A Teaching Assistant for Teacher-Student Learning: Knowledge Transfer from Skeleton to Inertial Sensing for Activity Recognition in Industrial Domains

Hongyin Qiao
Graduate School of Information Science and Technology
The University of Osaka
Osaka, Japan
u976812c@ecs.osaka-u.ac.jp

Qingxin Xia
Graduate School of Information Science and Technology
The University of Osaka
Osaka, Japan
xia.qingxin@ist.osaka-u.ac.jp

Hamada Rizk
Graduate School of Information Science and Technology
The University of Osaka, Osaka, Japan
RIKEN-CCS, Kobe, Japan
hamada_rizk@ist.osaka-u.ac.jp

Takuya Maekawa
Graduate School of Information Science and Technology
The University of Osaka
Osaka, Japan
maekawa@ist.osaka-u.ac.jp

***Abstract*—Human activity recognition (HAR) in industrial domains is important for workflow optimization, throughput estimation, and bottleneck detection. Skeleton-based models achieve high HAR accuracy by exploiting rich spatial and temporal cues, but they are difficult to deploy in industrial sites due to occlusions, camera placement, and privacy concerns. IMU sensors, especially smartwatches, are practical for deployment but lack spatial awareness, resulting in weaker performance. This work aims to enable robust HAR using only a wrist-worn IMU by distilling knowledge from richer modalities. Knowledge distillation allows transferring information from a skeleton teacher to a single-IMU student, but the large modality gap has limited the success of prior teacher–student approaches. To address this issue, we propose a *teacher-assistant-student* (TAS) learning framework, in which a multi-IMU assistant model bridges the skeleton-based teacher and the single-IMU student. To support TAS, we develop the following techniques: (i) *Dense temporal Contrastive Learning*, aligning structural representations of skeleton and IMU segments; (ii) *Spatial Relationship Learning*, guiding models to capture spatial priors from skeleton data; and (iii) *Temporal Attention Transfer*, distilling attention patterns for key atomic actions. We further boost the robustness to behavioral variation with motion-guided IMU data diversification using physics-based simulation. Experiments on industrial HAR sensor data show that our framework consistently improves single-IMU recognition across diverse operational scenarios, highlighting its potential for practical deployment.**

***Index Terms*—Industrial HAR, Knowledge Distillation**

I. Introduction

A. Background

Human activity recognition (HAR) in industrial domains has attracted growing interest, supported by the release of open datasets [1], [2] that have facilitated research within the pervasive computing research community. This surge in research is driven by practical applications of sensor-based recognition, such as workflow optimization, throughput estimation, and bottleneck detection in warehouse operations.

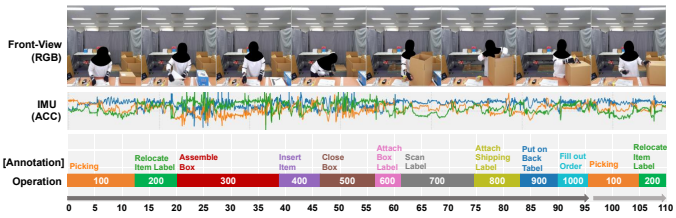

Fig. 1. Example of IMU time-series data and corresponding operation labels for a single work period in the logistics domain. For example, the "Picking" operation is composed of a sequence of atomic actions such as "Pick Up Sheet" and "Pick Up Item from Box."

Figure 1 shows an example from the OpenPack dataset [1], which provides time-series data from body-worn IMUs collected during human work activities, along with annotated *operation labels*. Each recording session consists of a series of *work periods*, during which the subject completes the packing of individual shipping boxes by sequentially performing predefined operations such as assembling, scanning, and labeling.

B. Problems and Goal

HAR in industrial settings is challenging due to the complexity of sensor data and the variability of human behavior. As shown in Figure 1, each operation consists of multiple atomic actions, leading to complex temporal patterns in the sensor data. In addition, worker behavior often changes between training and test conditions due to factors such as increased workload, growing task familiarity, or product variety, making it difficult to ensure sufficient coverage during training.

Within the pervasive computing community, skeleton- and inertial-based HAR methods have been actively studied [3]–[5]. Skeleton data offer rich spatial and temporal context, and are generally more robust to the aforementioned behavioral variations, yielding higher recognition performance. In contrast, IMU-based models lack spatial awareness and are more susceptible to changes in motion patterns, often resulting in lower accuracy.

Despite their advantages, skeleton-based systems are difficult to deploy in industrial sites due to occlusions caused by various obstacles in the industrial environment, tracking failures when workers move out of view, privacy concerns, and large data sizes of RGB images. In contrast, IMU sensors—particularly those in off-the-shelf smartwatches—are easy to wear, widely available, and capable of real-time transmission via a smartphone.

To address these challenges, we adopt a teacher-student framework that leverages synchronized skeleton and IMU data in a controlled training environment. A skeleton-based model is first trained as a teacher and used to guide training of a student model that relies only on single-IMU data from a smartwatch. This approach permits us to develop a robust, deployable IMU-based HAR model for real-world use.

However, transferring knowledge from a skeleton-based model to a single-IMU model is non-trivial due to the significant modality gap between them, and their latent feature spaces also differ substantially. While prior studies have explored teacher-student frameworks for HAR, few have addressed this challenge in cross-modal learning from skeleton to inertial data [6].

C. Approach

1) Teacher-assistant-student Learning: To bridge the gap between teacher and student, we propose a *teacher-assistant-student learning* (TAS learning) framework. In a controlled training environment, where both skeleton and multi-IMU data (i.e., IMUs attached to multiple body parts) are collected, we introduce a multi-IMU model as an intermediate *teaching assistant*. This model provides richer spatial information and has a latent representation naturally positioned between those of the skeleton-based teacher and the single-IMU student, thereby acting as a bridge that enables effective knowledge transfer from teacher to student.

To support this framework, we propose the following complementary techniques, each targeting a different aspect of complex industrial operations:

(i) Knowledge transfer with dense temporal contrastive learning aligns skeleton and multi-IMU representations by enforcing temporal level consistency through contrastive learning. Unlike prior contrastive methods for time-series [7], [8], which typically operate in point or window-based multimodal settings, we perform contrastive alignment *densely within each window*. Specifically, for each input window, we randomly sample a set of time points and apply an InfoNCE objective to align the assistant and teacher representations at the corresponding timestamps. This dense temporal sampling encourages the assistant to capture fine-grained temporal dynamics encoded in the teacher, and enables effective cross-modal knowledge transfer.

(ii) Spatial relationship learning compensates for the lack of explicit spatial structure in IMU data. While skeleton-based models naturally encode body-part relationships through keypoint graphs, IMU-based models lack this capability. To address this, we introduce auxiliary tasks during assistant/student training that estimate spatial relationships from IMU signals, guided by the teacher's skeleton data. In addition, the richer spatial information learned by the assistant is subsequently transferred to the student, enhancing its representation power.

(iii) Temporal attention transfer facilitates the sharing of temporal saliency cues across models. We observe that certain atomic actions are critical for identifying complex operations, and such key moments often manifest consistently across modalities. For example, an action of stretching packing tape when assembling shipping boxes is a characteristic and necessary action in the "assemble box" operation. By implementing temporal attention modules in the teacher, assistant, and student models, and transferring the teacher's attention patterns to the others, we guide the models to focus on discriminative time steps.

(iv) Feature distillation Due to the substantial modality gap, direct feature distillation across heterogeneous modalities is often suboptimal, as feature representations are highly modality-dependent. Consequently, we apply feature distillation exclusively between the multi-IMU assistant and the single-IMU student, leveraging their shared sensor modality to ensure more reliable knowledge transfer.

2) Motion-guided Diverse IMU Data Generation: In addition to the TAS learning framework, we address the challenge of behavioral variability in industrial tasks—such as changes in motion speed during busy seasons, growing task familiarity, and item diversity—by introducing a data diversification technique for training IMU data. While prior work on IMU data augmentation has explored methods such as jittering and scaling [9], these approaches are limited in simulating realistic variations in motion dynamics and joint movements.

To overcome this, we propose a physically plausible data diversification method, termed *motion-guided diverse IMU data generation*. Our approach leverages 3D body motion data extracted from skeleton data and applies controlled perturbations, including variations in motion speed and random joint-angle modifications, to generate diversified motion sequences. We then simulate IMU signals from these sequences using a physics-based model, assuming virtual sensors are attached to the surface of the 3D body. The synthesized IMU data are incorporated into the training set to improve the robustness of the single-IMU model.

D. Contributions

The research contributions of this study are as follows: (i)We propose a *cross-modal teacher-assistant-student learning* framework that, for the first time, introduces an assistant into teacher–student learning, enabling robust knowledge

transfer from skeleton-based models to single-IMU models via an intermediate multi-IMU assistant model, which is the main contribution of this study. (ii) We introduce *motion-guided diverse IMU data generation*, a physically plausible augmentation method that synthesizes diverse IMU signals using perturbed 3D body motion and physics-based simulation. (iii) We conduct comprehensive experiments using real-world industrial data to demonstrate the effectiveness of the proposed methods under behavioral variation.

II. Related Work

A. Activity Recognition in Industrial Domains

Sensor-based HAR in industrial environments has become increasingly important for streamlining manual processes in logistics and manufacturing. Unlike simple daily activities such as walking or sitting, industrial work involves complex, multi-step operations composed of atomic actions as shown in Figure 1. To deal with the complexities, prior studies have explored both supervised and unsupervised methods. In [10]–[12], motif-based approaches have been used to detect atomic actions in unsupervised settings. CNN-based models have also been applied to IMU data to recognize sequential operations [13]. Recently, lightweight temporal models and attention-based architectures have shown promise in segmenting and classifying complex work procedures [5], [14], [15].

Our work builds on this line of research by introducing a novel framework for wearable-sensor-based activity recognition that transfers knowledge from skeleton-based models to IMU-based models, improving robustness against behavior variations common in industrial domains.

B. Knowledge Transfer for Activity Recognition

Knowledge transfer has been actively studied in HAR to address the challenges of data scarcity and robustness. Primary directions include transfer across workers and across sensor modalities. Transfer across workers aims to adapt models trained on certain individuals to target users [16]–[18]. Transfer across modalities seeks to leverage a modality with rich information to improve recognition in another, more practical but limited modality [19], [20]. Knowledge transfer from skeleton-based models to IMU-based models has been relatively underexplored due to the large modality gap [6].

Recent studies have started to narrow the pose and inertial gap via multimodal representation learning and pretraining, e.g., by aligning motion features across modalities and leveraging large-scale or synthesized motion-sensor pairs [21]–[23]. While promising, these approaches mainly target general-purpose cross-modal embeddings and often rely on abundant paired data or additional modalities during pretraining, rather than explicitly tackling the extreme gap between full-body skeleton teachers and single-wrist IMU students in real industrial deployments. In contrast, our work addresses this challenge by introducing an intermediate multi-IMU model to bridge the modalities. This design is inspired by the teacher-assistant distillation paradigm, which introduces an intermediate assistant to ease knowledge transfer under large teacher–student gaps [24].

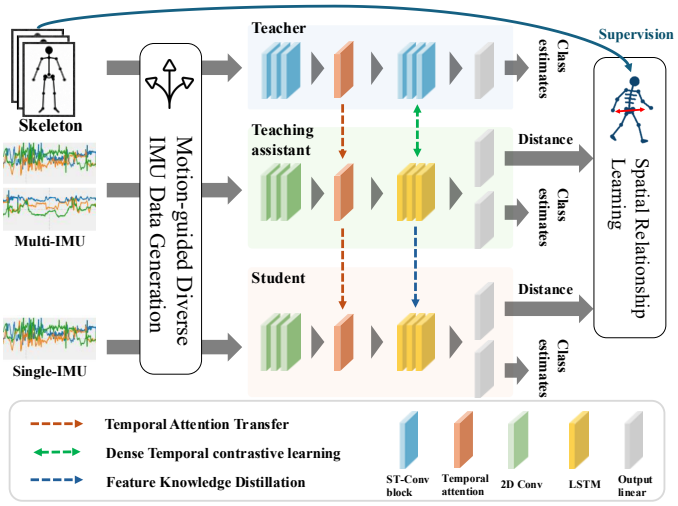

Fig. 2. Overview of the proposed method

C. Data Augmentation for Activity Recognition

Data augmentation is a common technique in IMU-based HAR to improve robustness under limited training data. Simple methods such as jittering, scaling, and time warping are widely used to simulate noise or minor motion variations [9]. However, they fall short in modeling realistic behavioral changes, such as varying movement speed or style. To overcome this, recent studies have explored physically plausible IMU augmentation using motion priors or physics-based simulation [25]. These methods aim to generate diverse yet realistic data by perturbing body motion while preserving physical consistency. In addition, recent studies on physically plausible IMU augmentation leverage simulation based on skeleton data. These methods first generate perturbed skeleton data, and then generate augmented IMU data with physics-based simulation [26]–[28]. While these studies mainly focus on perturbing joint angles in skeletons, this study also focuses on perturbing motion speed to deal with changes in motion speed during busy seasons and growing task familiarity.

III. Knowledge Transfer and Augmentation for IMU-based HAR

A. Overview

Figure 2 illustrates an overview of the training phase of the proposed framework. The training phase involves three models: a skeleton-based teacher, a multi-IMU-based teaching assistant, and a single-IMU-based student. In a controlled training environment, we collect synchronized skeleton and multi-IMU data.

To enhance data diversity, we first apply data augmentation to the collected skeleton sequences and then generate physically plausible IMU data using motion-guided simulation. These diversified IMU samples are added to the training data for the IMU model.

The teacher model is trained on the skeleton data. The assistant model is trained on the multi-IMU data and learns from the teacher via temporal attention transfer, spatial relationship learning and dense temporal contrastive learning. The student model is trained solely with single-IMU data and learns from the assistant through temporal attention transfer, spatial relationship learning, and feature distillation.

During inference, *only* the student model is used. It receives input from a single wrist-worn IMU sensor, such as a smartwatch, and predicts operation labels.

B. Teacher-Assistant-Student Learning (TAS Learning)

1) Overview: In our TAS learning framework, a skeleton-based *teacher*, a multi-IMU-based *teaching assistant*, and a single-IMU-based *student* are trained in a sequential manner to facilitate effective cross-modal knowledge transfer. We first train the teacher model using the skeleton data. The teacher captures rich spatial and temporal features of complex human activities and serves as a high-level guide for the other models.

Next, the assistant model is trained using multi-IMU data. To enable knowledge transfer from the teacher, we employ three main mechanisms: (i) *Temporal attention distillation:* The assistant learns to replicate the temporal attention weights generated by the teacher, helping it focus on key moments in the activity sequence. (ii) *Spatial relationship learning* to compensate for the lack of explicit spatial structure in IMU data. (iii) *Dense temporal contrastive learning:*We use contrastive learning to align the feature representations of the assistant with those of the teacher.

Finally, the student model is trained using only single-IMU data. The student benefits from the assistant model through three forms of knowledge transfer: (i) *Temporal attention distillation:* Similar to the teacher-to-assistant transfer, the student learns attention patterns from the assistant. (ii) *Spatial relationship learning* is also introduced during student training, enabling the model to capture lower-level spatial cues from a single IMU. (iii) *Feature distillation* to directly align the feature representations of the assistant and the student. Feature distillation is applied only between the assistant and the student, as distillation across similar modalities is more reliable than distillation between skeleton and multi-IMU representations.

2) Teacher, Assistant, and Student Models: Figure 3 shows architectures of the teacher, assistant, and student models. The teacher model takes as input a window of time-series skeleton data of length T_w seconds and predicts activity labels at a resolution of T_p seconds within that window. The model architecture consists of a stack of spatio-temporal graph convolutional (ST-GCN) blocks, followed by a temporal attention module, and another set of ST-GCN blocks. The ST-GCN block contains channel-wise attention using the CBAM module [29]. The final classification is performed using a 2D convolutional layer.

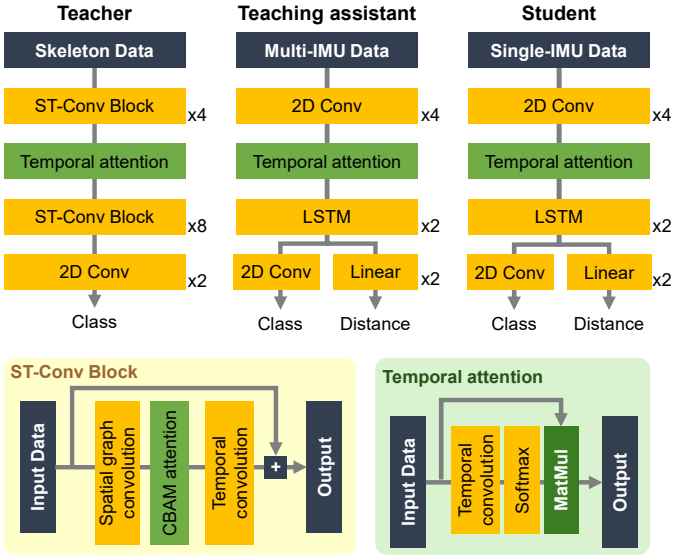

Fig. 3. Network structures of teacher, assistant, and student models

The teacher model is supervised by a standard cross-entropy loss:

$$\mathcal{L}_t = \mathcal{L}_c, \quad \mathcal{L}_c = -\frac{1}{T}\sum_{i=1}^{T}\sum_{k=1}^{K} y_{i,k}\log \hat{y}_{i,k}, \tag{1}$$

where T is the number of output time steps, K is the number of activity classes, $y_{i,k}$ is the ground-truth one-hot label for class k at time step i, and $\hat{y}_{i,k}$ is the predicted probability.

The assistant model receives multi-IMU input consisting of a T_w-second IMU window ($3 \times N_s$-axis acceleration, where N_s is the number of sensors). The architecture comprises 2D CNN layers, a temporal attention module, LSTM layers, and two output branches: one for activity classification and one for spatial relationship learning.

The assistant model is trained with a combined loss:

$$\mathcal{L}_a = \mathcal{L}_c + w_{at}\mathcal{L}_{\text{at}} + w_{sp}\mathcal{L}_{\text{spa}} + w_{cl}\mathcal{L}_{\text{cl}}, \tag{2}$$

where $w_\cdot$ is a weighting factor, $\mathcal{L}_{\text{at}}$ is the temporal attention distillation loss, and $\mathcal{L}_{\text{spa}}$ is the spatial relationship learning loss for assistants. $\mathcal{L}_{\text{cl}}$ is the dense temporal contrastive learning loss.

The student model processes a T_w-second IMU window (three-axis acceleration) as input and outputs activity labels. Its architecture mirrors that of the assistant, consisting of 2D CNN layers, a temporal attention module, LSTM layers, and two output branches.

The student is optimized using:

$$\mathcal{L}_s = \mathcal{L}_c + w_{at}\mathcal{L}_{\text{at}} + w_{sp}\mathcal{L}_{\text{sps}} + w_{kd}\mathcal{L}_{\text{kd}}, \tag{3}$$

where $\mathcal{L}_{\text{sps}}$ is the spatial relationship learning loss for students, and $\mathcal{L}_{\text{kd}}$ is the feature distillation loss.

3) Knowledge Transfer with Dense Temporal Contrastive Learning: Contrastive learning has been widely adopted for multimodal representation learning, aiming to embed features from different modalities into a shared latent space. In typical

self-supervised time-series frameworks [7], [8], multimodal contrastive learning aligns representations at the window or point level, often without relying on labels. However, in industrial HAR scenarios, fixed window sizes or point-wise alignment may be suboptimal: long windows can contain multiple atomic actions, while the point-wise alignment can mistakenly treat temporally adjacent samples from the same action as negative pairs. To address these challenges, we introduce *Dense Temporal Contrastive Learning*, which performs contrastive alignment within each time window by densely *sampling* time points. This design enables fine-grained temporal alignment across modalities, mitigates the ambiguity caused by long windows containing multiple atomic actions, and avoids assigning negative relations to temporally neighboring samples that likely correspond to the same action. Given synchronized representations from the teacher and assistant models, we randomly sample a set of time indices from each input window. For each sampled time step, the assistant and teacher features at the same temporal position are treated as positive pairs, while features from other time steps in the mini-batch serve as negatives. By enforcing temporal-level consistency, the assistant is encouraged to capture the fine-grained temporal dynamics encoded in the teacher representations. Formally, let $z^{(a)} \in \mathbb{R}^{B\times T\times C}$ and $z^{(t)} \in \mathbb{R}^{B\times T\times C}$ denote the latent representations of the assistant and teacher models, where B is the batch size, T is the window length, and C is the feature dimension. For each sample in the batch, we randomly select N time indices $\{t_1, \ldots, t_N\}$ from $\{1, \ldots, T\}$. The sampled features are flattened across the batch and time dimensions, and the InfoNCE loss is applied as:

$$\mathcal{L}_{\text{cl}} = -\frac{1}{BN}\sum_{i=1}^{BN}\log\frac{\exp\left(\text{sim}(z_i^{(a)}, z_i^{(t)})/\tau\right)}{\sum_{j=1}^{BN}\exp\left(\text{sim}(z_i^{(a)}, z_j^{(t)})/\tau\right)}, \quad (4)$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is a temperature parameter. As illustrated in Figure 2, this dense temporal contrastive loss is applied to the outputs of the teacher's final ST-GCN layer for skeleton data and the final LSTM layers of the assistant model for IMU data.

4) Spatial Relationship Learning: Inertial data from multiple body parts often lack explicit spatial structure, which can degrade the assistant model's ability to capture spatial information about body parts. To address this, we propose a spatial relationship learning module that infers inter-sensor distances, guided by the structural priors available in skeleton data. Specifically, we introduce an auxiliary task that requires the assistant model to predict the physical distances between sensor pairs at each time step, encouraging it to internalize spatial relationships during training.

Given a set of N_s IMU sensors, we define a set of N_p sensor pairs (e.g., left wrist and right wrist; Figure 4 left). For each sensor pair, the model learns to predict the physical distance between the sensors based on the observed acceleration data. During training, the assistant model is equipped with an additional output layer that produces an N_p-dimensional vector representing the predicted distances for each sensor pair at each time step.

The training objective minimizes the mean squared error (MSE) between the predicted and ground-truth distances:

$$\mathcal{L}_{\text{spa}} = \frac{1}{TN_p}\sum_{t=1}^{T}\sum_{p=1}^{N_p}\text{MSE}\left(d_{t,p}^{\text{GT}}, d_{t,p}^{\text{pred}}\right), \quad (5)$$

where $d_{t,p}^{\text{GT}}$ is the ground-truth physical distance at time step t for sensor pair p, and $d_{t,p}^{\text{pred}}$ is the corresponding predicted distance.

By minimizing $\mathcal{L}_{\text{spa}}$, the assistant model is encouraged to internalize spatial relationships among body parts, which complements its activity recognition capabilities and enhances downstream knowledge transfer to the student model.

Furthermore, we also introduce spatial relationship learning at the student level, guided by skeleton data. Since the student has access to only a single IMU sensor, it cannot infer inter-sensor distances as in the assistant. Instead, we define an auxiliary task in which the student classifies the discretized displacement of the wrist-mounted sensor over an interval of T_p time steps into C_D categories, using ground-truth motion derived from skeleton data as supervision. This auxiliary task is optimized with cross-entropy loss $\mathcal{L}_{\text{sps}}$.

This task enables the student to acquire lower-level spatial cues from a single IMU, complementing the higher-level inter-sensor relationships transferred from the assistant. By combining both sources of spatial knowledge, the student is able to improve its representation learning and achieve more effective activity recognition.

5) Feature Knowledge Distillation: To ensure that the single-IMU student model inherits the rich spatial-temporal representations from the multi-IMU assistant, we employ Feature Knowledge Distillation. Instead of directly regressing raw feature values, we align the latent distributions of the two models using Kullback-Leibler (KL) Divergence to capture the underlying relational structure of the feature space.

Specifically, let $\mathbf{f}_a$ and $\mathbf{f}_s$ denote the latent features from the assistant and student models, respectively. We first transform these features into softened probability distributions using a Softmax operation with a temperature parameter τ. The distillation loss L_{feat} is defined as:

$$\mathcal{L}_{\text{kd}} = \tau^2 \cdot \mathcal{D}_{\text{KL}}\left(\sigma\left(\frac{\mathbf{f}_s}{\tau}\right) \parallel \sigma\left(\frac{\mathbf{f}_a}{\tau}\right)\right) \quad (6)$$

where $\sigma(\cdot)$ denotes the Softmax function. This allows the student to learn knowledge and fine-grained correlations encoded in the assistant's feature maps. By minimizing this divergence, the student model is forced to mimic the assistant's feature activation patterns, effectively compensating for its lack of multi-sensor spatial awareness during inference.

6) Temporal Attention Transfer: To enhance the temporal focus of the student and assistant models, we transfer temporal attention weights from the teacher model, inspired by [5], [30]. We assume that the teacher model can better attend to

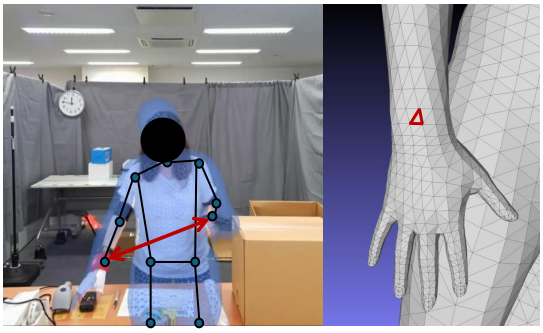

Fig. 4. Left: An example skeleton data and the distance between the two hands used in spatial relationship learning (red arrow). Right: An example extracted SMPL mesh and sensor attachment position (red polygon).

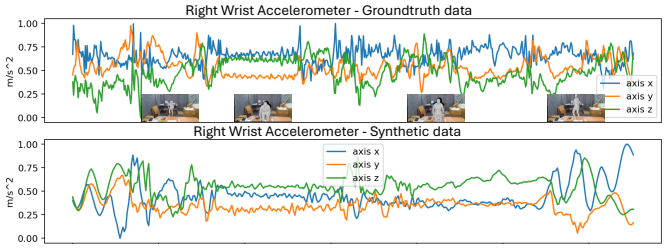

Fig. 5. Comparison between real and synthetic IMU. Key frames show that synthetic IMU generated under unoccluded views, or when the subject is close to the camera, aligns more closely with real wrist measurements than sequences affected by occlusion or long camera distance, highlighting the value of complementary on-body IMU sensing for HAR. Overall, The synthetic data, generated by our motion-guided simulation, capture realistic motion dynamics and overall temporal trends, while expected differences arise from occlusion and sensor placement. These complementary synthetic data add diversity to the training data and enhance model robustness.

key atomic actions within complex operations due to the rich spatial and contextual cues in skeleton data.

We denote the attention maps from the teacher and assistant as $A_t \in \mathbb{R}^{T\times T\times M}$ and $A_a \in \mathbb{R}^{T\times T\times M}$, respectively, where T is the number of time steps in a window and M is the number of attention heads. For each attention output, we then average along the temporal and head (M) dimensions to obtain a time series $\bar{A} \in \mathbb{R}^{T\times 1}$ indicating the importance of each time step. Finally, we minimize the mean squared error between these two time-series data to align their temporal focus:

$$\mathcal{L}_{\text{at}} = \text{MSE}(\bar{A}_t, \bar{A}_a), \tag{7}$$

where $\text{MSE}(\cdot,\cdot)$ denotes the mean squared error function. This encourages the assistant (and subsequently the student) to attend to temporally salient actions in a manner similar to the teacher. The same procedure is also applied between the assistant and student models to transfer temporal attention patterns.

C. Motion-Guided Diverse IMU Data Generation

1) Motion Diversification: To improve the generalization capability of the IMU models to behavioral variations, we propose a pipeline that generates diversified yet physically plausible IMU training data. The pipeline begins with temporal and spatial diversification of human motion sequences.

We first apply speed perturbation to the original skeleton sequences by randomly scaling the temporal axis, introducing variability in motion pace. These diversified skeletons are then converted to 3D mesh sequences using a pretrained SMPL reconstruction pipeline. Specifically, we employ WHAM [31], which reconstructs accurate 3D human motion from 2D keypoint trajectories in a world-grounded coordinate frame.

To further introduce realistic variability, we perturb joint angles in the SMPL mesh. For each time window, random angular offsets are applied to selected joints, simulating subtle but realistic changes in movement style and articulation.

2) IMU Data Simulation and Calibration: We generate virtual IMU signals from the diversified SMPL meshes using a physics-based simulation approach. Following the IMUSim framework [32], we attach a virtual IMU sensor to a specific body surface on the 3D mesh, specifying both position (polygon surface as shown in Figure 4) and orientation (sensor axes). Each sensor's local motion trajectory is extracted from the mesh, and the corresponding accelerometer readings are simulated using rigid body dynamics.

However, there is inevitably a misalignment between the orientation of the virtually attached sensors on the SMPL mesh and that of the real sensors used in practice. To resolve this, we perform a calibration step using short segments of paired skeleton and accelerometer data. We employ the Kabsch algorithm [33], which computes the optimal rotation matrix that minimizes the sum of squared errors between two sets of corresponding 3D points across time steps.

Formally, let $X = [x_1, \ldots, x_n]$ and $Y = [y_1, \ldots, y_n]$ be two sets of 3D points from the simulated and real IMU data, respectively, with their centroids subtracted. The covariance matrix is defined as $C = X^{\mathsf{T}}Y$, and its singular value decomposition is given by $C = U\Sigma V^{\mathsf{T}}$. The optimal rotation matrix is then:

$$R = VU^{\mathsf{T}}. \tag{8}$$

This pre-computed rotation R is applied to align the simulated IMU signals with the real sensor orientation, ensuring consistency between virtual and real measurements.

The resulting synthetic IMU signals reflect realistic motion patterns with diverse dynamics. Figure 5 illustrates examples of synthetic IMU signals alongside real IMU signals for comparison, showing that the generated data preserve realistic temporal structures. These data are added to the training set to improve the robustness of the IMU models against unseen behavior patterns.

IV. Evaluation

A. Datasets

Public industrial datasets containing both IMU and skeleton (or RGB) data are scarce; to our knowledge, only the OpenPack dataset [1] and the LARa dataset [2] meet these criteria. OpenPack includes data that simulate a variety of situations

(e.g., busy seasons), whereas LARa collects all periods under a single fixed setting and thus cannot evaluate the impact of sensor-data variation between training and testing. Therefore, we mainly use OpenPack for evaluation in this study.

1) OpenPack Dataset: The OpenPack dataset [1] is a recently released, publicly available dataset comprising 58 hours of multimodal recordings from 16 subjects performing packaging operations in a precisely simulated logistics workstation. This dataset contains 10 operations (activity classes) to recognize: Picking, Relocate Item Labels, Assemble Box, Insert Items, Close Box, Attach the Box Label, Scan Label, Attach Shipping Label, Put on the Back Table, and Fill out Order Sheet. The dataset includes four distinct work scenarios: **Scenario 1 (S1)**: The simplest setting, where workers followed the work instructions as accurately as possible. **Scenario 2**: In some logistics centers, workers are not required to strictly adhere to written instructions. In this scenario, subjects were encouraged to modify the workflow for efficiency, and the variety of items to be packed was increased compared to Scenario 1. **Scenario 3**: Building on Scenario 2, workers occasionally performed irregular actions depending on the items or surrounding conditions. **Scenario 4**: Extending Scenario 3, the workload was further increased to simulate seasonal events or flash sales. Subjects were prompted to work faster by introducing an auditory alarm to create a high-pressure environment. Further details of these scenarios are provided in [1].

We used accelerometer data collected by ATR TSND151 IMU sensors, placed on the subjects' right wrists, with a sampling rate of 30 Hz as inputs to the single-IMU student model. For the multi-IMU assistant model, accelerometer data from both the right and left wrists were used. For the teacher model, we used the provided skeleton data extracted from Kinect v2 devices installed as front-view cameras.

2) LARa Dataset: The LARa dataset [2] is a publicly available resource containing multiple data types, including accelerometer recordings from 14 subjects performing product picking and packaging in three different simulated scenarios. The accelerometer data were collected using MbientLab MMRL IMUs. For our experiments, we used data from the dominant (right) wrist of six workers who participated in Scenario 3, comprising a total of 57 work periods (1.9 hours of data). On average, each work period in these scenarios lasted approximately 120 seconds. These scenarios were selected due to their similarity to the actions involved in traditional packing processes in logistics centers. The dataset includes six activity classes to recognize: Standing, Walking, Cart, Handling (upward), Handling (centered), and Handling (downward). Similar to the OpenPack dataset, we used the right-wrist accelerometer data as input to the student model, and both-wrist data as input to the assistant model.

B. Evaluation Methodology

To evaluate the robustness of our method against changes in behavioral variations between training and testing, we first focus on worker-dependent evaluation where a model is trained on labeled data from workers containing a target worker.

The OpenPack dataset provides scenario-based data collection, enabling us to investigate changes in behavioral variations. We employ the first session, which contains twenty work periods, in Scenario 1 (the most basic setting) from all workers for training (with 10% of it further reserved for validation), and the remaining for testing. The test set thus consists of four subsets: the remaining periods of Scenario 1, Scenario 2, Scenario 3, and Scenario 4, each introducing increasing levels of variability as described in Section IV-A. The duration of each period is about 100 seconds.

For the LARa dataset, which does not provide scenario-based segmentation, we adopt a temporal split for each worker: the last 2 periods are used for testing and the remaining periods for training.

C. Comparative Methods

To evaluate the effectiveness of TAS learning, we compare it with the following methods including methods with state-of-the-art knowledge distillation (KD) objectives and tailored to KD from skeleton to IMU.

- **TS with Logits**: This method simply employs teacher-student learning from skeleton to IMU. The classical approach minimizes the Kullback–Leibler divergence between the softened output distributions of the teacher and student networks, thereby transferring class-level knowledge [34].
- **TS with Relational Knowledge Distillation (RKD)**: This method employs teacher-student learning from skeleton to single-IMU with RKD. RKD captures structural information by enforcing consistency in the pairwise distances and angles among samples in the feature space of teacher and student to preserve relational geometry [35].
- **TS with Contrastive Representation Distillation (CRD)**: This method employs teacher-student learning from skeleton to single-IMU with CRD. CRD leverages a contrastive InfoNCE loss to bring the student's representation closer to that of the teacher while simultaneously pushing away negative samples, encouraging discriminative embedding alignment [36].
- **TS with Probabilistic Knowledge Transfer (PKT)**: This method employs teacher-student learning from skeleton to IMU with PKT. PKT aligns teacher and student by minimizing the divergence between their pairwise similarity distributions in the feature space, thus capturing probabilistic relationships among samples [37].
- **TS with Self-Supervised Quantization-Aware Knowledge Distillation (SQAKD)**: This state-of-the-art method employs teacher-student learning from skeleton to single-IMU. SQAKD leverages the teacher's penultimate features and guides the student by minimizing the KL divergence between their softened feature distributions, enabling efficient transfer of representation-level knowledge [38].
- **TS with Customized Crossmodal Knowledge Distillation (C2KD)**: This state-of-the-art method employs teacher-student learning across skeleton and single-IMU. C2KD bridges the modality gap by introducing proxy networks and bidirectional distillation. The student not only mimics the teacher's predictions but also benefits from intra-modal proxy alignment, while

cross-modal knowledge transfer is adaptively filtered through on-the-fly selection using Kendall rank correlation. [39]

• **PSKD**: This is a state-of-the-art method tailored to KD from skeleton to IMU introduced in the related work section [6]. This is a Progressive Skeleton-to-Sensor Knowledge Distillation (PSKD) framework that transfers rich skeleton information to accelerometer-based HAR. In this framework, two teacher models, trained on skeleton-only and skeleton–sensor data, progressively guide the student model.

• **Multi-single with Logits**: This method employs KD from multi-IMU to single-IMU using the logit distillation. This is used to confirm the effectiveness of skeleton data.

In addition to the above methods, we prepared the following reference methods.

• **Skeleton only**: This method simply employs skeleton data for activity recognition. The network architecture is almost identical to our teacher model.

• **Multi-IMU only**: This method simply employs multi-IMU data for activity recognition. The network architecture is almost identical to our assistant model.

• **Single-IMU only**: This method simply employs single-IMU data for activity recognition. The network architecture is almost identical to our student model.

The above methods do not perform motion-guided diverse IMU data generation. The hyperparameters and experimental configurations are summarized in Table I. We employ an uncertainty-based learnable loss weighting scheme [40], which automatically adjusts the relative importance of each loss term. This approach avoids the need for manual weight tuning.

TABLE I
SUMMARY OF EXPERIMENTAL PARAMETERS AND TRAINING SETTINGS.

Parameter / Setting	Value / Description
Batch size(B)	32
Input window length (T_w)	450
Input window overlap	225
Resolution of activity label prediction (T_p)	1 s
Temperature parameter (τ)	5
Number of attention heads (M)	4
Number of distance classes (C_D)	3
Loss weighting scheme	Uncertainty-based [40]
Evaluation metric	Macro-averaged F1 score

D. Results

Table II shows the results. Before presenting the results of our proposed method, we first examine the recognition performance of the Skeleton only method. Scenario S1 achieves the highest accuracy, which gradually decreases through S2 and S3 and drops by about 0.17 in S4. In contrast, the single-IMU baseline starts roughly 0.077 lower than the skeleton model in S1 and suffers a dramatic decline in S4, performing more than 0.18 below the skeleton model and highlighting its lack of robustness.

Our proposed method substantially improves performance in all scenarios, surpassing the skeleton model in S2 and S3, consistently outperforming the multi-IMU model across all scenarios, and boosting the F1-score by about 21% over the single-IMU baseline in S4. It consistently achieves the best results across all scenarios and clearly outperforms state-of-the-art comparative methods.

TABLE II
COMPARISON OF OUR METHOD WITH STATE-OF-THE-ART BASELINES ON FOUR EVALUATION SETTINGS (S1–S4) IN THE OPENPACK DATASET. IMPROVEMENTS OVER THE SINGLE-IMU BASELINE ARE SHOWN IN PARENTHESES. ∗ DENOTES THE SINGLE-IMU BASELINE.

Method	S1	S2	S3	S4
Skeleton only	0.887	0.815	0.726	0.717
Multi-IMU only	0.820	0.812	0.713	0.592
Single-IMU only*	0.810	0.774	0.692	0.536
TS with Logits	0.808 (↓0.3%)	0.781 (↑0.9%)	0.705 (↑1.9%)	0.575 (↑7.3%)
TS with RKD [35]	0.788 (↓2.7%)	0.749 (↓3.2%)	0.654 (↓5.5%)	0.508 (↓5.2%)
TS with CRD [36]	0.809 (↓0.1%)	0.781 (↑0.9%)	0.708 (↑2.3%)	0.543 (↑1.3%)
TS with PKT [37]	0.815 (↑0.6%)	0.781 (↑0.9%)	0.695 (↑0.4%)	0.544 (↑1.5%)
TS with SQAKD [38]	0.823 (↑1.6%)	0.806 (↑4.1%)	0.719 (↑3.9%)	0.550 (↑2.6%)
TS with C2KD [39]	0.811 (↑0.1%)	0.795 (↑2.7%)	0.706 (↑2.0%)	0.539 (↑0.6%)
PSKD [6]	0.809 (↓0.1%)	0.799 (↑3.2%)	0.698 (↑0.9%)	0.529 (↓1.3%)
Multi-single with Logits	0.811 (↑0.1%)	0.789 (↑1.9%)	0.696 (↑0.6%)	0.553 (↑3.2%)
TAS (Ours)	**0.841** (↑3.8%)	**0.838** (↑8.3%)	**0.761** (↑9.9%)	**0.650** (↑21.3%)
TAS (w/o DA)	0.838 (↑3.5%)	0.816 (↑5.4%)	0.742 (↑7.2%)	0.611 (↑14.0%)

Traditional teacher–student learning with logits (TS with Logits) fails to effectively transfer knowledge because of the large modality gap between skeleton and single-IMU data.

Skeleton-to-IMU KD (SQAKD) produce modest gains over the single-IMU baseline in all scenarios, yet still fall short of the performance achieved by our method overall.

1) Ablation Study: To evaluate the contributions of the proposed method, we have prepared the following methods that do not perform data augmentation (motion-guided diverse data generation).

• **w/o Temporal Transfer**: This is a variant of our TAS method that does not perform temporal attention transfer.

• **w/o Spatial Learning**: This is a variant of our TAS method that does not perform spatial relationship learning.

• **w/o Feature KD**: This is a variant of our TAS method that does not perform feature knowledge distillation.

• **w/o Assistant**: This is a variant of our method that does not employ an assistant model. It directly transfers knowledge from the skeleton teacher to the single-IMU student by using the temporal attention, spatial relationship learning, and dense temporal contrastive learning.

• **w/o Teacher**: This is a variant of our method that does not use any skeleton data. It directly transfers knowledge from the assistant to the single-IMU student by leveraging temporal attention and feature knowledge distillation.

• **TAS (w/o DA)**: This is a variant of our TAS method. This method does not perform motion-guided diverse IMU data generation.

TABLE III
ABLATION STUDY: WORK OPERATION RECOGNITION PERFORMANCE (F1 SCORE) IN OPENPACK DATASET. IMPROVEMENTS COMPARED TO THE BASELINE ARE SHOWN IN PARENTHESES.

Method	S1	S2	S3	S4
Without DA (no virtual IMU data)				
Baseline	0.810	0.774	0.692	0.536
w/o Temporal Transfer	0.838 (↑3.5%)	0.814 (↑5.2%)	0.739 (↑6.8%)	0.591 (↑10.3%)
w/o Spatial Learning	0.828 (↑2.2%)	0.800 (↑3.4%)	0.719 (↑3.9%)	0.574 (↑7.1%)
w/o Feature KD	0.822 (↑1.5%)	0.797 (↑3.0%)	0.701 (↑1.3%)	0.579 (↑8.0%)
w/o Assistant	0.834 (↑3.0%)	0.813 (↑5.0%)	0.725 (↑4.8%)	0.571 (↑6.5%)
w/o Teacher	0.809 (↓0.1%)	0.776 (↑0.3%)	0.693 (↑0.1%)	0.544 (↑1.5%)
TAS (w/o DA)	0.838 (↑3.5%)	0.816 (↑5.4%)	**0.742** (↑7.2%)	0.611 (↑14.0%)
Without Teacher-Assistant-Student Learning				
Single w/ DA	0.816 (↑0.7%)	0.799 (↑3.2%)	0.700 (↑1.2%)	0.614 (↑14.6%)
No Ablation				
TAS (Ours)	**0.841** (↑3.8%)	**0.838** (↑8.3%)	0.761 (↑10.0%)	**0.650** (↑21.3%)

In addition to them, we have prepared an IMU-based method that does not perform TAS learning.

• **Single w/ DA**: This is a variant of Baseline employing a single-IMU model trained on the original and augmented data.

Table III shows the results. By comparing w/o Temporal Transfer, w/o Spatial Learning, and w/o Segment CL, we find that Feature KD contributes most significantly. Because direct feature alignment between similar modalities (i.e., multi-IMU to single-IMU) provides more precise guidance for the student model to capture universal behavioral representations.

Next, the spatial relationship learning shows clear benefits. The key difference between skeleton and accelerometer data is the lack of explicit spatial structure in accelerometer signals; skeleton-derived spatial knowledge was effectively transferred to the assistant and student models.

The effect of TAS learning is evident when comparing Single w/ DA with TAS: accuracy improves consistently across all scenarios, with especially large gains in S3 and S4.

The results of w/o Teacher show that without using skeleton data, the performance gap between multi-IMU and single-IMU models is relatively small, preventing the single-IMU model from learning useful knowledge from the multi-IMU model alone. This observation further validates the effectiveness of the proposed TAS framework.

Comparing w/o Assistant and TAS (w/o DA) also suggests the effectiveness of introducing the assistant across all scenarios, indicating that the multi-IMU assistant learns latent representations intermediate between those of the skeleton teacher and the single-IMU student and thereby bridges them.

The impact of motion-guided diverse data augmentation (DA) can be observed by comparing Baseline with Single w/ DA. DA yields pronounced improvements in S4, confirming that physically plausible augmentation of motion speed and joint angles enhances generalization.

TABLE IV
RESULTS OF WORKER-INDEPENDENT MODELS ON FOUR EVALUATION SETTINGS (S1–S4). IMPROVEMENTS COMPARED TO THE BASELINE ARE SHOWN IN PARENTHESES. ∗ DENOTES THE SINGLE-IMU BASELINE.

Method	S1	S2	S3	S4
Skeleton only	0.865	0.799	0.717	0.688
Multi-IMU only	0.830	0.829	0.737	0.608
Single-IMU only*	0.809	0.780	0.687	0.476
TS with Logits	0.799 (↓1.2%)	0.763 (↓2.2%)	0.672 (↓2.2%)	0.472 (↓0.8%)
TS with RKD [35]	0.782 (↓3.3%)	0.754 (↓3.3%)	0.658 (↓4.2%)	0.445 (↓6.5%)
TS with CRD [36]	0.805 (↓0.5%)	0.776 (↓0.5%)	0.694 (↑1.0%)	0.510 (↑7.1%)
TS with PKT [37]	0.816 (↑0.9%)	0.786 (↑0.8%)	0.692 (↑0.7%)	0.501 (↑5.3%)
TS with SQAKD [38]	0.831 (↑2.7%)	0.815 (↑4.5%)	0.729 (↑6.1%)	0.536 (↑12.6%)
TS with C2KD [39]	0.815 (↑0.7%)	0.780 (±0.0%)	0.689 (↑0.3%)	0.503 (↑5.7%)
PSKD [6]	0.801 (↓1.0%)	0.768 (↓1.5%)	0.677 (↓1.5%)	0.480 (↑0.8%)
Multi-single with Logits	0.801 (↓1.0%)	0.766 (↓1.8%)	0.675 (↓1.7%)	0.478 (↑0.4%)
TAS (Ours)	**0.866** (↑7.0%)	**0.838** (↑7.4%)	0.739 (↑7.6%)	**0.597** (↑25.4%)
TAS (w/o DA)	0.849 (↑4.9%)	0.812 (↑4.1%)	**0.743** (↑8.2%)	0.551 (↑15.8%)

Finally, comparing TAS (w/o DA) with Single w/ DA shows that introducing TAS brings greater overall benefits than DA. The assistant's advantage over DA is more marked in S3, which aligns with the previous finding that DA exerts its strongest effect in S4.

2) Worker-independent Models: Table IV shows the performance of the methods when labeled data of a target worker are excluded from the training data. Overall, the F1 scores for the worker-independent models are lower than those for the worker-dependent models. In particular, the F1 scores in S4 were notably poor in the worker-independent setting, indicating that in busy situations, i.e., S4, each user's movements exhibit strong individual characteristics. Surprisingly, most comparative methods perform worse than the single-IMU baseline. This degradation is largely driven by two interacting factors: the substantial modality gap between skeleton and IMU data and the distribution shift in sensor data between training and testing users. Notably, TAS is the only method that improves performance under this worker-independent setting. By mitigating the large modality gap between skeleton and IMU data, it likely suppresses the performance degradation caused by inter-user differences.

3) Performance in LARa Dataset: Table V shows the results of the methods and the ablation study. Because the training and test data were collected under the same condition, the data augmentation was not applied in this experiment.

The effectiveness of the proposed TAS method is also confirmed on the LARa dataset, where it achieved performance close to that of the multi-IMU model. The ablation study indicates that feature knowledge distillation contributes most significantly, which is consistent with the findings from

TABLE V
RESULTS OF WORKER-DEPENDENT MODELS IN LARA DATASET. IMPROVEMENTS COMPARED TO BASELINE ARE SHOWN IN PARENTHESES. $*$ DENOTES THE SINGLE-IMU BASELINE.

Method	F1	Method	F1
Skeleton only	0.527	TS with CRD [36]	0.430 (↑7.0%)
Multi-IMU only	0.453	TS with SQAKD [38]	0.387 (↓3.7%)
Single-IMU only $*$	0.402	TS with Logits	0.363 (↓9.7%)
TS with RKD [35]	0.387 (↓3.7%)	TS with PKT [37]	0.388 (↓3.5%)
TS with C2KD [39]	0.384 (↓4.5%)	PSKD [6]	0.373 (↓7.2%)
Multi-single with Logits	0.405 (↑0.7%)		
w/o Temporal Transfer	<u>0.443</u> (↑10.2%)	w/o Spatial Learning	0.439 (↑9.2%)
w/o Feature KD	0.423 (↑5.2%)	**TAS (Ours)**	**0.449** (↑11.7%)

TABLE VI
ABLATION STUDY: WORK OPERATION RECOGNITION PERFORMANCE (F1 SCORE) IN OPENPACK DATASET. IMPROVEMENTS COMPARED TO THE MULTI-IMU BASELINE ARE SHOWN IN PARENTHESES.

Method	S1	S2	S3	S4
Without DA (no virtual IMU data)				
Baseline	0.820	0.812	0.713	0.592
w/o Temporal Transfer	0.840 (↑2.4%)	0.833 (↑2.6%)	0.738 (↑3.5%)	0.635 (↑7.3%)
w/o Spatial Learning	0.840 (↑2.4%)	0.840 (↑3.4%)	<u>0.758</u> (↑6.3%)	0.642 (↑8.4%)
w/o Dense CL	0.839 (↑2.3%)	0.838 (↑3.2%)	0.742 (↑4.1%)	0.654 (↑10.5%)
TA (w/o DA)	<u>0.850</u> (↑3.7%)	<u>0.844</u> (↑3.9%)	0.756 (↑6.0%)	<u>0.663</u> (↑12.0%)
Without Teacher-Assistant Learning				
Multi w/ DA	0.841 (↑2.6%)	0.799 (↓1.6%)	0.744 (↑4.3%)	0.659 (↑11.3%)
No Ablation				
TA (Ours)	**0.852** (↑3.9%)	**0.850** (↑4.7%)	**0.774** (↑8.6%)	**0.678** (↑14.5%)

OpenPack. The temporal attention transfer proposed in [30] was not effective in this dataset, likely because the skeletons appear relatively small within the image frames, making it difficult to capture fine-grained motions.

4) Ablation Study of Assistant model: Table VI presents the ablation study of the Assistant model. Removing any individual component consistently degrades performance, indicating that each module contributes to the final performance in a complementary manner.

5) Computation Cost: Because the deployed single-IMU model processes data from only one accelerometer, it is computationally efficient. We measured the prediction time of the skeleton, multi-IMU, and single-IMU models on NVIDIA RTX PRO 6000 (Blackwell) GPUs. For an input data window of 15 s, the average prediction times were 4.6215 ms, 0.5542 ms, and 0.5480 ms, respectively. The single-IMU model's prediction time is substantially shorter than that of the skeleton model, confirming that it can run comfortably even on resource-constrained devices.

E. Discussion

One limitation of the proposed method is the cost of collecting multi-IMU data, since a single smartphone used for data storage or transmission can generally be paired with only one smartwatch. However, the collection cost is lower than that of skeleton data and is required only in the training environments. Since the testing phase requires only single-IMU data, the burden on actual work sites remains minimal.

Although the absolute performance on S4 in OpenPack remains in the low 65% range, this result is still valuable. S4 is the most challenging condition in the OpenPack dataset, yet our method achieves over 20% relative improvement compared to the baseline in the user dependent setting, showing it can capture meaningful signals even under difficult conditions. The consistent gains across S1–S4 further demonstrate robustness and generalizability, indicating that the method works reliably beyond favorable cases.

V. CONCLUSION

We presented a teacher–assistant–student learning framework that transfers rich spatio-temporal knowledge from skeleton data to a single-IMU model through a multi-IMU assistant. Dense temporal contrastive learning, spatial relationship learning, and motion-guided data augmentation jointly improved cross-modal alignment and robustness. Extensive experiments showed that our method consistently outperformed state-of-the-art baselines across all scenarios while requiring only a single IMU sensor during deployment. In future work, we plan to extend the framework to incorporate other sensing modalities as teachers to further enhance generalizability and robustness.

ACKNOWLEDGEMENT

This study is partially supported by JST K Program Grant Number JPMJKP25V2, and JSPS KAKENHI Grant Number JP21H05299 and JP25K22801, Japan.

We also sincerely thank Lala Shakti Swarup Ray (DFKI) for the helpful discussion and for suggesting the use of the WHAM model for pose extraction.

REFERENCES

[1] N. Yoshimura, J. Morales, T. Maekawa, and T. Hara, "Openpack: A large-scale dataset for recognizing packaging works in IoT-enabled logistic environments," in *2024 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 2024, pp. 90–97.

[2] F. Niemann, C. Reining, F. Moya Rueda, N. R. Nair, J. A. Steffens, G. A. Fink, and M. Ten Hompel, "Lara: Creating a dataset for human activity recognition in logistics using semantic attributes," *Sensors*, vol. 20, no. 15, p. 4083, 2020.

[3] Y. Zhang and T. Maekawa, "InterHandNet: Capturing two-hand interaction for robust hand-washing activity recognition," in *2025 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2025, pp. 13–24.

[4] T. Kumrai, J. Korpela, T. Maekawa, Y. Yu, and R. Kanai, "Human activity recognition with deep reinforcement learning using the camera of a mobile robot," in *2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 2020, pp. 1–10.

[5] J. Morales, N. Yoshimura, Q. Xia, A. Wada, Y. Namioka, and T. Maekawa, "Acceleration-based human activity recognition of packaging tasks using motif-guided attention networks," in *2022 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2022, pp. 1–12.

[6] J. Ni, A. H. Ngu, and Y. Yan, "Progressive cross-modal knowledge distillation for human action recognition," in *the 30th ACM International Conference on Multimedia (MM'22)*, 2022, pp. 5903–5912.

[7] H. Yang, Z. Ren, H. Yuan, Z. Xu, and J. Zhou, "Contrastive self-supervised representation learning without negative samples for multimodal human action recognition," *Frontiers in Neuroscience*, vol. 17, p. 1225312, 2023.

[8] S. Deldari, D. Spathis, M. Malekzadeh, F. Kawsar, F. D. Salim, and A. Mathur, "Crossl: Cross-modal self-supervised learning for time-series through latent masking," in *the 17th ACM International Conference on Web Search and Data Mining*, 2024, pp. 152–160.

[9] Z. Gao, H. Liu, and L. Li, "Data augmentation for time-series classification: An extensive empirical study and comprehensive survey," *Journal of Artificial Intelligence Research*, vol. 83, 2025.

[10] T. Maekawa, D. Nakai, K. Ohara, and Y. Namioka, "Toward practical factory activity recognition: Unsupervised understanding of repetitive assembly work in a factory," ser. UbiComp '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1088–1099.

[11] Q. Xia, A. Wada, J. Korpela, T. Maekawa, and Y. Namioka, "Unsupervised factory activity recognition with wearable sensors using process instruction information," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 2, p. 60, 2019.

[12] Q. Xia, J. Korpela, Y. Namioka, and T. Maekawa, "Robust Unsupervised Factory Activity Recognition with Body-worn Accelerometer Using Temporal Structure of Multiple Sensor Data Motifs," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 3, pp. 1–30, sep 2020.

[13] A. S. Syed, Z. S. Syed, and A. K. Memon, "Continuous human activity recognition in logistics from inertial sensor data using temporal convolutions in CNN," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 10, pp. 597–602, 2020.

[14] N. Yoshimura, T. Maekawa, T. Hara, A. Wada, and Y. Namioka, "Acceleration-based activity recognition of repetitive works with lightweight ordered-work segmentation network," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 6, no. 2, Jul 2022.

[15] Q. Xia, J. Morales, Y. Huang, T. Hara, K. Wu, H. Oshima, M. Fukuda, Y. Namioka, and T. Maekawa, "Self-supervised learning for complex activity recognition through motif identification learning," *IEEE Transactions on Mobile Computing*, vol. 24, no. 5, pp. 3779–3793, 2025.

[16] X. Qin, J. Wang, Y. Chen, W. Lu, and X. Jiang, "Domain generalization for activity recognition via adaptive feature fusion," *ACM Transactions on Intelligent Systems and Technology*, vol. 14, pp. 1 – 21, 2022.

[17] X. Zhang, L. Zhou, R. Xu, P. Cui, Z. Shen, and H. Liu, "Towards unsupervised domain generalization," in *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, June 2022, pp. 4910–4920.

[18] J. Morales, Q. Xia, N. Yoshimura, H. Oshima, M. Fukuda, Y. Namioka, and T. Maekawa, "Multilevel Transfer Learning for Complex Work Activity Recognition in Logistic Domain," in *2025 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. Los Alamitos, CA, USA: IEEE Computer Society, Mar. 2025, pp. 159–170.

[19] S. Tan, T. Nagarajan, and K. Grauman, "Egodistill: Egocentric head motion distillation for efficient video understanding," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023, pp. 33 485–33 498.

[20] E. S. Jeon, H. Choi, A. Shukla, Y. Wang, H. Lee, M. P. Buman, and P. Turaga, "Topological persistence guided knowledge distillation for wearable sensor data," *arXiv preprint arXiv:2407.05315*, 2024.

[21] S. G. Fritsch, C. Oguz, V. F. Rey, L. Ray, M. Kiefer-Emmanouilidis, and P. Lukowicz, "Mujo: Multimodal joint feature space learning for human activity recognition," in *2025 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2025, pp. 1–12.

[22] V. Fortes Rey, L. S. S. Ray, Q. Xia, K. Wu, and P. Lukowicz, "Enhancing inertial hand based har through joint representation of language, pose and synthetic imus," in *Proceedings of the 2024 ACM International Symposium on Wearable Computers*, ser. ISWC '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 25–31. [Online]. Available: https://doi.org/10.1145/3675095.3676609

[23] L. S. S. Ray, Q. Xia, V. F. Rey, K. Wu, and P. Lukowicz, "Improving imu based human activity recognition using simulated multimodal representations and a moe classifier," *Frontiers in Computer Science*, vol. Volume 7 - 2025, 2025.

[24] S.-I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 4, 2020, pp. 5191–5198.

[25] N. Oishi, D. Roggen, P. Birch, and P. Lago, "Simuaug: Variability-aware data augmentation for wearable imu using physics simulation," in *Activity, Behavior, and Healthcare Computing*. CRC Press, 2023, pp. 16–38.

[26] L. Huang and C. Xia, "Modifyaug: Data augmentation for virtual imu signal based on 3d motion modification used for real activity recognition," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '24. New York, NY, USA: Association for Computing Machinery, 2024.

[27] Z. Li, C. Yu, C. Liang, and Y. Shi, "Poseaugment: Generative human pose data augmentation with physical plausibility for imu-based motion capture," 2024. [Online]. Available: https://arxiv.org/abs/2409.14101

[28] C. Xia and Y. Sugiura, "Virtual imu data augmentation by spring-joint model for motion exercises recognition without using real data," in *the 2022 ACM International Symposium on Wearable Computers*, ser. ISWC '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 79–83.

[29] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[30] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *International Conference on Learning Representations (ICLR)*, 2017.

[31] S. Shin, J. Kim, E. Halilaj, and M. J. Black, "Wham: Reconstructing world-grounded humans with accurate 3d motion," in *the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2070–2080.

[32] A. D. Young, M. J. Ling, and D. K. Arvind, "Imusim: A simulation environment for inertial sensing algorithm design and evaluation," in *the 10th ACM/IEEE International Conference on Information Processing in Sensor Networks*. IEEE, 2011, pp. 199–210.

[33] W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A*, vol. 32, no. 5, pp. 922–923, 1976.

[34] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[35] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3967–3976.

[36] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," in *International Conference on Learning Representations (ICLR)*, 2020. [Online]. Available: https://openreview.net/forum?id=SkgpBJrtvS

[37] N. Passalis and A. Tefas, "Learning deep representations with probabilistic knowledge transfer," in *the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 268–284.

[38] K. Zhao and M. Zhao, "Self-supervised quantization-aware knowledge distillation," in *the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

[39] Y. Huo, Z. Jiang, W. Wang, Y. Liu, P. Gao, and M.-M. Cheng, "C2kd: Bridging the modality gap for cross-modal knowledge distillation," in *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 23 705–23 714.

[40] G. Y. . C. R. Kendall, A., "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics." in *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7482–7491.