Decoupled Heterogeneous Multi-Teacher Mutual Distillation for Wearable Human Activity Recognition

ZHIWEN XIAO, School of Computing and Artificial Intelligence, Southwest Jiaotong University, China Engineering Research Center of Sustainable Urban Intelligent Transportation, Ministry of Education, China
HUANLAI XING, School of Computing and Artificial Intelligence, Southwest Jiaotong University, China Engineering Research Center of Sustainable Urban Intelligent Transportation, Ministry of Education, China
RONG QU, School of Computer Science, University of Nottingham, UK
HUI LI, School of Mathematics and Statistics, Xi'an Jiaotong University, China
RUIBIN BAI, School of Computer Science, University of Nottingham Ningbo China, China
LEXI XU, Research Institute, China United Network Communications Corporation, China
JINCHENG PENG, School of Computing and Artificial Intelligence, Southwest Jiaotong University, China Engineering Research Center of Sustainable Urban Intelligent Transportation, Ministry of Education, China
LI FENG, School of Computing and Artificial Intelligence, Southwest Jiaotong University, China Engineering Research Center of Sustainable Urban Intelligent Transportation, Ministry of Education, China
QIAN WAN, School of Computing and Artificial Intelligence, Southwest Jiaotong University, China Engineering Research Center of Sustainable Urban Intelligent Transportation, Ministry of Education, China

This work was partially supported by the Natural Science Foundation of Sichuan Province (No. 2026NSFSC0433), the Ningbo Key Laboratory of Digital Port Technologies (No. NDPT2402), the University of Nottingham Ningbo China, the Doctoral Innovation Fund Program of Southwest Jiaotong University (No. CX-2025ZD09), Southwest Jiaotong University, China, the Key Laboratory of Detection and Application of Space Effect in Southwest Sichuan at Leshan Normal University, Education Department of Sichuan Province (No. YBXM202602001), and the Fundamental Research Funds for the Central Universities, P. R. China. The corresponding author is Huanlai Xing.
Authors' addresses: Zhiwen Xiao, School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 610031, China and Engineering Research Center of Sustainable Urban Intelligent Transportation, Ministry of Education, Chengdu 610031, China, xiao1994zw@163.com; Huanlai Xing, School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 610031, China and Engineering Research Center of Sustainable Urban Intelligent Transportation, Ministry of Education, Chengdu 610031, China, hxx@home.swjtu.edu.cn; Rong Qu, School of Computer Science, University of Nottingham, Nottingham NG8 1BB, UK, rong.qu@nottingham.ac.uk; Hui Li, School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China, lihui10@mail.xjtu.edu.cn; Ruibin Bai, School of Computer Science, University of Nottingham Ningbo China, Ningbo 315100, China, ruibin.bai@nottingham.edu.cn; Lexi Xu, Research Institute, China United Network Communications Corporation, Beijing 100048, China, davidlexi@hotmail.com; Jincheng Peng, School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 610031, China and Engineering Research Center of Sustainable Urban Intelligent Transportation, Ministry of Education, Chengdu 610031, China, jcpeng@my.swjtu.edu.cn; Li Feng, School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 610031, China and Engineering Research Center of Sustainable Urban Intelligent Transportation, Ministry of Education, Chengdu 610031, China, fengli@swjtu.edu.cn; Qian Wan, School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 610031, China and Engineering Research Center of Sustainable Urban Intelligent Transportation, Ministry of Education, Chengdu 610031, China, qianwan1990@outlook.com.

XXXX-XXXX/2026/9-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Wearable human activity recognition (HAR) requires compact models that can capture temporal dependencies under constrained computation, memory, and energy budgets. Knowledge distillation (KD) offers a practical route to this goal by transferring informative predictions and representations from high-capacity teachers to lightweight students. Existing KD methods for wearable HAR, however, remain limited when structurally different teachers are involved. They commonly rely on homogeneous architectures, use mainly teacher-to-student transfer, and provide limited support for heterogeneous multi-teacher learning. This work proposes a decoupled heterogeneous multi-teacher mutual distillation (DHMMD) framework for wearable HAR. DHMMD performs offline reciprocal optimization between multiple heterogeneous teachers and a compact student, while only the distilled student is retained for inference. The framework combines two complementary paths. Bidirectional decoupled semantic distillation transfers target-class and non-target-class probability information through both aggregated and teacher-specific guidance. Bidirectional contrastive feature distillation aligns intermediate representations across heterogeneous architectures using ensemble-level and pairwise feature supervision. Experiments on HAPT, WISDM, and UCI_HAR show that DHMMD outperforms 21 state-of-the-art KD baselines in terms of F_1 score. Under the MLP-student setting, DHMMD improves the F_1 score by 9.30 percentage points on WISDM. On-device latency measurements on Raspberry Pi 4 and PYNQ-Z2, together with board-power-based energy estimation and memory accounting, further show that the distilled students reduce inference cost relative to the heterogeneous teachers. These results indicate that DHMMD is a practical offline training framework for compact wearable HAR models on mobile and application-processor-class edge platforms.

Additional Key Words and Phrases: Data Mining, Human Activity Recognition, Knowledge Distillation, Model Compression, Wearable Sensors

ACM Reference Format:
Zhiwen Xiao, Huanlai Xing, Rong Qu, Hui Li, Ruibin Bai, Lexi Xu, Jincheng Peng, Li Feng, and Qian Wan. 2026. Decoupled Heterogeneous Multi-Teacher Mutual Distillation for Wearable Human Activity Recognition. 1, 1 (September 2026), 39 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Human activity recognition (HAR) entails the automated inference of individual actions by analyzing temporal sensor data that capture interactions with the environment [1]. It plays a critical role in diverse applications, including human motion tracking [38], gait recognition [29], and activity-aware energy management for battery-powered wearable systems, where recognized motion states can support adaptive sensing and computation policies [28]. The widespread adoption of mobile and wearable devices, such as smartphones and smartwatches, has enabled continuous, large-scale, and context-aware data acquisition, thereby advancing wearable HAR research [41]. These systems typically collect multivariate time series through embedded sensors (e.g., triaxial accelerometers), which encode motion signals along three spatial axes and exhibit both intra- and inter-dimensional dependencies. To ensure robust performance in complex, dynamic environments, HAR models must effectively capture both local temporal variations and global dependencies across sensor modalities [46].

Over the past decade, extensive efforts have been devoted to addressing the challenges in wearable HAR, encompassing both traditional machine learning and deep learning paradigms [41, 46]. Classical approaches typically rely on hand-crafted features and shallow models to extract salient characteristics from temporal sensor data. For instance, researchers proposed a hierarchical framework based on hidden Markov models to infer latent semantic dependencies among contextual cues, thereby enhancing interpretability [72]. In another hybrid scheme, principal component analysis (PCA), coordinate transformation, and support vector machines (SVM) were integrated to manage the intrinsic variability of wearable HAR [33].

In contrast, deep learning models autonomously learn hierarchical representations, enabling the discovery of temporal dependencies and semantic patterns. To this end, several advanced

architectures have emerged. Dong *et al.* [6] introduced a graph-based belief fusion framework combining convolutional neural networks (CNNs) and attention mechanisms to capture spatial relations at multiple semantic levels. SemiHAR [47] utilized a multi-task learning paradigm to jointly model user identity and activity semantics. FCNLSTMaN [48] integrated fully convolutional layers, long short-term memory (LSTM), and attention to simultaneously capture local fluctuations and global temporal dynamics. Further, MS-HARA [40] proposed a two-stream multi-scale framework optimized for both activity recognition and anticipation, enhancing adaptability to dynamic environments.

While effective, such task-specific deep models often incur high computational costs and limited transferability, rendering them suboptimal for deployment on resource-constrained devices. To mitigate this, Hinton *et al.* [16] introduced knowledge distillation (KD), wherein a lightweight student model is trained to replicate the behavior of a more expressive teacher model. Unlike traditional compression methods, such as pruning or low-rank factorization, KD transfers latent knowledge and also acts as a regularizer, improving generalization while preserving task-critical information. KD methodologies can be categorized into self-distillation, single-teacher, and multi-teacher distillation [9]. Self-distillation employs a model's internal knowledge as supervisory signals, as indicated in BYOT [67] and RefineSelf [22]. Single-teacher distillation transfers informative representations from a teacher to a student, either within-task or cross-domain, as exemplified by contrastive HAR distillation [54], FitNet [39], relation-based metric learning [10], and efficient HAR architectures [4]. Multi-teacher distillation integrates knowledge from multiple teachers to provide diverse and complementary supervision, thereby enhancing robustness. Representative works include ensemble-based strategies [59] and large kernel distillation frameworks [58]. Despite its potential, most multi-teacher distillation methods face several limitations:

- *Despite notable progress in multi-teacher distillation, its effectiveness often presupposes architectural homogeneity between teacher and student models. In heterogeneous scenarios, direct aggregation of knowledge from structurally diverse teachers may lead to conflicting representational biases, thereby hindering the student's capacity for effective assimilation and degrading performance.*
- *Most existing approaches emphasize unidirectional knowledge transfer from teachers to the student, overlooking the bidirectional nature of mutual learning [68, 57]. In DHMMD, the teacher networks are not fixed pretrained oracles during distillation. They are high-capacity HAR models jointly optimized with the lightweight student during offline training. This limitation reduces the opportunity for student feedback to refine teacher representations. Such refinement is not intended to make the teachers deployable. It adapts their semantic and feature spaces so that later teacher-to-student supervision becomes more stable and more compatible with the student. Mutual learning enables dynamic cross-model knowledge exchange, which is critical in heterogeneous multi-teacher contexts.*
- *Research on heterogeneous distillation for wearable HAR remains limited. While the recent framework in [51] addresses single-teacher heterogeneous distillation, the multi-teacher setting, particularly under architectural diversity, remains underexplored and warrants deeper investigation.*

To address these limitations, this work develops a decoupled heterogeneous multi-teacher mutual distillation (DHMMD) framework for wearable HAR. In DHMMD, "decoupled" refers to separating target-class and non-target-class output distributions for semantic distillation; "heterogeneous" denotes the use of structurally different teacher and student networks; and "mutual" indicates bidirectional offline optimization between teachers and the student. Unlike conventional methods focused on unidirectional target-class transfer, DHMMD enables reciprocal learning between

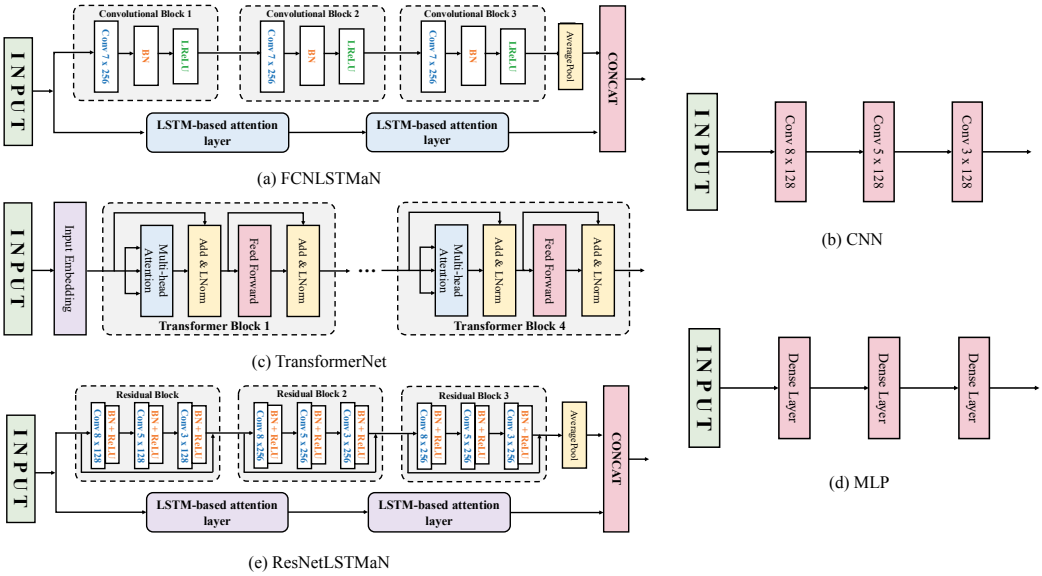


Fig. 1. Semantic representation of the teacher and student models. (a) FCNLSTMaN [48], composed of a fully convolutional network (FCN) with three convolutional blocks and an LSTM-based attention network (LSTMaN) featuring two stacked LSTM-attention layers. Each convolutional block integrates a 1-dimensional convolutional layer, batch normalization (BN), and a leaky rectified linear unit (ReLU) activation function. (b) CNN, serving as a lightweight student model, employs three convolutional neural networks. (c) TransformerNet [60], which includes an input embedding layer followed by four transformer blocks to capture long-range temporal dependencies. (d) MLP, a compact architecture consisting of three fully connected (i.e., dense) layers, designed for lightweight representation learning. (e) ResNetLSTMaN [53], combining a ResNet backbone with three residual blocks and an LSTMaN module with dual attention-based LSTM layers to fuse spatial and temporal features. Note: "Conv 7 × 128" denotes a 1-dimensional convolutional neural network with a kernel size of 7 and 128 output channels.

multiple heterogeneous teachers and a compact student. This reciprocal process is restricted to offline training. The lightweight student remains the inference model, whereas the teachers act as adaptive training participants that improve the quality of their supervision. DHMMD applies bidirectional decoupled semantic distillation at the output level to preserve target- and non-target-class information. It further uses bidirectional contrastive feature distillation to align intermediate representations across heterogeneous architectures.

Following recent wearable HAR distillation studies [51], DHMMD uses FCNLSTMaN [48], ResNetLSTMaN [53], and TransformerNet [60] as heterogeneous teachers. CNN and MLP are used as compact students. The three teachers provide complementary convolutional, residual, recurrent, attention-based, and transformer-based representations. This teacher-student configuration permits evaluation of whether DHMMD can transfer structurally diverse teacher knowledge into lightweight models.

The key contributions of this study can be outlined below:

- DHMMD is positioned as a framework-level contribution for wearable HAR rather than as a collection of independently new distillation primitives. Its novelty lies in reconfiguring heterogeneous multi-teacher mutual learning, attention-based teacher aggregation, decoupled

semantic transfer, and contrastive feature alignment into a unified framework that jointly exploits ensemble-level and teacher-specific supervision for compact student learning.
- A bidirectional decoupled semantic distillation strategy is designed for fine-grained transfer across heterogeneous networks. Attention-based aggregation first fuses target- and non-target-class probabilities from all teachers and aligns the aggregated guidance with student predictions. Teacher-specific bidirectional paths are then used to preserve individualized supervision from each teacher.
- A bidirectional contrastive feature distillation mechanism is further developed at the intermediate-representation level. Aggregated teacher features are aligned with student features through contrastive learning, while individual teacher-student feature pairs are aligned in parallel. This dual-path design improves representation discrimination and feature transfer across heterogeneous structures.
- DHMMD is evaluated on three widely used wearable HAR benchmarks: HAPT, WISDM, and UCI_HAR. With FCNLSTMaN, ResNetLSTMaN, and TransformerNet as heterogeneous teachers, and CNN or MLP as lightweight students, DHMMD outperforms 21 state-of-the-art KD approaches in terms of F_1 score across all datasets. Under the MLP-student setting, DHMMD improves the WISDM F_1 score by approximately 9.30 percentage points. On-device latency measurements on Raspberry Pi 4 and PYNQ-Z2, together with board-power-based energy estimation and analytical memory accounting, further indicate reduced inference latency, energy cost, and memory footprint relative to the heterogeneous teachers. These results support deployment on mobile and application-processor-class wearable edge platforms.

The remainder of the paper is organized as follows. Section 2 reviews wearable HAR and KD research. Section 3 presents the DHMMD framework and its main optimization components. Section 4 reports the experimental evaluation, including benchmark comparisons, ablation studies, representation analysis, and edge-device efficiency results. Section 5 discusses limitations and future work. Section 6 concludes the paper.

2 RELATED WORK

A critical examination of foundational and recent contributions in wearable HAR and KD is undertaken in this section.

2.1 Wearable HAR Methods

A diverse range of algorithms has been proposed to tackle the inherent challenges in wearable HAR, typically falling into two broad categories: traditional and deep learning-based methods [41, 46]. Traditional methods predominantly rely on statistical and machine learning techniques for high-level feature extraction in HAR. These approaches emphasize the identification of key patterns through a variety of algorithms, such as J48, decision trees, fuzzy-based window models, gradient boosting machines, k-nearest neighbors (KNN), logic-based reasoning, Bagging, Bayesian models, AdaBoost, collaborative approaches, Markov regression, SVM, PCA, logistic regression, random forests, and k-means clustering [72, 2, 20]. By leveraging these algorithms, traditional models aim to distill the most salient features of activity recognition, yet they often require extensive feature engineering and may struggle with the complexity of real-world scenarios.

In contrast, deep learning algorithms excel by building intricate representation hierarchies within neural network structures. These models effectively uncover latent connections within the HAR data and generate multi-level representations that capture subtle patterns and dependencies. For instance, Luo *et al.* [31] introduced the BinaryDilatedDenseNet, a binarized neural network model designed for feature extraction under low-latency and low-memory settings. Zhang and Xu [63]

proposed a multi-level network that integrates spatiotemporal attention and multi-scale embeddings to capture the underlying relationships within the time series. Prominent deep learning techniques in this domain include deep fused Lasso networks [71], SemiHAR [47], unsupervised weight learning-based domain adaptation [17], and CapLSTM [24]. Other notable methods include selective kernel convolution [7], perceptual space modeling [43], multi-head convolutional attention [61], kernel density estimation-based models [5], SensorGAN [21], deformable convolutional networks [55], GraphConvLSTM [14], CapMatch [50], GT-WHAR [74], inter-sensor correlations learning [34], STFNet [75], channel-equalization-HAR [18], dynamic weight optimization strategies [26], deep ensemble learning [19], Eff-WHAR [3], DiamondNet [73], and LoCATe-GAT [40]. Further innovations include iSMOTE [13], CNN-Mamba approach [30], and knowledge graph-based methods [56].

2.2 KD Methods

KD is a widely adopted regularization technique that enables the transfer of knowledge from a high-capacity model (the teacher) to a lower-capacity model (the student). KD approaches can be broadly categorized into three primary types based on the source of the student's supervision or the teacher–student paradigm: self-distillation, single-teacher distillation, and multi-teacher distillation [9].

In the self-distillation approach, the model leverages its own internal knowledge, effectively serving both as teacher and student, to enhance its intrinsic capabilities. For instance, the Be Your Own Teacher (BYOT) method [67] was introduced to enable knowledge transfer from the model's output to its lower layers. Another approach, detailed in [22], employs a self-distillation framework with feature refinement to promote the transfer of knowledge from the self-teacher network to a classifier. Additionally, the ProSelfLC framework [45] was proposed to minimize regularization entropy during knowledge transfer. In [66], transitive and densely connected self-distillation techniques were explored to support hierarchical knowledge transfer within the model. Furthermore, the self-bidirectional decoupled distillation [52] was applied to time series classification, offering a novel approach to enhance model performance through self-knowledge transfer.

Single-teacher distillation emphasizes the transfer of rich representations from a well-informed teacher model to a student, either within the same task or across domains, thereby bolstering the student's discriminative capabilities. This approach is exemplified by a variety of notable distillation models, including the pioneering vanilla response-based KD [16], contrastive distillation tailored for HAR [54], and relation-based metric learning [10]. Other significant models in this category include correlation congruence-based KD [36], and contrastive representation distillation [16], which aim to refine feature extraction and improve model performance. Additionally, expert embedding KD [23] and class attention transfer-based KD [12] offer more specialized strategies for knowledge transfer. The decoupled distillation technique [69] and mutual distillation [68], including their dense cross-layer variants [57], further enhance the richness of the transfer process. Notably, hybrid order relational KD [8] and collaborative KD [70] bring new dimensions to knowledge transfer by incorporating multi-model collaborations. Lightweight HAR models [4], and the one-for-all distillation approach [15] aim to optimize the trade-off between model efficiency and performance. Finally, models like FitNet [39] highlight the importance of maintaining computational efficiency while retaining high discriminative power in the student model.

Multi-teacher distillation leverages the expertise of multiple teacher models to provide complementary supervision, thereby enhancing the student's ability to generalize and exhibit greater robustness under diverse conditions. This approach has led to the development of various innovative frameworks, such as multi-teacher feature ensemble methods [59], large kernel distillation techniques [58], and meta-learning-based multi-teacher KD [62]. Noteworthy contributions also

include MKD-Cooper [27], which integrates cooperative learning among multiple teachers, and the use of teacher assistants in multiple roles [42]. Collaborative multi-teacher KD [37] further explores inter-teacher synergies, while MT4MTL-KD [11] introduces a multi-task learning approach to distillation. Additionally, BadCleaner [65] addresses the challenge of noise reduction in multi-teacher settings, ensuring more effective knowledge transfer.

Despite the advancements in multi-teacher distillation, several critical issues remain unaddressed: 1) the assumption of architectural homogeneity, which hinders effective knowledge transfer across structurally diverse models; 2) the predominance of unidirectional teacher-to-student distillation, which underexplores reciprocal adaptation; and 3) the limited study of heterogeneous multi-teacher distillation in wearable HAR scenarios. These gaps indicate that wearable HAR still lacks a unified distillation framework that can coordinate structurally diverse teachers, preserve both shared and teacher-specific knowledge, and support reciprocal optimization with compact students. Accordingly, the contribution of this work lies in a framework-level formulation rather than in treating each underlying distillation primitive as independently new. DHMMD extends prior HAR-specific heterogeneous distillation from a single-teacher setting to a heterogeneous multi-teacher setting, integrates aggregated and teacher-specific supervision, and couples bidirectional semantic and feature-level transfer within one offline optimization framework for compact student learning.

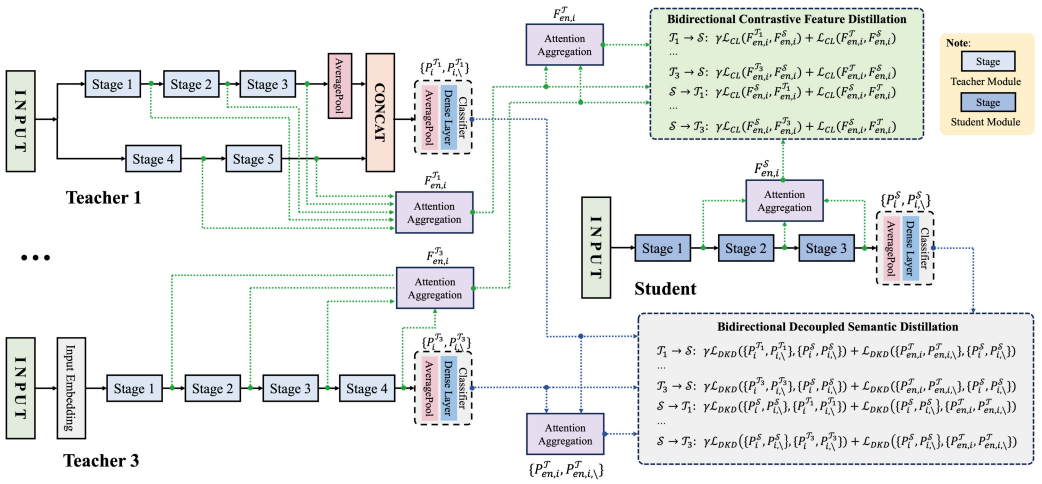


Fig. 2. Schematic representation of the DHMMD framework. The teacher and student modules are essential elements, denoting the neural network structures within the teacher and student models, respectively. For instance, in the TransformerNet teacher model, 'stage 1' through 'stage 4' correspond to 'transformer block 1' through 'transformer block 4'. In contrast, in the MLP student model, 'stage 1', 'stage 2', and 'stage 3' correspond to the first, second, and third fully connected (dense) layers, respectively.

3 PROPOSED DHMMD FRAMEWORK

This section presents the DHMMD framework and its main components: the teacher and student models, attention-based aggregation, bidirectional decoupled semantic distillation, bidirectional contrastive feature distillation, and the collective loss objective. These components define the heterogeneous multi-teacher mutual distillation process used for wearable HAR.

The novelty of DHMMD is defined at the framework level. The backbone models used in this study, including FCNLSTMaN, ResNetLSTMaN, TransformerNet, CNN, and MLP, are adopted

from prior work, and the decoupled KD and contrastive-learning objectives are not presented as standalone new losses. The contribution lies in how these components are reorganized for wearable HAR into a heterogeneous multi-teacher mutual distillation framework. Compared with HMKD [51], which studies single-teacher heterogeneous mutual distillation, DHMMD extends the setting to multiple heterogeneous teachers, introduces attention-based aggregation for ensemble-level guidance, and preserves teacher-specific supervision in both semantic and feature distillation paths.

3.1 DHMMD Overview

DHMMD enables mutual learning between heterogeneous teacher models and a compact student model for wearable HAR. Unlike conventional multi-teacher distillation, which mainly transfers knowledge from teachers to the student, DHMMD supports bidirectional exchange of target-class and non-target-class information. Bidirectional decoupled semantic distillation performs this exchange at the output-probability level. Bidirectional contrastive feature distillation complements it by aligning intermediate representations across heterogeneous architectures. Attention-based aggregation further produces ensemble-level guidance while preserving teacher-specific supervision. The framework structure is illustrated in Fig. 2.

Although the learning primitives in DHMMD are not inherently limited to HAR, their integration in this work is guided by the characteristics of wearable activity sensing. The input representations follow the format of each benchmark, using either multivariate inertial windows or released activity-window feature records. The teacher ensemble is selected to cover complementary temporal properties of wearable signals, including local motion fragments, long-range dependencies, and heterogeneous sequence representations across sensing channels. The decoupled semantic design further reflects the strong inter-class similarity in HAR, where target-class evidence and non-target-class relations are transferred separately rather than collapsed into a single probability signal. On the student side, the framework is constrained by efficient inference on mobile and application-processor-class edge platforms. DHMMD is therefore formulated as a wearable-HAR-oriented heterogeneous multi-teacher mutual distillation framework rather than a task-agnostic pipeline.

3.2 Teacher and Student Models

Following recent heterogeneous distillation studies for wearable HAR [51], DHMMD adopts FCNLSTMaN [48], ResNetLSTMaN [53], and TransformerNet [60] as the teacher ensemble, with CNN and MLP serving as compact student models. These networks are reused backbone architectures rather than new architectural contributions. The methodological contribution of DHMMD lies in how these heterogeneous models are organized within a multi-teacher mutual distillation framework.

The teacher ensemble is selected to provide complementary inductive biases for wearable HAR. FCNLSTMaN combines convolutional local-pattern extraction with recurrent attention-based temporal modeling. ResNetLSTMaN adds residual hierarchical abstraction while preserving recurrent temporal modeling. TransformerNet contributes global self-attention for long-range dependency modeling. This combination allows the teacher side to cover local motion fragments, hierarchical temporal abstractions, and broader activity-level dependencies. CNN and MLP are used as lightweight students to evaluate whether this heterogeneous teacher knowledge can be transferred to compact inference models. Detailed architectural descriptions of FCNLSTMaN, ResNetLSTMaN, TransformerNet, CNN, and MLP are provided in Appendix.

3.3 Attention-based Aggregation

Attention-based aggregation addresses the heterogeneity among the three teachers and the compact student. This module is specific to the multi-teacher setting. Its role is to generate ensemble-level

probability and feature representations from heterogeneous teachers, so that the student receives collective guidance together with teacher-specific supervision. Such aggregation is unnecessary in the single-teacher formulation of [51], but becomes essential when DHMMD combines three structurally different teachers.

3.3.1 Classification Probability Aggregation. Let $O_i^{\mathcal{T}_k} = [O_{i,1}^{\mathcal{T}_k}, O_{i,2}^{\mathcal{T}_k}, \ldots, O_{i,C}^{\mathcal{T}_k}] \in \mathbb{R}^{1\times C}$ denote the logit output of the k-th teacher model for input sample x_i, where $k \in \{1,2,3\}$, $i \in \{1,\ldots,N_{\mathrm{IS}}\}$, and C is the number of target categories. Let $m = y_i \in \{1,\ldots,C\}$ denote the ground-truth class index of sample x_i. The temperature-scaled softmax operation produces the predicted class probability distribution $P_i^{\mathcal{T}_k} \in \mathbb{R}^{1\times C}$. The target probability is denoted by $P_{i,m}^{\mathcal{T}_k}$, and the scalar non-target probability mass is denoted by $P_{i,\setminus m}^{\mathcal{T}_k}$. Their definitions are:

$$\begin{aligned} P_{i,m}^{\mathcal{T}_k} &= \frac{\exp(O_{i,m}^{\mathcal{T}_k}/t_{DKD})}{\sum_{h=1}^{C}\exp(O_{i,h}^{\mathcal{T}_k}/t_{DKD})}, \\ P_{i,\setminus m}^{\mathcal{T}_k} &= \frac{\sum_{q\neq m}\exp(O_{i,q}^{\mathcal{T}_k}/t_{DKD})}{\sum_{h=1}^{C}\exp(O_{i,h}^{\mathcal{T}_k}/t_{DKD})}, \end{aligned} \tag{1}$$

where m denotes the target class of sample x_i, while h and q are class-index variables used in the normalization and non-target summation, respectively. The temperature coefficient t_{DKD} is empirically fixed at 1.0 throughout all experimental trials. For a detailed analysis of its role, refer to Subsection 4.3.

To consolidate predictions from heterogeneous teacher networks, an attention-driven fusion strategy is adopted. This mechanism assigns dynamic weights to teacher-specific outputs via learnable vectors, thereby enabling selective emphasis on semantically richer knowledge. The aggregated teacher probability distribution and aggregated non-target probability mass, denoted by $P_{en,i}^{\mathcal{T}}$ and $P_{en,i,\setminus}^{\mathcal{T}}$, are computed as:

$$\begin{aligned} P_{en,i}^{\mathcal{T}} &= \sum_{k=1}^{3}\left(\frac{\exp(W_{CP}^{\top}P_i^{\mathcal{T}_k})}{\sum_{h=1}^{3}\exp(W_{CP}^{\top}P_i^{\mathcal{T}_h})}\right)P_i^{\mathcal{T}_k}, \\ P_{en,i,\setminus}^{\mathcal{T}} &= \sum_{k=1}^{3}\left(\frac{\exp(W_{CP,\setminus}P_{i,\setminus}^{\mathcal{T}_k})}{\sum_{h=1}^{3}\exp(W_{CP,\setminus}P_{i,\setminus}^{\mathcal{T}_h})}\right)P_{i,\setminus}^{\mathcal{T}_k}, \end{aligned} \tag{2}$$

where $W_{CP} \in \mathbb{R}^C$ and $W_{CP,\setminus} \in \mathbb{R}$ are trainable attention parameters responsible for adaptively modulating the contribution of each teacher.

Likewise, for the student model, let $O_i^{\mathcal{S}} = [O_{i,1}^{\mathcal{S}}, O_{i,2}^{\mathcal{S}}, \ldots, O_{i,C}^{\mathcal{S}}] \in \mathbb{R}^{1\times C}$ denote the predicted logits corresponding to sample x_i. Applying the same temperature-based normalization yields the student's class probability vector $P_i^{\mathcal{S}} \in \mathbb{R}^{1\times C}$ and scalar non-target probability mass $P_{i,\setminus m}^{\mathcal{S}}$:

$$\begin{aligned} P_{i,m}^{\mathcal{S}} &= \frac{\exp(O_{i,m}^{\mathcal{S}}/t_{DKD})}{\sum_{h=1}^{C}\exp(O_{i,h}^{\mathcal{S}}/t_{DKD})}, \\ P_{i,\setminus m}^{\mathcal{S}} &= \frac{\sum_{q\neq m}\exp(O_{i,q}^{\mathcal{S}}/t_{DKD})}{\sum_{h=1}^{C}\exp(O_{i,h}^{\mathcal{S}}/t_{DKD})}. \end{aligned} \tag{3}$$

3.3.2 Intermediate Feature Aggregation. Let $F_{j,i}^{\mathcal{T}_k}$ denote the activation of the j-th layer in the k-th teacher network when processing the input sample x_i, where $j = 1,2,\ldots,L^{\mathcal{T}_k}$, $k = 1,2,3$, and

$i = 1, 2, \ldots, N_{\text{IS}}$. The total number of layers in each teacher model is denoted by $L^{\mathcal{T}_k}$. Similarly, the student network's layer-wise feature output is represented as $F^{\mathcal{S}}_{j,i}$, with $j = 1, 2, \ldots, L^{\mathcal{S}}$.

To extract a compact yet informative representation from each teacher model, this work introduces an attention-based feature encoding strategy. Specifically, each intermediate feature $F^{\mathcal{T}_k}_{j,i}$ is first mapped to a shared embedding space via a learnable transformation $\phi_j(\cdot)$, yielding:

$$\tilde{F}^{\mathcal{T}_k}_{j,i} = \phi_j\left(F^{\mathcal{T}_k}_{j,i}\right). \tag{4}$$

For the i-th input sample, a set of layer-level attention scores $\alpha^{(k)}_{j,i}$ is computed to measure the contribution of each transformed layer:

$$\alpha^{(k)}_{j,i} = \frac{\exp\left(W^{\top}_{TE}\tilde{F}^{\mathcal{T}_k}_{j,i}\right)}{\sum_{l=1}^{L^{\mathcal{T}_k}}\exp\left(W^{\top}_{TE}\tilde{F}^{\mathcal{T}_k}_{l,i}\right)}, \tag{5}$$

where W_{TE} is a learnable attention vector shared across layers of the same teacher. The subscript i indicates that the layer attention is computed separately for each input sample.

The aggregated feature representation for the i-th sample in the k-th teacher model is then given by:

$$F^{\mathcal{T}_k}_{en,i} = \sum_{j=1}^{L^{\mathcal{T}_k}} \alpha^{(k)}_{j,i} \cdot \tilde{F}^{\mathcal{T}_k}_{j,i}. \tag{6}$$

After obtaining the integrated representation of each teacher, cross-teacher integration is performed to capture complementary knowledge. Specifically, the formulation defines the final teacher-side encoding $F^{\mathcal{T}}_{en,i}$ by aggregating the three teacher-specific representations using another attention mechanism:

$$F^{\mathcal{T}}_{en,i} = \sum_{k=1}^{3} \omega_{k,i} \cdot F^{\mathcal{T}_k}_{en,i}, \tag{7}$$

where the sample-specific teacher-level attention weights $\omega_{k,i}$ are computed as:

$$\omega_{k,i} = \frac{\exp\left(W^{\top}_{AE}F^{\mathcal{T}_k}_{en,i}\right)}{\sum_{q=1}^{3}\exp\left(W^{\top}_{AE}F^{\mathcal{T}_q}_{en,i}\right)}, \tag{8}$$

where W_{AE} is a learnable parameter vector that adaptively emphasizes the relative informativeness of each teacher model for the current input sample.

The student representation is then processed in the same aligned feature space. Given that feature dimensionalities may differ across layers, each student-layer feature $F^{\mathcal{S}}_{j,i}$ is aligned into a shared latent space through a transformation $\psi_j(\cdot)$, such that:

$$\tilde{F}^{\mathcal{S}}_{j,i} = \psi_j\left(F^{\mathcal{S}}_{j,i}\right). \tag{9}$$

To dynamically weight each transformed student-layer feature for sample x_i, attention coefficients $\beta_{j,i}$ are determined as:

$$\beta_{j,i} = \frac{\exp\left(W^{\top}_{SE}\tilde{F}^{\mathcal{S}}_{j,i}\right)}{\sum_{l=1}^{L^{\mathcal{S}}}\exp\left(W^{\top}_{SE}\tilde{F}^{\mathcal{S}}_{l,i}\right)}, \tag{10}$$

where W_{SE} denotes a learnable attention vector that governs sample-specific layer relevance on the student side.

The final encoded student representation is computed as:

$$F_{en,i}^{\mathcal{S}} = \sum_{j=1}^{L^{\mathcal{S}}} \beta_{j,i} \cdot \tilde{F}_{j,i}^{\mathcal{S}}. \tag{11}$$

3.4 Bidirectional Decoupled Semantic Distillation

The bidirectional decoupled semantic distillation strategy enables comprehensive knowledge transfer across both target-class and non-target-class categories between a heterogeneous multi-teacher ensemble and the student model. Its contribution is not the decoupled KD loss in isolation, but its integration into the heterogeneous multi-teacher mutual-learning setting of DHMMD. In this formulation, semantic supervision is organized through two complementary sources: teacher-specific guidance from each individual source model and ensemble-level guidance from the aggregated teacher representation. The former preserves discriminative cues that may be unique to a particular teacher, whereas the latter captures shared semantic consensus across the teacher ensemble. This design extends prior single-teacher heterogeneous distillation by allowing collective and individualized probability information to guide bidirectional learning within the same framework. Specifically, the decoupled semantic distillation strategy comprises two components: (1) teacher-to-student knowledge guidance through decoupled KD and (2) student-to-teacher knowledge guidance via decoupled KD.

To reduce redundancy, the decoupled semantic distillation loss associated with both directions can be unified under a general formulation. Specifically, the decoupled semantic distillation loss from model A to model B is defined as:

$$\begin{aligned}\mathcal{L}_{\mathrm{DSD}}^{A\to B} = \gamma\, \mathcal{L}_{DKD}\left(\{P_i^A, P_{i,\backslash}^A\}, \{P_i^B, P_{i,\backslash}^B\}\right) \\ + \mathcal{L}_{DKD}\left(\{P_{en,i}^{\mathcal{T}}, P_{en,i,\backslash}^{\mathcal{T}}\}, \{P_i^B, P_{i,\backslash}^B\}\right)\end{aligned} \tag{12}$$

where $A \in \{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3, \mathcal{S}\}$, and $B \in \{\mathcal{S}, \mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3\}$ depending on the transfer direction. Here, γ is a weighting coefficient that controls the strength of the source-specific reciprocal transfer term (empirically set to 0.1, see Subsection 4.3), and $\mathcal{L}_{DKD}$ denotes the decoupled KD loss. The first term models source-specific semantic transfer from model A to model B. When $A = \mathcal{T}_k$, it preserves the individualized supervisory signal of the k-th teacher; when $A = \mathcal{S}$, it conveys student feedback to the corresponding teacher during mutual learning. The second term aligns model B with the ensemble-level semantic consensus encoded by the aggregated teacher distributions $P_{en,i}^{\mathcal{T}}$ and $P_{en,i,\backslash}^{\mathcal{T}}$. Both terms are retained because the aggregated teacher captures shared knowledge across teachers, whereas the source-specific term preserves complementary discriminative cues that may be attenuated during attention-based aggregation. Thus, Eq. (12) jointly enforces collective guidance and individualized guidance, rather than allowing the ensemble term to replace the individual source-model term.

The distillation of semantic distributions uses Jensen–Shannon (JS) divergence rather than conventional KL divergence [69, 52], because JS divergence is symmetric and bounded. The decoupled loss function is accordingly defined as:

$$\mathcal{L}_{DKD}\left(\{P^A, P_{\backslash}^A\}, \{P^B, P_{\backslash}^B\}\right) = \mu\, \mathrm{JS}(P^A, P^B) + \lambda\, \mathrm{JS}(P_{\backslash}^A, P_{\backslash}^B) \tag{13}$$

where P^A and $P_{\backslash}^A$ denote the soft class probability distributions of the target and non-target classes predicted by model A, and likewise for B. The hyperparameters μ and λ control the relative emphasis on target and non-target class knowledge transfer, respectively. Following precedents in [69, 52], the value is fixed at $\mu = 1.0$, and empirically set $\lambda = 2.0$ to strengthen the influence of non-target

class alignment, thereby enhancing the student's generalization capabilities under heterogeneous supervision (refer to Subsection 4.3).

3.5 Bidirectional Contrastive Feature Distillation

The bidirectional contrastive feature distillation framework improves intermediate feature alignment between the heterogeneous multi-teacher ensemble and the student model, enabling reciprocal representation transfer during offline training. Its contribution is not the standalone InfoNCE objective, which follows established contrastive-learning practice, but the way this objective is reorganized within DHMMD for heterogeneous multi-teacher mutual distillation. The student is aligned with both teacher-specific feature representations and the ensemble-aggregated feature representation, while the teachers also receive student-guided feature feedback during reciprocal optimization. This design preserves individualized teacher cues, exploits shared ensemble-level structure, and supports feature transfer across heterogeneous architectures in wearable HAR. Specifically, this framework consists of two complementary and reciprocally structured distillation components: (1) teacher-to-student contrastive alignment, wherein the student aligns its encoded representations with both individual teacher outputs and the fused ensemble features; and (2) student-to-teacher contrastive feedback, which facilitates the refinement of teacher representations through student-guided contrastive learning.

To avoid redundant notation across transfer directions, the contrastive feature distillation loss from model A to model B is defined as:

$$\mathcal{L}_{\text{CFD}}^{A\to B} = \gamma\, \mathcal{L}_{CL}(F_{en,i}^{A}, F_{en,i}^{B}) + \mathcal{L}_{CL}(F_{en,i}^{\mathcal{T}}, F_{en,i}^{B}), \tag{14}$$

where $A \in \{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3, \mathcal{S}\}$, and $B \in \{\mathcal{S}, \mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3\}$ with their specific selection governed by the transfer direction. The parameter γ is a balancing coefficient that limits the magnitude of the source-specific reciprocal feature-alignment term (empirically set to 0.1; refer to Subsection 4.3), and $\mathcal{L}_{CL}$ represents the contrastive loss function.

Unlike conventional distance-based alignment paradigms, which often rely on fixed geometric proximity in feature space, this study employs an InfoNCE-based contrastive objective [35], owing to its superior capacity for enforcing instance-level discriminability through semantic separation in the latent space. This choice not only enhances alignment fidelity between heterogeneous representations but also facilitates the formation of compact, class-relevant clusters.

Formally, the contrastive feature distillation loss between feature embeddings F^A and F^B is defined as:

$$\mathcal{L}_{CL}(F^A, F^B) = -\log \frac{\exp(\text{sim}(F^A, F^B)/\tau)}{\exp(\text{sim}(F^A, F^B)/\tau) + \sum_{j=1}^{N_{\text{neg}}} \exp(\text{sim}(F^A, F_j^-)/\tau)}, \tag{15}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, τ is the temperature coefficient regulating distribution sharpness, and N_{neg} denotes the number of negative instances sampled from the current mini-batch. In this formulation, (F^A, F^B) forms the positive pair because the two representations correspond to the same input instance observed through the two models involved in the current distillation direction. The denominator therefore contains the matched positive term together with the negative terms, following the standard InfoNCE construction. Each negative feature F_j^- is formed from the representation of a different instance in the same mini-batch, excluding the matched positive pair. The proposed contrastive feature distillation is not implemented as a supervised contrastive loss; accordingly, same-class samples are not explicitly removed from the negative set. As long as an instance is not the matched positive partner of F^A, it may serve as a batch-level negative. This design is intentional because the role of $\mathcal{L}_{CL}$ in DHMMD is to align cross-model representations of the same instance while repelling them from other instance-level representations, rather than to

Algorithm 1 Offline Joint Training and Evaluation Procedure of DHMMD

Input: $\mathcal{D} = \{\mathcal{D}_{train}, \mathcal{D}_{val}, \mathcal{D}_{test}\}$; number of training epochs E; validation interval V
Output: Predictions of the trained models on $\mathcal{D}_{test}$, with the distilled student retained for deployment-oriented inference

Independently initialize teacher parameters $\theta_0^{\mathcal{T}_1}$, $\theta_0^{\mathcal{T}_2}$, $\theta_0^{\mathcal{T}_3}$ and student parameters $\theta_0^{\mathcal{S}}$ ▷ Default DHMMD trains all models jointly from scratch during offline optimization
Initialize the teacher losses $\mathcal{L}^{\mathcal{T}_k}$ for $k \in \{1, 2, 3\}$ and the student loss $\mathcal{L}^{\mathcal{S}}$
for $e = 1$ to E **do**
    Feed $\mathcal{D}_{train}$ into $\mathcal{T}_1$, $\mathcal{T}_2$, $\mathcal{T}_3$, and $\mathcal{S}$ ▷ Obtain logits, class-probability distributions, and intermediate feature representations
    Compute the bidirectional decoupled semantic distillation terms $\mathcal{L}_{\text{DSD}}^{A\to B}$ and bidirectional contrastive feature distillation terms $\mathcal{L}_{\text{CFD}}^{A\to B}$ for the required transfer directions ▷ Each term contains the source-specific transfer component and ensemble-level teacher guidance defined in Eqs. (12) and (14)
    Compute the teacher objectives:
    $\mathcal{L}^{\mathcal{T}_1} = \mathcal{L}_{\text{CFD}}^{\mathcal{S}\to\mathcal{T}_1} + \mathcal{L}_{\text{DSD}}^{\mathcal{S}\to\mathcal{T}_1} + \mathcal{L}_{\text{sup}}^{\mathcal{T}_1}$
    $\mathcal{L}^{\mathcal{T}_2} = \mathcal{L}_{\text{CFD}}^{\mathcal{S}\to\mathcal{T}_2} + \mathcal{L}_{\text{DSD}}^{\mathcal{S}\to\mathcal{T}_2} + \mathcal{L}_{\text{sup}}^{\mathcal{T}_2}$
    $\mathcal{L}^{\mathcal{T}_3} = \mathcal{L}_{\text{CFD}}^{\mathcal{S}\to\mathcal{T}_3} + \mathcal{L}_{\text{DSD}}^{\mathcal{S}\to\mathcal{T}_3} + \mathcal{L}_{\text{sup}}^{\mathcal{T}_3}$
    Compute the student objective:
    $\mathcal{L}^{\mathcal{S}} = \sum_{k=1}^{3} \left(\mathcal{L}_{\text{CFD}}^{\mathcal{T}_k\to\mathcal{S}} + \mathcal{L}_{\text{DSD}}^{\mathcal{T}_k\to\mathcal{S}} \right) + \mathcal{L}_{\text{sup}}^{\mathcal{S}}$
    ▷ The student receives teacher-specific guidance and ensemble-level teacher supervision
    Update each teacher using its supervised loss and the student-to-teacher distillation terms:
    $\theta_e^{\mathcal{T}_1} = \theta_{e-1}^{\mathcal{T}_1} - \eta^{\mathcal{T}_1} \nabla_{\theta_{e-1}^{\mathcal{T}_1}} \mathcal{L}^{\mathcal{T}_1}$
    $\theta_e^{\mathcal{T}_2} = \theta_{e-1}^{\mathcal{T}_2} - \eta^{\mathcal{T}_2} \nabla_{\theta_{e-1}^{\mathcal{T}_2}} \mathcal{L}^{\mathcal{T}_2}$
    $\theta_e^{\mathcal{T}_3} = \theta_{e-1}^{\mathcal{T}_3} - \eta^{\mathcal{T}_3} \nabla_{\theta_{e-1}^{\mathcal{T}_3}} \mathcal{L}^{\mathcal{T}_3}$
    Update the student using its supervised loss and the teacher-to-student distillation terms:
    $\theta_e^{\mathcal{S}} = \theta_{e-1}^{\mathcal{S}} - \eta^{\mathcal{S}} \nabla_{\theta_{e-1}^{\mathcal{S}}} \mathcal{L}^{\mathcal{S}}$
    if $e \bmod V = 0$ **then**
        Validate $\mathcal{T}_1$, $\mathcal{T}_2$, $\mathcal{T}_3$, and $\mathcal{S}$ on $\mathcal{D}_{val}$
        Record validation performance and retain the best student checkpoint
    end if
end for
Testing stage:
Obtain $Y^{\mathcal{T}_1}$, $Y^{\mathcal{T}_2}$, $Y^{\mathcal{T}_3}$, and $Y^{\mathcal{S}}$ on $\mathcal{D}_{test}$
Retain the trained student model $\mathcal{S}$ as the deployment-oriented inference model
return $Y^{\mathcal{T}_1}$, $Y^{\mathcal{T}_2}$, $Y^{\mathcal{T}_3}$, $Y^{\mathcal{S}}$

enforce class-level grouping through supervised contrastive learning. This formulation encourages the alignment of the positive pair (F^A, F^B) while separating it from the sampled negatives $\{F_j^-\}_{j=1}^{N_{\text{neg}}}$. The resulting objective promotes a more discriminative embedding space without relying on class-label-based contrastive grouping. Following empirical insights reported in [54, 44], the value is fixed at $\tau = 0.1$ to ensure a balanced trade-off between convergence stability and representation compactness.

3.6 Collective Loss Function

Following mutual learning frameworks [68, 57, 52], DHMMD adopts an offline co-optimization protocol in which the heterogeneous teachers and the student are trained jointly on the HAR training set. This protocol differs from conventional fixed teacher KD because the teachers are updated together with the student during training, while only the distilled student is retained for on-device inference. At the loss design level, the supervised cross-entropy term remains standard. The framework contribution lies in coupling the preceding modules into one objective: ensemble-aware semantic distillation, teacher-specific semantic distillation, ensemble-aware contrastive feature distillation, and teacher-specific contrastive feature distillation. This formulation reorganizes established learning operators into a coherent heterogeneous multi-teacher mutual distillation objective for wearable HAR.

The reciprocal pathway is designed as controlled co-adaptation rather than unrestricted student-driven correction of the teachers. The student-to-teacher losses update the teachers only during offline optimization, with the purpose of improving the consistency and usefulness of their training signals. Their influence is constrained by the relatively small coefficient γ on the source-specific reciprocal term, by the supervised cross-entropy loss that anchors each teacher to the HAR labels, and by the ensemble-level teacher guidance retained in both semantic and feature distillation. Therefore, early-stage student predictions do not dominate teacher learning. The bidirectional mechanism instead improves compatibility between the compact student and the heterogeneous teacher ensemble while preserving teacher discrimination.

The teacher-side and student-side objectives are optimized jointly so that semantic transfer, feature alignment, and supervised classification provide complementary training signals. For each teacher network $\mathcal{T}_k$, where $k \in \{1, 2, 3\}$, the objective contains the student-to-teacher contrastive feature distillation loss, the student-to-teacher decoupled semantic distillation loss, and the supervised teacher loss:

$$\mathcal{L}^{\mathcal{T}_k} = \mathcal{L}_{\mathrm{CFD}}^{\mathcal{S}\to\mathcal{T}_k} + \mathcal{L}_{\mathrm{DSD}}^{\mathcal{S}\to\mathcal{T}_k} + \mathcal{L}_{\mathrm{sup}}^{\mathcal{T}_k}, \quad k \in \{1, 2, 3\}. \tag{16}$$

The student objective aggregates teacher-to-student semantic and feature supervision from all three teachers, together with the supervised student loss:

$$\mathcal{L}^{\mathcal{S}} = \sum_{k=1}^{3} \left(\mathcal{L}_{\mathrm{CFD}}^{\mathcal{T}_k\to\mathcal{S}} + \mathcal{L}_{\mathrm{DSD}}^{\mathcal{T}_k\to\mathcal{S}} \right) + \mathcal{L}_{\mathrm{sup}}^{\mathcal{S}}. \tag{17}$$

For any model $\mathcal{M} \in \{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3, \mathcal{S}\}$, the supervised loss is computed by the standard cross-entropy criterion over the corresponding N_{IS} input instances, where $y_{i,c}$ denotes the one-hot label indicator for class c:

$$\mathcal{L}_{\mathrm{sup}}^{\mathcal{M}} = -\frac{1}{N_{\mathrm{IS}}} \sum_{i=1}^{N_{\mathrm{IS}}} \sum_{c=1}^{C} y_{i,c} \log \left(P_{i,c}^{\mathcal{M}} \right). \tag{18}$$

Joint stochastic gradient descent is then applied to all trainable models. Let $\theta_m^{\mathcal{M}}$ denote the parameters of model $\mathcal{M}$ at epoch m. Each model is updated according to its corresponding objective:

$$\theta_m^{\mathcal{M}} = \theta_{m-1}^{\mathcal{M}} - \eta^{\mathcal{M}} \nabla_{\theta_{m-1}^{\mathcal{M}}} \mathcal{L}^{\mathcal{M}}, \quad \mathcal{M} \in \{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3, \mathcal{S}\}. \tag{19}$$

where $\eta^{\mathcal{M}}$ is the learning rate of model $\mathcal{M}$, and $\nabla_{\theta_{m-1}^{\mathcal{M}}} \mathcal{L}^{\mathcal{M}}$ denotes the gradient of its objective with respect to the parameters from the previous epoch. In the default protocol, the three teachers and the student are initialized independently and trained jointly from scratch during offline optimization. The teachers are not pretrained and frozen in the main setting. At each epoch, every teacher is updated by its supervised classification loss and by the student-to-teacher semantic and feature

distillation terms, while the student is updated by its supervised classification loss and by the teacher-to-student distillation terms. Thus, bidirectional mutual learning denotes reciprocal parameter optimization during offline training. It does not imply joint deployment of the teacher ensemble and the student. After optimization, only the distilled student is retained for deployment-oriented inference. Algorithm 1 summarizes this offline training and testing workflow, and the appendix provides the corresponding mathematical analysis.

Table 1. Summary of the three wearable HAR benchmarks, including sensor channels and the model input representation used in this work.

Dataset	Instances	Sampling Rate	Classes	Sensor Channels	Input Used in This Work
HAPT	10,929	50Hz	12	triaxial linear acceleration + triaxial gyroscope	released 561-dimensional activity-window feature vector per instance
WISDM	1,098,207	20Hz	6	triaxial accelerometer	raw fixed-length window of 48 samples × 3 axes per instance
UCI_HAR	10,299	50Hz	6	triaxial linear acceleration + triaxial gyroscope	released 561-dimensional activity-window feature vector per instance

4 EXPERIMENTS

This section begins by outlining the experimental configuration and the evaluation criteria employed to assess model effectiveness. Subsequently, it investigates the sensitivity of key hyper-parameters through a comprehensive ablation analysis. The section concludes with a holistic evaluation of DHMMD's performance.

4.1 Experimental Setting

4.1.1 Dataset Description. In alignment with the most recent advancements in heterogeneous distillation for wearable HAR [51], the evaluation considers the effectiveness of the proposed DHMMD framework using three widely recognized wearable HAR benchmarks, i.e., HAPT [1], WISDM [2], and UCI_HAR [3]. A comprehensive summary of the key characteristics and configurations of these datasets is presented in Table 1, providing essential context for subsequent experimental analyses.

4.1.2 Data Preprocessing. Following prior wearable HAR studies [41, 48, 46, 72, 2, 51], the preprocessing procedure follows the released input format of each benchmark rather than enforcing a uniform raw-signal representation across all datasets. This distinction is necessary because HAPT and UCI_HAR are provided as activity-window feature records, whereas WISDM is processed from raw tri-axial accelerometer streams.

For HAPT and UCI_HAR, each sample is represented by a released 561-dimensional feature vector derived from smartphone inertial measurements. The value 561 therefore denotes the feature dimensionality of an activity window, not the length of a raw temporal sequence. No additional signal-level denoising or handcrafted feature extraction is applied beyond the released dataset representation. For WISDM, each sample is constructed from raw tri-axial accelerometer measurements collected at 20Hz. The continuous signal is normalized and segmented into fixed-length temporal windows, each containing 48 samples. Hence, the value 48 denotes the temporal

[1] https://archive.ics.uci.edu/dataset/341/smartphone+based+recognition+of+human+activities+and+postural+transitions
[2] https://www.cis.fordham.edu/wisdm/dataset.php
[3] https://archive.ics.uci.edu/dataset/240/human+activity+recognition+using+smartphones

window length used for the raw-sequence input, and no additional handcrafted feature-engineering stage is introduced before model input.

The resulting inputs are fixed-dimensional representations supplied to the teacher and student networks, although their origins differ across datasets. HAPT and UCI_HAR provide window-level feature vectors, while WISDM provides segmented raw accelerometer windows. DHMMD operates after this benchmark-specific input construction and is therefore compatible with both released feature representations and raw temporal-window representations.

For the raw-sequence setting used in WISDM, let $Data_h$ denote the sensor observation at the h-th timestamp, and let TWS denote the fixed time-window size. The i-th multivariate temporal input window is defined as

$$x_i = [Data_1, Data_2, \ldots, Data_{TWS}], \quad i = 1, 2, \ldots, N_{\mathrm{IS}}. \tag{20}$$

Equation (20) describes the raw temporal-window representation and applies directly to WISDM. For HAPT and UCI_HAR, x_i instead denotes the released 561-dimensional feature vector corresponding to the i-th activity window. This notation keeps the input definition consistent while preserving the dataset-specific difference between feature-vector inputs and raw-sequence inputs.

4.1.3 Data Partition. Following the evaluation protocols adopted in prior wearable HAR and distillation studies [41, 48, 46, 61, 64, 55, 7, 50, 51], each dataset is partitioned using a stratified sample-level strategy. The full dataset is first divided into a development subset and an independent test subset with a 7:3 ratio. The development subset is then split into training and validation subsets with an 8:2 ratio. This procedure yields training, validation, and test portions of 56%, 14%, and 30% of the original data, respectively. The test subset is excluded from model training, hyper-parameter tuning, and checkpoint selection.

The partition is stratified by activity class to preserve class-distribution consistency across the training, validation, and test subsets. It is performed at the sample level rather than at the subject level. Therefore, the present protocol does not correspond to leave-one-subject-out evaluation or another subject-independent setting, and samples from the same subject may appear in different subsets. This choice follows the benchmark regime used by the cited baselines, ensuring that all compared methods are evaluated under the same partitioning condition.

The reported results are interpreted under the stratified sample-level partition used throughout this study. Using the same partitioning regime as the compared methods keeps the comparison focused on methodological differences under matched data conditions. A subject-independent protocol addresses a different experimental setting because the test subjects are not represented in the training data, making it more directly suited to evaluating generalization across users. The present evaluation thus characterizes performance under the benchmark partition adopted in this study, whereas cross-user generalization requires a subject-independent experimental design.

Table 2. Parameters of MLP, CNN, FCNLSTMaN, ResNetLSTMaN, and TransformerNet across three wearable HAR benchmarks.

Model	HAPT	WISDM	UCI_HAR
MLP	172,837	**94,519**	169,579
CNN	**137,509**	135,187	**137,235**
FCNLSTMaN	788,013	576,507	785,007
ResNetLSTMaN	530,957	533,511	529,415
TransformerNet	871,777	745,521	864,049

4.1.4 Implementation Details. In both FCNLSTMaN and ResNetLSTMaN, each LSTM-based attention module is configured with 128 units to ensure sufficient temporal modeling capacity. For TransformerNet, a progressive architectural design is adopted across its four transformer layers. Specifically, the first transformer block comprises a 64-unit fully connected layer with an 8-head multi-head attention module. The subsequent transformer layers are configured with fully connected layers of 128, 192, and 256 units respectively, each maintaining 8 attention heads, thereby enabling hierarchical representation learning with increasing abstraction.

Table 2 reports the parameter counts of MLP, CNN, FCNLSTMaN, ResNetLSTMaN, and TransformerNet across the three HAR datasets. These statistics were computed using Python 3 and TensorFlow. The student models have substantially smaller parameter footprints than the teacher models. On HAPT, for example, MLP contains 172,837 parameters, whereas FCNLSTMaN, ResNetLSTMaN, and TransformerNet contain 788,013, 530,957, and 871,777 parameters, respectively.

All experiments are conducted on a computing platform running Ubuntu 18.04, equipped with an NVIDIA RTX 2080Ti GPU and an AMD R5 1400 CPU, supported by 32GB RAM. The optimization process employs the RMSProp algorithm, with a momentum coefficient of 0.9, an initial learning rate of 0.001, and a decay factor set to 0.9. The software stack includes Python 3, Scikit-learn, and TensorFlow, ensuring consistency and reproducibility throughout the experimental pipeline.

4.2 HAR Performance Metrics

In alignment with established evaluation practices in HAR research [41, 48, 46, 61, 64, 55, 50, 51], this study adopts two standard performance metrics, *Accuracy* and the macro-averaged *F-measure* (macro-F_1 score), to comprehensively assess model effectiveness. The macro-F_1 score is computed in a one-vs-rest manner for each activity class and then averaged uniformly across all classes. Their formal definitions are provided in Eqs. (21)–(24):

$$\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}}} \times 100\% \tag{21}$$

$$\begin{aligned} \text{Precision}^{(c)} &= \frac{TP^{(c)}}{TP^{(c)} + FP^{(c)}} \times 100\% \\ \text{Recall}^{(c)} &= \frac{TP^{(c)}}{TP^{(c)} + FN^{(c)}} \times 100\% \end{aligned} \tag{22}$$

$$F_1^{(c)} = \frac{2 \times \text{Precision}^{(c)} \times \text{Recall}^{(c)}}{\text{Precision}^{(c)} + \text{Recall}^{(c)}} \tag{23}$$

$$F_1^{\text{macro}} = \frac{1}{C} \sum_{c=1}^{C} F_1^{(c)} \tag{24}$$

where C denotes the number of activity classes, and $TP^{(c)}$, $FP^{(c)}$, and $FN^{(c)}$ are computed for the c-th class under the one-vs-rest setting. Unless otherwise stated, all F_1 values reported in the experimental section correspond to macro-F_1.

4.3 Hyper-parameter Sensitivity Analysis

To examine the sensitivity of DHMMD to key hyper-parameter configurations, we evaluate its performance across three representative wearable HAR benchmarks.

Table 3. F_1 scores achieved by DHMMD with various t_{DKD} values across three wearable HAR benchmarks.

Teacher	Student	t_{DKD}	HAPT (%)	WISDM (%)	UCI_HAR (%)
FCNLSTMaN ResNetLSTMaN TransformerNet	MLP	0.10	90.02	84.44	88.92
		0.50	91.15	85.69	91.43
		1.00	**92.96**	**88.21**	**92.81**
		2.00	90.29	86.02	90.89
		5.00	89.38	84.35	89.53
	CNN	0.10	90.29	90.02	89.49
		0.50	91.45	92.00	91.39
		1.00	**93.95**	**93.83**	**94.65**
		2.00	92.03	92.00	91.39
		5.00	90.29	89.19	89.02

4.3.1 With various t_{DKD} values. The parameter t_{DKD} controls the softness of the class-probability distributions used in decoupled KD. Table 3 shows that $t_{DKD} = 1.0$ gives the strongest F_1 scores across the three HAR datasets. This setting preserves informative dark knowledge while avoiding excessive smoothing, which can weaken class discrimination. The value $t_{DKD} = 1.0$ is therefore used in the remaining experiments.

Table 4. F_1 scores achieved by DHMMD with various γ values across three wearable HAR benchmarks.

Teacher	Student	γ	HAPT (%)	WISDM (%)	UCI_HAR (%)
FCNLSTMaN ResNetLSTMaN TransformerNet	MLP	0.10	**92.96**	**88.21**	**92.81**
		0.20	92.38	85.24	91.72
		0.30	91.68	86.02	91.02
		0.40	90.29	85.24	90.89
		0.50	91.15	84.93	90.21
		1.00	88.00	83.43	89.62
		2.00	88.54	83.93	88.94
	CNN	0.10	**93.95**	**93.83**	**94.65**
		0.20	92.77	91.85	92.34
		0.30	91.45	90.26	91.39
		0.40	91.90	90.85	90.89
		0.50	90.29	89.99	89.49
		1.00	90.73	90.02	90.89
		2.00	88.89	89.19	88.49

4.3.2 With various γ values. The hyper-parameter γ controls the strength of the source-specific reciprocal terms in both bidirectional DSD and bidirectional CFD. In this ablation, γ is kept fixed rather than dynamically scheduled, so that the effect of reciprocal transfer strength can be examined under a controlled optimization protocol. As shown in Table 4, $\gamma = 0.1$ achieves the best F_1 performance across the evaluated HAR benchmarks. When γ is too small, the reciprocal signal becomes weak and provides limited bidirectional adaptation. When γ is too large, the source-specific distillation terms can impose excessive regularization, especially when the student representation is still immature in the early training stage. The selected value therefore provides a stable balance between effective knowledge transfer and student adaptability, while preventing student-to-teacher feedback from dominating teacher updates.

4.3.3 With various decoupled KD loss functions. The decoupled KD loss is central to bidirectional decoupled semantic distillation. JS divergence is used as the basis of this loss because it provides

Table 5. F_1 scores achieved by DHMMD with various decoupled KD loss functions across three wearable HAR benchmarks.

Teacher	Student	Decoupled KD Loss	HAPT (%)	WISDM (%)	UCI_HAR (%)
FCNLSTMaN ResNetLSTMaN TransformerNet	MLP	JS	**92.96**	**88.21**	**92.81**
		KL	92.06	86.69	91.43
		L_1	90.02	84.44	89.62
		L_2	90.04	81.06	88.89
		CE	88.19	82.72	89.14
	CNN	JS	**93.95**	**93.83**	**94.65**
		KL	91.90	91.36	91.39
		L_1	90.73	90.02	90.89
		L_2	90.04	89.19	89.49
		CE	89.49	88.89	89.12

stable symmetric distribution alignment. Table 5 indicates that JS divergence outperforms L_1 loss, L_2 loss, KL divergence, and cross-entropy loss across the evaluated HAR tasks. This result supports the use of JS divergence for aligning teacher and student probability distributions with different levels of complexity.

Table 6. F_1 scores achieved by DHMMD with various teacher models across three wearable HAR benchmarks.

Teacher				Student	HAPT (%)	WISDM (%)	UCI_HAR (%)
FCNLSTMaN	ResNetLSTMaN	TransformerNet	MHA				
✓				MLP	92.43	84.02	90.18
	✓				92.74	85.72	91.72
		✓			91.32	83.92	89.14
			✓		90.04	82.39	88.89
✓	✓				92.69	85.93	91.88
	✓	✓			92.79	86.02	91.95
		✓	✓		92.77	85.88	91.43
✓			✓		92.60	85.34	91.39
✓	✓	✓			**92.96**	88.21	**92.81**
	✓	✓	✓		92.84	87.34	92.03
✓		✓	✓		92.89	87.58	92.13
✓	✓	✓	✓		92.92	**88.30**	92.68
✓				CNN	93.04	91.06	92.60
	✓				93.15	92.13	93.90
		✓			92.03	90.85	93.19
			✓		91.68	90.52	92.34
✓	✓				93.26	92.34	93.98
	✓	✓			93.34	92.58	94.02
		✓	✓		93.47	92.86	94.11
✓			✓		93.26	92.21	93.87
✓	✓	✓			**93.95**	**93.83**	**94.65**
	✓	✓	✓		93.59	93.63	94.21
✓		✓	✓		93.48	93.50	94.16
✓	✓	✓	✓		93.86	93.76	94.33

4.4 Ablation Study

The ablation study examines the effect of teacher selection, teacher-update protocol, and bidirectional distillation components across the three HAR datasets.

4.4.1 Rationale for selecting FCNLSTMaN, ResNetLSTMaN, and TransformerNet as teachers in DHMMD. The default teacher ensemble consists of FCNLSTMaN, ResNetLSTMaN, and TransformerNet. This selection combines convolutional, residual, LSTM, attention, and transformer

structures, following the multi-teacher distillation principle that architectural diversity can enrich transferable knowledge [62, 27].

An additional MHA teacher (multihead convolutional attention) [64] is included to evaluate whether simply increasing the number of teachers improves performance. Table 6 indicates that performance generally improves as the ensemble grows from one to three teachers. The four-teacher setting does not yield a consistent additional gain. Although it slightly improves the MLP result on WISDM, it lowers or marginally weakens several other settings. These results suggest that architectural diversity is useful only when the ensemble remains sufficiently coherent for effective knowledge transfer.

The FCNLSTMaN, ResNetLSTMaN, and TransformerNet ensemble is retained because it provides a favorable balance between architectural diversity and distillation stability across the evaluated settings, although it is not optimal in every individual case. With a larger pool of candidate teachers, exhaustive evaluation of all teacher combinations would become increasingly costly. Automated teacher selection or pruning of redundant teachers could instead identify a compact set of complementary models before joint distillation.

Table 7. Student F_1 scores achieved by DHMMD variants across three wearable HAR benchmarks, including teacher-update protocol and bidirectional distillation ablations.

Teacher	Evaluated Student	Method	HAPT (%)	WISDM (%)	UCI_HAR (%)
FCNLSTMaN ResNetLSTMaN TransformerNet	MLP	DHMMD-FrozenT	91.43	86.97	91.72
		DHMMD-PreT-Update	92.87	**89.01**	92.60
		DHMMD-w/o-BDSD	90.04	83.90	88.89
		DHMMD-TSUDSD	91.15	85.24	90.85
		DHMMD-STUDSD	90.29	84.35	91.72
		DHMMD-w/o-BCFD	89.49	83.93	89.14
		DHMMD-TSUCFD	90.73	85.24	90.21
		DHMMD-STUCFD	91.15	84.93	90.29
		DHMMD	**92.96**	88.21	**92.81**
	CNN	DHMMD-FrozenT	92.15	91.86	92.99
		DHMMD-PreT-Update	**94.02**	93.74	**94.89**
		DHMMD-w/o-BDSD	89.49	88.38	90.05
		DHMMD-TSUDSD	91.45	90.02	90.89
		DHMMD-STUDSD	91.90	90.26	91.39
		DHMMD-w/o-BCFD	89.49	89.99	89.12
		DHMMD-TSUCFD	92.03	91.85	90.89
		DHMMD-STUCFD	91.45	90.26	91.39
		DHMMD	93.95	**93.83**	94.65

4.4.2 Effect of Teacher Update Protocol. The treatment of teacher parameters during mutual learning is examined before analyzing the individual distillation modules. Table 7 includes two protocol-level variants that quantify the effect of different teacher-update strategies on the evaluated student model. DHMMD-FrozenT denotes a fixed-teacher distillation setting. In this protocol, FCNLSTMaN, ResNetLSTMaN, and TransformerNet are first trained independently using the supervised classification loss and are then frozen during student distillation. The scores reported for DHMMD-FrozenT in Table 7 are therefore not standalone teacher accuracies, but the final F_1 scores of the student trained under frozen teacher guidance. In this setting, only the student is updated, and the student-to-teacher pathway is disabled. DHMMD-PreT-Update also starts from independently pretrained teachers, but further updates the teachers together with the student under the full DHMMD objective. By contrast, the default DHMMD protocol initializes all teachers and the student independently and trains them jointly from scratch during offline optimization.

The comparison separates fixed teacher guidance from reciprocal teacher adaptation. Freezing the pretrained teachers consistently weakens the distilled student. For the MLP student, DHMMD-FrozenT trails the default DHMMD by 1.53, 1.24, and 1.09 percentage points on HAPT, WISDM, and UCI_HAR, respectively. For the CNN student, the corresponding drops are 1.80, 1.97, and 1.66 percentage points. These results indicate that fixed teacher guidance is less effective than reciprocal teacher-student adaptation. The gain is particularly evident for the CNN student, suggesting that teacher adaptation provides a more compatible feature-transfer process for compact convolutional representation learning.

DHMMD-PreT-Update remains close to the default DHMMD and slightly exceeds it in several cases, including WISDM with the MLP student and HAPT and UCI_HAR with the CNN student. The differences between these two jointly updated protocols, however, are small. This pattern suggests that teacher pretraining may offer a mild initialization benefit, but it is not essential for obtaining strong student performance. The default DHMMD protocol is retained as the main setting because it achieves comparable performance while avoiding a separate teacher-pretraining stage and preserving a unified end-to-end offline optimization procedure.

4.4.3 Impact of Bidirectional Decoupled Semantic Distillation. The contribution of bidirectional decoupled semantic distillation (BDSD) is evaluated through three ablated variants that remove or alter semantic-transfer directionality:

- *DHMMD-w/o-BDSD*: A baseline variant with BDSD removed entirely.
- *DHMMD-TSUDSD*: Employs a unidirectional semantic distillation path from teacher to student, replacing the bidirectional mechanism.
- *DHMMD-STUDSD*: Applies a reversed unidirectional distillation from student to teacher, in place of the original bidirectional formulation.

Table 7 indicates that the complete DHMMD objective outperforms the BDSD-ablated variants across all three datasets. The result supports the role of bidirectional semantic transfer in aligning heterogeneous teacher and student predictions. Removing BDSD or retaining only one direction weakens this alignment, because either collective teacher guidance or reciprocal student feedback is lost.

4.4.4 Impact of Bidirectional Contrastive Feature Distillation. The effect of bidirectional contrastive feature distillation (BCFD) is evaluated through three structurally modified variants:

- *DHMMD-w/o-BCFD*: Omits contrastive feature distillation altogether.
- *DHMMD-TSUCFD*: Restricts contrastive supervision to a unidirectional flow from teacher to student.
- *DHMMD-STUCFD*: Applies contrastive supervision solely in the student-to-teacher direction.

Table 7 indicates that DHMMD also outperforms the BCFD-ablated variants. Bidirectional contrastive alignment helps teacher and student features approach a shared embedding space while preserving model-specific structure. Removing BCFD or retaining only one contrastive direction produces asymmetric feature adaptation and weaker student performance.

The ablation results indicate that DHMMD benefits from both its training protocol and its distillation design. Reciprocal teacher-student adaptation is more effective than frozen teacher guidance. Bidirectional decoupled semantic distillation and bidirectional contrastive feature distillation further contribute complementary semantic and feature-level supervision. Together, the heterogeneous teacher ensemble, adaptive teacher updating, decoupled semantic transfer, and contrastive feature alignment form a coherent optimization scheme for compact wearable HAR models.

Table 8. F_1 scores achieved by various state-of-the-art KD algorithms across three wearable HAR benchmarks.

Student	Teacher	Method	HAPT (%)	WISDM (%)	UCI_HAR (%)	Student	Teacher	Method	HAPT (%)	WISDM (%)	UCI_HAR (%)
N/A	N/A	FCNLSTMaN	95.14	98.45	96.28	N/A	N/A	FCNLSTMaN	95.14	98.45	96.28
		ResNetLSTMaN	96.12	98.96	96.86			ResNetLSTMaN	96.12	98.96	96.86
		TransformerNet	96.05	98.83	96.75			TransformerNet	96.05	98.83	96.75
		MLP	83.58	78.91	85.07			CNN	85.88	88.00	87.05
MLP	N/A	BYOT	85.62	80.72	86.12	CNN	N/A	BYOT	86.89	88.74	88.39
		TSD	86.23	81.04	86.46			TSD	87.85	89.04	89.02
		SAD	86.93	81.45	87.04			SAD	87.34	88.92	88.49
		ProSelfLC	85.05	82.89	87.66			ProSelfLC	88.53	89.19	89.65
		Self-BiDecKD	88.23	83.63	88.04			Self-BiDecKD	89.49	89.75	89.86
	FCNLSTMaN	ResponseKD	84.86	80.13	85.86		FCNLSTMaN	ResponseKD	86.87	88.26	87.66
		FitNet	85.93	80.04	86.46			FitNet	88.19	88.92	88.49
		CC	86.87	80.32	86.89			CC	88.67	88.74	89.02
		CRD	88.53	80.58	86.46			CRD	87.85	89.04	89.12
		RelaKD	88.67	80.86	86.99			RelaKD	88.24	89.19	88.49
		DKD	88.19	79.49	85.96			DKD	88.89	89.14	88.57
		MD	88.24	74.17	88.49			MD	85.21	89.31	87.84
		DMD	88.89	80.86	86.46			DMD	89.49	88.38	88.94
		DIST	89.49	80.86	87.66			DIST	88.53	89.19	89.57
		OFA	90.04	81.06	88.89			OFA	90.97	89.99	90.18
		HMKD	92.43	83.90	90.21			HMKD	93.04	91.00	92.58
	ResNetLSTMaN	ResponseKD	85.05	80.92	85.59		ResNetLSTMaN	ResponseKD	86.39	88.26	87.84
		FitNet	85.83	80.93	86.35			FitNet	86.89	88.74	89.02
		CC	87.13	77.12	87.00			CC	88.19	89.31	89.57
		CRD	86.32	82.93	88.23			CRD	87.34	89.19	90.21
		RelaKD	88.19	82.72	89.14			RelaKD	89.49	87.84	89.12
		DKD	87.13	81.79	86.33			DKD	88.53	89.19	89.19
		MD	88.54	74.25	87.34			MD	88.89	88.38	88.49
		DMD	88.00	82.89	88.23			DMD	87.34	88.89	88.57
		DIST	88.89	83.93	88.94			DIST	90.04	89.99	89.49
		OFA	90.02	84.44	89.62			OFA	90.73	90.02	90.89
		HMKD	92.77	85.69	91.72			HMKD	93.15	92.13	93.90
	FCNLSTMaN ResNetLSTMaN TransformerNet	MeanTeacher	88.00	83.43	88.92		FCNLSTMaN ResNetLSTMaN TransformerNet	MeanTeacher	90.29	89.19	89.02
		AMTML	89.38	84.35	89.53			AMTML	89.49	88.74	91.02
		CMT-KD	91.15	84.93	90.85			CMT-KD	91.45	90.26	91.39
		DGDK	91.68	86.02	91.43			DGDK	92.03	92.00	93.55
		MTFE	90.29	85.24	90.89			MTFE	91.90	90.85	92.03
		DHMMD	**92.96**	**88.21**	**92.81**			DHMMD	**93.95**	**93.83**	**94.65**

4.5 Experimental Results and Discussion

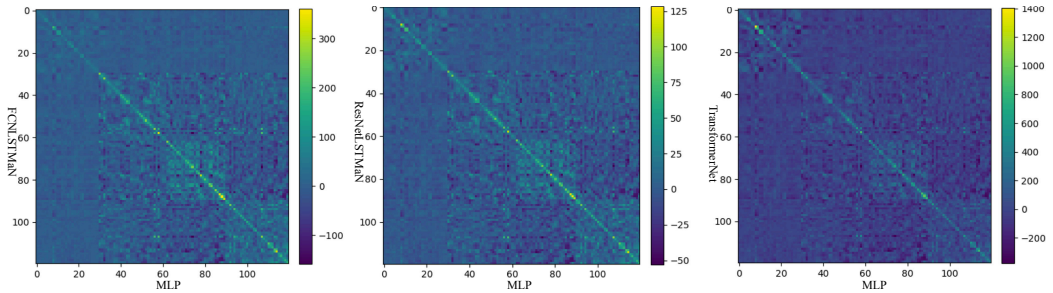
The effectiveness of DHMMD is evaluated against a broad set of KD algorithms using F_1 as the principal metric. The competing baselines encompass a diverse set of teacher-student configurations, including both classical and state-of-the-art KD paradigms. The following comparison defines the teacher-student backbones used in DHMMD and the baseline groups evaluated in the experiments.

- Five backbone models are considered: MLP, a pure multilayer perceptron with three fully connected layers; CNN, a pure convolutional network with three convolutional layers; FCNLSTMaN; ResNetLSTMaN; and TransformerNet. MLP and CNN serve as lightweight student candidates, whereas FCNLSTMaN, ResNetLSTMaN, and TransformerNet are used as teacher models;
- five self-distillation approaches are included as comparison baselines, namely BYOT [67], TSD [66], SAD [9], ProSelfLC [45], and Self-BiDecKD [52];
- eleven single-teacher KD methods are further considered, including ResponseKD [16], FitNet [39], CRD [16], and HMKD [51], which cover both representation-level and relation-level transfer strategies;
- five multi-teacher KD frameworks are also evaluated, including MeanTeacher [9], AMTML [62], and DGDK [42], in order to compare DHMMD against representative approaches that exploit supervision from multiple teacher sources.

Table 8 indicates that DHMMD achieves top-tier performance across the three wearable HAR benchmarks. With MLP as the student, DHMMD reaches an F_1 score of 92.81% on UCI_HAR and exceeds all evaluated baselines. HMKD remains competitive, but ResponseKD and other output-only alignment methods perform less effectively, indicating the value of deeper semantic and feature-level transfer.

The performance gains of DHMMD are consistent with its two framework-specific bidirectional distillation paths. First, the bidirectional decoupled semantic distillation module disaggregates class-relevant and class-irrelevant prediction knowledge, enabling refined alignment of probabilistic outputs from heterogeneous teachers. Second, the bidirectional contrastive feature distillation mechanism bridges internal feature spaces via both global and teacher-specific contrastive objectives, ensuring cross-model consistency at intermediate levels.

This architecture not only facilitates richer and more discriminative representation transfer but also alleviates the structural disparity challenge inherent in heterogeneous teacher-student setups. By contrast, traditional KD methods often fail to fully exploit multi-level supervisory signals, limiting their generalization capability, especially in high-variability tasks such as wearable HAR.



(a) CKA: FCNLSTMaN vs. MLP (b) CKA: ResNetLSTMaN vs. MLP (c) CKA: TransformerNet vs. MLP

Fig. 3. CKA similarity analysis between student MLP and teacher models (i.e., FCNLSTMaN, ResNetLSTMaN, and TransformerNet) in DHMMD on the UCI_HAR dataset.

4.6 Representation Visualization

To comprehensively evaluate what the DHMMD student models learn from the three heterogeneous teacher models (FCNLSTMaN, ResNetLSTMaN, and TransformerNet), this study employs the centered kernel alignment (CKA) [25] method. This technique offers insights into the feature alignment between the teacher and student models. As depicted in Fig. 3, the CKA similarity analysis between the student MLP and the teacher models on the UCI_HAR dataset reveals that the feature similarity consistently exceeds 0.7. Such high values indicate that the student model can effectively capture the learned features from the teacher models, thereby benefiting from the knowledge transfer facilitated by DHMMD. This observation underscores the robustness of the DHMMD framework in enabling effective feature learning across diverse model architectures.

Feature embeddings are visualized with t-distributed stochastic neighbor embedding (t-SNE) [32] to assess representation quality. Fig. 4 compares the embeddings learned by MLP, CNN, and DHMMD across the three HAR datasets. DHMMD produces more compact and better separated clusters than the standalone students. The WISDM visualizations in Figs. 4(f), 4(g), and 4(h) illustrate this pattern clearly. These results indicate that DHMMD improves representation discrimination rather than only changing the two-dimensional t-SNE projection.

The visual results in Fig. 5 illustrate a notable improvement in the performance of student models on multi-class classification tasks when guided by DHMMD. Specifically, on the HAPT dataset, DHMMD significantly boosts the accuracy of student models, particularly in distinguishing between static postures, such as Lie-to-Sit and Lie-to-Stand. This marked improvement underscores DHMMD's capacity to transfer knowledge effectively from heterogeneous teacher models to student models, enhancing the model's discriminative power even in complex, fine-grained classification

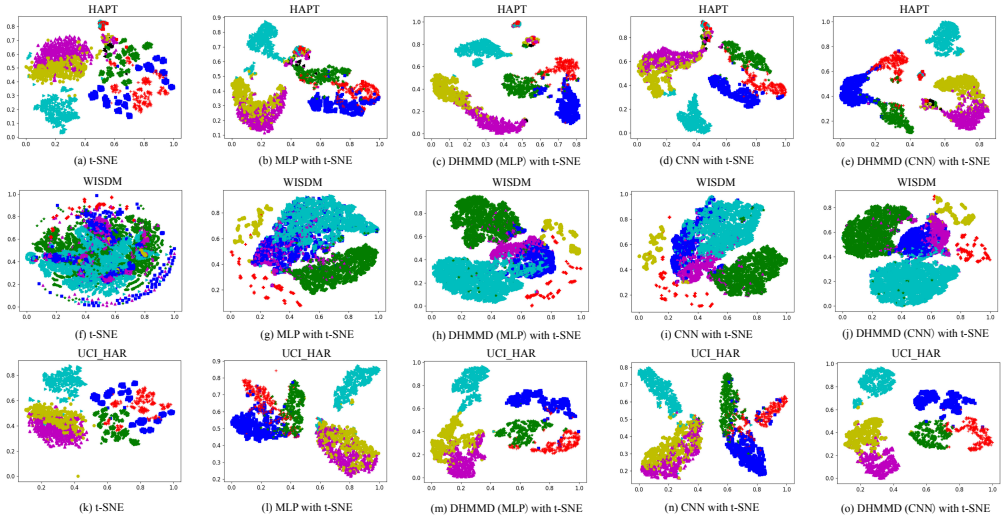


Fig. 4. Visualization of representations learned by t-SNE, MLP with t-SNE, CNN with t-SNE, DHMMD (MLP), and DHMMD (CNN) with t-SNE across three wearable HAR benchmarks. Note: "DHMMD (MLP)" refers to the DHMMD framework with MLP as the student model, while "DHMMD (CNN)" represents DHMMD with CNN as the student model.

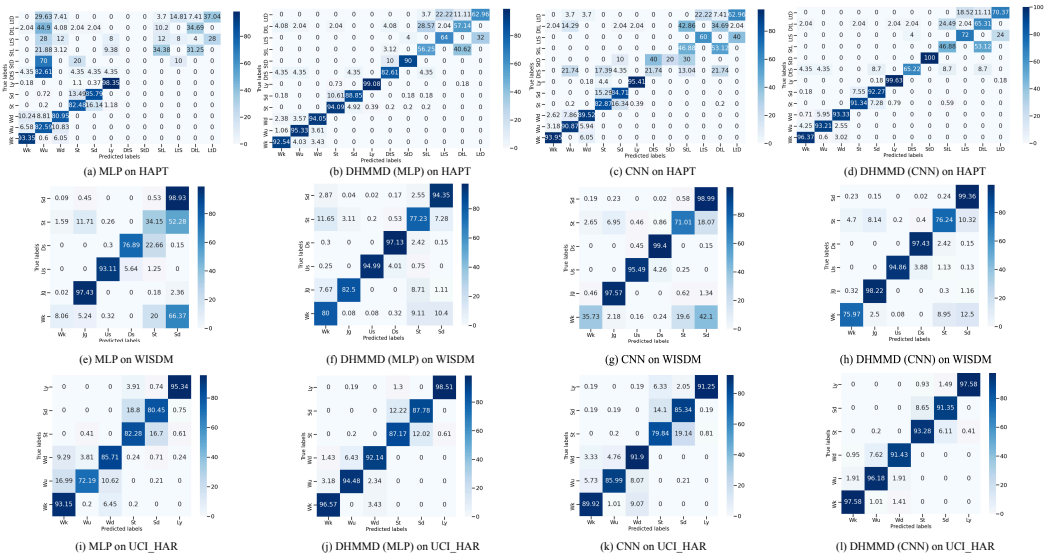


Fig. 5. Confusion matrices of MLP, CNN, DHMMD (MLP), and DHMMD (CNN) across three wearable HAR benchmarks. Note: "DHMMD (MLP)" refers to the DHMMD framework with MLP as the student model, while "DHMMD (CNN)" represents DHMMD with CNN as the student model.

scenarios. Such performance gains highlight the potential of DHMMD to improve the accuracy and reliability of HAR models across a wide range of applications.

By leveraging heterogeneous teacher models, bidirectional distillation strategies, and advanced feature learning mechanisms, DHMMD excels in refining the performance of student models, enabling more robust and generalized representations that are pivotal for accurate HAR.

4.7 Evaluation on Lightweight Edge Devices

The deployment efficiency of the distilled students is evaluated on two representative edge platforms: Raspberry Pi 4 with a quad-core Cortex-A72 CPU at 1.5 GHz, and PYNQ-Z2 with a dual-core Cortex-A9 CPU at 667 MHz. The purpose of this experiment is to characterize model-side inference efficiency after offline DHMMD training has been completed. Accordingly, the deployed objects in this subsection are the distilled student networks rather than the full DHMMD training framework or the teacher ensemble.

To characterize deployment efficiency, the evaluation reports the inference behavior of all teacher and student models on the two edge devices. The metrics in Table 9 are obtained through three procedures: latency is measured on device, energy is estimated from measured latency and average board power, and memory is derived analytically. Inference latency is recorded as the average forward-pass time per sample after the input representation has been constructed. Energy consumption per 100 samples is estimated as

$$E_{100} = P_{\text{dev}} \times \frac{100 \times L}{1000}, \tag{25}$$

where L denotes the measured per-sample latency in milliseconds and P_{dev} denotes the average board power of the target device. The values used are $P_{\text{dev}} = 7.0\,\text{W}$ for Raspberry Pi 4 and $P_{\text{dev}} = 2.43\,\text{W}$ for PYNQ-Z2. The reported energy values are computed directly from Eq. (25) and rounded to two decimal places. Memory usage is computed from model parameters and intermediate activations during inference. Table 9 therefore characterizes neural-network inference after input construction. It excludes preprocessing operations such as temporal window construction and dataset-specific signal preparation. Thus, the table reports measured latency, board-power-based energy estimation, and analytical memory accounting, not a direct hardware profile for every quantity.

The deployment scope is limited to mobile and embedded edge platforms with application-class processors and MB-level memory resources, including smartphones, Raspberry Pi-class boards, PYNQ-class system-on-chip (SoC) platforms, wearable gateways, and comparable embedded Linux devices. More resource-constrained endpoints, such as smart-ring-class devices, microcontroller unit (MCU)-based wearables, and platforms with sub-megabyte memory, are not covered by the current evaluation. For smartwatch-class devices, the reported results are most relevant to systems equipped with application-class processors rather than severely resource-limited controllers.

Table 9 shows a consistent efficiency gap between the distilled students and the heterogeneous teachers. The parameter counts are aligned with Table 2. Across the three datasets, the lightweight students contain 94,519 to 172,837 parameters, whereas the teacher models contain 529,415 to 871,777 parameters. The students require 6.12 to 8.64 ms per sample on Raspberry Pi 4 and 15.28 to 22.34 ms on PYNQ-Z2. The teachers require 18.82 to 31.74 ms and 46.93 to 76.15 ms, respectively. The same pattern appears in the energy estimates: the students consume 4.28 to 6.05 J per 100 samples on Raspberry Pi 4 and 3.71 to 5.43 J on PYNQ-Z2, while the teachers require 13.17 to 22.22 J and 11.40 to 18.50 J, respectively. Memory usage follows the same trend, with the student models remaining below 30 MB and the teacher models reaching up to 72.64 MB.

Within the evaluated hardware range, the distilled students require less inference time, estimated energy, and memory than the heterogeneous teacher models. The MLP and CNN used in the present experiments were selected as compact student backbones and do not define the student architecture of DHMMD. A purpose-built lightweight HAR network can be used in the same role when lower

Table 9. Parameters, measured inference latency, estimated energy consumption, and analytically derived memory usage of various HAR models across HAPT, WISDM, and UCI_HAR on Raspberry Pi 4 and PYNQ-Z2.

Model	Parameters	Latency (ms)	Estimated Energy (J/100 samples)	Memory (MB)
Raspberry Pi 4 (HAPT)				
MLP	172,837	6.12	4.28	14.80
CNN	137,509	8.29	5.80	26.89
FCNLSTMaN	788,013	18.82	13.17	44.50
ResNetLSTMaN	530,957	24.17	16.92	58.40
TransformerNet	871,777	30.85	21.60	72.50
Raspberry Pi 4 (WISDM)				
MLP	94,519	6.38	4.47	14.93
CNN	135,187	8.64	6.05	27.02
FCNLSTMaN	576,507	19.41	13.59	44.61
ResNetLSTMaN	533,511	24.98	17.49	58.53
TransformerNet	745,521	31.74	22.22	72.64
Raspberry Pi 4 (UCI_HAR)				
MLP	169,579	6.24	4.37	14.88
CNN	137,235	8.47	5.93	26.95
FCNLSTMaN	785,007	19.05	13.34	44.57
ResNetLSTMaN	529,415	24.56	17.19	58.48
TransformerNet	864,049	31.21	21.85	72.58
PYNQ-Z2 (HAPT)				
MLP	172,837	15.28	3.71	14.80
CNN	137,509	21.65	5.26	26.89
FCNLSTMaN	788,013	46.93	11.40	44.50
ResNetLSTMaN	530,957	59.84	14.54	58.40
TransformerNet	871,777	74.52	18.11	72.50
PYNQ-Z2 (WISDM)				
MLP	94,519	15.84	3.85	14.93
CNN	135,187	22.34	5.43	27.02
FCNLSTMaN	576,507	48.06	11.68	44.61
ResNetLSTMaN	533,511	61.27	14.89	58.53
TransformerNet	745,521	76.15	18.50	72.64
PYNQ-Z2 (UCI_HAR)				
MLP	169,579	15.56	3.78	14.88
CNN	137,235	21.98	5.34	26.95
FCNLSTMaN	785,007	47.52	11.55	44.57
ResNetLSTMaN	529,415	60.48	14.70	58.48
TransformerNet	864,049	75.31	18.30	72.58

inference cost is required, provided that its prediction outputs and selected intermediate features can be incorporated into the semantic and feature distillation objectives. In this case, the computational properties of the lightweight backbone are retained at inference because the teacher ensemble is used only during offline training. Devices with substantially tighter memory and power budgets, such as sub-MB wearable microcontrollers, may still require additional compression through quantization, structured pruning, activation-memory reduction, or hardware-aware compilation.

5 LIMITATIONS AND FUTURE WORK

The present study has several limitations that define the scope of the reported findings. The evaluation follows a sample-level benchmark protocol on three wearable HAR datasets and does not

yet examine subject-independent or cross-domain generalization. This setting supports controlled comparison with existing distillation baselines, but it does not fully capture variability caused by individual motion habits, sensor placement, device orientation, and data-collection conditions. Future work will evaluate DHMMD under leave-one-subject-out, cross-user, and cross-domain protocols to assess its robustness under more realistic deployment shifts. The edge-device analysis focuses on neural-network inference after input construction. Latency is measured directly on Raspberry Pi 4 and PYNQ-Z2, whereas energy consumption is estimated from latency and board-level power. A more complete deployment evaluation calls for external power profiling and end-to-end pipeline measurement, including sensing, preprocessing, inference, and data transfer.

The deployment scope is further limited to mobile and application-processor-class edge platforms with MB-level memory resources. More restrictive wearable devices, such as ultra-low-power microcontrollers and smart-ring-class platforms, require additional compression beyond the student architectures evaluated in this work. Future extensions will combine DHMMD with quantization, pruning, structured sparsity, and hardware-aware optimization to support these constrained devices. In addition, the teacher ensemble used here is an empirically effective heterogeneous configuration rather than the result of exhaustive architecture search. Further work will investigate structure-aware HAR models, sensor-topology information, and automated teacher selection to improve the diversity and compatibility of teacher guidance.

6 CONCLUSION

This study presents DHMMD, a heterogeneous multi-teacher mutual distillation framework for wearable HAR. DHMMD combines bidirectional decoupled semantic distillation with bidirectional contrastive feature distillation to transfer output-level and representation-level knowledge from structurally diverse teachers to compact students. Experiments on HAPT, WISDM, and UCI_HAR show that DHMMD improves student recognition performance over existing distillation baselines in terms of the F_1 score. The results indicate that aggregated teacher guidance and teacher-specific supervision provide complementary benefits for compact student learning. On-device latency measurements on Raspberry Pi 4 and PYNQ-Z2, together with board-power-based energy estimation and analytical memory accounting, further show that the distilled students reduce inference cost relative to the heterogeneous teachers. These deployment results apply to mobile and application-processor-class wearable edge platforms. More constrained smartwatch, smart-ring, and MCU-class devices require additional compression and hardware-aware optimization beyond the current student architectures. Taken together, the findings suggest that DHMMD offers an offline training strategy for improving compact wearable HAR models under edge-side inference constraints.

REFERENCES

[1] D. Anguita et al. “A public domain dataset for human activity recognition using smartphones”. In: *Proc. 21st Eur. Symposium Artif. Neural Netw., Comput. Intell. Mach. Learn.* 2013, pp. 437–442.

[2] Z. Chen et al. “Robust Human Activity Recognition Using Smartphone Sensors via CT-PCA and Online SVM”. In: *IEEE Trans. Ind. Inform.* 13.6 (2017), pp. 3070–3080. DOI: 10.1109/TII.2017.2712746.

[3] Zijie Chen et al. “Eff-WHAR: A Lightweight Design for Efficient Wearable Sensor-Based Human Activity Recognition”. In: *IEEE Sensors Journal* 25.2 (2025), pp. 3935–3948. DOI: 10.1109/JSEN.2024.3509961.

[4] Shizhuo Deng et al. “LHAR: Lightweight Human Activity Recognition on Knowledge Distillation”. In: *IEEE J. Biomed. Health. Inf.* 28.11 (2024), pp. 6318–6328. DOI: 10.1109/JBHI.2023.3298932.

[5] Yilin Dong et al. "Dezert-Smarandache Theory-Based Fusion for Human Activity Recognition in Body Sensor Networks". In: *IEEE Trans. Ind. Inform.* 16.11 (2020), pp. 7138–7149.

[6] Yilin Dong et al. "Graph-Structure-Based Multigranular Belief Fusion for Human Activity Recognition". In: *IEEE Trans. Neur. Net. Lear.* 35.10 (2024), pp. 13589–13603. DOI: 10.1109/TNNLS.2023.3270290.

[7] Wenbin Gao et al. "Deep Neural Networks for Sensor-Based Human Activity Recognition Using Selective Kernel Convolution". In: *IEEE Trans. Instrum. Meas.* 70 (2021), pp. 1–13. DOI: 10.1109/TIM.2021.3102735.

[8] Shiming Ge et al. "Look One and More: Distilling Hybrid Order Relational Knowledge for Cross-Resolution Image Recognition". In: *Proc. AAAI Conf. Artif. Intell.* 2020, pp. 10845–10852.

[9] J. Gou, B. Yu, et al. "Knowledge Distillation: A Survey". In: *Int. J. Comput. Vis.* 129 (2021), pp. 1789–1819.

[10] Jianping Gou et al. "Multilevel Attention-Based Sample Correlations for Knowledge Distillation". In: *IEEE Trans. Ind. Inform.* 19.5 (2023), pp. 7099–7109. DOI: 10.1109/TII.2022.3209672.

[11] Shuangchun Gui et al. "MT4MTL-KD: A Multi-Teacher Knowledge Distillation Framework for Triplet Recognition". In: *IEEE Transactions on Medical Imaging* 43.4 (2024), pp. 1628–1639. DOI: 10.1109/TMI.2023.3345736.

[12] Ziyao Guo et al. "Class Attention Transfer Based Knowledge Distillation". In: *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit. (CVPR).* 2023, pp. 11868–11877. DOI: 10.1109/CVPR52729.2023.01142.

[13] Rebeen Ali Hamad et al. "Self-Supervised Learning for Activity Recognition Based on Datasets With Imbalanced Classes". In: *IEEE Trans. Emerg. Top. Comput. Intell.* (2025), pp. 1–14. DOI: 10.1109/TETCI.2025.3526504.

[14] Jindong Han et al. "GraphConvLSTM: Spatiotemporal Learning for Activity Recognition with Wearable Sensors". In: *2019 IEEE Global Communications Conference (GLOBECOM).* 2019, pp. 1–6. DOI: 10.1109/GLOBECOM38437.2019.9013934.

[15] Zhiwei Hao et al. "One-for-All: Bridge the Gap Between Heterogeneous Architectures in Knowledge Distillation". In: *Proc. Adv. Neural Inf. Proces. Syst.* 2023, pp. 1–13.

[16] G. Hinton, O. Vinyals, and J. Dean. "Distillation the knowledge in a neural network". In: *arXiv preprint arXiv: 1503.02531* (2015).

[17] Rong Hu et al. "SWL-Adapt: An Unsupervised Domain Adaptation Model with Sample Weight Learning for Cross-User Wearable Human Activity Recognition". In: *Proc. AAAI Conf. Artif. Intell.* 2023, pp. 6012–6020.

[18] Wenbo Huang et al. "Channel-Equalization-HAR: A Light-weight Convolutional Neural Network for Wearable Sensor Based Human Activity Recognition". In: *IEEE Transactions on Mobile Computing* 22.9 (2023), pp. 5064–5077. DOI: 10.1109/TMC.2022.3174816.

[19] Wenbo Huang et al. "Deep Ensemble Learning for Human Activity Recognition Using Wearable Sensors via Filter Activation". In: *ACM Trans. Embed. Comput. Syst.* 22.1 (Oct. 2022). ISSN: 1539-9087.

[20] Dina Hussein and Ganapati Bhat. "CIM: A Novel Clustering-based Energy-Efficient Data Imputation Method for Human Activity Recognition". In: *ACM Trans. Embed. Comput. Syst.* 22.5s (Sept. 2023). ISSN: 1539-9087. DOI: 10.1145/3609111. URL: https://doi.org/10.1145/3609111.

[21] Dina Hussein and Ganapati Bhat. "SensorGAN: A Novel Data Recovery Approach for Wearable Human Activity Recognition". In: *ACM Trans. Embed. Comput. Syst.* 23.3 (May 2024). ISSN: 1539-9087.

[22] Mingi Ji et al. "Refine Myself by Teaching Myself: Feature Refinement via Self-Knowledge Distillation". In: *Proc IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit. (CVPR).* 2021, pp. 10659–10668. DOI: 10.1109/CVPR46437.2021.01052.

[23] Wei-Cheng Kao et al. "Specific Expert Learning: Enriching Ensemble Diversity via Knowledge Distillation". In: *IEEE Trans. Cybernetics* 53.4 (2023), pp. 2494–2505. DOI: 10.1109/TCYB.2021.3125320.

[24] Pritam Khan, Yogesh Kumar, and Sudhir Kumar. "CapsLSTM-Based Human Activity Recognition for Smart Healthcare With Scarce Labeled Data". In: *IEEE Trans. Comput. Social Syst.* 11.1 (2024), pp. 707–716. DOI: 10.1109/TCSS.2022.3223343.

[25] Simon Kornblith et al. "Similarity of Neural Network Representations Revisited". In: *International Conference on Machine Learning (ICML)*. Seminal work applying CKA to deep neural networks. 2019, pp. 3519–3529. URL: http://proceedings.mlr.press/v97/kornblith19a.html.

[26] Ying Li et al. "DWOSC: Dynamic Weight Optimization and Smoothness Constraint for Sensor-Based Human Activity Recognition". In: *IEEE Trans. Instrum. Meas.* 73 (2024), pp. 1–11. DOI: 10.1109/TIM.2024.3366277.

[27] Zhiyuan Li et al. "MKD-Cooper: Cooperative 3D Object Detection for Autonomous Driving via Multi-Teacher Knowledge Distillation". In: *IEEE Transactions on Intelligent Vehicles* 9.1 (2024), pp. 1490–1500. DOI: 10.1109/TIV.2023.3310580.

[28] Lukas Liedtke et al. "EStacker: Explaining Battery-Less IoT System Performance with Energy Stacks". In: *ACM Trans. Embed. Comput. Syst.* 25.1 (Jan. 2026). ISSN: 1539-9087.

[29] Ying Liu et al. "Wi-GPD Identification System Based on Gait Point Density". In: *ACM Trans. Embed. Comput. Syst.* 24.4 (July 2025). ISSN: 1539-9087.

[30] Fei Luo et al. "ActivityMamba: a CNN-Mamba hybrid neural network for efficient human activity recognition". In: *IEEE Transactions on Mobile Computing* (2025), pp. 1–15. DOI: 10.1109/TMC.2025.3544573.

[31] Fei Luo et al. "Binarized Neural Network for Edge Intelligence of Sensor-Based Human Activity Recognition". In: *IEEE Trans. Mobile Comput.* 22.3 (2023), pp. 1356–1368. DOI: 10.1109/TMC.2021.3109940.

[32] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *J. Mach. Learn. Res.* 9.11 (2008).

[33] Ihssene Menhour, M'hamed Bilal Abidine, and Belkacem Fergani. "A New Framework Using PCA, LDA and KNN-SVM to Activity Recognition Based SmartPhone's Sensors". In: *2018 6th International Conference on Multimedia Computing and Systems (ICMCS)*. 2018, pp. 1–5. DOI: 10.1109/ICMCS.2018.8525987.

[34] Shenghuan Miao et al. "Towards a Dynamic Inter-Sensor Correlations Learning Framework for Multi-Sensor-Based Wearable Human Activity Recognition". In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6.3 (Sept. 2022).

[35] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding". In: *arXiv preprint arXiv:1807.03748* (2018).

[36] Baoyun Peng et al. "Correlation Congruence for Knowledge Distillation". In: *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*. 2019, pp. 5006–5015. DOI: 10.1109/ICCV.2019.00511.

[37] Cuong Pham, Tuan Hoang, and Thanh-Toan Do. "Collaborative Multi-Teacher Knowledge Distillation for Learning Low Bit-width Deep Neural Networks". In: *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2023, pp. 6424–6432. DOI: 10.1109/WACV56688.2023.00637.

[38] Haotian Qiao et al. "Efficient Video Redaction at the Edge: Human Motion Tracking for Privacy Protection". In: *ACM Trans. Embed. Comput. Syst.* 24.5s (Sept. 2025). ISSN: 1539-9087.

[39] A. Romero et al. "FitNet: hints for thin deep nets". In: *Proc. Int. Conf. Learn. Represent.* 2015, pp. 1–13.

[40] Sandipan Sarma, Divyam Singal, and Arijit Sur. "LoCATe-GAT: Modeling Multi-Scale Local Context and Action Relationships for Zero-Shot Action Recognition". In: *IEEE Trans. Emerg. Top. Comput. Intell.* 9.4 (2025), pp. 2793–2805. DOI: 10.1109/TETCI.2024.3499995.

[41] F. Serpush et al. "Wearable Sensor-Based Human Activity Recognition in the Smart Healthcare System". In: *Comput. Intel. Neurosc.* (2022), pp. 1–31. DOI: 10.1155/2022/1391906.

[42] Wonchul Son et al. "Densely Guided Knowledge Distillation using Multiple Teacher Assistants". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 9375–9384. DOI: 10.1109/ICCV48922.2021.00926.

[43] Zhiyong Tao et al. "Wireless Perceptual Space Modeling Method for Cross-Domain Human Activity Recognition". In: *ACM Trans. Embed. Comput. Syst.* 24.3 (Apr. 2025). ISSN: 1539-9087.

[44] Yonglong Tian, Dilip Krishnan, and Phillip Isola. "Contrastive representation distillation". In: *Proc. Int. Conf. Learn. Repre.* 2020, pp. 1–19.

[45] Xinshao Wang et al. "ProSelfLC: Progressive Self Label Correction for Training Robust Deep Neural Networks". In: *Proc IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit. (CVPR)*. 2021, pp. 752–761. DOI: 10.1109/CVPR46437.2021.00081.

[46] Y. Wang, S. Cang, and H. Yu. "A survey on wearable sensor modality centred human activity recognition in health care". In: *Expert Syst. Appl.* 127 (2018), pp. 167–190.

[47] Chixuan Wei et al. "SemiHAR: Improving Semisupervised Human Activity Recognition via Multitask Learning". In: *IEEE Trans. Neur. Net. Lear.* 36.1 (2025), pp. 1884–1897. DOI: 10.1109/TNNLS.2023.3330879.

[48] Z. Xiao et al. "A federated learning system with enhanced feature extraction for human activity recognition". In: *Knowl.-Based Syst.* 229 (2021), pp. 1–14.

[49] Z. Xiao et al. "RTFN: A robust temporal feature network for time series classification". In: *Inf. Sci.* 571 (2021), pp. 65–86. DOI: 10.1016/j.ins.2021.04.053.

[50] Zhiwen Xiao et al. "CapMatch: Semi-supervised Contrastive Transformer Capsule with Feature-based Knowledge Distillation for Human Activity Recognition". In: *IEEE Trans. Neur. Net. Lear.* 36.2 (2025), pp. 2690–2704. DOI: 10.1109/TNNLS.2023.3344294.

[51] Zhiwen Xiao et al. "Heterogeneous Mutual Knowledge Distillation for Wearable Human Activity Recognition". In: *IEEE Trans. Neur. Net. Lear.* (2025), pp. 1–15. DOI: 10.1109/TNNLS.2025.3556317.

[52] Zhiwen Xiao et al. "Self-Bidirectional Decoupled Distillation for Time Series Classification". In: *IEEE Trans. Artif. Intell.* 5.8 (2024), pp. 4101–4110. DOI: 10.1109/TAI.2024.3360180.

[53] H. Xing et al. "SelfMatch: Robust semisupervised time-series classification with self-distillation". In: *Int. J. Intell. Syst.* (2022), pp. 1–28. DOI: 10.1002/int.22957.

[54] Qing Xu et al. "Contrastive Distillation With Regularized Knowledge for Deep Model Compression on Sensor-Based Human Activity Recognition". In: *IEEE Trans. Ind. Cyber-Phys. Syst.* 1 (2023), pp. 217–226. DOI: 10.1109/TICPS.2023.3320630.

[55] S. Xu et al. "Deformable Convolutional Networks for Multimodal Human Activity Recognition Using Wearable Sensors". In: *IEEE Trans. Instrum. Meas.* 71 (2022), pp. 1–14.

[56] Surong Yan, Kwei-Jay Lin, and Haosen Wang. "Toward Lightweight End-to-End Semantic Learning of Real-Time Human Activity Recognition for Enabling Ambient Intelligence". In: *EEE Trans. Knowl. Data Eng.* (2024), pp. 1–14. DOI: 10.1109/TKDE.2024.3386794.

[57] A. Yao and D. Sun. "Knowledge transfer via dense cross-layer mutual-distillation". In: *Proc. Lect. Notes Comput. Sci., ECCV*. 2020, pp. 294–311.

[58] Minghui Yao et al. "Long kernel distillation in human activity recognition". In: *Knowl.-Based Syst.* 316 (2025), pp. 566–570. DOI: 10.1016/j.knosys.2025.113397.

[59] Xin Ye et al. "Knowledge Distillation via Multi-Teacher Feature Ensemble". In: *IEEE Signal Proc. Let.* 31 (2024), pp. 566–570. DOI: 10.1109/LSP.2024.3359573.

[60] George Zerveas et al. "A Transformer-based Framework for Multivariate Time Series Representation Learning". In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. KDD '21. Virtual Event, Singapore: Association for Computing Machinery, 2021, 2114–2124. ISBN: 9781450383325. DOI: 10.1145/3447548.3467401. URL: https://doi.org/10.1145/3447548.3467401.

[61] H. Zhang et al. "A novel IoT-perceptive human activity recognition (HAR) approach using multihead convolutional attention". In: *IEEE Internet Things J.* 7.2 (2020), pp. 1072–1080.

[62] Hailin Zhang, Defang Chen, and Can Wang. "Adaptive Multi-Teacher Knowledge Distillation with Meta-Learning". In: *2023 IEEE International Conference on Multimedia and Expo (ICME)*. 2023, pp. 1943–1948. DOI: 10.1109/ICME55011.2023.00333.

[63] Haoran Zhang and Linhai Xu. "Multi-STMT: Multi-Level Network for Human Activity Recognition Based on Wearable Sensors". In: *IEEE Trans. Instrum. Meas.* 73 (2024), pp. 1–12. DOI: 10.1109/TIM.2024.3365155.

[64] Haoxi Zhang et al. "A Novel IoT-Perceptive Human Activity Recognition (HAR) Approach Using Multihead Convolutional Attention". In: *IEEE Internet Things J.* 7.2 (2020), pp. 1072–1080. DOI: 10.1109/JIOT.2019.2949715.

[65] Jiale Zhang et al. "BadCleaner: Defending Backdoor Attacks in Federated Learning via Attention-Based Multi-Teacher Distillation". In: *IEEE Transactions on Dependable and Secure Computing* 21.5 (2024), pp. 4559–4573. DOI: 10.1109/TDSC.2024.3354049.

[66] Linfeng Zhang, Chenglong Bao, and Kaisheng Ma. "Self-Distillation: Towards Efficient and Compact Neural Networks". In: *IEEE Trans. Pattern Anal. Intell.* 44.8 (2022), pp. 4388–4403. DOI: 10.1109/TPAMI.2021.3067100.

[67] Linfeng Zhang et al. "Be Your Own Teacher: Improve the Performance of Convolutional Neural Networks via Self Distillation". In: *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*. 2019, pp. 3712–3721. DOI: 10.1109/ICCV.2019.00381.

[68] Y. Zhang et al. "Deep mutual learning". In: *Proc IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit. (CVPR)*. 2018, pp. 4320–4328.

[69] Borui Zhao et al. "Decoupled Knowledge Distillation". In: *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit. (CVPR)*. 2022, pp. 11943–11952. DOI: 10.1109/CVPR52688.2022.01165.

[70] H. Zhao et al. "Highlight Every Step: Knowledge Distillation via Collaborative Teaching". In: *IEEE Trans. Cybernetics* 52.4 (2022), pp. 2070–2081.

[71] Yu Zhou et al. "Energy-Efficient and Interpretable Multisensor Human Activity Recognition via Deep Fused Lasso Net". In: *IEEE Trans. Emerg. Top. Comput. Intell.* 8.5 (2024), pp. 3576–3588. DOI: 10.1109/TETCI.2024.3430008.

[72] C. Zhu and W. Sheng. "Wearable Sensor-Based Hand Gesture and Daily Activity Recognition for Robot-Assisted Living". In: *IEEE Trans. Syst. Man Cybern. Part A Syst. Humans* 41.3 (2011), pp. 569–573.

[73] Yida Zhu et al. "DiamondNet: A Neural-Network-Based Heterogeneous Sensor Attentive Fusion for Human Activity Recognition". In: *IEEE Trans. Neur. Net. Lear.* 35.11 (2024), pp. 15321–15331. DOI: 10.1109/TNNLS.2023.3285547.

[74] Hailin Zou et al. "GT-WHAR: A Generic Graph-Based Temporal Framework for Wearable Human Activity Recognition With Multiple Sensors". In: *IEEE Trans. Emerg. Top. Comput. Intell.* 8.6 (2024), pp. 3912–3924. DOI: 10.1109/TETCI.2024.3378331.

[75] Hailin Zou et al. "STFNet: Enhanced and Lightweight Spatiotemporal Fusion Network for Wearable Human Activity Recognition". In: *IEEE Sensors Journal* 24.8 (2024), pp. 13686–13698. DOI: 10.1109/JSEN.2024.3373444.

A MATHEMATICAL ANALYSIS AND OPTIMIZATION THEORY OF DHMMD

This appendix provides an assumption-dependent analysis of the main optimization operators used in DHMMD. It does not introduce additional modules beyond the main framework. Appendix A.1 analyzes the bounded behavior of attention-based aggregation for heterogeneous teacher probabilities and intermediate features. Appendix A.2 studies the gradient behavior of bidirectional decoupled semantic distillation. Appendix A.3 examines the contrastive feature objective, and Appendix A.4 discusses the local stability interpretation of the collective co-optimization strategy.

A.1 Analytical Dynamics of Attention-Based Aggregation

The purpose of this subsection is to analyze the aggregation operators used in DHMMD rather than to state a generic result for attention mechanisms. In DHMMD, these operators combine heterogeneous teacher outputs and transformed intermediate features into ensemble-level representations, which are then used in semantic and feature distillation.

For the LSTMaN module, let $A = \mathrm{Softmax}(Z)$, where $Z = Query \cdot Key^{\top} \in \mathbb{R}^{T\times T}$ and T is the sequence length. The attention output is $O_{LSTMatt} = AValue$. To avoid ambiguity in tensor indexing, the derivative is written row-wise. For the r-th query position and the s-th key position,

$$\frac{\partial O_{LSTMatt,r,:}}{\partial Z_{r,s}} = A_{r,s}\left(Value_{s,:} - \sum_{\ell=1}^{T} A_{r,\ell} Value_{\ell,:}\right). \tag{A1}$$

This expression corresponds to the contraction between the softmax Jacobian and the value matrix. Under bounded value embeddings, the Jacobian norm of the attention operator is finite. Specifically, if $\|Value\|_2 \le C_V$, then there exists a finite constant L_{att} such that

$$\|\nabla_Z O_{LSTMatt}\|_2 \le L_{att} C_V. \tag{A2}$$

This bound gives a local stability interpretation for gradient propagation through the attention operator under the stated norm condition.

At the ensemble level, classification-probability aggregation uses temperature-scaled teacher distributions. Let

$$\alpha_{k,i} = \frac{\exp\left(W_{CP}^{\top} P_i^{\mathcal{T}_k}\right)}{\sum_{h=1}^{3} \exp\left(W_{CP}^{\top} P_i^{\mathcal{T}_h}\right)}, \qquad P_{en,i}^{\mathcal{T}} = \sum_{k=1}^{3} \alpha_{k,i} P_i^{\mathcal{T}_k}. \tag{A3}$$

The gradient of the aggregated probability with respect to W_{CP} can then be written as

$$\nabla_{W_{CP}} P_{en,i}^{\mathcal{T}} = \sum_{k=1}^{3} \alpha_{k,i} P_i^{\mathcal{T}_k} \otimes \left(P_i^{\mathcal{T}_k} - P_{en,i}^{\mathcal{T}}\right). \tag{A4}$$

Similarly, for the scalar non-target probability mass, let

$$\alpha_{k,i,\backslash} = \frac{\exp\left(W_{CP,\backslash} P_{i,\backslash}^{\mathcal{T}_k}\right)}{\sum_{h=1}^{3} \exp\left(W_{CP,\backslash} P_{i,\backslash}^{\mathcal{T}_h}\right)}, \qquad P_{en,i,\backslash}^{\mathcal{T}} = \sum_{k=1}^{3} \alpha_{k,i,\backslash} P_{i,\backslash}^{\mathcal{T}_k}. \tag{A5}$$

Its gradient follows

$$\nabla_{W_{CP,\backslash}} P_{en,i,\backslash}^{\mathcal{T}} = \sum_{k=1}^{3} \alpha_{k,i,\backslash} P_{i,\backslash}^{\mathcal{T}_k} \left(P_{i,\backslash}^{\mathcal{T}_k} - P_{en,i,\backslash}^{\mathcal{T}}\right). \tag{A6}$$

For layer-wise feature aggregation, the transformation ϕ_j maps teacher features into a shared feature space. For $\tilde{F}_{j,i}^{\mathcal{T}_k} = \phi_j(F_{j,i}^{\mathcal{T}_k})$ with parameters θ_ϕ, assume the local curvature is bounded as

$$\nabla_{\theta_\phi}^2 \tilde{F}_{j,i}^{\mathcal{T}_k} \preceq \Lambda_{\max}\mathbf{I}. \tag{A7}$$

The sample-specific teacher-level attention weights satisfy the simplex constraint $\sum_{k=1}^{3} \omega_{k,i} = 1$. Under this constrained weighting view, a natural-gradient expression for W_{AE} can be written as

$$\tilde{\nabla}_{W_{AE}} \mathcal{L} = F_\omega^{-1} \nabla_{W_{AE}} \mathcal{L}, \tag{A8}$$

where F_ω denotes the Fisher information matrix associated with the sample-specific attention weights:

$$F_\omega = \mathbb{E}\left[\left(\nabla_{W_{AE}} \log \omega_{k,i}\right) \left(\nabla_{W_{AE}} \log \omega_{k,i}\right)^\top \right]. \tag{A9}$$

For the student intermediate feature

$$F_{en,i}^{\mathcal{S}} = \sum_{j=1}^{L^{\mathcal{S}}} \beta_{j,i}\, \psi_j(F_{j,i}^{\mathcal{S}}), \tag{A10}$$

the same simplex weighting principle applies to $\beta_{j,i}$. If the transformed features and attention parameters are bounded, then the aggregated student feature has finite variance:

$$\mathrm{Tr}\Big(\mathrm{Cov}(F_{en,i}^{\mathcal{S}})\Big) < \infty. \tag{A11}$$

Thus, the aggregation operators used in DHMMD remain bounded under the stated assumptions. In framework terms, this supports the use of ensemble-level probability and feature signals without requiring a claim of global convergence or an independently optimal attention solution.

A.2 Gradient Dynamics of Bidirectional Decoupled Semantic Distillation

This subsection analyzes the JS-based decoupled probability loss used between teachers and the student. The goal is to interpret why reciprocal semantic supervision can be used in both directions without introducing unbounded local gradients under the stated assumptions.

The bidirectional decoupled semantic distillation relies on the JS divergence formulation

$$\mathcal{L}_{DKD}(\{P^A, P_{\setminus}^A\}, \{P^B, P_{\setminus}^B\}) = \mu\, \mathrm{JS}(P^A, P^B) + \lambda\, \mathrm{JS}(P_{\setminus}^A, P_{\setminus}^B).$$

Let $M^T = \frac{1}{2}(P^A + P^B)$ and $M^{NT} = \frac{1}{2}(P_{\setminus}^A + P_{\setminus}^B)$. The two JS terms are written as

$$\mathrm{JS}(P^A, P^B) = \frac{1}{2} D_{KL}(P^A \| M^T) + \frac{1}{2} D_{KL}(P^B \| M^T), \tag{A12}$$

$$\mathrm{JS}(P_{\setminus}^A, P_{\setminus}^B) = \frac{1}{2} D_{KL}(P_{\setminus}^A \| M^{NT}) + \frac{1}{2} D_{KL}(P_{\setminus}^B \| M^{NT}). \tag{A13}$$

For the student-side probability mapping,

$$\frac{\partial P_{i,c}^{\mathcal{S}}}{\partial O_{i,m}^{\mathcal{S}}} = \frac{1}{t_{DKD}} P_{i,c}^{\mathcal{S}} (\delta_{cm} - P_{i,m}^{\mathcal{S}}). \tag{A14}$$

The derivative of the target JS term with respect to $P^{\mathcal{S}}$ is

$$\nabla_{P^{\mathcal{S}}} \mathrm{JS}(P^{\mathcal{T}}, P^{\mathcal{S}}) = \frac{1}{2} \log\left(\frac{P^{\mathcal{S}}}{M^T}\right). \tag{A15}$$

Applying the chain rule, the target-class contribution to the gradient of the m-th logit becomes

$$\frac{\partial \mathcal{L}_{DKD}^{T}}{\partial O_{i,m}^{\mathcal{S}}} = \frac{\mu}{2t_{DKD}} \sum_{c=1}^{C} \log\left(\frac{2P_{i,c}^{\mathcal{S}}}{P_{i,c}^{\mathcal{T}} + P_{i,c}^{\mathcal{S}}}\right) P_{i,c}^{\mathcal{S}}(\delta_{cm} - P_{i,m}^{\mathcal{S}}). \tag{A16}$$

For the scalar non-target mass used in the main text, $P_{i,\setminus m}^{\mathcal{S}} = 1 - P_{i,m}^{\mathcal{S}}$. Therefore, for $q \neq m$,

$$\frac{\partial P_{i,\setminus m}^{\mathcal{S}}}{\partial O_{i,q}^{\mathcal{S}}} = \frac{1}{t_{DKD}} P_{i,q}^{\mathcal{S}} P_{i,m}^{\mathcal{S}}. \tag{A17}$$

The corresponding non-target contribution is

$$\frac{\partial}{\partial O_{i,q}^{\mathcal{S}}}\left[\lambda \,\mathrm{JS}(P_{\setminus}^{\mathcal{T}}, P_{\setminus}^{\mathcal{S}})\right] = \frac{\lambda}{2t_{DKD}} \log\left(\frac{2P_{i,\setminus m}^{\mathcal{S}}}{P_{i,\setminus m}^{\mathcal{T}} + P_{i,\setminus m}^{\mathcal{S}}}\right) P_{i,q}^{\mathcal{S}} P_{i,m}^{\mathcal{S}}, \qquad q \neq m. \tag{A18}$$

Combining the target and non-target components, the local curvature of the decoupled semantic distillation loss can be bounded under finite probabilities and bounded logits. Thus, for some finite constant $\kappa_{\max}$,

$$\mathrm{Tr}\left(\nabla_{O^{\mathcal{S}}}^{2} \mathcal{L}_{\mathrm{DSD}}^{A\to B}\right) \leq \frac{\gamma\mu + \lambda}{t_{DKD}^{2}} \kappa_{\max}. \tag{A19}$$

This bound gives a local curvature interpretation rather than a global convergence guarantee. Because JS divergence is symmetric, the teacher-to-student and student-to-teacher semantic objectives have comparable local curvature structures:

$$\nabla_{O^{\mathcal{T}_k}}^{2} \mathcal{L}_{\mathrm{DSD}}^{\mathcal{S}\to\mathcal{T}_k} \approx \nabla_{O^{\mathcal{S}}}^{2} \mathcal{L}_{\mathrm{DSD}}^{\mathcal{T}_k\to\mathcal{S}}. \tag{A20}$$

The result indicates that reciprocal semantic transfer can encourage closer teacher and student output distributions while keeping the local optimization behavior bounded under the stated assumptions.

A.3 Latent Space Uniformity in Bidirectional Contrastive Feature Distillation

The bidirectional contrastive feature distillation follows the InfoNCE form used in Eq. (15). The matched representations (F^A, F^B) form the positive pair, and the negative features are sampled from other instances in the same mini-batch. Let $s(A,B) = \mathrm{sim}(F^A, F^B)$. The loss can be written as

$$\mathcal{L}_{CL}(F^A, F^B) = -\frac{s(A,B)}{\tau} + \log\left[\exp\left(\frac{s(A,B)}{\tau}\right) + \sum_{j=1}^{N_{\mathrm{neg}}} \exp\left(\frac{s(A,F_j^-)}{\tau}\right)\right]. \tag{A21}$$

The expected risk can be interpreted through alignment and normalization terms:

$$\mathbb{E}[\mathcal{L}_{CL}] \approx -\frac{1}{\tau}\mathbb{E}_{(A,B)\sim P_{pos}}[s(A,B)] + \mathbb{E}_A\left[\log\left(\exp(s(A,B)/\tau) + \mathbb{E}_{F^-\sim P}[\exp(s(A,F^-)/\tau)]\right)\right]. \tag{A22}$$

The gradient with respect to F^A is

$$\nabla_{F^A}\mathcal{L}_{CL} = -\frac{1}{\tau}\nabla_{F^A}s(A,B) + \frac{1}{\tau}P_{pos}(B|A)\nabla_{F^A}s(A,B) + \frac{1}{\tau}\sum_{j=1}^{N_{\mathrm{neg}}} P_{neg}(j|A)\nabla_{F^A}s(A,F_j^-), \tag{A23}$$

where

$$P_{neg}(j|A) = \frac{\exp(s(A,F_j^-)/\tau)}{\exp(s(A,B)/\tau) + \sum_{q=1}^{N_{\mathrm{neg}}}\exp(s(A,F_q^-)/\tau)} \tag{A24}$$

and $P_{pos}(B|A)$ is defined by the same denominator with numerator $\exp(s(A,B)/\tau)$.

For cosine similarity

$$s(A,B) = \frac{(F^A)^\top F^B}{\|F^A\|\|F^B\|},$$

the partial derivative can be written as

$$\nabla_{F^A} s(A,B) = \frac{1}{\|F^A\|}\left(I - \frac{F^A(F^A)^\top}{\|F^A\|^2}\right)\frac{F^B}{\|F^B\|}. \tag{A25}$$

For the student-side feature objective

$$\mathcal{L}_{\mathrm{CFD}}^{\mathcal{T}_k\to\mathcal{S}} = \gamma\mathcal{L}_{CL}(F_{en,i}^{\mathcal{T}_k}, F_{en,i}^{\mathcal{S}}) + \mathcal{L}_{CL}(F_{en,i}^{\mathcal{T}}, F_{en,i}^{\mathcal{S}}),$$

the gradient with respect to the encoded student feature is

$$\nabla_{F_{en,i}^{\mathcal{S}}}\mathcal{L}_{\mathrm{CFD}}^{\mathcal{T}_k\to\mathcal{S}} = \gamma\nabla_{F^{\mathcal{S}}}\mathcal{L}_{CL}(F_{en,i}^{\mathcal{T}_k}, F^{\mathcal{S}}) + \nabla_{F^{\mathcal{S}}}\mathcal{L}_{CL}(F_{en,i}^{\mathcal{T}}, F^{\mathcal{S}}). \tag{A26}$$

If the normalized features are bounded and $\tau > 0$, then the contrastive gradient is also bounded. Therefore, for some finite constant C_{CL},

$$\left\|\nabla_{F_{en,i}^{\mathcal{S}}}\mathcal{L}_{\mathrm{CFD}}^{\mathcal{T}_k\to\mathcal{S}}\right\|_2 \le (1+\gamma)C_{CL}. \tag{A27}$$

The teacher-side feedback objective follows the same bounded-gradient form. This gives a local interpretation of contrastive feature transfer in DHMMD: the objective encourages matched cross-model representations to move closer while retaining separation from batch-level negatives.

The InfoNCE lower-bound interpretation can be written with the number of contrastive candidates $N_{\mathrm{neg}} + 1$:

$$I(F_{en,i}^{\mathcal{S}}; F_{en,i}^{\mathcal{T}}) \ge \log(N_{\mathrm{neg}}+1) - \mathcal{L}_{CL}(F_{en,i}^{\mathcal{S}}, F_{en,i}^{\mathcal{T}}), \tag{A28}$$

$$I(F_{en,i}^{\mathcal{S}}; F_{en,i}^{\mathcal{T}_k}) \ge \log(N_{\mathrm{neg}}+1) - \mathcal{L}_{CL}(F_{en,i}^{\mathcal{S}}, F_{en,i}^{\mathcal{T}_k}). \tag{A29}$$

These bounds do not prove that the mutual information is globally maximized. They indicate that minimizing the contrastive objective encourages a larger lower bound between the student representation and the teacher-side representations. This interpretation is consistent with the role of BCFD in improving cross-model feature compatibility across heterogeneous architectures.

A.4 Co-Optimization and Local Stability Interpretation

The collective loss objective and its local optimization behavior are analyzed below. Let the complete parameter state be

$$\Theta_m = \{\theta_m^{\mathcal{T}_1}, \theta_m^{\mathcal{T}_2}, \theta_m^{\mathcal{T}_3}, \theta_m^{\mathcal{S}}\}.$$

The unified objective is

$$\mathcal{J}(\Theta) = \sum_{k=1}^{3}\mathcal{L}^{\mathcal{T}_k} + \mathcal{L}^{\mathcal{S}}. \tag{A30}$$

The supervised component for any model $\mathcal{M}$ is

$$\mathcal{L}_{sup}^{\mathcal{M}} = -\frac{1}{N_{\mathrm{IS}}}\sum_{i=1}^{N_{\mathrm{IS}}}\sum_{c=1}^{C} y_{i,c}\log\left(P_{i,c}^{\mathcal{M}}\right). \tag{A31}$$

Assume that the collective objective is L-smooth, namely

$$\|\nabla\mathcal{J}(\Theta) - \nabla\mathcal{J}(\Theta')\|_2 \le L\|\Theta - \Theta'\|_2. \tag{A32}$$

For a learning rate $\eta < 2/L$, the descent lemma gives

$$\mathcal{J}(\Theta_m) \le \mathcal{J}(\Theta_{m-1}) - \eta\left(1 - \frac{L\eta}{2}\right)\|\nabla\mathcal{J}(\Theta_{m-1})\|_2^2. \tag{A33}$$

The coupling between teacher and student updates can be described through cross-curvature terms rather than by taking an inner product between gradients that may belong to different parameter spaces. For the contrastive component, a representative mixed derivative has the form

$$\frac{\partial^2 \mathcal{L}_{CL}}{\partial \theta^{\mathcal{S}} \partial \theta^{\mathcal{T}_k}} = \left(\frac{\partial F^{\mathcal{S}}_{en,i}}{\partial \theta^{\mathcal{S}}}\right)^{\top} \frac{\partial^2 \mathcal{L}_{CL}}{\partial F^{\mathcal{S}} \partial F^{\mathcal{T}_k}} \left(\frac{\partial F^{\mathcal{T}_k}_{en,i}}{\partial \theta^{\mathcal{T}_k}}\right). \tag{A34}$$

For the JS-based semantic component, the target-class cross-partial term is locally negative along each probability coordinate:

$$\frac{\partial^2}{\partial P_c^{\mathcal{S}} \partial P_c^{\mathcal{T}_k}} \mathrm{JS}(P^{\mathcal{S}}, P^{\mathcal{T}_k}) = -\frac{1}{2(P_c^{\mathcal{S}} + P_c^{\mathcal{T}_k})} \leq 0. \tag{A35}$$

This term is not a global convergence proof; it only indicates that the reciprocal semantic loss provides a bounded corrective interaction under finite probabilities.

Summing Eq. (A33) from $m = 1$ to E yields

$$\sum_{m=1}^{E} \eta \left(1 - \frac{L\eta}{2}\right) \|\nabla \mathcal{J}(\Theta_{m-1})\|_2^2 \leq \mathcal{J}(\Theta_0) - \mathcal{J}(\Theta_E). \tag{A36}$$

If $\mathcal{J}_{\inf}$ is a finite lower bound of the objective, then

$$\min_{1 \leq m \leq E} \|\nabla \mathcal{J}(\Theta_m)\|_2^2 \leq \frac{2(\mathcal{J}(\Theta_0) - \mathcal{J}_{\inf})}{E\eta(2 - L\eta)}. \tag{A37}$$

This is a standard sublinear stationarity bound under the smoothness assumption. It does not establish convergence to a global or strict local minimum for the non-convex deep models.

For a stochastic mini-batch update, let $\Delta\theta^{\mathcal{S}}$ and $\Delta\theta^{\mathcal{T}_k}$ denote one-step updates. The update directions are

$$\Delta\theta^{\mathcal{S}} = -\eta^{\mathcal{S}} \left(\nabla \mathcal{L}^{\mathcal{S}}_{sup} + \sum_{k=1}^{3} \left[\nabla \mathcal{L}^{\mathcal{T}_k \to \mathcal{S}}_{\mathrm{CFD}} + \nabla \mathcal{L}^{\mathcal{T}_k \to \mathcal{S}}_{\mathrm{DSD}}\right]\right), \tag{A38}$$

$$\Delta\theta^{\mathcal{T}_k} = -\eta^{\mathcal{T}_k} \left(\nabla \mathcal{L}^{\mathcal{T}_k}_{sup} + \nabla \mathcal{L}^{\mathcal{S} \to \mathcal{T}_k}_{\mathrm{CFD}} + \nabla \mathcal{L}^{\mathcal{S} \to \mathcal{T}_k}_{\mathrm{DSD}}\right). \tag{A39}$$

Assuming bounded stochastic-gradient variance,

$$\mathbb{E}\left[\|\hat{g}_m - \nabla \mathcal{J}(\Theta_m)\|^2\right] \leq \sigma^2. \tag{A40}$$

Taking expectation over the stochastic trajectory gives

$$\mathbb{E}[\mathcal{J}(\Theta_m)] \leq \mathbb{E}[\mathcal{J}(\Theta_{m-1})] - \eta \left(1 - \frac{\eta L}{2}\right) \mathbb{E}[\|\nabla \mathcal{J}(\Theta_{m-1})\|^2] + \frac{\eta^2 L \sigma^2}{2}. \tag{A41}$$

This inequality gives a bounded-variance interpretation of the stochastic co-optimization trajectory. It shows descent up to a variance-controlled residual term, not convergence to a strict local minimum.

The following inequality gives an upper-bound interpretation of how attention weighting enters a complexity-related term. If $\mathbf{K} \in \mathbb{R}^{3 \times N_{\mathrm{IS}}}$ denotes the teacher output space, the empirical Rademacher complexity can be bounded as

$$\hat{\mathfrak{R}}_S(\mathcal{H}) \leq \frac{1}{N_{\mathrm{IS}}} \sqrt{\mathrm{Tr}(\mathbf{K}^{\top}\mathbf{K})} \sqrt{\sum_{k=1}^{3} \omega_{k,i}^2}. \tag{A42}$$

Because the sample-specific attention weights lie on the simplex, $\sum_{k=1}^{3} \omega_{k,i} = 1$, and therefore $\sum_{k=1}^{3} \omega_{k,i}^2 \leq 1$. This property does not prove effective complexity control for the full non-convex

model. It only indicates that the aggregation-related factor in the above upper-bound view remains bounded under the stated attention weighting scheme.

The decoupled coefficients indicate that increasing λ strengthens the contribution of the non-target JS term, while increasing μ strengthens the target JS term:

$$\lim_{P^A \to P^B} \nabla_{P^A} \mathcal{L}_{DKD} = \mathbf{0}, \tag{A43}$$

$$\frac{\partial}{\partial \lambda} \mathcal{L}_{DKD}(\{P^A, P^A_{\setminus}\}, \{P^B, P^B_{\setminus}\}) = \mathrm{JS}(P^A_{\setminus}, P^B_{\setminus}) \geq 0, \tag{A44}$$

$$\frac{\partial}{\partial \mu} \mathcal{L}_{DKD}(\{P^A, P^A_{\setminus}\}, \{P^B, P^B_{\setminus}\}) = \mathrm{JS}(P^A, P^B) \geq 0. \tag{A45}$$

Substituting the partial gradients into the full parameter vector gives

$$\nabla_{\Theta} \mathcal{J}(\Theta) = \begin{bmatrix} \nabla_{\theta^{\mathcal{T}_1}} \mathcal{L}^{\mathcal{T}_1} \\ \nabla_{\theta^{\mathcal{T}_2}} \mathcal{L}^{\mathcal{T}_2} \\ \nabla_{\theta^{\mathcal{T}_3}} \mathcal{L}^{\mathcal{T}_3} \\ \nabla_{\theta^{\mathcal{S}}} \mathcal{L}^{\mathcal{S}} \end{bmatrix}. \tag{A46}$$

Under the stated smoothness and bounded-variance assumptions, the joint update in Algorithm 1 can be interpreted through a local stability condition. Let $H(\Theta) = \nabla^2_{\Theta} \mathcal{J}(\Theta)$ denote the local curvature matrix of the collective objective. If the learning rate is sufficiently small and the local update map satisfies

$$\rho(I - \eta H(\Theta)) < 1, \tag{A47}$$

then the update trajectory remains locally bounded around a stationary region. In the deterministic or noise-vanishing case, this condition reduces to the following step-size behavior:

$$\lim_{m \to \infty} \|\Theta_m - \Theta_{m-1}\| = 0. \tag{A48}$$

Equivalently, for each model component $\mathcal{M} \in \{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3, \mathcal{S}\}$,

$$\lim_{m \to \infty} \left\| \theta^{\mathcal{M}}_m - \theta^{\mathcal{M}}_{m-1} \right\| = 0. \tag{A49}$$

These relations provide a local stability interpretation of the offline co-optimization dynamics. They neither imply global convergence nor establish convergence to a strict local minimum. Under the stated assumptions and local curvature condition, the teacher and student updates can remain within a bounded stationary neighborhood during joint training.

B ARCHITECTURAL DETAILS OF TEACHER AND STUDENT MODELS

Building on recent advances in heterogeneous distillation for wearable HAR [51], this study adopts three structurally diverse teacher models, FCNLSTMaN [48], ResNetLSTMaN [53], and TransformerNet [60]. Together, they form a heterogeneous teacher ensemble that integrates convolutional, residual, LSTM, attention, and transformer architectures. These backbone networks are reused from prior work. The framework-specific contribution lies in their new configuration within DHMMD. Reference [51] is a prior single-teacher heterogeneous distillation study for wearable HAR, whereas DHMMD extends this line to a three-teacher setting with ensemble-level aggregation and bidirectional semantic and feature distillation paths. Guided by prior multi-teacher distillation strategies [62, 27], this setup expands architectural variety to strengthen knowledge diversity. Two lightweight students, a three-block CNN and a three-layer MLP, are used to evaluate distillation into compact architectures. No assertion is made that this teacher combination is theoretically optimal under exhaustive architecture search. It is selected as an empirically effective and complementary

ensemble for wearable HAR, where local motion primitives, hierarchical temporal abstractions, and long-range dependencies all play important roles.

The rationale for this ensemble is as follows. FCNLSTMaN couples convolutional local-pattern extraction with attention-enhanced recurrent modeling, and is therefore well suited to capturing fine-grained motion fragments together with sequential context. ResNetLSTMaN preserves this recurrent-attentive capability while introducing residual hierarchical convolutional blocks that improve optimization stability and promote deeper temporal abstraction. TransformerNet complements both LSTM-based teachers through self-attention, which is particularly effective for modeling long-range dependencies and global interactions across the input sequence. In wearable HAR, inertial signals often contain both short-duration motion cues and broader activity-level temporal structure. These three teachers therefore provide partially overlapping yet non-redundant inductive biases, and together they enrich the teacher-side knowledge space more effectively than any single architecture in isolation.

B.1 FCNLSTMaN

FCNLSTMaN integrates a FCN and an LSTMaN module in a parallel architecture, as illustrated in Fig. 1(a), enabling the joint modeling of spatial and temporal features.

FCN The FCN branch comprises three hierarchical convolutional blocks dedicated to local feature extraction. Each block is structured as a sequential composition of a 1-dimensional convolutional layer, followed by batch normalization and activation via the leaky ReLU function. This design progressively refines local temporal patterns while ensuring stable gradient propagation and non-linear transformation.

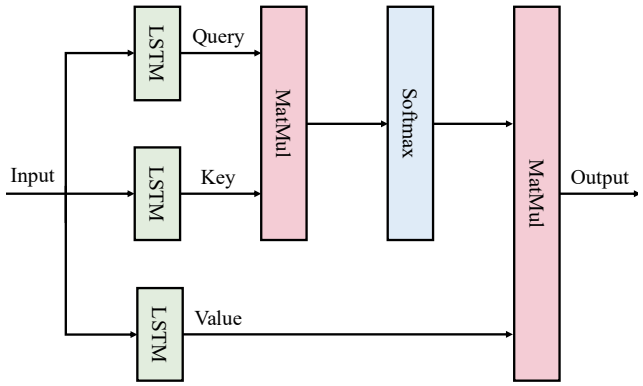


Fig. 6. Structural illustration of the LSTM-based attention mechanism [49]. The operation denoted by "MatMul" corresponds to matrix multiplication, while the "Softmax" function normalizes the resulting matrix into a probabilistic distribution over attention weights.

LSTMaN The LSTMaN module integrates two attention-enhanced LSTM layers to capture global temporal dependencies. Each layer embeds LSTM networks within an attention framework [49], as illustrated in Fig. 6, enabling the model to dynamically attend to salient temporal features. In this mechanism, a query vector $Query$ interacts with a set of key–value pairs $(Key, Value)$ to generate the output $O_{LSTMatt}$, which is formally defined in Eq. (A50). Specifically, $Query$, Key, and $Value$ are feature vectors derived from three separate LSTM encoders. The attention output is computed as:

$$O_{LSTMatt} = Softmax(Query \cdot Key^{\top}) \cdot Value \tag{A50}$$

where the *Softmax* function transforms the similarity scores into a normalized probability distribution, and $Key^{\top}$ denotes the transpose of the key matrix. This formulation allows the model to adaptively emphasize informative temporal patterns across the sequence.

B.1.1 ResNetLSTMaN. ResNetLSTMaN comprises a parallel combination of a ResNet and an LSTMaN module, as depicted in Fig. 1(e), aiming to capture both local and global temporal representations.

ResNet The ResNet component is constructed from three hierarchical residual blocks, namely, residual block 1 through 3, designed to extract local temporal features. Each block embeds three sequential 1-dimensional convolutional layers, followed by BN and ReLU activation. To counteract gradient vanishing and preserve feature integrity across layers, shortcut connections are introduced within each block, enabling effective information flow and deep feature propagation.

LSTMaN In parallel, the LSTMaN module, architecturally consistent with its counterpart in FCNLSTMaN, employs two stacked LSTM-based attention layers to model long-range dependencies. The initial layer focuses on capturing primary temporal correlations, while the subsequent layer hierarchically refines these representations by identifying higher-order interactions. This two-stage attention mechanism enhances the model's capacity to learn nuanced temporal structures, facilitating a richer semantic abstraction of sequential patterns.

B.2 TransformerNet

To further enrich architectural diversity, TransformerNet [60] (Fig. 1(c)) is incorporated as a third teacher model. This architecture begins with an input embedding layer that projects raw sequential data into a high-dimensional latent space, facilitating effective positional encoding. It is followed by a stack of four transformer blocks, transformer block 1 through 4, each comprising a multi-head self-attention mechanism, layer normalization, and a feedforward network. Through these components, each block progressively refines temporal representations by attending to varying time-step interactions, thereby capturing both local dependencies and long-range temporal dynamics. The sequential stacking of such blocks enables hierarchical abstraction, allowing the model to encode complex temporal patterns that are difficult to capture via conventional convolutional or recurrent structures. Within the DHMMD teacher ensemble, this global self-attention view complements the more locally and hierarchically biased representations learned by FCNLSTMaN and ResNetLSTMaN.

B.3 CNN and MLP

In this study, two fundamental student models are utilized: a CNN and an MLP, as illustrated in Fig. 1(b) and Fig. 1(d), respectively. The CNN model is designed with three sequential 1-dimensional convolutional layers: "Conv 8×128", "Conv 5×128", and "Conv 3×128". These layers progressively extract hierarchical features from the input data, with each convolution operation followed by a non-linear activation function, enabling the model to capture intricate spatial patterns. In contrast, the MLP model consists of three fully connected (dense) layers, each responsible for learning abstract representations of the input features. Through the dense connections between layers, the MLP model efficiently transforms input data into high-level features, facilitating accurate predictions for tasks like classification and regression. This teacher–student configuration is specifically aligned with wearable HAR: the teacher models are chosen to model complementary temporal characteristics of inertial activity signals, whereas the student models are chosen to reflect the efficiency requirements of mobile and application-processor-class wearable edge inference.