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

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

Abstract—Identifying 3D human walking poses in unconstrained environments has many applications such as enabling prosthetists and clinicians to access the amputees’ walking functions outside clinics and helping amputees obtain an optimal walking condition with predictive control. Thus, we propose the wearable motion capture problem of reconstructing and predicting 3D human poses from the wearable IMU sensors and wearable cameras. To solve this challenging 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 current pose and predict the future poses. 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 WearableMotionCapture dataset, and it also outperforms the current best methods on two public pose prediction datasets with IMU-only data: DIP-IMU and TotalCaputre. The source codes and the new dataset will be publicly available on <https://github.com/MoniruzzamanMd/Wearable-Motion-Capture>.

Index Terms—3D Human Pose Reconstruction, 3D Human Pose Prediction, Recurrent Neural Network, Wearable Sensing

I. INTRODUCTION

OVER 1.6 million people in the United States are living with lower-limb amputations. This number is expected to double by 2050, unfortunately. Individuals with lower limb amputation face multiple disadvantages while walking, such as increased strains on the lower back, increased metabolic cost, and gait asymmetry. Appropriate prosthetic devices can mitigate these disadvantages and improve quality of life. Therefore, there is a need to develop a smart prosthetic device with energy release that can provide continuous optimal prosthetic tuning parameters under different walking conditions. However, the development of such devices require 3D walking pose identification.

Most commonly, the motion capture systems [1] are used to achieve a highly accurate understanding of the human

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

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

M. Hossain is with the Department of Electrical and Computer Engineering, University of Central Florida (sanzid@knights.ucf.edu)

Z. Guo is with the Department of Electrical and Computer Engineering, University of Central Florida (e-mail: zsguo@ucf.edu)

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

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 amputees alone outdoors, we propose the on-body camera and IMU sensor solution to reconstruct and predict walking poses.

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.

Recently, several works [2], [3], [4], [5], [6] focused on the reconstruction of past or current human poses from third-person view RGB or RGB-D cameras, with wearable IMU sensors. However, these methods based on third-person views are not always possible for amputees alone outdoors. Thus, there is a need to reconstruct and predict human poses from wearable sensors only, so amputees can walk freely in their daily lives while being modeled and assisted.

Some works relied on 17 or more IMU sensors to obtain an accurate pose reconstruction [7], but wearing 17 sensors are very uncomfortable and impractical to use in daily living. A single chest-mounted wide-angle camera was used to roughly estimate the human motion [8], but the full-body motion is simplified to a classification task of a limited number of human actions. *When amputees walk in their daily lives, we really need a feasible and effective approach to accurately sense their poses with a small set of wearable sensors, so prosthetists and clinicians can access their patients’ walking functions outside clinics and help amputees obtain an optimal walking condition with prompt feedbacks.*

Therefore, we propose the fusion of a small number of wearable IMUs and wearable cameras to reconstruct the current and predict the future 3D poses of human subjects in unconstrained environment, overcoming the limitations of motion capture cameras and third-person view cameras, as shown in Fig 1. Reconstructing human poses during walking is for clinical diagnosis. Predicting human poses in the future is for real-time optimal control of the prosthetic device. Note, this collaborative project with Veteran Affairs aims to assist soldiers or patients with lower limb amputation, and the wearable cameras are used for environment sensing as well. In this paper, we mainly summarize how we integrate camera sensors and IMU sensors for the current pose reconstruction and the future pose prediction. Since we aim to use as less number of sensors as possible and focus on walking poses of people with lower limb amputations, we only use one sensor in the upper body and the remaining sensors in the lower body.

Currently, there is no public dataset that contains both wearable IMU and wearable camera data, along with its ground truth of 3D walking poses. In our assistive walking project, 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. We also compare the performance of our wearable motion capture method on two related datasets: DIP-IMU [9] (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 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. 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 on-body camera and IMU sensor solution to reconstruct and predict walking poses.
- 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 current pose and predict the future poses.
- 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 released to the community to promote the wearable motion capture.
- Our approach is able to generalize to both multi-modal and single-modal sensor input. Compared with two state-of-the-arts [9], [4] on their single-modal IMU-only datasets, our wearable motion capture shows its superior performance.

II. PROPOSED APPROACH

In this section, we present our proposed Attention-Oriented Recurrent Neural Network (AttRNet) for reconstructing the

current pose and predicting the future poses.

A. Problem Statement

Suppose that the observed IMU data are $\mathbb{I}_{[T-\Delta_1+1, T]} = [\mathbf{I}_{T-\Delta_1+1}, \dots, \mathbf{I}_T] \in \mathbb{R}^{\Delta_1 \times M \times D_{IMU}}$ and video data are $\mathbb{V}_{[T-\Delta_1+1, T]} = [\mathbf{V}_{T-\Delta_1+1}, \dots, \mathbf{V}_T] \in \mathbb{R}^{\Delta_1 \times N \times D_{Video}}$, 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 current pose and predict the future poses 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 and the reconstructed poses of the past two time steps for pose dynamics calculation. $\hat{\mathbb{P}}_{[T+1, T+\Delta_2]} = [\hat{\mathbf{P}}_{T+1}, \dots, \hat{\mathbf{P}}_{T+\Delta_2}] \in \mathbb{R}^{\Delta_2 \times J \times D_{Joints}}$ are the predicted future poses, J is the number of joints in the pose model, D_{Joints} is the feature 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 reconstruct the current pose and predict the future poses 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 device orientation as a 3×3 rotation matrix from raw IMU data, which is then flattened to a 9-dimensional vector as the orientation feature. Finally, the acceleration, angular velocity, and 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 $D_{IMU} = 15$.

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 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 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.

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 Encoder

Our attention-oriented 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 $\mathbf{I}_t = [\mathbf{I}_t^1; \dots; \mathbf{I}_t^M] \in \mathbb{R}^{M \times D_{IMU}}$ 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 output of the second fully-

connected layer is passed through a sigmoid function that enforces the attention scores in $\mathbf{u}_t^I$ to be between 0 and 1. Then, the attention weighted IMU features are computed as:

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

where $\odot$ 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. $\mathbf{I}_t^a$ is the attention weighted IMU feature at time t . Similarly, for the video features from N wearable cameras at time t , we compute the attention weighted video features $\mathbf{V}_t^a$:

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

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 fuse them in a single vector which will be used to reconstruct the current pose and decode the future poses. Given the attention-weighted IMU features $\mathbf{I}^a = [\mathbf{I}_{T-\Delta_1+1}^a; \dots; \mathbf{I}_T^a] \in \mathbb{R}^{\Delta_1 \times M \times D_{IMU}}$, the recurrent encoder passes these features through a Bidirectional Gated Recurrent Unit (B-GRU):

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

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 $\mathbf{V}^a = [\mathbf{V}_{T-\Delta_1+1}^a, \dots, \mathbf{V}_T^a] \in \mathbb{R}^{\Delta_1 \times N \times D_{Video}}$ are encoded as:

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

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.

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 goal of the recurrent decoder is to predict the future 3D poses $\hat{\mathbb{P}}_{[T+1, T+\Delta_2]}$. The recurrent decoder consists of GRU and the hidden state at each step is updated using the GRU update rules:

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

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. 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:

$$\hat{\mathbf{P}}_\tau = \mathbf{h}_\tau^d \mathbf{w}_\tau^P \quad (6)$$

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

Fig. 3. The 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.

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 $\mathbf{H}^{IMU} = [\mathbf{h}_{T-\Delta_1+1}^{IMU}, \dots, \mathbf{h}_T^{IMU}]$, 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, as follows:

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

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

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

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:

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

where $\mathcal{O}_{\tau-1}^{IMU} \in \mathbb{R}^{f \times 1}$ 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:

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

where $\mathcal{O}_{\tau-1}^{Video} \in \mathbb{R}^{f \times 1}$ 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

TABLE I
EXISTING 3D HUMAN POSE DATASET VS WEARABLEMOTIONCAPTURE DATASET.

Dataset	Wearable IMU	Wearable camera	Third-person view camera	#Minutes	#Subjects	#Pose sequences
TotalCapture [4]	✓		✓	50	5	46
DIP-IMU [9]	✓			92	10	64
WearableMotionCapture (ours)	✓	✓		327	10	140

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 $\mathcal{A}_{\tau-1} = (\hat{\mathbf{P}}_{\tau-1} - 2\hat{\mathbf{P}}_{\tau-2} + \hat{\mathbf{P}}_{\tau-3})$. 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 Eq. (5)), as follows:

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

G. Loss

The loss function in the proposed AttRNet is composed of two terms:

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

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 for the pose reconstruction, $\mathbb{P}_{[T+1,T+\Delta_2]} = [\mathbf{P}_{T+1}, \dots, \mathbf{P}_{T+\Delta_2}] \in \mathbb{R}^{\Delta_2 \times J \times D_{Joints}}$ are the ground truth poses for the future pose prediction, and $|| \cdot ||_2$ denotes the l_2 norm.

III. EXPERIMENTS

In this section, we describe the experimental results of our method. First, we introduce the datasets, and the implementation details of our proposed approach. Then we perform the experiments on our proposed WearableMotionCapture dataset. After that we compare our method with other state-of-the-art methods. Finally, we present some qualitative results, along with discussions and future works.

A. Dataset

We conduct experiments on wearable IMU+camera dataset collected by us: WearableMotionCapture, and two datasets of wearable IMUs: DIP-IMU [9] and TotalCapture [4].

WearableMotionCapture: There are some 3D human pose datasets from wearable IMU sensors and third-person view cameras, e.g. [4]. But, there is no public dataset to reconstruct or predict 3D human poses from wearable IMU and wearable video camera sensors yet. Thus, we collected such a dataset, named it as WearableMotionCapture in this paper.

We recorded data from 10 subjects (6 male and 4 female) wearing 8 IMU sensors (2 attached on the upper body and 6 on the lower body), and 2 video cameras (attached in the middle of the shank contour of each leg), as shown in Fig. 1(a).

Fig. 4. Walking scenarios of our WearableMotionCapture dataset using wearable IMU and camera sensors. (a) The subject walks on treadmill; (b) The subject walks on a round path; (c) The subject walks on ground; (d) The subject walks on stairs; (e) The subject walks on slope; and (f) The subject walks on ground and avoids the obstacles.

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 with 1080p resolution and wide-angle configuration. We down-sample the IMU data to have the same fps with the video data. Each subject was instructed to walk on several scenarios such as treadmill, ground, slope and stairs. 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 trials 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. The walking scenarios of our WearableMotionCapture dataset are shown in Fig. 4. 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 [10], for ground-truth collection. Three-dimensional marker trajectories were captured by motion capture cameras. We obtain the ground truth of joint angles using OpenSim [11], 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 588K video frames. The dataset overview is shown in Table I.

DIP-IMU [9]: 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 sensors 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. We use PyTorch to implement our proposed pose reconstruction and prediction model, and it takes about 60 minutes to train our network on WearableMotionCapture dataset on a single Tesla V100 GPU.

C. Evaluation on WearableMotionCapture

Experimental setup: The dataset contains 140 walking pose sequences from 10 subjects, and 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

TABLE II
EVALUATION OF POSE RECONSTRUCTION (RECO.), AND FUTURE POSE PREDICTION (PRED.), 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. THE PREDICTION MODELS ARE EVALUATED IN 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.

Methods	Half-Half	LOO
AttRNet (IMU+Video) (Reco.)	4.65	7.49
AttRNet (IMU+Video) (Pred.)	4.73	7.58
(i) w/o SAM	5.24	7.95
(ii) w/o SAM/D-TAM	5.70	8.51
(iii) w/o SAM/D-TAM/Pose dynamics	6.05	9.03
(iv) AttRNet (IMU) (Pred.)	7.09	10.86
(v) AttRNet (Video) (Pred.)	7.74	11.65

TABLE III
ABLATION STUDY OF ATTRNET ON DIFFERENT WALKING SCENARIOS. NUMBERS REPRESENT THE MEAN ABSOLUTE ERRORS OF JOINT ANGLES OVER 360 DEGREES FOR HALF AND HALF EVALUATION, WHERE THE PREDICTION MODELS ARE EVALUATED IN ONLINE MODE WITH 50 OBSERVED TIME STEPS TO PREDICT THE NEXT 10 TIME STEPS.

scenarios	Error	scenarios	Error
Treadmill (slow)	1.81	Ground (fast)	5.58
Treadmill (normal)	1.83	Ground (very fast)	5.71
Treadmill (fast)	1.95	Stair (trial-1)	5.77
Treadmill (very fast)	2.56	Stair (trial-2)	5.85
Round	4.69	Slope (trial-1)	5.09
Ground (slow)	5.11	Slope (trial-2)	5.52
Ground (normal)	5.43	Obstacles	9.28

half and half evaluation (i.e., the test set has 294K time steps from 70 pose sequences), we generate around $L_{test} = 294K - 70 \times (\Delta_1 + \Delta_2)$ 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 shown in Table II, for the pose reconstruction, our AttRNet achieves MAE of 4.65 degree for half and half evaluation, and 7.49 degree for leave-one-out evaluation. On the other hand, for the future pose prediction, our AttRNet achieves MAE of 4.73 degree and 7.58 degree for half and half evaluation and leave-one-out evaluation, respectively, slightly higher than the reconstruction error. In the following, we perform the ablation study and detailed evaluation on the prediction task only, which is more challenging than the reconstruction task.

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) We apply our AttRNet without Sensor-wise Attention Module (SAM); (ii) We perform the experiments without SAM or Dynamic Temporal Attention Module (D-TAM); (iii) We perform the experiment without

TABLE IV

PER JOINT PERFORMANCE EVALUATION. NUMBERS REPRESENT THE MEAN ABSOLUTE ERRORS OF JOINT ANGLES OVER 360 DEGREES FOR HALF AND HALF EVALUATION, WHERE THE PREDICTION MODELS ARE EVALUATED IN ONLINE MODE WITH 50 OBSERVED TIME STEPS TO PREDICT THE NEXT 10 TIME STEPS.

Joints	MAE	Joints	MAE
Pelvis-tilt	3.45	Hip-rotation-right	7.66
Pelvis-list	3.03	Knee-angle-left	4.74
Pelvis-rotation	9.68	Knee-angle-right	4.78
Hip-flexion-left	4.10	Ankle-angle-left	3.49
Hip-flexion-right	4.15	Ankle-angle-right	3.47
Hip-adduction-left	3.24	Lumber-extension	3.69
Hip-adduction-right	3.30	Lumber-bending	3.83
Hip-rotation-left	7.86	Lumber-rotation	5.26

TABLE V

PARAMETER ANALYSIS ON Δ_1 AND Δ_2 . (Δ_1, Δ_2) MEAN THAT THE PREDICTION MODELS ARE EVALUATED IN ONLINE MODE USING Δ_1 OBSERVED TIME STEPS TO PREDICT Δ_2 FUTURE TIME STEPS. NUMBERS REPRESENT THE MEAN ABSOLUTE ERRORS OF JOINT ANGLES OVER 360 DEGREES FOR HALF AND HALF EVALUATION.

(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

SAM, D-TAM or pose dynamics; (iv) and (v) We apply AttRNet with SAM, D-TAM and Pose Dynamics on IMU-only and Video-only, respectively. As shown in Table II, we can clearly see that each algorithm component is contributing to our AttRNet to improve the performance for both half and half and leave-one-out 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: We also perform the ablation studies on different walking scenarios, as shown in Table III. The prediction performance for walking on the treadmill is the best compared to other scenarios, while we get the lowest performance for walking on the ground while avoiding obstacles. The walking motions on the treadmill are repetitive, which makes it easier to predict the future poses. On the other hand, the walking motions are continuously changed when the subjects try to avoid the obstacles on the ground, which makes the future pose prediction more difficult. Over all scenarios in Table III, our maximal relative error is $9.28/360=2.58\%$.

Per joint evaluation: We perform the per joint evaluation, as shown in Table IV. The prediction performance for the joint ‘Pelvis list’ is the best compared to other joints, while we get the lowest performance for the joint ‘Pelvis rotation’. In our dataset, the subjects continuously try to rotate or change the walking direction. Therefore, predicting the joint angles related to rotation (e.g., ‘Pelvis rotation’, ‘Hip rotation left’, ‘Hip rotation right’ and ‘Lumber rotation’) becomes more difficult.

Parameter analysis: We perform experiments on different temporal intervals (Δ_1 and Δ_2) of the observed and future

TABLE VI

COMPARISON RESULTS ON DIP-IMU DATASET FOR FUTURE POSE PREDICTION. ERRORS ARE REPORTED AS BOTH POSITIONAL ERRORS IN CENTIMETERS AND JOINT ANGLE ERRORS IN DEGREES. NUMBERS IN PARENTHESIS (Δ_1, Δ_2) MEAN THAT THE MODELS ARE EVALUATED IN ONLINE MODE USING Δ_1 OBSERVED FRAMES TO PREDICT Δ_2 FUTURE FRAMES.

Methods	(Δ_1, Δ_2)	Error (cm)	Error (deg)
DIP [9] (IMU)	(20, 5)	6.63	18.49
AttRNet (IMU)	(20, 5)	3.59	9.33
DIP [9] (IMU)	(50, 5)	6.52	18.14
AttRNet (IMU)	(50, 5)	3.45	8.80

TABLE VII

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

time steps. In the left side of Table V, we show the prediction performance of our method for different temporal intervals Δ_1 of the observed sequences. The prediction error decreases with 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, which are sufficient for the prediction). On the other hand, in the right side of Table V, we show the performance for different temporal intervals Δ_2 in the future, where the prediction error increases for the far future pose prediction.

D. Comparison with the State-of-the-art

DIP-IMU evaluation: The DIP-IMU is designed to predict 3D human poses from wearable IMU sensors. Table VI shows the comparison results of our AttRNet with the current best results on DIP-IMU dataset [9] for the future pose prediction, where the results are compared for different intervals of observed sequences and different intervals of future predictions. Over all scenarios, our method establishes the new state-of-the-art results on DIP-IMU dataset for 3D human pose prediction, outperforming the current best by a large margin.

TotalCapture evaluation: The TotalCapture dataset is designed to reconstruct poses from the wearable IMUs and multiple third-person view video cameras. Table VII 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.

E. Qualitative Analysis

We present some qualitative results on the test samples of WearableMotionCapture dataset in Figure 5, where our AttRNet captures data from wearable IMU sensors and wearable cameras, and predicts 3D walking poses. It can be seen that the proposed method can successfully predict different walking poses on different scenarios (e.g., treadmill, stairs, slope,

Fig. 5. Visualization of our human pose prediction 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 predicts the 3D human poses based on the sensed data from wearable IMUs and wearable cameras.

and ground with obstacles). The related video demos can be accessed at <https://github.com/MoniruzzamanMd/Wearable-Motion-Capture>. The joint angle error of avoiding obstacles is the highest in Table III, but the visual evaluation in Figure 5(b) is not very obvious in individual time steps. We will work with prosthetists to find out what is the clinically-negligible angle error.

Though our project focuses on the lower limb amputees, the proposed method can be generalized to full body pose reconstruction and prediction using the 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 predicts fully-body 3D pose sequences. It can be seen that the proposed method also shows good performance on the fully-body pose prediction.

IV. 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, among which approximately 1600 are United States military with service-related amputations. However, the development of such a device requires an algorithm to reconstruct and predict walking poses in an uncontrolled daily-living environment, and an algorithm

Fig. 6. Visualization of our human pose prediction 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 predicts the fully-body 3D human poses based on the sensed data from wearable IMUs.

to sense the walker’s surrounding 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 on-body camera and IMU sensor solution to reconstruct and predict walking poses, which aids the prosthetic device control and clinicians’ diagnoses on amputees out of clinics.

For pose reconstruction and future pose prediction, different sensors on the human body have different capabilities to capture different joint movements at a specific time. Therefore, how to design a neural network that can compute sensor-wise attention scores at each time step to reconstruct the current pose and predict the future poses is an open research question. In this paper, we propose an Attention-Oriented Recurrent Neural Network (AttRNet) to reconstruct the current pose and predict the future poses in an online setting from both wearable IMU sensors and wearable cameras. In our AttRNet, we introduce an attention-oriented recurrent encoder, which performs sensor-wise attention at each time step and embeds features over different time steps of the observed input sequences. The performance improvement from our sensor-wise

attention-oriented recurrent encoder indicates that our network can effectively encode the most highly discriminative features from the observed input sequences. Intuitively, the observed poses at different time steps are highly related to different future poses. Therefore, how to design a neural network that can output a series of future poses by dynamically computing the relevant information from the encoded observed poses is another open research question. This motivates us to design an attention-oriented recurrent decoder, which is configured with Dynamic Temporal Attention Module (D-TAM). The further improvement in future pose prediction after using D-TAM in recurrent decoder indicates that the D-TAM can dynamically compute the relevant information from the encoded observed features.

Currently, the proposed WearableMotionCapture dataset is collected in the lab environment. We obtain the ground-truth poses from the marker-based motion capture system, which is difficult to set up outdoors. In the lab, we collected the dataset close to the real-world setting such as waking on ground, slope, stair, treadmill, round paths or random paths with obstacles. In the future, we will consider the wild environment

for testing, but the ground truth pose will be lack. On the other hand, in this paper, we mainly summarize how we integrate camera sensors and IMU sensors for the current pose reconstruction and the future pose prediction. In the future, in addition to our proposed AttRNet to reconstruct the current pose and predict the future poses, we will also develop a real-time multi-task learning framework for analyzing the surrounding walking environment as well.

V. RELATED WORKS

In this section, we review three types of motion capture methods: IMU-based, vision-based, and hybrid one.

IMU-based human motion capture: The wearable IMUs (e.g., Xsens [12]) show remarkable stability and accuracy in capturing bone orientation. Previously, Roetenberg et al. [7] introduced a motion tracking algorithm using IMU sensors. Recently, Huang et al. [9] 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. [13] introduced an RNN-ensemble approach for human pose estimation from IMU sensors. Most of these IMU-sensor based algorithms employ straight-forward recurrent neural networks to reconstruct human poses, without computing the highly discriminative features from the observed input sequences. *Differently, our AttRNet contains sensor-wise attention-oriented recurrent encoder, a reconstruction module, and a dynamic temporal attention-oriented recurrent decoder to reconstruct and predict 3D human poses.*

Vision-based human motion capture: The vision-based human motion capture can be divided into 2D pose estimation, and 3D future pose prediction. The state-of-the-arts of 2D human pose estimation has achieved impressive progresses. Previously, the heatmap-based approaches [14], [15], [16], [17], [18] and the regression-based algorithms [19], [20] were developed for 2D pose estimation. Newell et al. [21] introduced a deep hourglass model, which was later widely used as a backbone network [22], [23], [24], [25]. Recently, the research community pays a significant amount of attention to develop 3D pose estimation. Most of the methods [26], [27], [28], [29] on 3D pose estimation were originated from 2D pose estimation task. Some works [30], [31], [32], [33], [34], [35], [36] tried to develop multi-view based methods to get more accurate 3D pose estimation.

Despite the success of 2D/3D pose estimation, the future 3D pose prediction is yet under explored. Previously, some traditional methods such as hidden Markov model [37] and Gaussian model [38] were developed. Recently, some deep-learning based algorithms introduced recurrent networks [39], [40], [41], [42], [43], [44], [45] and feed forward networks [46], [47], [48] to predict future 3D poses. However, most of these vision-based pose analysis tasks relying on video data from third-person view cameras, which are not always possible in the daily life. 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 detection of objects [49],

[50], gaze [51], visible hands and arms [52]. *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. [6] 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. *Differently, we introduce an Attention-Oriented Recurrent Neural Network for reconstructing the current pose and predicting the future 3D poses from both wearable IMUs and wearable cameras.*

VI. CONCLUSIONS

In this paper, we proposed the wearable motion capture problem of reconstructing and predicting 3D human poses from the wearable IMU sensors and wearable cameras. 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 proposed approach also outperforms the current best methods on two IMU-only pose prediction datasets [4], [9].

ACKNOWLEDGMENT

This project was supported by National Science Foundation via Cyber-Physical Systems project CMMI-1646162, National Robotics Initiative project CMMI-1954548, and Human-Technology Frontier project ECCS-2025929.

REFERENCES

- [1] "Vicon. Vicon motion systems ltd." <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 *Winter Conference on Applications of Computer Vision*, 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 *International Conference on 3D Vision*, 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 *British Machine Vision Conference*, vol. 2, no. 5, 2017, pp. 1–13.
- [5] P. Hu, E. S.-I. Ho, and A. Munteanu, "3dbodynet: Fast reconstruction of 3d animatable human body shape from a single commodity depth camera," *IEEE Transactions on Multimedia*, 2021.

- [6] T. von 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 *European Conference on Computer Vision*, 2018, pp. 601–617.
- [7] D. Roetenberg, H. Luinge, and P. Slycke, "Xsens mvn: Full 6dof human motion tracking using miniature inertial sensors," *Xsens Motion Technologies BV, Tech. Rep.*, vol. 1, 2009.
- [8] H. Jiang and K. Grauman, "Seeing invisible poses: Estimating 3d body pose from egocentric video," in *Computer Vision and Pattern Recognition*, 2017, pp. 3501–3509.
- [9] 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 Transactions on Graphics*, vol. 37, no. 6, pp. 1–15, 2018.
- [10] M. P. Kadaba, H. Ramakrishnan, and M. Wooten, "Measurement of lower extremity kinematics during level walking," *Journal of orthopaedic research*, vol. 8, no. 3, pp. 383–392, 1990.
- [11] S. L. Delp, F. C. Anderson, A. S. Arnold, P. Loan, A. Habib, C. T. John, E. Guendelman, and D. G. Thelen, "Opensim: open-source software to create and analyze dynamic simulations of movement," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 11, pp. 1940–1950, 2007.
- [12] "Xsens. Inertial-sensor modules." <https://www.xsens.com/inertial-sensor-modules>.
- [13] D. Nagaraj, E. Schake, P. Leiner, and D. Werth, "An rnn-ensemble approach for real time human pose estimation from sparse imus," in *International Conference on Applications of Intelligent Systems*, 2020, pp. 1–6.
- [14] A. Kamel, B. Sheng, P. Li, J. Kim, and D. D. Feng, "Hybrid refinement-correction heatmaps for human pose estimation," *IEEE Transactions on Multimedia*, vol. 23, pp. 1330–1342, 2020.
- [15] M. Dantone, J. Gall, C. Leistner, and L. Van Gool, "Human pose estimation using body parts dependent joint regressors," in *Computer Vision and Pattern Recognition*, 2013, pp. 3041–3048.
- [16] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepcrut: A deeper, stronger, and faster multi-person pose estimation model," in *European Conference on Computer Vision*, 2016, pp. 34–50.
- [17] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards accurate multi-person pose estimation in the wild," in *Computer Vision and Pattern Recognition*, 2017, pp. 4903–4911.
- [18] J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," *arXiv preprint arXiv:1406.2984*, 2014.
- [19] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," in *Computer Vision and Pattern Recognition*, 2016, pp. 4733–4742.
- [20] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, "Multi-scale structure-aware network for human pose estimation," in *European Conference on Computer Vision*, 2018, pp. 713–728.
- [21] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*. Springer, 2016, pp. 483–499.
- [22] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *European Conference on Computer Vision*, 2016, pp. 717–732.
- [23] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Computer Vision and Pattern Recognition*, 2017, pp. 1831–1840.
- [24] F. Huang, A. Zeng, M. Liu, J. Qin, and Q. Xu, "Structure-aware 3d hourglass network for hand pose estimation from single depth image," *arXiv preprint arXiv:1812.10320*, 2018.
- [25] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in *International Conference on Computer Vision*, 2017, pp. 1281–1290.
- [26] C.-H. Chen and D. Ramanan, "3d human pose estimation= 2d pose estimation+ matching," in *Computer Vision and Pattern Recognition*, 2017, pp. 7035–7043.
- [27] H. Yasin, U. Iqbal, B. Kruger, A. Weber, and J. Gall, "A dual-source approach for 3d pose estimation from a single image," in *Computer Vision and Pattern Recognition*, 2016, pp. 4948–4956.
- [28] W.-Z. Nie, W.-W. Jia, W.-H. Li, A.-A. Liu, and S.-C. Zhao, "3d pose estimation based on reinforce learning for 2d image-based 3d model retrieval," *IEEE Transactions on Multimedia*, vol. 23, pp. 1021–1034, 2020.
- [29] S. Li and A. B. Chan, "3d human pose estimation from monocular images with deep convolutional neural network," in *Asian Conference on Computer Vision*, 2014, pp. 332–347.
- [30] J. Dong, W. Jiang, Q. Huang, H. Bao, and X. Zhou, "Fast and robust multi-person 3d pose estimation from multiple views," in *Computer Vision and Pattern Recognition*, 2019, pp. 7792–7801.
- [31] M. Kocabas, S. Karagoz, and E. Akbas, "Self-supervised learning of 3d human pose using multi-view geometry," in *Computer Vision and Pattern Recognition*, 2019, pp. 1077–1086.
- [32] A. Kadhodamohammadi and N. Padoy, "A generalizable approach for multi-view 3d human pose regression," *Machine Vision and Applications*, vol. 32, no. 1, pp. 1–14, 2021.
- [33] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Harvesting multiple views for marker-less 3d human pose annotations," in *Computer Vision and Pattern Recognition*, 2017, pp. 6988–6997.
- [34] H. Rhodin, J. Spörri, I. Katircioglu, V. Constantin, F. Meyer, E. Müller, M. Salzmann, and P. Fua, "Learning monocular 3d human pose estimation from multi-view images," in *Computer Vision and Pattern Recognition*, 2018, pp. 8437–8446.
- [35] Z. Liu, Z. Lin, X. Wei, and S.-C. Chan, "A new model-based method for multi-view human body tracking and its application to view transfer in image-based rendering," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1321–1334, 2017.
- [36] Y. Wu, D. Kong, S. Wang, J. Li, and B. Yin, "An unsupervised real-time framework of human pose tracking from range image sequences," *IEEE Transactions on Multimedia*, vol. 22, no. 8, pp. 2177–2190, 2019.
- [37] A. M. Lehrmann, P. V. Gehler, and S. Nowozin, "Efficient nonlinear markov models for human motion," in *Computer Vision and Pattern Recognition*, 2014, pp. 1314–1321.
- [38] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 283–298, 2007.
- [39] A. Gopalakrishnan, A. Mali, D. Kifer, L. Giles, and A. G. Ororbia, "A neural temporal model for human motion prediction," in *Computer Vision and Pattern Recognition*, 2019, pp. 12 116–12 125.
- [40] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *Computer Vision and Pattern Recognition*, 2017, pp. 2891–2900.
- [41] M. Li, S. Chen, Y. Zhao, Y. Zhang, Y. Wang, and Q. Tian, "Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction," in *Computer Vision and Pattern Recognition*, 2020, pp. 214–223.
- [42] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-rnn: Deep learning on spatio-temporal graphs," in *Computer Vision and Pattern Recognition*, 2016, pp. 5308–5317.
- [43] Z. Liu, S. Wu, S. Jin, Q. Liu, S. Lu, R. Zimmermann, and L. Cheng, "Towards natural and accurate future motion prediction of humans and animals," in *Computer Vision and Pattern Recognition*, 2019, pp. 10 004–10 012.
- [44] J. Walker, K. Marino, A. Gupta, and M. Hebert, "The pose knows: Video forecasting by generating pose futures," in *International Conference on Computer Vision*, 2017, pp. 3332–3341.
- [45] M. Li, Z. Zhou, and X. Liu, "Multi-person pose estimation using bounding box constraint and lstm," *IEEE Transactions on Multimedia*, vol. 21, no. 10, pp. 2653–2663, 2019.
- [46] C. Li, Z. Zhang, W. S. Lee, and G. H. Lee, "Convolutional sequence to sequence model for human dynamics," in *Computer Vision and Pattern Recognition*, 2018, pp. 5226–5234.
- [47] G. Ning, Z. Zhang, and Z. He, "Knowledge-guided deep fractal neural networks for human pose estimation," *IEEE Transactions on Multimedia*, vol. 20, no. 5, pp. 1246–1259, 2017.
- [48] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning trajectory dependencies for human motion prediction," in *International Conference on Computer Vision*, 2019, pp. 9489–9497.
- [49] M. Ma, H. Fan, and K. M. Kitani, "Going deeper into first-person activity recognition," in *Computer Vision and Pattern Recognition*, 2016, pp. 1894–1903.
- [50] K. Matsuo, K. Yamada, S. Ueno, and S. Naito, "An attention-based activity recognition for egocentric video," in *Computer Vision and Pattern Recognition Workshops*, 2014, pp. 551–556.
- [51] Y. Li, A. Fathi, and J. M. Rehg, "Learning to predict gaze in egocentric video," in *International Conference on Computer Vision*, 2013, pp. 3216–3223.
- [52] P. Morerio, L. Marcenaro, and C. S. Regazzoni, "Hand detection in first person vision," in *International Conference on Information Fusion*. IEEE, 2013, pp. 1502–1507.