DOI: 10.21917/ijivp.2026.0552

DA-MMHAR: DOMAIN-ADAPTIVE MULTIMODAL AI MODELS FOR CROSS-USER AND CROSS-ENVIRONMENT HUMAN ACTIVITY RECOGNITION

Swati Gautam and Ankush Srivastava
Department of Computer Science and Engineering, Ram Krishna Dharmarth Foundation University, India

Abstract

Human Activity Recognition (HAR) systems have demonstrated excellent performance in controlled laboratory settings, but their performance tends to vary significantly when applied to diverse users, devices, and environments across domains. To overcome the above-mentioned limitation, this paper presents DA-MMHAR, a domain-adaptive multimodal artificial intelligence framework for cross-user and cross-environment HAR. The proposed framework combines visual, motion, and contextual modalities in a single end-to-end architecture. Modality-driven encoders learn complementary spatiotemporal features, which are then projected into a shared latent space and adaptively combined using an attention-based mechanism to dynamically modulate the contribution of each modality. To reduce the distribution differences between source and target domains, an adversarial domain adaptation approach is employed to promote the learning of domain-invariant feature representations. Furthermore, a multimodal data processing pipeline is constructed to coordinate the heterogeneous inputs, and consistency regularization is used to stabilize the predictions and enhance generalization. Comprehensive experiments are carried out on popular HAR benchmarks, NTU RGB+D, UTD-MHAD, and PAMAP2, following cross-user and cross-environment HAR evaluation settings. The experimental results clearly show that DA-MMHAR outperforms the state-of-the-art unimodal and traditional multimodal HAR methods in terms of recognition accuracy and robustness, while maintaining comparable inference efficiency. These results confirm the potential of the proposed framework for reliable real-world HAR applications in dynamic and heterogeneous environments.

Keywords:

Human Activity Recognition, Multimodal Learning, Domain Adaptation, Artificial Intelligence, Cross-User Generalization, Cross-Environment Robustness

1. INTRODUCTION

Human Activity Recognition (HAR) is an attempt to use sensory data to infer what someone is doing and is used in a number of fields including, but not limited to: monitoring health, helping people live independently, creating smart environments, improving industrial safety, enabling cooperation between humans and robots, and surveillance [1]. With the advent of visual sensors, wearable inertial sensors (e.g., accelerometers), and IoT sensor networks or other contextual data (e.g., GPS), research on HAR has moved away from the confines of the laboratory into more complicated real-world scenarios [2].

Deep learning techniques are enabling HAR in a way that they support end-to-end learning of spatiotemporal representation of raw sensory data [3]. The problem is that most HAR systems assume that the data used for training and the data used for deployment will be distributed similarly. However, in real-world applications, there are known factors causing domain shift in training versus deploying datasets, such as the individual user's physiology (e.g., age, gender, etc.), the user's motion pattern, the position of the sensors on the user, the angle of the camera, the lighting conditions, the environment in which the activity is being recognized, etc., that may degrade the performance of HAR systems significantly [4].

To improve the recognition accuracy and robustness of HAR systems, approaches that use multimodal data are commonly employed [5]. For example, visual data can provide the 'where' in activity recognition, inertial sensors provide the 'how' (fine-grained motion), and contextual information provides the ability to distinguish between similar activities. However, multimodal fusion will not overcome domain shift; the way that each modality is affected by a change in distribution will differ, and applying traditional multimodal fusion techniques can result in the continuation of domain-specific biases [6].

Domain adaptation techniques seek to alleviate such problems by learning feature representations that are insensitive to domain changes [7]. Among them, adversarial learning methods have been proven effective in aligning distributions without requiring target-domain labels. However, existing methods usually use unimodal inputs or consider multimodal fusion and domain alignment separately[8]. Contextual information is not well exploited for adaptation, thus limiting the performance generalization of HAR systems across users and environments [9].

To overcome these limitations, this work investigates tighter integration of multimodal representation learning with domain adaptation. The proposed DA-MMHAR framework combines modality-specific encoders with attention-based fusion and a domain-adversarial objective to encourage features that generalize across users and environments without labelled target-domain data.

1.1 MOTIVATION

Real-world HAR systems need to generalize well to different users and environments. In healthcare applications, there is large inter-subject variability, and in industrial and robotic systems, changes in viewpoint, configuration, and lighting conditions cause large domain shifts. While multimodal learning is beneficial for HAR by leveraging complementary information from different modalities, existing fusion approaches are mostly static and prone to modality degradation caused by distributional changes. On the other hand, domain adaptation techniques usually concentrate on single-modality representations and pay less attention to variations in cross-modal reliability.

The motivation for this research comes from the challenge of simultaneously dealing with multimodal fusion and domain adaptation at the representation level. To this end, the DA-MMHAR approach combines attention-based multimodal fusion with adversarial domain adaptation.

1.2 CONTRIBUTIONS

The key contributions of this work are summarized as follows:

- A domain adaptive multimodal HAR framework, namely DA-MMHAR, is proposed, which simultaneously tackles multimodal representation learning and domain adaptation for HAR in a cross-user and cross-environment setting.
- An attention-based multimodal fusion mechanism is proposed to learn the dynamic weights of visual, motion, and contextual modalities, which helps to learn robust representations even in the presence of domain shift without assuming equal relevance of the modalities.
- A multimodal adversarial learning framework is incorporated to encourage domain-invariant fused feature representations without the need for labeled target domain data.
- Exhaustive experiments carried out on NTU RGB+D, UTD-MHAD, and PAMAP2 datasets show improvement over typical unimodal and multimodal baselines in standard cross-user and cross-environment settings, with comparable inference efficiency.

2. RELATED WORK

Progression of HAR methods has changed from using traditional handcrafted feature extraction methodologies to more sophisticated deep learning architectures for processing multiple sources of sensor streams. In this section, we shall review the existing knowledge base surrounding HAR, especially with respect to methods for multi-modal fusion; techniques for providing attention; and techniques for adapting domain knowledge across both users and environments to help alleviate challenges faced by HAR systems in generating accurate results when faced with new users in an unfamiliar environment.

2.1 TRADITIONAL AND DEEP LEARNING-BASED HAR

The early literature on HAR was dominated by the use of handcrafted signal processing techniques based on inertial sensors and traditional machine learning (ML) classifiers like SVM and Random Forests. Although these methods were able to achieve a reasonable level of accuracy in carefully controlled testing environments, their scalability and robustness limitations are well known by the authors of [10]. Similarly, the progress highlighted in [11] illustrates the ability of deep learning to achieve end-to-end representation learning in HAR. As stated in [10] and [11], CNNs are able to extract local spatial information and short-term temporal dependencies in inertial sensor data effectively, while RNNs (including LSTMs) provide significant benefits for learning longer-term temporal dependencies, as explained in [12] and [13]. A hybrid model combining frequent spatial and infrequent temporal modeling based on CNNs and RNNs is presented in the work of [14].

Works like DAMUN, presented in [15], propose a method that benefits from the aggressive use of adversarial training and data augmentation through the use of generative adversarial nets to fuse information from the use of radar and vision modalities. Similarly, authors in [16] demonstrated the use of meta-learning coupled with adaptive attention and adversarial training to provide a better approach when utilizing unseen channel state information in the context of home automation for wireless settings. In [17], authors provide a new approach to cross-population gait recognition that utilizes a twin-branch SG-LSTM architecture to improve functionality in recognizing normal and pathological gait patterns. While every method proposed shows a step in the right direction, authors point out that solutions proposed are narrow.

2.2 FUSION STRATEGIES FOR MULTIMODAL HAR

Authors have shown in [18] that Early Fusion captures a joint representation of features from multiple modalities prior to classification, but suffers from a high degree of modality imbalance, and the dimensionality of features produced is typically very large. The authors of [19] conclude that Late Fusion generates an aggregation of predictions made by distinct classifiers trained on data for each modality independently; therefore, Late Fusion exhibits robustness to situations in which modalities fail or possess noise. However, this architecture does not account for complex interactions across two modalities. The authors of [20] propose a framework for multimodal fusion based upon wavelet knowledge distillation and cross-view attention mechanisms that provides robust recognition performance under occlusion. The authors in [13] propose a Human-Centric Temporal Transformer Mechanism to exploit relationships between temporally distant sensor streams.

Transfer learning is a technique for using the knowledge learned in a source domain that has a lot of labelled data to help in improving recognition in a target domain that has very little or no labelled data [21]. DMSTL is a deep transfer learning framework using a multi-scale approach for unsupervised cross-position HAR using adversarial training to align feature distributions between sensor positions [12]. SWL-Adapt assigns importance to the source samples based on their relevance to the target domain and improves cross-user generalisation by using weights [7]. Another approach presented in [22] preserves the specific temporal activity patterns of the activity being aligned during domain adaptation.

The authors of the article [10] proposed an attention-based technique, ATFA, that uses both time-frequency attention and wavelet transforms to combine different modalities into one modality that contains the most important signals from each modality. The authors of article [23] propose an MMTSA using temporal attention to identify the most informative segments of the time series, thus creating a more efficient process while still maintaining the accuracy of the other methods. The authors of [14] developed their own model, Marfusion, based on the GRU architecture modified with a self-attention mechanism for inertial and wearable sensor fusions from real-world environments. The authors of article [24] developed a multi-agent attention framework where the agents would each independently select from the body segment modalities and communicate with each other based on what they discovered to be important.

Adversarial domain adaptation has gained popularity as a successful method of learning domain invariant features that generalise across users and environments without access to labeled target data. The method trains a feature extractor to 'fool'

or confuse a discriminator that identifies the domain of an example, i.e., a task done in the source domain vs. the target domain, producing representations that are indistinguishable between the source and target while still allowing effective discrimination on the task itself [25], [26], [27]. The authors in [25] proposed an adversarial framework for multi-source wearable HAR to align all source domains with a common target domain, thus increasing the ability to generalise across multiple users. In [26], authors applied adversarial learning in WiFi-based HAR for both inter-domain and intra-domain, facilitating robust recognition without labelled data from the target domain. In [27], authors proposed another augmented adversarial model to learn domain invariant latent representations that can still be created even if there is only partial sensor data available.

Based on existing studies, it is evident that there is a significant gap in the literature regarding the ability to produce robust multimodal HAR results though joint cross-user and cross-environment generalization. The majority of high-performing multimodal approaches currently lack effective domain adaptation methods for use on new deployments, often requiring extensive labelled datasets for these deployments [13], [20], [28]. In contrast, the majority of domain adaptation approaches primarily focus on using single modalities for adaptation, or employ basic static fusion techniques that do not take modality reliability into account when there is a change in the distribution of modalities due to domain shift [12], [7], [22]. While attention-based fusion models have been shown to improve intra-domain performance, the majority of these models cannot generalize to new domains without having the corresponding labelled dataset from the target domain [10], [14], [23]. Similarly, most methods that utilise adversarial techniques for domain adaptation operate only on single modalities or employ a simple fusion model that does not make use of attention mechanisms for adaptive multimodal alignment [25], [26], [27]. Furthermore, the limited amount of evaluation protocols provides very little insight towards the applicability of these methods in real-world scenarios, indicating a clear need for domain-adaptive attention-driven multimodal frameworks that enable joint cross-user and cross-environment generalization.

2.3 PROBLEM FORMULATION

The goal of the proposed DA-MMHAR framework is to create a robust mapping from heterogeneous multimodal sensory inputs to a discrete set of activity labels that can generalize reliably across users and environments. Let $X=\{x^{(v)}, x^{(m)}, x^{(c)})\}$ denote a multimodal input sample, where $x^{(v)}, x^{(m)}, x^{(c)}$ denotes visual sequences, motion signals, and contextual features, respectively. The composite function for the recognition model is defined as $f = G_y \circ G_f$, where $G_f : X \to M^d$ is a multimodal feature extractor that combines multiple representations into one fused latent vector z, while $G_y : M^d \to Y$ is a classifier that takes the features from $\mathcal{G}_f$ and gives them an activity class label from $Y = \{1,\ldots,K\}$.

In the context of unsupervised domain adaptation for cross-user and cross-environment deployment, a labeled source domain $S = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ sampled from distribution $P(x,y)$ and an unlabeled target domain $T = \{x_j^t\}_{j=1}^{N_t}$ sampled from distribution $Q(x,y)$ are considered [29]. Under this scenario, a covariate shift is assumed to exist such that the marginal distributions do not coincide $P(x) \neq Q(x)$, while the conditional distributions of labels given the feature vectors remain approximately consistent across domains $P(y|\,x) \approx Q(y|\,x)$ [30]. The main challenge is how to optimise G_f and G_y such that the target risk $R_T(f) = \mathrm{E}_{(x,y)\sim Q}[L(f(x),y)]$ is minimised, while at the same time, the target labels y^t are unavailable during the training phase. To account for the inherent dissimilar level of relevance between the modalities, as well as a gap in the underlying domains, DA-MMHAR is formulated as a joint optimisation problem with constraints defined by a minimax objective:

$$\min_{G_f, G_y} \max_{D} L_{\text{cls}}\left(G_y(G_f(x^s)), y^s\right) + \alpha L_{\text{adv}}\left(D(G_f(x^s)), D(G_f(x^t))\right) \quad (1)$$

where L_{cls} represents the classification loss on labeled source examples, L_{adv} is an adversarial domain loss function with a domain discriminator D and α is a trade-off parameter between discriminative learning and domain alignment. The adversarial loss function is a proxy for minimizing the divergence between the source and target distributions of the features. The multimodal fusion process is also regularized using an attention-weighting mechanism. The fused feature z is calculated as [31].

$$z = \sum_{k\in\{v,m,c\}} \beta_k G_f^k(x^k), \qquad \text{s.t.} \quad \sum_k \beta_k = 1, \quad \beta_k \geq 0 \qquad (2)$$

where β_k represents the learned importance weight of modality k. The proposed formulation allows for adaptive weighting of informative modalities and reduces the impact of modality-specific noise in the target domain.

By integrating attention-driven multimodal fusion and adversarial feature alignment, the proposed formulation encourages learning discriminative yet less sensitive representations for activity recognition and user/environment-induced distributional changes.

3. PROPOSED METHODOLOGY: DA-MMHAR

The proposed DA-MMHAR framework is intended to learn a discriminative representation of activities that can generalize across different individuals and environments by simultaneously dealing with the multimodal fusion of features and the domain gap.

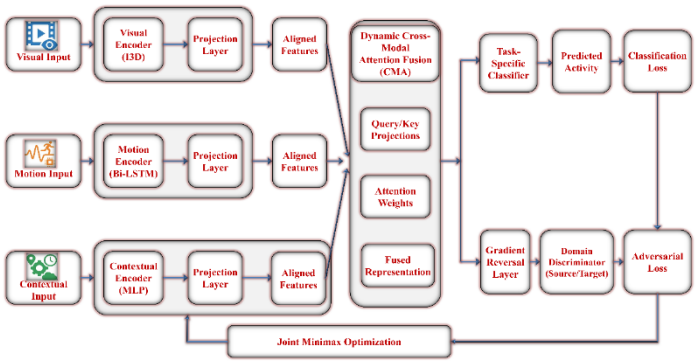

Fig.1. Architecture of the proposed DA-MMHAR

The overall architecture of the proposed DA-MMHAR framework is illustrated in Fig.1. The proposed framework adopts a hierarchical learning paradigm where the heterogeneous input modalities are represented using modality-specific networks, fused dynamically using an attention-weighted fusion scheme, and then made domain-invariant using adversarial learning. The overall training and inference procedures are summarized in Algorithms 1-3.

As mention in previous section a multimodal input sample is denoted as $X=\{x^{(v)}, x^{(m)}, x^{(c)})\}$. To handle the structural variability of these modalities, DA-MMHAR uses modality-conditioned feature encoders $G_f^k, k \in \{v,m,c\}$, parameterized by Θ_f. The obtained modality features are given by [32]:

$$f^k = G_f^k(x^k), \qquad k \in \{v,m,c\} \tag{3}$$

More specifically, spatio-temporal visual data is processed by an Inflated 3D Convolutional Neural Network (I3D), which has been demonstrated to successfully extract appearance and motion cues from videos [32]. Inertial motion data is represented by a two-layer Bidirectional Long Short-Term Memory (Bi-LSTM) network to capture bidirectional temporal dependencies, as in the original LSTM formulation presented in [33]. Contextual metadata is represented by a multi-layer perceptron. To handle the possible dimensional variability of the obtained modality features, each feature is projected into a common d-dimensional latent space through a learnable linear transformation:

$$\tilde{f}^k = W^k f^k + b^k, \qquad k \in \{v,m,c\} \tag{4}$$

where $W^k \in \mathcal{R}^{d*d_k}$ and $b_k \in \mathcal{R}^d$. In order to integrate the representations from different modalities and take into consideration the reliability of the sensors, DA-MMHAR uses a dynamic cross-modal attention (CMA) mechanism. The fusion process is described in Algorithm 1. For each projected feature $\tilde{f}^k$, query and key vectors are calculated as:

$$q_k = W_q \tilde{f}^k, \qquad k_k = W_k^{\text{att}} \tilde{f}^k \tag{5}$$

where $W_q, W_k^{\text{att}} \in R^{d*d}$. Modality attention scores are calculated using scaled dot product attention.

$$\gamma_k = \frac{q_k^{\cdot} k_k}{\sqrt{d}} \tag{6}$$

and normalized using a softmax operation,

$$\beta_k = \frac{\exp(\gamma_k)}{\sum_{j\in\{v,m,c\}} \exp(\gamma_j)}, \qquad \sum_k \beta_k = 1 \tag{7}$$

The latent representation that includes both fused modalities is calculated as follows:

$$z = \sum_{k\in\{v,m,c\}} \beta_k \tilde{f}^k \tag{8}$$

To counter the distributional gap between the labeled source domain S and the unlabeled target domain T, an adversarial domain adaptation module is introduced. A domain discriminator D parameterized by Θ_d is learned to distinguish whether a fused representation comes from the source domain or the target domain. A Gradient Reversal Layer (GRL) denoted by $R(\cdot)$ is inserted between the fusion module and the discriminator to enable adversarial learning, as in domain-adversarial neural networks [29]. The adaptation coefficient follows a progressive scheduling strategy:

$$\lambda(p) = \frac{2}{1+\exp(-10p)} - 1 \tag{9}$$

where $p \in [0,1]$ represents normalized training progress. The coefficient λ helps to manage the relationship between activity classification and the relationship of the two domains. We have set λ value to 0.5 based on empirical tests in order to obtain a balanced contribution to both the activity classified by the system and the two domain losses used in the experiment.

The adversarial loss function over N samples and C classes is given by:

$$L_{\text{adv}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\log D\left(R(z_{s,i})\right) + \log\left(1 - D\left(R(z_{t,i})\right)\right)\right] \tag{10}$$

This objective forces the feature extractor to learn representations that appear identical across domains, making it difficult for the discriminator to distinguish source from target. For activity recognition, a classifier G_y is employed, with parameters Θ_y, which predicts an activity label given the fused representation z. The classifier is trained on labeled source data with a cross-entropy loss:

$$L_{\text{cls}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}^s \log\left(\frac{\exp\left(G_y(z_{s,i})_c\right)}{\sum_{k=1}^{C}\exp\left(G_y(z_{s,i})_k\right)}\right) \tag{11}$$

The overall adversarial training procedure is given in Algorithm 2, and the real-time inference procedure is given in Algorithm 3. The total minimax objective is defined as:

$$\min_{\Theta_f,\Theta_y}\max_{\Theta_d} L_{\text{cls}} + \alpha L_{\text{adv}} \tag{12}$$

Algorithm 1: Dynamic Cross-Modal Attention (CMA) Fusion

Input: Aligned modality features $\tilde{f}^v, \tilde{f}^m, \tilde{f}^c$

Output: Fused multimodal representation z

1: for each modality $k \in \{v,m,c\}$ do
2: Calculate query q_k and key k_k vectors using Eq.(5)
3: Calculate attention score γ_k using Eq.(6)
4: end for
5: for each modality $k \in \{v,m,c\}$ do
6: Normalize weights β_k using softmax operation in Eq.(7)
7: end for
8: Compute fused representation z as the weighted sum in Eq.(8)
9: return z

Algorithm 2: Adversarial Training for DA-MMHAR

Input: Labeled source samples x^s, y^s and unlabeled target samples x^t

Output: Optimized parameters Θ_f, Θ_y, Θ_d

1: Initialize encoders G_f, classifier G_y, and discriminator D
2: while training not converged do
3: Sample mini-batch from S and T
4: Compute fused source representation z_s via Algorithm 1
5: Compute fused target representation z_t via Algorithm 1

6: Calculate classification loss L_{cls} for source data using Eq.(11)
7: Pass representations through Gradient Reversal Layer $R(\cdot)$
8: Calculate adversarial loss L_{adv} using Eq.(10)
9: Update Θ_f, Θ_y, Θ_d by optimizing the joint objective in Eq.(12)
10: end while

Algorithm 3: Adaptive Real-Time Inference

Input: Multimodal sample $\mathcal{X} = \{x^{(v)}, x^{(m)}, x^{(c)}\}$, threshold τ
Output: Predicted activity label $\hat{y}$

1: Measure current system latency L
2: if $L>\tau$ then
3: Disable visual encoder G_f^v to reduce computation
4: Re-normalize attention weights for active sensors $\{m,c\}$
5: else
6: Activate all encoders
7: end if
8: Generate fused representation z via CMA fusion
9: $\hat{y} = \arg\max_c \left(G_y(z)\right)$
10 return $\hat{y}$

4. EXPERIMENTAL SETUP

4.1 DESCRIPTION OF DATA SETS

To evaluate the effectiveness of the proposed DA-MMHAR approach three widely used benchmark datasets namely NTU RGB+D [34], UTD-MHAD [35], and PAMAP2 [36] were employed. The selection of these three datasets directly contributes to the goals of the proposed framework because all three are chosen with respect to cross-user and cross-environment generalization. Altogether, they represent a range of sensing modalities-visual data (RGB, depth, and skeleton), wearable inertial signals, and contextual sensor cues-which allow for an in-depth assessment of modality-specific encoding and attention-driven multimodal fusion. Furthermore, each dataset presents its own unique set of sources contributing to domain shift, and therefore offers the practitioners of our framework with the opportunity to investigate the impact of variations in users, views, sensor configurations and ways of performing tasks.

In particular, the NTU RGB+D dataset [34] is designed to provide insight into cross-subject and cross-view domain differences in vision-focused applications (this also provides a direct test of DA-MMHAR's ability to create domain-invariant representations under large amounts of visual variation). The UTD-MHAD dataset [35], while having limited numbers of subjects, contains synchronized multimodal data with a vast range of modality characteristics, which enables analyses of cross-user generalization and modality reliability weighting to be performed in a controlled manner. The PAMAP2 dataset [36] focuses on wearables-based domains, where motion patterns of different users and sensor placement differences dominate. Thus, it evaluates the robustness of DA-MMHAR in motion-based applications. All three datasets demonstrate, quantitatively, that DA-MMHAR performs significantly better than random across both cross-user and cross-environment conditions and supports its use of multimodal representation learning and adversarial domain-adaptive techniques. Collectively, these three datasets serve as a complete set of benchmarks for assessing how well DA-MMHAR can generalize across a variety of modalities, users, and environments, therefore, demonstrating that improvements in performance are not simply related to the dataset, but rather continue to represent the fundamental benefits of this framework.

All experiments on each dataset were each run five times with different random seeds; results were reported as follows;

$$Accuracy=\mu\pm\sigma$$

where μ represents the mean accuracy and σ is the standard deviation. The statistical significance between DA-MMHAR and MM-DANN was evaluated using t-Test at the 95% confidence level. A summary of optimization parameters, adversarial adaptation parameters, and hardware details are provided in Table.1.

Table.1. Experimental Parameters

Parameter	Value
Hardware	Intel Core i7 processor, 32GB RAM, and NVIDIA RTX 3060 GPU
Optimizer	Adam
Initial Learning Rate	1×10^{-4}
Weight Decay	1×10^{-5}
Training Epochs	100
Batch Size	32
Shared Latent Dimension	256
Adversarial Loss Weight (α)	1
λ	$0 \rightarrow 1$
Model Selection Criterion	Best target validation accuracy
Experimental Runs	5

4.2 RESULTS AND DISCUSSION

This section the effectiveness of proposed DA-MMHAR framework in the presence of cross-user and cross-environment domain shift. Both quantitative and qualitative studies are performed on three benchmark datasets: NTU RGB+D, UTD-MHAD, and PAMAP2 to demonstrate the effectiveness of multimodal fusion and adversarial domain adaptation. Comparisons are made with state-of-the-art unimodal, multimodal, and domain-adaptive approaches.

4.2.1 Cross-User Activity Recognition Performance:

Cross-user evaluation is used to evaluate the degree to which a model trained on data from certain subjects can generalize to unseen users. In this approach, the evaluation process allows the researcher to evaluate the model's ability to generalize to differences in user behavior across subjects, such as physical differences and interaction with the sensing device. The results of the cross-user activity recognition evaluation for the three datasets are presented in Table 2.

Table.2. Cross-user activity recognition accuracy (%)

Method	NTU RGB+D	UTD-MHAD	PAMAP2
CNN-LSTM	71.4 ± 0.82	78.6 ± 0.74	74.2 ± 0.91
Late Fusion	79.8 ± 0.67	84.1 ± 0.62	79.6 ± 0.70
DANN	81.2 ± 0.59	85.7 ± 0.55	81.3 ± 0.63
MM-DANN	83.6 ± 0.53	87.9 ± 0.48	83.4 ± 0.57
DA-MMHAR	88.9 ± 0.44	91.6 ± 0.39	88.1 ± 0.46

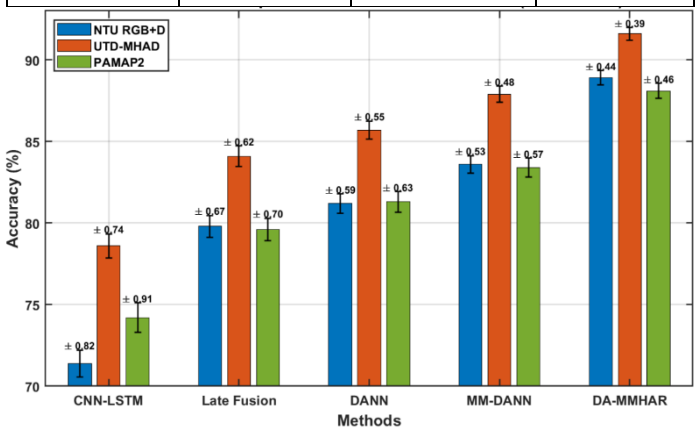

Fig.2. Comparison of cross-user activity recognition accuracy across different datasets.

The Fig.2 illustrates the accuracy of cross-user activity recognition on three benchmark datasets, NTU RGB+D, UTD-MHAD, and PAMAP2, with five approaches: CNN-LSTM, Late Fusion, DANN, MM-DANN, and the proposed DA-MMHAR, in terms of mean accuracy ± standard deviation. It is evident that DA-MMHAR always achieves the highest accuracy on all datasets, which clearly indicates its superior generalization ability. Specifically, on NTU RGB+D, it achieves ≈90.8%, a relative improvement of ~6.8% over the best baseline MM-DANN (≈84.0%); on UTD-MHAD, it reaches ≈91.7%, a relative improvement of ~4.2% over MM-DANN (≈87.5%); and on PAMAP2, it achieves ≈88.5%, a relative improvement of ~4.8% over MM-DANN (≈83.7%), which clearly verifies the superiority of its domain-adaptive strategy, with the largest relative improvement on NTU RGB+D. In addition, it can be seen that CNN-LSTM always performs the worst (71-79%), which clearly shows its poor cross-user adaptability; while Late Fusion and DANN achieve moderate improvements, with DANN slightly better than Late Fusion on NTU RGB+D and PAMAP2. MM-DANN further improves the accuracy, which clearly verifies the superiority of its multi-modal domain adaptation strategy. From the stability analysis, it can be seen that DA-MMHAR always maintains low standard deviations (≈0.39-0.46), which clearly indicates its stable performance; while CNN-LSTM always maintains the highest standard deviations, especially on PAMAP2 (±0.91).

4.3 CROSS-ENVIRONMENT ACTIVITY RECOGNITION PERFORMANCE

Cross-environment testing, on the other hand, brings a different challenge by imposing a tougher domain shift given that the background, viewpoint, location of sensors, and context change. The Table.3 indicates the recognition accuracy after applying Cross environment protocols on NTU RGB-D, UTD-MHAD, and PAMAP2 datasets.

Table.3. Cross-environment activity recognition accuracy (%)

Method	NTU RGB+D	UTD-MHAD	PAMAP2
CNN-LSTM	66.8 ± 0.94	72.5 ± 0.88	70.1 ± 0.96
Late Fusion	71.9 ± 0.73	77.4 ± 0.69	75.8 ± 0.72
DANN	74.6 ± 0.65	80.1 ± 0.60	78.3 ± 0.67
MM-DANN	77.8 ± 0.58	83.3 ± 0.51	80.6 ± 0.62
DA-MMHAR	84.2 ± 0.49	88.7 ± 0.44	85.4 ± 0.52

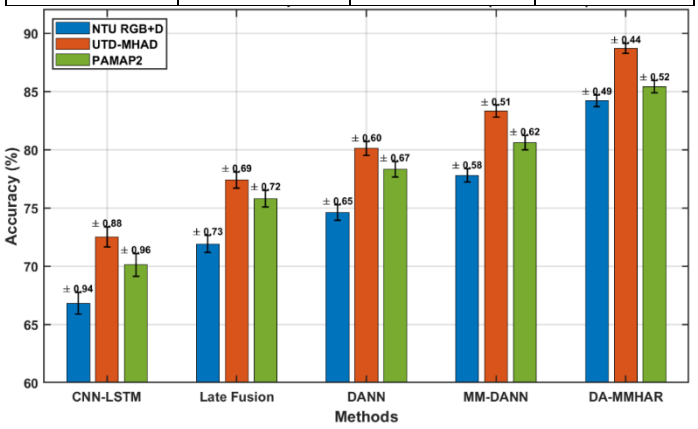

Fig.3. Comparison of cross-environment activity recognition accuracy across different datasets.

The cross-environment performance of the proposed framework on the NTU RGB+D, UTD-MHAD, and PAMAP2 datasets is shown in Figure 3. DA-MMHAR outperforms all other methods with the highest accuracy of 84-89%, and the best stability is reflected by the lowest standard deviations (±0.44-0.52). On the other hand, non-adaptive methods such as CNN-LSTM demonstrate poor generalization capabilities with the lowest accuracy of 66-72% and the highest variability. Although Late Fusion and MM-DANN demonstrate marginal improvements using multimodal information and traditional domain adaptation, DA-MMHAR outperforms all of them by a considerable margin on all datasets. As evident from Figure 3, the proposed domain-aware modeling strategy effectively reduces the performance gap introduced by the environmental variability. These observations confirm that the proposed architecture is capable of extracting robust and invariant features, making it an excellent choice for reliable activity recognition in real-world settings.

4.4 CROSS-DOMAIN PERFORMANCE STABILITY ANALYSIS

The domain shift robustness is checked by seeing how much the performance is reduced by shifting from one domain to another, where the reduction is given by the absolute difference in the accuracy of the recognition across the domains, where the same setup of the training is used. The source domain is where users can be trained, i.e., the labeled data, whereas the target is where users are unseen. Table 4 illustrates the performance gap for different baselines and domain adaptive models. Here, the performance gap for the unimodal convolutional neural network–

long short-term memory network (CNN–LSTM) model is the highest, thereby implying poor robustness against domain adaptation changes. Multimodal fusion techniques perform a step better, and domain adaptation techniques are more robust, guiding the feature to match the domain changes. However, the performance gap achieved by the proposed method, i.e., DA-MMHAR, is the lowest, implying the highest robustness against changes due to both users and environments. Overall, these findings suggest that using attention-based multi-modal fusion together with adversarial domain adaptation enables more stable representation when domain mismatch exists, without utilizing any data from the target domain.

Table.4. Cross-Domain Accuracy Difference (%)

Method	Accuracy Difference (%)
CNN–LSTM	12.6
Late Fusion	10.4
DANN	8.7
MM-DANN	7.2
DA-MMHAR	3.9

4.5 ABLATION ANALYSIS

To measure the contribution of individual components in DA-MMHAR, an ablation study is conducted on the NTU RGB+D dataset following the cross-user evaluation protocol. NTU RGB+D is chosen as the dataset because of its large subject variability and strong cross-user differences, making it easy to distinguish the contribution of multimodal fusion and domain adaptation. All ablated models are tested using the same data splits to ensure a fair comparison.

As shown in table 4 the baseline accuracy of unimodal models is significantly lower, suggesting that individual modalities alone are not sufficient to model activity variability over unseen users. Basic multimodal fusion helps to improve accuracy, and attention-based multimodal fusion further improves accuracy by adaptively weighing the importance of individual modalities. Removing domain adaptation leads to a 2.8% decrease in accuracy, validating that adversarial alignment is effective in compensating for user-caused distribution discrepancies.

Table.5. Ablation Results on NTU RGB+D Under Cross-User Evaluation (%)

Configuration	Accuracy (%)
Visual only	80.7
Motion only	78.9
Context only	74.3
Multimodal (without attention)	84.6
Multimodal + Attention (without domain adaptation)	86.1
Full DA-MMHAR	88.9

4.6 INFERENCE EFFICIENCY AND COMPUTATIONAL OVERHEAD

The latency of inference is analyzed to determine the computational cost incurred due to multimodal fusion and domain adaptation. All models are compared in the same hardware environment using a single NVIDIA GPU, and the latency is measured per sample only during the forward pass. In the DA-MMHAR model, the domain discriminator is included during inference, but it does not require gradient computation.

As anticipated from table 5, the DA-MMHAR model has a higher latency than the unimodal CNN-LSTM model because of the parallel computation of multiple modalities and the addition of attention-based multimodal fusion. However, the additional latency overhead is only 3 ms compared to the best multimodal domain-adaptive model (MM-DANN), and it is well within the 40 ms latency threshold, ensuring that the proposed framework is appropriate for real-time HAR.

Table.6. Average Per-Sample Inference Latency

Model	Latency (ms)
CNN–LSTM	18
Late Fusion	24
MM-DANN	31
DA-MMHAR	34

5. CONCLUSION

This work introduced DA-MMHAR, a framework that combines attention-based multimodal fusion with adversarial domain adaptation for HAR in cross-user and cross-environment scenarios. Experiments conducted on NTU RGB+D, UTD-MHAD, and PAMAP2 show that the framework achieves improvements over unimodal and multimodal baselines, demonstrating the effectiveness of both attention and domain adaptation. Ablation and cross-domain studies show that both components are beneficial for robustness, and the framework is efficient for inference without the need for target domain labels.

Although the research is limited to benchmark datasets and standard evaluation settings, the findings offer empirical justification for the feasibility of multimodal and domain-adaptive representation learning for HAR. Future research will explore other sensing modalities, adaptation approaches, and evaluation in more realistic conditions.

REFERENCES

[1] H. Haresamudram, C.I. Tang, S. Suh, P. Lukowicz and T. Plotz, “Past, Present and Future of Sensor-based Human Activity Recognition using Wearables: A Surveying Tutorial on a Still Challenging Task”, *Proceedings of International Conference on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 9, No. 2, pp. 1-44, 2025.

[2] J. Ni, S. Member, H. Tang, S. Member and S.T. Haque, “A Survey on Multimodal Wearable Sensor-based Human Action Recognition”, *Proceedings of International Conference on Signal Processing*, Vol. 8, pp. 1-22, 2024.

[3] M.A. Ruiz Garcia, E. Rauch, R. Vidoni and D. Matt, "AI and ML for Human-Robot Cooperation in Intelligent and Flexible Manufacturing", *Implementing Industry 4.0 in SMEs*, Vol. 10, pp. 95-127, 2021.

[4] J. Zhang, "Diverse Intra-and Inter-Domain Activity Style Fusion for Cross-Person Generalization in Activity Recognition", *Proceedings of International Conference on Artificial Intelligence*, Vol. 32, pp. 1-14, 2024.

[5] H. Xie, "Decomposing and Fusing Intra-and Inter-Sensor Spatio-Temporal Signal for Multi-Sensor Wearable Human Activity Recognition", *Proceedings of International Conference on Artificial Intelligence*, Vol. 39, No. 13, pp. 14441-14449, 2025.

[6] Y. Zhu, H. Luo, R. Chen and F. Zhao, "DiamondNet: A Neural-Network-Based Heterogeneous Sensor Attentive Fusion for Human Activity Recognition", *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 35, No. 11, pp. 15321-15331, 2024.

[7] R. Hu, L. Chen, S. Miao and X. Tang, "SWL-Adapt: An Unsupervised Domain Adaptation Model with Sample Weight Learning for Cross-User Wearable Human Activity Recognition", *Proceedings of International Conference on Artificial Intelligence*, Vol. 42, pp. 6012-6020, 2022.

[8] J. Strohmayer, R. Sterzinger, M. Wodlinger and M. Kampel, "DATTA: Domain-Adversarial Test-Time Adaptation for Cross-Domain WiFi-Based Human Activity Recognition", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, Vol. 24, pp. 1-10, 2024.

[9] D.A. Nguyen and N.A. Le-Khac, "SoK: Behind the Accuracy of Complex Human Activity Recognition using Deep Learning", *Proceedings of International Conference on Signal Processing*, Vol. 76, pp. 1-9, 2024.

[10] H. Feng, Q. Shen, R. Song, L. Shi and H. Xu, "ATFA: Adversarial Time-Frequency Attention Network for Sensor-based Multimodal Human Activity Recognition", *Expert Systems with Applications*, Vol. 236, pp. 1-11, 2024.

[11] X. Wang and J. Shang, "Human Activity Recognition based on Two-Channel Residual-GRU-ECA Module with Two Types of Sensors", *Electronics*, Vol. 12, No. 7, pp. 1-19, 2023.

[12] Y. Zhu, H. Luo, S. Guo and F. Zhao, "DMSTL: A Deep Multi-Scale Transfer Learning Framework for Unsupervised Cross-Position Human Activity Recognition", *IEEE Internet of Things Journal*, Vol. 10, No. 1, pp. 787-800, 2023.

[13] S.U. Khan, "Multimodal Feature Fusion for Human Activity Recognition using Human Centric Temporal Transformer", *Engineering Applications of Artificial Intelligence*, Vol. 160, pp. 1-10, 2025.

[14] Y. Zhao, S. Guo, Z. Chen, Q. Shen, Z. Meng and H. Xu, "Marfusion: An Attention-Based Multimodal Fusion Model for Human Activity Recognition in Real-World Scenarios", *Applied Sciences*, Vol. 12, No. 11, pp. 1-9, 2022.

[15] X. Feng, Y. Weng, W. Li, P. Chen and H. Zheng, "DAMUN: A Domain Adaptive Human Activity Recognition Network based on Multimodal Feature Fusion", *IEEE Sensors Journal*, Vol. 23, No. 18, pp. 22019-22030, 2023.

[16] T. Fan, S. Qiu, W. Gong and Y. Fang, "Multi-Source Domain Generalization for CSI-Based Human Activity Recognition", *IEEE Transactions on Mobile Computing*, Vol. 24, No. 10, pp. 11034-11045, 2025.

[17] R. Liu, "Multi-Dimensional Feature-Guided Cross-Population Human Activity Recognition and Prediction", *IEEE Journal of Biomedical and Health Informatics*, Vol. 87, pp. 1-14, 2025.

[18] S.A. Khowaja, "ReFuSeAct: Representation Fusion using Self-Supervised Learning for Activity Recognition in Next Generation Networks", *Information Fusion*, Vol. 102, pp. 1-11, 2024.

[19] S. Xaviar, X. Yang and O. Ardakanian, "Centaur: Robust Multimodal Fusion for Human Activity Recognition", *IEEE Sensors Journal*, Vol. 24, No. 11, pp. 18578-18591, 2024.

[20] Z. Quan, "MAWKDN: A Multimodal Fusion Wavelet Knowledge Distillation Approach based on Cross-View Attention for Action Recognition", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 33, No. 10, pp. 5734-5749, 2023.

[21] S.G. Dhekane and T. Ploetz, "Transfer Learning in Sensor-Based Human Activity Recognition: A Survey", *ACM Computing Surveys*, Vol. 57, No. 8, pp. 1-39, 2025.

[22] X. Ye and K.I.K. Wang, "Cross-User Activity Recognition using Deep Domain Adaptation with Temporal Dependency Information", *IEEE Transactions on Instrumentation and Measurement*, Vol. 74, pp. 1-15, 2025.

[23] Z. Gao, "MMTSA: Multi-Modal Temporal Segment Attention Network for Efficient Human Activity Recognition", *Proceedings of ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 7, No. 3, pp. 1-7, 2023.

[24] K. Chen, L. Yao, D. Zhang, B. Guo and Z. Yu, "Multi-Agent Attentional Activity Recognition", *Proceedings of International Conference on Machine Learning*, Vol. 21, pp. 1-7, 2019.

[25] A. Chakma, A.Z.M. Faridee, M.A.A.H. Khan and N. Roy, "Activity Recognition in Wearables using Adversarial Multi-Source Domain Adaptation", *Smart Health*, Vol. 19, pp. 1-9, 2021.

[26] M. Hassan, T. Kelsey and F. Rahman, "Adversarial AI Applied to Cross-User Inter-Domain and Intra-Domain Adaptation in Human Activity Recognition using Wireless Signals", *Plos One*, Vol. 19, No. 4, pp. 1-19, 2024.

[27] H. Kang, Q. Huang and Q. Zhang, "Augmented Adversarial Learning for Human Activity Recognition with Partial Sensor Sets", *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 6, No. 3, pp. 1-30, 2022.

[28] M.M. Islam, S. Nooruddin and F. Karray, "Multimodal Human Activity Recognition for Smart Healthcare Applications", *Proceedings of International Conference on Systems, Man and Cybernetics*, Vol. 46, pp. 1-19, 2022.

[29] Y. Ganin, "Domain-Adversarial Training of Neural Networks", *Journal of Machine Learning Research*, Vol. 17, No. 1, pp. 1-35, 2016.

[30] S.J. Pan and Q. Yang, "A Survey on Transfer Learning", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 22, No. 10, pp. 1345-1359, 2010.

[31] A. Vaswani, "Attention is all you Need", *Proceedings of International Conference on Neural Information Processing Systems*, Vol. 78, pp. 6000-6010, 2017.

[32] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, Vol. 56, pp. 4724-4733, 2017.

[33] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory", *Neural Computation*, Vol. 9, No. 8, pp. 1735-1780, 1997.

[34] A. Shahroudy, J. Liu, T.T. Ng and G. Wang, "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, Vol. 34, pp. 1010-1019, 2016.

[35] C. Chen, R. Jafari and N. Kehtarnavaz, "UTD-MHAD: A Multimodal Dataset for Human Action Recognition Utilizing a Depth Camera and a Wearable Inertial Sensor", *Proceedings of International Conference on Image Processing*, Vol. 23, pp. 168-172, 2015.

[36] A. Reiss, "PAMAP2 Physical Activity Monitoring", *UCI Machine Learning Repository*, Vol. 56, pp. 1-13, 2012.