sensors

MDPI

Article

Camera Motion Agnostic Method for Estimating 3D Human Poses

Seong Hyun Kim [1], Sunwon Jeong [2], Sungbum Park [2] and Ju Yong Chang [1,*]

[1] Department of Electronics and Communications Engineering, Kwangwoon University, Seoul 01897, Korea
[2] Vision AI Lab, AI Center, NCSOFT, Seongnam 13494, Korea
* Correspondence: jychang@kw.ac.kr; Tel.: +82-2-940-5136

Abstract: Although the performance of 3D human pose and shape estimation methods has improved considerably in recent years, existing approaches typically generate 3D poses defined in a camera or human-centered coordinate system. This makes it difficult to estimate a person's pure pose and motion in a world coordinate system for a video captured using a moving camera. To address this issue, this paper presents a camera motion agnostic approach for predicting 3D human pose and mesh defined in the world coordinate system. The core idea of the proposed approach is to estimate the difference between two adjacent global poses (i.e., global motion) that is invariant to selecting the coordinate system, instead of the global pose coupled to the camera motion. To this end, we propose a network based on bidirectional gated recurrent units (GRUs) that predicts the global motion sequence from the local pose sequence consisting of relative rotations of joints called global motion regressor (GMR). We use 3DPW and synthetic datasets, which are constructed in a moving-camera environment, for evaluation. We conduct extensive experiments and prove the effectiveness of the proposed method empirically.

Keywords: 3D human pose estimation; 3D human shape reconstruction; statistical shape model

Citation: Kim, S.H.; Jeong, S.; Park, S.; Chang, J.Y. Camera Motion Agnostic Method for Estimating 3D Human Poses. *Sensors* **2022**, *22*, 7975. https://doi.org/10.3390/s22207975

Academic Editors: Roberto Vezzani, Mohamed Daoudi, Guido Borghi, Marcella Cornia, Claudio Ferrari, Federico Becattini and Andrea Pilzer

Received: 6 September 2022
Accepted: 18 October 2022
Published: 19 October 2022

1. Introduction

3D human pose estimation [1–11] is an important topic in computer vision that can be applied to many applications, such as virtual/augmented reality, human action recognition, and human behavior understanding. Various sensors, such as multi-view cameras with markers [12,13], depth cameras [14], and inertial measurement units (IMUs) [15], can be used for 3D human pose estimation. Despite its high accuracy, the marker-based method using multi-view cameras has disadvantages in that its hardware is expensive and setup is complicated. Also, the depth camera-based method generally does not work well outdoors, and the IMUs-based method suffers from heading drift. On the other hand, 3D human pose estimation based on a monocular color camera does not require markers, is relatively low-cost, has high flexibility, and thus has recently received much attention. The performance of 3D human pose estimation based on a monocular camera has improved thanks to advances in deep learning remarkably.

The majority of 3D human pose estimation methods reconstruct 3D poses defined in the camera or human-centered coordinate system. The estimated 3D human pose is coupled to the camera pose. Therefore, reconstructing intrinsic human poses for a video sequence captured by a moving camera is challenging. Our paper addresses this problem and proposes a method to estimate the *intrinsic human pose* independent of camera motion. Figure 1 shows the difference between 3D human pose sequences reconstructed using the proposed and existing methods.

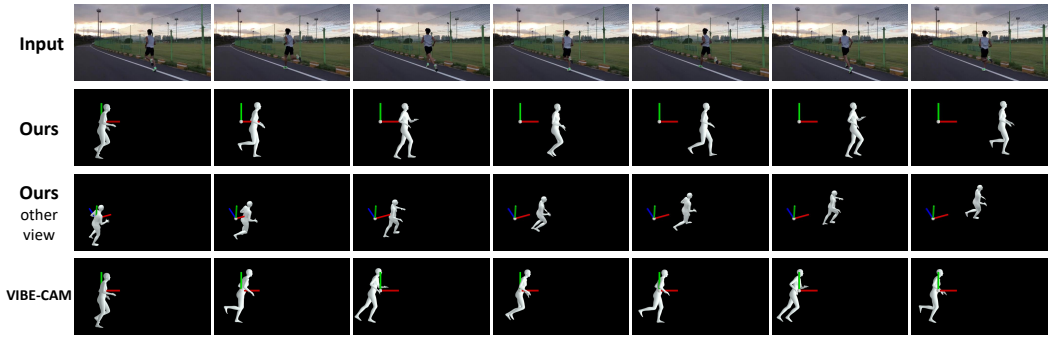

Figure 1. Given a runner video (first row), the proposed framework correctly reconstructs 3D running path (second and third rows), while VIBE-CAM, the combination of state-of-the-art human pose estimation methods [11,16,17], fails to reconstruct the 3D global pose of the runner (fourth row). The global pose represents the orientation and location of the entire body. The visualized reference frame is defined as being aligned with the person in the first frame. VIBE-CAM is detailed in Section 4.5.

In the kinematic chain model for human body or the statistical human shape model, such as SMPL [18], 3D human poses can be decomposed into a *local pose* that represents the orientation of rigid body parts and a *global pose* that represents the orientation and location of the entire body, as shown in Figure 2. The local pose is represented hierarchically through relative rotations of rigid body parts from the rest pose (i.e., zero pose) and defined in the generic coordinate system [18]. Therefore, the local pose is independent of the selection of the reference coordinate system. However, the global pose is dependent on the selection of the reference coordinate system. The global pose is generally defined on the basis of the camera coordinate system in existing methods [5,6,11,19,20]; thus, the estimated 3D human pose is coupled to the camera motion. Our basic idea is to estimate the difference in the global pose in adjacent frames (i.e., global motion) invariant to the selection of the reference coordinate system instead of the global pose coupled to the camera motion.

So, how can we estimate the global motion decoupled from the camera motion? We speculate that the *global motion* (i.e., global pose displacement between neighboring frames) can be predicted from the local pose sequence, as shown in Figure 2. Suppose a person makes a jump to the left. We can easily infer that the person jumps to the left, as shown in Figure 2 (bottom row) from the local pose sequence in Figure 2 (top row). Therefore, this study aims to design a deep network that estimates the global human motion sequence from the local human pose sequence. Specifically, the local pose sequence is reconstructed from a video using an existing 3D human pose estimation method, such as VIBE [11]. We model the mapping function from the input local pose sequence to the output global motion sequence through a temporal network called global motion regressor (GMR) and train the network using the large-scale motion capture dataset AMASS [21].

We evaluate the proposed method using the 3DPW dataset [15]. We also synthesize an animated 3D human pose dataset using CMU sequences in the AMASS dataset to allow camera movement in the synthetic video. Both datasets are used for qualitative and quantitative evaluations of the proposed method.

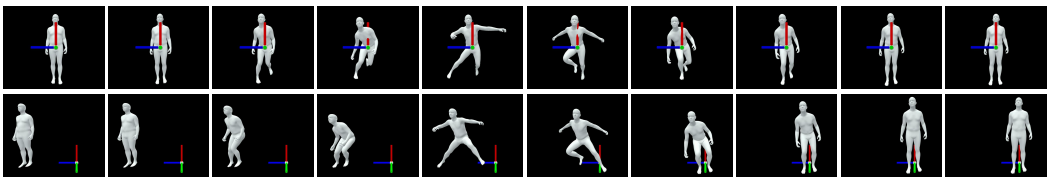

Figure 2. The top row shows the image sequence rendered using only the local pose without the global pose. Here, the relative orientations between rigid body parts (i.e., local pose) change, but the entire body's orientation and location (i.e., global pose) remain unchanged. The bottom row shows the rendering result for the case where the global pose is further included. Please note that the main purpose of the paper is to estimate the global pose sequence from the local pose sequence.

The main contributions of this paper are presented as follows:

- We propose a deep learning-based framework for predicting a pure human pose independent of camera motion. We demonstrate that it is possible to estimate the human pose sequence in the world coordinate system without camera calibration from a video including camera motion.
- We propose a model based on gated recurrent units (GRUs) [22] that transforms the local human pose sequence into the global motion sequence invariant to the selection of the reference coordinate system. The proposed model can be combined with any human pose estimation method that predicts local human poses.
- We propose new metrics for the evaluation of the proposed method. Moreover, we train the proposed model for various input/output rotation representations and rotation loss functions and quantitatively compare them using the proposed evaluation metrics to determine the optimal rotation representation and loss function.

2. Related Works

This section first reviews general methods for reconstructing 3D human poses and shapes simultaneously, which are related but do not have the same goal as our study. Then, an explanation of how existing methods can be utilized to achieve the goal of our study is provided.

3D human pose and shape estimation from a single image. The method for estimating the 3D human pose and shape from a single image can be divided into model-based and model-free approaches. Model-based approaches commonly use the statistical body shape model SMPL [18] to reconstruct the human shape and allow the network to predict parameters of the SMPL model. Meanwhile, the model-free approach performs 3D human shape reconstruction by directly estimating a 3D human mesh instead of predicting SMPL parameters. [5,6,19,23,24] belong to the model-based approach. Kanazawa et al. [5] introduced an adversarial training method to obtain an anthropometrically plausible 3D shape and proposed a discriminator network. Pavlakos et al. [19] used keypoints and silhouettes as an intermediate representation for predicting SMPL parameters. Omran et al. [23] utilized body part segmentation. Kolotouros et al. [6] proposed a method that combines feedforward regression step and SMPLify-based optimization step [25] into a loop structure to combine advantages of regression-based and optimization-based methods. Kocabas et al. [24] estimated body-part-guided attention masks and used them for 3D human pose and shape estimation robust to occlusion. The following references [9,10,26–28] belong to the model-free approach. Varol et al. [26] proposed a network that directly predicts a 3D human mesh in volumetric space and used keypoints, segmentation, and 3D pose as the intermediate representation for this. Kolotouros et al. [9] proposed a graph convolutional network for 3D human mesh reconstruction. Their network takes rest poses and image features as inputs and directly regresses the 3D human mesh. Moon et al. [10] proposed the image-to-lixel prediction network that predicts vertex coordinates of the 3D human mesh through 1D heatmaps. Lin et al. [27] proposed a transformer-based network that simultaneously reconstructs human pose and shape by modeling vertex-vertex and vertex-joint interactions. Lin et al. [28] combined graph convolutional neural networks with their existing transformer-based method to model both local and global interactions simultaneously.

3D human pose and shape estimation from a video. Kanazawa et al. [29] proposed a method for predicting not only 3D meshes that correspond to a single input image but also those that correspond to frames in the past and future through learning using video data. Arnab et al. [30] proposed a bundle-adjustment-based algorithm that temporally and consistently refines initial per-frame SMPL estimates. Sun et al. [31] proposed a transformer-based temporal model. In that study, in order for the network to learn temporal information better, the order of shuffled frames can be predicted, and an unsupervised adversarial training method for this was proposed. Kocabas et al. [11] proposed a temporal model based on GRU. In that study, a motion discriminator network was proposed to allow the

network to generate a plausible 3D human motion. Luo et al. [20] proposed a two-stage model for human motion estimation. Overall coarse motion is predicted using variational motion estimation in the first stage of the model and then further improved through motion residual regressor in the second stage. Choi et al. [32] proposed a method that reconstructs a temporally consistent human shape using temporal information of past and future frames. Wan et al. [33] proposed a multi-level attention-based framework in which three intrinsic relations (i.e., spatial, temporal, and human joint relations) are jointly modeled.

3D human pose estimation in the world coordinate system. All the methods reviewed above reconstruct 3D human pose and shape in the camera coordinate system. The result reconstructed by these methods from a video captured by a fixed camera can be considered to be defined in the world coordinate system. However, assuming the general environment where no extrinsic camera parameters are given, it is difficult to convert the reconstruction result from a video with camera motion into the pure 3D human pose defined in the world coordinate system. One possible method is to extract camera motions from the video using a structure-from-motion (SfM) method such as COLMAP [34], and use them to transform the human pose in the camera coordinate system into the world coordinate system. However, SfM methods often fail to achieve successful results in videos containing dynamic objects. Although foreground-background segmentation can be used for removal of dynamic foreground objects, a fully automated method for camera motion estimation is still unavailable. Our goal is to obtain the pure 3D human pose sequence in the world coordinate system without camera calibration from a video with any camera motion. To achieve that goal, we propose a deep-learning-based method to reconstruct 3D human poses in the world coordinate system from a video captured by a moving camera. The baseline for evaluating the proposed method is constructed by combining the existing 2D human pose estimation [17] and 3D human pose estimation [11,16] methods, and the detailed procedure for it is described in Section 4.5.

3. Proposed Method

3.1. Overall Approach

Figure 3 shows the overall framework of the proposed method. First, we use a human pose estimation network to determine the local pose sequence $\mathbf{L} = \{L_i\}_{i=1}^{T}$ given an input video $\mathbf{V} = \{V_i\}_{i=1}^{T}$ with length T, where $L_i \in \mathbb{R}^{92}$ represents the relative rotations of 23 joints in an unit-quaternion form. Second, bidirectional GRU-based temporal encoder outputs a latent feature containing temporal information of this sequence from the local pose sequence $\mathbf{L}$. We obtain the global motion sequence $\Delta\mathbf{G} = \{\Delta G_i\}_{i=1}^{T}$ from the latent feature through the motion regression layer. A global motion ΔG_i consists of an orientation motion $\Delta A_i \in \mathbb{R}^3$ in an axis-angle form and a translation motion $\Delta T_i \in \mathbb{R}^3$. Third, we accumulate estimated global motions starting with an initial global pose to compute a global pose sequence $\mathbf{G} = \{G_i\}_{i=1}^{T}$. Finally, the computed global pose sequence $\mathbf{G}$ and the input local pose sequence $\mathbf{L}$ are converted into a global human mesh sequence $\mathbf{M}^g = \{M_i^g\}_{i=1}^{T}$ defined in the world coordinate system through the SMPL model [18].

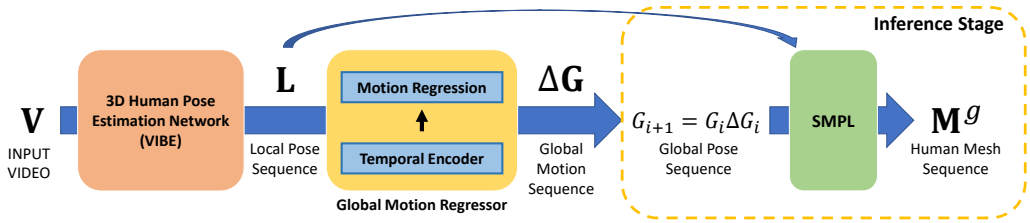

Figure 3. Overall framework of the proposed method. Given an input video, the existing 3D human pose estimation network outputs a local human pose sequence. The proposed global motion regressor generates a global motion sequence from the local pose sequence. In the inference stage, the global motion is accumulated into a global pose, and finally, the SMPL reconstructs a human mesh sequence with the global pose defined in the world coordinate system.

3.2. SMPL Representation

SMPL [18] represents human pose and shape using the pose parameter $\theta \in \mathbb{R}^{72}$ and the shape parameter $\beta \in \mathbb{R}^{10}$. The pose parameter is parameterized by global 3D rotation and the relative 3D rotations of 23 joints in an axis-angle representation. The shape parameter is parameterized using the first 10 principal component coefficients of the human shape space. SMPL provides a differentiable function that generates the 3D human mesh $M(\theta, \beta) \in \mathbb{R}^{6890\times 3}$ from the pose parameter θ and the shape parameter β. Relative rotations of the 23 joints of the pose parameter that correspond to the local pose become the input to GMR. However, since GMR uses the local pose represented in an unit quaternion form as an input, the local pose represented in the unit quaternion form is first transformed to an axis-angle form, which is then used as an input of the SMPL model. Global rotation corresponds to the global pose's orientation, which is the output of GMR. The shape parameter in this study is obtained using the existing 3D human pose estimation method [11]. Unlike existing methods [5,6,11,19,20], the proposed method generates a global human mesh defined in the world coordinate system by adding the translation to the 3D human mesh M as follows:

$$M^g(\theta, \beta, T) = M(\theta, \beta) + T, \tag{1}$$

where $T \in \mathbb{R}^3$ denotes the global translation, which is one of the outputs of the proposed method.

3.3. Global Motion Regressor (GMR)

The proposed network estimates the global motion sequence, that is, the deviation of global poses between two adjacent frames from the local pose sequence $\mathbf{L}$. Various temporal neural architectures have been proposed to address these types of sequence data in recent years. We model GMR using bidirectional GRU [22] to encode long-term information effectively in this work. Figure 4 shows the architecture of the proposed GMR network. First, the local pose sequence $\mathbf{L} = \{L_i\}_{i=1}^T$ is fed into the temporal encoder that consists of bidirectional GRUs and a linear projection layer. Each bidirectional GRU forwards the local pose sequence to the GRU layer in forward and reverse directions and concatenates their results to generate hidden states $\mathbf{H} = \{H_i\}_{i=1}^T$, where $H_i \in \mathbb{R}^{4096}$. Second, the dimension of output hidden states $\mathbf{H}$ is reduced by the linear projection layer and the linear projection layer then generates the latent feature $\mathbf{F} = \{F_i\}_{i=1}^T$, where $F_i \in \mathbb{R}^{2048}$. Finally, the motion regression layer regresses the global motion sequence $\Delta\mathbf{G} = \{\Delta G_i\}_{i=1}^T$ from the latent feature $\mathbf{F}$. The global motion $\Delta G_i = (\Delta A_i, \Delta T_i)$ consists of orientation ΔA_i and translation ΔT_i motions between i-th and $(i+1)$-th frames. ΔA_i represented in the axis-angle form is transformed to a 3×3 rotation matrix ΔR_i through the Rodrigues' rotation formula [35]. Then, the global motion $\Delta G_i \in SE(3)$ can be written using $\Delta R_i \in SO(3)$ and $\Delta T_i \in \mathbb{R}^3$. Moreover, the global pose $G_i \in SE(3)$ can be represented using $R_i \in SO(3)$ and $T_i \in \mathbb{R}^3$. The following equations hold between the global pose G_i and the global motion ΔG_i:

$$G_{i+1} = G_i \Delta G_i = \begin{bmatrix} R_i & T_i \\ \mathbf{0}^T & 1 \end{bmatrix} \begin{bmatrix} \Delta R_i & \Delta T_i \\ \mathbf{0}^T & 1 \end{bmatrix}, \tag{2}$$

$$R_{i+1} = R_i \Delta R_i, \tag{3}$$

$$T_{i+1} = R_i \Delta T_i + T_i. \tag{4}$$

Finally, through the SMPL model [18], we reconstruct the global human mesh $M_i^g = M^g([A_i, L_i], \beta_i, T_i)$ defined in the world coordinate system from the obtained global poses G_i and input local poses L_i, where A_i is the axis-angle form of R_i and $[\cdot, \cdot]$ denotes the concatenation.

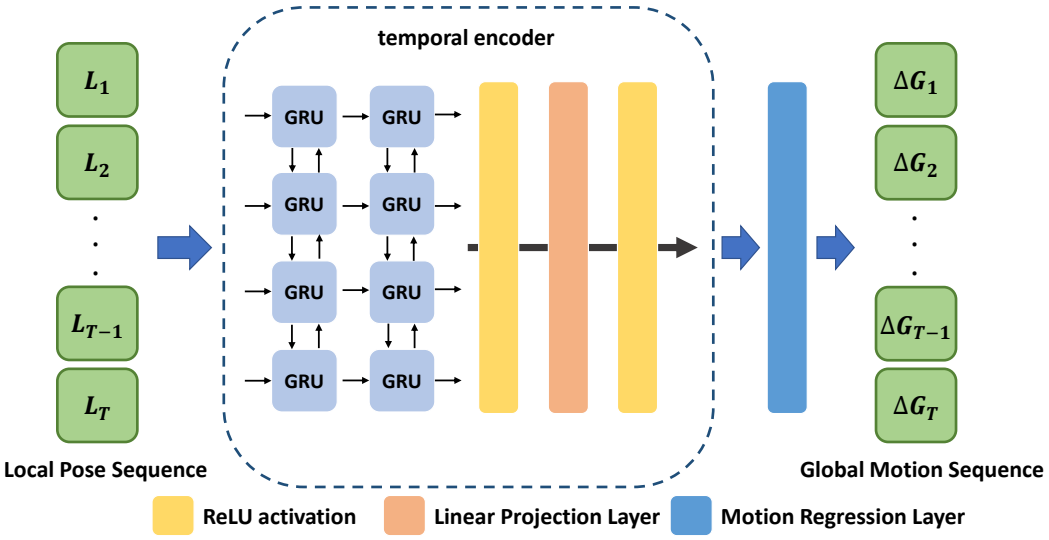

Figure 4. Architecture of **G**lobal **M**otion **R**egressor (GMR).

3.4. Loss Function

We train the proposed GMR using the following loss function:

$$\begin{aligned}\mathcal{L}_{total} &= w_{ori}\mathcal{L}_{ori} + w_{trans}\mathcal{L}_{trans} \\ &+ w_{vertex}\mathcal{L}_{vertex} + w_{smooth}\mathcal{L}_{smooth},\end{aligned} \tag{5}$$

where $\mathcal{L}_{ori}$, $\mathcal{L}_{trans}$, $\mathcal{L}_{vertex}$, and $\mathcal{L}_{smooth}$ are orientation, translation, vertex, and smoothness losses, respectively; and w_{ori}, w_{trans}, w_{vertex}, and w_{smooth} denote weights of losses and are set to 1, 1, 1, and 10^{-2}, respectively. We must carefully define the loss function $\mathcal{L}_{ori}$ to supervise the predicted orientation motion ΔR_i because the 3D rotation belongs to $SO(3)$, not the Euclidean space. Hartley et al. [36] described various distance measures that can be used for the elements of $SO(3)$. Taking them into account, we test the angular loss $\mathcal{L}_{angular}$, the chordal loss $\mathcal{L}_{chordal}$, and the axis-angle loss $\mathcal{L}_{axis\text{-}angle}$, which are based on the commonly used distance measures for $SO(3)$, defined as follows:

$$\mathcal{L}_{angular} = \sum_{i=1}^{T} \| \log(\Delta R_i \Delta R_i^{*T}) \|_2^2, \tag{6}$$

$$\mathcal{L}_{chordal} = \sum_{i=1}^{T} \| \Delta R_i - \Delta R_i^* \|_F^2, \tag{7}$$

$$\mathcal{L}_{axis\text{-}angle} = \sum_{i=1}^{T} \| \log(\Delta R_i) - \log(\Delta R_i^*) \|_2^2, \tag{8}$$

where $\Delta R_i \Delta R_i^{*T}$, ΔR_i, and ΔR_i^* are mapped to an axis-angle form through the logarithm map, and $*$ indicates the ground-truth. We also define the translation loss $\mathcal{L}_{trans}$ using the Euclidean distance between the predicted translation motion ΔT_i and its ground-truth as follows:

$$\mathcal{L}_{trans} = \sum_{i=1}^{T} \| \Delta T_i - \Delta T_i^* \|_2^2. \tag{9}$$

For vertex-wise loss on the reconstructed 3D mesh surface, we further define the vertex loss $\mathcal{L}_{vertex}$ on the basis of the L1 distance as follows:

$$\mathcal{L}_{vertex} = \sum_{i=1}^{T} \sum_{j=1}^{N} \| \Delta M_i^g[j] - \Delta M_i^{g*}[j] \|_1, \tag{10}$$

where $\Delta M_i^g[j]$ denotes the j-th row vector of matrix ΔM_i^g, that is, the coordinates of the j-th vertex, and $N = 6890$ is the total number of vertices. GMR predicts the global motion, which is the temporal deviation of global poses between two adjacent frames. Therefore, instead of directly supervising the global human mesh M_i^g, we apply the vertex loss to the global human mesh offset $\Delta M_i^g = M^g([\Delta A_i, L_i], \beta_i, \Delta T_i)$. Finally, we use the smoothness loss $\mathcal{L}_{smooth}$ to generate a smooth global motion:

$$\mathcal{L}_{smooth} = \sum_{i=1}^{T-1} \|\Delta R_i - \Delta R_{i+1}\|_F^2, \tag{11}$$

which is based on the Frobenius norm between orientation motions in adjacent frames and helps to reconstruct temporally coherent global orientations.

3.5. Flip Augmentation

We use the large-scale motion capture dataset AMASS [21] to train the proposed GMR. The AMASS dataset provides large amounts of sequence data from a wide range of human actions. However, its diversity is still limited compared with the variation of real human action. Therefore, we randomly flip sequences of the AMASS dataset in the temporally reverse direction and use them for learning. The used data augmentation process allows the network to utilize additional diverse training data. In this work, we call it flip augmentation, which uses both original and flipped datasets for training.

3.6. Inference

Given an input video of the frame length T, we estimate the local pose sequence $\mathbf{L}$ using the existing human pose estimation network [11]. GMR then estimates the global motion sequence $\Delta\mathbf{G}$ from $\mathbf{L}$. We assume that a person moves from the origin of the world coordinate system, and the orientation and the translation of the initial global pose G_1 are defined as an identity matrix and a 3D zero vector, respectively. Thus, the initial global human pose is denoted $G_1 = I_{4\times4}$. The global human pose sequence $\mathbf{G} = \{G_1, \ldots, G_T\}$ is subsequently calculated by repeatedly applying Equations (3) and (4) starting with the initial global human pose G_1. Finally, we place the global human pose sequence $\mathbf{G}$ and the local pose sequence $\mathbf{L}$ into SMPL [18] and obtain the global human mesh sequence $\mathbf{M}^g = \{M_1^g, \ldots, M_T^g\}$ defined in the world coordinate system using Equation (1).

4. Experimental Results

In this section, we present various experimental results to prove the effectiveness of the proposed method. The evaluation of the proposed GMR requires the local pose sequence for the input video, and in our experiment, it is obtained through VIBE, one of the existing SMPL-based human pose estimation methods. VIBE requires bounding box information for the target human subject. We assume that such bounding box information is given, and utilize the information provided by the datasets used in our experiments. This bounding box information can be obtained by various detectors [37,38] and trackers [39].

4.1. Implementation Details

We set the sequence length and frame rate of the input video to 64 and 10 fps, respectively, to train GMR. However, GMR can work for input sequences of arbitrary length. We use VIBE [11] in the test stage to obtain the local pose sequence. VIBE outputs SMPL pose parameters consisting of global orientations and local poses. However, we only use the local pose from the VIBE output, discard the global orientation, and reconstruct the new global orientation using the proposed method. This is because the global orientation generated by VIBE is defined in the camera coordinate system, so it fails to provide a 3D human pose in the world coordinate system. The bidirectional GRU of the temporal encoder consists of four layers with 2048 neurons, and the linear projection layer consists of one linear layer with 2048 neurons. The motion regression layer consists of one linear layer that

outputs the global motion. The weights of GMR are initialized using a uniform distribution $\mathcal{U}(-\sqrt{k}, \sqrt{k})$, where k represents the size of the hidden feature and the size of the input feature for the GRU and linear layers, respectively. We use the Adam optimizer [40] to optimize the loss function and set the learning rate to 5×10^{-5}. We set the mini-batch size to 32 and train the network using one Nvidia RTX3090 GPU. The number of epochs is set to 100, and GMR training takes about 5 h. Figure 5 shows the curves for loss and train/test errors in GMR training, which are the results for the best model found in ablation experiments in Section 4.4. PyTorch [41] is used to implement our code.

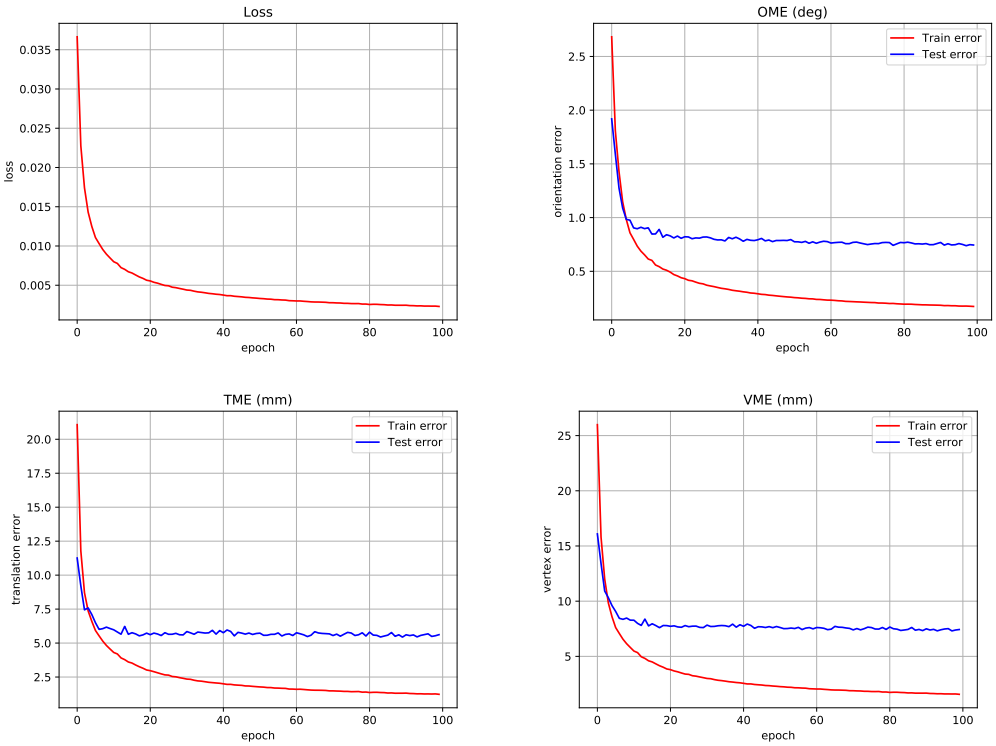

Figure 5. Curves for our loss and errors in the training process.

4.2. Datasets

We use the AMASS [21] dataset for training. The AMASS dataset consists of sequences of publicly available datasets, such as CMU MoCap [42] and TotalCapture [43], and provides SMPL parameters extracted using MoSh++. We sample each sequence of the AMASS dataset at a rate of 10 fps and use them for training. The AMASS dataset consists of 11,352 videos, and the total number of frames after sampling is about 145M.

We use three datasets for evaluation. The first dataset, Human3.6M [44], is widely used in 3D human pose estimation research. The Human3.6M dataset provides 3.6M video frames composed of images captured from fixed cameras. We use SMPL parameters extracted via MoSh [12] for quantitative evaluation, and S9 and S11 of seven subjects are used for evaluation. We use the Human3.6M for ablation experiments and utilize the ground-truth local pose sequence as the input to GMR in this case.

The second dataset, 3DPW [15], contains 60 sequences captured outdoors. The 3DPW dataset provides global human poses for evaluating the proposed method. However, the provided global poses are difficult to use for evaluation due to severe drift. For evaluation on the 3DPW dataset, we acquire camera poses from the 3DPW dataset using the existing structure-from-motion method, COLMAP [34], and use them to generate pseudo-ground-truth global human poses. The 3DPW dataset provides relatively accurate 3D human poses defined in the camera coordinate system. Therefore, we convert the 3D human pose defined

in the camera coordinate system into the world coordinate system using the camera pose obtained through COLMAP as follows:

$$\begin{bmatrix} R_w & T_w \\ \mathbf{0}^T & 1 \end{bmatrix} = \begin{bmatrix} R_{col} & T_{col} \\ \mathbf{0}^T & 1 \end{bmatrix} \begin{bmatrix} R_c & T_c \\ \mathbf{0}^T & 1 \end{bmatrix}, \tag{12}$$

$$R_w = R_{col} R_c, \tag{13}$$

$$T_w = R_{col} T_c + T_{col}, \tag{14}$$

where R_w and T_w denote the pseudo-ground-truth global human pose in the world coordinate system, R_{col} and T_{col} denote the camera pose obtained through COLMAP, and R_c and T_c denote the orientation and translation of the human subject defined in the camera coordinate system. The 3DPW dataset provides intrinsic camera parameters, which can be utilized for camera calibration. Since the 3DPW dataset contains dynamic objects, simply applying COLMAP often fails to obtain successful results. Therefore, we mask out dynamic objects using the existing segmentation method, Mask R-CNN [38], so that COLMAP extracts features only from static regions. After automatic reconstruction through COLMAP, we manually filter out sequences that fail to reconstruct successful results. Also, frames with severe drift in the reconstructed sequence are manually discarded. As a result, we obtain global human poses for 17 sequences and perform evaluations on these sequences. Table 1 shows the details of the processed 3DPW dataset. We divide the types of camera motion into "Small", "Linear", and "Panning". "Small" indicates a sequence with little camera motion. "Linear" denotes the linear camera motion. And, "Panning" means that the camera moves horizontally around a fixed position.

Table 1. Details of the processed 3DPW dataset.

Sequence Name	Frame Range	Camera Motion Type
courtyard_basketball_00	00000.jpg−00467.jpg	Small
courtyard_basketball_01	00000.jpg−00957.jpg	Small
courtyard_bodyScannerMotions_00	00000.jpg−01256.jpg	Small
courtyard_box_00	00000.jpg−01040.jpg	Small
courtyard_captureSelfies_00	00300.jpg−00696.jpg	Small
courtyard_golf_00	00000.jpg−00603.jpg	Small
courtyard_rangeOfMotions_00	00000.jpg−00600.jpg	Small
courtyard_rangeOfMotions_01	00000.jpg−00586.jpg	Small
downtown_arguing_00	00000.jpg−00897.jpg	Small
downtown_crossStreets_00	00000.jpg−00587.jpg	Panning
downtown_runForBus_00	00000.jpg−00207.jpg	Linear
downtown_sitOnStairs_00	00000.jpg−00477.jpg	Linear & Panning
downtown_walkBridge_01	00042.jpg−00234.jpg	Panning
downtown_walkDownhill_00	00132.jpg−00435.jpg	Panning
downtown_walkUphill_00	00000.jpg−00285.jpg	Panning
downtown_windowShopping_00	00048.jpg−00327.jpg	Panning
downtown_windowShopping_00	00972.jpg−01542.jpg	Linear

Although the 3DPW dataset contains various scenes, the camera motion is limited. Therefore, we additionally build an animated synthetic video dataset based on general 3D animation production methods and use them for evaluation. In the Blender tool (https://www.blender.org/, accessed on 1 December 2021), we import the CMU motion BVH data. We also import a 3D human model that can generate 3D human animation sequences from the Adobe Mixamo character repository (https://www.mixamo.com/, accessed on 1 December 2021). 3D animation sequences are created by the Blender tool. Finally, we include the camera motion in animation sequences to obtain synthetic videos with the camera motion. In addition to the camera motion in the 3DPW dataset, we adopt circular camera motion to construct the synthetic dataset. We observed that it is more challenging than "Linear" or "Panning" camera motions. We use CMU sequences of the

AMASS to create these synthetic videos. The CMU dataset of AMASS consists of 106 subjects. We use 50 sequences for 16 subjects to create synthetic videos, and the remaining sequences are included in the training set.

4.3. Evaluation Metrics

Our proposed method predicts global human motion to obtain the intrinsic 3D human pose decoupled from camera motion. To the best of our knowledge, there is no metric for quantitatively evaluating the estimated global motion by the proposed method. Therefore, we newly propose the following metrics for evaluating the proposed method. The first evaluation metric is the orientation motion error (OME) and is defined as follows:

$$E_{orien} = \frac{1}{T}\sum_{i=1}^{T}\|log(\Delta R_i^* \Delta R_i^T)\|_2, \tag{15}$$

where $\Delta R_i \in SO(3)$ satisfies $\Delta R_i^T \Delta R_i = I_{3\times3}$. If the network prediction is correct, $\Delta R_i^* \Delta R_i^T = I_{3\times3}$ should hold. We transform $\Delta R_i^* \Delta R_i^T$ to $\mathbb{R}^3$ through the logarithm map and apply L2 norm to its result to calculate the angular error. The second evaluation metric is the translation motion error (TME) which is defined as follows:

$$E_{trans} = \frac{1}{T}\sum_{i=1}^{T}\|\Delta T_i - \Delta T_i^*\|_2. \tag{16}$$

The translation motion error computes the Euclidean distance between the prediction and its ground-truth for the translation motion in $\mathbb{R}^3$. The last evaluation metric is the vertex motion error (VME) and is defined as follows:

$$E_{vertex} = \frac{1}{TN}\sum_{i=1}^{T}\sum_{j=1}^{N}\|\Delta M_i^g[j] - \Delta M_i^{g*}[j]\|_2. \tag{17}$$

Since the network predicts human motion, we define the distance between the prediction and its ground-truth for the global human mesh offset as the vertex motion error. The units of orientation, translation, and vertex motion errors are degree, mm, and mm, respectively. We quantitatively evaluate the proposed method using these three evaluation metrics.

4.4. Ablation Experiments

Analysis of GMR input and output representation. Table 2 presents the quantitative comparison of nine possible combinations of 3D rotation representations for the input local pose $\mathbf{L}$ and the output orientation motion $\Delta\mathbf{A}$. The number of layers and hidden units of GRU are set to 2 and 512, respectively, in all ablation experiments for simplicity. In this experiment, the network is trained using only the vertex loss $\mathcal{L}_{vertex}$. We conduct experiments using axis-angle, 6D [45], and unit-quaternion forms, which are widely used to represent the 3D rotation in existing human pose estimation methods. Using the 6D rotation form as the output of the network can achieve satisfactory performance due to its continuity in angular representation [45]. In our GMR, however, the orientation motion has a small magnitude and causes a relatively less continuity problem than other pose estimation cases. In our experiments, the quaternion/axis-angle combination outperforms other combinations, proving that the proposed method is relatively free from discontinuity problems.

Table 2. Ablation results for GMR input and output representations on Human3.6M. The row and the column correspond to the input local pose and the output orientation motion in GMR, respectively. Numbers denote the VME. The best results are shown in bold.

In/Out	Axis-Angle	6D	Quaternion
Axis-angle	10.48	10.83	11.15
6D	9.83	10.07	10.14
Quaternion	**9.46**	9.48	9.91

Analysis of orientation losses. We attempt to find the optimal orientation loss from three candidates in Section 3.4 to improve the GMR training. GMR is trained using the final loss function in Equation (5) for fair comparison. Table 3 shows the quantitative comparison results. We demonstrate that chordal loss $\mathcal{L}_{chordal}$ defined by the Frobenius norm of the 3×3 rotation matrix shows better performance than others. From these results, we observe that applying a loss function to the rotation matrix produces a better global motion in the proposed method. Similar to our observation, state-of-the-art human pose estimation methods [6,11] also incorporate the chordal loss. We use the chordal loss as the orientation loss according to the experimental results.

Table 3. Ablation results for orientation losses on Human3.6M. The best results are shown in bold.

Loss Type	OME	TME	VME
Axis-angle	1.05	6.98	9.51
Angular	1.06	7.03	9.46
Chordal	**1.01**	**6.82**	**9.28**

Analysis of loss components. The effect of each loss component is presented in Table 4. When the orientation loss is added to the vertex loss, the orientation motion estimation performance is improved as we expected. When the smoothness loss is added, the translation and the vertex motion errors are reduced, while the orientation motion error increases. The smoothness loss forces the model to generate a smooth orientation motion, but it also causes the orientation motion to be estimated in the wrong direction. Finally, V+O+S+T outperforms V+O+S for all evaluation metrics. Although V+O+S+T shows lower performance in the orientation motion error than V+O, the effect is trivial. Therefore, we use V+O+S+T as the final loss function.

Table 4. Comparison results for adding loss components on Human3.6M. V: Vertex loss, O: Orientation loss, S: Smoothness loss, T: Translation loss. The best results are shown in bold.

Losses	OME	TME	VME
V	1.02	7.01	9.46
V+O	**0.99**	7.05	9.34
V+O+S	1.03	6.88	9.31
V+O+S+T	1.01	**6.82**	**9.28**

Analysis of GRU structure. In Table 5, we present ablation results for the GRU structure. The deeper and wider the structure of the proposed network, the better its performance without overfitting. GRU with 4 layers and 2048 hidden units shows the best performance among the candidates, so we use it as the final model.

Table 5. Ablation results for GRU structure on Human3.6M. The best results are shown in bold.

Layers/Hidden Units	512			1024			2048		
	OME	TME	VME	OME	TME	VME	OME	TME	VME
2	1.01	6.82	9.28	0.93	6.33	8.56	0.85	5.92	7.92
3	0.92	6.28	8.58	0.86	5.91	7.96	0.79	5.56	7.47
4	0.85	5.70	7.74	0.80	5.48	7.37	**0.76**	**5.14**	**7.01**

Effect of flip augmentation. The results of quantitative analysis on the effect of flip augmentation are presented in Table 6. The flip augmentation can produce physically impossible motions that can harm the performance of the proposed method. However, according to Table 6, the flip augmentation enhances the performance of all quantitative evaluation metrics. These results show that flipping many sequences in the AMASS dataset is physically plausible and thus the use of flipped sequences helps the learning of GMR by increasing the diversity of training data. Even a small number of non-reversible actions can positively affect the performance by regularizing the model.

Table 6. Comparison result for flip augmentation on Human3.6M. The best results are shown in bold.

	OME	TME	VME
w/o flip augmentation	0.76	5.14	7.01
w/ flip augmentation	**0.70**	**4.78**	**6.47**

Analysis of sampling rate. This paragraph provides an analysis of the sampling rate. For the sampling rate experiments, the AMASS dataset is sampled at rates of 5 fps, 10 fps, and 15 fps. The sampled AMASS dataset is split into TotalCapture sequences for evaluation and the remaining sequences for training. Ground-truth local and global pose sequences are used for learning and evaluation of GMR. Figure 6 shows the results of the reconstructed global pose over time. The reconstructed global pose is derived from the accumulation and transformation of the global motion sequence, described in Section 3.6. A higher sampling rate results in a smaller vertex error. Sequences with a higher sampling rate provide more information per unit time, thus enabling more accurate global motion estimation. However, a higher sampling rate requires a longer sequence, thus increasing the amount of computation. This shows the tradeoff between estimation accuracy and computational complexity by the sampling rate. Note that the sequence data used in all experiments in this paper except for this paragraph is sampled at a rate of 10 fps.

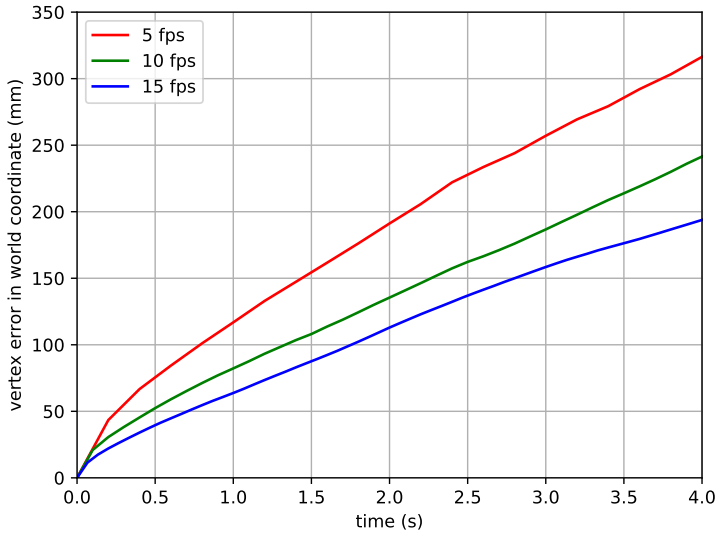

Figure 6. Vertex errors on training and test data acquired using different sampling rates. The numbers in the graph represent the vertex error over time.

Analysis of the sequential framework. The proposed framework can be considered a sequential combination of the existing 3D human pose estimation network VIBE [11] and the proposed GMR. To justify our sequential framework, we perform a quantitative comparison between the non-sequential and proposed frameworks. The non-sequential baseline can be simply constructed by reformulating VIBE to output both local pose and global motion. Unlike the proposed sequential framework, it can be learned end-to-end, which requires 2D videos and their corresponding ground-truth local poses and global motions. AMASS [21] does not provide videos, and end-to-end learning is not feasible with AMASS. For the end-to-end learning, we extracted pseudo-ground-truth human pose parameters from Human3.6M [44] and MPI-INF-3DHP [46] datasets by fitting the SMPL model to the ground-truth 3D joints in the world coordinate system using SMPLify-X [47]. Table 7 shows the quantitative comparison results on the 3DPW dataset. As a result, the proposed sequential framework outperforms the non-sequential baseline for all metrics. We believe that it is because local pose estimation and global motion estimation are not highly correlated so jointly training them makes training harder, resulting in lower performance.

Table 7. Quantitative comparison between the non sequential framework and the proposed framework on the 3DPW dataset. The best results are shown in bold.

Method	OME	TME	VME
Non-sequential	3.90	45.48	126.83
Ours	**3.67**	**38.55**	**120.37**

4.5. Comparison with Existing Method

Baseline. Compared with existing pose estimation methods, we present quantitative and qualitative evaluation results that show the advantages and limitations of our new framework. Specifically, we combine existing methods [11,16,17]. We first reconstruct a 3D human pose and shape sequence in the human-centered coordinate system from an input video using VIBE [11]. We then obtain a 2D human pose sequence by applying the 2D human pose tracking method STAF [17] to the input video. The global alignment module in DeepCap [16] computes the translation of the subject through the alignment process between 3D and 2D human poses from VIBE [11] and STAF [17], respectively. The overall procedure provides a 3D human mesh sequence in the camera coordinate system. We call this baseline VIBE-CAM and use the baseline for comparison.

Quantitative results. The quantitative comparison with VIBE-CAM for the 3DPW dataset is presented in Table 8. The resultant global motion of VIBE-CAM is very different from the ground-truth motion because it yields global human poses in the camera coordinate system. Therefore, the proposed GMR significantly outperforms VIBE-CAM in all metrics. The results are further improved when the ground-truth local pose is used as the input of GMR. This shows that there remains a lot of room for performance improvement through better local pose estimation. The quantitative comparison results for the synthetic dataset are presented in Table 9. The proposed framework outperforms VIBE-CAM, except for the orientation motion error, in the camera-motion-off case. However, when camera moves, VIBE-CAM reconstructs 3D human poses in the camera coordinate system, resulting in a global motion estimate significantly different from the ground-truth motion. Therefore, the proposed GMR significantly outperforms VIBE-CAM in all metrics. All quantitative results demonstrate that the proposed scheme provides an intrinsic global human motion independent of any camera motion embedded in the input video.

Table 8. Quantitative comparison between the proposed method and VIBE-CAM on the 3DPW dataset. Ours (GT input) indicates that the ground-truth local pose is used as the input of GMR. The best results are shown in bold.

Method	OME	TME	VME
VIBE-CAM	3.88	49.83	127.07
Ours	**3.67**	**38.55**	**120.37**
Ours(GT input)	1.60	27.55	29.39

Table 9. Quantitative comparison between the proposed method and VIBE-CAM on the synthetic dataset. Camera-motion-off indicates the synthetic video created without camera motion, while Camera-motion-on means the synthetic video with camera motion. The best results are shown in bold.

	Camera-Motion-Off			Camera-Motion-On		
Method	**OME**	**TME**	**VME**	**OME**	**TME**	**VME**
VIBE-CAM	**3.77**	58.01	117.28	4.66	81.70	132.63
Ours	3.80	**36.37**	**105.11**	**4.01**	**39.27**	**108.08**

Analysis on accumulated motion. Figure 7 shows the comparison of our reconstructed global pose with VIBE-CAM results over time. First, the vertex error (blue line) of VIBE-CAM increases significantly because VIBE-CAM estimates the pose in the camera coordinate system. In the case of proposed method, while the vertex error still increases over time, however, the graph (green line) shows a relatively lower error than VIBE-CAM. The error increase is unavoidable because the motion error is also accumulated in the global pose reconstruction step. We believe that this error accumulation problem can be alleviated through the use of constraints, such as loop closure in methods for simultaneous localization and mapping [48]. The proposed method shows a significantly lower vertex error graph (red line) when we use the ground-truth local pose in our GMR network. It demonstrates that the proposed GMR model is well-trained and effectively regresses the global motion sequence from the local pose sequence.

Results for Mannequin Challenge dataset. Figure 8 shows an interesting result on the Mannequin Challenge dataset [49]. The dataset consists of videos that include static people in a moving camera environment, as shown in Figure 8 (top row). Therefore, the 3D pose of a person reconstructed through the proposed method should not change temporally in the world coordinate system. In Figure 8 (middle row), VIBE-CAM predicts the human pose in the camera coordinate system and shows unexpected human movement with respect to the camera motion in the video. In Figure 8 (bottom row), however, the reconstructed human pose in our framework shows no movements as the original Mannequin Challenge scenario says. Therefore, the proposed framework effectively predicts the intrinsic human pose regardless of camera movement.

Qualitative results. Figure 9 shows the qualitative results of the proposed method and VIBE-CAM for the 3DPW dataset. In the input video, a person is walking down a hill, and the camera is rotating to follow the person. The human pose sequence reconstructed by VIBE-CAM is defined in the camera coordinate system. Therefore, the camera's rotation makes the result reconstructed by VIBE-CAM not represent the human walking motion. On the other hand, the proposed method estimates global human motion independent of camera motion. The human pose sequence reconstructed from the global motion sequence correctly represents the walking motion of the person regardless of camera rotation. These results show that the proposed method effectively reconstructs the intrinsic human pose independent of camera motion. Additional results on the 3DPW, synthetic, and Mannequin Challenge datasets are available in the Supplementary Material.

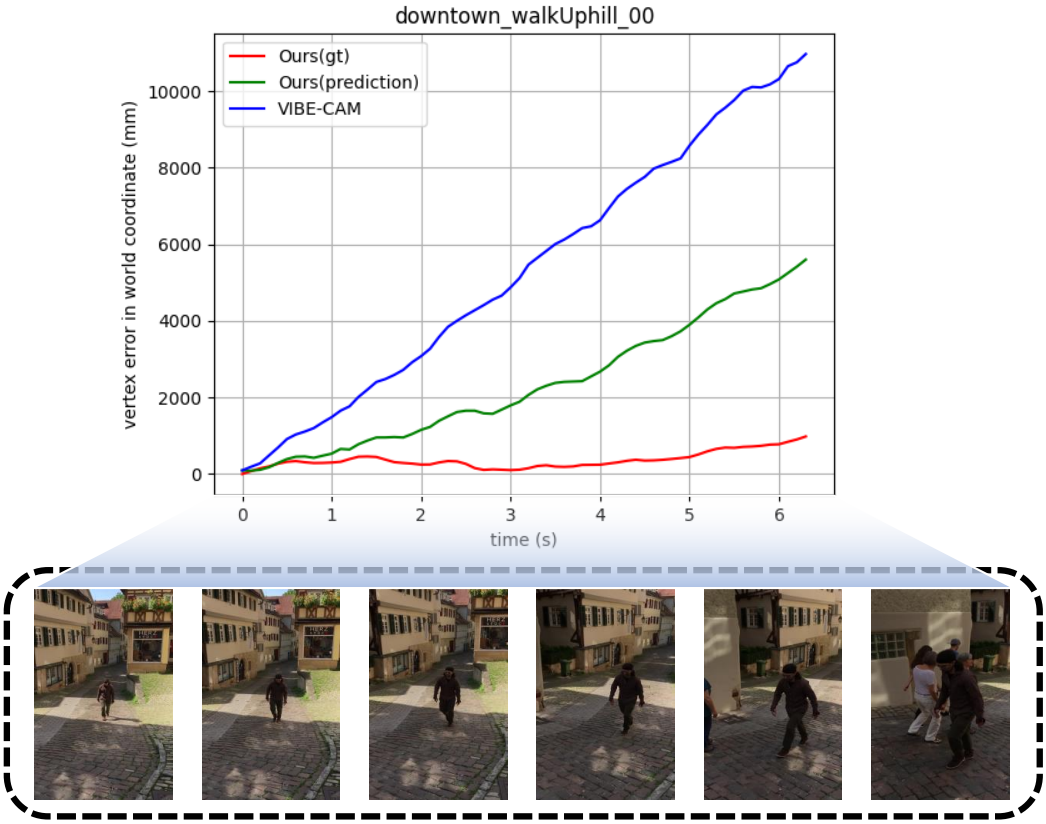

Figure 7. Vertex error over time. The numbers in the graph represent the vertex error between the predicted human mesh and its ground-truth in the world coordinate system.

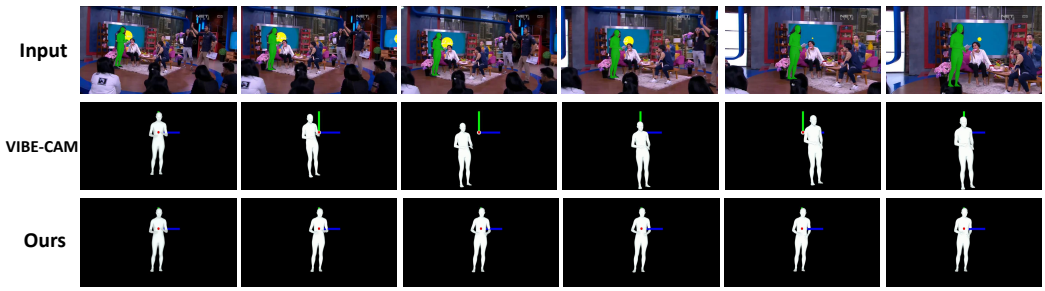

Figure 8. Qualitative comparison on the Mannequin Challenge dataset. The proposed method provides static human poses while VIBE-CAM reconstructs unexpected global human poses with respect to the camera movement in the input video. Note that the reference coordinate systems of VIBE-CAM is aligned with that of the proposed method for easy comparison.

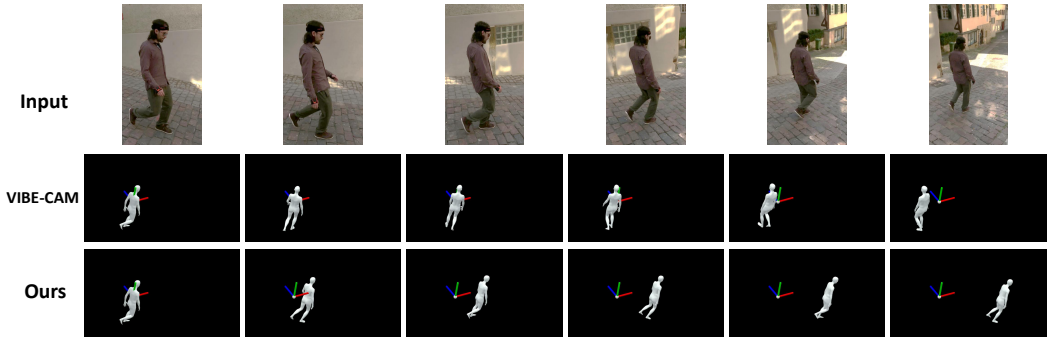

Figure 9. Qualitative results on the 3DPW dataset. The downtown_walkDownhill_00 sequence is used as input to VIBE-CAM and our method.

4.6. Limitation of Proposed Method

We argued in this paper that the proposed method can generate global human poses defined in the world coordinate system. However, strictly speaking, it is over-claiming. For example, if a person moves inside a train running at a constant speed, then the global human pose computed by the proposed method is defined based on the running train rather than the ground on which the world coordinate system is usually based. Therefore, in the proposed method, it can only be argued that the global human pose is computed in *a single coordinate system that is consistent with the overall motion of the entire sequence*. Nevertheless, 3D human poses reconstructed in this coordinate system are still independent of camera motion and can provide valuable information for various applications. We refer to this coordinate system as the world coordinate system in this study for convenience.

In order to overcome the above limitation, the camera pose obtained by calibrating the camera to the world should be utilized. For this, we have to rely on the existing SfM-based camera motion estimation, which is fragile for videos containing dynamic foreground objects, as mentioned in Section 2. We believe that human motion acquired through the method proposed in our study can provide constraints for robust camera motion estimation. Our future work is to combine the existing camera motion and 3D human pose estimation approach with the method proposed in this study to investigate this idea further.

5. Conclusions

A camera motion agnostic method for estimating 3D human poses in the world coordinate system is presented in this study. The majority of 3D human pose estimation methods estimate 3D poses defined in the camera coordinate system, so it is difficult to obtain a pure human pose from a video with camera motion. To address this issue, we propose a network that generates the global motion sequence invariant to the selection of the coordinate system from the local pose sequence. Our method can reconstruct the global human mesh defined in the world coordinate system in the inference stage. We generate a pseudo ground-truth global human pose dataset from 3DPW and construct a synthetic video dataset to evaluate the proposed method. We conduct thorough experiments for quantitative and qualitative evaluation, and prove the effectiveness of the proposed method.

Supplementary Materials: The supporting information is available online: https://zenodo.org/record/7053434. The supplementary video shows the qualitative results of the proposed method.

Author Contributions: Conceptualization, S.H.K., S.J., S.P. and J.Y.C.; methodology, S.H.K. and J.Y.C.; software, S.H.K.; validation, S.H.K.; formal analysis, S.H.K.; investigation, S.H.K.; resources, S.J. and S.P.; data curation, S.H.K.; writing—original draft preparation, S.H.K.; writing—review and editing, S.P. and J.Y.C.; visualization, S.H.K.; supervision, J.Y.C.; project administration, J.Y.C.; funding acquisition, J.Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by NCSOFT; in part by the Samsung Research Funding Center of Samsung Electronics under Project SRFCIT-1901-06; in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-00348, Development of A Cloud-based Video Surveillance System for Unmanned Store Environments using Integrated 2D/3D Video Analysis); and in part by the Excellent Researcher Support Project of Kwangwoon University in 2021.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Huang, Y.; Bogo, F.; Lassner, C.; Kanazawa, A.; Gehler, P.V.; Romero, J.; Akhter, I.; Black, M.J. Towards accurate marker-less human shape and pose estimation over time. In Proceedings of the International Conference on 3D Vision (3DV), Qingdao, China, 10–12 October 2017.

2. Pavlakos, G.; Zhou, X.; Derpanis, K.G.; Daniilidis, K. Coarse-to-fine volumetric prediction for single-image 3D human pose. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
3. Martinez, J.; Hossain, R.; Romero, J.; Little, J.J. A simple yet effective baseline for 3d human pose estimation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
4. Pavllo, D.; Feichtenhofer, C.; Grangier, D.; Auli, M. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
5. Kanazawa, A.; Black, M.J.; Jacobs, D.W.; Malik, J. End-to-end recovery of human shape and pose. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
6. Kolotouros, N.; Pavlakos, G.; Black, M.J.; Daniilidis, K. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.
7. Kocabas, M.; Karagoz, S.; Akbas, E. Self-supervised learning of 3d human pose using multi-view geometry. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
8. Guler, R.A.; Kokkinos, I. Holopose: Holistic 3D human reconstruction in-the-wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
9. Kolotouros, N.; Pavlakos, G.; Daniilidis, K. Convolutional mesh regression for single-image human shape reconstruction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
10. Moon, G.; Lee, K.M. I2L-MeshNet: Image-to-Lixel Prediction Network for Accurate 3D Human Pose and Mesh Estimation from a Single RGB Image. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020.
11. Kocabas, M.; Athanasiou, N.; Black, M.J. Vibe: Video inference for human body pose and shape estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
12. Loper, M.; Mahmood, N.; Black, M.J. MoSh: Motion and shape capture from sparse markers. *ACM TOG* **2014**, *33*, 220. [CrossRef]
13. Han, S.; Liu, B.; Wang, R.; Ye, Y.; Twigg, C.D.; Kin, K. Online optical marker-based hand tracking with deep labels. *ACM TOG* **2018**, *37*, 166. [CrossRef]
14. Haque, A.; Peng, B.; Luo, Z.; Alahi, A.; Yeung, S.; Fei-Fei, L. Towards viewpoint invariant 3D human pose estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 160–177.
15. von Marcard, T.; Henschel, R.; Black, M.J.; Rosenhahn, B.; Pons-Moll, G. Recovering accurate 3D human pose in the wild using imus and a moving camera. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
16. Habermann, M.; Xu, W.; Zollhofer, M.; Pons-Moll, G.; Theobalt, C. Deepcap: Monocular human performance capture using weak supervision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
17. Raaj, Y.; Idrees, H.; Hidalgo, G.; Sheikh, Y. Efficient online multi-person 2D pose tracking with recurrent spatio-temporal affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
18. Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; Black, M.J. SMPL: A skinned multi-person linear model. *ACM TOG* **2015**, *34*, 248. [CrossRef]
19. Pavlakos, G.; Zhu, L.; Zhou, X.; Daniilidis, K. 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 (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
20. Luo, Z.; Golestaneh, S.A.; Kitani, K.M. 3D Human motion estimation via motion compression and refinement. In Proceedings of the Asian Conference on Computer Vision (ACCV), Kyoto, Japan, 30 November–4 December 2020.
21. Mahmood, N.; Ghorbani, N.; Troje, N.F.; Pons-Moll, G.; Black, M.J. AMASS: Archive of motion capture as surface shapes. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.
22. Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014. [CrossRef]
23. Omran, M.; Lassner, C.; Pons-Moll, G.; Gehler, P.; Schiele, B. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In Proceedings of the International Conference on 3D Vision (3DV), Verona, Italy, 5–8 September 2018.
24. Kocabas, M.; Huang, C.H.P.; Hilliges, O.; Black, M.J. PARE: Part attention regressor for 3D human body estimation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021.
25. Bogo, F.; Kanazawa, A.; Lassner, C.; Gehler, P.; Romero, J.; Black, M.J. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016.
26. Varol, G.; Ceylan, D.; Russell, B.; Yang, J.; Yumer, E.; Laptev, I.; Schmid, C. Bodynet: Volumetric inference of 3D human body shapes. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.

27. Lin, K.; Wang, L.; Liu, Z. End-to-end human pose and mesh reconstruction with transformers. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021.
28. Lin, K.; Wang, L.; Liu, Z. Mesh graphormer. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021.
29. Kanazawa, A.; Zhang, J.Y.; Felsen, P.; Malik, J. Learning 3D human dynamics from video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
30. Arnab, A.; Doersch, C.; Zisserman, A. Exploiting temporal context for 3D human pose estimation in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
31. Sun, Y.; Ye, Y.; Liu, W.; Gao, W.; Fu, Y.; Mei, T. Human mesh recovery from monocular images via a skeleton-disentangled representation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.
32. Choi, H.; Moon, G.; Chang, J.Y.; Lee, K.M. Beyond static features for temporally consistent 3D human pose and shape from a video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021.
33. Wan, Z.; Li, Z.; Tian, M.; Liu, J.; Yi, S.; Li, H. Encoder-decoder with multi-level attention for 3D human shape and pose estimation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021.
34. Schonberger, J.L.; Frahm, J.M. Structure-from-motion revisited. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
35. Gallego, G.; Yezzi, A. A compact formula for the derivative of a 3-D rotation in exponential coordinates. *J. Math. Imaging Vis.* **2015**, *51*, 378–384. [CrossRef]
36. Hartley, R.; Trumpf, J.; Dai, Y.; Li, H. Rotation averaging. *Int. J. Comput. Vis.* **2013**, *103*, 267–305. [CrossRef]
37. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
38. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
39. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016.
40. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
41. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2019; pp. 8024–8035.
42. De la Torre, F.; Hodgins, J.; Bargteil, A.; Martin, X.; Macey, J.; Collado, A.; Beltran, P. *Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) Database*; Robotics Institute, Carnegie Mellon University: Pittsburgh, PA, USA, 2008.
43. Trumble, M.; Gilbert, A.; Malleson, C.; Hilton, A.; Collomosse, J.P. Total capture: 3D human pose estimation fusing video and inertial sensors. In Proceedings of the British Machine Vision Conference (BMVC), London, UK, 4–7 September 2017.
44. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 1325–1339. [CrossRef] [PubMed]
45. Zhou, Y.; Barnes, C.; Lu, J.; Yang, J.; Li, H. On the continuity of rotation representations in neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
46. Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; Theobalt, C. Monocular 3D human pose estimation in the wild using improved cnn supervision. In Proceedings of the International Conference on 3D Vision (3DV), Qingdao, China, 10–12 October 2017.
47. Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A.A.; Tzionas, D.; Black, M.J. Expressive body capture: 3D hands, face, and body from a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
48. Strasdat, H.; Davison, A.J.; Montiel, J.M.; Konolige, K. Double window optimisation for constant time visual SLAM. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011.
49. Li, Z.; Dekel, T.; Cole, F.; Tucker, R.; Snavely, N.; Liu, C.; Freeman, W.T. Learning the depths of moving people by watching frozen people. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.