IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 October 2025, accepted 13 October 2025, date of publication 22 October 2025, date of current version 29 October 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3624393

RESEARCH ARTICLE

Gait Inertial Poser (GIP): Gait-Aware Human Motion Capture Using Shoe-Embedded IMUs

RYOSUKE HORI[1,2], **(Student Member, IEEE), HIROYUKI DEGUCHI**[1,2], **(Student Member, IEEE), TSUBASA MARUYAMA**[1], **MITSUNORI TADA**[1], **AND HIDEO SAITO**[2], **(Senior Member, IEEE)**

[1]National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064, Japan
[2]Graduate School of Science and Technology, Keio University, Yokohama 223-8522, Japan

Corresponding author: Ryosuke Hori (hori-rysk@keio.jp)

This work was supported by Japan Society for the Promotion of Science (JSPS) Research Fellowship.

ABSTRACT Gait is a fundamental aspect of human mobility, and disruptions in normal gait can significantly reduce quality of life (QOL). Although recent advances in 3D gait analysis (3DGA) enable precise, quantitative assessments, these methods are typically confined to controlled laboratory environments and thus fail to accurately capture natural gait variability. Conversely, wearable IMU sensors offer cost-effective, portable solutions for capturing movements across diverse settings but face challenges such as invasiveness and sensor drift. In this study, we propose "Gait Inertial Poser (GIP)," a novel method estimating 3D full-body pose during straight walking on flat ground, using only two shoe-embedded IMU sensors. GIP initially estimates personalized body shapes from user attributes (height, weight, age, gender) and then employs a Transformer-based module to infer gait motion parameters from IMU data. To ensure temporal continuity and smoothness of the estimated motion, we further introduce a smoothing module based on a Variational Autoencoder (VAE), which further incorporates a specialized loss function that explicitly enforces kinematic constraints during foot-ground contact, thereby improving the overall estimation accuracy. Comprehensive experiments conducted on two public datasets quantitatively and qualitatively demonstrate that GIP achieves high accuracy in straight-line walking. This approach overcomes limitations of traditional laboratory-based methods, opening new opportunities for real-time monitoring and remote rehabilitation in everyday environments. The code will be available at https://github.com/RyosukeHori/GaitInertialPoser

INDEX TERMS Human motion capture, 3D pose estimation, gait analysis, IMU, deep learning.

I. INTRODUCTION

Gait is the most fundamental mode of human locomotion, encompassing critical information about a person's health and physical capabilities, such as gait phase, stride length, and muscle strength. Quantitative analysis of this information, known as gait analysis, is widely utilized in medical fields for diagnosing patients, monitoring disease progression, and evaluating treatment and rehabilitation effectiveness [1], [2]. It also plays a significant role in sports science and rehabilitation research. Traditionally, clinical gait analysis has relied heavily on subjective evaluations through visual observation by healthcare professionals and self-reports by patients. However, these methods are prone to variability among evaluators and human error. While Instrumented Gait Analysis (IGA), utilizing optical motion capture systems and force plates, provides objective and precise measurements [3], it is confined to laboratory settings and often fails to accurately reflect natural walking behaviors in daily life [4], [5], [6].

Recently, there has been growing interest in gait analysis methods employing Inertial Measurement Units (IMUs) to overcome these limitations [7], [8]. IMUs are compact, lightweight, cost-effective, and easily wearable, making them suitable for continuous gait measurement in various everyday environments. A conventional IMU-based method reconstructs foot trajectories by double integrating acceleration data [9], [10], [11]. Additionally, numerous

The associate editor coordinating the review of this manuscript and approving it for publication was Vincenzo Conti.

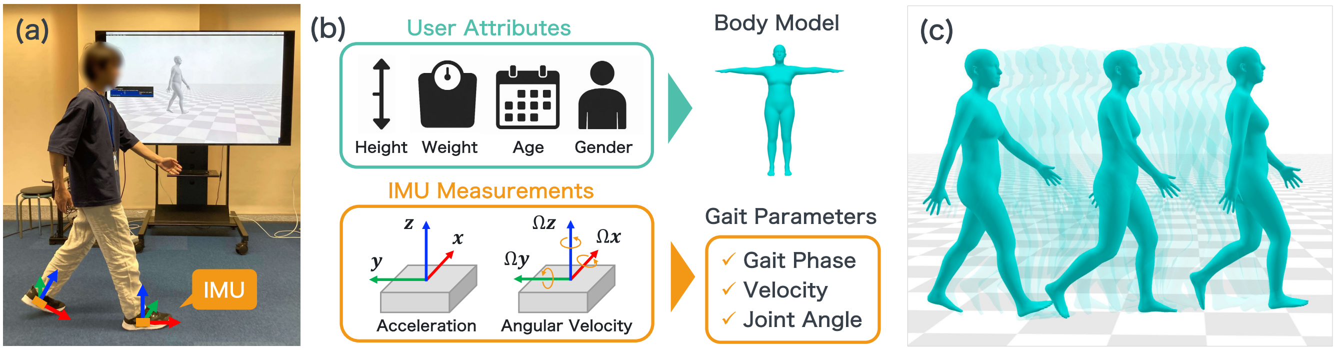


FIGURE 1. Overview of the proposed Gait Inertial Poser (GIP). (a) User wearing minimal inertial measurement units (IMUs) attached only to shoe soles, enabling unobtrusive daily gait measurement. (b) Input data consisting of easily obtainable user attributes (height, weight, age, gender) and IMU measurements (acceleration and angular velocity). These inputs allow estimation of the user's personalized body model and gait parameters, including gait phase, velocity, and joint angles. (c) Final output visualizing smooth, accurate full-body walking motions tailored to the user's characteristics.

machine learning-based approaches have been proposed to estimate lower-limb or full-body joint angles during gait from IMU data [12], [13], [14], [15]. More recently, methods have emerged capable of estimating not only joint angles but also full-body motion trajectories using only a few body-worn IMUs [16], [17], [18], [19]. However, these methods suffer from reduced estimation accuracy due to drift issues, including cumulative trajectory errors resulting from integration, and inaccuracies in global orientation information input into neural networks. These drift issues arise from the lack of absolute positioning in IMU sensors, as well as the inherent noise and measurement errors in IMU data.

To address these issues, we propose a novel method named "Gait Inertial Poser (GIP)." As illustrated in Figure 1, (a) GIP employs only two IMU sensors embedded in shoes, enabling non-invasive and highly accurate full-body gait motion estimation during straight walking on flat ground, suitable for unobtrusive measurements in daily life. (b) Additionally, the method leverages individual user attributes such as height, weight, age, and gender, enabling precise motion estimation tailored to each user's body shape. Specifically, user-specific body models are estimated from these attributes, and detailed gait parameters, including gait phases, walking velocity, and joint angles, are derived from IMU acceleration and angular velocity data via deep learning models. (c) By integrating these estimations, our approach enables the reconstruction of natural and precise full-body gait poses reflecting individual characteristics.

A key feature of our proposed method is intentionally limiting motion estimation to linear gait patterns. This design choice is critical to enabling highly accurate and practical gait motion estimation. It achieves this using only two shoe-embedded IMUs and requires no additional wearable devices or global positioning, making the approach realistic and widely applicable. Such a focus is well aligned with IGA, which primarily aims to quantify gait motion itself rather than absolute travel direction, and reflects typical clinical protocols (e.g., 10 m Walk Test) where straight and level walkways are standard. By restricting motion estimation to straight-line walking, we can avoid the drift issues typically observed in the conventional IMU-based full-body motion estimation methods. Unlike these methods, our approach does not require integrating angular velocities for global orientation or integrating accelerations to reconstruct displacement, and instead relies solely on IMU measurements within the sensor's local coordinate system, thereby avoiding integration-induced drift accumulation and enabling highly accurate gait motion estimation. Furthermore, we incorporate a Transformer-based deep learning model and a loss function inspired by Zero Velocity Update (ZUPT), achieving stable and accurate gait pattern estimation. We also employ a Variational Autoencoder (VAE) to enhance the temporal smoothness of the estimated motions.

To validate the effectiveness of our proposed method, we conducted quantitative and qualitative experiments using publicly available datasets such as the AIST Gait Database 2019 [20] and UnderPressure dataset [21]. Additionally, we developed a practical demonstration system utilizing commercially available IMU-equipped shoes to confirm the practical applicability of our method in real-world environments. In particular, we adopted a consumer-grade IMU device, ORPHE CORE [22], embedded in sneakers designed specifically for this device. This setup allows consistent and vibration-resistant sensor placement inside the shoe sole, mitigating variability due to IMU mounting positions or shoe type.

The main contributions of this study are summarized as follows:

1) We propose a non-invasive and highly accurate gait motion estimation method (Gait Inertial Poser) that naturally integrates into daily life using only two IMUs embedded in shoes.
2) Our model comprises multiple modules that estimate personalized body models from user attributes and accurately predict gait parameters from IMU data. It also incorporates a VAE to ensure smooth motion reconstruction.

3) The effectiveness and practicality of the proposed method are experimentally validated using publicly available datasets and a demonstration system with commercial IMU devices.

II. RELATED WORKS

A. MOTION SENSING FOR GAIT ANALYSIS

Gait analysis plays a critical role in evaluating and preventing declines in quality of life (QOL) caused by deviations from normal gait patterns [23], [24]. Specifically, gait parameters such as joint angles, velocities, and stride lengths provide quantitative and objective measures that are valuable for medical diagnostics and rehabilitation assessments. Traditional clinical approaches typically rely on subjective assessments, including visual observation by medical professionals or self-reports by patients. However, these subjective methods can lead to inter-observer variability and human error [25], [26], [27].

IGA, which employs optical motion capture (MoCap) systems and force plates, offers objective and highly accurate measurement capabilities [3]. Despite their precision, these systems are restricted to specialized laboratory environments. Consequently, concerns have been raised regarding the potential discrepancy between laboratory-measured gait and natural daily walking patterns due to factors such as the Hawthorne effect—where subjects alter their behavior under observation—and the absence of complex real-world conditions like uneven surfaces or obstacles [4], [5], [6].

To overcome these limitations, recent research has actively explored the use of inertial measurement units (IMUs) for gait analysis [7], [8], [28], [29], [30]. IMUs, characterized by their compact size, lightweight, ease of attachment, and affordability, enable the collection of gait data in natural daily settings across extended periods. Wearable IMU-based systems also facilitate patient-driven, long-term monitoring and remote rehabilitation programs.

Common IMU-based gait analysis methods [9], [10], [11] reconstruct foot trajectories by integrating acceleration data obtained from IMUs attached to the feet. However, the inherent noise in IMU measurements accumulates through double integration, causing significant drift in positional and velocity estimates, especially during prolonged measurements. To mitigate this drift problem, correction methods such as the zero-velocity update (ZUPT)—which resets velocity to zero when the foot is stationary—are widely employed [31], [32], [33].

Furthermore, recent studies have proposed numerous approaches [12], [13], [14], [29], [34] leveraging machine learning techniques to estimate joint angles during various movements, including gait, from IMU data. These approaches typically involve attaching multiple IMUs to the lower limbs and incorporating human body models to achieve precise estimation of body segment positions and motions. While such methods enable more detailed gait analysis, the increase in the number of sensors can make the setup cumbersome and raise invasiveness concerns for users. Thus, minimizing the number of IMU sensors is essential to reduce invasiveness and facilitate long-term usage in daily life.

To address this, our method adopts an approach using only two IMUs embedded in shoes, significantly minimizing invasiveness, and aims to accurately capture natural gait motions during daily activities.

B. FULL-BODY HUMAN MOTION CAPTURE

In gait analysis, considering full-body poses—including arm swing and trunk movements in addition to lower limb motions—enables more comprehensive and accurate assessments and diagnoses.

One of the methods widely used for accurately capturing full-body human poses is the optical MoCap system [35], [36], [37], [38]. This system involves attaching small reflective markers to the subject's body and tracking their movement with multiple infrared cameras. While optical MoCap systems achieve high precision and frame rates, they come with significant constraints, including high costs and the requirement to attach markers to the body. To address these challenges, markerless MoCap methods using RGB cameras or depth cameras have been proposed, significantly reducing the cost and complexity of the setup [39], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53]. However, even these methods face challenges such as the synchronization and calibration of multiple camera systems.

Moreover, with advances in machine learning technology, 3D human pose estimation (HPE) methods based on monocular RGB cameras have also been proposed [54], [55], [56], [57], [58], [59], [60], [61], [62], [63]. These methods allow for relatively accurate pose capture in scenarios where camera pre-calibration or marker attachment is impractical. However, these third-person camera methods still face issues such as occlusion caused by obstacles between the camera and the subject, as well as the limitation that the subject's range of motion is confined within the camera's field of view. In response to these challenges, egocentric pose estimation methods [64], [65], [66], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76], [77], [78], [79], [80], [81] using wearable cameras have been developed, enabling pose estimation over a wide range of user activities. However, due to the nature of wearing the camera on the body, there are inherent difficulties, such as the need to use special fisheye cameras and the possibility that large portions of the body may fall outside the field of view depending on the pose. Additionally, a common issue across all camera-based methods is the concern for privacy when measuring movements in various daily life scenarios.

Motion capture using IMU sensors is considered an effective solution to this problem. Unlike camera-based systems, IMUs are not constrained by occlusions or the field of view of the camera, offering a broader capture range and avoiding privacy issues. However, high-end commercial products [82], [83], [84] require a large number of IMUs to be attached to the body, making them unsuitable for capturing motion in everyday life. Recently, research on full-body pose estimation using a small number of wearable sensors, leveraging machine learning techniques, has been advancing [16], [17], [18], [85], [86], [87], [88], [89], [90], [91]. For example, while traditional commercial products like Xsens require 17 IMUs, these new methods propose innovative approaches that estimate full-body motion with only 6 IMUs. Additionally, methods aiming to minimize the number of IMUs have been proposed [15], [92], [93]. These approaches estimate upper-body or full-body poses using only one to three sensors embedded in everyday items such as mobile phones, earbuds, watches, clothing, or shoes.

Some of these methods partially mitigate drift issues by leveraging foot-ground contacts and anatomical or kinematic constraints of the human body. Nevertheless, they still encounter accumulated errors from IMU orientation data in the global coordinate system, as these data are directly input to neural networks, leading to decreased estimation accuracy over prolonged measurements.

In this study, in contrast to previous methods that require IMU orientation data in a global reference frame to estimate diverse motions, we deliberately restrict our target to linear walking motions. This approach allows highly accurate gait motion estimation using solely IMU data within the sensor coordinate system, effectively avoiding errors associated with global coordinate transformations.

III. METHOD

We propose a novel approach, **Gait Inertial Poser (GIP)**, enabling non-invasive and accurate full-body motion estimation using a minimal number of IMU sensors naturally integrated into daily life. This method aims to enable gait analysis in diverse everyday environments beyond the constraints of laboratory settings. As illustrated in Figure 2, our framework consists primarily of three modules: the **Body Module**, the **Gait Module**, and the **Smoothing Module**. (a) First, subject-specific attributes, such as height, weight, age, and gender, along with acceleration and angular velocity data obtained from IMUs, serve as inputs. (b) The Body Module estimates personalized body shape parameters based on these individual attributes. (c) The Gait Module employs a Transformer-based deep learning model to simultaneously estimate multiple gait-related parameters from IMU data, including gait phase, joint angles, root joint velocity, root joint height, and foot height. (d) The Smoothing Module utilizes a Variational Autoencoder (VAE) to enhance the temporal smoothness and continuity of the estimated motion sequences. (e) Integrating these modules yields natural and smooth full-body gait motions.

In the following subsections, we describe the sensor configuration used in the proposed method, detailed network architecture including each module, and the loss functions employed.

A. SENSOR CONFIGURATION

IMUs are widely used for gait analysis and full-body MoCap due to their compatibility with human motion measurement, stemming from their wearability, low cost, high sampling rate, and real-time capabilities. An IMU consists of an accelerometer, gyroscope, and magnetometer, which together allow for tracking the velocity, position, and orientation of the body part to which the sensor is attached. The accelerometer measures the linear acceleration of an object along three axes, with a high capability to detect minute changes in acceleration. However, when acceleration data is double-integrated to determine position, noise and bias can accumulate, resulting in drift. The gyroscope measures the rotational velocity of an object along three axes and can respond immediately to dynamic movements to measure angular velocity, but it is prone to bias accumulation over time. The magnetometer detects the orientation of an object by measuring the Earth's magnetic field, but it is highly sensitive to magnetic disturbances (e.g., interference from electronic devices or metal), which can cause significant drift compared to the other two sensors. Therefore, in this study, we exclude magnetometer data and use the accelerometer and gyroscope values as inputs to the network, addressing their potential drift issues through network and loss function design.

Furthermore, sensor placement is also a critical factor in gait motion measurement. To capture natural gait patterns during everyday life, it is necessary to select a non-invasive sensor configuration that does not interfere with users' daily activities. The widely used commercial IMU sensor system by Xsens [83] requires attaching 17 IMUs across the body for accurate full-body motion capture, making it unsuitable for long-term measurement in daily life or public settings. Recent advancements in machine learning-based pose estimation methods using IMU data have enabled diverse motion estimation with fewer IMUs, such as six sensors placed on the head, waist, hands, and legs [16], [17], [18], [86], [87], [88], [89], [90], [91], or just three sensors located on the wrist (smartwatch), pocket (smartphone), and head (earbuds) [19], [92]. In contrast, this study adopts an even more practical sensor configuration for daily use specifically for gait analysis, utilizing only two IMU sensors embedded in shoe soles. As smart insoles and smart shoes are becoming commercially available, embedding IMU sensors into regular footwear is increasingly feasible, enabling non-invasive, long-term monitoring of everyday movements without imposing additional user burden.

In this study, the acceleration data are denoted as $\boldsymbol{A} = \{\boldsymbol{A}^t\}_{t=0}^{T}$, and angular velocity data as $\boldsymbol{R} = \{\boldsymbol{R}^t\}_{t=0}^{T}$, where t represents the t-th frame in a sequence of length T. Given that our method uses two IMUs embedded in the shoe soles,

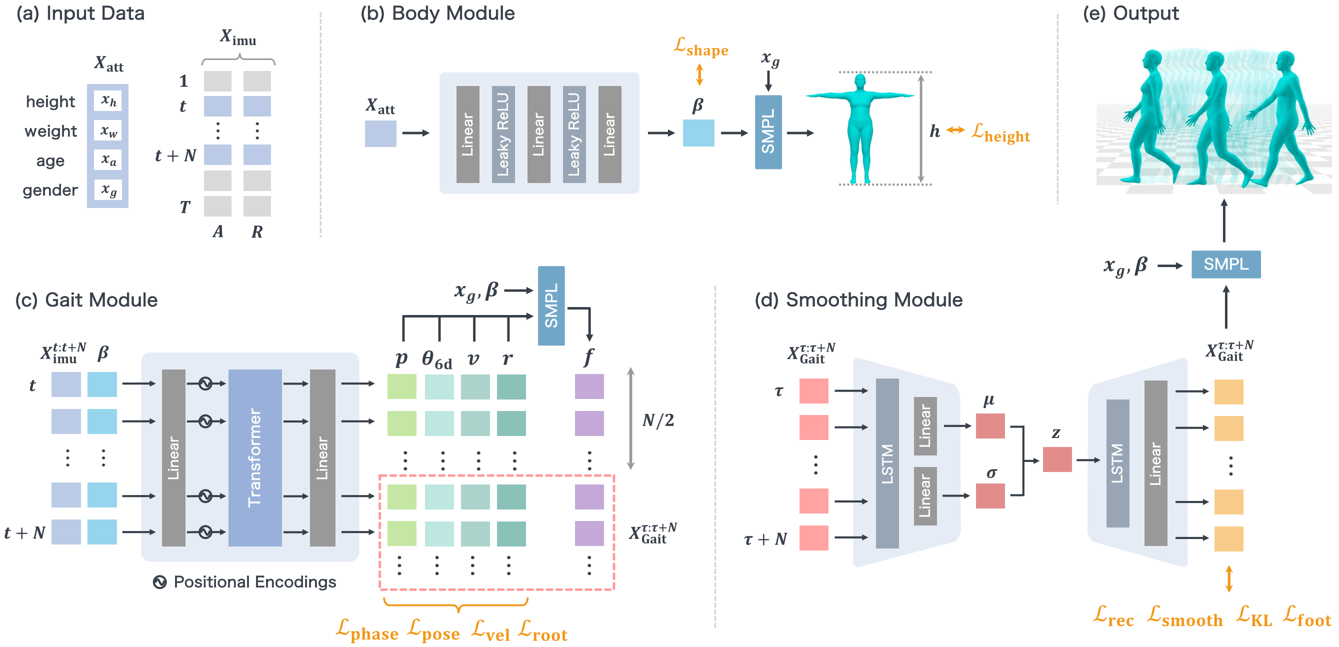


FIGURE 2. Detailed network architecture of the proposed method. (a) Input includes user attributes (height, weight, age, gender) and IMU measurements (acceleration, angular velocity). (b) The Body Module estimates personalized body shape parameters. (c) The Gait Module, based on a Transformer, simultaneously estimates gait phase, joint angles, velocities, root positions, and foot trajectories using IMU data. (d) The Smoothing Module employs a Variational Autoencoder (VAE) to enhance temporal smoothness and stability of the estimated motion, effectively reducing IMU drift through loss functions inspired by zero velocity updates (ZUPT). (e) Final output is a smooth, stable full-body walking motion reconstruction.

the data at each frame are represented as $\boldsymbol{A}^t \in \mathbb{R}^{2\times3}$ and $\boldsymbol{R}^t \in \mathbb{R}^{2\times3}$, respectively.

B. NETWORK DETAILS

The proposed method, **Gait Inertial Poser (GIP)**, is a deep learning-based framework designed for full-body motion estimation using a minimal number of IMU sensors. As illustrated in Figure 2, the network comprises three main modules: the **Body Module**, the **Gait Module**, and the **Smoothing Module**. The Body Module takes attribute data $\boldsymbol{X}_{\text{att}}$ as input and estimates the body shape parameters β of a parametric human mesh model. The Gait Module uses IMU data $\boldsymbol{X}_{\text{imu}}$ (acceleration $\boldsymbol{A}$ and angular velocity $\boldsymbol{R}$) and body shape parameters β to estimate multiple gait parameters simultaneously, including the gait phase $\boldsymbol{p}$, 6D joint angle representation $\theta_{6\text{d}}$, root joint velocity (walking speed) $\boldsymbol{v}$, root joint height $\boldsymbol{r}$, and foot height $\boldsymbol{f}$. The Smoothing Module takes the gait parameters $\boldsymbol{X}_{\text{Gait}}^{\tau:\tau+N}$ estimated by the Gait Module using a sliding-window approach as input, incorporating overlapping frames from the subsequent window based on stride $N/2$ to reduce discontinuities at window boundaries. Subsequently, a Variational Autoencoder (VAE) is employed to enhance the temporal smoothness of the estimated motion sequences. Details of each module are provided in the following subsections.

1) BODY MODULE

The Body Module estimates body shape from simple attribute information without requiring a full-body scan. IMU data do not contain direct information about body shape; thus, IMU-based pose estimation methods typically use a common default body shape model for all subjects. However, differences in individual body shapes affect walking motions, making the use of personalized body models essential for accurate motion reconstruction. In particular, limb lengths influence stride length and IMU measurements. Accurately reproducing these individual differences is crucial for improving the precision of motion estimation.

To represent user-specific body shapes, we adopt the SMPL (Skinned Multi-Person Linear) model [94], a parametric human body model widely utilized in recent pose and shape estimation studies. This model effectively represents complex and diverse human poses and shapes using a limited number of parameters. The SMPL parameters consist of pose parameters $\theta \in \mathbb{R}^{24\times3}$, describing the relative rotations of 23 joints and the global rotation of the root joint, and shape parameters $\beta \in \mathbb{R}^{10}$, reflecting individual characteristics such as height, weight, and limb proportions. Using these parameters, the SMPL regression model estimates triangular mesh vertices $\boldsymbol{M} \in \mathbb{R}^{6890\times3}$ and 3D joint positions $\boldsymbol{J} \in \mathbb{R}^{24\times3}$. Additionally, it incorporates the parameter $\boldsymbol{d} \in \mathbb{R}^3$ to capture the global translation of the person.

The Body Module estimates the body shape parameters β from subject attributes $\boldsymbol{X}_{\text{att}} = \{x_h, x_w, x_a, x_g\}$ (height, weight, age, and gender) using a multi-layer perceptron (MLP). The SMPL model includes three variants: male, female, and neutral, and the appropriate variant is selected based on the gender attribute. The shape parameters β form a 10-dimensional vector comprising the top 10 principal components extracted via principal component analysis from a large-scale body scan dataset. By inputting these estimated body shape parameters along with IMU data into subsequent modules, our approach achieves highly accurate motion estimation that accounts for individual body differences.

2) GAIT MODULE

The Gait Module simultaneously estimates multiple gait-related parameters such as gait phases, joint angles, walking speed, and foot trajectories from the body shape parameters and IMU sensor data. The network structure employs a Transformer-based deep learning model, effectively capturing temporal variations in IMU data for highly accurate motion estimation.

Human gait is characterized as periodic and repetitive movements of body segments, commonly described through "gait phases." Incorporating gait phases facilitates a deeper understanding and analysis of the periodic gait mechanism. The gait cycle primarily consists of two phases: the stance phase, which accounts for approximately 62% of the cycle when the foot contacts the ground, and the swing phase, comprising about 38% when the foot is off the ground. For detailed analysis, the stance phase can be further divided into initial contact, loading response, mid-stance, terminal stance, and pre-swing phases. Similarly, the swing phase can be subdivided into initial swing, mid-swing, and terminal swing. This study specifically focuses on four key phases segmented by the following gait events: heel contact, toe contact, heel-off, and toe-off. The Gait Module identifies these gait phases using the input IMU sequential data and outputs the corresponding gait phase logits $\boldsymbol{p} \in \mathbb{R}^4$. The final gait phase is determined by selecting the maximum value from these logits.

In addition to gait phases, accurate gait motion estimation requires consideration of kinematic parameters such as joint movements, walking speed, and foot trajectories. Therefore, the module leverages the body shape parameter β estimated by the Body Module to simultaneously predict gait phase $\boldsymbol{p}$, 6-dimensional joint angle representation θ_{6d}, root joint velocity $\boldsymbol{v}$, and root joint height $\boldsymbol{r}$, based on the SMPL human body model. The 6D representation (θ_{6d}) is a continuous and rotation-invariant representation of joint angles [95], widely used in pose estimation tasks to mitigate issues like discontinuity and ambiguity inherent in other representations such as Euler angles or axis-angle. Parameters such as walking speed and stride length are particularly influenced by subject-specific body attributes like limb length, thus improving estimation accuracy through incorporating individual body shape parameters. Furthermore, the Transformer architecture is specifically designed to extract temporal features, ensuring stable temporal estimations of these parameters, inspired by prior studies [17], [96] that adopted similar architectures for pose estimation from IMU data. Subsequently, by solving forward kinematics (FK) using a pre-trained SMPL joint estimation model with the predicted joint angles and body shape parameters, the module obtains full-body joint positions and derives foot trajectories $\boldsymbol{f}$. The estimated gait parameters are then input into the subsequent Smoothing Module for further motion refinement.

3) SMOOTHING MODULE

The Smoothing Module is designed to reduce temporal discontinuities in the motion parameter sequences estimated by the Gait Module, resulting in smooth and natural motion estimation. Because the Gait Module performs motion estimation using a sliding-window approach, discontinuities may arise at the boundaries between estimation windows. To address this issue, the Smoothing Module processes overlapping frames between adjacent windows by using a stride $N/2$, enabling continuous correction of motion near these boundaries.

Specifically, when utilizing the gait parameter sequence $\boldsymbol{X}_{\text{Gait}}^{\tau:\tau+N}$ estimated by the Gait Module, frames from the latter half of each estimation window overlap with frames from the first half of the subsequent window, thereby seamlessly connecting otherwise discontinuous estimations. Moreover, we introduce a loss function inspired by the classical Zero Velocity Update (ZUPT) algorithm, a widely used technique in IMU-based gait analysis that reduces integration-induced positional drift by resetting sensor velocity to zero during foot-ground contact. By leveraging the gait phase information estimated by the Gait Module, our proposed loss function explicitly imposes kinematic constraints on foot velocity and height during identified foot-ground contact periods, thereby effectively mitigating IMU drift and improving motion estimation accuracy. The details of this loss function are described in subsequent sections.

The module employs a network architecture based on a Variational Autoencoder (VAE). Sequential input data are first processed by an LSTM to estimate the mean μ and standard deviation σ of latent variables, from which a latent representation z is sampled. This latent representation is then fed into a decoder to reconstruct a continuous and smooth motion sequence. This design ensures temporal stability in the estimated motions, resulting in smooth gait patterns well-suited for practical applications in real-world environments.

C. LOSS FUNCTIONS

Our network architecture allows each module to be trained separately, ensuring stable learning without mutual interference and facilitating efficient convergence. Below, we detail the loss functions employed for individual module training.

1) BODY MODULE

The loss function $\mathcal{L}_{\text{body}}$ used for training the Body Module consists of two terms: a shape loss $\mathcal{L}_{\text{shape}}$ and a height loss $\mathcal{L}_{\text{height}}$, as described in Equation 1. Here, λ_{shape} and λ_{height} are weighting coefficients used as hyperparameters to balance each loss component effectively.

$$\mathcal{L}_{\text{body}} = \lambda_{\text{shape}}\mathcal{L}_{\text{shape}} + \lambda_{\text{height}}\mathcal{L}_{\text{height}} \tag{1}$$

$$\mathcal{L}_{\text{shape}} = |\beta - \hat{\beta}|_1 \tag{2}$$

$$\mathcal{L}_{\text{height}} = |\boldsymbol{h} - \hat{\boldsymbol{h}}|_2^2 \tag{3}$$

The shape loss $\mathcal{L}_{\text{shape}}$ calculates the mean absolute error (MAE) between the estimated shape parameters β and the ground truth $\hat{\beta}$. Since each dimension of β corresponds to principal components obtained via PCA on body shape data, dimensions may have different variance scales. Therefore, using MAE ensures stable training by preventing disproportionately large contributions from dimensions with higher variance.

Although $\mathcal{L}_{\text{shape}}$ effectively optimizes multiple aspects such as limb length and body thickness, height is particularly crucial in gait analysis due to its significant influence on stride length and joint angle estimations. Thus, we introduce an additional height-specific loss term, $\mathcal{L}_{\text{height}}$, which explicitly guides the model to more accurately predict body height. Here, body height is defined as the distance from the soles of the feet to the top of the head when applying the estimated and ground truth shape parameters (β and $\hat{\beta}$) to the SMPL model in the default T-pose. This loss term calculates the mean squared error (MSE) between the predicted height $\boldsymbol{h}$ and the ground truth height $\hat{\boldsymbol{h}}$.

In practice, the height loss is computed by generating a body mesh using the estimated β parameters with gender-specific SMPL models (male or female) and measuring the vertical distance between the vertex at the top of the head and the midpoint of the left and right heel vertices. This calculated height is then compared with the ground truth to compute the height loss. This formulation ensures accurate and stable estimation of individual-specific body shapes, contributing to precise gait analysis.

2) GAIT MODULE

The loss function $\mathcal{L}_{\text{gait}}$ for training the Gait Module combines multiple components, each measuring the accuracy of different gait-related parameters:

$$\mathcal{L}_{\text{gait}} = \lambda_{\text{phase}}\mathcal{L}_{\text{phase}} + \lambda_{\text{pose}}\mathcal{L}_{\text{pose}} + \lambda_{\text{ori}}\mathcal{L}_{\text{ori}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{root}}\mathcal{L}_{\text{root}}, \tag{4}$$

where each λ represents a hyperparameter balancing the contribution of each component.

The gait phase loss $\mathcal{L}_{\text{phase}}$ evaluates the accuracy of gait phase estimation through cross-entropy loss for a multi-class classification task, defined as follows:

$$\mathcal{L}_{\text{phase}} = -\frac{1}{2T}\sum_{t=1}^{T}\left(\log p_{\text{left},t}^{(\hat{y}_{\text{left},t})} + \log p_{\text{right},t}^{(\hat{y}_{\text{right},t})}\right), \tag{5}$$

where $p_{\text{left},t}^{(c)}$ and $p_{\text{right},t}^{(c)}$ represent the predicted probabilities for class c of the left and right feet, respectively, at frame t. The ground truth labels for the left and right feet at frame t are denoted as $\hat{y}_{\text{left},t}$ and $\hat{y}_{\text{right},t}$, respectively, each indicating one of the classes. T denotes the total number of frames.

The remaining losses (pose, orientation, velocity, and root height) are computed using mean squared error (MSE) between predicted and ground truth values:

$$\mathcal{L}_{\text{pose}} = \frac{1}{TK}\sum_{t=1}^{T}\sum_{k=1}^{K}\|\theta_t^k - \hat{\theta}_t^k\|_2^2, \tag{6}$$

$$\mathcal{L}_{\text{ori}} = \frac{1}{T}\sum_{t=1}^{T}\|\theta_t^0 - \hat{\theta}_t^0\|_2^2, \tag{7}$$

$$\mathcal{L}_{\text{vel}} = \frac{1}{T}\sum_{t=1}^{T}\|\boldsymbol{v}_t - \hat{\boldsymbol{v}}_t\|_2^2, \tag{8}$$

$$\mathcal{L}_{\text{root}} = \frac{1}{T}\sum_{t=1}^{T}\|\boldsymbol{r}_t^{(z)} - \hat{\boldsymbol{r}}_t^{(z)}\|_2^2, \tag{9}$$

where K is the number of joints, θ_t^k represents the estimated 6-dimensional joint angle representation for the k-th joint at frame t, θ_t^0 is the estimated 6-dimensional orientation of the root joint at frame t, $\boldsymbol{v}_t$ denotes the estimated root joint velocity (walking speed) at frame t, and $\boldsymbol{r}_t^{(z)}$ indicates the estimated root joint height at frame t. Corresponding ground truth values are denoted with hats $(\hat{\cdot})$. By integrating these multiple loss components, the Gait Module accurately estimates the diverse and crucial parameters essential for detailed gait analysis.

3) SMOOTHING MODULE

The loss function for the Smoothing Module, $\mathcal{L}_{\text{smooth}}$, is designed based on a Variational Autoencoder (VAE) framework and comprises multiple components. These components include reconstruction loss, smoothing loss, translation loss, foot-ground contact constraints, and the KL divergence term. Specifically, it is defined as follows:

$$\mathcal{L}_{\text{smooth}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}} + \lambda_{\text{fpos}}\mathcal{L}_{\text{fpos}} + \lambda_{\text{fvel}}\mathcal{L}_{\text{fvel}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}}, \tag{10}$$

where each λ represents a weighting hyperparameter to balance the contribution of each term.

The reconstruction loss $\mathcal{L}_{\text{rec}}$ is calculated as the Mean Squared Error (MSE) between the predicted gait parameters (joint angles, orientations, velocities, root joint height, foot joint height) and their ground truth values, ensuring accurate reproduction of the input motion sequences:

$$\mathcal{L}_{\text{rec}} = \frac{1}{TD}\sum_{t=1}^{T}\sum_{d=1}^{D}\|z_{t,d} - \hat{z}_{t,d}\|_2^2, \tag{11}$$

where $z_{t,d}$ and $\hat{z}_{t,d}$ denote the predicted and ground truth motion parameters at frame t and dimension d, respectively. D represents the total number of gait parameter dimensions.

The smoothing loss $\mathcal{L}_{\text{smooth}}$ is defined by minimizing the second-order temporal derivative of the predicted motion parameters. This encourages temporal smoothness and continuity in motion estimations:

$$\mathcal{L}_{\text{smooth}} = \frac{1}{TD} \sum_{t=2}^{T-1} \sum_{d=1}^{D} \|\ddot{\boldsymbol{x}}_{t,d}\|_2^2, \tag{12}$$

where the second derivative term $\ddot{\boldsymbol{x}}_t$ is approximated via finite differences as follows:

$$\ddot{\boldsymbol{x}}_t \approx \frac{\boldsymbol{x}_{t+1} - 2\boldsymbol{x}_t + \boldsymbol{x}_{t-1}}{\Delta t^2}. \tag{13}$$

Here, $\boldsymbol{x}_t \in \mathbb{R}^{K\times 3}$ denotes 3D joint positions at frame t, which are obtained via the forward kinematics function $\mathrm{FK}(\beta, \theta_t, \boldsymbol{r}_t)$ using the pre-trained SMPL joint regressor.

Moreover, inspired by the classical Zero Velocity Update (ZUPT) algorithm, we introduce constraints during periods identified as foot-ground contact phases, leveraging gait phase information $\boldsymbol{p}$ from the Gait Module. Specifically, we define two loss functions: a foot height loss, $\mathcal{L}_{\text{fpos}}$, and a foot velocity loss, $\mathcal{L}_{\text{fvel}}$.

The foot height loss $\mathcal{L}_{\text{fpos}}$ minimizes the vertical displacement of foot sole vertices, promoting accurate foot-ground contact constraints:

$$\mathcal{L}_{\text{fpos}} = \frac{1}{|\mathcal{C}|} \sum_{t\in\mathcal{C}} \|\boldsymbol{f}_t^{(z)}\|_2^2, \tag{14}$$

where $\boldsymbol{f}_t^{(z)}$ is the predicted vertical position of the foot sole at frame t, and $\mathcal{C}$ is the set of foot-ground contact frames and vertices identified by gait phase predictions.

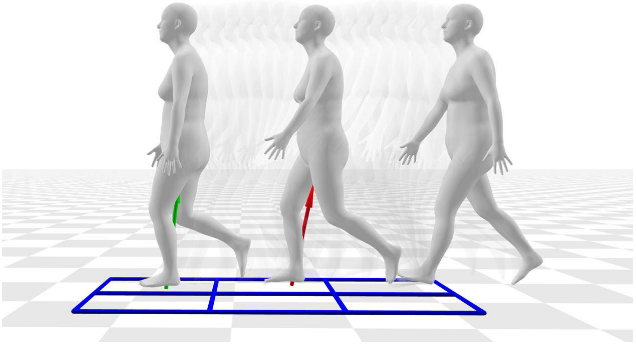


FIGURE 3. Example of the gait motion data from the AIST Gait Database. Subjects walked naturally along a straight path, stepping on force plates (indicated in blue), allowing simultaneous capture of motion and ground reaction force (GRF) data (indicated in red and green). The GRF data were used to define gait phases such as heel-strike and toe-off.

Similarly, the foot velocity loss $\mathcal{L}_{\text{fvel}}$ suppresses horizontal and vertical foot movements during identified ground-contact periods:

$$\mathcal{L}_{\text{fvel}} = \frac{1}{|\mathcal{C}|} \sum_{i\in\mathcal{C}} \|\dot{\boldsymbol{f}}_i\|_2^2, \tag{15}$$

where $\dot{\boldsymbol{f}}_i$ denotes the velocity of foot sole at frame t.

Lastly, the KL divergence term $\mathcal{L}_{\text{KL}}$ regularizes the latent variable distribution to approximate a standard normal distribution, following the standard VAE objective:

$$\mathcal{L}_{\text{KL}} = -\frac{1}{2} \sum_{l=1}^{L} \left(1 + \log \sigma_l^2 - \mu_l^2 - \sigma_l^2\right), \tag{16}$$

where L is the dimensionality of the latent space, and μ_l, σ_l^2 represent the mean and variance of the latent distribution, respectively.

This combined loss function ensures accurate reconstruction, temporal smoothness, robust velocity integration, effective drift suppression during foot contact, and stable latent representations, thereby significantly enhancing the accuracy and smoothness of gait motion estimation.

IV. EXPERIMENTS

A. DATASETS

Our proposed approach is a data-driven motion estimation framework utilizing machine learning models. Thus, in this study, we used the following two existing gait datasets for model training and evaluation.

The first dataset is the "AIST Gait Database 2019" [20], published by the National Institute of Advanced Industrial Science and Technology (AIST). This database comprises gait motion data collected according to protocols approved by the AIST Human Ergonomics Experiment Committee. After excluding data with confirmed measurement errors, the dataset consists of 3,424 walking trials conducted by a total of 352 healthy adults, including 1,609 trials from 167 males and 1,815 trials from 185 females. The participants span a wide demographic range, with ages from 20 to 78 years (mean 51.1 ± 18.7), heights from 138 to 185 cm (mean 162.9 ± 8.4), and body weights from 34 to 100 kg (mean 59.5 ± 10.2). The data were captured at 200 FPS, amounting to a total of 1 million frames. Measurements were performed using a Vicon optical MoCap system, with subjects instructed to walk naturally along a straight line at their preferred walking speed. Additionally, the dataset includes ground reaction force (GRF) data captured using force plates. During data collection, participants walked over force plates located near the center of the laboratory setup, allowing simultaneous recording of motion data and GRF data (Figure 3).

Gait phase labels used for training our method were derived using this GRF data. The gait cycle was segmented into four phases based on gait events: "heel-strike," "toe-contact," "heel-off," and "toe-off." Heel-strike and toe-off timings were defined based on the initiation and cessation of ground reaction forces, respectively. In contrast, toe-contact and heel-off phases were labeled when the velocity of specific mesh vertices on the foot soles fell below a predefined threshold. This dataset is highly suitable for evaluating gait estimation methods due to the diversity of its subjects. However, because the dataset relies on force plates, each trial contains only a very short sequence of a few seconds and does not include actual IMU sensor measurements.

To address these limitations, we supplemented the evaluation using synthesized IMU data (described later) and also utilized another dataset, which contains longer gait sequences and actual IMU measurement data.

The second dataset used is the "UnderPressure dataset" [21], originally created to address the foot-skating issue common in computer graphics-based motion estimation tasks. This dataset contains motion data recorded from 10 participants wearing insoles equipped with IMU sensors and foot-pressure sensors. The participants consist of 8 males and 2 females, aged 21–55 years (mean 31.4 ± 11.7), with heights ranging from 167 to 184 cm (mean 176.4 ± 7.7 cm) and weights from 65 to 91 kg (mean 77.9 ± 9.3 kg). Each participant performed motion sequences with an average duration of about 1.5 minutes. It includes diverse movements such as walking at slow, medium, and fast speeds, walking with random directional changes, interacting with objects, and stair climbing. Since our experiments focus on straight-line gait motions, we excluded sequences involving interactions with objects and extracted segments where the yaw angular velocity of the root joint remained below a specified threshold. The total duration of these extracted segments amounted to 170K frames (28.5 minutes at 100 FPS). For gait phase labeling, we defined the four phases mentioned above based on pressure detection using the 16 foot-pressure sensors per foot, divided into front and rear sections.

B. DATA SYNTHESIS

Creating extensive motion capture datasets requires significant effort. Consequently, many IMU-based approaches typically employ synthesized IMU data in addition to real data for training and evaluation. Following this common practice, we generated synthetic acceleration and angular velocity data for cases in the AIST Gait Database and UnderPressure datasets where actual IMU data was unavailable.

Accurate synthesis of IMU data requires reproducing not only skeletal poses but also individual-specific body shapes. Therefore, we generate mesh models as pseudo-ground-truth data to represent the actual body shape of each subject. For the AIST Gait Database, we reproduced individual body shapes using "Mosh++" [97], [98], an optimization method that fits the SMPL human mesh model to optical marker trajectory data. Prior to optimization, we manually established correspondences between optical marker positions and SMPL mesh vertices to accurately generate subject-specific mesh models. In contrast, the UnderPressure dataset contains motion data recorded using commercial high-performance IMU sensors (Xsens) without optical markers. Therefore, we fitted the SMPL model by solving an inverse kinematics (IK) algorithm based on skeletal landmarks common to both the Xsens model and the SMPL model.

Next, we synthesized IMU data from the obtained SMPL mesh models. Specifically, to evaluate our proposed method and baseline methods described later, we created synthetic IMU data for five locations: the left wrist, left ear, right pocket, and both foot soles, as shown in Figure 4. The IMU coordinate systems were defined based on selected triangular mesh faces by using the vectors of triangle edges, the normal vector to the triangle face, and their cross product. Using this IMU coordinate system, we computed the orientation and acceleration in the global coordinate system during motions and subsequently transformed these data into synthetic acceleration and angular velocity data in the IMU sensor coordinate system.

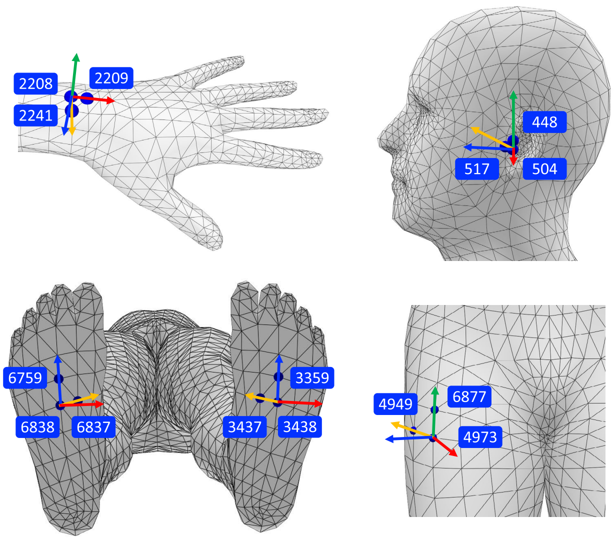


FIGURE 4. Synthetic IMU sensor positions defined on SMPL mesh vertices (blue labels indicate SMPL vertex IDs). The foot soles correspond to the proposed method, while positions on the left wrist, left ear, and right thigh simulate IMU signals for baseline methods (smartwatch, earphone, smartphone in pocket). The yellow arrows indicate vectors used to define the mesh face normal for establishing IMU coordinate systems.

C. IMPLEMENTATION DETAILS

We implemented our proposed and baseline methods using PyTorch as the deep learning framework. Model training and evaluation were conducted on a Linux server equipped with an NVIDIA RTX A6000 GPU. Specifically, the Body Module consisted of a three-layer multi-layer perceptron (MLP) with hidden layers of 128, 256, and 128 units, respectively. The Gait Module employed a Transformer-based encoder with 8 attention heads and 4 encoder layers, a latent dimension of 256, and an input window size of 50 frames. The Smoothing Module was implemented as a Variational Autoencoder (VAE) with a latent dimension of 64 and an LSTM dimension of 256.

For training, we employed the Adam optimizer [99] with an initial learning rate of 1×10^{-3}, trained for 500 epochs. The dataset was divided into training, validation, and test subsets with a ratio of 8:1:1, and a mini-batch size of 32 was used. Model training and testing were conducted at 200 fps for the AIST Gait Database and at 100 fps for the UnderPressure dataset. Loss function weights were empirically optimized through preliminary experiments and set as follows:

$$(\lambda_{\text{shape}}, \lambda_{\text{height}}) = (0.1, 100),$$
$$(\lambda_{\text{phase}}, \lambda_{\text{pose}}, \lambda_{\text{ori}}, \lambda_{\text{vel}}, \lambda_{\text{root}}) = (1, 100, 100, 10, 10),$$
$$(\lambda_{\text{rec}}, \lambda_{\text{smooth}}, \lambda_{\text{fpos}}, \lambda_{\text{fvel}}, \lambda_{\text{KL}}) = (1000, 1 \times 10^5, 1, 0.1).$$

To identify foot-ground contact phases in the AIST Gait Database, we set the velocity threshold of the toes and heels on the foot sole to 0.1 m/s. For the UnderPressure dataset, straight-walking segments were extracted by analyzing the root joint velocity direction. The method compares the mean horizontal heading over 15-past and future windows and labels frames as non-straight when the heading change exceeds 2°, also excluding a margin of neighboring 10 frames—all parameter values were empirically determined. During inference, the root joint trajectory was calculated by integrating the estimated root joint velocity, starting from the origin on the horizontal plane, combined with the estimated root joint height.

Finally, following Zhang et al. [17], we applied low-pass filtering to both synthetic and real IMU data to reduce differences in waveform characteristics. Specifically, a 4th-order Butterworth low-pass filter with a cutoff frequency of 10 Hz and a sampling frequency of 100 Hz was employed.

D. BASELINE METHODS

To evaluate the performance of our proposed method, we employed the following four baseline methods for comparison:

- **Integration:** The simplest baseline method, which estimates trajectories by directly integrating acceleration and angular velocity data from shoe-embedded IMU sensors. The initial position and orientation are assumed to be known. However, it is widely recognized that this method suffers from cumulative integral drift, significantly degrading accuracy over time.
- **Integration + ZUPT:** An enhanced integration method that incorporates Zero-Velocity Update (ZUPT) to mitigate drift. Using gait-phase information estimated by our proposed method, foot-ground contact periods are identified. In addition to setting velocities to zero during ground contacts, the method also removes accumulated velocity drift between successive ground contacts. Specifically, drift correction is applied by linearly interpolating velocities between the start of each ground contact and the end of the next, subtracting the interpolated drift from the original velocity data. Furthermore, positional drift along the vertical axis is corrected by adjusting trajectories so that the foot positions align with the ground plane at the beginning and end of each swing phase.
- **IMUPoser** [92]: A pose estimation method utilizing consumer-grade wearable devices such as smartphones, smartwatches, and earbuds. In this experiment, IMU data from three locations (left wrist, right pocket, and left ear) were used. While IMUPoser accurately estimates root-relative body pose and segment movements, it does not estimate the absolute position (translation) of the root joint. Following prior work [89], we employed the Versatile Quaternion-based Filter (VQF), a high-precision orientation estimation method that accounts for gravity, to transform 6-DoF IMU data (acceleration and angular velocity) into orientation and acceleration in the global coordinate system.
- **MobilePoser** [19]: An extension of IMUPoser [92], employing the same IMU sensor configuration but additionally estimating the absolute position (translation) of the root joint. Thus, MobilePoser estimates both pose and overall body translation, enhancing its applicability to a broader range of environments. As with IMUPoser, we utilized VQF to convert IMU data into the global coordinate frame.

E. EVALUATION METRICS

We employed the following metrics to evaluate the performance of the proposed method and baseline methods:

- **Pose-G, Pose-L**: These metrics evaluate the accuracy of pose estimation. Pose-G measures the mean per joint position error (MPJPE), calculated as the average Euclidean distance between the estimated and ground truth (GT) joint positions in global coordinates over all joints and frames. Pose-L, or Pelvis-MPJPE (PEL-MPJPE), evaluates joint errors after aligning the root joint positions of predictions and GT. The unit is centimeters (cm), computed as:

$$\text{MPJPE} = \frac{1}{TK}\sum_{t=1}^{T}\sum_{k=1}^{K}\|\boldsymbol{J}_{t,k} - \hat{\boldsymbol{J}}_{t,k}\|_2 \tag{17}$$

  where T is the number of frames, K is the number of joints, and $\boldsymbol{J}_{t,k}$ and $\hat{\boldsymbol{J}}_{t,k}$ are the predicted and GT joint coordinates for frame t and joint k.
- **Mesh-P, Mesh-T**: These metrics assess SMPL mesh vertex prediction accuracy. They calculate the mean per vertex error (MPVPE), averaging the Euclidean distances between estimated and GT mesh vertices across all vertices and frames. Mesh-P evaluates errors with aligned root coordinates, thus incorporating pose information ("P" for Pose). Since Mesh-P combines shape and pose errors, Mesh-T evaluates the pure shape estimation accuracy by measuring errors at the default T-pose. Units are centimeters (cm), calculated as:

$$\text{MPVPE} = \frac{1}{TQ}\sum_{t=1}^{T}\sum_{q=1}^{Q}\|\boldsymbol{M}_{t,v} - \hat{\boldsymbol{M}}_{t,v}\|_2 \tag{18}$$

  where Q is the number of vertices, and $\boldsymbol{M}_{t,q}$ and $\hat{\boldsymbol{M}}_{t,q}$ represent the predicted and GT vertex coordinates for frame t and vertex q.
- **Joint Angle**: This metric measures joint angle estimation accuracy. Although SMPL uses axis-angle representation, angles are converted to Euler angles

TABLE 1. Performance comparison of foot trajectory and pose estimation methods across different datasets (AIST Gait Database and UnderPressure Dataset) and evaluation metrics. Bold values indicate the highest accuracy achieved by each metric within each dataset.

Dataset	Method	Pose-G [cm] (↓)	Pose-L [cm] (↓)	Mesh-P [cm] (↓)	Mesh-T [cm] (↓)	JointAngle [degree] (↓)	InterFoot [cm] (↓)	FootTraj [cm] (↓)	FootVel [cm/s] (↓)	RootVel [cm/s] (↑)	Phase (↓) [F1 Score]
AIST Gait DB	Integration	-	-	-	-	-	8.25	11.44	18.63	-	-
	Integration + ZUPT	-	-	-	-	-	7.26	8.04	**17.93**	-	**0.96**
	IMUPoser (3IMUs)	-	3.06	3.69	4.40	2.69	5.30	-	-	-	-
	MobilePoser (3IMUs)	9.11	**2.90**	**3.60**	4.40	2.74	4.12	9.59	33.07	19.19	-
	Ours (2IMUs)	**5.73**	3.18	4.36	**1.58**	**2.26**	**3.11**	**5.32**	22.84	**8.25**	**0.96**
UnderPressure	Integration	-	-	-	-	-	24.47	28.20	36.15	-	-
	Integration + ZUPT	-	-	-	-	-	14.83	19.91	37.23	-	**0.89**
	IMUPoser (3IMUs)	-	4.26	4.90	4.37	3.10	8.08	-	-	-	-
	MobilePoser (3IMUs)	28.83	4.14	**4.71**	4.37	3.12	6.99	31.74	44.58	26.89	-
	Ours (2IMUs)	**10.51**	**3.87**	4.84	**2.06**	**2.29**	**4.25**	**12.35**	**26.70**	**8.45**	**0.89**

(ZYX order) for evaluation. Angle differences between predicted and GT values are computed for each axis, joint, and frame, and averaged. Units are degrees (°):

$$\text{Angle Error} = \frac{1}{3TK}\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{a\in\{x,y,z\}} |\theta_{t,k,a} - \hat{\theta}_{t,k,a}| \tag{19}$$

where $\theta_{t,k,a}$ and $\hat{\theta}_{t,k,a}$ denote the predicted and GT joint angles for frame t, joint k, and axis a.

- **InterFoot**: This metric evaluates stride length accuracy by measuring the difference between the predicted and GT distances between feet, averaged over all frames. Units are centimeters (cm):

$$\text{InterFoot Error} = \frac{1}{T}\sum_{t=1}^{T}\left| \|\boldsymbol{f}_t^L - \boldsymbol{f}_t^R\|_2 - \|\hat{\boldsymbol{f}}_t^L - \hat{\boldsymbol{f}}_t^R\|_2 \right|, \tag{20}$$

where $\boldsymbol{f}_t^L, \boldsymbol{f}_t^R$ represent the predicted positions of the left and right feet at frame t, and $\hat{\boldsymbol{f}}_t^L, \hat{\boldsymbol{f}}_t^R$ are the corresponding ground-truth positions.

- **FootTraj**: This metric evaluates the accuracy of foot trajectory estimation by measuring the Euclidean distance between the predicted and ground truth 3D foot positions for both left and right feet. The result is averaged over all frames and both feet. Units are centimeters (cm):

$$\text{FootTraj Error} = \frac{1}{2T}\sum_{t=1}^{T}\sum_{s\in\{\text{left,right}\}} \|\boldsymbol{f}_{t,s} - \hat{\boldsymbol{f}}_{t,s}\|_2 \tag{21}$$

where $\boldsymbol{f}_{t,s}$ and $\hat{\boldsymbol{f}}_{t,s}$ represent the predicted and ground truth positions of foot s (left or right) at frame t.

- **FootVel**: This metric evaluates the accuracy of foot velocity by comparing the temporal derivatives of the predicted and ground truth foot trajectories. The velocity is computed using finite differences, and the error is defined as:

$$\text{FootVel Error} = \frac{1}{2(T-1)}\sum_{t=1}^{T-1}\sum_{s\in\{\text{left,right}\}} \|\dot{\boldsymbol{f}}_{t,s} - \dot{\hat{\boldsymbol{f}}}_{t,s}\|_2, \tag{22}$$

where $\dot{\boldsymbol{f}}_{t,s}$ denotes the velocity of foot s (left or right) at frame t, approximated as $\dot{\boldsymbol{f}}_{t,s} \approx (\boldsymbol{f}_{t+1,s} - \boldsymbol{f}_{t,s})/\Delta t$.

- **RootVel**: This metric evaluates the accuracy of translational velocity of the root joint. Similar to FootVel, velocity is computed using finite differences, and the error is defined as:

$$\text{RootVel Error} = \frac{1}{T-1}\sum_{t=1}^{T-1} \|\dot{\boldsymbol{r}}_t - \dot{\hat{\boldsymbol{r}}}_t\|_2, \tag{23}$$

where $\dot{\boldsymbol{r}}_t$ denotes the root joint velocity at frame t, approximated by $\dot{\boldsymbol{r}}_t \approx (\boldsymbol{r}_{t+1} - \boldsymbol{r}_t)/\Delta t$.

- **Phase**: This metric evaluates the accuracy of gait phase classification using the macro-averaged F1 score across the four gait phases. At each frame, the predicted gait phase label is compared with the ground truth label, and the F1 score is computed for each class individually across all frames. The final metric is the macro-average over all classes:

$$\text{Macro-F1} = \frac{1}{C}\sum_{c=1}^{C} \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}, \tag{24}$$

where $C = 4$ is the number of gait phase classes, and Precision_c and Recall_c represent the precision and recall for class c, calculated across the entire sequence of frames.

F. COMPARISON WITH BASELINE METHODS

In this section, we present quantitative and qualitative evaluation results from comparative experiments conducted using the AIST Gait Database and the UnderPressure dataset to compare our proposed method against baseline methods.

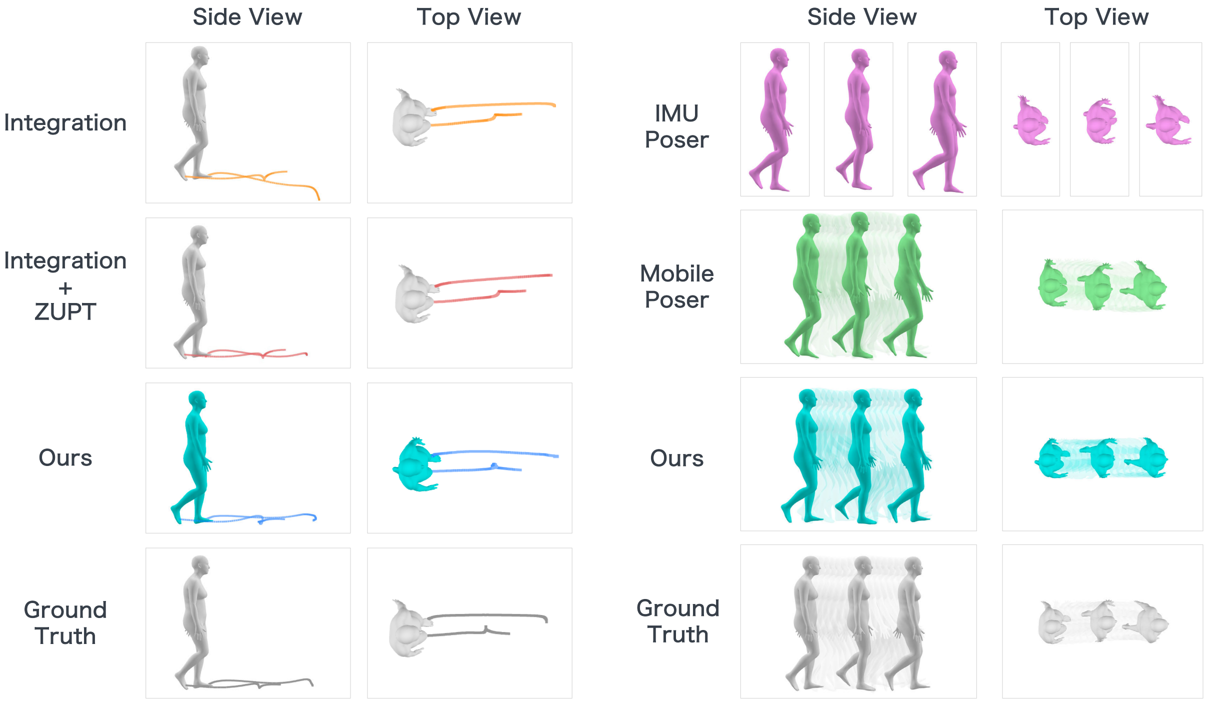


FIGURE 5. Qualitative comparison of foot trajectory (left) and full-body pose estimation (right) results using the AIST Gait DB. The results show side and top views for baseline methods (Integration, Integration + ZUPT, IMUPoser, MobilePoser) and our method (Ours), compared with the Ground Truth. Our method demonstrates accurate and stable trajectory and pose estimations, effectively mitigating drift and reconstructing natural walking motions.

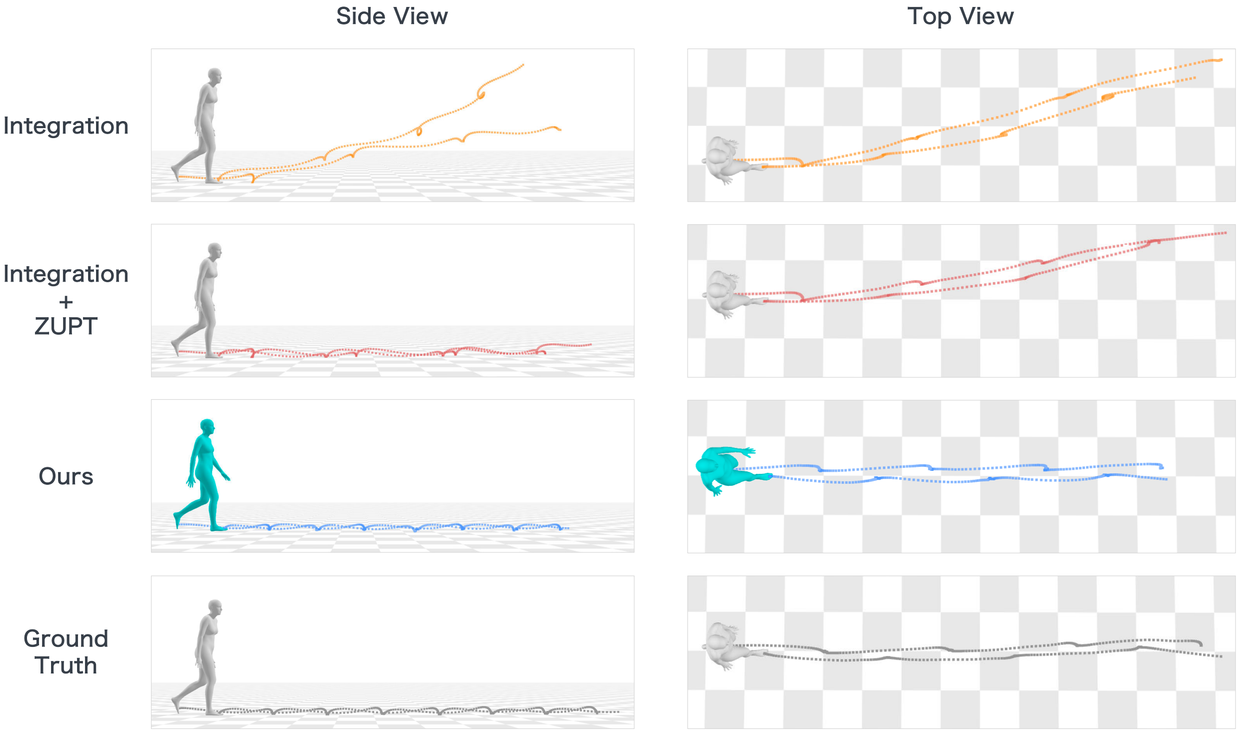


FIGURE 6. Qualitative comparison of foot trajectory estimation results on the UnderPressure. The side and top views illustrate estimated foot trajectories for Integration, Integration + ZUPT, Ours, and the Ground Truth. Our method effectively suppresses drift, accurately reconstructing stable foot trajectories.

1) QUANTITATIVE EVALUATION

Table 1 summarizes the quantitative evaluation results. Firstly, our proposed method exhibited superior performance across both datasets for Pose-G, Mesh-T, and JointAngle metrics, demonstrating its effectiveness in accurately estimating absolute joint positions and reconstructing body shapes.

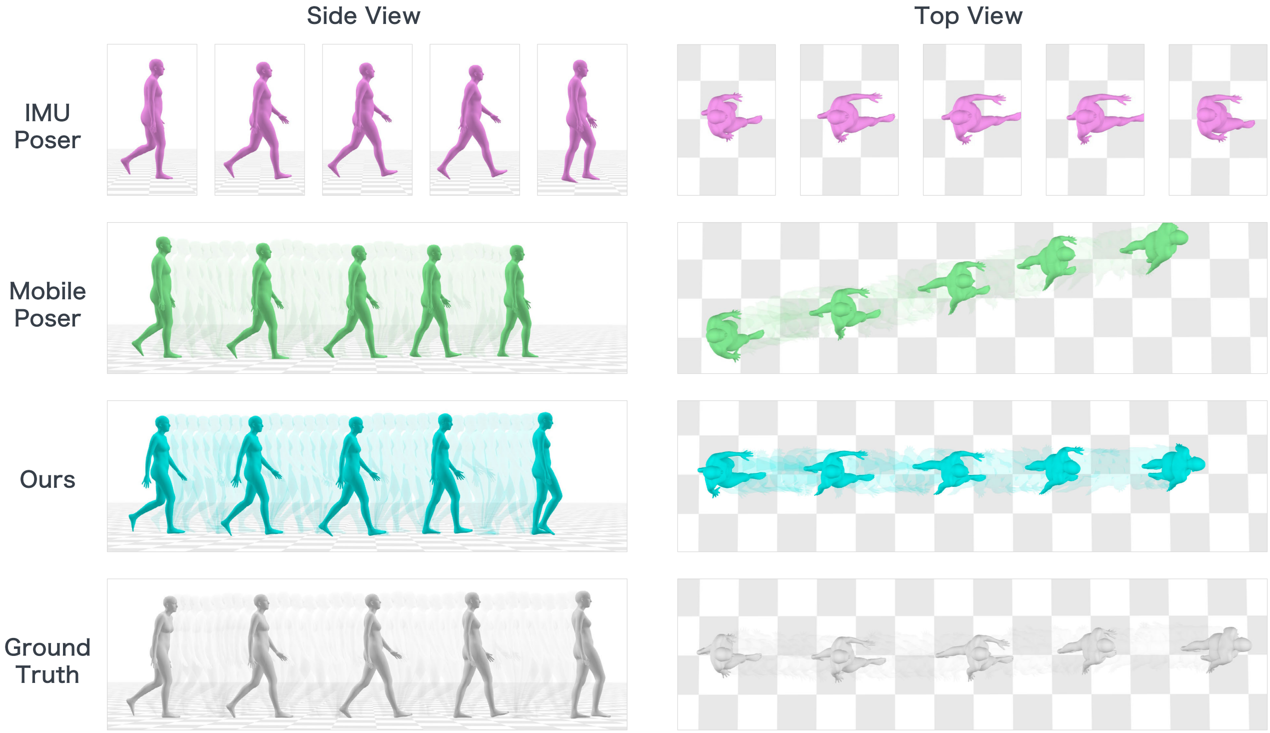


FIGURE 7. Qualitative comparison of full-body pose estimation results on the UnderPressure. Side and top views show results from IMUPoser, MobilePoser, Ours, and the Ground Truth. Compared to baseline methods, our approach achieves more accurate and natural full-body pose reconstructions, demonstrating stable performance even with real IMU data.

The improved Pose-G accuracy compared to MobilePoser is primarily attributed to two factors. First, the estimation accuracy of the root joint velocity (RootVel) during straight walking significantly affects the accuracy of joint positions. Our proposed method estimates velocity more accurately than MobilePoser, resulting in root joint positions that are closer to the ground truth (GT) and consequently reducing the relative joint position error (Pose-G). Second, MobilePoser utilizes IMU data in the global coordinate system, potentially causing drift errors during extended measurements. Consumer-grade devices often include 6DoF IMUs without magnetometers, making global orientation estimation challenging and susceptible to drift over time. Following previous research [89], we applied the Versatile Quaternion-based Filter (VQF) for global coordinate transformations. However, VQF accumulates errors for data recordings longer than several tens of seconds. In contrast, our method utilizes IMU data directly in the sensor coordinate system, thereby significantly mitigating drift issues and maintaining stable accuracy during straight walking motions.

Conversely, our method slightly underperformed IMUPoser and MobilePoser in the Pose-L and Mesh-P metrics. This minor discrepancy likely arises because IMUPoser employs more IMUs (three sensors) compared to our two sensors. However, the observed differences were minimal, on the order of a few millimeters. Furthermore, the strong performance in Mesh-T indicates the efficacy of our method's personalized body shape reconstruction approach, considering individual attributes such as height and weight. Regarding JointAngle evaluation, our method achieved the best performance, although the differences among all methods were minor. This suggests that even with fewer IMUs, accurate pose estimation is feasible for simple motions like walking.

Additionally, our method demonstrated high accuracy across both datasets in metrics like InterFoot (inter-foot distance) and FootTraj (foot trajectory), highlighting its ability to reconstruct personalized body shapes and effectively suppress drift. Regarding FootVel (foot velocity), Integration+ZUPT performed best with the synthetic IMU data of the AIST Gait Database, whereas our method was superior with the real, noisy IMU data from the UnderPressure dataset. This indicates that the quality of the IMU data influences the accuracy of the foot velocity estimation.

For gait phase estimation (F1 score), our method showed strong performance overall, particularly excelling with synthetic IMU data from the AIST Gait Database. This outcome suggests synthetic IMU data more distinctly represents features such as ground contact states.

2) QUALITATIVE EVALUATION

Figures 5, 6 and 7 illustrate the qualitative evaluation results. Figures 5 and 6 shows that the Integration and

TABLE 2. Performance comparison under identical IMU placement (two IMUs on foot soles). Bold indicates the best-performing methods per dataset. "Ours B+G+S" is the full model with Body, Gait, and Smoothing Modules; "Ours G+S" excludes the Body Module; and "Ours B+G" excludes the Smoothing Module, enabling evaluation of the contribution of each module.

Dataset	Method	Pose-G [cm] (↓)	Pose-L [cm] (↓)	Mesh-P [cm] (↓)	Mesh-T [cm] (↓)	JointAngle [degree] (↓)	InterFoot [cm] (↓)	FootTraj [cm] (↓)	FootVel [cm/s] (↓)	RootVel [cm/s] (↓)	Phase [F1 Score] (↑)
AIST Gait DB	IMUPoser (2IMUs)	-	3.40	4.45	4.40	2.83	4.84	-	-	-	-
	MobilePoser (2IMUs)	16.58	3.37	4.39	4.40	2.93	4.37	17.46	55.84	26.10	-
	Ours G+S (2IMUs)	6.93	4.13	5.17	3.10	2.23	3.70	8.48	24.45	8.92	**0.96**
	Ours B+G (2IMUs)	**4.79**	**3.18**	**4.35**	**1.58**	**2.22**	3.32	**4.28**	50.18	8.84	**0.96**
	Ours B+G+S (2IMUs)	5.73	**3.18**	4.36	**1.58**	2.26	**3.11**	5.32	**22.84**	**8.25**	**0.96**
UnderPressure	IMUPoser (2IMUs)	-	7.36	9.22	4.37	3.79	5.92	-	-	-	-
	MobilePoser (2IMUs)	47.36	6.67	8.18	4.37	3.63	6.39	49.60	42.97	34.02	-
	Ours G+S (2IMUs)	11.38	4.60	5.41	2.97	2.38	6.89	15.24	29.33	**8.18**	**0.90**
	Ours B+G (2IMUs)	11.21	4.27	5.37	**2.06**	2.48	4.74	13.28	93.12	10.58	0.89
	Ours B+G+S (2IMUs)	**10.51**	**3.87**	**4.84**	**2.06**	**2.29**	**4.25**	**12.35**	**26.70**	8.45	0.89

Integration+ZUPT methods resulted in unrealistic trajectories due to integration drift, such as floating above the ground or unnatural deviations. While ZUPT correction addressed unnatural vertical deviations and floating, it was insufficient in mitigating lateral drift relative to the walking direction. In contrast, our proposed method generated trajectories remarkably close to the ground truth, confirming stable trajectory estimation. This demonstrates that our method effectively employs anatomical constraints via human body modeling, significantly suppressing drift.

In Figure 7, full-body pose estimations from IMUPoser (without root translation), MobilePoser (with root translation), and our method (with root translation) were compared. All methods performed reasonably well in pose estimation; however, MobilePoser displayed noticeable directional errors, causing lateral drift in trajectories. Conversely, our method accurately reproduced the actual walking direction and trajectories, demonstrating superior directional estimation performance. This advantage arises because our method uses IMU data directly in the sensor coordinate system, thereby avoiding cumulative errors typically encountered with global coordinate transformations, particularly evident during extended measurements such as those captured by the UnderPressure dataset.

Overall, the quantitative and qualitative evaluations validate that our proposed method, utilizing only two IMUs placed on the foot soles, can achieve highly accurate gait motion estimation compared to existing methods. Clear advantages include effective drift suppression and stable estimation of walking direction, speed, and pose.

3) RELATIONSHIP BETWEEN JOINT POSITION ERROR AND IMU PLACEMENT

Figure 8 illustrates the distribution of joint position errors for each method in the UnderPressure dataset. Observing the heatmaps at the top, we can clearly identify a general trend: the joint position error is smaller in areas where IMUs are attached (indicated by pink stars), and increases as the distance from IMU attachment sites grows. This trend is consistent across all methods (IMUPoser, MobilePoser, and Ours), demonstrating that the direct motion information obtained from IMU-attached areas significantly improves joint position estimation accuracy.

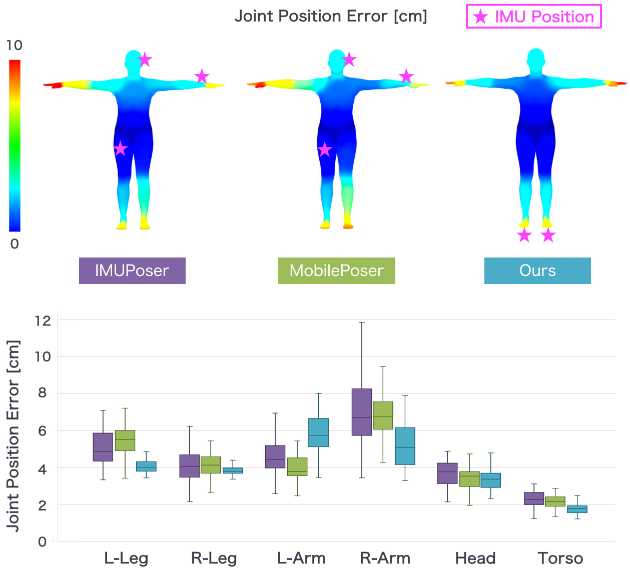


FIGURE 8. Visualization of joint position error distributions. (Top) Heatmaps illustrating that joints closer to IMU attachment sites (indicated by pink stars) exhibit lower positional errors. (Bottom) Box plots detailing joint position errors for different body segments. Our method, utilizing IMUs attached to the soles of both feet, achieves notably higher accuracy in the lower limbs, essential for gait analysis.

Examining the box plots at the bottom of Figure 8 more specifically, IMUPoser and MobilePoser exhibit relatively higher accuracy for the left arm where an IMU is attached, while accuracy tends to degrade for the right arm and lower limbs. In contrast, our proposed method, which attaches IMUs to the soles of both feet, significantly reduces joint position errors in the lower limbs, crucial for gait analysis. This confirms that attaching IMUs to both feet effectively captures lower limb motion more directly and accurately. As for the torso and head regions, there was little difference in accuracy among methods, suggesting that for straight-line walking movements, these regions are less affected by IMU placement.

These results clearly highlight the significant influence of IMU sensor placement on the estimation accuracy of limb joint positions. Particularly for gait-centered analyses, attaching IMUs to the foot soles proves highly effective, indicating that the sensor configuration adopted in our proposed method provides superior performance for lower-body motion estimation.

G. COMPARATIVE EXPERIMENT UNDER IDENTICAL IMU PLACEMENT CONDITIONS

In this section, we conducted experiments under conditions where all methods used IMUs attached at two locations on the soles of the feet, enabling a fair comparison between our proposed method and baseline methods. The results are summarized in Table 2. Additionally, we conducted an ablation study on the proposed method, which consists of three modules (Body, Gait, and Smoothing). Specifically, we compared three configurations: the full model (Ours B+G+S), a model without the Body Module (Ours G+S), and a model without the Smoothing Module (Ours B+G), to examine the contribution of each module.

From Table 2, it is clear that our proposed method, especially the full model (Ours B+G+S), consistently outperformed baseline methods on both the AIST Gait Database and the UnderPressure dataset. Specifically, the proposed method achieved the best performance across metrics related to pose and shape estimation, such as Pose-G, Pose-L, Mesh-P, Mesh-T, and JointAngle, demonstrating high accuracy in both pose estimation and body shape reconstruction. On the other hand, the lower accuracy observed for MobilePoser in the Pose-G and FootTraj metrics on the UnderPressure dataset can be attributed to the use of actual IMUs embedded within insole sensors. These IMUs are of the 6DoF type, lacking magnetometers, and thus cannot provide absolute orientation information in the Yaw (horizontal rotation) direction. Although the VQF was employed to transform IMU data into a global coordinate system, it was not able to completely eliminate noise and drift in the Yaw direction. Consequently, positional errors accumulated, degrading estimation accuracy. In contrast, our proposed method uses IMU data directly in the sensor coordinate system, thus avoiding error accumulation associated with coordinate transformations and effectively overcoming this issue.

The configuration without the Body Module (Ours G+S) exhibits reduced accuracy in metrics related to body shape and pose estimation—such as Pose, Mesh, and InterFoot—indicating that estimating and incorporating a personalized body model from user attributes plays a crucial role in accurately determining body-part positions and stride length. In addition, removing the Smoothing Module (Ours B+G) caused a marked degradation of performance, particularly in the FootVel and RootVel metrics, demonstrating that the Smoothing Module effectively enhances the stability of velocity estimation. Figure 9 shows the Foot Velocity Error of the estimated poses from the Gait Module and the Smoothing Module, plotted over two consecutive windows. In this experiment, the Gait Module estimates gait motion in 50-frame segments from the IMU data of the UnderPressure dataset, while the Smoothing Module processes 50-frame segments shifted by half the window length (25 frames) as the stride width. As a result, around the 50th frame at the center of the plot, the Gait Module alone exhibits discontinuities in pose estimation due to inter-window motion estimation errors, which appear as large peaks in the velocity error. By contrast, applying the Smoothing Module eliminates such discontinuities and, combined with the loss-function design described in Section III-C3, further reduces the velocity error across the entire sequence. These results demonstrate that the Smoothing Module plays an essential role in producing smoother and more accurate motion outputs in the proposed method.

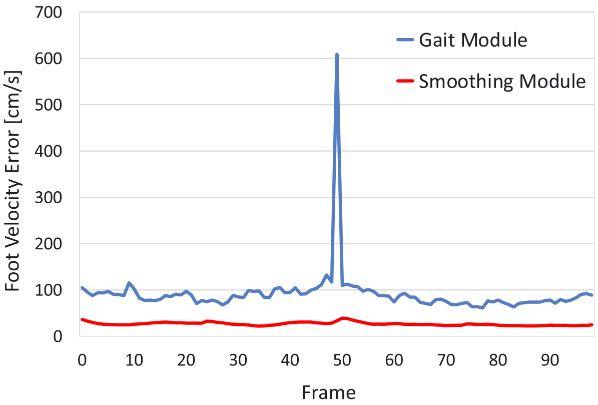


FIGURE 9. Foot Velocity Error of estimated poses over two consecutive windows, comparing the Gait Module (blue) and Smoothing Module (red). The results are computed on the UnderPressure dataset.

Overall, these results confirm that our method demonstrates superior performance even under identical conditions with IMUs fixed to the soles of the feet. Specifically, the design of our method, which utilizes IMU data in the sensor coordinate system, effectively resolves potential drift issues in the Yaw direction when employing real IMU sensor data. Moreover, the combination of the Gait Module with the Body Module and Smoothing Module notably improves the stability of velocity estimations during gait analysis.

V. DISCUSSION AND FUTURE WORK

A. PRACTICAL DEMONSTRATION SYSTEM

To demonstrate the practical applicability of our proposed method, we constructed a demonstration system

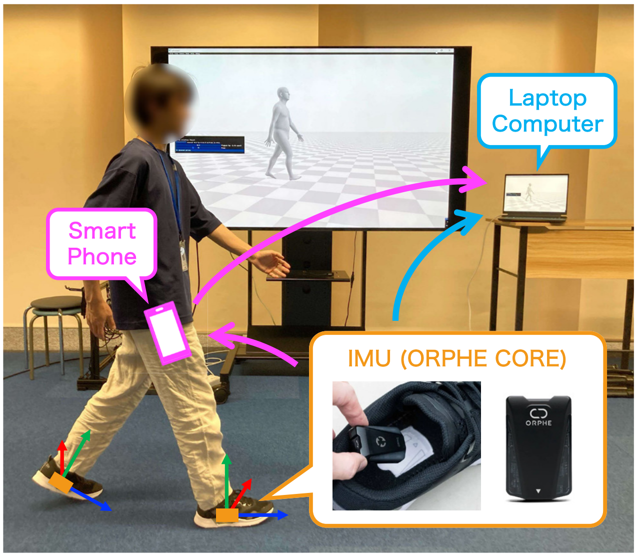


FIGURE 10. Demonstration system of our proposed method using consumer-grade IMU devices (ORPHE CORE). IMU data captured by sensors embedded in shoe soles are transmitted either directly to a laptop computer via Bluetooth (blue arrow), or relayed through a smartphone (magenta arrows) to minimize BLE interference. This setup ensures reliable gait pose estimation in practical scenarios.

for gait motion estimation using consumer-grade IMU devices (Figure 10). Specifically, we adopted the ORPHE CORE [22], a commercial device consisting of specialized shoes with IMUs embedded in the soles. In our demonstration system, the IMU data captured by the ORPHE CORE sensors are transmitted via Bluetooth Low Energy (BLE) to a laptop PC, allowing instantaneous estimation of gait poses. However, BLE communication is highly susceptible to interference from environmental factors, and transmitting data directly from the IMU sensors to a remote PC can be challenging due to limited communication range. To address this, assuming users typically carry their smartphones in hand or in a pocket, we developed an Android application to relay IMU data. As the smartphone remains in close proximity to the ORPHE CORE sensors, it significantly reduces data loss due to BLE interference. The collected data is then transmitted from the smartphone to the PC via Wi-Fi, enabling reliable operation even when the PC is not near the user.

TABLE 3. Quantitative comparison of estimation accuracy across various motions, including straight walking at multiple speeds (Normal, Slow, Fast), non-straight walking, and jogging in the UnderPressure dataset.

Method	Pose-G [cm] (↓)	Pose-L [cm] (↓)	JointAngle [degree] (↓)	RootVel [cm/s] (↓)
Normal Walk	9.20	3.68	2.24	8.16
Slow Walk	7.66	3.87	2.47	7.67
Fast Walk	13.86	3.80	2.28	9.27
Non-Straight Walk	50.18	11.97	4.24	36.09
Jogging	73.87	9.66	10.24	48.48

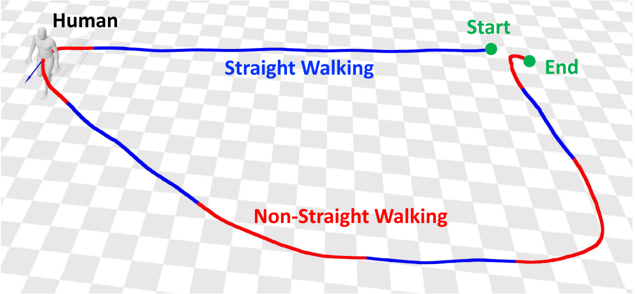


FIGURE 11. Example of straight-walking segmentation in the UnderPressure dataset. Blue indicates straight-walking segments and red indicates non-straight segments. Only the straight-walking segments were used for training and evaluation of the proposed method.

Beyond ensuring robust data transmission, it is equally important for real-world applications that the estimation can be performed under limited computational resources. The proposed model contains 3.05 M parameters and requires about 1.94 M FLOPs per frame, achieving an inference speed of 255 FPS on a workstation running Ubuntu 22.04.5 with an AMD EPYC 7453 (28 cores) CPU and an NVIDIA RTX A6000 GPU. By contrast, the baseline method MobilePoser has 6.67 M parameters, requires about 6.70 M FLOPs per frame, and achieves 93.2 FPS. These results demonstrate that the proposed method is lighter and faster than the baseline. An important future direction is to further reduce the model size and improve efficiency so that real-time operation can be achieved even on smartphones and other mobile or low-power devices.

B. HANDLING STRAIGHT-LINE WALKING CONSTRAINTS

Our proposed method accurately estimates gait pose from minimal IMU data with instantaneous computation. However, the current model is trained specifically on straight-line walking. Consequently, when a user performs motions other than straight-line walking, unnatural movements may be estimated. Table 3 presents the estimation accuracy for three straight-walking speeds (Normal, Slow, and Fast) as well as for non-straight walking and jogging. For the three straight-walking speeds, although some speed-dependent variation is observed in metrics related to walking speed such as Pose-G and RootVel, the estimation maintains reasonable accuracy, confirming that the proposed method can adequately handle changes in walking speed. In contrast, for non-straight walking—examples of which are illustrated by the red trajectories in Figure 11—accuracy drops substantially, particularly in metrics other than joint angle estimation, and jogging shows decreased accuracy across all metrics. These results indicate that turning motions, such as those represented by the red line that exceed the range of body sway observed during straight walking (blue line) in Figure 11, as well as jogging motions, produce acceleration and angular velocity waveforms that fall outside the distribution of the training data, and thus the current model does not yet support such movements.

Moreover, during the training phase, the straight-line segments of the walking data must be manually extracted, and the demonstration system currently lacks a mechanism to identify straight-line walking segments in real-time. To address this limitation, future development should focus on implementing real-time classification of IMU waveform data to automatically detect straight-line walking segments. By incorporating this functionality, we could naturally and continuously collect long-term gait data in everyday life without requiring conscious effort from the user, ultimately facilitating detailed analysis of long-term gait changes due to aging or medical conditions.

C. ADDRESSING DEVICE AND FOOTWEAR VARIABILITY

Practical deployment of our method requires robustness to real-world factors that can affect sensing and estimation accuracy. Variations in footwear type and IMU attachment position may introduce measurement inconsistencies even for identical gait patterns. Possible alternative IMU placements—such as on the instep (e.g., shoelaces), shoe sides, or the heel—would lead to different sensor orientations and measurements, potentially degrading model performance due to inconsistencies in the input–output relationship. Moreover, attaching IMUs to soft shoe parts can introduce additional sensor noise from foot contact-induced vibrations. To mitigate these effects, we intentionally adopted the ORPHE CORE device embedded inside the shoe sole, ensuring a completely fixed and vibration-resistant position that minimizes sensitivity to placement variability.

Similarly, footwear type can influence IMU signals through differences in material stiffness, geometry, and fit, causing waveform variations even for the same motion. To reduce such variability and ensure consistent evaluation, we standardized footwear to sneakers, which are common in everyday use and specifically designed to house the ORPHE CORE device for robust and reproducible IMU placement. While this standardized setup enables highly accurate motion estimation, real-world usage may deviate from these conditions. Future work should therefore explore adaptive calibration strategies and robust model generalization techniques to maintain reliable performance across diverse footwear and sensor placement conditions.

D. TOWARD REAL-WORLD GAIT ANALYSIS

Our proposed method has demonstrated that gait motion can be accurately estimated from only two foot-mounted IMUs under controlled conditions of straight-line walking on level ground. However, comprehensive gait analysis in real-world conditions requires validating the proposed method in naturalistic settings, such as outdoor walkways, residential spaces, and other daily-life environments, to ensure reliable operation beyond controlled laboratory conditions. Long-term, in-the-wild evaluations will be critical to assess performance across ground surfaces of varying compliance (e.g., asphalt, concrete, grass, sand), thereby bridging the gap between laboratory demonstrations and everyday application.

In addition, comprehensive gait analysis in real-world conditions necessitates capturing and analyzing gait in diverse environments, including stairs and slopes. Developing robust algorithms capable of accurately estimating gait pose in these non-flat environments is an important future challenge. Leveraging multiple wearable sensors tailored to the specific movements and environmental contexts under investigation could further enhance motion estimation accuracy and versatility. Promising future directions include integrating insole foot-pressure sensors to perform biomechanical-informed pose estimation, employing smartwatches for detailed full-body motion reconstruction, and using first-person wearable cameras for combined environmental reconstruction and analysis of human interactions with surroundings. By integrating these complementary technologies, we anticipate the development of more comprehensive, versatile gait analysis methods that can be effectively applied across various aspects of everyday life.

E. TOWARD CLINICAL APPLICATIONS

In this study, we proposed a method aimed at accurately reproducing normal gait motions of healthy individuals, but for future clinical applications two aspects will be crucial: constructing detailed, personalized human body models that represent individual patients, and accurately estimating pathological or otherwise abnormal gait patterns.

In this work we employed the SMPL model, a statistical body model built from the body shapes of healthy individuals, which assumes normal body morphology and therefore struggles to represent patients whose body structures have changed. For example, patients with knee osteoarthritis (KOA) are known to exhibit unique skeletal and gait characteristics caused by joint degeneration, pain, and muscular atrophy—features that standard statistical models may fail to capture. To address such clinical scenarios, it will be important to reconstruct high-resolution 3D personalized human body models from medical imaging data such as CT or MRI scans, allowing more accurate reproduction of pathological body structures and motions. Incorporating such personalized models will enable quantitative analysis of disease-specific gait patterns and is expected to contribute to clinical applications such as diagnosis, longitudinal monitoring, and treatment evaluation.

Furthermore, acquiring the ability to estimate abnormal gait unique to non-healthy populations is equally indispensable. Examples include asymmetric gait caused by hemiparesis, shuffling gait associated with Parkinson's disease, and spastic gait observed in cerebral palsy. To capture these conditions, it will be necessary to build and train on datasets that reflect the characteristic motion patterns of each disease. Since these pathological gaits often differ markedly from those of healthy individuals in foot-pressure distribution and muscle activation patterns, combining complementary sensing modalities such as insole pressure sensors or electromyography is also expected to be effective. Advancing such data-driven model extensions and sensor integration will enable high-accuracy joint motion estimation and quantitative characterization of disease-specific gait, further supporting diagnostic assistance and the evaluation of rehabilitation outcomes in clinical practice.

In addition, collecting large, well-annotated datasets of disease-specific gait is inherently challenging in clinical settings. To mitigate this data scarcity, it will be important to draw on strategies surveyed by Alzubaidi et al. [100]—including domain-specific transfer learning, self-supervised representation learning, deep generative approaches such as GANs and DeepSMOTE, and physics-informed neural networks (PINNs). Building on these insights to design new learning frameworks tailored to wearable sensor data, such as IMU and insole pressure signals, will be crucial for robustly modeling patient-specific body structures and pathological gait patterns and for translating our method to future clinical applications.

VI. CONCLUSION

In this paper, we introduced "Gait Inertial Poser (GIP)", a novel method for accurately estimating full-body human poses using only two shoe-embedded IMUs. Our gait-aware deep learning framework effectively addresses common IMU-based issues such as integration drift and orientation inaccuracies, without relying on global coordinate transformations. Experimental results on the AIST Gait Database and UnderPressure datasets demonstrated that our approach consistently outperforms baseline methods across multiple metrics. Additionally, an ablation study confirmed the significant contribution of our Smoothing Module in enhancing velocity estimation stability. We further developed a demonstration system utilizing commercially available IMU shoes (ORPHE CORE), verifying the real-world applicability of our method through stable and near-instantaneous gait estimation in practical conditions. Future work includes extending the method to recognize and handle diverse walking patterns and terrains, as well as integrating additional wearable sensors to further enhance the applicability and versatility of gait analysis in everyday life.

ACKNOWLEDGMENT

This article is based on results obtained from a project, Programs for Bridging the gap between research and development and the IDeal Society (society 5.0) and Generating Economic and social value (BRIDGE) / Practical Global Research in the AI X Robotics Services, implemented by the Cabinet Office, Government of Japan.

REFERENCES

[1] T. A. L. Wren, G. E. Gorton, S. Õunpuu, and C. A. Tucker, "Efficacy of clinical gait analysis: A systematic review," *Gait Posture*, vol. 34, no. 2, pp. 149–153, Jun. 2011.

[2] M. Bonanno, A. M. De Nunzio, A. Quartarone, A. Militi, F. Petralito, and R. S. Calabrò, "Gait analysis in neurorehabilitation: From research to clinical practice," *Bioengineering*, vol. 10, no. 7, p. 785, Jun. 2023.

[3] T. A. L. Wren, C. A. Tucker, S. A. Rethlefsen, G. E. Gorton, and S. Õunpuu, "Clinical efficacy of instrumented gait analysis: Systematic review 2020 update," *Gait Posture*, vol. 80, pp. 274–279, Jul. 2020.

[4] L. Carcreff, C. N. Gerber, A. Paraschiv-Ionescu, G. De Coulon, C. J. Newman, K. Aminian, and S. Armand, "Comparison of gait characteristics between clinical and daily life settings in children with cerebral palsy," *Sci. Rep.*, vol. 10, no. 1, p. 2091, Feb. 2020.

[5] V. V. Shah, J. McNames, M. Mancini, P. Carlson-Kuhta, R. I. Spain, J. G. Nutt, M. El-Gohary, C. Curtze, and F. B. Horak, "Laboratory versus daily life gait characteristics in patients with multiple sclerosis, Parkinson's disease, and matched controls," *J. NeuroEngineering Rehabil.*, vol. 17, no. 1, p. 159, Dec. 2020.

[6] A. C. Schmitt, S. T. Baudendistel, A. L. Lipat, T. A. White, T. E. Raffegeau, and C. J. Hass, "Walking indoors, outdoors, and on a treadmill: Gait differences in healthy young and older adults," *Gait Posture*, vol. 90, pp. 468–474, Oct. 2021.

[7] A. Saboor, T. Kask, A. Kuusik, M. M. Alam, Y. Le Moullec, I. K. Niazi, A. Zoha, and R. Ahmad, "Latest research trends in gait analysis using wearable sensors and machine learning: A systematic review," *IEEE Access*, vol. 8, pp. 167830–167864, 2020.

[8] Y. Hutabarat, D. Owaki, and M. Hayashibe, "Recent advances in quantitative gait analysis using wearable sensors: A review," *IEEE Sensors J.*, vol. 21, no. 23, pp. 26470–26487, Dec. 2021.

[9] D. Laidig, A. J. Jocham, B. Guggenberger, K. Adamer, M. Fischer, and T. Seel, "Calibration-free gait assessment by foot-worn inertial sensors," *Frontiers Digit. Health*, vol. 3, Nov. 2021.

[10] K. Hori, Y. Mao, Y. Ono, H. Ora, Y. Hirobe, H. Sawada, A. Inaba, S. Orimo, and Y. Miyake, "Inertial measurement unit-based estimation of foot trajectory for clinical gait analysis," *Frontiers Physiol.*, vol. 10, Jan. 2020. [Online]. Available: https://www.frontiersin.org/journals/physiology/articles/10.3389/fphys.2019.01530

[11] H. Uchitomi, Y. Hirobe, and Y. Miyake, "Three-dimensional continuous gait trajectory estimation using single shank-worn inertial measurement units and clinical walk test application," *Sci. Rep.*, vol. 12, no. 1, p. 5368, Mar. 2022, doi: 10.1038/s41598-022-09372-w.

[12] M. Mundt, A. Koeppe, F. Bamer, S. David, and B. Markert, "Artificial neural networks in motion analysis—Applications of unsupervised and heuristic feature selection techniques," *Sensors*, vol. 20, no. 16, p. 4581, Aug. 2020, doi: 10.3390/s20164581.

[13] M. Mundt, A. Koeppe, S. David, T. Witter, F. Bamer, W. Potthast, and B. Markert, "Estimation of gait mechanics based on simulated and measured IMU data using an artificial neural network," *Frontiers Bioengineering Biotechnol.*, vol. 8, p. 41, Feb. 2020. [Online]. Available: https://www.frontiersin.org/journals/bioengineering-and-biotechnology/articles/10.3389/fbioe.2020.00041

[14] M. Sharifi Renani, A. M. Eustace, C. A. Myers, and C. W. Clary, "The use of synthetic IMU signals in the training of deep learning models significantly improves the accuracy of joint kinematic predictions," *Sensors*, vol. 21, no. 17, p. 5876, Aug. 2021, doi: 10.3390/s21175876.

[15] Y. Kumano, S. Kanoga, M. Yamamoto, H. Takemura, and M. Tada, "Estimating whole-body walking motion from inertial measurement units at wrist and heels using deep learning," *Int. J. Autom. Technol.*, vol. 17, no. 3, pp. 217–225, May 2023.

[16] 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, Aug. 2021.

[17] Y. Jiang, Y. Ye, D. Gopinath, J. Won, A. Winkler, and C. K. Liu, "Transformer inertial poser: real-time human motion reconstruction from sparse IMUs with simultaneous terrain generation," *Proc. SIGGRAPH Asia*, Daegu, Republic of Korea, 2022, doi: 10.1145/3550469.3555428.

[18] X. Yi, Y. Zhou, and F. Xu, "Physical non-inertial poser (PNP): Modeling non-inertial effects in sparse-inertial human motion capture," in *Proc. Special Interest Group Comput. Graph. Interact. Techn. Conf. Conf. Papers*. New York, NY, USA: Association for Computing Machinery, Jul. 2024, pp. 1–11, doi: 10.1145/3641519.3657436.

[19] V. Xu, C. Gao, H. Hoffmann, and K. Ahuja, "MobilePoser: real-time full-body pose estimation and 3D human translation from IMUs in mobile consumer devices," in *Proc. 37th Annu. ACM Symp. User Interface Softw. Technol.* New York, NY, USA: Association for Computing Machinery, Oct. 2024, pp. 1–11, doi: 10.1145/3654777.3676461.

[20] Y. Kobayashi, N. Hida, K. Nakajima, M. Fujimoto, and M. Mochimaru. (2019). *Aist Gait Database 2019*. [Online]. Available: https://unit.aist.go.jp/harc/ExPART/GDB2019.html

[21] L. Mourot, L. Hoyet, F. L. Clerc, and P. Hellier, "UnderPressure: Deep learning for foot contact detection, ground reaction force estimation and footskate cleanup," *Comput. Graph. Forum*, vol. 41, no. 8, pp. 195–206, Dec. 2022.

[22] Y. Uno, I. Ogasawara, S. Konda, N. Yoshida, N. Otsuka, Y. Kikukawa, A. Tsujii, and K. Nakata, "Validity of spatio-temporal gait parameters in healthy young adults using a motion-sensor-based gait analysis system (ORPHE ANALYTICS) during walking and running," *Sensors*, vol. 23, no. 1, p. 331, Dec. 2022, doi: 10.3390/s23010331.

[23] H. Hörder, I. Skoog, and K. Frändin, "Health-related quality of life in relation to walking habits and fitness: A population-based study of 75-year-olds," *Qual. Life Res.*, vol. 22, no. 6, pp. 1213–1223, Aug. 2013.

[24] J. Park and T.-H. Kim, "The effects of balance and gait function on quality of life of stroke patients," *NeuroRehabilitation*, vol. 44, no. 1, pp. 37–41, Feb. 2019.

[25] S. Barker, R. Craik, W. Freedman, N. Herrmann, and H. Hillstrom, "Accuracy, reliability, and validity of a spatiotemporal gait analysis system," *Med. Eng. Phys.*, vol. 28, no. 5, pp. 460–467, Jun. 2006.

[26] M. Saleh and G. Murdoch, "In defence of gait analysis. Observation and measurement in gait assessment," *J. Bone Joint Surgery. Brit. volume*, vol. 67, no. 2, pp. 237–241, Mar. 1985.

[27] 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.

[28] L. C. Benson, C. A. Clermont, E. Bošnjak, and R. Ferber, "The use of wearable devices for walking and running gait analysis outside of the lab: A systematic review," *Gait Posture*, vol. 63, pp. 124–138, Jun. 2018.

[29] R. Mason, L. T. Pearson, G. Barry, F. Young, O. Lennon, A. Godfrey, and S. Stuart, "Wearables for running gait analysis: A systematic review," *Sports Med.*, vol. 53, no. 1, pp. 241–268, Jan. 2023.

[30] I. Caciula, G. M. Ionita, H. G. Coanda, N. Angelescu, D. Hagiescu, and F. Albu, "Low cost sensor-based gait monitoring system," in *Proc. 15th Int. Conf. Electron., Comput. Artif. Intell. (ECAI)*, Jun. 2023, pp. 01–04.

[31] H. Li, H. Liu, Z. Li, C. Li, Z. Meng, N. Gao, and Z. Zhang, "Adaptive threshold-based ZUPT for single IMU-enabled wearable pedestrian localization," *IEEE Internet Things J.*, vol. 10, no. 13, pp. 11749–11760, Jul. 2023.

[32] R. P. Suresh, V. Sridhar, J. Pramod, and V. Talasila, "Zero velocity potential update (ZUPT) as a correction technique," in *Proc. 3rd Int. Conf. Internet Things: Smart Innov. Usages (IoT-SIU)*, Feb. 2018, pp. 1–8.

[33] J. Wahlström and I. Skog, "Fifteen years of progress at zero velocity: A review," *IEEE Sensors J.*, vol. 21, no. 2, pp. 1139–1151, Jan. 2021.

[34] E. Rapp, S. Shin, W. Thomsen, R. Ferber, and E. Halilaj, "Estimation of kinematics from inertial measurement units using a combined deep learning and optimization framework," *J. Biomechanics*, vol. 116, Feb. 2021, Art. no. 110229. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021929021000099

[35] Vicon. *Vicon Motion Capture Systems*. Accessed: Sep. 1, 2024. [Online]. Available: https://www.vicon.com/

[36] OptiTrack. *Optitrack Motion Capture Systems*. Accessed: Sep. 1, 2024. [Online]. Available: https://www.optitrack.com/

[37] Qualisys. *Qualisys Motion Capture Systems*. Accessed: Sep. 1, 2024. [Online]. Available: https://www.qualisys.com/

[38] M. Analysis. *Motion Analysis Motion Capture Systems*. Accessed: Sep. 1, 2024. [Online]. Available: https://www.motionanalysis.com/

[39] Theia3D. *Theia3D Markerless Motion Capture Systems*. Accessed: Sep. 1, 2024. [Online]. Available: https://www.theiamarkerless.ca/

[40] S. D. Uhlrich, A. Falisse, Ł. Kidziński, J. Muccini, M. Ko, A. S. Chaudhari, J. L. Hicks, and S. L. Delp, "OpenCap: Human movement dynamics from smartphone videos," *PLOS Comput. Biol.*, vol. 19, no. 10, Oct. 2023, Art. no. e1011462. [Online]. Available: https://www.biorxiv.org/content/early/2022/07/10/2022.07.07.499061

[41] C. Bregler and J. Malik, "Tracking people with twists and exponential maps," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Sep. 1998, pp. 8–15.

[42] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun, "Performance capture from sparse multi-view video," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 1–10, Aug. 2008.

[43] L. Sigal, A. O. Balan, and M. J. Black, "HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *Int. J. Comput. Vis.*, vol. 87, nos. 1–2, pp. 4–27, Mar. 2010.

[44] L. Sigal, M. Isard, H. Haussecker, and M. J. Black, "Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propagation," *Int. J. Comput. Vis.*, vol. 98, no. 1, pp. 15–48, May 2012.

[45] C. Stoll, N. Hasler, J. Gall, H.-P. Seidel, and C. Theobalt, "Fast articulated motion tracking using a sums of Gaussians body model," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 951–958.

[46] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund, "Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 5, pp. 538–552, Sep. 2012.

[47] M. Burénius, J. Sullivan, and S. Carlsson, "3D pictorial structures for multiple view articulated pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Nov. 2013, pp. 3618–3625.

[48] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic studio: A massively multiview system for social motion capture," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3334–3342.

[49] A. Elhayek, E. de Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt, "Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3810–3818.

[50] H. Rhodin, N. Robertini, C. Richardt, H.-P. Seidel, and C. Theobalt, "A versatile scene model with differentiable visibility applied to generative pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 765–773.

[51] N. Robertini, D. Casas, H. Rhodin, H.-P. Seidel, and C. Theobalt, "Model-based outdoor performance capture," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 166–175.

[52] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Harvesting multiple views for marker-less 3D human pose annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1253–1262.

[53] L. Xu, Y. Liu, W. Cheng, K. Guo, G. Zhou, Q. Dai, and L. Fang, "FlyCap: Markerless motion capture using multiple autonomous flying cameras," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 8, pp. 2284–2297, Aug. 2018.

[54] S. Li and A. B. Chan, "3D human pose estimation from monocular images with deep convolutional neural network," in *Proc. Asian Conf. Comput. Vis.*, 2015, pp. 332–347.

[55] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3D human pose estimation in the wild using improved CNN supervision," in *Proc. Int. Conf. 3D Vis. (3DV)*, Oct. 2017, pp. 506–516.

[56] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-Fine volumetric prediction for single-image 3D human pose," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1263–1272.

[57] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, "VNect: Real-time 3D human pose estimation with a single RGB camera," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, Aug. 2017.

[58] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis, "Sparseness meets deepness: 3D human pose estimation from monocular video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4966–4975.

[59] C.-H. Chen and D. Ramanan, "3D human pose estimation = 2D pose estimation + matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5759–5767.

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

[61] E. Jahangiri and A. L. Yuille, "Generating multiple diverse hypotheses for human 3D pose consistent with 2D joint detections," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 805–814.

[62] J. Wang, S. Tan, X. Zhen, S. Xu, F. Zheng, Z. He, and L. Shao, "Deep 3D human pose estimation: A review," *Comput. Vis. Image Underst.*, vol. 210, Sep. 2021, Art. no. 103225.

[63] Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: A survey of deep learning-based methods," *Comput. Vis. Image Understand.*, vol. 192, Mar. 2020, Art. no. 102897.

[64] T. Shiratori, H. S. Park, L. Sigal, Y. Sheikh, and J. K. Hodgins, "Motion capture from body-mounted cameras," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 1–10, Jul. 2011, doi: 10.1145/2010324.1964926.

[65] H. Rhodin, C. Richardt, D. Casas, E. Insafutdinov, M. Shafiei, H.-P. Seidel, B. Schiele, and C. Theobalt, ''EgoCap: Egocentric marker-less motion capture with two fisheye cameras,'' *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–11, Nov. 2016, doi: 10.1145/2980179.2980235.

[66] Y.-W. Cha, T. Price, Z. Wei, X. Lu, N. Rewkowski, R. Chabra, Z. Qin, H. Kim, Z. Su, Y. Liu, A. Ilie, A. State, Z. Xu, J.-M. Frahm, and H. Fuchs, ''Towards fully mobile 3D face, body, and environment capture using only head-worn cameras,'' *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 11, pp. 2993–3004, Nov. 2018.

[67] Y.-W. Cha, H. Shaik, Q. Zhang, F. Feng, A. State, A. Ilie, and H. Fuchs, ''Mobile. Egocentric human body motion reconstruction using only eyeglasses-mounted cameras and a few body-worn inertial sensors,'' in *Proc. IEEE Virtual Reality 3D User Interfaces (VR)*, Mar. 2021, pp. 616–625.

[68] H. Jiang and K. Grauman, ''Seeing invisible poses: Estimating 3D body pose from egocentric video,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3501–3509.

[69] Y. Yuan and K. Kitani, ''3D ego-pose estimation via imitation learning,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 763–778.

[70] E. Ng, D. Xiang, H. Joo, and K. Grauman, ''You2Me: Inferring body pose in egocentric video via first and second person interactions,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9887–9897.

[71] Y. Yuan and K. Kitani, ''Ego-pose estimation and forecasting as real-time PD control,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10081–10091.

[72] H. Jiang and V. K. Ithapu, ''Egocentric pose estimation from human vision span,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10986–10994.

[73] W. Xu, A. Chatterjee, M. Zollhöfer, H. Rhodin, P. Fua, H.-P. Seidel, and C. Theobalt, ''Mo2Cap2: real-time mobile 3D motion capture with a cap-mounted fisheye camera,'' *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 5, pp. 2093–2101, May 2019.

[74] T. Hu, K. Sarkar, L. Liu, M. Zwicker, and C. Theobalt, ''EgoRenderer: Rendering human avatars from egocentric camera images,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14508–14518.

[75] D. Tome, P. Peluse, L. Agapito, and H. Badino, ''XR-EgoPose: Egocentric 3D human pose from an HMD camera,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7727–7737.

[76] D. Tome, T. Alldieck, P. Peluse, G. Pons-Moll, L. Agapito, H. Badino, and F. de la Torre, ''SelfPose: 3D egocentric pose estimation from a headset mounted camera,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 6794–6806, Jun. 2023, doi: 10.1109/TPAMI.2020.3029700.

[77] J. Wang, L. Liu, W. Xu, K. Sarkar, and C. Theobalt, ''Estimating egocentric 3D human pose in global space,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11480–11489.

[78] D.-H. Hwang, K. Aso, Y. Yuan, K. Kitani, and H. Koike, ''MonoEye: Multimodal human motion capture system using a single ultra-wide fisheye camera,'' in *Proc. 33rd Annu. ACM Symp. User Interface Softw. Technol.*, Oct. 2020, pp. 98–111.

[79] R. Hori, R. Hachiuma, M. Isogawa, D. Mikami, and H. Saito, ''Silhouette-based 3D human pose estimation using a single Wrist-mounted 360° camera,'' *IEEE Access*, vol. 10, pp. 54957–54968, 2022.

[80] H. Akada, J. Wang, S. Shimada, M. Takahashi, C. Theobalt, and V. Golyanik, ''UnrealEgo: A new dataset for robust egocentric 3D human motion capture,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 1–17.

[81] H. Akada, J. Wang, V. Golyanik, and C. Theobalt, ''3D human pose perception from egocentric stereo videos,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 767–776.

[82] RE Inc. *Rokoko Imu Motion Capture Systems*. Accessed: Sep. 1, 2024. [Online]. Available: https://www.rokoko.com/

[83] Movella. *Xsens Imu Motion Capture Systems*. Accessed: Sep. 1, 2024. [Online]. Available: https://www.movella.com/products/xsens

[84] RE Inc. *Noitom IMU Motion Capture Systems*. Accessed: Sep. 1, 2024. [Online]. Available: https://www.noitom.com/

[85] T. von 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, May 2017.

[86] 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, Dec. 2018.

[87] X. Yi, Y. Zhou, M. Habermann, S. Shimada, V. Golyanik, C. Theobalt, and F. Xu, ''Physical inertial poser (PIP): Physics-aware real-time human motion tracking from sparse inertial sensors,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13157–13168.

[88] Y. Zhang, S. Xia, L. Chu, J. Yang, Q. Wu, and L. Pei, ''Dynamic inertial poser (DynaIP): part-based motion dynamics learning for enhanced human pose estimation with sparse inertial sensors,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 1889–1899.

[89] R. Armani, C. Qian, J. Jiang, and C. Holz, ''Ultra inertial poser: Scalable motion capture and tracking from sparse inertial sensors and ultra-wideband ranging,'' in *Proc. ACM SIGGRAPH Conf. Papers*. New York, NY, USA: Association for Computing Machinery, 2024, pp. 1–11, doi: 10.1145/3641519.3657465.

[90] T. Van Wouwe, S. Lee, A. Falisse, S. Delp, and C. K. Liu, ''DiffusionPoser: Real-time human motion reconstruction from arbitrary sparse sensors using autoregressive diffusion,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 2513–2523.

[91] X. Xiao, J. Wang, P. Feng, A. Gong, X. Zhang, and J. Zhang, ''Fast human motion reconstruction from sparse inertial measurement units considering the human shape,'' *Nature Commun.*, vol. 15, no. 1, p. 2423, Mar. 2024, doi: 10.1038/s41467-024-46662-5.

[92] V. Mollyn, R. Arakawa, M. Goel, C. Harrison, and K. Ahuja, ''IMUPoser: Full-body pose estimation using IMUs in phones, watches, and earbuds,'' in *Proc. CHI Conf. Human Factors Comput. Syst.*, Apr. 2023, pp. 1–12.

[93] C. Zuo, Y. Wang, L. Zhan, S. Guo, X. Yi, F. Xu, and Y. Qin, ''Loose inertial poser: Motion capture with IMU-attached loose-wear jacket,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 2209–2219.

[94] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, ''SMPL: A skinned multi-person linear model,'' *ACM Trans. Graph.*, vol. 34, no. 6, pp. 851–866, 2015.

[95] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, ''On the continuity of rotation representations in neural networks,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5738–5746.

[96] J. Lee and H. Joo, ''Mocap everyone everywhere: Lightweight motion capture with smartwatches and a head-mounted camera,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 1091–1100.

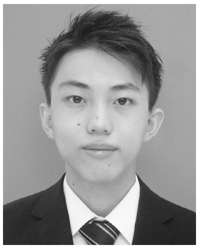
[97] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. Black, ''AMASS: Archive of motion capture as surface shapes,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5441–5450.

[98] M. M. Loper, N. Mahmood, and M. J. Black, ''MoSh: Motion and shape capture from sparse markers,'' in *Proc. SIGGRAPH Asia*, 2014, vol. 33, no. 6, pp. 220:1–220:13.

[99] D. P. Kingma and J. Ba, ''Adam: A method for stochastic optimization,'' in *Proc. Int. Conf. Learn. Represent.*, 2014.

[100] L. Alzubaidi, J. Bai, A. Al-Sabaawi, J. Santamaría, A. S. Albahri, B. S. N. Al-dabbagh, M. A. Fadhel, M. Manoufali, J. Zhang, A. H. Al-Timemy, Y. Duan, A. Abdullah, L. Farhan, Y. Lu, A. Gupta, F. Albu, A. Abbosh, and Y. Gu, ''A survey on deep learning tools dealing with data scarcity: Definitions, challenges, solutions, tips, and applications,'' *J. Big Data*, vol. 10, no. 1, Apr. 2023, doi: 10.1186/s40537-023-00727-2.

RYOSUKE HORI (Student Member, IEEE) received the B.E. and M.Sc.Eng. degrees in information and computer science from Keio University, Japan, in 2021 and 2022, respectively, where he is currently pursuing the Ph.D. degree in science and technology. His research interests include 3D human pose and shape estimation and neuromorphic vision.



HIROYUKI DEGUCHI (Student Member, IEEE) received the B.E. degree in information and computer science from Keio University, Japan, in 2024, where he is currently pursuing the M.Sc.Eng. degree in science and technology. His research interests include 3D human pose and shape estimation and 3D reconstruction.

TSUBASA MARUYAMA received the M.S. and Ph.D. degrees in information science from Hokkaido University, Japan, in 2014 and 2017, respectively. In 2017, he joined the National Institute of Advanced Industrial Science and Technology (AIST), Japan. His research interests include motion measurements, digital twins, and 3D environmental modeling.

MITSUNORI TADA received the B.S. and M.S. degrees in mechanical engineering from The University of Tokyo, Tokyo, Japan, in 1997 and 1999, respectively, and the Ph.D. degree in engineering from Nara Institute of Science and Technology, Nara, Japan, in 2002. Since then, he has been with the National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, where he has been leading the Research Team, Artificial Intelligence Research Center, since 2018. His research interests include realizing human-centered cyber-physical systems. To realize such systems, his research includes motion measurement using wearable sensors, motion analysis with digital human models, and physical function enhancement through assist suits.

HIDEO SAITO (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Keio University, Japan, in 1992. Since 1992, he has been with the Faculty of Science and Technology, Keio University. From 1997 to 1999, he joined the Virtualized Reality Project, Robotics Institute, Carnegie Mellon University, as a Visiting Researcher. Since 2006, he has been a Full Professor with the Department of Information and Computer Science, Keio University. His research interests include computer vision and pattern recognition, and their applications to augmented reality, virtual reality, and human–robotic interaction. His recent activities in academic conferences include being the Program Chair of ACCV 2014, the General Chair of ISMAR 2015, and the Program Chair of ISMAR 2016.