IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 4, 2021, accepted August 6, 2021, date of publication August 18, 2021, date of current version August 27, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3105801

Learning Sensor Interdependencies for IMU-to-Segment Assignment

TOMOYA KAICHI[1], TSUBASA MARUYAMA[2], MITSUNORI TADA[2], AND HIDEO SAITO[1], (Senior Member, IEEE)

[1]Graduate School of Science and Technology, Keio University, Yokohama 223-8522, Japan
[2]Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology, Koto-ku 135-0064, Japan

Corresponding author: Tomoya Kaichi (kaichi@keio.jp)

This work was supported in part by the Grant-in-Aid for JSPS Fellows under Grant 19J22153, and in part by JST AIP-PRISM Grant JPMJCR18Y2.

ABSTRACT Due to the recent technological advances in inertial measurement units (IMUs), many applications for the measurement of human motion using multiple body-worn IMUs have been developed. In these applications, each IMU has to be attached to a predefined body segment. A technique to identify the body segment on which each IMU is mounted allows users to attach inertial sensors to arbitrary body segments, which avoids having to remeasure due to incorrect attachment of the sensors. We address this IMU-to-segment assignment problem and propose a novel end-to-end learning model that incorporates a global feature generation module and an attention-based mechanism. The former extracts the feature representing the motion of all attached IMUs, and the latter enables the model to learn the dependency relationships between the IMUs. The proposed model thus identifies the IMU placement based on the features from global motion and relevant IMUs. We quantitatively evaluated the proposed method using synthetic and real public datasets with three sensor configurations, including a full-body configuration mounting 15 sensors. The results demonstrated that our approach significantly outperformed the conventional and baseline methods for all datasets and sensor configurations.

INDEX TERMS Inertial measurement units, IMU-to-segment assignment, attention mechanism, convolutional neural network, recurrent neural network.

I. INTRODUCTION

Inertial measurement units (IMUs) are a prominent option for analyzing human motion. IMUs measure 3D acceleration, angular velocity, and magnetic field, and they calculate their 3D orientation. Body-worn IMUs can be used to estimate rotational and, sometimes, translational motion of the attached segment, which help estimate the required motion parameters. As the sensors operate at a high frame rate with low latency, they can be introduced in real-time applications for motion analysis, such as full-body motion capture [1]–[3] and navigation [4], [5]. Furthermore, recent technological advances have dramatically reduced the size and price of IMUs, making them the most promising technology for the continuous tracking of human movements in daily life [6]–[8]. Due to recent improvements that have enabled easier configuration, non-expert (but trained) users can collect motion data with IMUs. A clinic's doctors or their assistants can use the inertial sensors to track patients' motions to assist in rehabilitation or disease diagnosis [9]–[11]. Some studies have collected data from many participants wearing IMUs during everyday life for an action recognition task [12]–[14].

For a detailed and robust motion analysis, many IMU-based applications derive data from multiple sensors mounted on multiple body segments. The conventional approach to gait analysis attaches six IMUs to the upper and lower legs and feet [15]. Some IMU-based full-body motion analyses require more than 10 inertial sensors to track one subject [2], [16], [17]. Such configurations are prone to errors because each sensor must be attached to a predefined body segment. If an IMU is mounted on the wrong segment, remeasurement will be required. This problem can be an obstacle for general users' ability to measure motion with IMUs. Hence, a technique to identify the segment to which each sensor is attached based on the sensor signals is

The associate editor coordinating the review of this manuscript and approving it for publication was Agustin Leobardo Herrera-May.

desired, as it would make IMU attachment easier and quicker. This identification task is called an IMU-to-segment (I2S) assignment [18].

In this paper, we address an I2S assignment: the task of classifying IMU data into classes corresponding to the body segments on which IMUs are mounted. With the assignment framework proposed in this paper, although only one IMU needs to be attached to the predetermined segment, the other IMUs can be mounted on arbitrary segments because our framework automatically assigns the sensors to the segments to which they are attached based on the sensors' measurements during a few seconds of walking. The classical approaches to I2S assignments involve manually designing features for discriminating IMU placements [19]–[21]. Recent work has proposed extraction for features using deep neural networks (DNNs) [18]. Although these approaches achieved high assignment accuracy in well-controlled settings (e.g., the approximate angle of the sensor to the segment in the test set is the same as those of the training set), their accuracy has decreased in trials that did not meet these conditions.

To mitigate these limitations and robustly perform the I2S assignment, we propose an approach that merges features across all body-worn IMUs and learns the global dependencies between these IMUs. Unlike conventional methods that classify sensors one by one, our approach assigns locations to all body-worn IMUs at once through a DNN. The proposed model classifies each IMU based on a global feature that represents the motions of all sensor-attached segments of a body. Additionally, the model learns the dependency relationships between IMUs, which enables it to perform assignments based on the data from relevant IMUs (e.g., IMUs attached to the adjacent segment). To implement this feature fusion and dependency learning, we present a new DNN architecture that incorporates a global feature generation module and an attention-based mechanism.

We experimentally evaluated our method using synthetic and real datasets in three sensor configurations. The results demonstrated that the proposed approach significantly outperformed those of the conventional work and baselines in assignment accuracy. Also, the ablation studies and attention maps generated by the intermediate layer of the proposed model suggested that our model captured the dependency relationships between IMUs. The results obtained with the real IMU dataset validated the robustness of our method. Our contributions are summarized as follows:

- We propose a novel I2S assignment model that generates a global feature representing the motion of all body segments to which IMUs are attached and learns pairwise dependencies between the IMUs.
- We demonstrate that merging features extracted from multiple body-worn IMUs can benefit the identification of a segment where each IMU is mounted.
- We show that the proposed method outperforms the conventional and baseline methods in three sensor configurations on synthetic and real public datasets.

II. RELATED WORK

A. IMU-TO-SEGMENT ASSIGNMENT

A line of research on placement recognition of inertial sensors has aimed to define effective feature representations based on signals from IMUs. The early work applied hand-crafted feature descriptors, such as root mean square and amplitudes of accelerations and classical classification algorithms, including support vector machines and decision trees [19]–[21]. The feature descriptors of these approaches are designed based on the intuition and experience of the researchers, with no agreement regarding the most suitable features for I2S assignments.

A recent study for I2S assignment proposed an approach that combines convolutional neural networks (CNNs) and recurrent networks [18]. This combined network was trained in an end-to-end manner without the need to manually design features. This approach assumes that IMUs are attached to the lower limbs and assigns IMUs one by one, ignoring the signals from other IMUs. The proposed method assigns IMUs mounted on the full-body segments by using the signals from all body-worn IMUs. Our method generates a global feature that represents multi-segment motions, which allows the model to assign an IMU of interest based on its relative motion to all the IMUs. In addition, the proposed model learns dependency relationships between the IMUs. Intuitively, when assigning a sensor on the left tibia, the model should pay attention to the data from the IMUs attached to the left femur and the left foot as well. Global feature extraction and dependency learning are incorporated into the proposed model using the techniques introduced in the following Secs. II-B and II-C, respectively. To the best of our knowledge, our work is the first deep learning approach that assigns each IMU to a segment using the aggregated global feature and the sensor interdependencies.

B. GLOBAL FEATURE EXTRACTION

The proposed module to generate a global feature that represents the motion of all segments to which IMUs are attached is inspired by a technique used in point cloud semantic segmentation: the task of separating a point cloud into multiple regions according to the semantic meanings of points [22]. Because a 3D point in a point cloud, which has only positional data, has little information, recently developed approaches have successfully handled point clouds by aggregating local features and obtaining global features [23]–[25]. The feature aggregation module incorporated in the proposed model allows the model to use the global motion of the body segments for the assignment of IMUs.

Pointnet [23] is the pioneering work in applying neural networks to learn over general point sets. It takes raw point clouds as an input and obtains a global feature through a pooling layer that follows individual feature extractors composed of a simple multi-layer-perceptron (MLP). The pooling aggregator is widely used in various tasks against various data structures [26]–[28] due to its simple implementation and the permutation invariance of the inputs. The proposed

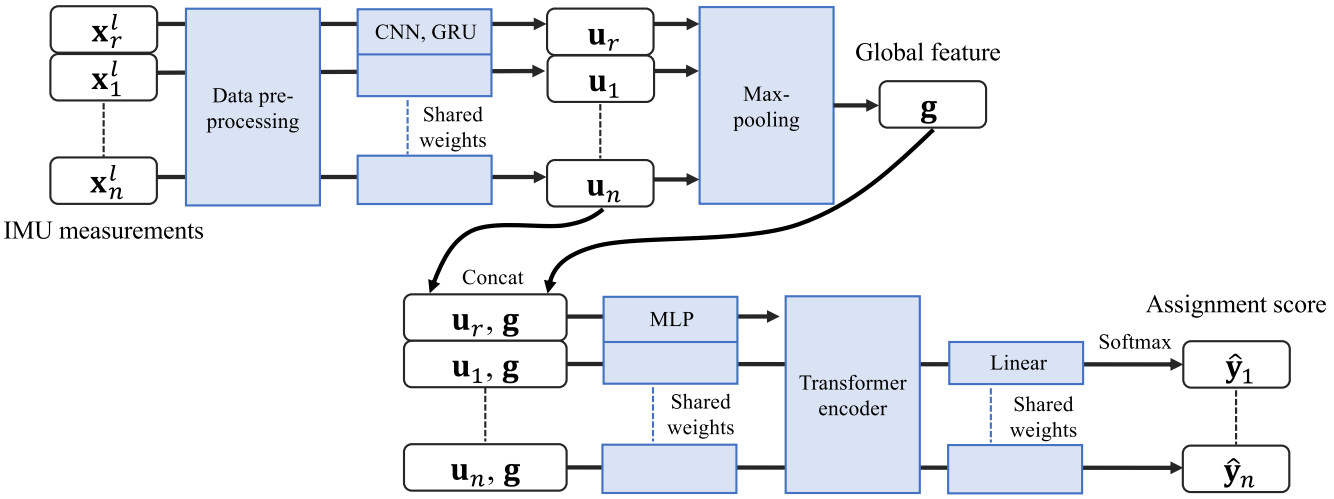


FIGURE 1. An overview of the proposed framework.

assignment model generates a global feature using the pooling aggregator to merge individual features from the IMU data that are input in a random order.

C. ATTENTION MECHANISM

Attention-based neural networks have been successfully applied to a wide variety of fields, such as natural language [29], [30], image [31], [32], and speech processing [33]. The studies report that learning the dependencies among the intermediate features through the attention mechanism improves recognition accuracy. The learned attention also helps interpret the reasoning behind the machine prediction and improves the explainability of the DNN models [34], [35].

Transformer is one of the most promising approaches for learning global dependencies using the attention mechanism [29]. Transformer has been proposed for use in the task of natural language processing and has been quickly adopted for a variety of tasks, such as image classification [32] and object detection [31]. The self-attention operator in Transformer explores the dependencies of input feature vectors. We incorporate the Transformer encoder into our model to obtain the dependency relationships between body-worn IMUs. We expect the attention mechanism to capture the pairwise dependencies of the sensors, which enables the assignment of an IMU that relies on the features extracted from the dependent IMUs.

III. METHODS

A. PROBLEM SETTING

We address the I2S assignment, which involves identifying a segment to which each IMU is mounted, based only on the IMU signals without relying on external sensors. We construct a DNN-based model to learn the discriminant features and classify the IMUs into the attached segments. In the proposed framework, a user processes the assignment following the three steps below:

1) The user selects a root IMU from a set of IMUs to be mounted and attaches it to the predetermined root segment of a subject.
2) The user mounts the remaining IMUs on the arbitrary body segments of the subject.
3) The proposed model provides assignment predictions using the data from all body-worn IMUs while the subject walks for a few seconds.

In our problem setup, only one sensor is placed on the predetermined segment, which dramatically reduces the risk of replacement and the effort required from the user to attach the sensors. Unlike with the conventional methods [18], the user can mount IMUs at any angle. The position of the sensors needs to be known (e.g., an arbitrary sensor should be mounted on the middle of a bone); however, this constraint is satisfied in most practical situations [19]. The role of the root IMU and the difference in assignment accuracy depending on the selected segment as a root are mentioned in Secs. III-C and VI-B, respectively. When we mount 15 sensors on each segment, the I2S assignment can be regarded as a task to classify the sequence data of 14 IMUs into 14 classes associated with the segments, except for the root segment.

B. METHOD OVERVIEW

The proposed I2S assignment framework, as illustrated in Fig. 1, consists of data preprocessing, IMU-wise feature extraction, global feature generation, and attention learning modules. Our model takes as input the accelerations and angular velocities of n target IMUs to be classified and one root IMU and provides n predicted classification scores associated with all segments except the root. Note that the data from the root IMU are placed at the top of the input matrix; however, the data from the n target IMUs are stored in the input matrix in a random order to train the model for the assignment task.

In the data preprocessing module, accelerations and angular velocities in the sensor-local coordinates are converted to the root sensor coordinates, and noise is added to the accelerations for data augmentation. Then, the discriminant features are extracted from the IMU signals in a one-by-one manner, and these features are merged in the global feature generation module. In the final step, global dependencies are learned in the Transformer encoder [29], and the model then provides classification scores through linear transformation with softmax activation.

C. DATA PREPROCESSING

Coordinate transformation and data augmentation are performed in the data preprocessing modules for better generalization and convergence of the proposed assignment model. In this section and Fig. 2, the accelerations, angular velocities, and orientations refer to the values at a specific time step t ($1 \leq t \leq T$), where T is the window size of the IMU data; however, the notation of time step t is eliminated for simplicity.

At first, the raw sensor signals w.r.t. the sensor-local coordinates F_S^i ($1 \leq i \leq n$), where n is the number of IMUs to be assigned, are transformed into the root sensor coordinate frame F_R. The transformation makes the inputs invariant to the walking direction of the subject; this means the representation of the sensor signals can be the same when the subject is walking north and south, which facilitates the training of the model. The transformation matrix $\mathbf{R}_{RS}^i$ that maps F_S^i to F_R can be obtained via

$$\mathbf{R}_{RS}^i = \mathbf{R}_{WR}{}^{\mathsf{T}} \mathbf{R}_{WS}^i, \tag{1}$$

where $\mathbf{R}_{WR}$ and $\mathbf{R}_{WS}^i$ represent the orientation of the root sensor and the i-th sensor w.r.t. the world coordinate frame F_W, respectively. Fig. 2 depicts an example of coordinate transformation when the lower back is chosen as a root segment. Then, 3D acceleration $\mathbf{a}_i$ w.r.t. F_R is calculated by a simple dot product with $\mathbf{R}_{RS}^i$ and the sensor-local acceleration $\mathbf{a}_i^l$ expressed as

$$\mathbf{a}_i = \mathbf{R}_{RS}^i \mathbf{a}_i^l. \tag{2}$$

Given $\mathbf{R}_{RS}^i$ and the sensor-local angular velocity, 3D angular velocity $\boldsymbol{\omega}_i$ w.r.t. F_R is obtained by applying the classical method [36].

Data augmentation is executed to avoid over-fitting and to stabilize the performance of the trained model. Following the methods of successful studies that have applied DNNs to IMU data [18], [37], we augment the sensor signals by adding zero-mean Gaussian noise to the accelerations. The i-th IMU data after the above data preprocessing is referred to as $\mathbf{x}_i \in \mathbb{R}^{T \times 6}$, which stacks T frames of $\mathbf{a}_i$ and $\boldsymbol{\omega}_i$.

D. IMU-WISE FEATURE EXTRACTION AND FEATURE AGGREGATION

The proposed DNN-based assignment model starts with IMU-wise feature extraction. Inspired by the conventional architectures applied to IMU accelerations and angular velocities [18], [37], we construct the feature extractor with CNN layers and a recurrent network layer.

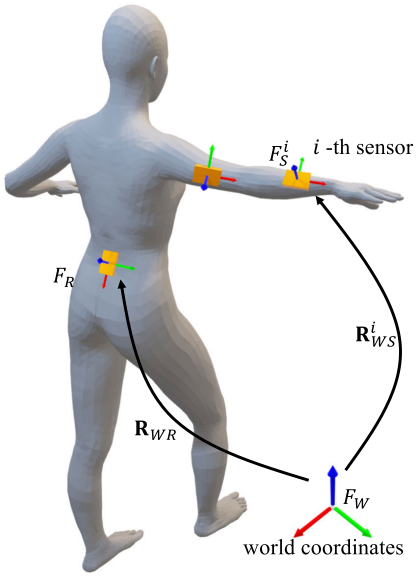


FIGURE 2. The relations among the coordinate systems.

The main difference between previous work and ours is the step-by-step change in kernel size for each CNN layer. As shown in Fig. 3, the kernel size and strides of the first convolution along the height are three. This operator explicitly extracts features from accelerations and angular velocities separately, and the next convolution layer with kernel height $k_h = 2$ fuses both features. Another convolution layer follows to acquire deeper merged features. This feature extraction architecture is inspired by those in the previous literature that report high recognition accuracy in multi-modal fusion tasks using multi-stream feature extraction and fusion modules [38], [39]. In our model, batch normalization and non-linear activation follow each convolution operation. We use ReLU activation $\rho(\cdot)$ for the activation function that computes $\rho(x) = \max(0, x)$.

We incorporate the recurrent units after the convolution layers. We adopt gated recurrent units (GRU) [40] following the results presented in the previous work that performed I2S assignments [18]. The feature map from the last CNN layer $\mathbf{m} \in \mathbb{R}^{T_L \times d_L}$ is divided into T_L one-dimensional features $\mathbf{m}_j \in \mathbb{R}^{d_L}$, where ($1 \leq j \leq T_L$). The feature $\mathbf{m}_j$ is recurrently processed by GRU, and the output at the last time step T_L is returned. Finally, we obtain the IMU-wise feature representation $\mathbf{u}_i$, which is extracted from $\mathbf{x}_i$.

The IMU-wise features individually extracted by the CNNs and the recurrent layer are aggregated to generate a global feature that represents the global motion of the segments to which the IMUs are attached. The architecture chosen for feature merging follows the recent success of the pooling aggregator proposed in [23]. The aggregated feature $\mathbf{g}$ is described as

$$\mathbf{g}(p, q) = \max(\mathbf{u}_r(p, q), \mathbf{u}_1(p, q), \cdots, \mathbf{u}_n(p, q)), \tag{3}$$

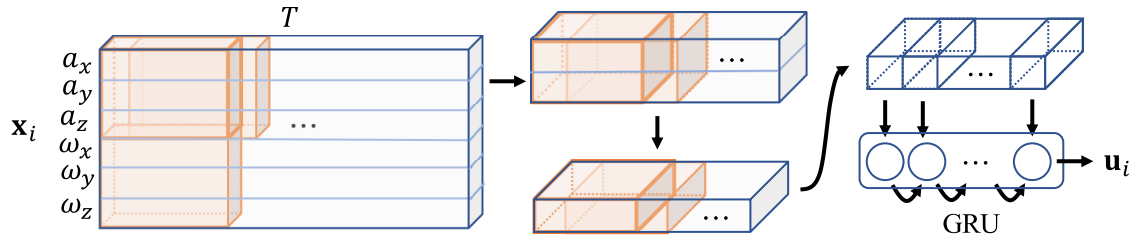


FIGURE 3. Illustration of the proposed convolution operator. The orange boxes in the blue blocks represent the convolution kernels. The kernel size changes for each operation. (a_x, a_y, a_z) and $(\omega_x, \omega_y, \omega_z)$ represent the accelerations $\mathbf{a}_i$ and angular velocities $\boldsymbol{\omega}_i$, respectively.

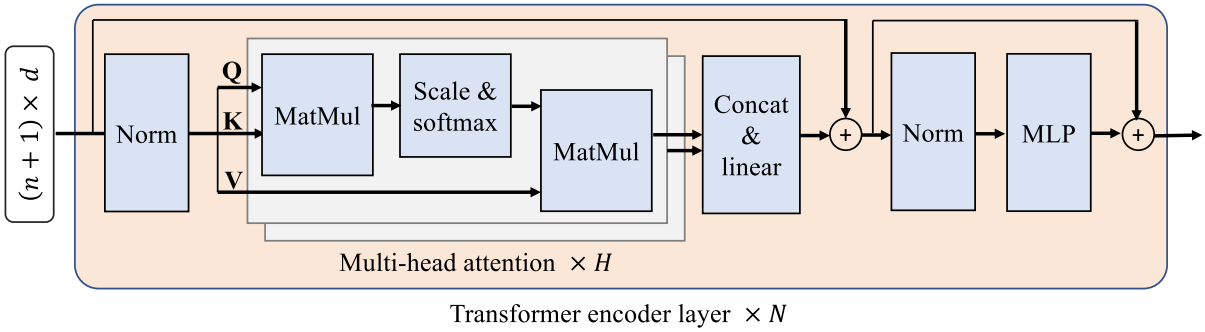


FIGURE 4. The architecture of the transformer encoder layer. The differences from the original are the position of the normalization operator and the lack of position embeddings.

where $\mathbf{g}(p, q)$ and $\mathbf{u}_i(p, q)$ denote the values of $\mathbf{g}$ and $\mathbf{u}_i$ at position (p, q), respectively, and $\mathbf{u}_r$ represents the feature extracted from the root IMU data. The global feature $\mathbf{g}$ forms the same shape as $\mathbf{u}_i$. The features $\mathbf{g}$ and $\mathbf{u}_i$ are concatenated to describe the feature of the i-th IMU, which contains the global feature extracted from all the body-mounted IMUs.

E. ATTENTION-BASED ARCHITECTURE

Transformer learns the dependency relationships between the feature vectors and obtain discriminant feature representations [29]. The IMU-wise features concatenated with the global feature $(\mathbf{u}_i, \mathbf{g})$ are projected to d-dimensional vectors through the MLP with d nodes. The $(n + 1)$ d-dimensional features form a matrix $\mathbf{U} \in \mathbb{R}^{(n+1)\times d}$, which is input to the Transformer layer, as shown in Fig. 4.

The architecture within the attention learning layer is designed to be similar to that of the original Transformer encoder [29]; however, there are two differences between the original and ours. One is the position at which layer normalizations (LNs) are applied. LNs are applied before the multi-head attention module and before MLP, following the method used by recent works that modified the Transformer and improved its recognition accuracy [32], [41]. Another difference is the lack of position embeddings because our model solves an assignment problem that assumes the order of the input is unknown.

A given input $\mathbf{U}$ to the attention learning module is normalized by LN. The normalized $\mathbf{U}$ is projected H times into queries $\mathbf{Q}_h \in \mathbb{R}^{(n+1)\times d_k}$, keys $\mathbf{K}_h \in \mathbb{R}^{(n+1)\times d_k}$, and values $\mathbf{V}_h \in \mathbb{R}^{(n+1)\times d_v}$ by three learnable matrices $\mathbf{W}_h^q, \mathbf{W}_h^k \in \mathbb{R}^{d\times d_k}$, and $\mathbf{W}_h^v \in \mathbb{R}^{d\times d_v}$, where $1 \leq h \leq H$. Using $\mathbf{Q}_h$ and $\mathbf{K}_h$, the attention matrix $\mathbf{A}_h$ is calculated by

$$\mathbf{A}_h = \mathrm{softmax}\left(\frac{\mathbf{Q}_h\mathbf{K}_h}{\sqrt{d}}\right). \tag{4}$$

The H outputs from the multi-head attention, $\mathbf{A}_h\mathbf{V}_h$ are concatenated, linearly projected, and undergo LN. Then, the layer produces an output of the same shape as the input through the IMU-wise MLP. The residual connections are applied before the second LN operator and after the IMU-wise MLP. The attention-based module is composed of a stack of N identical attention learning layers. From the output of the last layer, n feature vectors (except that of the root IMU) are linearly projected with softmax activation, resulting in n probabilities $\hat{\mathbf{y}}_i \in \mathbb{R}^n$.

In the training phase, we use the cross-entropy loss between $\hat{\mathbf{y}}_i$ and the one-hot true label $\mathbf{y}_i \in \mathbb{R}^n$ which is associated with the input $\mathbf{x}_i$ as an objective function. The proposed model is trained in an end-to-end manner.

In the test phase, we found that defining an objective function from the probability distribution $\hat{\mathbf{y}}_i$ and assigning the IMUs to maximize the function improves the accuracy, rather than classifying them directly into the segment indicated by the maximum value of $\hat{\mathbf{y}}_i$. Specifically, the prediction matrix $\mathbf{Y} \in \mathbb{R}^{n\times n}$ is defined by

$$\mathbf{Y}^\mathsf{T} = (\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \cdots, \hat{\mathbf{y}}_n). \tag{5}$$

Let $\mathbf{B} \in \mathbb{R}^{n\times n}$ be a boolean matrix, where $\mathbf{B}(i, j) = 1$ if row i is assigned to column j. Only one of the elements in a

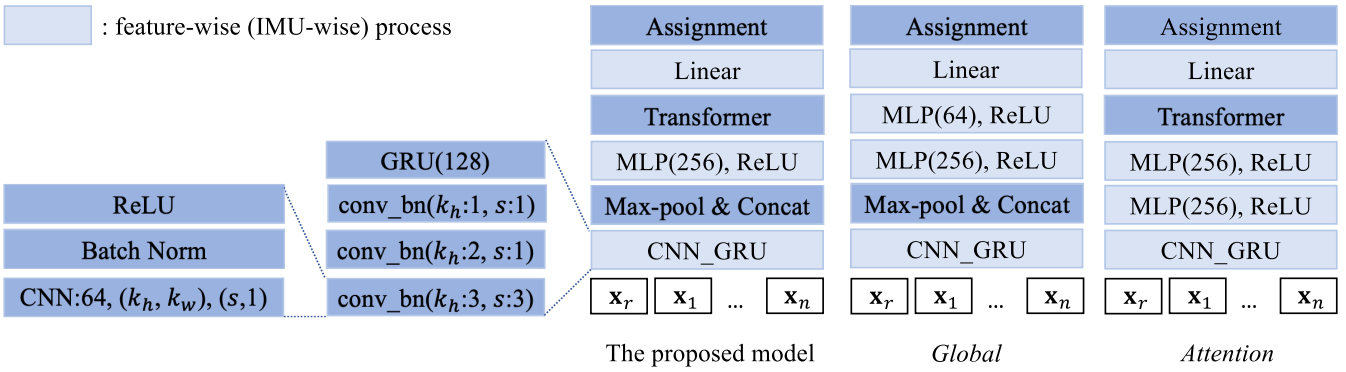


FIGURE 5. The architecture and the hyperparameters of the networks. The three blocks on the left show the proposed method, and the two on the right depict the implemented baselines.

row has 1, and the others must have 0. Then, the assignment algorithm seeks $\hat{\mathbf{B}}$ by solving the following optimization:

$$\hat{\mathbf{B}} = \arg\max_{\mathbf{B}} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{B}(i,j)\mathbf{Y}(i,j). \tag{6}$$

We solved the optimization using the 2D rectangle assignment algorithm [42] implemented in the SciPy library [43]. In the experiments, this optimization was applied to the proposed method and all comparison approaches, which contributed to the improved accuracy of all methods, including the conventional method.

IV. EXPERIMENTAL SETUP

A. IMPLEMENTATION DETAILS

The left three blocks in Fig. 5 illustrate the architecture and hyperparameters of the proposed model. The architecture of each block is detailed in Sec. III. The algorithm based on the Tree-structured Parzen Estimator was used to seek the hyperparameter values, such as the learning rate, the batch size, and the number of kernels and GRU nodes. We divided the dataset into training, validation, and test set (see Appendix A for details); the validation set was then used for parameter tuning, and the values found are described in Appendix B. The parameters are fixed through all the experiments.

B. BASELINES

The assignment accuracy of the proposed model was compared to that of the conventional method [18], referred to as *one-by-one*, which applied DNN to identify IMU placement and infer the I2S orientation alignment of the IMU in a one-by-one manner. Since our work focuses on the I2S assignment, the branch layers for the alignment in *one-by-one* were pruned.

To validate the contribution of the feature aggregation module and the attention-based mechanism, we implemented the two baseline methods. The two models, *Global* and *Attention*, are depicted as the right two blocks in Fig. 5. *Global* is composed of IMU-wise feature extraction and global feature aggregation by the max-pooling layer. *Global* is a model made by removing the attention-based learning module from the proposed architecture. In contrast, *Attention* handles the features extracted from each IMU data to learn the dependency relationships without aggregating the IMU-wise features. The hyperparameters, the dataset division, and the coordinate frame of the input are consistent for the proposed, conventional, and baseline models across all the experiments.

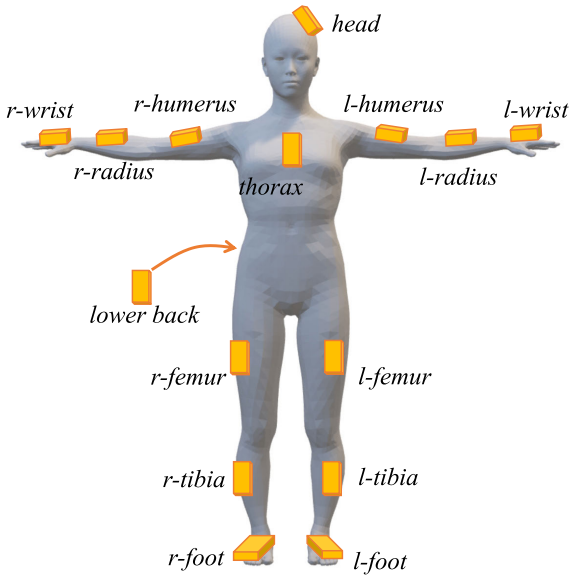


FIGURE 6. Sensor placement in the full-body configuration.

C. DATASET

We quantitatively evaluate the performance of our approach on the synthetic and real IMU datasets: CMU-MoCap [44] and TotalCapture [45]. The sensor arrangement of the CMU-MoCap is shown in Fig. 6. Assuming that the proposed framework is utilized not only for full-body motion analysis but also for the measurement of body parts, we evaluated the model on lower-, upper-, and full-body configurations. The sensor placements are defined as follows:

- lower body (7): *lower back*, *l-femur*, *r-femur*, *l-tibia*, *r-tibia*, *l-foot*, and *r-foot*
- upper body (9): *head*, *thorax*, *lower back*, *l-humerus*, *r-humerus*, *l-radius*, *r-radius*, *l-wrist*, and *r-wrist*

- full body (15): segments on both lower and upper body (*lower back* is duplicated),

where *l*- and *r*- represent left and right body segments, and the figures in $(\cdot)$ denote the number of the segments. Then, since the root segment is determined a priori, the I2S assignment in lower-, upper-, and full-body configurations can be regarded as the task of classifying the time-series signals of the IMUs into 6, 8, and 14 classes, respectively. We selected *lower back* as a root segment through all experiments, excluding Sec. VI-B.

CMU-MoCap is the public human motion dataset captured with the marker-based optical motion capture system (MoCap) [44]. We generated the synthetic IMU data assuming that the IMU was attached to the segments of the body measured in CMU-MoCap. The generation algorithm is described in Appendix C. We selected the same scene used in [18] (42 subjects performing different walking styles). The models were trained with IMU signals from 26 subjects in the training set and 7 subjects in the validation set, and they were tested with the remaining 9 subjects' data (detailed in Appendix A).

TotalCapture is a public dataset providing 60 frame-per-second (fps) of all-synchronized IMU data, HD videos, and ground-truth human poses measured by optical MoCap [45]. Since our approach uses only IMU signals for the I2S assignment, real IMU data were utilized for the training and evaluation of the models. The number of IMUs was 13, and the sensor arrangement was the same as with CMU-MoCap, with the *l-wrist* and *r-wrist* sensors removed. TotalCapture has five subjects with a variety of motions measured. We selected the walking scenes and used three subjects' data for training, one subject's data for validation, and the rest for testing. The period during which the subjects took a calibration pose (the first and last two seconds) and walked backward were manually removed from the dataset. TotalCapture is a challenging dataset in three aspects. First, the number of subjects in the training data is small, which easily causes over-fitting. Second, it contains a variety of walking styles, including many twists and turns and slow and fast walking. Finally, the positions and angles of the sensors attached to the body change slightly depending on the subject because TotalCapture is not a dataset intended for evaluating I2S assignment but for pose estimation. Through the experiments on TotalCapture, we evaluated the versatility of the proposed method.

The window size of the input IMU data was two seconds (i.e., the number of frames $T = 120$ in 60 fps input data), and the windows were always shifted by 0.25 seconds. CMU-MoCap and TotalCapture provide 120 fps and 60 fps IMU signals, respectively, and we used them at the original frame rate.

V. RESULTS

A. ASSIGNMENT ACCURACY

The experimental results obtained using the setup described in Sec. IV are shown in Table 1. As seen in this table, the proposed method outperformed the other methods on both datasets for three configurations of sensor attachment, showing that I2S assignment training in the proposed approach yields better feature representations to discriminate the segment to which each IMU is attached.

TABLE 1. Assignment accuracy on the two datasets in the three configurations. All figures represent percentages.

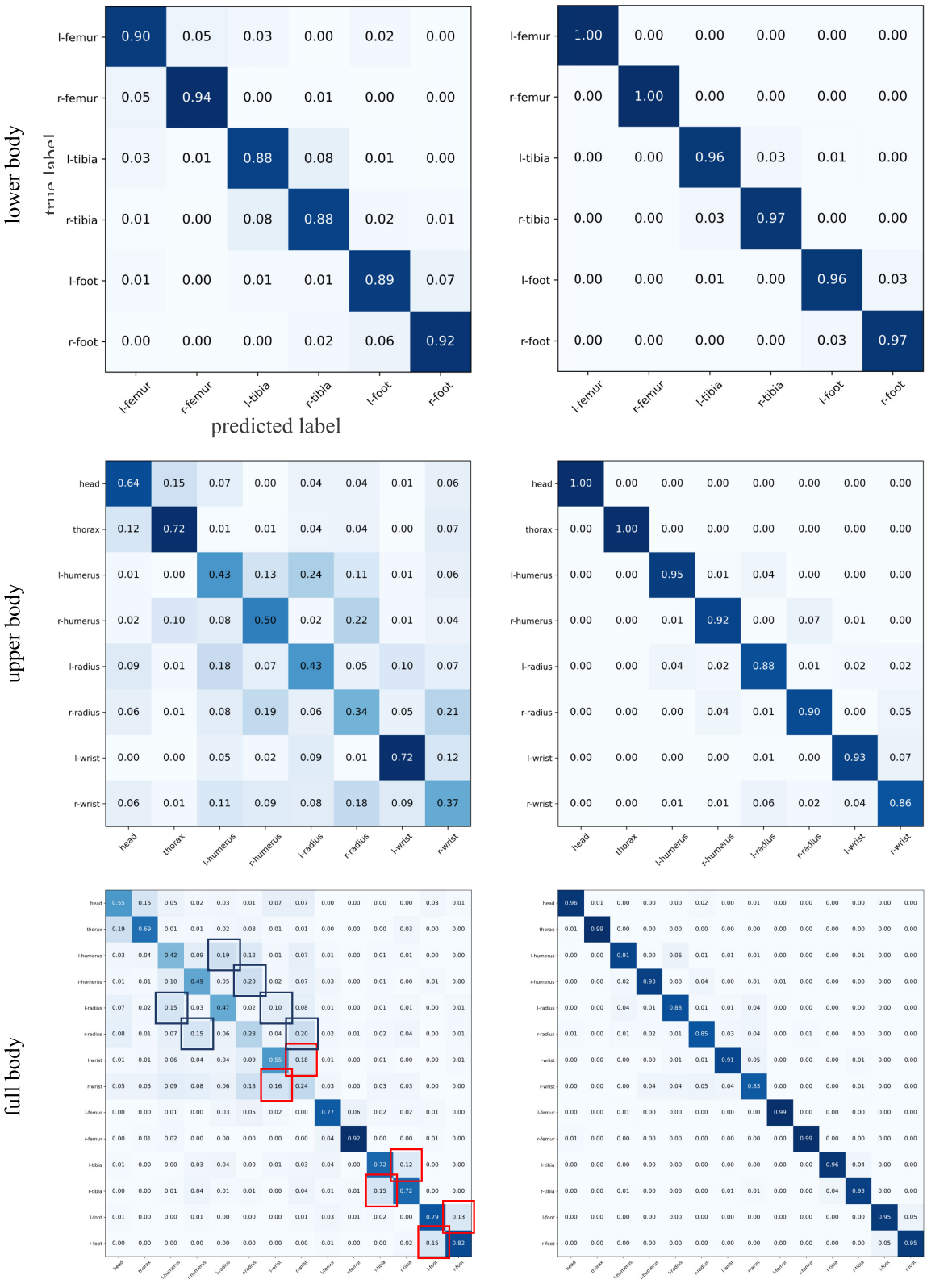
	CMU-MoCap [44]			**TotalCapture** [45]		
	lower	upper	full	lower	upper	full
One-by-one [18]	90.0	51.7	60.2	93.6	80.2	83.1
Global	97.3	81.8	88.1	96.6	89.9	89.9
Attention	97.6	89.7	91.9	91.7	91.0	90.1
Ours	97.8	93.0	93.1	96.7	93.5	91.6

The assignment results on the CMU-MoCap [44] are visualized using confusion matrices in Fig. 7. The matrices show that the assignment errors are caused by two main types of mistakes: left/right switch (l/r switch) and intra-limb misassignment (intra-misassignment). The l/r switch indicates an incorrect assignment to the opposite side of the actually attached segments (e.g., the IMU mounted on the *l-wrist* is classified into the *r-wrist* class). The intra-misassignment denotes that the IMU attached to a part of the limb is misclassified to another part of the same limb (e.g., the IMU mounted on the *l-wrist* is assigned to the *l-radius* or the *l-humerus* class). We highlighted some of the l/r switches and intra-misassignments in the confusion matrix at the lower left part of Fig. 7 with red and blue squares, respectively. The figure shows that the proposed method reduced both mistakes and significantly improved the assignment accuracy.

B. ABLATION STUDIES

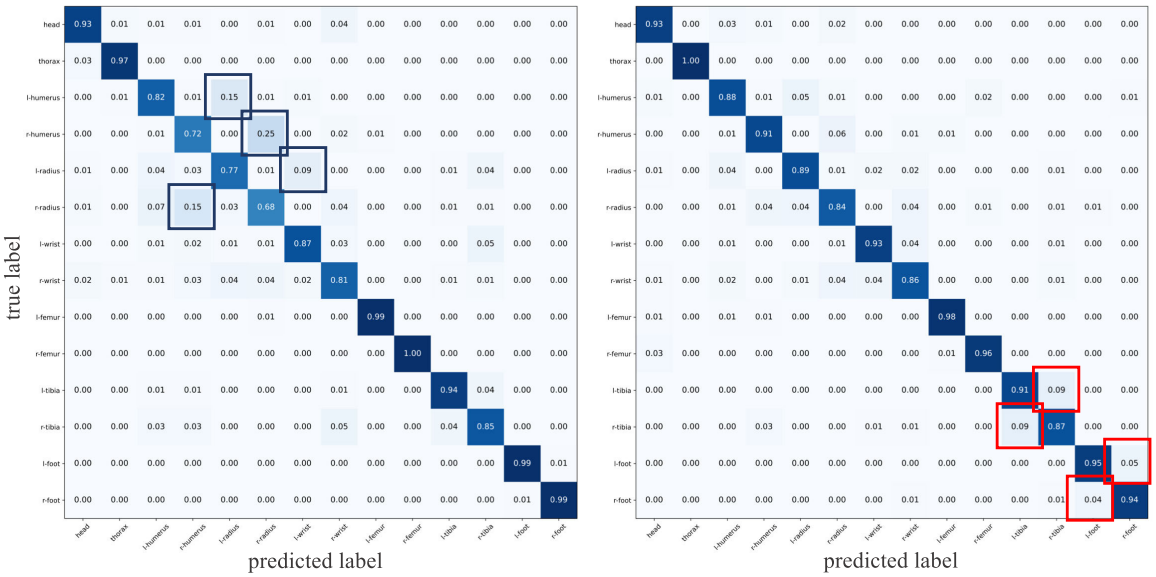
To analyze the contribution of each module in our model to mitigate the l/r switch and intra-misassignment problems, we visualized the confusion matrices of *Global* and *Attention* in Fig. 8 and computed the error rate caused by each mistake. On the CMU-MoCap dataset, the average l/r switch rates (the number of l/r switches divided by the total number of assignments) and the intra-misassignment rates for all three configurations were 2.2% and 5.5%, respectively, for *Global* and 3.1% and 2.4% for *Attention*. The lower l/r switch rates of *Global* and the lower intra-misassignment rate of *Attention* can be observed in the confusion matrices shown in Fig. 8 as well.

The results suggest that the global feature aggregation alleviates the l/r switch problem. This could be because the aggregation allows the network to model the motion of all body segments and capture the motion of each IMU relative to the global body motion, thus enabling the model to discriminate between left and right. The results also suggest that the attention module reduces intra-misassignment errors. This could be because the model with the attention learning architecture classifies the IMU data with consideration of the information from the relevant IMUs, such as IMUs attached to adjacent and opposite segments. For example, as can be seen in Fig. 10(b) (see Sec. II-C for an explanation



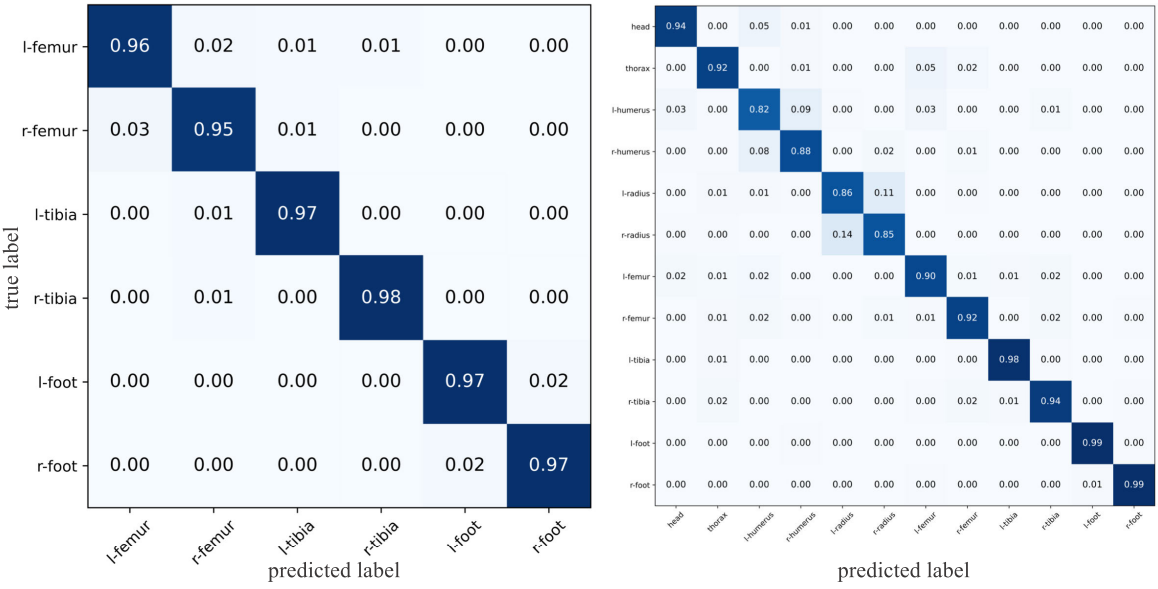
(a) *one-by-one* [16] (b) ours

FIGURE 7. Some results on CMU-MoCap [44] in terms of confusion matrices. The left column represents the assignment results of the conventional work [18], and the right column represents the results of the proposed method. The red and blue rectangles on the lower-left confusion matrix highlight the left/right switches and intra-limb misassignments, respectively.



(a) *Global* (b) *Attention*

FIGURE 8. Comparison between the two baselines on the CMU-MoCap [44] in the full-body configuration. The major cause of incorrect assignment was intra-limb misassignments in (a) ***Global***. On the other hand, (b) ***Attention*** suffered from left/right switches.



(a) lower body (b) full body

FIGURE 9. The assignment accuracy of the proposed method on the TotalCapture dataset [45] in terms of confusion matrices.

of the figure), when assigning an IMU mounted on *l-tibia*, the self-attention architecture devotes much attention to *l-femur*, *l-foot*, and *r-tibia*. The assignment prediction relying on the IMUs on the segments in the same limb should prevent intra-misassignment.

C. RESULTS ON A CHALLENGING DATASET

The results on the TotalCapture dataset [45], as presented in Table 1 and Fig. 9, revealed that the proposed approach is robust to different walking styles and slight changes in the IMU positions depending on the subjects. Our model took the same period of data as an input regardless of the change in walking speed, but the method achieved high accuracy in all the sensor configurations.

The accuracy in assigning the arm segments was lower than that of the other segments for two main reasons. One is a variety of movements not found in a normal gait in the training dataset, such as touching a head or face and raising

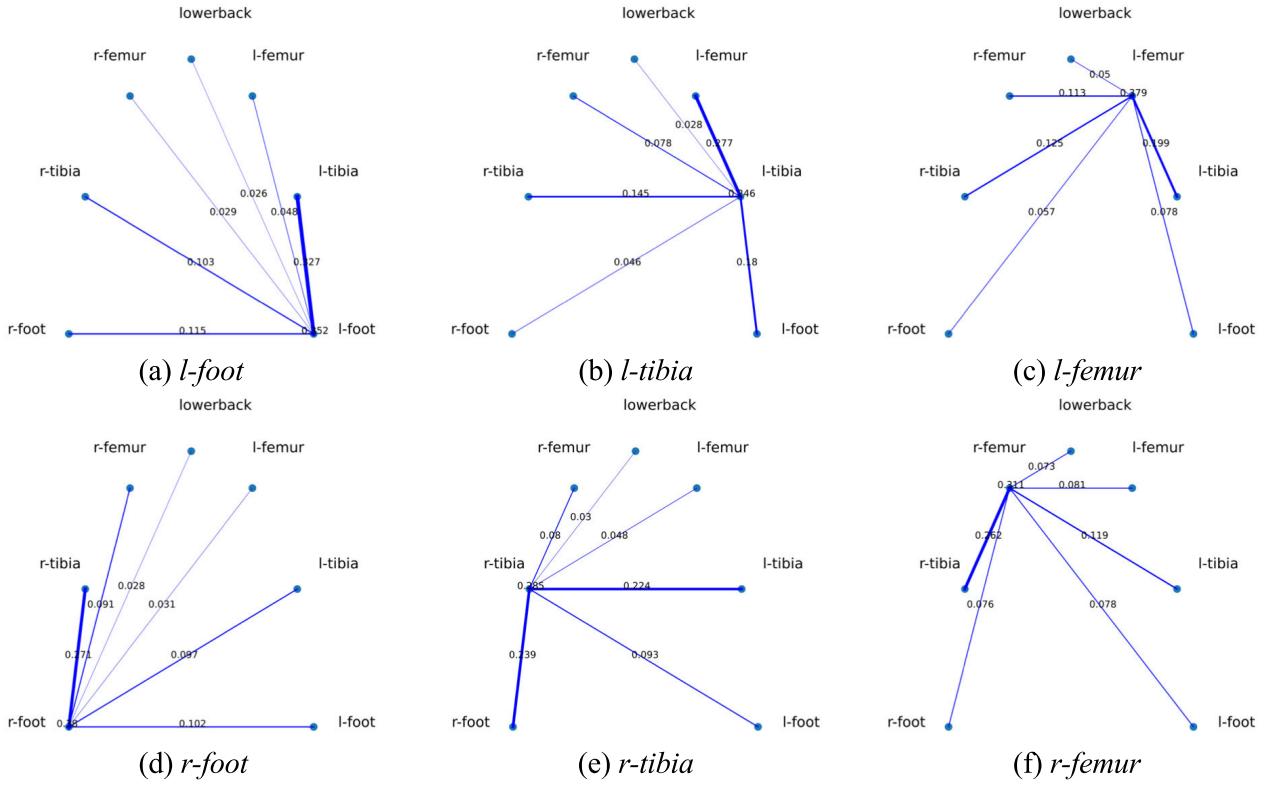


FIGURE 10. Visualization of the self-attention matrices of the proposed model. The figures on the lines denote the attention scores. The higher the score, the darker the color of the line. The training and test were performed on the CMU-MoCap [44] in the lower-body configuration.

clenched fists. The other is that the subject in the test set walked without moving his arms for a few seconds. The trained model could not distinguish between the IMU movement on the arms, head, and chest in the scene. Specifically, the mean assignment accuracy in the three seconds of the test scene in which the subject walked slowly without waving his arms (from 55 to 58 seconds in S5-W2 in TotalCapture [45]) was 61.1% in the full-body setting.

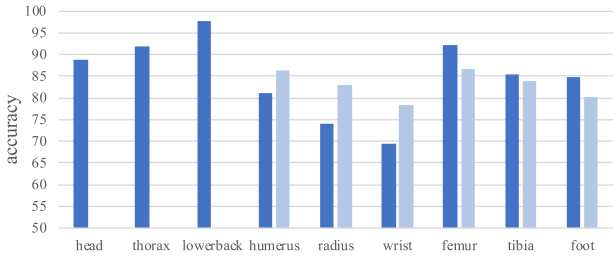


FIGURE 11. Assignment accuracy depending on the root segment on CMU-MoCap [44] in full-body configuration. The graphs in lighter blue represent the right side of the segment (e.g., *r-humerus* and *r-radius*).

VI. DISCUSSION

A. ATTENTION MAPS VISUALIZATION

An attention mechanism can be used to improve the explainability of deep learning models [35], [46], [47]. Explainability, in this context, refers to a better understanding for humans of why the models behave as they do. The explainability of a model helps users make decisions based on the model and allows researchers to understand what input and intermediate features affect the results of the model. The attention learning architecture used in our model can capture the pairwise relationships between the IMUs and explain what dependencies the predicted assignments rely highly on.

To visualize the dependencies between the IMUs, we calculated the mean attention matrix, representing the average of the self-attention matrix (calculated by Eq. (4)) from all the H heads and all the N Transformer encoder layers. Each row of the attention matrix represents the dependencies between the IMUs associated with the columns. The pairwise dependencies are separately visualized for each body segment in Fig. 10. A high attention score suggests high dependency. For example, Fig. 10(a) describes the degree of dependence on the IMU attached to each segment when the model performed the assignment for the IMU mounted on *l-foot*. For all the segments from (a) to (f) in Fig. 10, it can be seen that much attention is devoted to the adjacent segments and the opposite segments even in the test phase, during which the model has no prior knowledge of which segment each IMU is mounted on.

B. ROOT SEGMENT SELECTION

The accuracy of the I2S assignment according to the root segment is shown in Fig. 11. The results suggest that the segments that stably and faithfully follow the body orientation

TABLE 2. Dataset division. The figures represent the ID of the subjects.

	Train	Validation	Test
CMU-MoCap [44]	2, 6, 7, 10, 15, 16, 32, 36, 37, 38, 39, 45, 81, 91, 93, 103, 104, 105, 114, 120, 132, 133, 139, 141, 143, 144	3, 8, 43, 69, 113, 136, 137	5, 12, 26, 27, 29, 46, 49, 55, 111
TotalCapture [45]	1, 2, 3	4	5

(e.g., *lower back*, *thorax*, *head*, and *femur*) are suitable for the root. In contrast, when we chose the segments on the arms that have great freedom of movement during walking, the assignment accuracy decreased significantly.

VII. CONCLUSION

We have presented an approach that identifies the segment on which each IMU is mounted by merging the features of all the body-worn IMUs and by learning the dependency relationships between the sensors. A pooling aggregator was incorporated to obtain a feature that represents the global motion of the body. In addition, a self-attention learning architecture was implemented to allow the model to perform an IMU assignment relying on the signals from the relevant IMUs. The proposed model was quantitatively evaluated on simulated and real IMU datasets, which validated our method, showing that it accurately and robustly performed the I2S assignment. Ablation studies suggested that the global feature fusion and attention mechanism reduced left/right switches and intra-limb misassignments.

Our I2S assignment framework assumes that the sensor configuration is known a priori and that one of the sensors is placed on the predetermined segment. These limitations do not significantly impair practicality; however, further studies to relax them are needed.

APPENDIX A
DIVISION OF THE DATASET

The data in a dataset are divided into training, validation, and test sets. In this paper, both the synthetic dataset CMU-MoCap [44] and the real dataset TotalCapture [45] are divided into the three sets on the basis of the subject (i.e., all the trials (scenes) of a subject are put into one of the three sets). The specific division of each dataset is summarized in Table 2. In CMU-MoCap [44], we selected subjects performing simple "walking" and had at least 600 frames in every scene as a test set.

APPENDIX B
HYPERPARAMETERS FOR MODEL TRAINING

We adopted the hyperparameters described in this section. Fig. 5 also visualizes the architecture and parameters of the network.

In IMU-wise feature extraction, three CNNs with different kernel sizes (3, k_w), (2, k_w), and (1, k_w), where k_w represents the kernel width, were utilized in this order. The strides of these kernels were (3,1), (1,1), and (1,1), respectively: k_w varies to scale the size of the convolution operator, depending on the input frame T. Specifically, k_w was set to $[T/15 + 1]$ in the experiments. The number of nodes in GRU following the CNNs was 128. After the max pooling and the concatenations of the vectors, MLP with the number of nodes $d = 256$ mapped each feature to be the input of the Transformer encoder layer [29]. The hyperparameters in the Transformer encoder were as follows: The embedding dimensions of the query d_k, key d_k, and value d_v were 256. The number of the MLP nodes after the second LN operator was set to 768. The number of the attention heads H and of encoder layers N was fixed to 4.

Our network was implemented in TensorFlow [48] and trained for 1000 epochs with a batch size 128. Early stopping with patience 400 was performed, and the model that achieved the lowest loss on the validation set was utilized for the test. RMSProp with a fixed learning rate 0.001 was applied to optimize the model.

APPENDIX C
SIMULATED DATA GENERATION

The public human motion dataset named CMU-MoCap [44] provides much 3D kinematics data which are measured using the optical MoCap. We used the human joint position $\mathbf{p}^t_{WJ}$ and orientation $\mathbf{R}^t_{WJ}$ w.r.t. the world coordinates F_W at time step t. We virtually attached an IMU to a bone by defining a rotation matrix $\mathbf{R}_{JS}$ and translation vector $\mathbf{t}_{JS}$, which represent the orientation and position of the virtual sensor w.r.t. the joint coordinate frame F_J. They are kept fixed in F_J during movement, assuming that the IMU is attached to a rigid human body and its motion is perfectly linked to the associated joint motion. In the experiments, the IMU was placed at the midpoint of the bone, and the joint position data was preprocessed with a zero lag Butterworth filter [49] of order 8 and a cutoff frequency of 10 Hz, following the previous work [18]. Then, the IMU position $\mathbf{p}^t_{WS}$ and orientation $\mathbf{R}^t_{WS}$ w.r.t. F_W at time step t were obtained via

$$\mathbf{p}^t_{WS} = \mathbf{p}^t_{WJ} + \mathbf{R}^t_{WJ}\mathbf{t}_{JS} \tag{7}$$

$$\mathbf{R}^t_{WS} = \mathbf{R}^t_{WJ}\mathbf{R}_{JS}. \tag{8}$$

The angular velocity of the IMU w.r.t. F_W is computed by [36] using the sensor orientation in the current frame R^t_{WS} and the next frame $R^{(t+1)}_{WS}$. The IMU acceleration at t time step $\mathbf{a}^t_{WS}$ w.r.t. F_W is calculated by

$$\mathbf{a}^t_{WS} = \frac{\mathbf{p}^{(t+1)}_{WS} - 2\mathbf{p}^t_{WS} + \mathbf{p}^{(t-1)}_{WS}}{\Delta t^2}, \tag{9}$$

where Δt denotes the period of the time step. Since all the IMU accelerations and angular velocities are transformed

to the root sensor coordinate before they are input to the assignment model, these values are invariant to the IMU orientations w.r.t. F_J. Therefore, unlike the previous work [18], we did not generate IMU data with various orientations relative to F_J.

REFERENCES

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

[2] C. Malleson, A. Gilbert, M. Trumble, J. Collomosse, A. Hilton, and M. Volino, "Real-time full-body motion capture from video and IMUs," in *Proc. Int. Conf. 3D Vis. (3DV)*, Oct. 2017, pp. 449–457.

[3] T. Maruyama, M. Tada, and H. Toda, "Riding motion capture system using inertial measurement units with contact constraints," *Int. J. Autom. Technol.*, vol. 13, no. 4, pp. 506–516, Jul. 2019.

[4] R. Gonzalez and P. Dabove, "Performance assessment of an ultra low-cost inertial measurement unit for ground vehicle navigation," *Sensors*, vol. 19, no. 18, p. 3865, Sep. 2019.

[5] F. Jamil, N. Iqbal, S. Ahmad, and D.-H. Kim, "Toward accurate position estimation using learning to prediction algorithm in indoor navigation," *Sensors*, vol. 20, no. 16, p. 4410, 2020.

[6] L. Sy, M. Raitor, M. D. Rosario, H. Khamis, L. Kark, N. H. Lovell, and S. J. Redmond, "Estimating lower limb kinematics using a reduced wearable sensor count," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 4, pp. 1293–1304, Apr. 2021.

[7] S. Tedesco, J. Barton, and B. O'Flynn, "A review of activity trackers for senior citizens: Research perspectives, commercial landscape and the role of the insurance industry," *Sensors*, vol. 17, no. 6, p. 1277, Jun. 2017.

[8] S. Qiu, Z. Wang, H. Zhao, and H. Hu, "Using distributed wearable sensors to measure and evaluate human lower limb motions," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 4, pp. 939–950, Apr. 2016.

[9] C. Caramia, D. Torricelli, M. Schmid, A. Muñoz-Gonzalez, J. Gonzalez-Vargas, F. Grandas, and J. L. Pons, "IMU-based classification of Parkinson's disease from gait: A sensitivity analysis on sensor location and feature selection," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 6, pp. 1765–1774, Nov. 2018.

[10] N. M. Rad, T. Van Laarhoven, C. Furlanello, and E. Marchiori, "Novelty detection using deep normative modeling for IMU-based abnormal movement monitoring in Parkinson's disease and autism spectrum disorders," *Sensors*, vol. 18, no. 10, p. 3533, 2018.

[11] H. Zhao, Z. Wang, S. Qiu, Y. Shen, and J. Wang, "IMU-based gait analysis for rehabilitation assessment of patients with gait disorders," in *Proc. 4th Int. Conf. Syst. Informat. (ICSAI)*, Nov. 2017, pp. 622–626.

[12] P. Kasebzadeh, G. Hendeby, C. Fritsche, F. Gunnarsson, and F. Gustafsson, "IMU dataset for motion and device mode classification," in *Proc. Int. Conf. Indoor Positioning Indoor Navigat. (IPIN)*, Sep. 2017, pp. 1–8.

[13] F. Haider, F. A. Salim, D. B. W. Postma, R. van Delden, D. Reidsma, B.-J. van Beijnum, and S. Luz, "A super-bagging method for volleyball action recognition using wearable sensors," *Multimodal Technol. Interact.*, vol. 4, no. 2, p. 33, Jun. 2020.

[14] P. V. Rouast, H. Heydarian, M. T. P. Adam, and M. E. Rollo, "OREBA: A dataset for objectively recognizing eating behavior and associated intake," *IEEE Access*, vol. 8, pp. 181955–181963, 2020.

[15] T. Seel, J. Raisch, and T. Schauer, "IMU-based joint angle measurement for gait analysis," *Sensors*, vol. 14, no. 4, pp. 6891–6909, Jan. 2014.

[16] T. Kaichi, T. Maruyama, M. Tada, and H. Saito, "Resolving position ambiguity of IMU-based human pose with a single RGB camera," *Sensors*, vol. 20, no. 19, p. 5453, Sep. 2020.

[17] G. Marta, F. Simona, C. Andrea, B. Dario, S. Stefano, V. Federico, B. Marco, B. Francesco, M. Stefano, and P. Alessandra, "Wearable biofeedback suit to promote and monitor aquatic exercises: A feasibility study," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 4, pp. 1219–1231, Apr. 2020.

[18] T. Zimmermann, B. Taetz, and G. Bleser, "IMU-to-segment assignment and orientation alignment for the lower body using deep learning," *Sensors*, vol. 18, no. 1, p. 302, 2018.

[19] D. Weenk, B.-J. F. Van Beijnum, C. T. Baten, H. J. Hermens, and P. H. Veltink, "Automatic identification of inertial sensor placement on human body segments during walking," *J. Neuroeng. Rehabil.*, vol. 10, no. 1, p. 31, 2013.

[20] K. Kunze, P. Lukowicz, H. Junker, and G. Tröster, "Where am I: Recognizing on-body positions of wearable sensors," in *Proc. Int. Symp. Location-Context-Awareness*. Cham, Switzerland: Springer, 2005, pp. 264–275.

[21] N. Amini, M. Sarrafzadeh, A. Vahdatpour, and W. Xu, "Accelerometer-based on-body sensor localization for health and medical monitoring applications," *Pervas. Mobile Comput.*, vol. 7, no. 6, pp. 746–760, 2011.

[22] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun, "Deep learning for 3D point clouds: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jun. 29, 2020, doi: 10.1109/TPAMI.2020.3005434.

[23] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 652–660.

[24] Q. Hu, B. Yang, L. Xie, S. Rosa, Y. Guo, Z. Wang, N. Trigoni, and A. Markham, "Learning semantic segmentation of large-scale point clouds with random sampling," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, May 25, 2021, doi: 10.1109/TPAMI.2021.3083288.

[25] L. Ge, Z. Ren, and J. Yuan, "Point-to-point regression pointnet for 3D hand pose estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 475–491.

[26] D. Xu, D. Anguelov, and A. Jain, "PointFusion: Deep sensor fusion for 3D bounding box estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 244–253.

[27] P.-H. Huang, K. Matzen, J. Kopf, N. Ahuja, and J.-B. Huang, "DeepMVS: Learning multi-view stereopsis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2821–2830.

[28] P. T. Komiske, E. M. Metodiev, and J. Thaler, "Energy flow networks: Deep sets for particle jets," *J. High Energy Phys.*, vol. 2019, no. 1, p. 121, Jan. 2019.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*. [Online]. Available: http://arxiv.org/abs/1706.03762

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186.

[31] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*. [Online]. Available: http://arxiv.org/abs/2010.11929

[33] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," 2015, *arXiv:1506.07503*. [Online]. Available: http://arxiv.org/abs/1506.07503

[34] X. Chen, H. Chen, H. Xu, Y. Zhang, Y. Cao, Z. Qin, and H. Zha, "Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2019, pp. 765–774.

[35] S. Yang, Z. Quan, M. Nie, and W. Yang, "TransPose: Keypoint localization via transformer," 2020, *arXiv:2012.14214*. [Online]. Available: http://arxiv.org/abs/2012.14214

[36] F. Barnes, "Stable member equations of motion for a three-axis gyro stabilized platform," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-7, no. 5, pp. 830–842, Sep. 1971.

[37] F. J. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.

[38] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, Nov. 2017, pp. 163–171.

[39] D. H. Kim, M. K. Lee, D. Y. Choi, and B. C. Song, "Multi-modal emotion recognition using semi-supervised learning and multiple neural networks in the wild," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, Nov. 2017, pp. 529–535.

[40] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*. [Online]. Available: http://arxiv.org/abs/1412.3555

[41] S. Narang, H. Won Chung, Y. Tay, W. Fedus, T. Fevry, M. Matena, K. Malkan, N. Fiedel, N. Shazeer, Z. Lan, Y. Zhou, W. Li, N. Ding, J. Marcus, A. Roberts, and C. Raffel, "Do transformer modifications transfer across implementations and applications?" 2021, *arXiv:2102.11972*. [Online]. Available: http://arxiv.org/abs/2102.11972
[42] D. F. Crouse, "On implementing 2D rectangular assignment algorithms," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 52, no. 4, pp. 1679–1696, Aug. 2016.
[43] P. Virtanen *et al.*, "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, pp. 261–272, Feb. 2020.
[44] *CMU Graphics Lab Motion Capture Database*. Accessed: Apr. 13, 2021. [Online]. Available: http://mocap.cs.cmu.edu
[45] M. Trumble, A. Gilbert, C. Malleson, A. Hilton, and J. Collomosse, "Total capture: 3D human pose estimation fusing video and inertial sensors," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–13.
[46] Y. Zhang, S. Qian, Q. Fang, and C. Xu, "Multi-modal knowledge-aware hierarchical attention network for explainable medical question answering," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1089–1097.
[47] Y. Zhang, P. Zhao, Y. Guan, L. Chen, K. Bian, L. Song, B. Cui, and X. Li, "Preference-aware mask for session-based recommendation with bidirectional transformer," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 3412–3416.
[48] M. Abadi *et al.* (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. [Online]. Available: https://www.tensorflow.org/
[49] S. Butterworth, "On the theory of filter amplifiers," *Exp. Wireless Wireless Eng.*, vol. 7, pp. 536–541, Oct. 1930.

TOMOYA KAICHI received the B.E. and M.Sc.Eng. degrees in information and computer science from Keio University, Japan, in 2017 and 2018, respectively, where he is currently pursuing the Ph.D. degree in science and technology. His research interests include human motion analysis, optical-inertial sensor fusion, and hyperspectral image analysis.

TSUBASA MARUYAMA received the M.S. and Ph.D. degrees in information science from Hokkaido University, Japan, in 2014 and 2017, respectively. He joined the National Institute of Advanced Industrial Science and Technology (AIST), in 2017. His research interests include the digital human, human motion measurements, human-centered digital twin, and three dimensional (3D) environment modeling using 3D laser scanning and photogrammetry.

MITSUNORI TADA received the Ph.D. degree in engineering from Nara Institute of Science and Technology, Japan, in 2002. Since 2002, he has been with the National Institute of Advanced Industrial Science and Technology (AIST). Since 2018, he has been a Research Team Leader with the Artificial Intelligence Research Center, AIST. His research interests include real-time human motion measurement and analysis for realizing human-centered cyber physical systems with human digital twins.

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 was a Visiting Researcher with The Robotics Institute, Carnegie Mellon University, through the Virtualized Reality Project. 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 in regard 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.