DisCovHAR: Contrastive Attention for Human Activity Recognition Under Distribution Shifts

Luke Chen, Mohanad Odema, *Student Member, IEEE*, and Mohammad Abdullah Al Faruque, *Senior Member, IEEE*

***Abstract*—Advances in Internet of Things (IoT) wearable sensors and edge-artificial intelligence (Edge-AI) have enabled practical realizations of machine learning (ML)-enabled mobile sensing applications like human activity recognition (HAR). The effective deployment of these data-driven models necessitates learning robust representations capable of handling prevalent distribution shifts (DS), including new users, device positions, rotations, and more. In that respect, contrastive learning (CL) has shown promise in learning transformation-invariant features, outperforming traditional HAR methods. However, recent findings reveal that the contrastive loss induces shrinkage and expansion of the feature space which may limit the generalization capacity of the model. To address this, we propose *DisCovHAR*, a contrastive attention method to selectively apply the contrastive loss to a subset of the feature space through the transformer encoder attention mechanism. Extensive experiments on three HAR datasets (DSADS, PAMAP2, and USCHAD) demonstrate its superiority over state-of-the-art methods. Specifically, our approach yields up to 4.47% and 7.82% average accuracy improvements in subject-wise and position-wise generalization settings. Furthermore, *DisCovHAR* demonstrates up to 5.07% increased robustness compared to prior methods under multivariate distribution shift scenarios.**

***Index Terms*—Contrastive learning (CL), cyber–physical Systems, distribution shifts (DS), eHealth and mHealth, human activity recognition (HAR), mobile and ubiquitous systems, transformer.**

I. Introduction

THE CONVERGENCE of Internet of Things (IoT) wearable sensor technologies and edge-artificial intelligence (Edge-AI) has paved the way for innovative applications. In this paradigm, machine learning (ML) models are deployed on resource-constrained devices such as mobile phones and augmented/virtual reality (AR/VR) devices, leading to the practical realization of advanced mobile sensing applications [1], [2]. One notable application is human activity recognition (HAR), which spans a wide range of sectors. In healthcare, it facilitates health monitoring and physiotherapy, while in interactive entertainment, it enables immersive experiences in AR/VR. Additionally, in the realm of fitness, HAR empowers activity tracking and personalized coaching, revolutionizing how individuals engage with their well-being [3], [4], [5], [6], [7]. ML plays an instrumental role in advancing HAR across various contexts and settings. Its success can be attributed to the inherent ability to learn representations that capture and model complex patterns from raw sensor data [8].

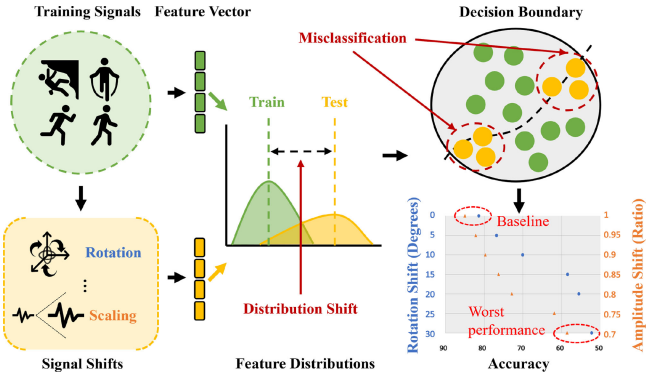


Fig. 1. DS cause misclassification and thus performance drop on a supervised HAR model.

While existing ML-based HAR traditionally relies on supervised learning with cross-entropy loss for end-to-end training, a potential limitation is the learned features' sensitivity toward dataset biases due to spurious correlations which lead to poor generalization performance [9]. This is especially a problem when the data statistics differ from training due to new users, diverse devices, varying sensor orientations, and/or positions, etc. as a result of distribution shifts (DS). Even minor deviations from the training data distribution can result in significant performance degradation from misclassification of activities in fitness applications to misdiagnosis of critical medical conditions, underscoring the urgency of addressing this challenge [10], [11]. Fig. 1 illustrates the impact of DS on HAR performance. A supervised model, when subjected to common DS in HAR, experiences a worst-case accuracy drop of 20%–30%, highlighting the necessity for generalizable HAR solutions. Realistically, the performance may be worse when exposed to multivariate DS (MV-DS), i.e., new user characteristics and device orientation variations. The need to consider MV-DS when evaluating model robustness is highlighted in [12] where they demonstrated that no evaluated methods consistently improve robustness under MV-DS.

To address this problem, an emerging trend in HAR involves the adoption of contrastive learning (CL). The contrastive loss aims to generalize features whereby similar (dissimilar) sample features are pulled together (pushed away). In the

Received 18 August 2024; revised 6 December 2024; accepted 4 March 2025. Date of publication 14 March 2025; date of current version 9 June 2025. This work was partially supported by the National Science Foundation (NSF) under Award CCF-2140154. *(Corresponding author: Luke Chen.)*

The authors are with the Department of Electrical Engineering and Computer Science, University of California at Irvine, Irvine, CA 92697 USA (e-mail: panwangc@uci.edu; modema@uci.edu; alfaruqu@uci.edu).

Digital Object Identifier 10.1109/JIOT.2025.3551263

context of DS generalization, contrasting samples are generated through transformation functions that reflect DS in the wild. The goal is to remove the effect of data/DS augmentations and learn features that are domain-invariant [11]. The intuition is that data augmentations cover the DS space (user, device, sensor differences, etc.), which is often irrelevant to the downstream tasks. Two popular CL approaches are the self-supervised SimCLR [13] and supervised CL (SupCon) [14] which have demonstrated notable generalization improvements over prior methods [15], [16], [17], marking a promising avenue for more generalized and robust HAR models. Typically, a CL model constitutes an encoder followed by a projection head that maps the feature embeddings to a lower dimensional space to apply the contrastive loss. One phenomenon observed in the SimCLR paper is that the encoder features provided superior performance over the projection head features. Appalaraju et al. [18] hypothesized that the projection head learns transformation invariant features that are useful for contrastive loss, and filters information-rich features useful for downstream tasks leading to information loss. More recently, Gui et al. [19] demystified the role of the projection head and finds that the contrastive loss induces shrinkage and expansion on the encoder features, which directly correlates to generalization performance. A key finding is that the positive rank deficit between the encoder features and the projected features correlated to better generalization performance. Similarly, Gupta et al. [20] founded that the projection head implicitly learns to apply the contrastive loss on a subset of the feature space, and by explicitly enforcing the subspace selection as part of the optimization, they showed better generalization performance.

In summary, the contrastive loss aims to learn invariant features that are generalizable but induce information loss useful for downstream tasks. This can be explained by the shrinkage and expansion effects observed through the encoder and projection features. Lastly, applying the contrastive loss on a subset of the feature space can help minimize the adverse effects. Inspired by the prior observations, we propose *DisCovHAR*, a contrastive attention method to address the shortcomings of the contrastive loss and improve the representation learning capacity of the supervised CL method. We empirically show that *DisCovHAR* results in improved features leading to better generalization, robustness, and adaptability against DS in HAR. This work provides the following key contributions.

1) We perform an in-depth study on two popular CL techniques SimCLR and Supervised CL alongside prior art methods with experiments to assess generalizability, robustness, and adaptability against univariate and multivariate DS.
2) We propose *DisCovHAR*, a Distribution-invariant contrastive-attention framework for human activity recognition to improve the representation learning capabilities of CL by modifying the standard CL framework in two ways.
   a) We address the shortcomings of the contrastive loss and its effects on feature generalization for the downstream task through feature subspace attention by utilizing an attention-based projection head.
   b) We integrate an instance normalization (IN) layer to explicitly capture shift-invariant features without additional feature engineering preprocessing.
3) DisCovHAR achieves up to 4.47% and 7.82% average accuracy improvements under subject-induced and device position-induced DS along with up to 5.07% increased robustness compared to prior methods under multivariate distribution shift settings.

By addressing these research objectives, we aim to provide meaningful insights for the application of CL in crafting robust and adaptable HAR models under diverse DS.

TABLE I
CL AND TRANSFORMERS IN HAR

Related Works	Enc./Proj.	Loss Function	DS
SimCLRHAR [15] (2020)	CNN(BN)/MLP	SSCL	Subject
SupCon [14] (2020)	CNN(BN)/MLP	SupCon.	Images
CSSHAR [21] (2021)	XFRMR/MLP	SSCL	Subject
SemiC-HAR [22] (2021)	CNN(BN)/MLP	Semi-Sup.	Subject
ColloSSL [23] (2022)	CNN(BN)/MLP	Multi-view	Device
Calda [24] (2023)	CNN(BN)/MLP	Advers.CL	Subject
ClusterHAR [25] (2023)	CNN(BN)/MLP	Cluster-SSCL	Subject
TASKED [26] (2023)	XFRMR/NA	Advers.CL	Subject
STAR [27] (2023)	STAR-FRMR/NA	Supervised	Subject
DisCovHAR (This Work)	CNN(B+**IN**)/**XFRMR**	SupCon.	**MV-DS**

II. RELATED WORKS

A. Contrastive Learning for Generalizable HAR

CL has gained significant attention in recent years owing to its discriminative representation learning capabilities that encourage similar data points to be closer while pushing dissimilar data points apart [13], [14]. One of the first works applying CL to HAR [15] uses the popular SimCLR framework [13] which demonstrated superior generalization capabilities compared to traditional supervised learning and multitask self-supervised learning approaches [28]. This introductory work focused on the effects of time-series transformations on generalizing to unseen subjects (subject-induced shifts) but did not consider other forms of DS. Many works have since adapted CL to HAR and improved its learning capacity through architectural and/or loss optimizations. Table I compares and contrasts recent works in HAR to our work. Liu and Abdelzaher [22] proposed semi-supervised CL (SSCL) by combining SimCLR [13] with SupCon loss [14] to improve encoder pretraining against subject shifts by leveraging both labeled and unlabeled data. However, SSCL does not address the contrastive loss features subspace selection problem which limits its generalization capacity and robustness [20]. Jain et al. [23] introduced a multiview contrastive loss which extends [13] to a multidevice setting. While it improves generalization against device position-induced shifts, its applicability is limited to this specific setting. Wilson et al. [24] incorporated adversarial learning with [13] to learn domain-invariant features for domain adaptation toward subjective-induced shifts. Domain adaptation assumes access to unlabeled test data during training, whereas our work focuses on the zero-shot generalization

setting without access to test data during training. Wang et al. [25] addressed the negative selection problem by incorporating an unsupervised clustering module (K-means, BIRCH, etc.) to improve negative labels for the SimCLR loss. However, its effectiveness depends on the dataset, i.e., the availability of hard negative samples and the variance in dataset performance of the clustering method not to mention the added complexity of the clustering method.

The above works address DS implicitly by improving the generalization capacity of CL mainly by modifying the loss function. They only evaluate for the subject-induced shifts or only evaluate a single type of shift (devices). Similarly, we aim to improve CL generalization capacity and robustness toward DS, however, our work focuses on architectural modifications for addressing DS and improving generalization. Specifically, compared to other CNN encoder approaches that only use batch normalization (BN), our method incorporates IN together with BN to introduce an information channel that explicitly captures the spatial configuration of the input or the signal shape [29]. This is important because signal morphology is the primary feature in discriminating between different time-series representations while ignoring the effects of style, i.e., amplitude shifts, rotation shifts, and task-irrelevant patient-specific biometric features. Additionally our use of the transformer as the projection head, unlike the traditional use as the feature encoder, addresses the contrastive loss feature subspace selection problem thereby improving the generalization capacity and robustness of CL toward DS.

B. Transformers for HAR

The feature encoder is a crucial component of ML given that model performance hinges on its ability to extract features that can generalize well to unseen data under DS. The introduction of the transformer and its self-attention mechanism has greatly improved feature extractor capabilities given its ability to adaptively focus on important features [30]. Khaertdinov et al. [21] is one of the first HAR works to integrate a transformer-based encoder in combination with CL. They replaced the traditional CNN encoder with a transformer as a more powerful feature extractor and demonstrated more robust performance and transfer-learning capabilities. Suh et al. [26] utilized transformers to capture spatiotemporal features useful for time-series data and adversarial training to enforce the features to be subject-invariant by focusing on generalizing for differences between subject behaviors. Ahn et al. [27] combined transformers with different attention mechanisms (zigzag, binary, etc.) to extract spatiotemporal cross-modal features from video and skeletal sequences for video and image-based action recognition. All of these works combine transformers with other components to improve the learned representations to construct more generalizable HAR models. However, in pursuit of performance, they neglect the added computational complexity that transformers have over traditional CNNs. This can greatly reduce a device's energy efficiency making it less practical for wearable applications. We propose a transformer-based *projection head* which provides two distinct advantages over prior works: 1) as the projection head is discarded during inference, no additional computational cost is incurred and 2) since the transformer projection head is used during training, the CNN encoder learns to attend to important features similar to performing cross-architectural knowledge distillation [31].

C. Contrastive Learning Projection Head

Given the proliferation of CL in HAR, there has been limited exploration into the various aspects of CL that contribute to its generalization capabilities, specifically the projection head. The original SimCLR paper offered initial insights into the role of the projection head, revealing that a nonlinear projection outperforms a linear one. Interestingly, the representation before the nonlinear projection consistently exhibited superior performance compared to the representation after the projection. This suggests that the hidden layer preceding the nonlinear projection retains more information, whereas the projected features experience a loss of information induced by the contrastive loss [13]. Another study [18] investigated the importance of the nonlinear projection head. Surprisingly, even with random initializations, the inclusion of the projection head contributed to better representations compared to cases without. Visual evidence from feature inversion, based on deep images prior, showed that features before the projection head retained a wide array of information, including color, shape, location, and orientation. In contrast, features after the projection head primarily preserved discriminative information intended for classification tasks. Further insights come from [20], which made two noteworthy observations. First, the projection head is a low-rank mapping, and second, the null space of the projection head contributes to generalization. They hypothesize that the projection head implicitly learns to select a subspace of features to apply the contrastive loss, effectively addressing the limitations of the contrastive loss. In the most recent study by [19], they identified two key phenomena: 1) the contrastive loss induced on the projectors causes the signal directions in the representations learned by the encoders to either expand or shrink. Using a linear projector, they demonstrate the impact of expansion and shrinkage on the downstream task generalization, where shrinkage (expansion) of the encoder hinders (improves) generalization. These observations emphasize the critical role of the projection head in influencing the generalizability of learned features. This underscores the necessity of a projection head capable of selectively applying the contrastive loss to a specific feature subspace. In essence, the role of the projection head should be to identify only the features that most effectively minimize the contrastive loss. Prior CL-based HAR works have so far overlooked the aforementioned issue, with [20] being the initial work to address it through feature subspace loss optimization. In contrast, our approach explicitly enforces the feature subspace selection by incorporating a Transformer-based projection head that selectively applies the contrastive loss to the feature subspace through its attention mechanism.

III. Problem Formulation

A. Contrastive Loss Feature Subspace Selection

CL architectures involve an encoder $h = f_\theta(x)$ that maps an input x to features h described by the feature space $\mathcal{H} \in \mathbb{R}^m$.

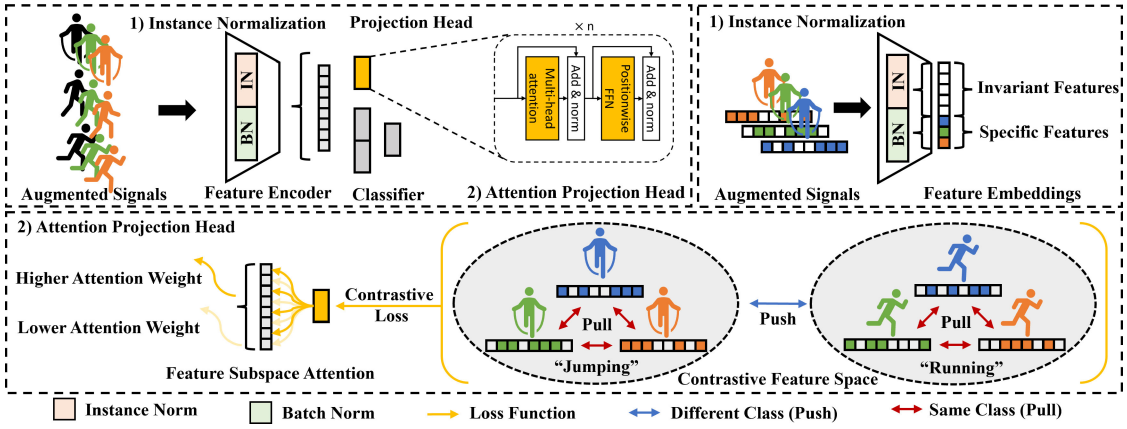


Fig. 2. *DisCovHAR* framework with IN for improved feature learning and transformer head for contrastive loss feature subspace selection. (a) Eigen value spectrum of covariance matrix of projection head outputs (z). (b) Eigen value spectrum of covariance matrix of encoder outputs (h). (c) Multiclass Receiver Operating Characteristic (One-vs-Rest).

The features h are then projected as embeddings z, described by the embedding space $\mathcal{Z} \in \mathbb{R}^n$, using a projection head $z = g_\phi(h)$. By convention, the contrastive loss is applied on g_ϕ, which is discarded after training, but the features of f_θ are used for classification [13]. As such, the contrastive loss is minimized on $\mathcal{Z}$, which is optimized to learn transformation-invariant features. However, this results in $\mathcal{Z}$ being low-rank compared to $\mathcal{H}$, indicating that the contrastive loss induces information loss of information-rich and signal-specific features in favor of invariant features [20]. To minimize these effects, the projection head must select the optimal feature *subspace* for applying the contrastive loss, learning invariant features, and retaining signal-specific features useful for the task. The contrastive loss optimization over the encoder and projection head is expressed as follows:

$$\min_{f_\theta, g_\phi} \mathcal{L}\big(g_\phi \circ f_\theta; W_\theta, W_\phi\big) \tag{1}$$

where $\mathcal{L}$ is the contrastive loss function, W_θ and W_ϕ are parameters for the encoder and projection head, respectively.

Specifically, the projection head g_θ performs a matrix multiplication of weights W_ϕ with the output features h as

$$z = g_\phi(h) = W_\phi \cdot h. \tag{2}$$

Equation (1) *implicitly* performs the feature subspace selection via back-propagation through $g_\phi(h)$, which is suboptimal. We want to *explicitly* select the feature subspace to apply the contrastive loss, in other words, find the best features that minimize the contrastive loss. This could be done by rewriting the loss function as an iterative optimization as in [20]

$$\min_{f_\theta} = \min_{f_\theta} \min_{g_\phi} \mathcal{L}\big(g_\phi \circ f_\theta; W_\theta, W_\phi\big). \tag{3}$$

Intuitively, this can also be achieved by replacing $g_\phi(h)$ with a transformer and leveraging its attention mechanism [30]. Rewriting (2) with the self-attention mechanism gives

$$z^{\text{Attn}} = g_{\text{Attn}}(h) = \text{Softmax}\left(\frac{W_\mathcal{Q} h (W_\mathcal{K} h)^T}{\sqrt{d_k}}\right) W_\mathcal{V} h \tag{4}$$

where $W_\mathcal{Q}$, $W_\mathcal{K}$, $W_\mathcal{V}$ are the learnable weights for Query ($\mathcal{Q}$), Key ($\mathcal{K}$), and Value ($\mathcal{V}$), respectively, and d_k is the dimensionality of $\mathcal{K}$. We expand in Section IV-A on how the attention mechanism performs the feature subspace selection.

Ultimately, the goal is for the encoder f_θ to capture both signal-specific and invariant features. We can encode this behavior by enforcing a subset of W_θ's weights to favor signal-specific features and another to favor invariant features. We can achieve this simply through normalization layers, specifically BN and IN by leveraging their architectural propensity for these types of features, respectively [29]. We represent these weights as follows:

$$W_{\theta^*} = W_{BN} + W_{IN}. \tag{5}$$

Finally, we rewrite the optimization (1) by combining the attention projection with the BN + IN parameters as

$$\min_{f_{\theta^*}, g_{\text{Attn}}} \mathcal{L}(g_{\text{Attn}} \circ f_{\theta^*}; W_{\theta^*}, W_{\text{Attn}}) \tag{6}$$

where $W_{\text{Attn}} = \{W_\mathcal{Q}, W_\mathcal{K}, W_\mathcal{V}\}$.

IV. DisCovHAR Framework

We introduce *DisCovHAR*, a distribution-invariant contrastive-attention framework to address the contrastive loss's shortcomings and improve CL representation learning capacity for HAR. As shown in Fig. 2 *DisCovHAR* does so in two ways: 1) a learnable IN layer to automatically capture invariant and generalizable features and 2) an attention projection head to selectively apply the CL loss on the feature subspace. We further explain each component's contribution to *DisCovHAR* as follows.

A. Attention-Based Projection Head

The transformer proposed by [30] provides a self-attention mechanism to dynamically attend to the most relevant parts of an input sequence. This naturally aligns with the role of the projection head which is to find the most relevant features that best minimize the contrastive loss. Assume some time-series feature embeddings h of dimensions $B \times C \times 1 \times T$ denoting batch, channel, and time dimensions, respectively. The expected input of the attention layer is $B \times S \times F$ where S denotes the sequence of tokens to attend over. By

aligning the channel with the sequence dimension, we can pay attention to the feature space, since the channel dimension represents different input features captured by the model. Specifically, attention scores are calculated using (4) through a dot product between $\mathcal{Q}$ and $\mathcal{K}$, scaled by the square root of the dimensionality of the key vectors $\sqrt{d_k}$ [30]. The dot product between $\mathcal{Q}$ and $\mathcal{K}$ represents the relevance of each token with each other and w.r.t. the contrastive loss optimization objective. The attention scores are then multiplied with $\mathcal{V}$, representing the inputs, to obtain an attention-weighted representation z^{Attn}. Multiheaded attention is used to capture different aspects of the feature space. As such, multiple attention scores are computed in parallel and concatenated to obtain $z^{\text{Attn}}_{\text{concat}}$. Finally, a linear output layer, represented by some weights W_O is used to map the multiheaded attention to the original feature dimensions to obtain a holistic attention representation over the feature space

$$z^* = W_O \cdot z^{\text{Attn}}_{\text{concat}}. \tag{7}$$

The multiheaded attention projection head effectively selects feature subspaces of h that are deemed most relevant for minimizing the contrastive loss. Less relevant dimensions receive lower attention weights, effectively reducing the overall impact of the contrastive loss on the final encoded representation. Given the nature of the contrastive loss which induces transformation invariance, consequently, only the feature subspaces in which the attention mechanism assigns more importance will become more invariant. Conversely, the feature subspaces considered less important are less affected by the contrastive loss, allowing them to maintain a degree of specificity.

B. Instance Normalization

Normalization layers are traditionally used for network stabilization and convergence [32]. However, different types of normalization have shown to capture different kinds of features [29]. The most common BN normalizes activations over each channel using mini-batch data [33]. Because BN retains batch statistics, it can encode features specific to the signal distributions. IN performs the same operation but on a per data-instance basis [34]. Since IN does not rely on mini-batch data, it focuses on capturing statistics that best normalize each instance independently resulting in distribution-invariant features. IN is known in computer vision tasks to learn features invariant to changes in color, style, and appearance, which are captured by BN, to promote generalization [34]. For time-series data like HAR, IN mimics z-normalization which is commonly used in time-series forecasting and classification [32]. Z-normalization helps models to focus on structural similarities and dissimilarities rather than amplitude-driven features which can be influenced by noise and other factors [35]. Similarly, IN has been found to capture the shape and trend of time-series sensor data, which is beneficial for recognizing different activities and improves generalization toward unseen domains [36], [37]. Given an encoder $f_\theta(x)$ follows the construct of convolution (conv), normalization (BN/IN), activation (act), and pooling (pool). We modify it to output the signal-specific and invariant features as follows:

$$\text{feats} = \text{conv}(x) \tag{8}$$

$$\text{concat} = BN(\text{feats}) + \text{IN}(\text{feats}) \tag{9}$$

$$\text{output} = \text{pool}(\text{act}(\text{concat})) \tag{10}$$

where the conv(x) output features are half of the original to maintain computation complexity with the added IN channels.

V. Experimental Setup

A. Dataset

We describe each dataset's processing steps, focusing on data segmentation and domain labeling. For specific dataset details, please refer to their respective papers.

DSADS [38]: Daily and Sports Activities Dataset includes 19 activities performed by eight subjects. Each data segment is a nonoverlapping five-second window sampled at 25 Hz.

USC-HAD [39]: USC Human Activity Dataset includes 12 activities performed by 14 subjects. Each data segment is a 1.26-s window sampled at 100 Hz with 50% overlap.

PAMAP2 [40]: Physical Activity Monitoring dataset includes 18 activities performed by nine subjects. Each data segment is a 1.27-s window sampled at 100 Hz with 50% overlap. We only retain common activity labels shared across all subjects (1, 2, 3, 4, 12, 13, 16, and 17).

Subject Domains: We split all datasets into four domains denoted by [0, 1, 2, 3] following the constructs of [41]. We split *DSADS* into [p1, p2], ..., [p7, p8]. We split *USC-HAD* into [Subject2, Subject12, Subject3, Subject1], [Subject7, Subject4, Subject10, Subject6], [Subject8, Subject14, Subject9, Subject11], and [Subject5, Subject13]. We split *PAMAP2* into [subject101, subject102], ..., [subject107, subject108].

Position Domains: Only *DSADS* and *PAMAP2* have sensors worn on different positions. Therefore, we split *DSADS* into five domains [0, 1, 2, 3, 4] corresponding to [torso (*t*), right-arm (*ra*), left-arm (*la*), right-leg (*rl*), and left-leg (*ll*)]. For *PAMAP2* we split into three domains [0, 1, 2] corresponding to [hand (*h*), chest (*c*), and ankle (*a*)].

B. Model

For fairness, we construct the same backbone model for all comparison works which consists of two convolution blocks, each having one convolution, batch-norm, and pooling layer with (1, 2) kernel and stride of two. The channel dimensions for all models are conv_blk1:(*input*, 16), conv_blk2:(16, 32) where *input* is dataset-dependent. The kernel size for *DSADS* and *PAMAP2* are (1, 9) and (1, 6) for *USCHAD*. The MLP projection head consists of two fully connected layers and the attention projection head is a transformer encoder with 4 attention heads using default settings. The classifiers are two fully connected layers with ReLU activation and dropout at 0.2. We train each model using stochastic gradient descent with Adam optimizer fine-tuned over learning rates [0.0001, ..., 0.1] with a fixed batch size of 256 for 500 epochs or until convergence on a validation set split from 20% of the training set. To prevent overfitting, we employed dropout (0.2) and weight-decay (1e-4) regularizations. The contrastive-learning hyper-parameters are kept at the same settings as in [14].

TABLE II
SUBJECT-WISE GENERALIZATION CLASSIFICATION ACCURACY (%)±(STD. DEV.)

Dataset	DSADS					USCHAD					PAMAP2				
Target	0	1	2	3	AVG	0	1	2	3	AVG	0	1	2	3	AVG
SimCLRHAR [15]	82.31	75.18	80.82	80.07	79.60±2.88	76.67	79.86	71.97	71.79	75.07±1.36	67.51	83.17	74.52	74.20	74.85±2.89
SupCon [14]	91.23	87.52	88.41	85.80	88.24±1.88	82.37	86.21	78.50	79.20	81.57±0.90	62.48	90.34	81.61	71.90	76.98±3.37
CSSHAR [21]	81.27	74.50	79.46	77.18	78.10±2.55	68.96	72.58	63.09	62.94	66.89±2.40	65.94	74.95	72.03	70.19	70.78±3.93
SemiC-HAR [22]	84.01	88.44	87.68	87.49	86.91±1.83	80.76	85.80	77.25	78.40	80.55±0.91	66.27	89.45	79.60	74.80	77.53±2.27
ClusterHAR [25]	79.31	79.54	81.45	83.36	80.91±2.31	77.72	80.58	73.09	73.38	76.19±1.56	70.76	85.41	78.19	74.39	77.19±2.22
Calda [24]	91.34	87.60	89.00	86.42	88.59±1.46	82.46	86.23	78.59	78.08	81.34±1.18	63.17	90.42	81.91	77.02	78.13±2.01
DisCovHAR	93.09	88.05	89.51	89.30*	**89.99**±1.29	83.60	86.64	77.03	81.49*	**82.19**±1.07	75.95*	91.94	82.53	76.98	**81.45**±1.38

C. Data Augmentation and Distribution Shifts

Tang et al. [15] proposed a set of transformation functions for CL in HAR. These include random Gaussian noise, amplitude scaling, 3-D rotation, signal inversion, time reversal, signal scrambling, time warping, and channel shuffling. After fine-tuning, they observed models trained with rotation only, and rotation + amplitude scaling resulted in the top-performing models. Similarly, Hussein et al. [11] considered real-world DS due to user movement or error causing changes in sensor orientation (rotation) and sensor location resulting in fluctuations in signal strength (amplitude). Fixed transformations like rotation and signal amplification can improve the model's robustness to specific variations encountered in HAR, though this may limit the universality of data augmentations. One can include additional domain-specific augmentations as in [15], or apply more universal augmentations for general time series data that are not limited to a specific domain [11], [42]. Although data augmentations are a fundamental part of CL, the scope of this article focuses on architectural modifications and their impact on performance. Given prior works' insights, we only applied rotation and amplitude scaling transformations. Specifically, we follow the *rotations* (0°–30°) and *amplitude* shifts (0.5-1.0 ratios) that had the most significant impact on performance as indicated in [11]. Rotation shift applies a random rotation to any of the x, y, z axes or combination thereof. Given a random rotation value r, and 3 axis acceleration values x, y, and z, we can induce rotation shift by applying the 3-D vector rotation equations for each axis [28], [43].

Rotation about the x-axis gives

$$x_{\text{new}} = x \tag{11}$$

$$y_{\text{new}} = y \cdot (r) - z \cdot \sin(r) \tag{12}$$

$$z_{\text{new}} = y \cdot \sin(r) + z \cdot \cos(r). \tag{13}$$

Rotation about the y-axis gives

$$x_{\text{new}} = x \cdot \cos(r) + z \cdot \sin(r) \tag{14}$$

$$y_{\text{new}} = y \tag{15}$$

$$z_{\text{new}} = -x \cdot \sin(r) + z \cdot \cos(r). \tag{16}$$

Rotation about the z-axis gives

$$x_{\text{new}} = x \cdot \cos(r) - y \cdot \sin(r) \tag{17}$$

$$y_{\text{new}} = x \cdot \sin(r) + y \cdot \cos(r) \tag{18}$$

$$z_{\text{new}} = z. \tag{19}$$

Amplitude shift uses the ratio $r = (x/x')$ where x' is the transformed signal of the original signal x given r. For example, given amplitude shift $r = 1.25$, $x_{\text{new}} = x/1.25$, where the amplitude of x has been scaled down by 1.25 [28], [43].

VI. EXPERIMENTS

All results presented are taken from averaging across five seed runs unless otherwise stated. We also included the standard deviation (std. dev.) across these runs to show the statistical significance of our method. We evaluate all models under various DS including subject-wise and position-wise under univariate (UV-DS) and multivariate (MV-DS) distribution shift scenarios. All models are trained using the leave-one-target-out scheme where one of the targets is left out while the remaining are used for training. We broadly categorize potential DS encountered in HAR under two types. *Intrinsic Distribution Shifts:* Derived/originating internally, i.e., patient and device position shifts which ultimately depend on the intrinsic characteristics of the measured target. *Extrinsic Distribution Shifts:* Derived/originating externally, i.e., external influences, such as rotation and amplitude shifts.

A. Subject-Wise Distribution Shifts

We first evaluate the generalization performance of *DisCovHAR* compared to prior art methods under the subject-wise distribution shift setting. This setting evaluates the scenario where a trained model experiences DS due to new and unseen subjects whose input signal characteristics may inherently vary from interpersonal variations in biometrics. From Table II, we see that *DisCovHAR* achieves the best average performance compared to all other state-of-the-art methods across all three datasets as indicated by the *bolded* value. Compared to SupCon, we see an average performance increase of 1.75%, 0.62%, and 4.47% for DSADS, USCHAD, and PAMAP2, respectively. Specifically, we indicate the target domain with the highest performance improvement by the asterisk symbol *. We see up to 3.5%, 2.29%, and 10.47% improvement for DSADS (target 3), USCHAD (target 3), and PAMAP2 (target 0), respectively. These results indicate that the learned representations from the contrastive loss alone are limited in their generalization capacity which may be due to the shrinkage and expansion effects of the contrastive loss directly impacting generalization performance. More analysis and justification of these effects are shown in Section VI-E.

B. Position-Wise Distribution Shifts

Next, we compare the generalization performance under the position-wise distribution shift in Table III. This setting evaluates DS settings due to previously unseen signal characteristics

TABLE III
POSITION-WISE GENERALIZATION CLASSIFICATION ACCURACY (%)±(STD. DEV.)

Dataset	DSADS						PAMAP2			
Target	T	RA	LA	RL	LL	AVG	H	C	A	AVG
SimCLRHAR [15]	46.61	47.50	46.36	42.41	39.65	44.51±3.07	30.26	40.34	19.26	29.95±2.60
SupCon [14]	60.88	53.76	64.01	49.54	44.37	54.51±2.32	33.50	40.71	23.17	32.46±1.56
CSSHAR [21]	49.95	44.50	48.00	43.18	40.70	45.27±3.69	29.67	35.88	17.04	27.53±4.34
SemiC-HAR [22]	60.70	49.98	59.85	43.05	36.58	50.03±2.48	31.53	38.07	23.00	30.87±2.71
ColloSSL [23]	51.43	48.81	53.57	40.33	38.56	46.54±4.28	31.93	37.08	20.98	30.00±4.22
ClusterHAR [25]	49.58	46.23	51.00	42.79	37.67	45.45±3.56	31.41	39.01	23.11	31.18±1.66
Calda [24]	60.22	55.94	66.13	53.53	45.19	56.20±2.33	34.00	41.56	22.29	32.62±1.99
DisCovHAR	59.64	62.10	70.47	60.08	59.37*	**62.33**±1.91	38.04*	43.76	20.77	**34.19**±1.84

caused by new sensor device placements which may occur as a result of user intent or error. The position shift is much more difficult to generalize due to the larger discrepancy in the signal waveform between different body parts for the same activity. For example, the signal distribution of running differs significantly when measured from the arms as compared to the legs. We see this from the overall lower average performance when compared to the subject-wise experiments. Again, *DisCovHAR* outperforms all other methods on average. Specifically, we see a significant performance disparity compared to SupCon, with 15% and 4.54% increases for the target left-leg and hand for DSADS and PAMAP2, respectively. In particular, PAMAP2 lacks position diversity where each position, hand (*h*), chest (*c*), and arm (*a*) have limited transferability which makes it harder to learn generalizable features as compared to DSADS which has right arm, left arm, right leg, and left leg correlations. Still, the performance of *DisCovHAR* indicates that the learned representations have significantly better generalization capabilities, even under the more difficult position shifts.

C. Multivariate Distribution Shifts

Most works in HAR evaluate model generalization only on univariate DS [13], [14], [21], [24]. We want to assess a model's robustness toward multivariate DS occurring naturally under realistic IoT-based wearable HAR settings. Specifically, we assess the robustness of models under increasing levels of extrinsic shifts. As such, for each intrinsic shift (subject or position), we induce an additional extrinsic shift in the form of rotational and/or amplitude shifts. Fig. 3 considers the different combinations (per graph) of intrinsic with extrinsic shifts evaluated on the DSADS dataset. Each point in a graph represents the average model performance overall leave-one-target-out test cases (i.e., the AVG column of Tables II and III). The first (left-most) point in each graph indicates the baseline performance without the influence of extrinsic shifts. Naturally, there is a negative correlation between increasing shifts (both in magnitude and in number of shifts) with decreasing performance which can be observed across all graphs. Specifically, we observe the worst-case performance drop of 19.73%, 19.36%, and 11.86% w.r.t. the baseline for SimCLR, SupCon, and *DisCovHAR*, respectively, [Fig. 3(c)].

This underscores the importance of considering multivariate DS when evaluating HAR as it depicts a much more drastic worst-case scenario for deployed models. Apart from the worst-case performance drop, what is equally important is the rate at which performance degrades which indicates the robustness of the model toward increasing extrinsic shifts. This is visualized by the slope of the trendlines in the graphs, where the steeper the slope indicates worse robustness and vice versa. We quantitatively show this by computing the average performance drop in Table IV by taking the average difference between the baseline and each extrinsic shift scenario [*X*-axis of Fig. 3]. For each distribution shift combination (DS Comb.) we see that *DisCovHAR* has the lowest average performance drops as indicated by the bolded values, and up to *5.07%* more robust when compared to SupCon, indicated by *, under the multivariate distribution shift setting SRA. Additionally, we note that rotational shifts induce higher performance drops compared to amplitude shifts, this is expected as rotations change the signal waveform, whereas amplitude retains the signal shape but scales up/down which can be more easily addressed by learning scale-invariant features.

TABLE IV
AVERAGE ACCURACY DROP (%)±(STD. DEV.) UNDER MV-DS

DS Comb.	SimCLRHAR [15]	SupCon [14]	*DisCovHAR*
SR	12.80±0.95	12.67±0.98	**8.37**±0.55
SA	1.06±0.08	0.65±0.06	**0.45**±0.04
SRA	15.02±1.83	14.74*±1.83	**9.67**±1.11
PR	7.68±0.48	5.87±0.48	**5.51**±0.32
PA	2.12±0.13	1.92±0.15	**1.86**±0.10
PRA	4.86±1.19	4.11±1.15	**3.13**±0.74

TABLE V
DATA AUGMENTATION EFFECTS ON MODEL ACCURACY (%)±(STD. DEV.)

Eval. Set	Data Aug.	SimCLRHAR [15]	SupCon [14]	*DisCovHAR*
Val.	Amp.	74.14±2.56	90.04±0.51	90.68±0.40
	Rot.	85.16±0.76	90.54±0.81	90.95±0.36
Test	Amp.	64.33±2.45	72.74±2.39	80.50±2.13
	Rot.	74.70±2.15	78.34±1.31	81.02±1.69

D. Data Augmentation Effects

Table V showcases an experiment on the individual impact of rotation and amplitude augmentations on HAR model performance. To illustrate this, we trained separate models for each augmentation on the USCHAD dataset and tested them on subject domain 0. We evaluate the validation and test set to analyze the effects of unseen *in-distribution* (validation data drawn from the same distribution (subjects) as training but not used in training) performance versus unseen *out-of-distribution* (test data is drawn from a different distribution than training and not previously seen) performance. The results

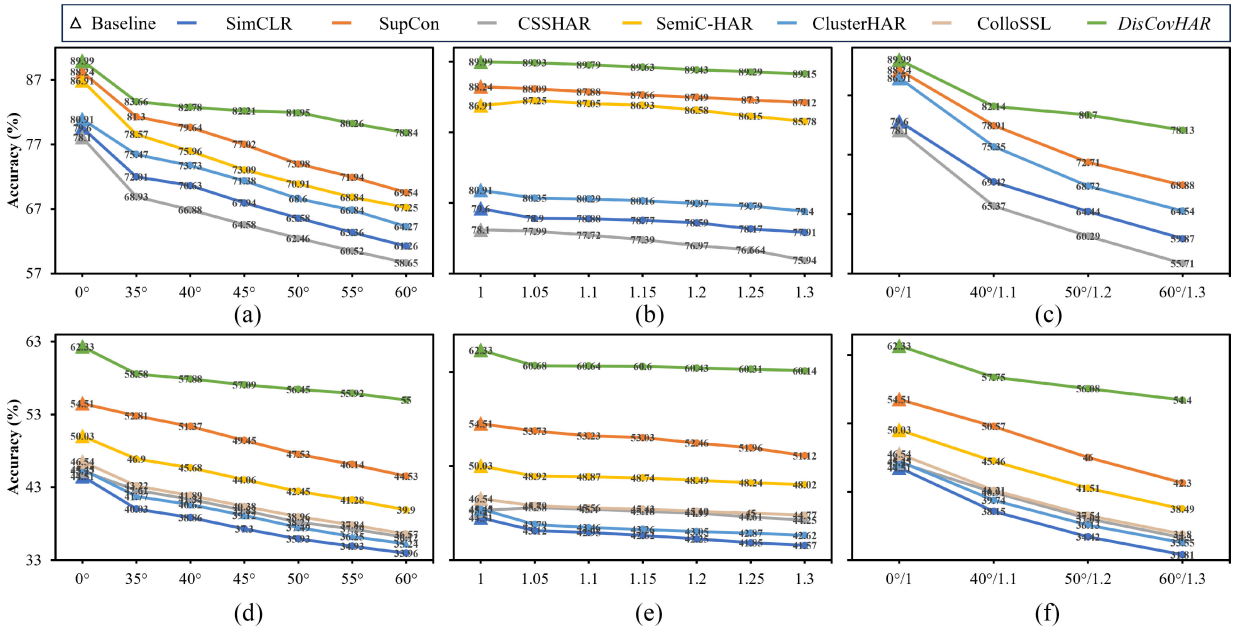


Fig. 3. Multivariate distribution shift experiments to measure the impact of increasing DS in both strength and number of shifts. (a) Subject + Rotation (SR). (b) Subject + Amplitude (SA). (c) Subject + Rotation + Amplitude (SRA). (d) Position + Rotation (PR). (e) Position + Amplitude (PA). (f) Position + Rotation + Amplitude (PRA).

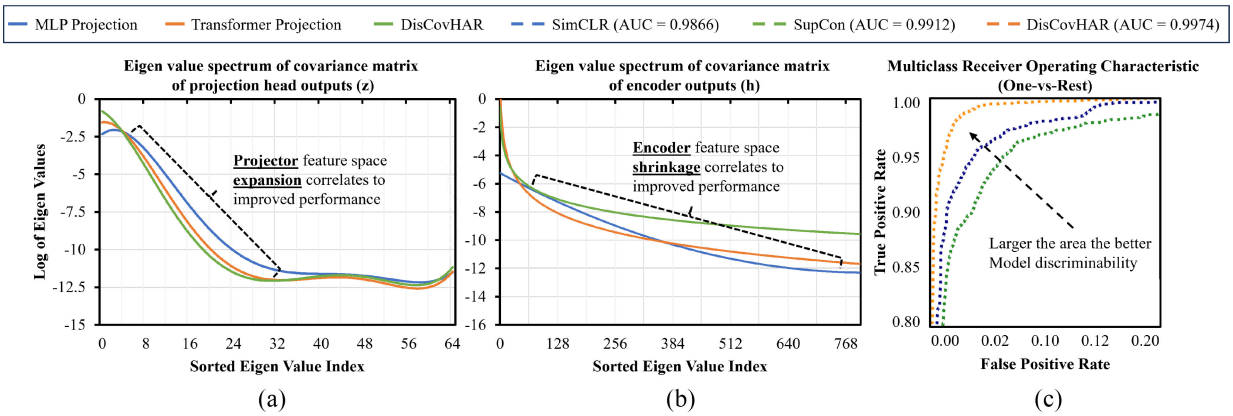


Fig. 4. Projection head feature analysis. (covariance matrix computed using torch.cov and eigenvalues computed using torch.linalg.eig).

show that rotation provides the most benefit across all three methods compared to having amplitude shifts. What is more note-worthy is our method's ability to generalize better while being *less dependent on the augmentation used*, particularly in the out-of-distribution test case. Specifically, there is a 10.4% and 5.6% discrepancy in generalization performance between models trained using only amplitude augmentations versus only rotation augmentations for SimCLRHAR and SupCon, respectively. While our method only has a 0.52% difference. This indicates that our method can better capture general features without heavily relying on the augmentation quality.

E. Projection Head Representation Analysis

We aim to uncover the reasoning behind why our attention-based projection head outperforms the standard MLP projection head by examining the eigenvalue spectrum of the covariance matrix of the learned representations for: 1) projection head (64-dimension output); 2) feature encoder (800-dimension output); and 3) the T-SNE plots of the penultimate layer before the classification output for each type of projection head. A model's intermediate feature outputs form a set of feature vectors, whose covariance matrix reveals how feature dimensions vary together. The eigenvalues indicate the amount of variance in the feature dimensions along specific directions (principal components) in the feature space. Large eigenvalues signify high variability/significance, while small eigenvalues indicate low variability/significance. A wider eigenvalue spectrum from the attention-projection head indicates richer and more diverse representations. In contrast, the MLP projection head's narrower spectrum shows limited feature representation indicating information loss induced by the contrastive loss. As can be seen in Fig. 4(a), we reaffirm the hypothesis claimed in [19] where projection heads exhibiting expansion characteristics lead to better generalization performance. Additionally, we performed the

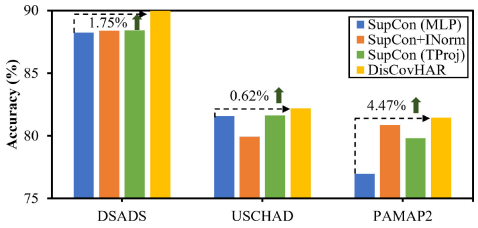


Fig. 5. Subject-wise model ablation of *DisCovHAR* components.

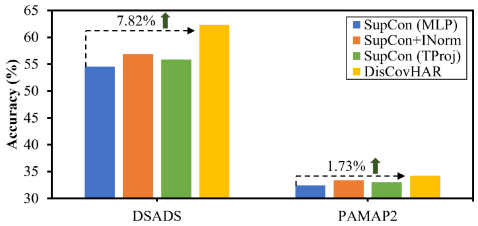


Fig. 6. Position-wise model ablation of *DisCovHAR* components.

analysis for the encoder outputs in Fig. 4(b) and found that the opposite effect is desired where the encoders exhibiting shrinkage characteristics lead to better generalization performance. The encoder output of the attention projection head has a narrower eigenvalue spectrum, reflecting the attention to important and less redundant features compared to the broader spectrum seen with the MLP projection head.

VII. Ablation Study

We perform an ablation study analyzing the contributions of each added component within *DisCovHAR*. We compare four variations, (1) SupCon, (2) SupCon + Instance Normalization (INorm), (3) SupCon + Transformer Projection Head (TProj), and (4) *DisCovHAR* which uses both INorm and TProj. Figs. 5 and 6 illustrate the average classification accuracy across the three experimental datasets for the subject-wise and position-wise scenarios, respectively. We note that depending on the dataset, the performance contributions between INorm compared to TProj vary. However, we see that by incorporating both the INorm and TProj components, *DisCovHAR* consistently achieves the best overall performance across all three datasets and scenarios.

This indicates three things: 1) the style-invariance properties introduced by the INorm layer improve the representation capacity of the encoder when compared to just using batch norm; 2) the TProj head helps mitigate the issues with the contrastive loss; and 3) the two synergizes well to further enhances the overall performance.

A. t-SNE Plots

t-distributed stochastic neighbor embedding (t-SNE) plot visualizes multivariate data. We visualize the feature space of the penultimate layer before the classification output to show how the learned representations differ in the feature space in terms of class discriminability. As can be seen in Fig. 7, the feature representations learned by SimCLR (a) are more closely clustered compared to both SupCon (b) and *DisCovHAR* (c). The closer proximity exhibits a phenomenon known as feature-collapse which leads to poor class discriminability and an indicator of model performance plateau. Compared to both (a) and (b), *DisCovHAR*'s feature clusters (c) are farther apart and more distinct resulting in better generalization.

B. Sensitivity Analysis

Finally, we performed a sensitivity analysis of the different components of the Attention-based Projection Head to analyze the effects of these components on the downstream task performance. Specifically, we investigate the number of attention heads, feedforward network dimension, and dropout ratio. We found that four attention heads on average performed the best across all three datasets. For the feedforward network dimension, we found that increasing past 2048 did not provide major performance improvements while decreasing led to performance degradation and instability. We did not observe major performance differences across different dropout ratios.

VIII. Discussion and Future Work

We applied *DisCovHAR* to the self-supervised SimCLR method. However, we found that when adding the IN layer, the training and validation contrastive loss converges rapidly ($\leq$ 5 epochs) to an extremely low value but performs worse on the test set. Although this may seem like an obvious case of model over-fitting, empirical sweeps by reducing the projection head complexity (dim_feedforward: 2048 $\rightarrow$ 128) as well as increasing its regularization strength (dropout: 0.1 $\rightarrow$ 0.5) exhibited the same behavior. This is likely due to the attention head placing a heavy focus on the invariant features introduced by the IN layer, as such, the model may have learned shortcut features that align well with the contrastive loss objective leading to suboptimal models. However, the opposite effect is observed in the supervised contrastive loss, where incorporating the instance layer improves overall generalization performance for the pretrained representations. This is likely because the addition of class label information in the contrastive loss optimization introduces a regularization that balances between invariance and task-specific information. We leave the investigation of the self-supervised SimCLR method together with a transformer-based projection head as future work.

IX. Conclusion

Recent works observed that the contrastive loss induces information loss of rich signal-specific features as a side-effect of invariant feature learning which ultimately limits generalizability. We proposed *DisCovHAR* to address the information loss by: 1) explicitly enforcing the encoder's learnable parameters to favor both invariant features and signal-specific features through instance and BN, respectively and 2) leveraging transformer-based projection head to selectively apply the contrastive loss on the optimal feature subspace, thereby minimizing information loss by focusing only on the features that best minimize the contrastive objective and retaining more information-rich features useful for the task. Extensive experiments show *DisCovHAR*, compared

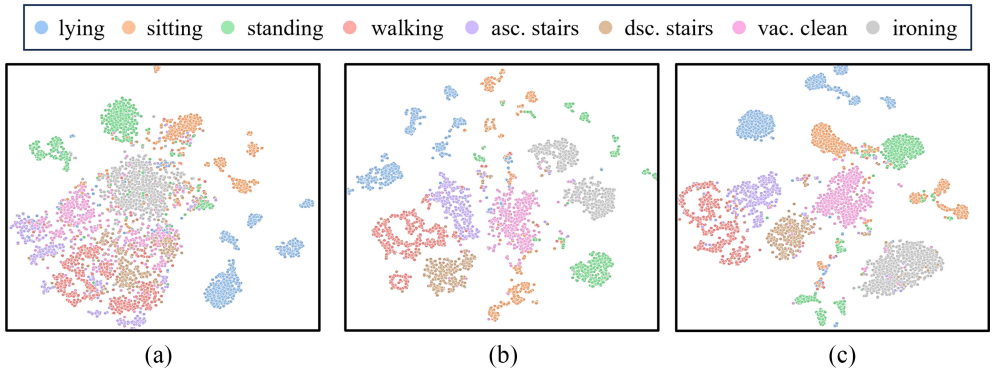


Fig. 7. t-SNE plots visualizes the feature space class separability. Axis are omitted as they are multidimensional features reduced to two arbitrary dimensions. (a) SimCLR. (b) SupCon. (c) DisCovHAR.

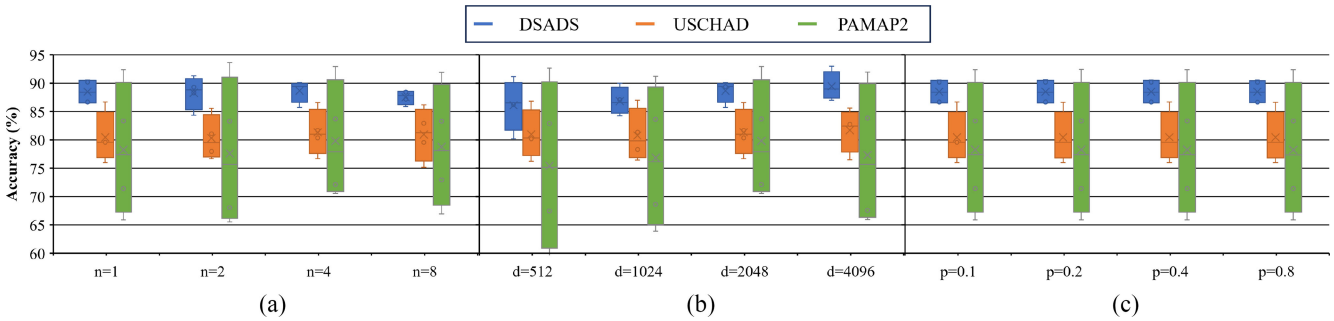


Fig. 8. Sensitivity analysis of the transformer encoder layer varying the (a) number of attention heads, (b) feedforward dimension, and (c) dropout ratio.

to prior works, improves up to 4.47% and 7.82% better generalization under subject-wise and position-wise univariate distribution shift settings, respectively. Under multivariate DS, *DisCovHAR* shows up to 5.07% improved robustness. Additionally, we verified the recent observation that expansion (shrinkage) of the projection (encoder) features leads to better generalization for HAR. Although we showed our method on HAR application, using an attention projection head for CL is independent of the application and can be generally applied to other CL frameworks.

REFERENCES

[1] N. Rashid, B. U. Demirel, and M. A. Al Faruque, "AHAR: Adaptive CNN for energy-efficient human activity recognition in low-power edge devices," *IEEE Internet Things J.*, vol. 9, no. 15, pp. 13041–13051, Aug. 2022.

[2] N. Rashid, M. Dautta, P. Tseng, and M. A. Al Faruque, "HEAR: Fog-enabled energy-aware online human eating activity recognition," *IEEE Internet Things J.*, vol. 8, no. 2, pp. 860–868, Jan. 2021.

[3] O. Barut, L. Zhou, and Y. Luo, "Multitask LSTM model for human activity recognition and intensity estimation using wearable sensor data," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8760–8768, Sep. 2020.

[4] X. Zhou, W. Liang, K. I.-K. Wang, H. Wang, L. T. Yang, and Q. Jin, "Deep-learning-enhanced human activity recognition for Internet of Healthcare Things," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6429–6438, Jul. 2020.

[5] M. Abdel-Basset, H. Hawash, R. K. Chakrabortty, M. Ryan, M. Elhoseny, and H. Song, "ST-DeepHAR: Deep learning model for human activity recognition in IoHT applications," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4969–4979, Mar. 2021.

[6] M. Abdel-Basset, H. Hawash, V. Chang, R. K. Chakrabortty, and M. Ryan, "Deep learning for heterogeneous human activity recognition in complex IoT applications," *IEEE Internet Things J.*, vol. 9, no. 8, pp. 5653–5665, Apr. 2022.

[7] T. Huynh-The, C.-H. Hua, N. A. Tu, and D.-S. Kim, "Physical activity recognition with statistical-deep fusion model using multiple sensory data for smart health," *IEEE Internet Things J.*, vol. 8, no. 3, pp. 1533–1543, Feb. 2021.

[8] A. Jordao, A. C. Nazare Jr., J. Sena, and W. R. Schwartz, "Human activity recognition based on wearable sensor data: A standardization of the state-of-the-art," 2019, *arXiv:1806.05226*.

[9] P. Izmailov, P. Kirichenko, N. Gruver, and A. G. Wilson, "On feature learning in the presence of spurious correlations," in *Proc. 36th NeurIPS*, 2022, pp. 38516–38532.

[10] A. Subbaswamy and S. Saria, "From development to deployment: Dataset shift, causality, and shift-stable models in health AI," *Biostatistics*, vol. 21, no. 2, pp. 345–352, 2020.

[11] D. Hussein, T. Belkhouja, G. Bhat, and J. R. Doppa, "Reliable machine learning for wearable activity monitoring: Novel algorithms and theoretical guarantees," in *Proc. 41st IEEE/ACM ICCAD*, 2022, pp. 1–9. [Online]. Available: https://doi.org/10.1145/3508352.3549430

[12] D. Hendrycks et al., "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *Proc. IEEE/CVF ICCV*, 2021, pp. 8340–8349.

[13] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th ICML*, 2020, pp. 1597–1607.

[14] P. Khosla et al., "Supervised contrastive learning," in *Proc. 34th NeurIPS*, 2020, pp. 18661–18673.

[15] C. I. Tang, I. Perez-Pozuelo, D. Spathis, and C. Mascolo, "Exploring contrastive learning in human activity recognition for healthcare," in *Proc. NeurIPS*, 2020, pp. 1–6.

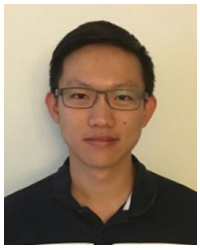
[16] H. Qian, T. Tian, and C. Miao, "What makes good contrastive learning on small-scale wearable-based tasks?" in *Proc. 28th ACM SIGKDD Conf. Knowl. Discov. Data Min.*, 2022, pp. 3761–3771.

[17] D. Schneider, S. Sarfraz, A. Roitberg, and R. Stiefelhagen, "Pose-based contrastive learning for domain agnostic activity representations," in *Proc. IEEE/CVF CVPR*, 2022, pp. 3433–3443.

[18] S. Appalaraju, Y. Zhu, Y. Xie, and I. Fehérvári, "Towards good practices in self-supervised representation learning," 2020, *arXiv:2012.00868*.

[19] Y. Gui, C. Ma, and Y. Zhong, "Unraveling projection heads in contrastive learning: Insights from expansion and shrinkage," 2023, *arXiv:2306.03335*.

[20] K. Gupta, T. Ajanthan, A. van den Hengel, and S. Gould, "Understanding and improving the role of projection head in self-supervised learning," 2022, *arXiv:2212.11491*.
[21] B. Khaertdinov, E. Ghaleb, and S. Asteriadis, "Contrastive self-supervised learning for sensor-based human activity recognition," in *Proc. IEEE IJCB*, 2021, pp. 1–8.
[22] D. Liu and T. Abdelzaher, "Semi-supervised contrastive learning for human activity recognition," in *Proc. 17th DCOSS*, 2021, pp. 45–53.
[23] Y. Jain, C. I. Tang, C. Min, F. Kawsar, and A. Mathur, "ColloSSL: Collaborative self-supervised learning for human activity recognition," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 6, no. 1, pp. 1–28, 2022.
[24] G. Wilson, J. R. Doppa, and D. J. Cook, "CALDA: Improving multi-source time series domain adaptation with contrastive adversarial learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 14208–14221, Dec. 2023.
[25] J. Wang, T. Zhu, L. L. Chen, H. Ning, and Y. Wan, "Negative selection by clustering for contrastive learning in human activity recognition," *IEEE Internet Things J.*, vol. 10, no. 12, pp. 10833–10844, Jun. 2023.
[26] S. Suh, V. F. Rey, and P. Lukowicz, "TASKED: Transformer-based adversarial learning for human activity recognition using wearable sensors via self-knowledge distillation," *Knowl. Based Syst.*, vol. 260, Jan. 2023, Art. no. 110143.
[27] D. Ahn, S. Kim, H. Hong, and B. C. Ko, "Star-transformer: A spatio-temporal cross attention transformer for human action recognition," in *Proc. IEEE/CVF WACV*, 2023, pp. 3330–3339.
[28] A. Saeed, T. Ozcelebi, and J. Lukkien, "Multi-task self-supervised learning for human activity detection," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 3, no. 2, pp. 1–30, 2019.
[29] H. Nam and H.-E. Kim, "Batch-instance normalization for adaptively style-invariant neural networks," in *Proc. 32nd NeurIPS*, 2018, pp. 1–10.
[30] A. Vaswani et al., "Attention is all you need," in *Proc. 31st NeurIPS*, 2017, pp. 1–11.
[31] Y. Liu, J. Cao, B. Li, W. Hu, J. Ding, and L. Li, "Cross-architecture knowledge distillation," in *Proc. ACCV*, 2022, pp. 3396–3411.
[32] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller , "Deep learning for time series classification: A review," *DMKD J.*, vol. 33, no. 4, pp. 917–963, 2019.
[33] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd ICML*, 2015, pp. 448–456.
[34] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2017, *arXiv:1607.08022*.
[35] D. Q. Goldin and P. C. Kanellakis, "On similarity queries for time-series data: Constraint specification and implementation," in *Proc. Int. Conf. Princ. Pract. Constraint Programm.*, 1995, pp. 137–153.
[36] Z. Hong et al., "CrossHAR: Generalizing cross-dataset human activity recognition via hierarchical self-supervised pretraining," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 8, no. 2, pp. 1–26, 2024.
[37] T. Kim, J. Kim, Y. Tae, C. Park, J. H. Choi, and J. Choo, "Reversible instance normalization for accurate time-series forecasting against distribution shift," in *Proc. ICLR*, 2021, pp. 1–25.
[38] B. Barshan and M. C. Yüksek, "Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units," *Comput. J.*, vol. 57, no. 11, pp. 1649–1667, Nov. 2014.
[39] M. Zhang and A. A. Sawchuk, "USC-HAD: A daily activity dataset for ubiquitous activity recognition using wearable sensors," in *Proc. ACM UbiComp*, 2012, pp. 1036–1043.
[40] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in *Proc. 16th ISWC*, 2012, pp. 108–109.
[41] W. Lu, J. Wang, H. Li, Y. Chen, and X. Xie, "Domain-invariant feature exploration for domain generalization," *Trans. Mach. Learn. Res.*, pp. 1–20, Jul. 2022. [Online]. Available: https://openreview.net/forum?id=0xENE7HiYm
[42] B. U. Demirel and C. Holz, "Finding order in chaos: A novel data augmentation method for time series in contrastive learning," in *Proc. 37th NeurIPS*, 2024, pp. 1–34.
[43] T. T. Um et al., "Data augmentation of wearable sensor data for Parkinson's disease monitoring using convolutional neural networks," in *Proc. 19th ACM ICMI*, 2017, pp. 216–220.



Luke Chen received the B.S. and M.S. degrees in computer engineering from the University of California at Irvine, Irvine, CA, USA, in 2019 and 2023, respectively, where he is currently pursuing the Ph.D. degree in computer engineering.

His current research focuses on distribution shifts, multimodal dynamic neural networks, and edge-split computing to build robust, adaptable, and energy-efficient ML models for various cyber–physical systems, including mobile healthcare and autonomous systems.

Mohanad Odema (Student Member, IEEE) received the B.Sc. degree in communications and electronics engineering and the M.Sc. degree in computer engineering from Ain Shams University, Cairo, Egypt, in 2014 and 2018, respectively. He is currently pursuing the Ph.D. degree in computer engineering with the University of California at Irvine, Irvine, CA, USA.

His current research interests are focused on HW/SW co-design solutions for deep learning, bridging model- and system-level optimizations to enhance the performance of core industry AI applications in vision and NLP, targeting domains, including Edge-AI, autonomous systems, and tinyML.

Mohammad Abdullah Al Faruque (Senior Member, IEEE) received the Ph.D. degree in computer science from Karlsruhe Institute of Technology, Karlsruhe, Germany, in 2009.

He is currently with the University of California at Irvine, Irvine, CA, USA, as a Full Professor and Directing the Embedded and Cyber–Physical Systems Laboratory. His current research is focused on the system-level design of embedded and cyber–physical systems with a special interest in low-power design, CPS security, and data-driven CPS design.

Prof. Al Faruque has received four Best Paper awards, such as ACSAC 2022, DATE 2016, DAC 2015, and ICCAD 2009, and many Best Paper Award nominations. He has been an IEEE CEDA Distinguished Lecturer since 2022 and an ACM Senior Member.