Uncertainty-Aware Human Mesh Recovery from Video by Learning Part-Based 3D Dynamics

Gun-Hee Lee[1], Seong-Whan Lee[1,2]
[1]Department of Computer and Radio Communications Engineering, Korea University, Seoul, South Korea
[2]Department of Artificial Intelligence, Korea University, Seoul, South Korea
{gunhlee, sw.lee}@korea.ac.kr

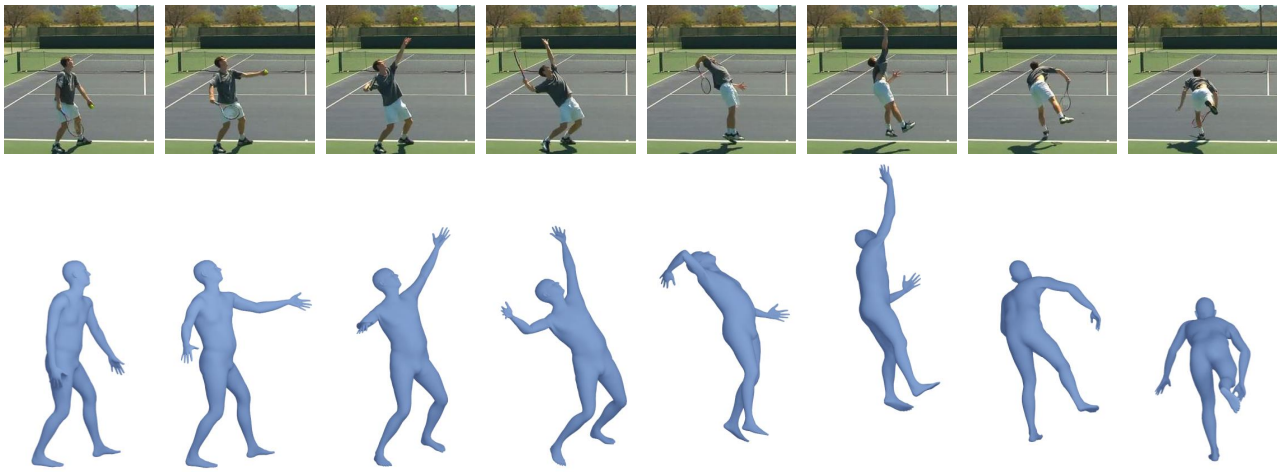

Figure 1: We propose a method that estimates 3D human pose and shape from video using uncertainty information and part-based 3D dynamics. Our method is able to recover the accurate and smooth 3D motion, achieving the state-of-the-art performance on standard benchmarks.

Abstract

Despite the recent success of 3D human reconstruction methods, recovering the accurate and smooth 3D human motion from video is still challenging. Designing a temporal model in the encoding stage is not sufficient enough to settle the trade-off problem between the per-frame accuracy and the motion smoothness. To address this problem, we approach some of the fundamental problems of 3D reconstruction tasks, simultaneously predicting 3D pose and 3D motion dynamics. First, we utilize the power of uncertainty to address the problem of multiple 3D configurations resulting in the same 2D projections. Second, we confirmed that dividing the body into local regions shows outstanding results for estimating 3D motion dynamics. In this paper, we propose (i) an encoder that makes two different estimations: a static feature that presents 2D pose feature as distribution and a dynamic feature that includes optical flow information and (ii) a decoder that divides the body into five different local regions to estimate the 3D motion dynamics of each region. We demonstrate how our method recovers the accurate and smooth motion and achieves the state-of-the-art results for both constrained and in-the-wild videos.

1. Introduction

Reconstructing a 3D human mesh can be used for many applications, including motion analysis, virtual and augmented reality, gaming, and biometrics. However, estimating 3D human pose and shape from a single image or video is a challenging problem because of the limited 3D scan data and the ambiguity that multiple 3D configurations can result in the same 2D projection. To address the above difficulties, Loper *et al.* [27] introduced a parametric 3D human mesh model, SMPL, that was learned from thousands of 3D body scans. Recently, many studies have been proposed to directly regress the model parameters from the input image by utilizing the power of the DCNN, which have shown im-

pressive results [3, 12, 28, 32, 11, 36]. However, these single image-based methods tend to produce temporally inconsistent and unsmooth 3D motion when applied to a video.

Several methods [2, 5, 37, 29, 38, 18] have been proposed to effectively extend single image-based methods to video cases. They have introduced the concept of temporal network to SMPL. This network makes a model learn directly from a video to better capture temporal information. However, these methods are still not capable of recovering the accurate and smooth 3D human motion. Among the above studies, contrary to other methods that showed limitations in recovering smooth 3D motion, the model proposed in [2] succeeded in reducing the temporal inconsistency by learning 3D motion dynamics but showed a low per-frame accuracy. To address this problem, we approach some of the fundamental problems of 3D reconstruction tasks, simultaneously learning 3D pose and 3D motion dynamics.

The main reason that the 3D reconstruction task is challenging derives from the existence of ambiguity in that various 3D poses can be projected into the same or similar 2D poses. The estimated 3D meshes can be completely wrong even though they are closely matched with input images when projected into 2D space. However, existing studies have not addressed this problem directly. We found that we could improve robustness of the model on such ambiguity by utilizing the power of uncertainty in the embedding step [15]. As 2D poses have an inherent ambiguity, it is difficult to represent 2D poses through a deterministic mapping, which previous 3D human reconstruction methods use in the latent feature space. Unlike previous methods, we propose employing a view-invariant probabilistic encoder for a static feature that presents 2D pose features as distribution to inform the decoder of the uncertainty information in 2D space. As an ideal model reconstructs a view-invariant 3D human mesh, the uncertainty concept plays an important role in 3D human reconstruction task. Furthermore, we introduce a novel method to optimize the decoding process using uncertainty-aware pose loss, which further helps the model to reconstruct an accurate 3D pose. Apart from the static feature taking into account the uncertainty of the 2D pose, we also estimate the dynamic feature including optical flow information from the video. This dynamic feature is effective for estimating 3D motion changes in a short period of time. The encoding method we suggested makes two different estimations; a static feature and a dynamic feature from the video show a significant effect on the decoder to recover the accurate and smooth 3D motion.

Additionally, we confirmed that dividing the body into local regions shows outstanding results for estimating 3D motion dynamics. Estimating 3D motion dynamics for all joints together is difficult, as the deformations of the local body regions are different. The nearby joints have strong dependency, while the dependency of the distant joints is weak. We propose to estimate 3D dynamics by dividing the entire body into five local body regions: torso, left arm, right arm, left leg, and right leg. Unlike existing methods that ignore spatial relationships between features by using a fully connected layer (FCN) to estimate 3D pose and 3D motion dynamics, we model the spatial relationships between local regions in the decoding process. This allows the model to consider the independent characteristics of different local body regions while making the joints in the same local body region more dependent. This decoding method also enables our network to better infer about uncommon global poses by learning the distribution of local body poses instead of the distribution of global body poses.

In this paper, we propose a 3D human reconstruction method that can estimate the accurate and smooth motion from video. Qualitative and quantitative results show that our method outperforms previous state-of-the-art methods for both constrained and in-the-wild videos. The contributions of the paper can be summarized as follows:

- We propose to estimate two different features from video: a static feature and a dynamic feature for simultaneously predicting 3D pose and motion dynamics.
- We propose employing a view-invariant probabilistic encoder that presents 2D pose features as distribution for considering uncertainty in 2D space. Furthermore, we introduce an uncertainty loss in the decoding process.
- We present a decoder that divides the body into five different local regions to estimate the 3D motion dynamics of each region.

2. Related Work

3D pose and shape from a single image. The first concept of SMPL, a statistical body shape model, was built by Loper *et al.* [27]. Since then, many efforts have been made to improve the model-based approach for 3D pose and shape estimation, which predicts SMPL parameters from an input image. Bogo *et al.* [10] proposed the first end-to-end approach, SMPLify, which fits the SMPL model to the output of an off-the-shelf keypoint detector [26]. Lassner *et al.* [7] used silhouettes along with keypoints in the fitting procedure. Recently, with rapidly developing power in neural networks, several attempts have been made to use DCNN to directly regress the SMPL parameters from pixels [33, 3, 28, 12, 22, 20, 13, 32]. They used a 2D keypoint reprojection loss [3, 22, 20], body/part segmentation [28, 12] as cues for weak supervision. Kanazawa *et al.* [3] proposed an end-to-end trainable human mesh recovery system that penalizes a statistically implausible 3D human mesh through adversarial training. Kolotouros *et al.* [32] introduced a collaboration between regression-based and optimization-based

methods for a self-improving system. On the other hand, a few model-free approaches [14, 31, 16, 19] have been presented to regress mesh vertex coordinates directly. Varol *et al.* [14] proposed BodyNet, which estimates the 3D human shape in 3D volumetric space. Kolotouros *et al.* [31] proposed a Graph CNN architecture that takes a SMPL template mesh as input and estimates the 3D vertex coordinates using image features from ResNet [23]. Moon and Lee [16] introduced a lixel-based heatmap to localize mesh vertices in a fully convolutional manner. Choi *et al.* [19] proposed the recovery of a 3D human mesh from a 2D pose using a Graph CNN. However, these single image-based methods tend to produce jitter when applied to a video.

3D pose and shape from video. Several methods have been conducted to exploit temporal information for estimating 3D pose and shape from video [4, 2, 37, 5, 29, 38, 18]. Arnab *et al.* [4] presented a bundle adjustment method to improve HMR for temporally consistent fits of the SMPL model. Kanazawa *et al.* [2] proposed learning 3D human dynamics to reduce the 3D prediction's temporal inconsistency. Doersch *et al.* [5] used 2D keypoint heatmaps and a sequence of optical flow to train their network. Sun *et al.* [37] proposed a skeleton-disentangling based framework that divides 3D human pose and shape estimation task into multi-level spatial and temporal granularity. They enforced the network with an unsupervised adversarial training strategy, temporal shuffles and order recovery. Kocabas *et al.* [29] proposed using a bi-directional gated recurrent unit (GRU) [25] to extract a temporal feature from video and feed it to a SMPL parameter regressor. A motion discriminator was introduced to encourage the regressor to produce plausible 3D human motion. Luo *et al.* [38] estimated 3D human motion in two stages. They first capture the coarse overall motion using a variational motion estimator and then refine the pose using the motion refinement regressor. Choi *et al.* [18] proposed to remove the residual connection between the static and temporal features and forecast the current temporal features from the past and future frames for temporally consistent motion. However, these methods are still not capable of recovering the accurate and smooth 3D human motion.

3. Proposed Method

Our method covers two challenging tasks: (i) estimating an uncertainty-aware temporal feature from video and (ii) recovering the accurate and smooth 3D motion by dividing the body into five different local regions to estimate part-based dynamics. The overall framework of our method is shown in Figure 2.

3.1. Problem Setup

Given an input video $V = \{I_t\}_{t=1}^{T}$ of length T, where I_t denotes the t^{th} frame, our goal is to recover human motion sequences $M = \{\Theta_t\}_{t=1}^{T}$ where each Θ_t represents the SMPL [27] parameters for the t^{th} frame. The SMPL parameters Θ consists of the pose $\theta \in \mathbb{R}^{24\times 3}$ and shape $\beta \in \mathbb{R}^{10}$ parameters. While θ models the global body rotation and the relative rotation of 23 joints in axis-angle format, β models the body shape as captured by the first 10 coefficients of a PCA shape space. Given θ and β, SMPL defines a function $\mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{6890\times 3}$ that outputs a 3D human mesh. The SMPL 3D joint locations $X(\Theta) = W\mathcal{M}(\theta, \beta)$ are defined as a linear combination of the mesh vertices via a pre-trained linear regressor, W. To project the 3D joints X back to 2D space, we use a weak perspective camera model with scale and translation parameters $[s, t], t \in \mathbb{R}^2$. We denote $x \in \mathbb{R}^{j\times 2} = s\Pi(RX(\Theta)) + t$ as the 2D projection of the 3D joints, where $R \in \mathbb{R}^3$ is the global rotation matrix and Π represents the orthographic projection.

3.2. Uncertainty-aware Temporal Feature

Our temporal encoder extracts an uncertainty-aware temporal feature that includes uncertainty and dynamics information from video. First, we encode image features into a temporal feature, following Kocabas *et al.* [29]. Given a sequence of frames $I_1, \ldots, I_T$, ResNet, pre-trained by Kolotouros *et al.* [32], extracts an image feature per frame. Then, a global average is applied to the ResNet outputs, which become $\mathbf{f}_1, \ldots, \mathbf{f}_T$, where $\mathrm{f}_t \in \mathbb{R}^{2048}$. These are sent to a GRU layer that yields a temporal feature $\mathbf{g}_1, \ldots, \mathbf{g}_T$ based on the previous frames, where $\mathrm{g}_t \in \mathbb{R}^{2048}$. This feature is concatenated with a static feature that considers the uncertainty in 2D space and a dynamic feature that includes optical flow information, which are described as follows.

Uncertainty-aware static feature. An ideal embedding vector z for 2D pose should remain consistent across camera views. However, human poses in 2D space have an inherent ambiguity in that various 3D poses can be projected into the same 2D pose and deterministic mapping in encoding stage that does not consider the ambiguity hinders the performance in 3D reconstruction task. Inspired by Sun *et al.* [21], we propose employing an uncertainty-aware pose encoder using probabilistic embedding which encodes a 2D pose feature based on a Gaussian distribution, $p(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\sigma}^2\mathbf{I})$. We use 2D pose outputs from HRNet [24]. For a pair of input 2D poses $(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $p(m \mid \boldsymbol{x}_i, \boldsymbol{x}_j)$ is defined as the probability that their corresponding 3D poses $(\boldsymbol{y}_i, \boldsymbol{y}_j)$ match. Probabilistic embedding can be used to this matching probability as $p(m \mid \boldsymbol{x}_i, \boldsymbol{x}_j) = \int p(m \mid \boldsymbol{z}_i, \boldsymbol{z}_j)\, p(\boldsymbol{z}_i \mid \boldsymbol{x}_i)\, p(\boldsymbol{z}_j \mid \boldsymbol{x}_j)\, \mathrm{d}\boldsymbol{z}_i\, \mathrm{d}\boldsymbol{z}_j$. It can be approximated using Monte-Carlo sampling with K samples drawn from each distribution as

$$p(m \mid \boldsymbol{x}_i, \boldsymbol{x}_j) \approx \frac{1}{K^2} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} p\left(m \mid \boldsymbol{z}_i^{(k_1)}, \boldsymbol{z}_j^{(k_2)}\right). \quad (1)$$

A combination of triplet ratio loss and positive pairwise loss

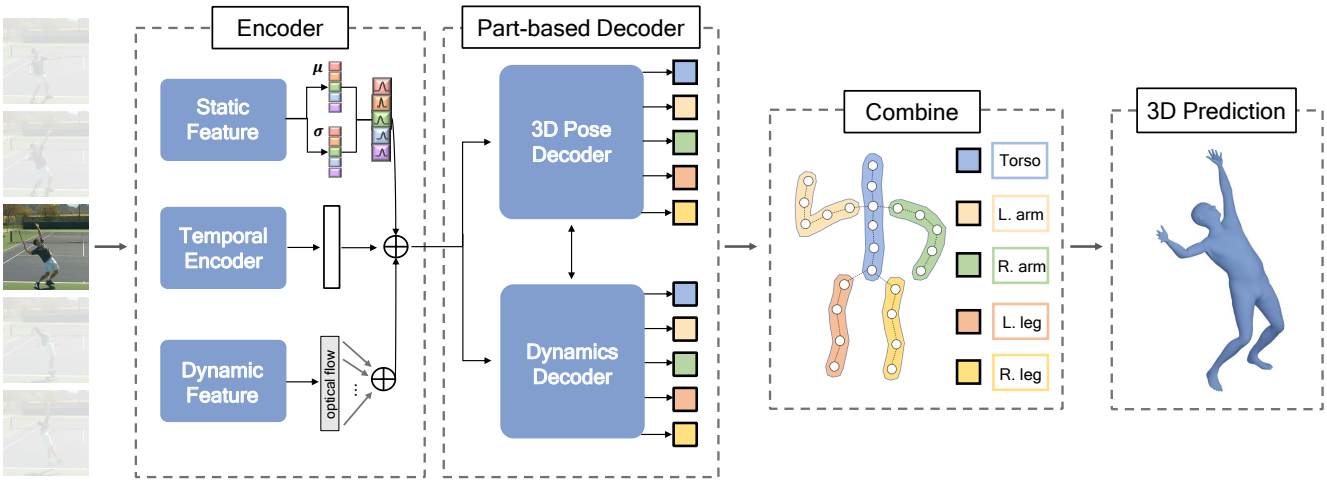

Figure 2: The overall framework of the proposed method. Given a temporal sequence of images, the model extracts an uncertainty-aware temporal feature that includes uncertainty in 2D space and optical flow information. Then, the decoder simultaneously predicts 3D pose and part-based motion dynamics to recover accurate and smooth motion.

is used for the training process of estimating this uncertainty value. The triplet ratio loss pushes together/pulls apart 2D poses corresponding to similar/dissimilar 3D poses. Given batch size N and input triplet $(\boldsymbol{x}_i, \boldsymbol{x}_{i^+}, \boldsymbol{x}_{i^-})$, triplet ratio loss using distance kernel D_m is defined as

$$\mathcal{L}_{\text{ratio}} = \sum_{i=1}^{N} \max(0, D_m(\boldsymbol{z}_i, \boldsymbol{z}_{i^+}) - D_m(\boldsymbol{z}_i, \boldsymbol{z}_{i^-}) + \alpha). \tag{2}$$

A positive pairwise loss is applied to increase the matching probability of similar poses and defined as

$$\mathcal{L}_{\text{positive}} = \sum_{i=1}^{N} -\log p\left(m \mid \boldsymbol{z}_i, \boldsymbol{z}_{i^+}\right). \tag{3}$$

The overall loss function for uncertainty-aware embedding model is the combination of $\mathcal{L}_{\text{ratio}}$ and $\mathcal{L}_{\text{positive}}$ as following:

$$\mathcal{L}_{\text{uncertainty}} = \mathcal{L}_{\text{ratio}} + \mathcal{L}_{\text{positive}}. \tag{4}$$

The model outputs the mean $\mu_t \in \mathbb{R}^{32}$ and covariance $\sigma_t \in \mathbb{R}^{32}$ of the 2D pose. We then combine them into an uncertainty-aware static feature $u_t \in \mathbb{R}^{64}$. The estimated σ value can denote the uncertainty and be applied to our loss function for recovering human motion sequences.

Dynamic feature. Optical flow has strong cues for motion dynamics, which can be a key to solving the problem that the previous methods produce jitter for fast motion, resulting in temporally inconsistent motion. We extract the dynamic feature including optical flow information between each successive frame, following [9, 17]. The stack of homographies can be used to represent the optical flow within the interval. We estimate the homography from flow correspondences by solving a homogeneous linear equation via SVD. Then, the output 3×3 homography matrix is normalized by the top-left corner element. For a frame I_t, the optical flow information is constructed by calculating the homographies between successive frames within the interval $[I_{t-15}, I_t]$. We combine the homographies into a $d_t \in \mathbb{R}^{135}$ vector. This dynamic feature shows effective results for predicting our part-based motion dynamics in the decoding process.

3.3. Learning Part-based 3D Human Dynamics

We propose to simultaneously predict 3D pose and part-based 3D motion dynamics. The auxiliary loss for learning to predict 3D motion dynamics helps in estimating temporally consistent and smooth motion. Furthermore, we propose dividing the entire body into five local body regions: torso, left arm, right arm, left leg, and right leg is effective for estimating 3D dynamics because the deformations of the local body regions are different. The joint positions within each group are highly correlated, while the joint positions between the groups are significantly less related. However, previous methods ignore the spatial relationships by using FCN layers to predict 3D pose and dynamics. Each predicted pose/dynamics and each intermediate feature is connected to all of the input features indiscriminately. To address this problem, we use split-and-recombine model [1] to estimate our part-based motion dynamics. We divide the body into five groups and the FCN layers are divided into groups accordingly. Low-Dimensional Global Conetxt (LDGC) is incorporated in a group connected layer to account for global information while largely preserving local feature independence. It coarsely represents informa-

tion from the less relevant joints, and is brought back to the local group in a manner that limits disruption to the local pose modeling while allowing the local group to account for non-local dependencies. With this split-and-recombine approach, the g_{th} group features in layer $l+1$, $\mathbf{f}^{l+1}\left[\mathcal{G}_g^{l+1}\right]$, can be expressed as the following:

$$\mathbf{f}^{l+1}\left[\mathcal{G}_g^{l+1}\right] = \Theta_g^l\left(\mathbf{f}^l\left[\mathcal{G}_g^l\right] \circ \mathcal{M}\mathbf{f}^l\left[\mathcal{G}^l \backslash \mathcal{G}_g^l\right]\right), \quad (5)$$

where Θ_g^l is the fully-connected weight matrix, $\mathbf{f}^l\left[\mathcal{G}^l \backslash \mathcal{G}_g^l\right]$ is the global context for the g^{th} group and $\mathcal{M}$ is the mapping function that defines how the global context is represented. The split-and-recombine based model is used to learn a dynamics decoder that predicts the change in SMPL parameters at time $t \pm \Delta t$. The dynamics predictor is trained such that the predicted pose in the new timestep $\theta_{t\pm\Delta t} = \theta_t \pm \Delta\theta$ minimizes the generator loss at time frame $t \pm \Delta t$. The part-based dynamics learning enforces the model to better recover smooth motion and infer about uncommon 3D poses.

3.4. Loss Functions

We propose a loss function consisting of the generator loss, $\mathcal{L}_{\mathcal{G}}$, dynamics loss, $\mathcal{L}_{\Delta t}$, and uncertainty loss, $\mathcal{L}_{unc}$. As long as the respective data are available, our network is trained with the loss function as following:

$$\mathcal{L} = \lambda_{\mathcal{G}}\mathcal{L}_{\mathcal{G}} + \lambda_{\Delta t}L_{\Delta t} + \lambda_{unc}\mathcal{L}_{unc}, \quad (6)$$

where we weight each of loss terms with λ parameters. **Generator loss.** Generator loss consists of three L2 losses between the predicted and groud-truth 2D/3D joint positions and SMPL parameters. Specifically:

$$L_{\mathcal{G}} = L_{2D} + L_{3D} + L_{SMPL}, \quad (7)$$

where each term is calculated as:

$$L_{2D} = \sum_{t=1}^{T} \|x_t - \hat{x}_t\|_2, \quad (8)$$

$$L_{3D} = \sum_{t=1}^{T} \|X_t - \hat{X}_t\|_2, \quad (9)$$

$$L_{SMPL} = \|\beta - \hat{\beta}\|_2 + \sum_{t=1}^{T} \|\theta_t - \hat{\theta}_t\|_2. \quad (10)$$

Dynamics loss. The dynamics predictor is trained such that the predicted pose in the new timestep $\theta_{t\pm\Delta t} = \theta_t \pm \Delta\theta$ minimizes the generator loss at time frame $t \pm \Delta t$. Camera parameters are required to estimate the 2D joints for calculating the 2D loss. The optimal scale s and translation $\vec{t}$ parameters align the orthographically projected 3D joints $x_{\text{orth}} = X[:, :2]$ with the visible ground-truth 2D joints x_{gt} : $\min_{s,\vec{t}}\|(sx_{\text{orth}} + \vec{t}) - x_{gt}\|_2$. We solve this problem following [2] and use the optimal camera parameters $\Pi^* = [s^*, \overrightarrow{t^*}]$ to compute the 2D joint positions loss at times $t \pm \Delta t$.

Uncertainty loss. Simply optimizing for similarity of 2D poses can fool the network to output a completely wrong 3D mesh that is closely matched with input image when projected into 2D space. The similarity between the features from the uncertainty-aware encoder of the input image and the reconstructed result can help our method to be robust to inherent ambiguity in 2D space. Based on this concept, we propose an uncertainty loss as follows:

$$\mathcal{L}_{unc} = \sum_{t=1}^{T} \|\frac{\sigma_t^2}{\sum \sigma_t^2}\mu_t - \frac{\hat{\sigma}_t^2}{\sum \hat{\sigma}_t^2}\hat{\mu}_t\|_2. \quad (11)$$

4. Experimental Results

4.1. Datasets and Evaluation Metrics

Datasets. We use Human 3.6M [6], MPI-INF-3DHP [8], and 3DPW [34] for 3D datasets. Human 3.6M consists of motion capture sequences of actors performing tasks in a controlled lab environment. MPI-INF-3DHP is a dataset captured with a multi-view setup mostly in indoor environments. 3DPW contains 61 sequences of indoor and outdoor activities. On the other hand, we use InstaVariety [2], Penn Action [35], and PoseTrack [30] for 2D video datasets. InstaVariety dataset contains annotated pseudo ground truth 2D keypoints paired with video sequences. There are in total 28,272 videos with varying length. Penn Action consists of 15 sports actions, with 1,257 training videos and 1,068 test videos. PoseTrack is a benchmark for multi-person pose estimation and tracking in videos and contains 1,337 videos.
Evaluation metrics. We report several metrics on 3DPW, MPI-INF,3DHP, and Human3.6M datasets. For the per-frame accuracy evaluation, we use Procrustes-aligned mean per joint position error (PA-MPJPE), mean per joint position error (MPJPE), and mean per vertex position error (MPVPE). The position errors are measured in millimeter between the estimated and ground-truth 3D vertices after aligning the root joint. For the motion smoothness evaluation, we report the acceleration error. The acceleration error computes an average of the difference in acceleration between the estimated and ground-truth 3D joints in (mm/s^2).

4.2. Comparisons to the State-of-the-art

We show our model's power to recover accurate and smooth 3D human motion from video by comparing the results with previous state-of-the art frame-based [3, 31, 32, 16, 19] and temporal [2, 5, 37, 29, 18] methods. As shown in Table 1, our method achieves the state-of-the-art performance in terms of per-frame accuracy with significantly

	Models	3DPW				MPI-INF-3DHP			Human3.6M		
		PA-MPJPE ↓	MPJPE ↓	MPVPE↓	Accel↓	PA-MPJPE↓	MPJPE↓	Accel↓	PA-MPJPE↓	MPJPE↓	Accel↓
Frame-based	Kanazawa *et al.* [3]	76.7	130.0	-	37.4	89.8	124.2	-	56.8	88	-
	Kolotouros *et al.* [31]	70.2	-	-	-	-	-	-	50.1	-	-
	Kolotouros *et al.* [32]	59.2	96.9	116.4	29.8	67.5	105.2	-	41.1	-	18.3
	Moon *et al.* [16]	57.7	93.2	110.1	30.9	-	-	-	41.1	**55.7**	13.4
	Choi *et al.* [19]	58.3	**88.9**	106.3	22.6	-	-	-	46.3	64.9	23.9
Temporal	Kanazawa *et al.* [2]	72.6	116.5	139.3	15.2	-	-	-	56.9	-	-
	Doersch *et al.* [5]	74.7	-	-	-	-	-	-	-	-	-
	Sun *et al.* [37]	69.5	-	-	-	-	-	-	42.4	59.1	-
	Kocabas *et al.* [29]	56.5	95.8	113.4	27.1	63.4	97.7	29.0	41.5	65.9	18.3
	Choi *et al.* [18]	55.8	95.0	111.5	7.0	62.8	97.4	**8.0**	41.1	62.3	**5.3**
	Ours	**52.2**	92.8	**106.1**	**6.8**	**59.4**	**93.5**	9.4	**38.4**	58.4	6.1

Table 1: Evaluation of state-of-the-art methods on 3DPW, MPI-INF-3DHP, and Human3.6M datasets. The best results are highlighted in bold and "–" shows the results that are not available. Our method achieves the state-of-the-art performance in terms of per-frame accuracy with significantly low acceleration error.

low acceleration error in an indoor dataset Human3.6M and challenging in-the-wild datasets 3DPW and MPI-INF-3DHP. Figure 3 also shows that our method achieves significant improvements in acceleration error. These results validate our hypothesis that exploiting uncertainty information and 3D dynamics of local body regions is important for improving both per-frame accuracy and motion smoothness. We briefly compare the quantitative results of our approach with those of previous state-of-the-art methods. Kanazawa *et al.* [3] proposed an adversarial prior that constrains the 3D human pose to lie in the manifold of real human poses. Kolotouros *et al.* [32] proposed combining regression-based and optimization-based methods in a collaborative fashion by using SMPLify in the training loop. [3, 32] improved the per-frame accuracy of the model-based approach. On the other hand, several studies on model-free approaches [31, 16, 19] have been proposed to better estimate in-the-wild 3D poses, and they showed improved accuracy. However, the acceleration errors shown in the above studies are significantly high and these methods yield jittery when applied to video. The work most related to our method is [2]. They showed a large performance improvement in terms of acceleration error, but their aggressive smoothing resulted in poor accuracy in fast motion or extreme poses. Kocabus *et al.* [29] proposed a temporal model and AMASS motion discriminator to approach the trade-off between per-frame accuracy and smoothness, and they showed performance improvement. However, they still produce many jitters and estimated motions are unsmooth. Recently, Choi *et al.* [18] showed remarkable performance improvement in terms of both per-frame accuracy and smoothness, but performance improvement is still needed in both aspects. Our method significantly surpasses the existing methods in terms of per-frame accuracy and simultaneously shows remarkable improvement in motion smoothness. The qualitative comparisons in Figure 4, 5 show the superiority of our method to previous state-of-the-art, VIBE [29]. Our method

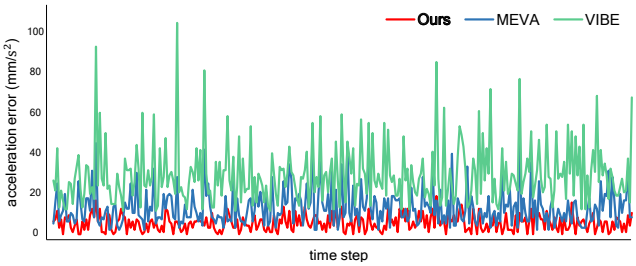

Figure 3: Comparison of the acceleration errors for our method, MEVA [38], and VIBE [29]. Our method shows clearly lower acceleration errors than previous methods.

is able to produce accurate 3D meshes for difficult poses and we can confirm this especially in the arms and legs.

4.3. Ablation Study

In this study, we show how each component of the two tasks that we propose leads to improvements in per-frame accuracy and motion smoothness, respectively.

Uncertainty-aware temporal feature. Table 2 shows the performance of the models with different temporal encoders. We use the original VIBE [29] as a baseline feature extractor(w/o both) and add an uncertainty-aware static feature and a dynamic feature to the temporal feature from baseline to identify performance changes accordingly. The per-frame accuracy remarkably increased with uncertainty-aware static feature embedding. This result verifies that the deterministic embedding of the 2D pose hinders the 3D mesh recovery because there is an uncertainty in the 2D pose. We deliver 2D pose uncertainty information to the decoder using view-invariant probabilistic embedding, and this information is helpful in estimating the accurate 3D poses. In addition, estimating a dynamic feature in the embedding step shows a considerable effect in terms of motion smoothness by lowering the acceleration errors of the baseline. This proves that extracting a dynamic feature includ-

Figure 4: Qualitative comparison with VIBE [29]. For each sequence, the top row shows the input images, the middle row shows our results (blue), and the bottom row shows the results of VIBE (gray). Our method can produce accurate 3D poses.

Figure 5: Qualitative comparison on 3DPW dataset with VIBE [29]. The output meshes from VIBE and our method are rendered in pink and blue, respectively. Our method (blue) is able to produce accurate 3D meshes for difficult poses.

	3DPW			
	PA-MPJPE↓	MPJPE↓	MPVPE↓	Accel↓
w/o both	56.8	97.4	113.7	9.8
w/o u_t	56.3	96.6	112.6	7.8
w/o d_t	53.0	93.8	107.2	9.4
Ours	**52.2**	**92.8**	**106.1**	**6.8**

Table 2: Comparison between different temporal encoders. The results show the importance of the static features u_t and dynamic features d_t.

ing optical flow information is helpful to estimate 3D motion changes. Our embedding method is effective in terms of both per-frame accuracy and motion smoothness.

Part-based 3D dynamics prediction. Table 3 shows the acceleration error significantly reduced with learning part-based 3D dynamics. This result verifies that dividing the entire body into five local body regions is effective for estimating 3D motion changes as the deformations of the local body regions are different. Additionally, it showed a considerable effect in estimating an accurate 3D mesh by lowering the per-frame accuracy error. Our method better infer about uncommon global poses by learning the distribution of local body poses instead of the distribution of global body pose, which is also shown in qualitative results. Through the above comparisons, we confirmed that our decoding method is effective in terms of both per-frame accuracy and motion smoothness.

	3DPW			
	PA-MPJPE↓	MPJPE↓	MPVPE↓	Accel↓
w/o d_t, dynamics	56.1	96.3	112.8	21.6
w/o dynamics	55.7	95.8	111.4	12.7
Ours	**52.2**	**92.8**	**106.1**	**6.8**

Table 3: Ablation study on part-based dynamics predictor.

5. Conclusion

In this paper, we present an uncertainty-aware human mesh recovery method that uses uncertainty information in a 2D pose to directly address a fundamental problem of 3D reconstruction tasks. Our method ensures the decoder to see the uncertain features that can further improve robustness of the model on inherent ambiguity in 2D space. We also propose to divide the body into five different local regions to estimate the 3D motion dynamics of each region. Our method outperforms previous state-of-the-art methods and this work can be a step forward in finding effective feature embedding techniques for 3D human reconstruction.

Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-00079, Artificial Intelligence Graduate School Program (Korea University), No. 2019-0-01371, Development of brain-inspired AI with human-like intelligence).

References

[1] Ailing Zeng, Xiao Sun, Fuyang Huang, Minhao Liu, Qiang Xu, and Stephan Lin. Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In *ECCV*, 2020. 4

[2] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *CVPR*, 2019. 2, 3, 5, 6

[3] Angjoo Kanazawa, Michael J Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 2, 5, 6

[4] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *CVPR*, 2019. 3

[5] Carl Doersch and Andrew Zisserman. Sim2real transfer learning for 3d human pose estimation: motion to the rescue. In *NeurIPS*, 2019. 2, 3, 5, 6

[6] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. In *TPAMI*, 2014. 5

[7] 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*, 2017. 2

[8] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3DV*, 2017. 5

[9] Evonne Ng, Donglai Xiang, Han-Byul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. In *CVPR*, 2020. 4

[10] 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*, 2016. 2

[11] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Kosecka, and Ziyan Wu. Hierarchical kinematic human mesh recovery. In *ECCV*, 2020. 2

[12] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *CVPR*, 2018. 2

[13] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. Texturepose: Supervising human mesh estimation with texture consistency. In *ICCV*, 2019. 2

[14] 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*, 2018. 3

[15] Gun-Hee Lee and Seong-Whan Lee. Uncertainty-aware mesh decoder for high fidelity 3d face reconstruction. In *CVPR*, 2020. 2

[16] Gyeong-Sik 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*, 2020. 3, 5, 6

[17] Hao Jiang and Kristen Grauman. Seeing invisible poses: Estimating 3d body pose from egocentric video. In *CVPR*, 2017. 4

[18] Hong-Suk Choi, Gyeong-Sik Moon, and Kyoung-Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *arXiv*, 2020. 2, 3, 5, 6

[19] Hong-Suk Choi, Gyeong-Sik Moon, and Kyoung-Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *ECCV*, 2020. 3, 5, 6

[20] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadak. Self-supervised learning of motion capture. In *NIPS*, 2017. 2

[21] Jennifer J. Sun, Jiaping Zhao, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, and Ting Liu. View-invariant probabilistic embedding for human pose. In *ECCV*, 2020. 3

[22] Jun Kai Vince Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3d human shape and pose prediction. In *BMVC*, 2017. 2

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[24] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 3

[25] Kyung-Hyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*, 2014. 3

[26] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, 2016. 2

[27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. In *ACM TOG*, 2015. 1, 2, 3

[28] Mohamed Omran, Christoph Lassner, Gerard PonsMoll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *3DV*, 2018. 2

[29] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 2, 3, 5, 6, 7, 8

[30] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *CVPR*, 2018. 5

[31] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. 3, 5, 6

[32] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 2, 3, 5, 6

[33] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *CVPR*, 2019. 2

[34] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 5

[35] Weiyu Zhang, Menglong Zhu, and Konstantinos G. Derpanis. From actemes to action: A strongly supervised representation for detailed action understanding. In *ICCV*, 2013. 5

[36] Yu Rong, Ziwei Liu, Cheng Li, Kaidi Cao, and Chen Change Loy. Delving deep into hybrid annotations for 3d human recovery in the wild. In *ICCV*, 2019. 2

[37] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, YiLi Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *ICCV*, 2019. 2, 3, 5, 6

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