subsequently learn two different 1×1 convolution layers Φ_1 and Φ_2 . This self-attention on the camera parameters and the body-aware features help us adaptively find the range which is important to capture the temporal smoothness.

Finally, the feature aggregation block first concatenates all the attention and NSSM maps to get a $W \times W \times 5$ tensor and later resize it to $W \times W$ matrix using a 1×1 convolution layer (Φ_6). This $W \times W$ represents the consolidated similarity between frames across the window. This feature is subsequently multiplied with the down-sampled spatial features (of 1024 dimension obtained by Φ_5) and the result is then uplifted (using convolution layer Φ_7) to get $Y \in \mathbb{R}^{W \times 2048}$. Thus, together, they yield spatio-temporal aggregated features for every frame by considering the remaining past and future frames inside the window. The per-frame temporally aggregated feature Y_i is finally added to the spatial features to get the spatio-temporally aggregated features Z_i for i^{th} frame as:

$$Z_i = H_i + Y_i. \quad (2)$$

3.3. Motion Estimation & Refinement

Once we have the spatio-temporal features Z_i , we obtain an independent coarse pose/shape, and camera estimation for each frame using predictor network (g)

$$\Theta_i^{coarse}, \omega_i^{pred} = g(Z_i) \quad (3)$$

where g predicts the SMPL parameters i.e., $\Theta_i^{coarse} \in \mathbb{R}^{85}$ and the camera parameters i.e., $\omega_i^{pred} \in \mathbb{R}^3$ for frame F_i .

We propose to further refine these estimated independent coarse poses and shapes (obtained using spatiotemporally aggregated features) using an LSTM [10] based joint *residual* prediction. The LSTM ζ takes as input the features Z_i and coarse SMPL pose estimates Θ_i^{coarse} and predicts the residual $\Theta_i^{res} \in \mathbb{R}^{85}$, which is subsequently added to Θ_i^{coarse} in order to recover the refined pose and shape parameters Θ_i^{pred} .

$$\Theta_i^{res} = \zeta(Z_i, \Theta_i^{coarse}) \quad (4)$$

$$\Theta_i^{pred} = \Theta_i^{coarse} + \Theta_i^{res} \quad (5)$$

To ensure temporally consistent predictions at the window boundaries, we average the pose and shape parameter estimates of bordering frames across the neighboring windows.

3.4. Loss Functions

Similar to existing literature [13, 16, 38], we adopt loss functions on body pose and shape (L_{SMPL}), 3D joint coordinates (L_{3D}), and 2D joint coordinates (L_{2D}) obtained with predicted weak-perspective camera parameters (ω^{pose}). These loss functions are briefly explained below.

$$\begin{aligned} L_{SMPL} &= \lambda_{shape} \|\hat{\beta}_i - \beta_i\|_2 \\ &+ \lambda_{pose} \|\{\hat{R}_i, \hat{\theta}_i\} - \{R_i, \theta_i\}\|_2 \end{aligned} \quad (6)$$

where β_i and $\{R_i, \theta_i\}$ respectively are the predicted pose and shape parameters for the i^{th} frame, and $\hat{\beta}_i, \{\hat{R}_i, \hat{\theta}_i\}$ are the corresponding ground-truths.

$$L_{3D} = \|\hat{J}_i^c - J_i^c\|_2 \quad (7)$$

where J_i^c represents predicted the 3D joint coordinates for the i^{th} frame and \hat{J}_i^c are the corresponding ground-truth 3D joint coordinates.

$$L_{2D} = \|\hat{x}_i - \Pi(J_i^c)\|_2 \quad (8)$$

where \hat{x}_i represents the ground-truth 2D keypoints for the i^{th} frame and Π represents the 3D-2D projection obtained from the predicted camera parameters ω^{pred} .

The final loss function is a linear combination of these losses defined as:

$$L_{final} = \lambda_1 L_{SMPL} + \lambda_2 L_{3D} + \lambda_3 L_{2D} \quad (9)$$

It is important to note that our model is trained in an end-to-end trainable fashion where L_{final} is applied on the final predicted pose and shape parameters obtained from LSTM ζ . There is no separate training performed for the coarse estimation predictor g .

4. Experiments and Results

Datasets Details: We adopt the same train/test splits of Human3.6M [11], 3DPW [36] and MPI-INF-3DHP [27] datasets used by existing work [2, 16, 38]. *Human3.6M* is a large scale dataset containing video sequences with corresponding 3D pose annotations of various subjects performing different actions like discussion, smoking, talking on the phone etc. Similar to existing work [2, 16, 38], we use the sub-sampled dataset (25 FPS) for our experiments. *MPI-INF-3DHP* contains 8 subjects with 16 videos per subject. It is captured in a combination of indoor and outdoor settings with actions ranging from walking and sitting, to complex dynamic actions like exercising. It is captured by a markerless motion capture system using a multi-view camera setup. *3DPW* is an in-the-wild dataset, captured with a moving cell-phone camera. It uses inertial measurement unit (IMU) sensors patched to the human body parts to calculate the ground-truth SMPL [23] parameters. It contains 60 video sequences with 18 3D models in different clothing, performing daily-life activities like walking, buying vegetables etc. Further, in order to evaluate the generalization ability of our method to unseen data, we use three additional datasets: Fitness-AQA [31], PROX [7] and i3DB [28]. These datasets contain sequences having actions/motion fairly different from our training datasets. Fitness-AQA contains videos of subjects lifting weights in a gym, which leads to self-occlusion and complex body poses, while PROX and i3DB contain video sequences of

Method	Human3.6M [11]			3DPW [36]			MPI-INF-3DHP [27]			
	PA-MPJPE↓	MPJPE↓	ACC-ERR↓	PA-MPJPE↓	MPJPE↓	MPVPE↓	ACC-ERR↓	PA-MPJPE↓	MPJPE↓	ACC-ERR↓
VIBE [16]	41.4	65.6	-	51.9	82.9	99.1	23.4	64.6	96.6	-
MEVA [24]	53.2	76.0	15.3	54.7	86.9	-	11.6	65.4	96.4	11.1
Uncertainty-Aware [19]	38.4	58.4	6.1	52.2	92.8	106.1	6.8	59.4	93.5	9.4
TCMR [2]	52.0	73.6	3.9	52.7	86.5	102.9	7.1	63.5	97.3	8.5
MPS-Net [38]	47.4	69.4	<u>3.6</u>	52.1	84.3	99.7	7.4	62.8	96.7	9.6
HUMOR [34]	47.3	69.3	4.2	51.9	74.8	<u>81.4</u>	<u>6.3</u>	63.2	98.1	<u>8.4</u>
GLAMR [39]*	48.3	72.8	6.0	51.7	<u>72.9</u>	86.6	8.9	60.17	96.2	8.9
D&D [20]†	<u>35.5</u>	<u>52.5</u>	6.1	<u>42.7</u>	73.7	88.6	7.0	-	-	-
Our Method	31.0	41.3	3.3	39.2	63.5	61.8	5.3	53.2	88.7	8.1

Table 1. Quantitative Comparison of our method with other monocular video-based methods. Best is in **bold** and second best is underlined. (*: GLAMR uses Human3.6M, 3DPW and AMASS [25] as 3D datasets | †: D&D performs individual training on Human3.6M and 3DPW)

humans interacting with objects in an indoor setting like a room/office.

Evaluation Metrics: We use the standard evaluation metrics used in existing literature [2, 16, 24, 38] to evaluate our method’s performance. Specifically, we use the mean per joint position error (MPJPE), Procrustes-aligned mean per joint position error (PA-MPJPE), mean per vertex position error (MPVPE) and acceleration error (ACC-ERR). MPJPE is defined as the mean of the Euclidean distances between the ground truth and the predicted joint positions. PA-MPJPE is defined as the MPJPE computed after using Procrustes alignment (PA) to solve for translation, scale and rotation between the estimated body and the ground truth. MPVPE is given by the mean of the Euclidean distances between the ground truth and the predicted vertex positions of each vertex in the SMPL [23] body model constructed using the predicted pose/shape parameters. Finally, (ACC-ERR) is defined as the mean difference between the accelerations of the ground truth and predicted 3D joints. Specifically, the change in position of the 3D joints in unit time (i.e. across two consecutive frames) gives us the velocity of the joints, and the change in velocity in unit time gives us the acceleration. Acceleration error is then measured by finding the difference between the groundtruth and predicted accelerations. MPJPE, PA-MPJPE and MPVPE are measured in millimeters (*mm*) and express the fidelity of the estimated 3D pose and shape. While ACC-ERR, measured in mm/t^2 (where t denotes unit time - the time interval between two consecutive frames) expresses the temporal consistency of the estimation.

Method	Fitness-AQA [31]		PROX [7]		i3dB [28]	
	MPJPE ↓	ACC-ERR↓	MPJPE↓	ACC-ERR↓	MPJPE↓	ACC-ERR↓
TCMR [2]	89.3	7.6	29.7	2.3	47.1	3.1
MPS-Net [38]	64.7	6.1	22.1	1.9	35.5	2.7
Our Method	43.5	5.3	18.3	1.6	21.6	2.1

Table 2. Generalization results on unseen datasets.

Implementation Details: We obtain the body aware features and per-frame pose/camera initializations using the pre-trained CSE [29] and HMR2.0 [4] models, respectively.

Similar to existing work [2, 16, 38], we initialize our pose, shape, and camera predictor in the motion estimation and refinement module with the pre-trained SPIN [18] checkpoint. In the same module, the LSTM has 3 layers and uses 2048 as the hidden feature size. Training is performed for 35 epochs with a mini-batch size of 32 and an initial learning rate of 5×10^{-5} . The learning rate is reduced by a factor of 10 every time the 3D pose accuracy does not improve for the 5 consecutive epochs. Adam Solver [15] is used for optimization. For our experiments, we use a window size of 16 (see Table 5 for discussion on choice of window size). We set the coefficients for L_{SMPL} , L_{3D} and L_{2D} to 300.0, 0.06 and 60.0 respectively. Training is done for 35 epochs and takes about 7 hours using 3 NVIDIA RTX A-6000 GPUs.

4.1. Quantitative Results

Table 1 provides a quantitative comparison between our method and existing SOTA monocular video-based methods. All methods use the same 3D datasets as ours with standard training/test split, except GLAMR [39] (which uses Human3.6M, 3DPW and AMASS datasets) and D&D [20] (which trains individually on Human3.6M and 3DPW). Some of these methods also utilize additional 2D datasets for training and we used their pre-trained model for comparison. However, we only rely on the 3D datasets for training. It can be observed from Table 1 that our method significantly outperforms existing methods over all metrics across datasets, demonstrating the superiority of our method.

4.2. Qualitative Results

Figure 4 visualize qualitative comparison with monocular video-based SOTA methods. More specifically, row-1 shows a complex pose under poor illumination. It can be observed that while all other methods fail to recover the body pose, our method successfully recovers the body pose reliably. In rows 2-4, it can be observed that while other methods are also able to recover the pose, our method provides more accurate SMPL fitting. In the last two rows, we demonstrate results on even more challenging cases involv-

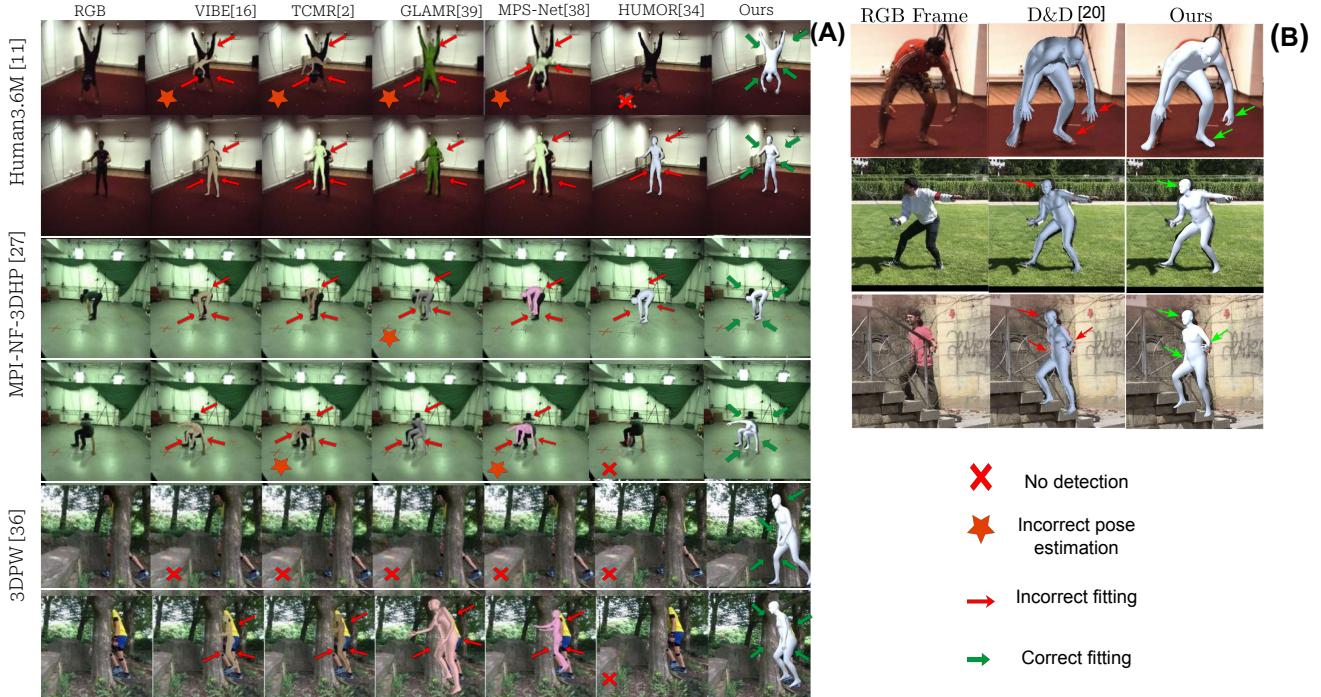


Figure 4. (A) We show the estimated pose overlaid on the frames of the sample videos from Human3.6M [11], 3DPW [36], and MPI-INF-3DHP datasets [27]. (B) Similar results are shown in comparison to D&D [20].



Figure 5. Qualitative comparison across frames on a sequence of 3DPW dataset. The green arrows show the improved regions compared to the red ones. Our method is able to achieve more accurate and more temporally consistent SMPL fitting.

ing significant occlusion, where most existing methods fail even to detect the human. However, our method not only

detects the human but also provides a reasonable SMPL fitting. Additionally, in Figure 5, we demonstrate the results



Figure 6. Qualitative comparison on unseen datasets.

Configuration	Human3.6M [11]			3DPW [36]			MPI-INF-3DHP [27]		
	PA-MPJPE ↓	MPJPE ↓	ACC-ERR ↓	PA-MPJPE ↓	MPJPE ↓	ACC-ERR ↓	PA-MPJPE ↓	MPJPE ↓	ACC-ERR ↓
1. Ours w/o Body-Aware Features (i.e., w/o H)	34.8	45.0	3.5	41.6	67.3	5.3	56.4	91.8	8.3
2. Ours w/o Per-Frame initialization (i.e., w/o $\{R^{init}, \theta^{init}\}$ and w/o ω^{init})	43.8	71.9	4.2	49.8	78.5	5.5	61.4	93.4	8.4
3. Ours w/o pose initialization (i.e., w/o $\{R^{init}, \theta^{init}\}$)	41.5	53.2	3.6	48.1	74.7	5.5	58.6	93.2	8.4
4. Ours w/o camera initialization (i.e., w/o ω^{init})	37.3	49.1	3.5	47.3	73.8	5.4	57.2	91.8	8.3
5. Ours w/o LSTM based refinement on coarse estimates (i.e., w/o ζ)	32.7	42.8	3.5	40.3	69.3	5.6	54.6	90.3	8.4
6. Our Final Method	31.0	41.3	3.3	39.2	63.5	5.3	53.2	88.7	8.1
7. Ours + AM on pose (i.e., AM on $\{\hat{R}^{init}, \hat{\theta}^{init}\}$)	33.8	44.8	3.4	42.7	68.0	5.3	56.8	90.7	8.3
8. Ours w. LSTM on Feature Space (followed by refined estimation of shape/pose)	39.2	47.2	4.0	44.3	72.9	5.8	58.7	92.6	8.3

Table 3. Ablation study on our method’s performance in different configurations. (Best is in **bold**.)

of MPS-Net [38], GLAMR [39] and our method across multiple frames of a video sequence where the person is trying to lift the bag. It can be observed that the head orientation and the proximity of the hand to the bag are more temporally consistent in our method compared to other methods. Further, the SMPL fitting is also better for our method.

4.3. Generalization to Unseen Datasets

We also test the generalization ability of our method by evaluating its performance on completely unseen Fitness-AQA [31], PROX [7] and i3DB [28] datasets. These datasets contain diverse scenarios and were not seen during training done on Human3.6M, 3DPW, and MPI-INF-3DHP. As reported in Table 2, our method significantly outperforms existing SOTA on unseen datasets, demonstrating our method’s generalization ability. We also show a qualitative comparison for the same in Figure 6. It can be seen that our method estimates pose and shape more accurately than other SOTA methods.

4.4. Ablation Study

We perform a detailed ablative study to analyze the contributions of different components of our method. Table 3 provides the quantitative ablative results, which list our final method’s performance in row-6. We sequentially re-

moved each component of our method and reported the performance drop in rows 1-5. More specifically, row-1 reports the results where we replace our body-aware feature encoder with generic ResNet. This leads to a drop in performance, demonstrating the contribution of the body aware features to our overall performance. In row-2, we train our network without using the per-frame pose and camera initialization. This too leads to a drop in the model performance. In row-3 & row-4, we report the performance of the model by individually removing the pose initialization and camera initialization. The results demonstrate that both pose initialization and camera initialization contribute individually to our method’s performance. In row-5, we report the performance by removing the LSTM-based motion refinement component, and once again find a drop in performance, especially in the ACC-ERR metric, demonstrating the contribution of the motion refinement module. We also report two additional ablative results in the last two rows of Table 3 as modifications to our proposed method. Specifically, row-7 reports the performance of the modified method by adding the self-attention on the body pose initialization to our method. However, unlike self-attention on body-aware features and camera pose, we empirically find that self-attention on body pose leads to a degradation in performance. One possible explanation for this degra-

Method	Human3.6M [11]			3DPW [36]			MPI-INF-3DHP [27]		
	PA-MPJPE↓	MPJPE↓	ACC-ERR↓	PA-MPJPE↓	MPJPE↓	ACC-ERR↓	PA-MPJPE↓	MPJPE↓	ACC-ERR↓
PARE [17]	53.8	72.8	6.9	46.5	74.5	7.1	63.1	95.7	9.3
Ours w. PARE	42.6	66.4	4.1	49.3	67.3	5.4	58.7	93.3	8.1
CLIFF [22]	32.7	47.1	6.7	43.0	69.0	7.3	62.4	96.7	9.3
Ours w. CLIFF	31.2	42.7	5.7	39.3	65.1	6.7	59.3	94.2	8.3
HMR 2.0 [4]	33.8	45.3	3.8	44.4	69.8	5.6	57.1	91.9	8.4
Our Method (uses HMR2.0 initialization)	31.0	41.3	3.3	39.2	63.5	5.3	53.2	88.7	8.1

Table 4. Evaluation of our method with different per-frame initializers. (Best is in **bold**.)

Window size	Human3.6M			3DPW			MPI-INF-3DHP			
	PA-MPJPE ↓	MPJPE ↓	ACC-ERR ↓	PA-MPJPE ↓	MPJPE ↓	MPVPE ↓	ACC-ERR ↓	PA-MPJPE ↓	MPJPE ↓	ACC-ERR
8 frames	31.3	41.7	3.4	40.1	64.4	62.3	5.4	54.7	89.0	8.2
16 frames	31.0	41.3	3.3	39.2	63.5	61.8	5.3	53.2	88.7	8.1
32 frames	32.0	42.0	3.8	40.7	65.8	63.1	5.4	55.3	89.2	8.2

Table 5. Ablation study on performance of our method with different temporal window sizes. (Best is in **bold**).

dation is that self-attention to body poses can sometimes be misleading due to the frequently repeating body poses in a temporal window (e.g. walking involves very similar body poses). Nevertheless, we observed that using self-similarity (NSSM) on body pose helps as it exploits the spatio-temporal ordering (see row-3 & row-6). Finally, in row-8, we report the performance of an alternate setup for temporal refinement where we use the LSTM to aggregate temporal features before passing them to the pose/shape predictor, thereby eliminating the coarse prediction step. However, this leads to a drop in performance. As an explanation to this, we hypothesize that learning pose/shape corrections is more conducive to LSTM and hence our method provides a better estimate of body pose.

In addition, we evaluate the performance of our method with different per-frame initialization methods and report results in Table 4. It can be observed that our method consistently improves over the per-frame initialization methods (especially in terms of acceleration errors).

Furthermore we examine the performance of our method by varying the choice of temporal window sizes. These results are reported in Table 5. Similar to existing works [2, 16, 38], we find that a temporal window of size 16 provides optimal performance.

5. Discussion

Recovering from Bad Initialization: The spatio-temporal feature aggregation (STA) provides our method temporal context by considering the remaining past and future frames. This allows our method to recover accurate pose and shape even when CSE [29] and HMR2 [4] are not able to provide good initializations. We show few such results in Figure 7.

Limitations and Future Work: As shown in Figure 8, our method can fail in scenarios containing humans with extremely loose clothing as it is difficult to localize the under-

lying body in such scenarios. We plan to explore extension of our work to loose clothing in the future.

6. Conclusion

We proposed a novel method for recovering temporally consistent 3D human pose and shape from monocular video. Our method utilizes body-aware spatial features along with initial per-frame SMPL pose parameters to learn spatio-temporally aggregated features over a window. These features are then used to predict the coarse SMPL and camera parameters which are then further refined using a joint prediction of motion with LSTM. We demonstrate that our method consistently outperforms the SOTA methods both qualitatively and quantitatively. We also reported detailed ablative studies to establish relevance of key components of proposed method. As part of future work, it will be interesting to see extension of this work for humans with very loose garments (e.g., robes/abaya).

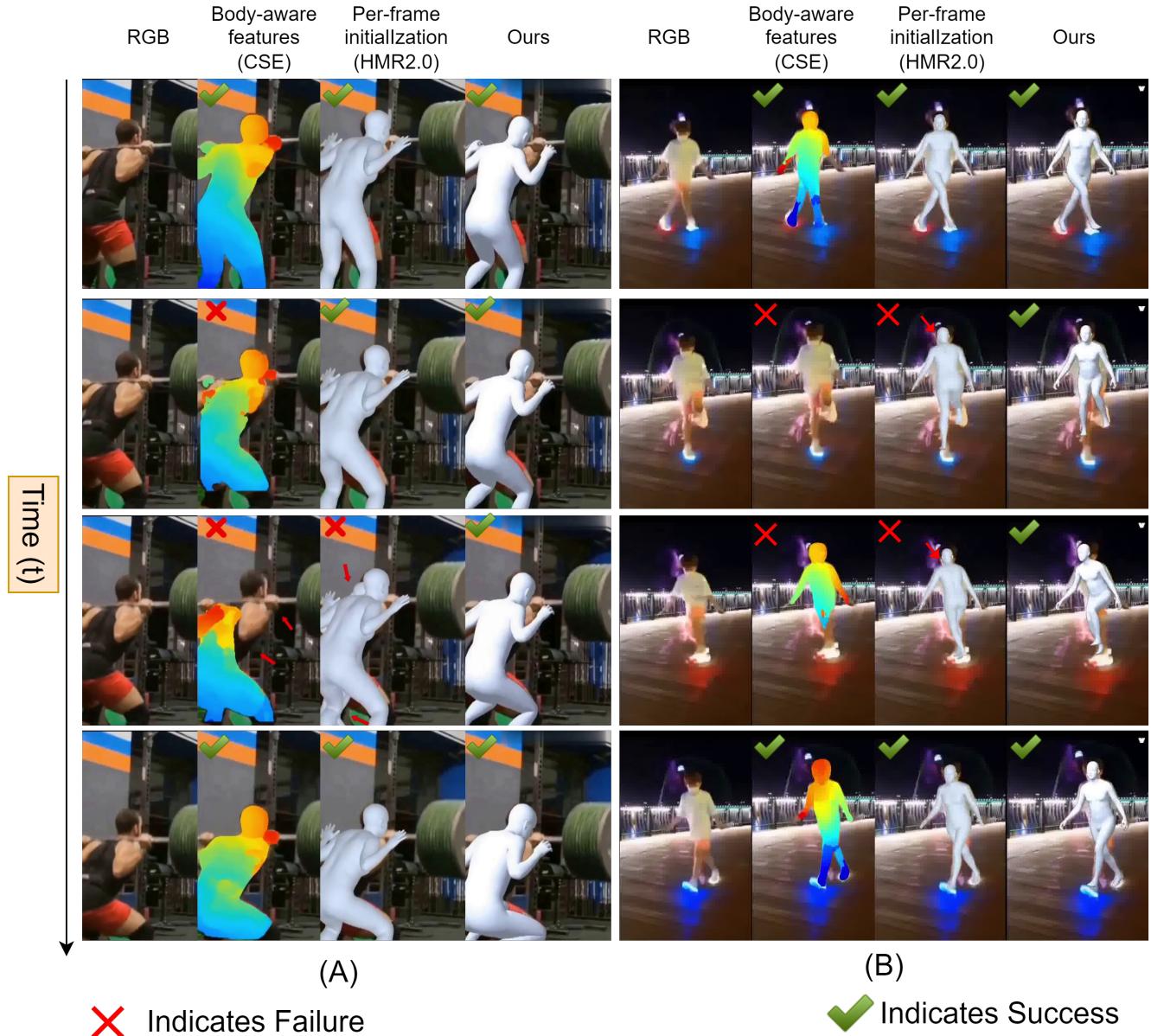


Figure 7. Temporal context provided by STA module allows our method to recover accurate pose and shape even when CSE/HMR2.0 are unable to provide a good initialization. Notice that in the 2nd and 3rd rows of (B), the HMR2.0 prediction is oriented wrongly.

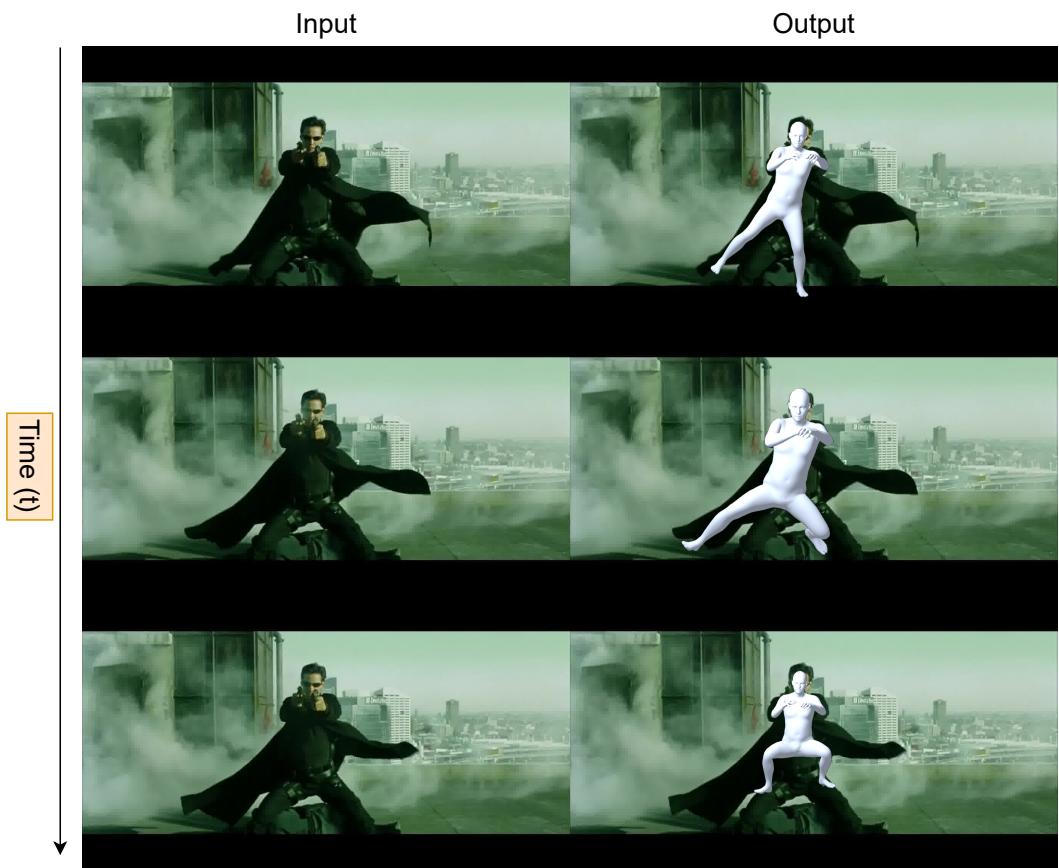


Figure 8. Our method can fail for humans in extremely loose clothing as it is difficult to localize the underlying body in such scenarios.

References

- [1] Fabien Baradel, Romain Brégier, Thibault Groueix, Philippe Weinzaepfel, Yannis Kalantidis, and Grégory Rogez. Posebert: A generic transformer module for temporal 3d human modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3
- [2] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1964–1973, 2021. 1, 2, 3, 5, 6, 9
- [3] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Košeká, and Ziyang Wu. Hierarchical kinematic human mesh recovery. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 768–784. Springer, 2020. 2
- [4] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa*, and Jitendra Malik*. Humans in 4D: Reconstructing and tracking humans with transformers. In *International Conference on Computer Vision (ICCV)*, 2023. 2, 3, 4, 6, 9
- [5] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10884–10894, 2019. 3
- [6] Moon Gyeongsik and Kyoung. Mu Lee. Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *Proceedings of the IEEE/CVF European conference on computer vision*, 2020. 3
- [7] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019. 5, 6, 8
- [8] Chengan He, Jun Saito, James Zachary, Holly Rushmeier, and Yi Zhou. Nemf: Neural motion fields for kinematic animation. In *Advances in Neural Information Processing Systems*, 2022. 2
- [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2
- [10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 5
- [11] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 1, 5, 6, 7, 8, 9
- [12] Umar Iqbal, Kevin Xie, Yunrong Guo, Jan Kautz, and Pavlo Molchanov. Kama: 3d keypoint aware body mesh articulation. In *2021 International Conference on 3D Vision (3DV)*, pages 689–699. IEEE, 2021. 3
- [13] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 2, 3, 5
- [14] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5614–5623, 2019. 3
- [15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [16] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020. 2, 3, 5, 6, 9
- [17] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11127–11137, 2021. 3, 9
- [18] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2252–2261, 2019. 2, 3, 6
- [19] Gun-Hee Lee and Seong-Whan Lee. Uncertainty-aware human mesh recovery from video by learning part-based 3d dynamics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12375–12384, 2021. 2, 6
- [20] Jiefeng Li, Siyuan Bian, Chao Xu, Gang Liu, Gang Yu, and Cewu Lu. D&d: Learning human dynamics from dynamic camera. In *European Conference on Computer Vision*, 2022. 4, 6, 7
- [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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3383–3393, 2021. 3
- [22] 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*, 2022. 9
- [23] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 2, 3, 5, 6
- [24] Zhengyi Luo, S Alireza Golestaneh, and Kris M Kitani. 3d human motion estimation via motion compression and refinement. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2, 3, 6
- [25] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 6
- [26] Mathieu Marsot, Stefanie Wuhrer, Jean-Sébastien Franco, and Stephane Durocher. A structured latent space for human

- body motion generation. *arXiv preprint arXiv:2106.04387*, 2021. 2
- [27] 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 *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017. 5, 6, 7, 8, 9
- [28] Aron Monszpart, Paul Guerrero, Duygu Ceylan, Ersin Yumer, and Niloy J Mitra. imapper: interaction-guided scene mapping from monocular videos. *ACM Transactions On Graphics (TOG)*, 38(4):1–15, 2019. 5, 6, 8
- [29] Natalia Neverova, David Novotny, Marc Szafraniec, Vasil Khalidov, Patrick Labatut, and Andrea Vedaldi. Continuous surface embeddings. *Advances in Neural Information Processing Systems*, 33:17258–17270, 2020. 2, 4, 6, 9
- [30] Mohamed Omran, Christoph Lassner, Gérard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*, pages 484–494. IEEE, 2018. 2, 3
- [31] Paritosh Parmar, Amol Gharat, and Helge Rhodin. Domain knowledge-informed self-supervised representations for workout form assessment. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, pages 105–123. Springer, 2022. 5, 6, 8
- [32] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 459–468, 2018. 2
- [33] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2
- [34] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11488–11499, 2021. 2, 3, 6
- [35] Shashank Tripathi, Siddhant Ranade, Ambrish Tyagi, and Amit Agrawal. Posenet3d: Learning temporally consistent 3d human pose via knowledge distillation. In *2020 International Conference on 3D Vision (3DV)*, pages 311–321. IEEE, 2020. 2
- [36] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gérard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018. 5, 6, 7, 8, 9
- [37] X. Wang, Ross B. Girshick, Abhinav Kumar Gupta, and Kaiming He. Non-local neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2017. 4
- [38] Wen-Li Wei, Jen-Chun Lin, Tyng-Luh Liu, and Hong-Yuan Mark Liao. Capturing humans in motion: temporal-attentive 3d human pose and shape estimation from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13211–13220, 2022. 1, 2, 3, 4, 5, 6, 8, 9
- [39] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11038–11049, 2022. 1, 2, 3, 4, 6, 8
- [40] 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 *European Conference on Computer Vision*. Springer, 2022. 2, 3
- [41] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 4
- [42] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3