Wearable Motion Capture: Reconstructing and Predicting 3D Human Poses From Wearable Sensors

Md Moniruzzaman, *Member, IEEE*, Zhaozheng Yin, *Senior Member, IEEE*, Md Sanzid Bin Hossain, *Graduate Student Member, IEEE*, Hwan Choi, *Member, IEEE*, and Zhishan Guo, *Senior Member, IEEE*

***Abstract*—Reconstructing and predicting 3D human walking poses in unconstrained measurement environments have the potential to use for health monitoring systems for people with movement disabilities by assessing progression after treatments and providing information for assistive device controls. The latest pose estimation algorithms utilize motion capture systems, which capture data from IMU sensors and third-person view cameras. However, third-person views are not always possible for outpatients alone. Thus, we propose the wearable motion capture problem of reconstructing and predicting 3D human poses from the wearable IMU sensors and wearable cameras, which aids clinicians' diagnoses on patients out of clinics. To solve this problem, we introduce a novel Attention-Oriented Recurrent Neural Network (AttRNet) that contains a sensor-wise attention-oriented recurrent encoder, a reconstruction module, and a dynamic temporal attention-oriented recurrent decoder, to reconstruct the 3D human pose over time and predict the 3D human poses at the following time steps. To evaluate our approach, we collected a new WearableMotionCapture dataset using wearable IMUs and wearable video cameras, along with the musculoskeletal joint angle ground truth. The proposed AttRNet shows high accuracy on the new lower-limb WearableMotionCapture dataset, and it also outperforms the state-of-the-art methods on two public full-body pose datasets: DIP-IMU and TotalCaputre.**

***Index Terms*—3D pose reconstruction, 3d pose prediction, recurrent neural network, wearable sensing.**

Manuscript received 24 January 2023; revised 12 July 2023; accepted 24 August 2023. Date of publication 4 September 2023; date of current version 7 November 2023. This work was supported by the NSF Projects under Grants CMMI-1954548, ECCS-2025929, SHF-2028481, and CMMI-2246673. *(Corresponding author: Zhaozheng Yin.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by The Institutional Review Board (IRB) of the University of Central Florida (UCF) approved the study's protocol (IRB ID: STUDY00002011). All participants provided informed written consent before participating in the experiment.

Md Moniruzzaman is with the Department of Computer Science, Stony Brook University, Stony Brook, NY 11794 USA (e-mail: mmoniruzzama@cs.stonybrook.edu).

Zhaozheng Yin is with the Department of Computer Science and Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY 11794 USA (e-mail: zyin@cs.stonybrook.edu).

Md Sanzid Bin Hossain is with the Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL 32816 USA (e-mail: sanzid@knights.ucf.edu).

Hwan Choi is with the Department of Mechanical and Aerospace Engineering, University of Central Florida, Orlando, FL 32816 USA (e-mail: hwan.choi@ucf.edu).

Zhishan Guo is with the Department of Computer Science, North Carolina State University, Raleigh, NC 27695 USA (e-mail: zguo32@ncsu.edu).

Digital Object Identifier 10.1109/JBHI.2023.3311448

I. Introduction

PEOPLE with movement disorders face multiple disadvantages while walking, such as increased strains on the lower back, increased metabolic cost, and gait asymmetry. Appropriately monitoring the progression of walking can mitigate these disadvantages and prevent secondary issues such as joint arthritis, risk of falls, and vascular diseases by having timely follow-up treatments from frequent assessment. Current monitoring procedures are only available at the clinical site. However, due to the absence of feasible technologies, it is extremely challenging to monitor the progress of the treatments after outpatient discharge. Thus, there is a need to assess the walking poses outside the clinic, which will not only significantly save medical expenditure by preventing unnecessary visits, but also enable patients to have the appropriate treatments without delay between regular visits.

Challenges: Most commonly, the motion capture systems [1] are used to achieve a highly accurate understanding of the human pose, but the numerous wearable markers and extra setup of motion capture cameras in the laboratory make this approach infeasible in an unconstrained daily environment.

Several works [2], [3], [4], [5], [6], [7] focused on the reconstruction of human poses from third-person view RGB or RGB-D cameras. However, these methods based on third-person views are not always possible for outpatients alone. Thus, there is a need to reconstruct and predict human poses from wearable sensors only, so discharged patients can walk freely in their daily lives.

Some works relied on numerous IMU sensors (e.g., 17 or more) to obtain an accurate human pose reconstruction [8], but wearing many sensors is very uncomfortable and impractical to use in daily living. Recently, several works [9], [10], [11], [12] used a reduced set of IMU sesnors for human pose reconstruction. However, motion capture from sparse inertial sensors is inherently ambiguous and challenging.

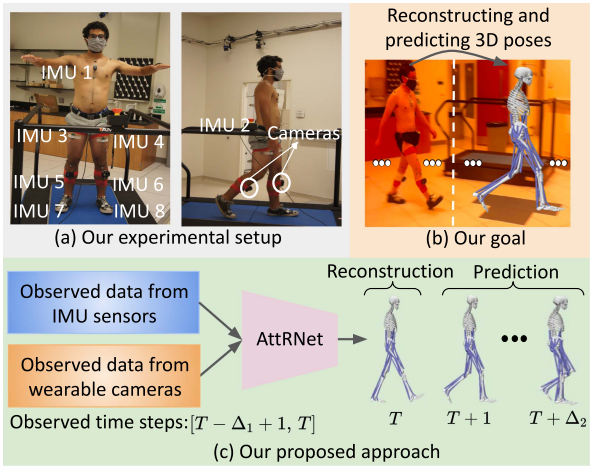

Fig. 1. Illustration of our proposed approach: (a) Our experimental setup; (b) Our goal; and (c) Input and output of our proposed Attention-Oriented Recurrent Neural Network (AttRNet). Note, in contrast to prior works that require vision data from third-person views for pose reconstruction which are not always possible for outpatients alone, we propose the on-body camera and IMU sensor solution to reconstruct and predict walking poses in a daily living environment.

Research question: The above challenges lead to a research question: *When discharged patients walk in their daily lives, how to design a feasible and effective approach to accurately sense their poses with a small set of wearable sensors, so clinicians can access their patients' walking functions outside clinics and researchers can design intelligent prosthetic devices to assist outpatients with real-time optimal control?*

Our contributions: As shown in Fig. 1, we propose to handle the ***wearable motion capture*** problem with two tasks: (1) ***reconstructing*** 3D human pose outdoors over time for clinical diagnosis; and (2) ***predicting*** 3D human poses at the following time steps for real-time assistive device controls.

Currently, there is no public dataset that contains both wearable IMU and wearable camera data, along with its ground truth of 3D walking poses. To develop the wearable motion capture algorithm, we collected a dataset under varying walking conditions (e.g., on the treadmill, on the ground, slope, stairs) from 10 subjects with different walking speeds. Though our collaborative research projects aim at people with lower limb amputations, our wearable motion capture can be applied to reconstruct and predict upper limb and full-body poses too. Hence, we also compare the performance of our wearable motion capture method on two related full-body pose datasets: DIP-IMU [13] (contains IMU-only data for 10 subjects) and TotalCapture [4] (contains IMU data and videos from third-person view cameras on 5 subjects).

Our main contributions have four folds:

- We propose to handle the wearable motion capture problem of reconstructing and predicting 3D human poses from the wearable IMU sensors and wearable cameras, which aids clinicians' diagnoses on people with movement disabilities. Prior works that require vision data from third-person views for pose reconstruction are not always possible for outpatients alone, thus we propose to reconstruct and predict walking poses via on-body camera and IMU sensors.
- We propose a novel Attention-Oriented Recurrent Neural Network (AttRNet) that contains a sensor-wise attention-oriented recurrent encoder, a reconstruction module, and a dynamic temporal attention-oriented recurrent decoder, to reconstruct the 3D human pose over time and predict the 3D human poses at the following few time steps.
- We introduce a new dataset containing data from wearable IMU and wearable camera sensors with the 3D human pose ground truth. To our best knowledge, no prior work was done on 3D human pose reconstruction or prediction from the fusion of both wearable IMU and wearable camera sensors. This dataset will be available on[1]:
- Our approach is able to generalize to both multi-modal and single-modal sensor input, and it can be applied to both lower limb pose and full-body pose analysis. Our proposed approach outperforms the state-of-the-art methods on two full-body pose datasets [4], [13].

II. Related Works

IMU-based human motion capture: The wearable IMUs (e.g., Xsens [15]) show remarkable stability and accuracy in capturing human motion [16], [17], [18], [19], [20]. Previously, Roetenberg et al. [8] introduced a motion tracking algorithm using IMU sesnors. Recently, Huang et al. [13] proposed a Recurrent Neural Network (RNN) based algorithm to reconstruct human poses from sparse inertial measurements, and also introduced an IMU-based human motion capture dataset. More recently, Nagaraj et al. [19] introduced an RNN-ensemble approach for human pose estimation from IMU sensors. Most of these IMU-sensor-based algorithms employ straightforward recurrent neural networks to reconstruct human poses. However, since different sensors on the human body have different capabilities to capture different joint movements at a specific time and the feature at the past time step is highly related to a certain future time step, straightforward recurrent neural networks might not be sufficient to compute the discriminative features and predict future poses from the observed input sequences. *Differently, we introduce a novel Attention-Oriented Recurrent Neural Network (AttRNet), which contains a sensor-wise attention-oriented recurrent encoder, a reconstruction module, and a dynamic temporal attention-oriented recurrent decoder, to reconstruct and predict 3D human poses by encoding the highly discriminative features from the observed input sequences.*

Vision-based human motion capture: The vision-based human motion capture can be divided into 2D and 3D pose estimation, and 3D future pose prediction. The state-of-the-art of 2D human pose estimation has achieved impressive progresses. Previously, the heatmap-based approaches [21], [22], [23], [24] and the regression-based algorithms [25], [26] were developed for 2D pose estimation. Newell et al. [14] introduced a deep hourglass model, which was later widely used as a backbone network [27], [28], [29]. Recently, the research community pays a significant amount of attention to develop 3D pose estimation. Most of the methods [30], [31], [32] on 3D pose estimation were originated

[1][Online]. Available: https://github.com/MoniruzzamanMd/Wearable-Motion-Capture

TABLE I
EXISTING MOTION CAPTURE VS. OUR WEARABLE MOTION CAPTURE

Method	Wearable IMUs	Wearable cameras	Third-person view cameras
Existing IMU-based motion capture (e.g., [13])	✓	✗	✗
Existing vision-based motion capture (e.g., [14])	✗	✗	✓
Existing hybrid (IMU+vision)-based motion capture (e.g., [4])	✓	✗	✓
Wearable motion capture (Ours)	✓	✓	✗

from 2D pose estimation task. Some works [33], [34], [35], [36], [37] tried to develop multi-view based methods to get more accurate 3D pose estimation.

Despite the success of 2D/3D pose estimation, 3D pose prediction is yet under-explored. Previously, some traditional methods such as the hidden Markov model [38] and the Gaussian model [39] were developed. Recently, some deep-learning-based algorithms introduced recurrent networks [40], [41], [42], [43], [44] and feed-forward networks [45], [46], [47] to predict future 3D poses. However, most of these vision-based pose analysis tasks rely on video data from third-person view cameras, which are not always possible for outpatients alone. Thus, there is a need to reconstruct and predict human poses from wearable sensors.

Most works with egocentric vision focused on objects and activities in front of cameras such as the detection of objects [48], [49], gaze [50], visible hands and arms [51]. *Differently, we are interested in the movement information of the wearable cameras for reconstructing and predicting 3D walking poses of the camera-carrying subject.*

Hybrid approaches for human motion capture: The hybrid approach mainly fuses the IMU and the vision modalities to learn richer features for human motion capture. Previously, Malleson et al. [3] proposed a real-time optimization approach to fuse multi-view data and IMU data to perform real-time motion capture. Recently, Trumble et al. [4] introduced an algorithm for fusing multi-view videos with IMU sensor data to estimate 3D human poses. Marcard et al. [5] proposed a graph-based optimization approach that jointly optimizes vision and IMU data on a SMPL model. More recently, DeepFuse [2] introduced an IMU-aware network for real-time 3D human pose estimation from multi-view images. Most of these hybrid approaches require vision data from third-person view cameras. However, the third-person views for pose reconstruction are not always possible for patients alone outdoors. *Differently, we introduce our AttRNet to reconstruct the 3D pose over time and predict the 3D poses at the following time steps from both wearable IMUs and wearable cameras.*

Note that different from the existing motion capture problem, in our wearable motion capture, both the IMU and camera sensors are worn on the human body, as shown in Fig. 1 and summarized in Table I.

III. PROPOSED APPROACH

A. Problem Statement

Suppose that the observed IMU data are $\mathbb{I}_{[T-\Delta_1+1,T]} = [\mathbf{I}_{T-\Delta_1+1}, \ldots, \mathbf{I}_T] \in \mathbb{R}^{\Delta_1 \times M \times D_{IMU}}$ and video data are $\mathbb{V}_{[T-\Delta_1+1,T]} = [\mathbf{V}_{T-\Delta_1+1}, \ldots, \mathbf{V}_T] \in \mathbb{R}^{\Delta_1 \times N \times D_{Video}}$, where M and N are the number of IMU and camera sensors, respectively, Δ_1 represents the temporal interval in which we can go back to the past from the current time step T, and D_{IMU} and D_{Video} are the feature dimensions extracted from each IMU and camera sensor at each time step, respectively. The goal of our proposed approach is to reconstruct the 3D human pose over time and predict the 3D human poses at the following time steps given the observed IMU and video data. Mathematically, we aim to obtain the following reconstruction and predictions functions:

Reconstruction: $(\mathbb{I}_{[T-\Delta_1+1,T]}, \mathbb{V}_{[T-\Delta_1+1,T]}) \rightarrow \hat{\mathbb{P}}_{[T-2,T]}$

Prediction: $(\mathbb{I}_{[T-\Delta_1+1,T]}, \mathbb{V}_{[T-\Delta_1+1,T]}) \rightarrow \hat{\mathbb{P}}_{[T+1,T+\Delta_2]}$,

where $\hat{\mathbb{P}}_{[T-2,T]} = [\hat{\mathbf{P}}_{T-2}, \hat{\mathbf{P}}_{T-1}, \hat{\mathbf{P}}_T] \in \mathbb{R}^{3 \times J \times D_{Joints}}$ denotes the reconstructed pose at the current time T ($\hat{\mathbf{P}}_T$) and the reconstructed poses of the past two time steps ($\hat{\mathbf{P}}_{T-2}$ and $\hat{\mathbf{P}}_{T-1}$) for pose dynamics calculation. $\hat{\mathbb{P}}_{[T+1,T+\Delta_2]} = [\hat{\mathbf{P}}_{T+1}, \ldots, \hat{\mathbf{P}}_{T+\Delta_2}] \in \mathbb{R}^{\Delta_2 \times J \times D_{Joints}}$ are the predicted future poses, J is the number of joints in the pose model, D_{Joints} is the coordinate dimension of each joint, and Δ_2 is the temporal interval in which we aim to predict the future poses.

B. Method Overview

We propose a novel Attention-Oriented Recurrent Neural Network (AttRNet) to jointly reconstruct the 3D pose over time and predict the 3D poses at the following time steps in an online setting from both wearable IMU sensors and wearable cameras, as shown in Fig. 2. In our AttRNet, we introduce an attention-oriented recurrent encoder-decoder and a reconstruction module. Our attention-oriented recurrent encoder performs sensor-wise attention at each time step and embeds features over different time steps of the observed input sequences. On the other hand, our attention-oriented recurrent decoder outputs a series of future poses by dynamically computing the relevant information from the encoded observed features using a dynamic temporal attention module. In our AttRNet, we also design a reconstruction module that reconstructs the current pose to support the initial pose prediction of the decoder. In the following sections, we discuss the IMU and video features, our proposed attention-oriented recurrent encoder, reconstruction module, and attention-oriented recurrent decoder in details.

C. IMU and Video Features

IMU features: An IMU returns 3-channel acceleration and 3-channel angular velocity from the accelerometer and the gyroscope, respectively. We calculate the orientation as a 3×3 rotation matrix from raw IMU data, which is then flattened

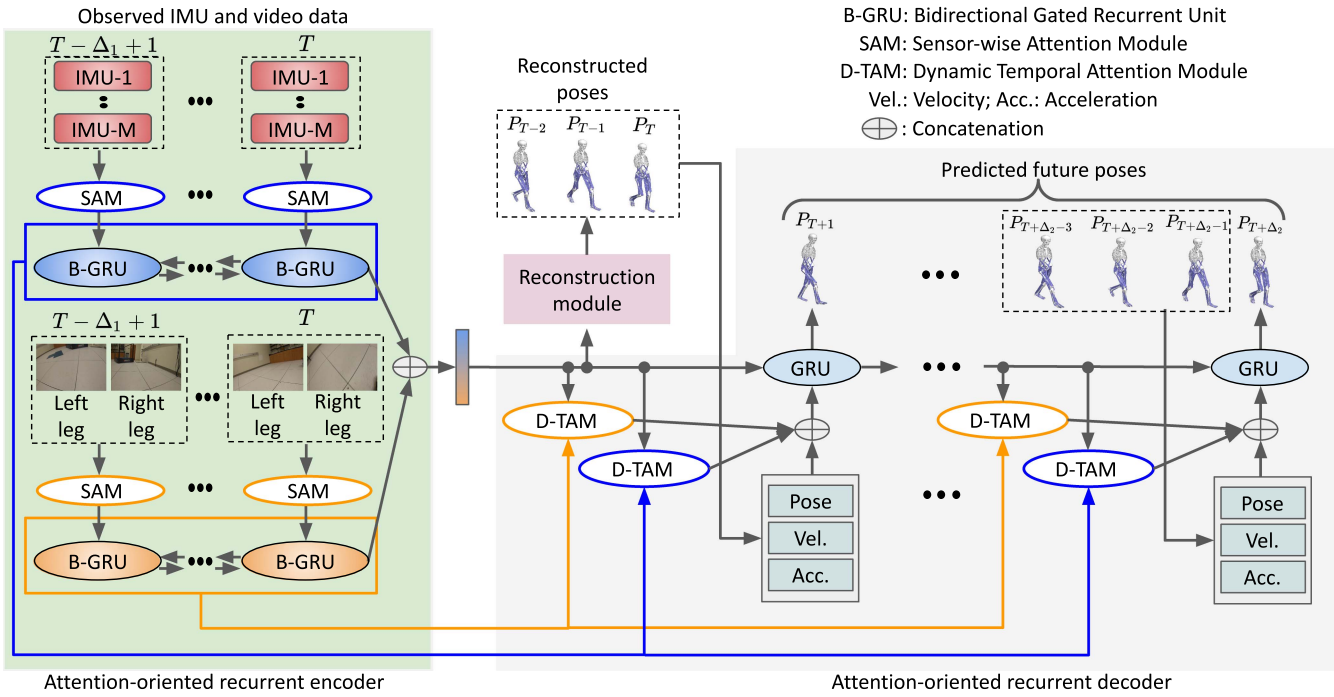

Fig. 2. Illustration of our proposed Attention-oriented Recurrent Neural Network (AttRNet). The AttRNet contains attention-oriented encoder-decoder and a reconstruction module. Our attention-oriented recurrent encoder contains Sensor-wise Attention Modules (SAM) to compute sensor-wise attention scores at each time step and bidirectional GRUs to embed attention-weighted features over different time steps of the observed sequences. The attention-oriented recurrent decoder is configured with Dynamic Temporal Attention Modules (D-TAM) that output a series of future poses by dynamically computing the relevant information from the encoded features. The reconstruction module reconstructs the three recent poses, which are also the initial inputs of the future pose prediction. The three recent poses are used to compute the pose velocity and acceleration.

to a 9-dimensional vector as the orientation feature. Finally, the 3-channel acceleration, 3-channel angular velocity, and 9-dimensional orientation features of the m-th IMU sensor at time t are concatenated into the feature representation $\mathbf{I}_t^m \in \mathbb{R}^{1\times D_{IMU}}$, where $D_{IMU} = 15$.

Video features: For the videos from wearable cameras on legs, we compute the histogram of optical flow features, where the optical flow vectors are quantified into different orientation bins and the magnitude of a bin is computed from the aggregation of the magnitudes of the flow vectors inside that bin. Formally, at time t, the histogram of optical flow feature of the n-th camera sensor is $\mathbf{V}_t^n \in \mathbb{R}^{1\times D_{Video}}$, where D_{Video} is the number of bins.

D. Attention-Oriented Recurrent Encoder

Our attention-oriented recurrent encoder contains sensor-wise attention modules to compute sensor-wise attention scores at each time step and recurrent encoders to embed attention-weighted features over different time steps of the observed input sequences.

Sensor-wise attention module: Since different sensors on the human body (e.g., Fig. 1(a)) have different capabilities to capture different joint movements at a specific time, we are motivated to compute sensor-wise attentions for both IMU and video features, i.e., we introduce Sensor-wise Attention Modules (SAM) to learn attention scores for different sensors and update the sensor-wise features with those attention scores. Formally, for a specific time t, the SAM loads the IMU features from M IMUs $\mathbf{I}_t = [\mathbf{I}_t^1; \ldots; \mathbf{I}_t^M] \in \mathbb{R}^{M\times D_{IMU}}$ to compute the attention score vector, $\mathbf{u}_t^I \in [0,1]^{M\times 1}$. The SAM consists of two fully-connected layers and a ReLU layer located between them. The second fully-connected layer outputs attention score vector, which are then passed through a sigmoid function that enforces the attention scores to be between 0 and 1. The attention-weighted IMU feature at time t, $\mathbf{I}_t^a$, is:

$$\mathbf{I}_t^a = \mathbf{I}_t \otimes \mathbf{u}_t^I \quad \text{where} \quad \mathbf{I}_t^a \in \mathbb{R}^{M\times D_{IMU}} \tag{1}$$

where $\otimes$ represents the element-wise multiplication, i.e., each column of the $\mathbf{I}_t$ matrix is multiplied with the column vector $\mathbf{u}_t^I$ element-wisely. Similarly, for the video features from N wearable cameras at time t, we compute the attention weighted video features features $\mathbf{V}_t^a$:

$$\mathbf{V}_t^a = \mathbf{V}_t \otimes \mathbf{u}_t^V \quad \text{where} \quad \mathbf{V}_t^a \in \mathbb{R}^{N\times D_{Video}} \tag{2}$$

where $\mathbf{V}_t$ is the video feature from N camera sensors and $\mathbf{u}_t^V$ is the corresponding attention score vector.

Recurrent encoders: The recurrent encoders separately encode the attention-weighted IMU and video features of different time steps and then fuse them into a single vector which will be used to reconstruct the pose over time and decode poses at the following time steps. Given the attention-weighted IMU features $\mathbf{I}^a = [\mathbf{I}_{T-\Delta_1+1}^a, \ldots, \mathbf{I}_T^a] \in \mathbb{R}^{\Delta_1\times M\times D_{IMU}}$, the recurrent encoder passes these features through a Bidirectional Gated

Recurrent Unit (B-GRU):

$$\mathbf{h}_t^{IMU} = \text{B-GRU}(\text{flatten}(\mathbf{I}_t^a), \mathbf{h}_{t-1}^{IMU}), t = T - \Delta_1 + 1, \ldots, T \tag{3}$$

where $\mathbf{h}_t^{IMU}$ and $\mathbf{h}_{t-1}^{IMU}$ are the hidden states for the IMU features at time t and $t-1$, respectively. The hidden state at T, $\mathbf{h}_T^{IMU}$ encodes the observed IMU features.

Similarly, the attention-weighted video features $\mathbf{V}^a = [\mathbf{V}_{T-\Delta_1+1}^a, \ldots, \mathbf{V}_T^a] \in \mathbb{R}^{\Delta_1 \times N \times D_{Video}}$ are encoded as:

$$\mathbf{h}_t^{Video} = \text{B-GRU}(\text{flatten}(\mathbf{V}_t^a), \mathbf{h}_{t-1}^{Video}), t = T - \Delta_1 + 1, \ldots, T \tag{4}$$

where $\mathbf{h}_t^{Video}$ and $\mathbf{h}_{t-1}^{Video}$ are the hidden states for the video features at time t and $t-1$, respectively. The hidden state at time T, $\mathbf{h}_T^{Video}$ encodes the observed video features. Finally, we concatenate the last hidden state's output of two modalities to get a single encoded vector, $\mathbf{h}_T^e = [\mathbf{h}_T^{IMU}, \mathbf{h}_T^{Video}]$, which will be fed into the reconstruction module and the decoder.

E. Reconstruction Module

We design a reconstruction module to reconstruct the current pose and the poses in the past two time steps to compute the pose dynamics. Formally, the reconstruction module loads the encoded feature vector $\mathbf{h}_T^e$ and predicts $\hat{\mathbb{P}}_{[T-2,T]} \in \mathbb{R}^{3 \times J \times D_{Joints}}$. The reconstruction module consists of two fully-connected layers and a ReLU layer between them. After the final fully connected layer, the output layer is reshaped to $\hat{\mathbb{P}}_{[T-2,T]}$, which represents the poses of the last three time steps of the observed sequences. Since this module reconstructs the current pose and the poses in the past two-time steps, we call this module as a reconstruction module.

Note, although both the sensor-wise attention module and the reconstruction module consist of two fully-connected layers and a ReLU layer between them, the detailed designs of these two modules are different since they do not share parameters and their goals are different.

F. Attention-Oriented Recurrent Decoder

Our attention-oriented recurrent decoder outputs a series of future poses by dynamically computing the relevant information from the encoded features of different time steps using the dynamic temporal attention module.

Recurrent decoder: The recurrent decoder aims to predict the future 3D poses $\hat{\mathbb{P}}_{[T+1,T+\Delta_2]}$. More specifically, from the time step $T+1$ to $T+\Delta_2$, our recurrent decoder predicts the future poses over different time steps, i.e., at time step $T+1$, our recurrent decoder predicts the pose $\hat{\mathbf{P}}_{T+1}$; at time step $T+2$, our recurrent decoder predicts the pose $\hat{\mathbf{P}}_{T+2}$; and so on. The recurrent decoder consists of GRU and the hidden state at each step is updated using the GRU update rules:

$$\mathbf{h}_\tau^d = \text{GRU}(\mathbf{z}_{\tau-1}, \mathbf{h}_{\tau-1}^d), \tau = T+1, \ldots, T+\Delta_2 \tag{5}$$

where the input $\mathbf{z}_{\tau-1}$ is computed from the output of the dynamic temporal attention modules, the previous predicted pose, and the pose dynamics of the previous predicted poses (the details are explained later shortly). The hidden states $\mathbf{h}_\tau^d$ and $\mathbf{h}_{\tau-1}^d$ are the current and previous hidden states of the decoder, respectively.

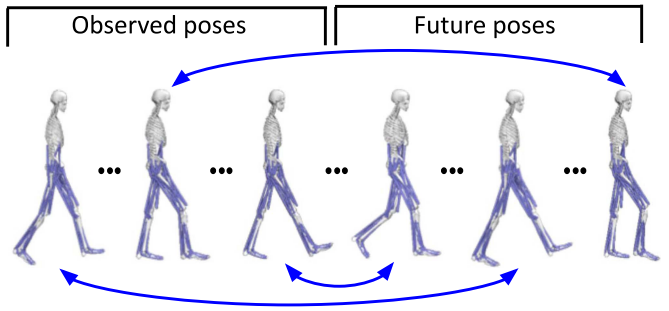

Fig. 3. Importance of Dynamic Temporal Attention Module (D-TAM). The observed poses at different time steps are highly related to different future poses. The blue arrows indicate that the corresponding poses are highly related.

During the first step of the decoder, the output of the recurrent encoder $\mathbf{h}_T^e$ is used as the previous hidden state.

Given the hidden state $\mathbf{h}_\tau^d$, the pose $\mathbf{P}_\tau$ is predicted using a fully-connected layer, as follows:

$$\hat{\mathbf{P}}_\tau = \mathbf{h}_\tau^d \mathbf{w}_\tau^P \tag{6}$$

where $\mathbf{w}_\tau^P$ is a trainable parameter. In the following, we discuss how we compute $\mathbf{z}_{\tau-1}$ from the D-TAM and pose dynamics in details.

Dynamic temporal attention module (D-TAM): The recurrent encoder encodes the observed input sequences and outputs a global representation as a single vector (e.g., $\mathbf{h}_T^{IMU}$ as the global feature representation of all observed IMU data). However, the global representation of the encoder might not be sufficient to predict future poses, due to the variations of human poses. Since humans walk with repetitive patterns, as shown in Fig. 3, the encoded feature at a past time step is highly related to a certain future time step. Thus, we configure the recurrent decoder with a Dynamic Temporal Attention Module (D-TAM) that dynamically computes the relevant information from the encoded features to predict future poses at different time steps. We employ D-TAM for both encoded IMU and video features, as described below.

Given the hidden state of the decoder $\mathbf{h}_{\tau-1}^d$ at time $\tau-1$ and the encoded IMU features of all the observed time steps $\mathbf{H}^{IMU} = [\mathbf{h}_{T-\Delta_1+1}^{IMU}, \ldots, \mathbf{h}_T^{IMU}]$, first we generate the query $(\mathbf{q}_{\tau-1}^{IMU})$, the keys $(\mathbf{K}_{\tau-1}^{IMU})$ and the values $(\mathbf{X}_{\tau-1}^{IMU})$ for the dynamic temporal attention on the encoded IMU features:

$$\mathbf{q}_{\tau-1}^{IMU} = \mathbf{h}_{\tau-1}^d \mathbf{W}_{\tau-1}^{\mathbf{q},IMU}, \text{ where } \mathbf{q}_{\tau-1}^{IMU} \in \mathbb{R}^{1 \times f} \tag{7}$$

$$\mathbf{K}_{\tau-1}^{IMU} = \mathbf{H}^{IMU} \mathbf{W}_{\tau-1}^{\mathbf{K},IMU} \text{ where } \mathbf{K}_{\tau-1}^{IMU} \in \mathbb{R}^{\Delta_1 \times f} \tag{8}$$

$$\mathbf{X}_{\tau-1}^{IMU} = \mathbf{H}^{IMU} \mathbf{W}_{\tau-1}^{\mathbf{X},IMU} \text{ where } \mathbf{X}_{\tau-1}^{IMU} \in \mathbb{R}^{\Delta_1 \times f} \tag{9}$$

where f is the feature dimension. $\mathbf{W}_{\tau-1}^{\mathbf{q},IMU}$, $\mathbf{W}_{\tau-1}^{\mathbf{K},IMU}$ and $\mathbf{W}_{\tau-1}^{\mathbf{X},IMU}$ are the trainable parameters. Then, we compute the output using a weighted sum of the values:

$$\mathcal{O}_{\tau-1}^{IMU} = \left(\mathbf{q}_{\tau-1}^{IMU} \left(\mathbf{K}_{\tau-1}^{IMU}\right)^\top\right) \left(\mathbf{X}_{\tau-1}^{IMU}\right) \tag{10}$$

TABLE II
EXISTING 3D HUMAN POSE DATASET VS OUR WEARABLEMOTIONCAPTURE DATASET

Dataset	#Wearable IMUs	#Wearable cameras	#Third-person view cameras	#Minutes	#Subjects	#Pose sequences
TotalCapture [4]	13	✗	8	50	5	46
DIP-IMU [13]	17	✗	✗	92	10	64
WearableMotionCapture (ours)	8	2	✗	327	10	140

where $\mathcal{O}_{\tau-1}^{IMU} \in \mathbb{R}^{1\times f}$ is the output of the D-TAM computed from the encoded IMU features and the current hidden state of the decoder.

Similarly, we apply D-TAM on video features and compute the output using a weighted sum of the values, as follows:

$$\mathcal{O}_{\tau-1}^{Video} = \left(\mathbf{q}_{\tau-1}^{Video}\left(\mathbf{K}_{\tau-1}^{Video}\right)^{\top}\right)\left(\mathbf{X}_{\tau-1}^{Video}\right) \tag{11}$$

where $\mathcal{O}_{\tau-1}^{Video} \in \mathbb{R}^{1\times f}$ is the output of the D-TAM computed from the encoded video features and the current hidden state of the decoder.

Pose dynamics: Since the first-order and second-order pose motions such as velocity and acceleration carry important motion dynamics, we use them in addition to the pose at time $\tau-1$ to predict the pose at time τ. For the pose dynamics at time $\tau-1$, we compute the velocity as $\mathcal{V}_{\tau-1} = (\hat{\mathbf{P}}_{\tau-2} - \hat{\mathbf{P}}_{\tau-1})$ and the acceleration as $\mathcal{A}_{\tau-1} = (\hat{\mathbf{P}}_{\tau-1} - 2\hat{\mathbf{P}}_{\tau-2} + \mathbf{P}_{\tau-3})$. Finally, we concatenate the output of each D-TAM, the pose, the velocity, and the acceleration to generate the input for the decoder (i.e., $\mathbf{z}_{\tau-1}$ in (5)), as follows:

$$\mathbf{z}_{\tau-1} = \left[\mathcal{O}_{\tau-1}^{IMU}, \mathcal{O}_{\tau-1}^{Video}, \hat{\mathbf{P}}_{\tau-1}, \mathcal{V}_{\tau-1}, \mathcal{A}_{\tau-1}\right] \tag{12}$$

G. Loss

The proposed AttRNet is trained with its hidden state output at each step supervised. The loss function in the proposed AttRNet is composed of two terms:

$$\begin{aligned}\mathcal{L}_{Total} &= \mathcal{L}_{Reco} + \mathcal{L}_{Pred} \\ &\triangleq ||\mathbb{P}_{[T-2,T]} - \hat{\mathbb{P}}_{[T-2,T]}||_2 \\ &\quad + ||\mathbb{P}_{[T+1,T+\Delta_2]} - \hat{\mathbb{P}}_{[T+1,T+\Delta_2]}||_2\end{aligned} \tag{13}$$

where $\mathcal{L}_{Reco}$ and $\mathcal{L}_{Pred}$ are the reconstruction and prediction loss, respectively. $\mathbb{P}_{[T-2,T]} = [\mathbf{P}_{T-2}, \mathbf{P}_{T-1}, \mathbf{P}_T] \in \mathbb{R}^{3\times J\times D_{Joints}}$ are the ground truth poses at time step T for the pose reconstruction, $\mathbb{P}_{[T+1,T+\Delta_2]} = [\mathbf{P}_{T+1}, \ldots, \mathbf{P}_{T+\Delta_2}] \in \mathbb{R}^{\Delta_2\times J\times D_{Joints}}$ are the ground truth poses from the time step $T+1$ to $T+\Delta_2$ for the future pose prediction, and $||\cdot||_2$ denotes the l_2 norm.

IV. EXPERIMENTS

A. Dataset

We conduct experiments on the wearable IMU+camera dataset on lower limbs which was collected by us: WearableMotionCapture, and two public full-body pose datasets: DIP-IMU [13] and TotalCapture [4]. The datasets are summarized in Table II.

WearableMotionCapture: We recorded data from 10 subjects (6 male, 4 female, age: 23.9±2.91 years, height: 1.65±0.06 m, weight: 63.41±6.81 kg) wearing 8 Avanti wireless IMU (Delsys, Boston, MA) sensors (2 attached on the upper body and 6 on the lower body), and 2 GoPro video cameras (attached in the middle of the shank contour of each leg), as shown in Fig. 1(a). All participants provided informed written consent before participating in the experiment. The Institutional Review Board (IRB) of the University of Central Florida (UCF) approved the study's protocol (IRB ID: STUDY00002011). The IMU data were recorded with a sampling frequency of 148 Hz, while the video data were collected at 30 fps. We down-sample the IMU data to have the same fps with the video data. Each subject was instructed to walk on several scenarios (Fig. 4). Each participant walked on both treadmill and ground with four different speeds (slow, normal, fast, and very fast). For each participant, we also recorded two trials for walking on stairs, and two trails for walking on slope. Furthermore, we collected two trials for each participant, where the participant walked on a round path and walked in a random path while avoiding two obstacles placed on the walking path. Overall, we collected 14 walking scenarios for each subject. Thirty-two reflective markers were placed on the participant based on a modified Helen-Hayes marker set [52], for ground-truth collection. Three-dimensional marker trajectories were captured by motion capture cameras. We obtain the ground truth of joint angles using OpenSim [53], an open source musculoskeletal analysis tool, on the marker tracking data from motion capture cameras. We use the musculoskeletal model to represent the 3D human walking pose (Fig. 1(c)). Musculoskeletal model is a skeleton model consisting of bones that are connected by joints. In total, we collected 140 pose sequences from 10 subjects, which are 327 minutes of IMU data along with 588 K video frames.

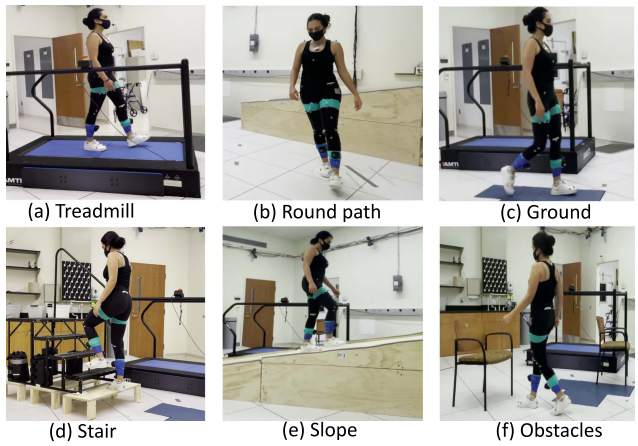

Fig. 4. Walking scenarios of our WearableMotionCapture dataset using wearable IMU and camera sensors: (a) Treadmill; (b) Round path; (c) Ground; (d) Stairs; (e) Slope; and (f) Obstacles.

DIP-IMU [13]: The DIP-IMU dataset consists of 10 subjects (9 male, 1 female), each performing motions in five different categories, including controlled motion of the experiments (arms, legs), locomotion, natural full-body activities (e.g., jumping jacks, boxing), and interaction tasks with everyday objects. These motions are recorded from 17 IMU sensors. We follow the train-test splits provided by the dataset to evaluate our method.

TotalCapture [4]: The TotalCapture dataset consists of 5 subjects (4 male and 1 female), each performing several activities such as walking, acting, range of motions and freestyle motions, which are recorded using 13 wearable IMU sesors and 8 third-person view RGB-cameras. Since we aim to reconstruct and predict human poses from wearable sensors, we only use the wearable IMU sensor data for comparisons in our experiment. We follow the train-test splits provided by the dataset to evaluate our method.

B. Implementation Details

Our recurrent encoders are constructed using the Bidirectional GRU (B-GRU). The hidden state's dimension of each B-GRU is set to 256. The decoder is configured with Dynamic Temporal Attention Modules (D-TAM) and GRUs with hidden state of dimension 512. The channel numbers between two fully-connected layers are set as 64 and 1024 for the sensor-wise attention and reconstruction modules, respectively. We use PyTorch to implement our proposed pose reconstruction and prediction model, and it takes about 60 minutes to train our network on the WearableMotionCapture dataset on a single Tesla V100 GPU.

C. Evaluation on WearableMotionCapture

Experimental setup: We consider two evaluation mechanisms:

- *Half and half evaluation:* We first randomly shuffle the samples of each subject, and then consider one half of the dataset to train the model and the other half is kept for the testing.
- *Leave-one-out evaluation:* We use the samples from 9 out of 10 subjects for training, and the samples of the left one subject are reserved for testing. We repeat this process 10 times for 10 different testing subjects and report the average.

Evaluation metric: We employ the Mean Absolute Error (MAE) as the evaluation metric for the pose reconstruction at the current time T as:

$$E_{reco} = \frac{1}{L_{test}} \frac{1}{D_{Joints}} \sum_{l=1}^{L_{test}} \sum_{j=1}^{D_{Joints}} |\mathbf{P}_{T,l}^{j} - \hat{\mathbf{P}}_{T,l}^{j}| \quad (14)$$

where E_{reco} is the MAE of the reconstructed pose, $\hat{\mathbf{P}}_{T,l}^{j}$ is the predicted angle of joint j at time T from the l-th testing sample and $\mathbf{P}_{T,l}^{j}$ is the ground truth. L_{test} is the number of testing samples generated from all pose sequences of the testing data by sliding temporal windows with stride 1. For example, for half and half evaluation (i.e., the test set has 294 K time steps from 70 pose sequences), we generate around $L_{test} = 294\ K - 70 \times (\Delta_1 + \Delta_2)$ testing samples (i.e., each sample is defined as observing Δ_1 time steps to predict the following Δ_2 time steps). Similarly, we employ the MAE as the evaluation metric for the future pose prediction of Δ_2 time steps as:

$$E_{pred} = \frac{1}{L_{test}} \frac{1}{\Delta_2} \frac{1}{D_{Joints}} \sum_{l=1}^{L_{test}} \sum_{\tau=1}^{\Delta_2} \sum_{j=1}^{D_{Joints}} |\mathbf{P}_{\tau,l}^{j} - \hat{\mathbf{P}}_{\tau,l}^{j}| \quad (15)$$

Reconstruction and prediction performances: As summarized on the first row of Table III, for the pose reconstruction, our AttRNet achieves MAE of 4.65 ° for half and half evaluation, and 7.49 ° for leave-one-out evaluation. On the other hand, for the future pose prediction, our AttRNet achieves MAE of 4.73 ° and 7.58 ° for half and half evaluation and leave-one-out evaluation, respectively, slightly higher than the reconstruction error.

From Table I, we observe that if a testing subject provides some calibration dataset, the pose sensing algorithm can perform better, as shown by the half-half experiment where half of the data from all subjects are used for training and the rest half is for testing. On the other hand, if a testing subject exhibits new pose patterns beyond the current training dataset, the pose sensing algorithm will have a slightly larger error, as shown by the leave-one-out experiment where a testing subject does not provide any calibration dataset, and the training is performed on other subjects. Since humans have variations on walking patterns, developing a one-size-fits-all algorithm to sense human poses could be challenging in real-world applications. To remedy this, in the future, we plan to investigate the customized sensing algorithms for individuals, which includes the investigation of transfer-learning algorithms to adapt a pose sensing algorithm trained on our dataset to individuals using a small calibration set of these individuals captured in labs.

Effect of our method on subjects with different body shapes: Although the kinematics of people tend to differ a lot, there is no direct relation with the body shape. For example, two people with the same height and weight will not walk with the same strategy. The main reason, which will cause the method to perform differently is the different walking patterns of different subjects. As a result, we see that our model performs differently when validating with the leave-one-out experiment compared to the half-and-half experiment. As the testing dataset is absent during the training, model does not have any prior knowledge of the walking pattern for the specific test subject. This causes performance degradation compared to half-and-half evaluation.

Ablation studies on different modules and modalities: To systematically evaluate our method and study the contribution of each algorithm component, we perform a number of ablation experiments: (i) our AttRNet without Sensor-wise Attention Module (SAM); (ii) our AttRNet without SAM or Dynamic Temporal Attention Module (D-TAM); (iii) our AttRNet without SAM, D-TAM or pose dynamics; (iv) and (v) our AttRNet with SAM, D-TAM and Pose Dynamics on IMU-only and Video-only, respectively. As shown in Table III, we can see that each algorithm component is contributing to our AttRNet to improve the performance for both half and half and leave-one-out

TABLE III
EVALUATION OF POSE RECONSTRUCTION AND FUTURE POSE PREDICTION, AND THE ABLATION STUDY ON DIFFERENT MODULES OF THE PROPOSED APPROACH ON OUR WEARABLEMOTIONCAPTURE DATASET FOR BOTH HALF AND HALF, AND LEAVE-ONE-OUT (LOO) EVALUATION

Method	Reconstruction		Prediction	
	Half-Half	LOO	Half-Half	LOO
AttRNet (IMU+Video)	4.65	7.49	4.73	7.58
(i) w/o SAM	5.14	7.87	5.24	7.95
(ii) w/o SAM/D-TAM	5.63	8.42	5.70	8.51
(iii) w/o SAM/D-TAM/Pose dynamics	5.96	8.93	6.05	9.03
(iv) AttRNet (IMU)	6.99	10.78	7.09	10.86
(v) AttRNet (Video)	7.65	11.56	7.74	11.65

The prediction models are evaluated in the online mode with 50 observed time steps to predict the next 10 time steps. Numbers represent the mean absolute errors of joint angles over 360 degrees.

TABLE IV
ABLATION STUDY OF POSE RECONSTRUCTION AND FUTURE POSE PREDICTION ON DIFFERENT WALKING SCENARIOS

scenarios	Reconstruction	Prediction	scenarios	Reconstruction	Prediction
Treadmill (slow)	1.78	1.81	Ground (fast)	5.49	5.58
Treadmill (normal)	1.79	1.83	Ground (very fast)	5.60	5.71
Treadmill (fast)	1.90	1.95	Stair (trial-1)	5.67	5.77
Treadmill (very fast)	2.52	2.56	Stair (trial-2)	5.74	5.85
Round	4.61	4.69	Slope (trial-1)	5.01	5.09
Ground (slow)	5.03	5.11	Slope (trial-2)	5.43	5.52
Ground (normal)	5.33	5.43	Obstacles	9.16	9.28

The prediction models are evaluated in the online mode with 50 observed time steps to predict the next 10 time steps. Numbers represent the mean absolute errors of joint angles over 360 degrees for half and half evaluation.

TABLE V
PER JOINT RECONSTRUCTION AND PREDICTION PERFORMANCE EVALUATION

Joints	Reconstruction	Prediction	Joints	Reconstruction	Prediction
Pelvis-tilt	3.39	3.45	Hip-rotation-right	7.53	7.66
Pelvis-list	2.98	3.03	Knee-angle-left	4.65	4.74
Pelvis-rotation	9.55	9.68	Knee-angle-right	4.69	4.78
Hip-flexion-left	4.03	4.10	Ankle-angle-left	3.43	3.49
Hip-flexion-right	4.07	4.15	Ankle-angle-right	3.42	3.47
Hip-adduction-left	3.19	3.24	Lumber-extension	3.63	3.69
Hip-adduction-right	3.23	3.30	Lumber-bending	3.76	3.83
Hip-rotation-left	7.72	7.86	Lumber-rotation	5.17	5.26

The prediction models are evaluated in the online mode with 50 observed time steps to predict the next 10 time steps. Numbers represent the mean absolute errors of joint angles over 360 degrees for half and half evaluation.

evaluations. Our AttRNet achieves the best performance from the fusion of IMU and video data, compared to the single-sensor approaches.

Ablation studies on different walking scenarios: The ablation studies on different walking scenarios are shown in Table IV. Since the walking patterns on the treadmill are repetitive, the pose reconstruction and prediction for walking on the treadmill are relatively easier and the performances are better compared to other scenarios. The walking motions are continuously changed when the subjects try to avoid obstacles on the ground, making the pose reconstruction and prediction more difficult. For example, if the obstacle is close, people reduce the step length and the walking path is going to be more acute. On the other hand, if the obstacle is far and enough time to deviate the obstacle, the path is going to be greater and step lengths may be similar before initiating the dodging the obstacle. However, even though the walking patterns are continuously changed in most of the walking scenarios, our AttRNet still reconstructs and predicts the poses well. Overall scenarios in Table IV, our maximal relative reconstruction and prediction errors are $9.16/360 = 2.54\%$ and $9.28/360 = 2.58\%$, respectively.

Per joint evaluation: The pose reconstruction and prediction errors regarding different joints are as shown in Table V. The reconstruction and prediction performances for the joint 'Pelvis list' are the best compared to other joints, while we get the lowest performances for the joint 'Pelvis rotation'. In our dataset, the subjects continuously try to rotate or change the walking direction. Therefore, reconstructing and predicting the joint angles related to rotation (e.g., 'Pelvis rotation', 'Hip rotation left', 'Hip rotation right' and 'Lumber rotation') become more difficult.

Parameter analysis: We perform experiments on different temporal intervals (Δ_1 and Δ_2) of the observed and future time steps. In the left side of Table VI, we show the prediction performance of our method for different temporal intervals Δ_1 of the observed sequences. The prediction error decreases with

TABLE VI
PARAMETER ANALYSIS ON Δ_1 AND Δ_2. (Δ_1, Δ_2) MEAN THAT THE PREDICTION MODELS ARE EVALUATED IN THE ONLINE MODE USING Δ_1 OBSERVED TIME STEPS TO PREDICT Δ_2 FUTURE TIME STEPS

(Δ_1, Δ_2)	Error (deg)	(Δ_1, Δ_2)	Error (deg)
(10, 10)	5.60	(50, 5)	4.64
(20, 10)	5.16	(50, 10)	4.73
(30, 10)	4.89	(50, 20)	4.97
(50, 10)	4.73	(50, 30)	5.21
(60, 10)	4.73	(50, 50)	5.60

Numbers represent the mean absolute errors of joint angles over 360 degrees for half and half evaluation.

TABLE VII
COMPARING OUR ATTRNET WITH OTHER BASELINE MODELS ON THE WEARABLEMOTIONCAPTURE DATASET

Model	Reconstruction	Prediction
RNN	7.84	7.92
B-RNN	6.73	6.82
LSTM	7.97	8.04
B-LSTM	6.88	6.98
GRU	7.01	7.12
B-GRU	5.96	6.05
AttRNet	4.65	4.73

Numbers represent the mean absolute errors of joint angles over 360 degrees for half and half evaluation.

the increased number of observed sequences and it saturates at $\Delta_1 = 50$ (we believe the reason is because it covers a few full cycles of human walking gaits in the recent past, which are sufficient for the prediction at the following time steps). On the other hand, the right side of Table VI shows the performance for different temporal intervals Δ_2 in the future, where the prediction error increases for the far future pose prediction.

Comparison with baseline models: We compare our AttRNet with some basline models, as shown in Table VII. Since we introduce our AttRNet based on recurrent networks to reconstruct the pose over time and predict the poses at the following time steps, we compare our AttRNet with some recurrent network-based baseline models such as Recurrent Neural Network (RNN), Bidirectional RNN (B-RNN), Long Short-Term Memory (LSTM), Bidirectional LSTM (B-LSTM), Gated Recurrent Unit (GRU), and Bidirectional GRU (B-GRU). Our AttRNet achieves superior performance compared to these recurrent network-based baseline models. The performance improvement from our AttRNet compared to the baseline models validates that our sensor-wise attention-oriented recurrent encoder can effectively encode the most highly discriminative features from the observed input sequences, and our dynamic temporal attention-oriented recurrent decoder can dynamically compute the most relevant information from the encoded observed features.

D. Comparison With the State-of-the-Art

DIP-IMU evaluation: The DIP-IMU is designed to predict 3D full-body human poses from wearable IMU sensors. Following the literature [13], [54], we adopt two different modes: (1) the offline mode where the full sequence is available; and (2) the online mode where our AttRNet observes past 20 time steps, and predicts 1 current time step and 5 future time steps in a sliding window manner. Table VIII shows the comparison results

TABLE VIII
3D HUMAN POSE PREDICTION PERFORMANCE COMPARISON WITH OTHER STATE-OF-THE-ART METHODS ON THE DIP-IMU DATASET FOR BOTH THE OFFLINE AND ONLINE MODES

Method	Offline		Online	
	Error (cm)	Error (deg)	Error (cm)	Error (deg)
SOP [55]	8.17	9.83	-	-
SIP [55]	6.66	8.77	-	-
DIP [13]	6.98	14.41	7.33	15.16
TransPose [54]	4.90	7.62	5.95	8.85
AttRNet (Ours)	**3.45**	**6.98**	**4.84**	**8.12**

Errors are reported as both positional errors in centimeters and joint angle errors in degrees.

TABLE IX
COMPARISON RESULTS REGARDING MEAN JOINT POSITIONAL ERROR ON TOTALCAPTURE DATASET. TSP: TEMPORAL SEQUENCE PREDICTION

Methods	Error (mm)
TotalCapture-IMU [4]	107.9
TotalCapture-IMU + TSP [4]	91.0
AttRNet (IMU)	**87.2**

of our AttRNet with other methods on DIP-IMU dataset [13] for 3D human pose prediction, where the results are compared for both the offline and online modes. Over all scenarios, our method achieves superior performance and establishes the new state-of-the-art results on DIP-IMU dataset for 3D human pose prediction.

TotalCapture evaluation: The TotalCapture dataset is designed to reconstruct full-body poses from the wearable IMUs and multiple third-person view video cameras. Table IX shows AttRNet outperforms the current best results on the IMU dataset of TotalCapture [4]. Note, for fair comparison, we only compare on wearable sensors, and TotalCapture only evaluates on positional errors. We follow the train-test splits provided by the corresponding dataset to evaluate our method.

E. Qualitative Analysis

We present some qualitative results on the test samples of the WearableMotionCapture dataset in Fig. 5, where our AttRNet captures data from wearable IMUs and wearable cameras, and reconstructs 3D walking poses. It can be seen that the proposed method can successfully reconstruct different walking poses on different scenarios (e.g., treadmill, stairs, slope, and ground with obstacles). The related video demos can be accessed at.[2] The joint angle error of avoiding obstacles is the highest in Table IV, but the visual evaluation in Fig. 5(b) is not very obvious in individual time steps.

Though our project focuses on the lower limb which has the potential to apply to people with movement disabilities such as stroke survivors, lower limb amputees, and children with cerebral palsy, the proposed method can also be generalized to full-body pose reconstruction using wearable sensors. We present some qualitative results on the test samples of walking, acting, freestyle and fighting pose sequences from the TotalCapture dataset, as shown in Fig. 6, where our AttRNet captures data from wearable IMU sensors and reconstructs

[2][Online]. Available: https://github.com/MoniruzzamanMd/Wearable-Motion-Capture

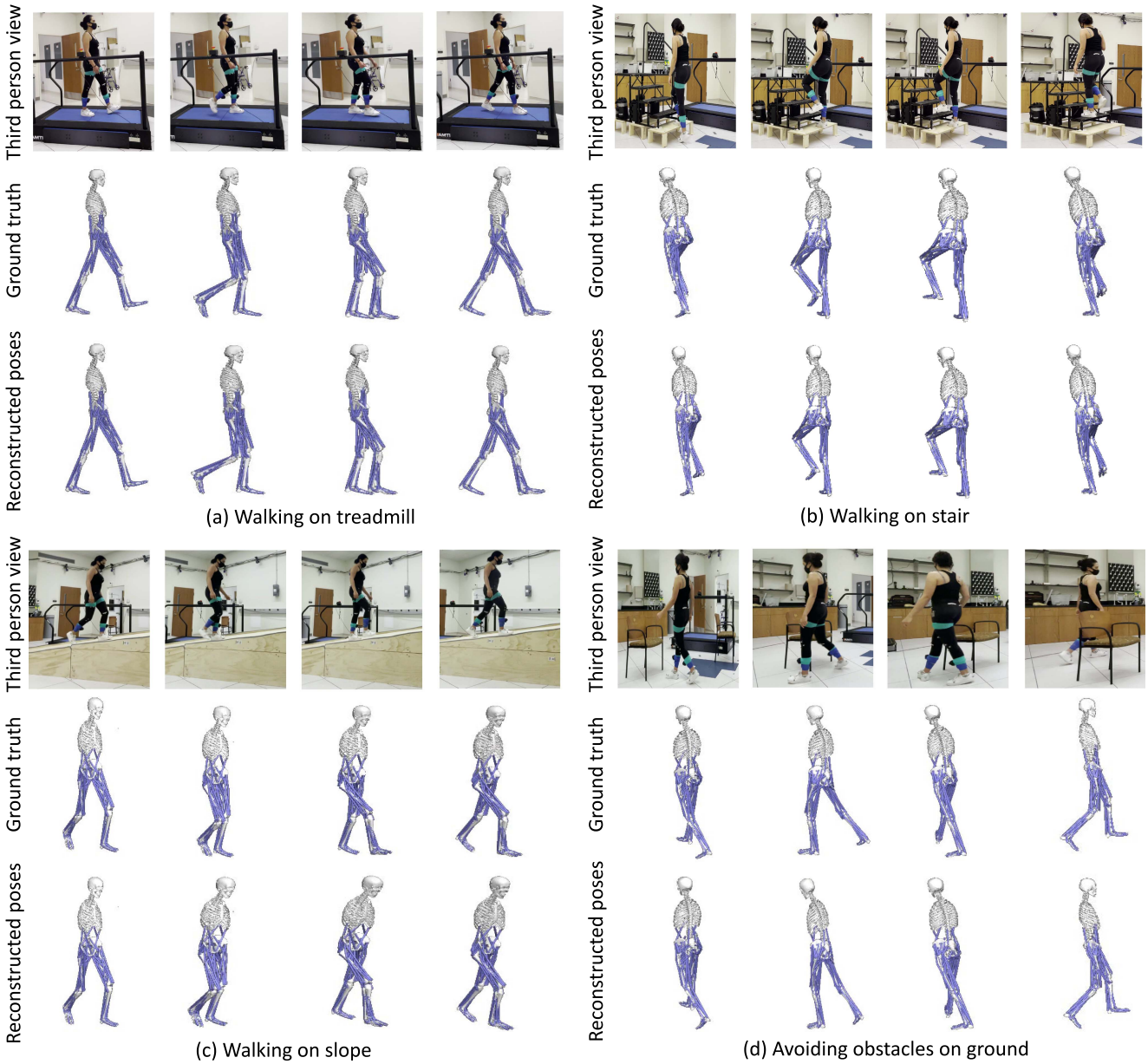

Fig. 5. Visualization of our human pose reconstruction over time on test samples of WearableMotionCapture dataset. (a) The subject walks on a treadmill; (b) The subject walks on stairs; (c) The subject walks on slope; and (d) The subject walks on ground and avoids the obstacles. Our AttRNet well reconstructs the 3D human poses based on the sensed data from wearable IMUs and wearable cameras.

full-body 3D pose sequences. It can be seen that the proposed method also shows good performance on the fully-body pose reconstruction.

V. Discussions and Future Works

In our assistive walking project, we aim to create a proactive prosthetic device that can positively affect the lives of the 1.6 million people with amputation. However, the development of such a device requires an algorithm to reconstruct and predict walking poses in an uncontrolled daily-living environment. Prior works that require vision data from third-person views for pose reconstruction are not always possible for amputees alone outdoors, thus we propose the wearable motion capture problem of reconstructing and predicting 3D human poses from the wearable IMU sensors and wearable cameras, which aids the prosthetic device control and clinicians' diagnoses on amputees out of clinics. For this challenging problem, we collected a new WearableMotionCapture dataset and proposed a novel Attention-Oriented Recurrent Neural Network (AttRNet) to reconstruct the 3D human pose over time and predict the 3D human poses at the following time steps.

Although our AttRNet achieves the promising performances from the fusion of IMU and video data, wearing cameras might raise privacy concerns from amputees. However, the usage of IMU sensors does not have any camera-related privacy

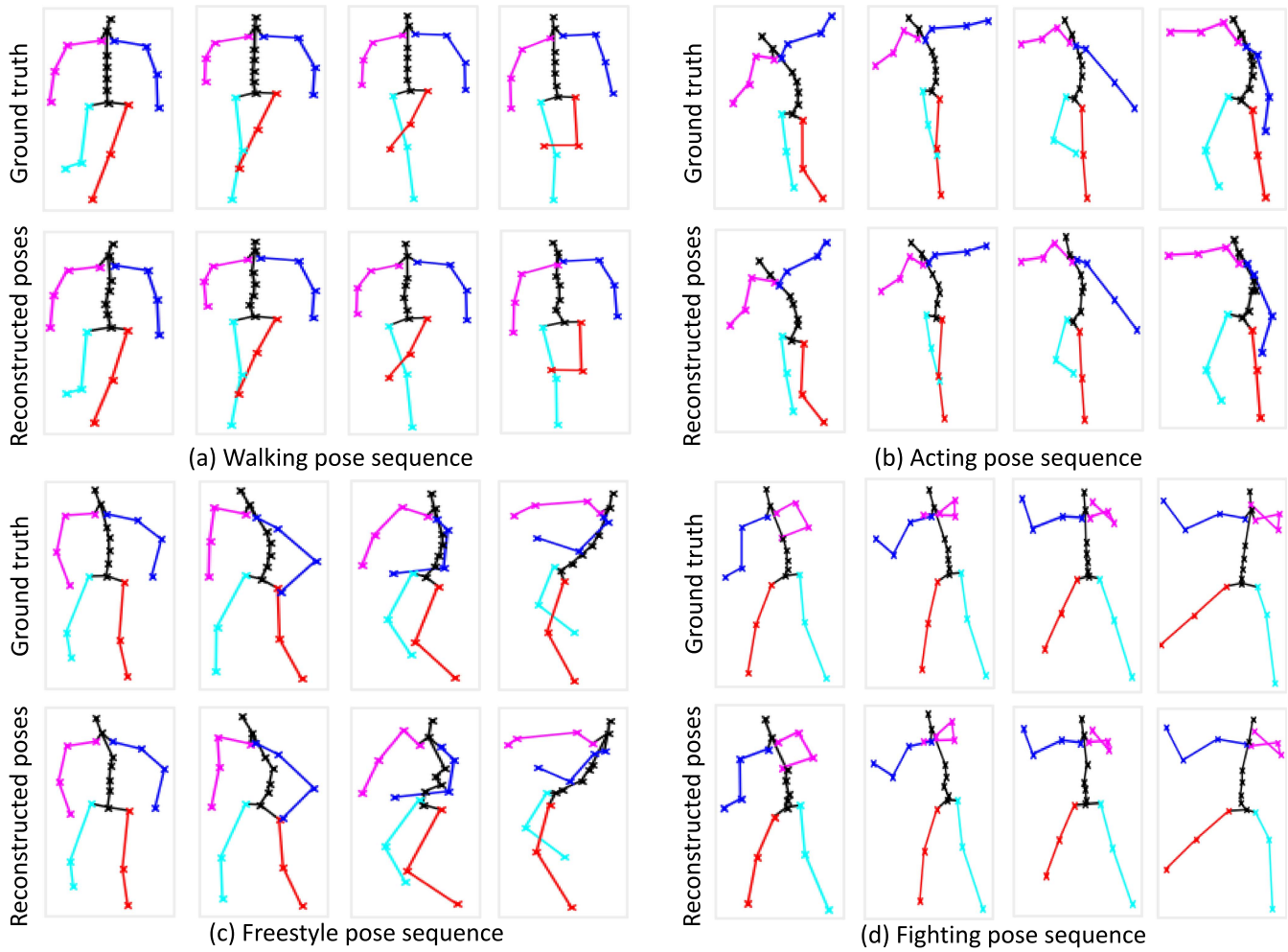

Fig. 6. Visualization of our human pose reconstruction over time on test samples of TotalCapture dataset. The four examples represent the full-body of 'walking', 'acting', 'freestyle,' and 'fighting', pose sequences, respectively. Our AttRNet well reconstructs the full-body 3D human poses based on the sensed data from wearable IMUs.

concern. Therefore, in the future, we plan to develop a teacher-student learning-based Knowledge Distillation (KD) mechanism to transfer the knowledge from the multi-modal (wearable cameras + wearable IMUs) teacher network to the single-modal (wearable IMUs only) student network. During the test, we only use the trained single-modal (IMU only) student network for pose reconstruction and prediction.

VI. Conclusion

In this article, we proposed the wearable motion capture problem of reconstructing and predicting 3D human poses from the wearable cameras and IMUs. We developed a novel Attention-Oriented Recurrent Neural Network (AttRNet) to solve the wearable motion capture problem, which contains a sensor-wise attention-oriented recurrent encoder, a reconstruction module, and a dynamic temporal attention-oriented recurrent decoder, to reconstruct the current pose and predict the future poses. The extensive experiments on a newly collected WearableMotionCapture dataset show the effectiveness of each module of our AttRNet and the fusion of two sensor modalities. Our AttRNet also outperforms the current best methods on two full-body pose datasets [4], [13].

References

[1] Vicon, "Vicon Motion Systems Ltd.," [Online]. Available: http://www.vicon.com/
[2] F. Huang, A. Zeng, M. Liu, Q. Lai, and Q. Xu, "DeepFuse: An IMU-aware network for real-time 3D human pose estimation from multi-view image," in *Proc. Winter Conf. Appl. Comput. Vis.*, 2020, pp. 429–438.
[3] C. Malleson, A. Gilbert, M. Trumble, J. Collomosse, A. Hilton, and M. Volino, "Real-time full-body motion capture from video and IMUs," in *Proc. Int. Conf. 3D Vis.*, 2017, pp. 449–457.
[4] M. Trumble, A. Gilbert, C. Malleson, A. Hilton, and J. P. Collomosse, "Total capture: 3D human pose estimation fusing video and inertial sensors," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–13.
[5] T. v. Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3D human pose in the wild using IMUs and a moving camera," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 601–617.
[6] Z.-P. Bian, J. Hou, L.-P. Chau, and N. Magnenat-Thalmann, "Fall detection based on body part tracking using a depth camera," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 2, pp. 430–439, Mar. 2015.
[7] D. Leightley, J. S. McPhee, and M. H. Yap, "Automated analysis and quantification of human mobility using a depth sensor," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 4, pp. 939–948, Jul. 2017.

[8] D. Roetenberg, H. Luinge, and P. Slycke, "Xsens MVN: Full 6DOF human motion tracking using miniature inertial sensors," *Xsens Motion Technol. BV*, Enschede, Netherlands, Tech. Rep. 1, 2009.

[9] D. Yang, D. Kim, and S.-H. Lee, "LoBSTr: Real-time lower-body pose prediction from sparse upper-body tracking signals," in *Computer Graphics Forum*, vol. 40, Hoboken, NJ, USA: Wiley Online Library, 2021, pp. 265–275.

[10] V. Guzov, A. Mir, T. Sattler, and G. Pons-Moll, "Human POSEitioning system (HPS): 3D human pose estimation and self-localization in large scenes from body-mounted sensors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4318–4329.

[11] X. Yi et al., "Physical inertial poser (PIP): Physics-aware real-time human motion tracking from sparse inertial sensors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 13167–13178.

[12] Y. Jiang, Y. Ye, D. Gopinath, J. Won, A. W. Winkler, and C. K. Liu, "Transformer inertial poser: Real-time human motion reconstruction from sparse IMUs with simultaneous terrain generation," in *Proc. SIGGRAPH Asia Conf. Papers*, 2022, pp. 1–9.

[13] Y. Huang, M. Kaufmann, E. Aksan, M. J. Black, O. Hilliges, and G. Pons-Moll, "Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time," *ACM Trans. Graph.*, vol. 37, no. 6, pp. 1–15, 2018.

[14] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 483–499.

[15] Xsens, "Inertial-sensor modules.," [Online]. Available: https://www.xsens.com/inertial-sensor-modules

[16] S. Chen, J. Lach, B. Lo, and G.-Z. Yang, "Toward pervasive gait analysis with wearable sensors: A systematic review," *IEEE J. Biomed. Health Informat.*, vol. 20, no. 6, pp. 1521–1537, Nov. 2016.

[17] M. S. B. Hossain, J. Dranetz, H. Choi, and Z. Guo, "DeepBBWAE-Net: A CNN-RNN based deep superlearner for estimating lower extremity sagittal plane joint kinematics using shoe-mounted IMU sensors in daily living," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 8, pp. 3906–3917, Aug. 2022.

[18] P. Kang, J. Li, B. Fan, S. Jiang, and P. B. Shull, "Wrist-worn hand gesture recognition while walking via transfer learning," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 3, pp. 952–961, Mar. 2022.

[19] D. Nagaraj, E. Schake, P. Leiner, and D. Werth, "An RNN-ensemble approach for real time human pose estimation from sparse IMUs," in *Proc. Int. Conf. Appl. Intell. Syst.*, 2020, pp. 1–6.

[20] P. Müller, M.-A. Bégin, T. Schauer, and T. Seel, "Alignment-free, self-calibrating elbow angles measurement using inertial sensors," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 2, pp. 312–319, Mar. 2017.

[21] M. Dantone, J. Gall, C. Leistner, and L. V. Gool, "Human pose estimation using body parts dependent joint regressors," in *Proc. Comput. Vis. Pattern Recognit.*, 2013, pp. 3041–3048.

[22] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "DeeperCut: A deeper, stronger, and faster multi-person pose estimation model," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 34–50.

[23] G. Papandreou et al., "Towards accurate multi-person pose estimation in the wild," in *Proc. Comput. Vis. Pattern Recognit.*, 2017, pp. 4903–4911.

[24] J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Proc. Neural Inf. Process. Syst.*, 2014, pp. 1799–1807.

[25] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," in *Proc. Comput. Vis. Pattern Recognit.*, 2016, pp. 4733–4742.

[26] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, "Multi-scale structure-aware network for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 713–728.

[27] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 717–732.

[28] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proc. Comput. Vis. Pattern Recognit.*, 2017, pp. 1831–1840.

[29] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 1281–1290.

[30] C.-H. Chen and D. Ramanan, "3D human pose estimation=2D pose estimation matching," in *Proc. Comput. Vis. Pattern Recognit.*, 2017, pp. 7035–7043.

[31] H. Yasin, U. Iqbal, B. Kruger, A. Weber, and J. Gall, "A dual-source approach for 3D pose estimation from a single image," in *Proc. Comput. Vis. Pattern Recognit.*, 2016, pp. 4948–4956.

[32] S. Li and A. B. Chan, "3D human pose estimation from monocular images with deep convolutional neural network," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 332–347.

[33] J. Dong, W. Jiang, Q. Huang, H. Bao, and X. Zhou, "Fast and robust multi-person 3D pose estimation from multiple views," in *Proc. Comput. Vis. Pattern Recognit.*, 2019, pp. 7792–7801.

[34] M. Kocabas, S. Karagoz, and E. Akbas, "Self-supervised learning of 3D human pose using multi-view geometry," in *Proc. Comput. Vis. Pattern Recognit.*, 2019, pp. 1077–1086.

[35] A. Kadkhodamohammadi and N. Padoy, "A generalizable approach for multi-view 3D human pose regression," *Mach. Vis. Appl.*, vol. 32, no. 1, pp. 1–14, 2021.

[36] Z. Guo et al., "Vision-based finger tapping test in patients with Parkinson's disease via spatial-temporal 3D hand pose estimation," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 8, pp. 3848–3859, Aug. 2022.

[37] H. Rhodin et al., "Learning monocular 3D human pose estimation from multi-view images," in *Proc. Comput. Vis. Pattern Recognit.*, 2018, pp. 8437–8446.

[38] A. M. Lehrmann, P. V. Gehler, and S. Nowozin, "Efficient nonlinear Markov models for human motion," in *Proc. Comput. Vis. Pattern Recognit.*, 2014, pp. 1314–1321.

[39] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 283–298, Feb. 2008.

[40] A. Gopalakrishnan, A. Mali, D. Kifer, L. Giles, and A. G. Ororbia, "A neural temporal model for human motion prediction," in *Proc. Comput. Vis. Pattern Recognit.*, 2019, pp. 12116–12 125.

[41] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *Proc. Comput. Vis. Pattern Recognit.*, 2017, pp. 2891–2900.

[42] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-RNN: Deep learning on spatio-temporal graphs," in *Proc. Comput. Vis. Pattern Recognit.*, 2016, pp. 5308–5317.

[43] Z. Liu et al., "Towards natural and accurate future motion prediction of humans and animals," in *Proc. Comput. Vis. Pattern Recognit.*, 2019, pp. 10004–10012.

[44] J. Walker, K. Marino, A. Gupta, and M. Hebert, "The pose knows: Video forecasting by generating pose futures," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 3332–3341.

[45] C. Li, Z. Zhang, W. S. Lee, and G. H. Lee, "Convolutional sequence to sequence model for human dynamics," in *Proc. Comput. Vis. Pattern Recognit.*, 2018, pp. 5226–5234.

[46] G. Ning, Z. Zhang, and Z. He, "Knowledge-guided deep fractal neural networks for human pose estimation," *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1246–1259, May 2018.

[47] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning trajectory dependencies for human motion prediction," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 9489–9497.

[48] M. Ma, H. Fan, and K. M. Kitani, "Going deeper into first-person activity recognition," in *Proc. Comput. Vis. Pattern Recognit.*, 2016, pp. 1894–1903.

[49] K. Matsuo, K. Yamada, S. Ueno, and S. Naito, "An attention-based activity recognition for egocentric video," in *Proc. Comput. Vis. Pattern Recognit. Workshops*, 2014, pp. 551–556.

[50] Y. Li, A. Fathi, and J. M. Rehg, "Learning to predict gaze in egocentric video," in *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 3216–3223.

[51] P. Morerio, L. Marcenaro, and C. S. Regazzoni, "Hand detection in first person vision," in *Proc. IEEE Int. Conf. Inf. Fusion*, 2013, pp. 1502–1507.

[52] M. P. Kadaba, H. Ramakrishnan, and M. Wootten, "Measurement of lower extremity kinematics during level walking," *J. Orthopaedic Res.*, vol. 8, no. 3, pp. 383–392, 1990.

[53] S. L. Delp et al., "Opensim: Open-source software to create and analyze dynamic simulations of movement," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 11, pp. 1940–1950, Nov. 2007.

[54] X. Yi, Y. Zhou, and F. Xu, "Transpose: Real-time 3D human translation and pose estimation with six inertial sensors," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 1–13, 2021.

[55] T. V. Marcard, B. Rosenhahn, M. J. Black, and G. Pons-Moll, "Sparse inertial poser: Automatic 3D human pose estimation from sparse IMUs," *Comput. Graph. Forum*, vol. 36, no. 2, pp. 349–360, 2017.