Generalizable Sensor-Based Activity Recognition via Categorical Concept Invariant Learning

Di Xiong[1]*, Shuoyuan Wang[2]*, Lei Zhang[1]†, Wenbo Huang[3], Chaolei Han[3]

[1]Nanjing Normal University, Nanjing 210023, Jiangsu, China
[2]Southern University of Science and Technology, Shenzhen 518055, Guangdong, China
[3]Southeast University, Nanjing 211189, Jiangsu, China
{221812013, leizhang}@njnu.edu.cn, claytonwang0205@gmail.com, {wenbohuang1002, chaoleihan}@seu.edu.cn

Abstract

Human Activity Recognition (HAR) aims to recognize activities by training models on massive sensor data. In real-world deployment, a crucial aspect of HAR that has been largely overlooked is that the test sets may have different distributions from training sets due to inter-subject variability including age, gender, behavioral habits, etc., which leads to poor generalization performance. One promising solution is to learn domain-invariant representations to enable a model to generalize on an unseen distribution. However, most existing methods only consider the feature-invariance of the penultimate layer for domain-invariant learning, which leads to suboptimal results. In this paper, we propose a Categorical Concept Invariant Learning (CCIL) framework for generalizable activity recognition, which introduces a concept matrix to regularize the model in the training stage by simultaneously concentrating on feature-invariance and logit-invariance. Our key idea is that the concept matrix for samples belonging to the same activity category should be similar. Extensive experiments on four public HAR benchmarks demonstrate that our CCIL substantially outperforms the state-of-the-art approaches under cross-person, cross-dataset, cross-position, and one-person-to-another settings.

Introduction

Sensor-based human activity recognition (HAR) aims to train models using massive data collected from wearable sensors such as accelerometers and magnetometers. HAR has wide applications in many areas, including personal fitness, elderly-care, human-machine interaction, sports tracking, etc (Huang et al. 2022). Despite significant progress, current HAR study still faces critical challenges that prevent practical deployment while rendering the performance suboptimal on never-seen-before data (Qian, Pan, and Miao 2021). As shown in Figure 1, the distribution of sensor signals is typically influenced by various factors such as age, gender, and deployment locations. For instance, due to inter-subject variability, a model that recognizes the activities of an adult does not generalize well on new unseen data from an elderly person, because they may have different behavioral patterns that make their data distributions highly diverse. Therefore, simply generalizing a model trained on existing data to new unseen data may not work due to such distribution shift problem in sensor signals.

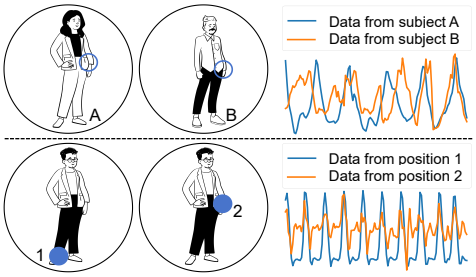


Figure 1: Domain shift: sensor readings collected from different subjects or different locations of the same subject.

In practice, the real-world sensor samples are typically restricted to access during training. Taking elderly fall detection as an example, it is rather unrealistic to aggregate training data from the elderly people based on safety concerns. However, it is feasible to collect training data from young subjects while ensuring enough safe conditions. We have to expect a model trained on the data collected from young subjects to be readily extensible to elderly users with no need of training data collected from them. To mitigate this issue, DG has been a popular technique to reduce the distribution discrepancy between the source and target domains with no need of direct access to never-seen-before target data during training. Though many research efforts have been devoted to computer vision applications, they may be incompatible with time series data. Until now, there has been limited research attention targeted at wearable sensor data including feature disentanglement (Qian, Pan, and Miao 2021), data augmentation (Zhang et al. 2018), gradient operation (Huang et al. 2020), and domain-invariant representation learning (Lu et al. 2024; Du et al. 2021), which focus on explicitly or implicitly regularizing the models based on the analysis of features. Despite notable achievements in DG, it still remains a major challenge that is far from being solved on sensor data. A recent study (Gulrajani and Lopez-Paz 2021) have empirically shown that most current state-of-the-art approaches are even inferior to the baseline empirical risk minimization (ERM) algorithm. These findings clearly

*These authors contributed equally.
†Corresponding author.

highlight the necessity of innovative and effective models that can ensure robustness across domains.

Following this cue, as well as the uniqueness of each person's activity characteristics, this paper takes a different perspective toward obtaining robust outputs through emphasizing the logit-invariance of the classifier weights in a deep learning model, instead of only concentrating on the feature-invariance. As we know, most existing models typically calculate the final output logits through multiplying the classifier weights with the penultimate layer's feature, where every product term can be viewed as a contribution to the corresponding logit. While organizing all these contributions for logits from all activity classes as a matrix (also named concept matrix) based on input sensor samples, we conjecture that the matrices induced by samples belonging to the same activity class should be similar for a well-generalized model. Based on this intuition, we introduce a new regularization loss term, which aims to enforce similarity between the concept matrix of samples belonging to the same activity category and their corresponding mean value. A dynamic momentum update strategy is used to update the category-wise mean concept matrix during each training iteration. On one hand, different from most existing feature-based regularization (Cha et al. 2022), our approach also takes into full consideration the effect of the classifier weights, avoiding biased estimation for feature importance. On the other hand, to overcome the drawback of logit-based regularization that only has a coarse value, our approach provides a fine-grained characterization of cross-domain activity recognition by considering the varying influence of every contribution to final classification results. In summary, the main contributions of this paper are three-fold:

- **New perspective:** In this paper, we propose a Categorical Concept Invariant Learning framework named CCIL, which is mainly built on the category-wise mean concept matrix. We provide new insights from both logit-invariance and feature-invariance perspectives to explain the rationale behind our generalizable activity recognition algorithm.
- **Simple algorithm:** A new regularization term is introduced to enforce similarity between the concept matrix of samples from the same activity class and their mean value, while a dynamic momentum update strategy is used to update the matrix during each training iteration. Our approach is simple, which adds only a few lines of code upon the standard ERM baseline.
- **Superior performance:** Comprehensive experiments on four public sensor-based HAR datasets demonstrate that our proposed CCIL consistently beat the state-of-the-art baselines while evaluated under the rigorous cross-domain settings. These results highlight the effectiveness and universality of our concept matrix invariance regularization, despite its simplicity.

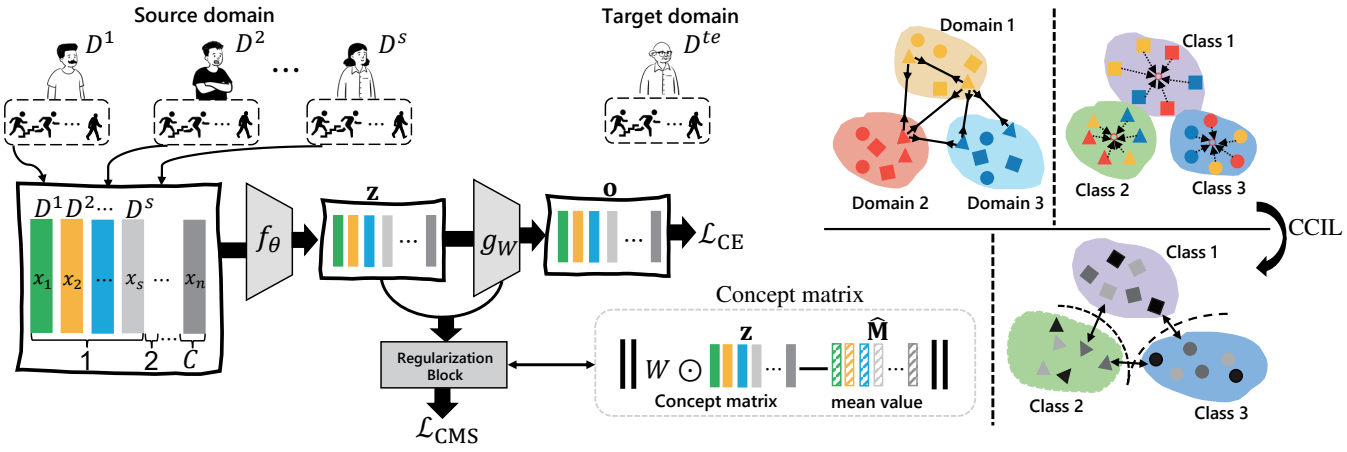
Related Work

Human Activity Recognition

Human activity recognition (HAR) mainly attempts to recognize activities of daily living that are performed by different persons. Based on data type, it can be roughly grouped into two categories: vision-based HAR and sensor-based HAR (Dang et al. 2020): The former collects activity data by cameras or other optical devices, which would often encounter severe privacy leaking problems (Sun et al. 2022). For example, sensitive personal data like facial information will be accidentally released on cameras. Moreover, cameras may not work in HAR when a person is beyond their coverage range (Kong and Fu 2022); The latter collects activity data through ambient sensors deployed in smart environment or wearable sensors attached to different body parts (Gu et al. 2021). Due to small size and low price, there has been the wide popularity of inertial sensors embedded in wearable devices like smart phones and watches, that makes them convenient and practical to record activity data for offering smart user services (Chen et al. 2021). In contrast to video-based action recognition, there is relative limited research attention on HAR using wearable sensor data. Thus, this paper mainly concentrates on wearable sensor-based HAR problem. To resolve sensor-based HAR, deep learning models have recently been widely applied to automatically extract features from raw sensor signals for activity recognition (Qian, Pan, and Miao 2021; Hammerla, Halloran, and Plötz 2016; Qian et al. 2019; Wang et al. 2024). Despite remarkable progress, these activity recognition models are typically trained based on the assumption that the training and testing data have independently identical distributions, ignoring the fact the sensor data collected from different persons may follow different distributions due to their unique characteristics in body shapes, behavior patterns, or other biological factors.

Domain-Invariant Learning for Generalization

To address distribution-shift problem, domain adaptation (DA) has offered a popular solution to bridge domain gaps (Kouw and Loog 2019; Wang et al. 2018; Yu and Lin 2023). However, DA has a notable limitation in that it more or less requires to access target domain data during training, which renders DA infeasible in many real-world HAR situations (Qian, Pan, and Miao 2021). Moreover, while there are multiple target domains, the model has to be re-trained on every target domain (Lu et al. 2021), which is time-consuming and inefficient. DG has been recently proposed to handle such challenging situations (Wang et al. 2018). It aims to learn a robust and generalizable model from one or more different but related source domains, that can perform well on the never-seen-before target domains. In computer vision community, existing DG-related literatures may be coarsely categorized into three research streams (Zhou et al. 2022; Wang et al. 2022): Learning strategy (Huang et al. 2020; Sagawa et al. 2019), Data manipulation (Zhang et al. 2018; Zhou et al. 2021), Representation learning (Parascandolo et al. 2020; Du et al. 2021; Qian, Pan, and Miao 2021; Lu et al. 2024; Chen et al. 2023; Sun and Saenko 2016). Though the past five years have witnessed its remarkable success in the computer vision community, there has been very limited research attention for human activity recognition using wearable sensor data. To the best of our knowledge, the latest work to solve such DG problem for HAR is DIVERSITY



(a) CCIL Architecture

(b) Categorical Concept Invariance

Figure 2: (a) An overview of our CCIL framework based on the concept matrix. (b) CCIL learns domain-invariant representation by mapping the latent representation of the same-class samples close together.

(Lu et al. 2024), which leverages the latent distributions to minimize the distribution divergence for time series out-of-distribution detection and generalization. There now exists very few DG-related works that explore the effect of classifier weights on series time data for cross-domain activity recognition. Different from prior arts, our work is the first to simultaneously concentrate on the feature and logit invariance regularizations, highlighting the effectiveness of the concept matrix.

Methodology

Problem Formulation

Given a set of source domains utilized as the training dataset $\mathcal{D}^{tr} = \{\mathcal{D}^1, \mathcal{D}^2, ..., \mathcal{D}^S\}$, for the i-th source domain $\mathcal{D}^i$ we use $P^i(\boldsymbol{x}, y)$ on $\mathcal{X} \times \mathcal{Y}$ to represent the joint distribution, where $\boldsymbol{x} \in \mathcal{X}$ denotes the sensor input obtained by sliding window [1], $y \in \mathcal{Y} = \{1, ..., C\}$ denotes the label space with total C activity categories. We perform training from the source domain $\mathcal{D}^{tr}$ dataset to obtain the cross-domain activity recognition model $h : \mathcal{X} \to \mathcal{Y}$, which can generalize well on the never-seen-before target domain $\mathcal{D}^{te}$ dataset. It is important to note that the target domain $\mathcal{D}^{te}$ can only be accessed at inference. Although the source and target domains have the same input and output spaces, the source and target domains have different distributions P^{te}, i.e., $P^i(\boldsymbol{x}, y) \neq P^j(\boldsymbol{x}, y), \forall i, j \in \{1, 2, ..., S, te\}$. In a word, we hope that the model h trained on the source domain D^{tr} can minimize the average prediction error ϵ_t on the target domain D^{te}:

$$\epsilon_t = \mathbb{E}_{(\boldsymbol{x},y)\sim P^{te}(\boldsymbol{x},y)} \mathcal{L}(h(\mathbf{x}), y). \quad (1)$$

Motivation Based on Feature v.s. Logit

In this section, we answer the potential motivation behind the proposed CCIL and provide in-depth insights into our algorithm design. To learn domain-invariant representations for sensor-based activity recognition, a well-generalized model should be stable under cross-domain scenarios. Most existing DG-based works concentrate on implicitly or explicitly regularizing the model based on the notation of feature-invariance (Lu et al. 2024; Du et al. 2021; Cha et al. 2022). However, such feature-invariance leaning strategy still poses a serious limitation. To be specific, while only focusing on the feature-invariance, it fails to take into full consideration the classifier weights, which are in charge of determining the importance of different feature elements, thereby resulting in a biased estimate for feature importance. For example, while a feature element has a big value, it might correspond to a small value in the classifier, leading to a lower effect on final activity classification results. Solely considering the feature-invariance would be biased or misleading, thus undermining the generalization ability of the model. Therefore, the influence of the classifier weights should be fully considered so as to avoid such biased estimation of feature importance.

To mitigate this issue in cross-domain situation, instead of only concentrating on the feature-invariance, the logit may implicitly take into account the effect of classifier weights to a certain extent. However, the logit is only able to provide a coarse value, that lacks a fine-grained perspective to interpret the rationale behind generalizable cross-domain activity recognition process. As a consequence, focusing only on logit-invariance may lead to ineffectiveness in generating robust feature representations, which will be verified in later visualizing analysis. Our concept matrix aims to overcome the two drawbacks by simultaneously concentrating on both feature-invariance or logit-invariance, which learn domain-invariant representations from a more fine-grained perspective while taking into account the influence of classifier weights. A new regularization loss term is introduced based on the concept matrix to capture both feature-invariance and logit-invariance representations for generalizable cross-domain activity recognition.

[1] Due to its implementational simplicity, a popuar sliding window strategy is used to divide continous time series data into fixed-size windows as sensor inputs here (Ordóñez and Roggen 2016).

Framework Overview

This section presents a comprehensive description of our newly proposed DG approach. As illustrated in Figure 2, we propose Categorical Concept Invariant Learning abbreviated as CCIL to learn both feature-invariance and logit-invariance for generalizable cross-domain activity recognition. CCIL takes inputs from multiple different but related source domains for model training, while the target domain data is only utilized for model test. Subsequently, after going through a common feature extractor, the output features are multiplied with the classifier weights, which can then be used to form the concept matrix. While the same activity performed by different persons (domains) tend to have similar activity semantics, we may leverage the invariance across domains to regularize the model to facilitate domain generalization. To ensure robust results, the concept matrix of samples belonging to the same activity class should align with their corresponding mean value, which is very reasonable such the causal factors for invariance-learning are usually stable patterns to persist across domains (Lu et al. 2024; Chen et al. 2023). On this basis, we introduce a new regularization loss term that allows the model to explore more invariance.

Concept Matrix In most existing HAR models, the backbone architecture h is typically comprised of a feature extractor and an activity classifier. The feature $\mathbf{z} \in \mathbb{R}^D$ can be produced through a feature extractor f (i.e., $\mathbf{z} = f(x)$) parameterized by θ, which contains two convolution layers and one pooling layer (Lu et al. 2024). Assuming that there are total C activity classes in $\mathcal{Y}$ after applying the classifier g comprised of one fully-connected layer on $\mathbf{z}$, we can obtain the final logits $\mathbf{o} = \mathbf{W}^\top \mathbf{z} \in \mathbb{R}^C$ (i.e., $\mathbf{o} = g(\mathbf{z})$), where $\mathbf{W} \in \mathbb{R}^{D \times C}$ is the weights of the classifier g. For simplicity, we omit the bias in the classifier. On this basis, the concept matrix can be constructed based on the output feature $\mathbf{z}$ and classifier weights $\mathbf{W}$. In practice, every logit value is equivalent to the summation of element-wise multiplications between the feature elements and the corresponding weights in the classifier. Without loss of generality, o_c (i.e., the c-th dimension of $\mathbf{o}$) can be formulated as follows:

$$o_c = \mathbf{W}_{\{,c\}}^\top = \sum_{j=1}^{D} W_{\{j,c\}} z_j, \tag{2}$$

where the logit value on the c-th activity class is a simple addition of all $W_{\{j,c\}} z_j$. Intuitively, it can be seen as an element-wise contribution to o_c. Therefore, we are able to aggregate all $W_{\{j,c\}} z_j$ to form the concept matrix, that may be mathematically denoted as follows:

$$\mathbf{M} = \begin{bmatrix} W_{\{1,1\}} z_1 & W_{\{1,2\}} z_1 & \dots & W_{\{1,C\}} z_1 \\ W_{\{2,1\}} z_2 & W_{\{2,2\}} z_2 & \dots & W_{\{2,C\}} z_2 \\ \vdots & \vdots & \ddots & \vdots \\ W_{\{D,1\}} z_D & W_{\{D,2\}} z_D & \dots & W_{\{D,C\}} z_D \end{bmatrix}. \tag{3}$$

Since the Softmax function is implemented on all the logits to produce the final posterior probability. It is important to note that the concept matrix $\mathbf{M}$ should be constructed from all classes. That is to say, the final posterior probability will be affected by the logits from all activity classes.

Categorical Concept Invariant Learning Our key idea is that the concept matrix for activity samples of the same category should align with their corresponding mean value, implying that the concept matrix for the same activity category should be similar regardless of domains. To achieve this goal, we introduce a regularization term based on the concept matrix similarity (abbreviated as CMS) in training phase. Such regularization term may be formulated as:

$$\mathcal{L}_{\text{CMS}} = \frac{1}{N_b} \sum_{c} \sum_{\{i|y_i=c\}} \|\mathbf{M}_i - \hat{\mathbf{M}}_c\|^2, \tag{4}$$

where N_b is the number of samples in one mini-batch, $\mathbf{M}_i$ denotes the concept matrix of the i-th sample, $\hat{\mathbf{M}}_c$ denotes the mean matrix of the concept matrix corresponding to the c-th class, and $\|\cdot\|$ is l_2 norm. The $\hat{\mathbf{M}}_c$ in the above equation requires averaging the activity samples in all domains, which is impractical and computationally expensive. Since it is unrealistic to directly calculate the concept matrix of all sensor samples, we perform a dynamic momentum update to adapt the concept matrix during each training iteration, which can greatly ease the computational burden. To be specific, inspired by previous work (He et al. 2020; Chen et al. 2023) we can utilize momentum updating for $\hat{\mathbf{M}}_c$ online:

$$\hat{\mathbf{M}}_c^t = (1-\lambda) \times \hat{\mathbf{M}}_c^{t-1} + \lambda \times \frac{1}{|y_i = c|} \sum_{\{i|y_i=c\}} \mathbf{M}_i, \tag{5}$$

where λ is the positive momentum value, t is the iteration index, $|y_i = c|$ denotes the sample corresponding to the c-th class of activity identification, and $\hat{\mathbf{M}}_c$ is initialized from the first iteration to compute the processed concept matrix.

Learning Objective The overall learning objective can be written as follows:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \alpha \mathcal{L}_{\text{CMS}}, \tag{6}$$

where $\mathcal{L}_{\text{CE}}$ denotes the standard cross-entropy loss, $\mathcal{L}_{\text{CMS}}$ is the loss of CCIL, and α is a positive weight coefficient. As can be seen in Eq. 6, our CCIL is very simple, that only requires adding only a few lines of code upon the vanialla ERM training pipeline.

Experiments

Experimental Setup

Dataset and Model Architecture The widely-employed sliding window strategy is first used to segment time series data, while maintaining the same window length and overlap rate as in previous works (Ordóñez and Roggen 2016; Anguita et al. 2013; Wang et al. 2019). We directly follow the model architecture in (Lu et al. 2024) to conduct the experiments. The backbone architecture consists of two modules: the feature extractor and activity classifier. The feature extractor includes two convolutional layers followed by max-pooling operation for feature extraction, while the classifier contains a fully connected layer for final predictions. We evaluate our method on four public sensor-based HAR benchmark: DSADS (Altun, Barshan, and Tunçel 2010), PAMAP2 (Reiss and Stricker 2012), USC-HAD (Zhang and Sawchuk 2012) and UCI-HAR (Anguita et al. 2013).

Method	Target (DSADS)					Target (USC-HAD)					Target (PAMAP2)				
	0	1	2	3	AVG	0	1	2	3	AVG	0	1	2	3	AVG
ERM	83.1	79.3	87.8	71.0	80.3	81.0	57.7	74.0	65.9	69.7	90.0	78.1	55.8	84.4	77.1
DANN	89.1	84.2	85.9	83.4	85.6	81.2	57.9	76.7	70.7	71.6	82.2	78.1	55.8	87.3	75.7
CORAL	91.0	85.8	86.6	78.2	85.4	78.8	58.9	75.0	53.7	66.6	86.2	77.8	49.0	87.8	75.2
Mixup	89.6	82.2	89.2	86.9	87.0	80.0	64.1	74.3	61.3	69.9	89.4	80.3	58.4	87.7	79.0
GroupDRO	91.7	85.9	87.6	78.3	85.9	80.1	55.5	74.7	60.0	67.6	85.2	77.7	56.2	85.0	76.0
RSC	84.9	82.3	86.7	77.7	82.9	81.9	57.9	73.4	65.1	69.6	87.1	76.9	60.3	87.8	78.0
ANDMask	85.0	75.8	87.0	77.6	81.4	79.9	55.3	74.5	65.0	68.7	86.7	76.4	43.6	85.6	73.1
GILE	81.0	75.0	77.0	66.0	74.7	78.0	62.0	77.0	63.0	70.0	83.0	68.0	42.0	76.0	67.5
AdaRNN	80.9	75.5	90.2	75.5	80.5	78.6	55.3	66.9	73.7	68.6	81.6	71.8	45.4	82.7	70.4
DIVERSIFY	90.4	86.5	90.0	86.1	88.2	82.6	63.5	78.7	71.3	74.0	91.0	84.3	60.5	87.7	80.8
Ours	**94.7**	**88.2**	**92.5**	**87.5**	**90.7**	**85.2**	**66.5**	**79.3**	**77.0**	**77.0**	**93.8**	**87.2**	**63.8**	**93.2**	**84.5**

Table 1: Accuracy on cross-person generalization. We use 0,1,2,3 to denotes the unseen test set. **Bold** means the best while underline means the second-best.

Method	Target (DSADS)					
	0	1	2	3	4	AVG
ERM	41.5	26.7	35.8	21.4	27.3	30.6
DANN	45.4	25.3	38.1	28.9	25.1	32.6
CORAL	33.2	25.2	25.8	22.3	20.6	25.4
Mixup	48.8	34.2	37.5	29.5	29.9	36.0
GroupDRO	27.1	26.7	24.3	18.4	24.8	24.3
RSC	46.4	27.4	35.9	27.0	29.8	33.3
ANDMask	47.5	31.1	39.2	30.2	29.9	35.6
DIVERSIFY	47.7	32.9	**44.5**	**31.6**	30.4	37.4
Ours	**49.6**	**35.6**	44.2	31.4	**32.6**	**38.7**

Table 2: Accuracy on cross-position generalization. We use 0,1,2,3,4 to denotes the unseen test set. **Bold** means the best while underline means the second-best.

Cross-Domain Settings Cross-domain Settings are divided into the following four categories. *Cross-person* setting [2]. In the DSADS dataset, there are a total of 8 subjects. We divide the 8 subjects into 4 domains, each of which contains two subjects. We use the sliding window technique with a window size of 125 and an overlap rate of 50%. The final processed sample size is (45, 1, 125), where 45 represents sensors from 5 positions, with each position having 3 different sensors, and each sensor being 3-axis. In the USC-HAD dataset, there are a total of 14 subjects. We roughly divide them into four domains, where three of four domains with each containing four subjects are used as source domain, while the rest domain containing two subjects is utilized as target domain. We use the sliding window technique with a window size of 200 and an overlap rate of 50%. The final processed sample size is (6, 1, 200), where 6 represents sensors from one position, with this position having 2 different sensors, and each sensor being 3-axis. In the PAMAP2 dataset, there are a total of 9 subjects with subject IDs 0–8. We divide them into four domains: domains: (2, 3, 8), (1, 5), (0, 7), (4, 6). We use the sliding window technique with a window size of 200 and an overlap rate of 50%. The final processed sample size is (27, 1, 200), where 27 represents sensors from 3 positions, with each position having 3 different sensors, and each sensor being 3-axis; *Cross-position* setting. We utilize the DSADS dataset for cross-position experiments, dividing it into five domains based on position. The sliding window size and overlap rate are consistent with those in the cross-person setting. The final processed sample size is (9, 1, 125), where 9 represents three sensors from a single position, with each sensor capturing three-axis data; *Cross-dataset* setting. We merge four datasets, which are then roughly divided into four domains. We select six common activities across all four datasets which come from two sensors at similar or identical positions in each dataset. The high-frequency datasets such as USC-HAD are downsampled to ensure consistent sensor sampling frequencies for alignment. The sliding window size and overlap rate are the same as those in the cross-person setting, resulting in a final processed sample size of (6, 1, 50); *One-person-to-another* setting. We utilize the DSADS, USC-HAD, and PAMAP2 datasets, selecting four pairs of subjects to generalize from one subject to another. Specifically, the pairs are (0, 1), (2, 3), (4, 5), and (6, 7), where we generalize from the second subject in each pair to the first. The sliding window size, overlap rate, and final processed sample size are consistent with those in the cross-person setting.

Comparative Methods We compare our approach with three recent methods: GILE (Qian, Pan, and Miao 2021), AdaRNN (Du et al. 2021), and DIVERSIFY (Lu et al. 2024). We will also compare it with seven commonly used domain generalization (DG) methods: EMR (Vapnik 1991), DANN (Ganin et al. 2016), CORAL (Sun and Saenko 2016), Mixup (Zhang et al. 2018), GroupDRO (Sagawa et al. 2019), RSC (Huang et al. 2020), and ANDMask (Parascandolo et al. 2020). For a fair comparison, all methods, except GILE and AdaRNN, use the same network architecture.

Implementation Details The maximum training period was set to 150 epochs and an Adam optimizer with a weight decay of 5×10^{-4} was used. All methods utilized a learning rate of 10^{-2} or 10^{-3}. In all experiments, the batch size was set to 32. Some DG methods require domain labels to be known during training, whereas ours do not, making our approach both more challenging and more practical. For meth-

[2]Since the baselines for UCI-HAR are already good enough, we do not run cross-person experiments on it.

Method	Target				
	0	1	2	3	AVG
ERM	26.4	29.6	44.4	32.9	33.3
DANN	29.7	45.3	46.1	43.8	41.2
CORAL	39.5	41.8	39.1	36.6	39.2
Mixup	37.3	47.4	40.2	23.1	37.0
GroupDRO	51.4	36.7	33.2	33.8	38.8
RSC	33.1	39.7	45.3	45.9	41.0
ANDMask	41.7	33.8	43.2	40.2	39.7
DIVERSIFY	48.7	46.9	49.0	**59.9**	51.1
Ours	**52.1**	**48.5**	**50.3**	59.6	**52.6**

Table 3: Accuracy on cross-dataset generalization. We use 0,1,2,3 to denotes the unseen test set. 0 represents DSADS, 1 represents USC-HAD, 2 represents UCI-HAR, and 3 represents PAMAP2. **Bold** means the best while underline means the second-best.

Method	Target			
	0	1	2	AVG
ERM	51.3	46.2	53.1	50.2
Mixup	62.7	46.3	58.6	55.8
GroupDRO	51.3	48.0	53.1	50.8
RSC	59.1	49.0	59.7	55.9
ANDMask	57.2	45.9	54.3	52.5
DIVERSIFY	67.6	55.0	62.5	61.7
Ours	**70.2**	**57.5**	**63.7**	**63.8**

Table 4: Accuracy on one-person-to-another generalization. We use 0,1,2 to denotes the unseen test set. 0 represents DSADS, 1 represents USC-HAD, and 2 represents PAMAP2. **Bold** means the best while underline means the second-best.

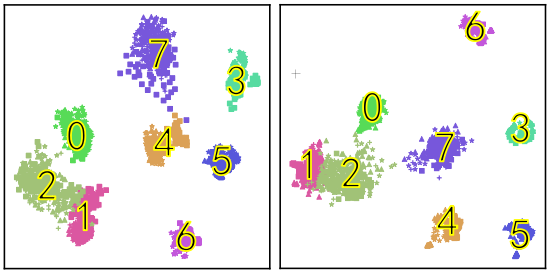
ods that require domain labels, we assigned domain labels in batches. Following the generalization setup of HAR in DIVERSIFY (Lu et al. 2024), we employed a source-domain validation strategy. The source domain data was split into training and validation sets with a ratio of 8:2. All methods were adjusted to report the average best performance over three trials. The experiments were conducted on a server equipped with a GeForce 3090 GPU.

Experimental Results

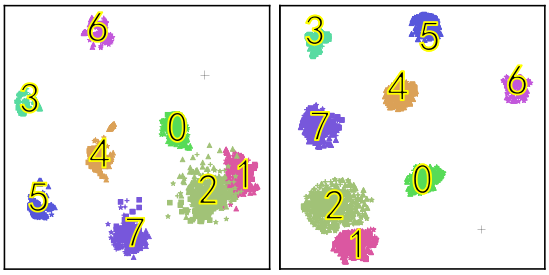
The classification results of our method for HAR under cross-person, cross-dataset, cross-position, and one-person-to-another generalization settings are presented in Tables 1-4. We draw some conclusions from these results: (1) As listed in Table 1, in the term of average accuracy, we note that the naïve ERM baseline achieves favorable performance against compared arts. Most existing strategies cannot consistently improve ERM while evaluated under the rigorous settings, and some DG methods perform even worse than ERM on certain tasks, which are in well line with previous observations in (Gulrajani and Lopez-Paz 2021). This may be due to the inability of these methods to reduce the distribution discrepancy in time series sensor data. Therefore, it is crucial to explore domain-invariant knowledge that can effectively reduce distribution discrepancy in sensor data for HAR; (2) Overall, our proposed approach consistently demonstrates superior performance against other state-of-the-art baselines under cross-person setting, where ours is ranked first place on all three benchmarks. Specifically, in terms of average accuracy, our CCIL significantly surpass the baseline ERM by large margins of 10.4%, 7.3%, and 7.4% on DSADS, USC-HAD, and PAMAP respectively. In fact, domain generalization is a challenging task, and it is often difficult to achieve an improvement over 1%. As can be seen in Table 1, the second-best baseline only has a slight improvement compared to the third one. In contrast to the best baseline DIVERSIFY, our approach still achieves further improvements of 2.5%, 3.0%, and 3.7% on all three benchmarks. The observations validate the effectiveness of our approach compared against existing baselines; (3) As aforementioned, other methods, such as CORAL, GroupDRO, and ANDMask, achieve competitive results on some tasks, but perform worse on others. This inconsistency may stem from their overlook for domain-invariant knowledge, potentially ignoring latent information between diverse distributions. DANN is another method for domain-invariant learning through adversarial training. It outperforms ERM in scenarios with a large number of classes, such as in the DSADS dataset. However, in cases with fewer classes, it performs even worse compared to ERM, as observed in the PAMAP2 dataset. Our method demonstrates robust performance regardless of the number of categories; (4) As shown in Tables 2-4, it can be seen that our CCIL still consistently achieves the matched or better performance under cross-position, cross-dataset, and one-person-to-another settings. For instance, it is well known that cross-position is more challenging than other two cases. In this case, CCIL beat the best baseline by 1.3% while achieving an improvement with about 8.1% compared to ERM under the cross-position setting. The results demonstrate our approach has a good generalization ability for time series classification under various domain generalization evaluation settings. Importantly, our approach is very simple, that require only adding a few lines of code upon the naïve ERM baseline. Moreover, it is model-agonistic and can be easily integrated with other network structures for cross-domain activity recognition. Detailed results are provided in supplementary materials.

Ablation Study

In addition to the baseline ERM method, we compare our suggested approach with the following variants to access their independent impact of each component: (1) The feature-invariance constraint abbreviated as 'W/fea', where Eq. 4 is replaced as: $\mathcal{L}_{\text{CMS}} = \frac{1}{N_b}\sum_c \sum_{\{i|y_i=c\}} \|\mathbf{z}_i - \hat{\mathbf{z}}_c\|^2$; (2) The logit-invariance constraint abbreviated as 'W/log', where Eq. 4 is replaced as: $\mathcal{L}_{\text{CMS}} = \frac{1}{N_b}\sum_c \sum_{\{i|y_i=c\}} \|\mathbf{o}_i - \hat{\mathbf{o}}_c\|^2$; (3) Ours with $\lambda = 0$ (i.e., 'W/$\lambda = 0$'), where $\hat{\mathbf{M}}$ equals the mean value dynamical calculated from current batch; (4) Ours with $\lambda = 1$ (i.e., 'W/$\lambda = 1$'), where $\hat{\mathbf{M}}$ is kept fixed from the initial pretrained model. We observe that either feature-invariance constraint or logit-invariance constraint can substantially beat the baseline ERM method. In contrast to them, our ap-



(a) ERM (b) Logit-invariance



(c) Feature-invariance (d) Ours

Figure 3: Visualization of t-SNE embedding for the DSADS dataset. Here, different colors represent different classes. Different shapes indicate different domains. Best viewed in color and zoom in.

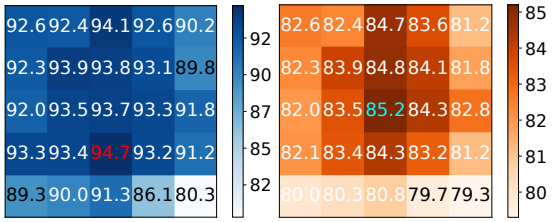
Model	Invariance			Target (DSADS)				AVG
	F	L	C	0	1	2	3	
ERM	✗	✗	✗	83.1	79.3	87.8	71.0	80.3
W/Fea	✓	✗	✗	92.3	87.3	89.1	85.5	88.6
W/Log	✗	✓	✗	86.5	83.6	88.2	79.7	84.5
W/λ=0	✗	✗	✓	89.1	86.5	88.4	78.0	85.5
W/λ=1	✗	✗	✓	92.0	87.9	90.5	86.9	89.3
Ours	✗	✗	✓	**94.7**	**88.2**	**92.5**	**87.5**	**90.7**

Table 5: Main ablation study on DSADS dataset, where 'F', 'L', and 'C' respectively indicates the feature-invariance, logits-invariance, and our concept matrix invariance constraints, while λ denotes the momentum value.

proach works the best, indicating the effectiveness of the concept matrix invariance constraint. Though the setting of $\lambda = 1$ is inferior to our optimal design, it still significantly outperforms all other variants, suggesting the necessity of dynamic momentum update.

T-SNE Visualization

To better understanding our invariance regularization, we provide a t-SNE visualization illustration on DSADS dataset, as plotted in Figure 3. In contrast to the other three strategies, it can be seen the clusters from our proposed invariance regularization are more distinctly separated, indicating its effectiveness while generalizing on an unseen distribution. This is in well line with the results reported in Table 5. Meanwhile, we observe that the logit-invariance constraint alone performs only slightly better than the baseline ERM method, both of which are inferior to the feature-invariance constraint. This is not surprising that since the logit only can provide a coarse value, which is incapable of capturing fine-grained domain-invariant representations. Therefore, the feature-invariance constraint can provide a more subtle representation compared to both of them. However, due to ignoring the effect of classifier weights, the feature-invariance constraint possibly causes the model to concentrate on unimportant features. In contrast, our CCIL can produce a more robust and stable clustering results.



(a) DSADS (b) USC-HAD

Figure 4: Parameters sensitivity analysis of α and λ. The horizontal axis signifies $\alpha \in \{0.1, 0.5, 1, 5, 10\}$, while the vertical axis denotes $\lambda \in \{0, 0.9, 0.99, 0.999, 0.9999\}$.

Parameter Sensitivity Analysis

We focus on the momentum coefficient λ and the parameter α in CCIL, which are empirically evaluated for their sensitivities by choosing values from $\{0, 0.9, 0.99, 0.999, 0.9999\}$ and $\{0.1, 0.5, 1, 5, 10\}$, respectively. The results are shown in Figure 4. It can be seen that the CCIL method has robust performance across a wide range of hyperparameters on the DSADS and USC-HAD datasets. From the results, we observed that the performance is inferior when $\lambda = 0$ compared to values such as $\lambda = 0.9$. The best performance is achieved when $\lambda = 0.9$ and $\alpha = 1$. This indicates that they play a crucial role in generalization performance, necessitating the use of a momentum update strategy for updating the concept matrix.

Conclusion

In this paper, we propose CCIL, a new regularization approach for sensor-based cross-domain activity recognition. While there exist diverse distributions in time series activity data across domains, e.g., different persons, CCIL addresses this problem by learning domain-invariant knowledge. To ensure robust outputs, the key idea of our algorithm is to capture domain-invariant knowledge by enforcing similarity between the concept matrix of samples from the same activity category and their corresponding mean value. Different from prior most works, our approach takes a different path by taking into full consideration the classifier weights (i.e., the logit-invariance), rather than only concentrating on feature-invariance. Experiments on multiple public datasets demonstrate the superiority of our CCIL approach across various cross-domain settings.

Acknowledgements

The authors would like to appreciate all participants of peer review. The work was supported in part by the National Natural Science Foundation of China under Grant 62373194, in part by the Excellent Ph.D Training Program (SEU).

References

Altun, K.; Barshan, B.; and Tunçel, O. 2010. Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recognition*.

Anguita, D.; Ghio, A.; Oneto, L.; Parra, X.; Reyes-Ortiz, J. L.; et al. 2013. A public domain dataset for human activity recognition using smartphones. In *Esann*.

Cha, J.; Lee, K.; Park, S.; and Chun, S. 2022. Domain generalization by mutual-information regularization with pretrained models. In *ECCV*.

Chen, K.; Zhang, D.; Yao, L.; Guo, B.; Yu, Z.; and Liu, Y. 2021. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Computing Surveys*.

Chen, L.; Zhang, Y.; Song, Y.; Van Den Hengel, A.; and Liu, L. 2023. Domain generalization via rationale invariance. In *ICCV*.

Dang, L. M.; Min, K.; Wang, H.; Piran, M. J.; and Moon, H. 2020. Sensor-based and vision-based human activity recognition: A comprehensive survey. *Pattern Recognition*.

Du, Y.; Wang, J.; Feng, W.; Pan, S.; Qin, T.; Xu, R.; and Wang, C. 2021. Adarnn: Adaptive learning and forecasting of time series. In *CIKM*.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*.

Gu, F.; Chung, M. H.; Chignell, M.; Valaee, S.; and Liu, X. 2021. A Survey on Deep Learning for Human Activity Recognition. *ACM Computing Surveys*.

Gulrajani, I.; and Lopez-Paz, D. 2021. In search of lost domain generalization. In *ICLR*.

Hammerla, N. Y.; Halloran, S.; and Plötz, T. 2016. Deep, convolutional, and recurrent models for human activity recognition using wearables. In *IJCAI*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*.

Huang, W.; Zhang, L.; Wu, H.; Min, F.; and Song, A. 2022. Channel-Equalization-HAR: a light-weight convolutional neural network for wearable sensor based human activity recognition. *IEEE Transactions on Mobile Computing*.

Huang, Z.; Wang, H.; Xing, E. P.; and Huang, D. 2020. Self-challenging improves cross-domain generalization. In *ECCV*.

Kong, Y.; and Fu, Y. 2022. Human action recognition and prediction: A survey. *International Journal of Computer Vision*.

Kouw, W. M.; and Loog, M. 2019. A review of domain adaptation without target labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Lu, W.; Chen, Y.; Wang, J.; and Qin, X. 2021. Cross-domain activity recognition via substructural optimal transport. *Neurocomputing*.

Lu, W.; Wang, J.; Sun, X.; Chen, Y.; Ji, X.; Yang, Q.; and Xie, X. 2024. Diversify: A General Framework for Time Series Out-of-distribution Detection and Generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ordóñez, F. J.; and Roggen, D. 2016. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*.

Parascandolo, G.; Neitz, A.; ORVIETO, A.; Gresele, L.; and Schölkopf, B. 2020. Learning explanations that are hard to vary. In *ICLR*.

Qian, H.; Pan, S. J.; Da, B.; and Miao, C. 2019. A Novel Distribution-Embedded Neural Network for Sensor-Based Activity Recognition. In *IJCAI*.

Qian, H.; Pan, S. J.; and Miao, C. 2021. Latent independent excitation for generalizable sensor-based cross-person activity recognition. In *AAAI*.

Reiss, A.; and Stricker, D. 2012. Introducing a new benchmarked dataset for activity monitoring. In *ISWC*.

Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2019. Distributionally Robust Neural Networks. In *ICLR*.

Sun, B.; and Saenko, K. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*.

Sun, Z.; Ke, Q.; Rahmani, H.; Bennamoun, M.; Wang, G.; and Liu, J. 2022. Human action recognition from various data modalities: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Vapnik, V. 1991. Principles of risk minimization for learning theory. *Advances in neural information processing systems*.

Wang, J.; Chen, Y.; Hao, S.; Peng, X.; and Hu, L. 2019. Deep learning for sensor-based activity recognition: A survey. *Pattern recognition letters*.

Wang, J.; Chen, Y.; Hu, L.; Peng, X.; and Philip, S. Y. 2018. Stratified transfer learning for cross-domain activity recognition. In *PerCom*.

Wang, J.; Lan, C.; Liu, C.; Ouyang, Y.; Qin, T.; Lu, W.; Chen, Y.; Zeng, W.; and Philip, S. Y. 2022. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*.

Wang, S.; Wang, J.; Xi, H.; Zhang, B.; Zhang, L.; and Wei, H. 2024. Optimization-Free Test-Time Adaptation for Cross-Person Activity Recognition. In *IMWUT/Ubicomp*.

Yu, Y.-C.; and Lin, H.-T. 2023. Semi-supervised domain adaptation with source label adaptation. In *CVPR*.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *ICLR*.

Zhang, M.; and Sawchuk, A. A. 2012. USC-HAD: A daily activity dataset for ubiquitous activity recognition using wearable sensors. In *IMWUT/Ubicomp*.

Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; and Loy, C. C. 2022. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhou, K.; Yang, Y.; Qiao, Y.; and Xiang, T. 2021. Domain Generalization with MixStyle. In *ICLR*.