ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

On the Components That Enable Robust Generalization in HAR Models

Otávio Napoli and Edson Borin *

Institute of Computing - State University of Campinas
Av. Albert Einstein, 1251 - Cidade Universitária, Campinas - SP, 13083-889 - Brazil

Abstract. Generalizing human activity recognition (HAR) models across datasets remains challenging due to variations in sensors, environments, and user behavior. Domain Generalization (DG) methods attempt to address these shifts through objective-level modifications, architecture-level augmentations, and model pretraining strategies, but prior HAR studies often evaluate these components in isolation using suboptimal baselines. We systematically assess the contribution of each DG component across multiple HAR architectures, from CNNs to Transformers, using the DAGHAR benchmark. Our results show that pretraining the model with a self-supervised learning technique provides the most substantial and consistent gains in cross-dataset generalization, while architecture-level augmentations offer complementary improvements, and objective-level methods alone yield limited benefits across architectures. This suggests that DG studies should treat model pretraining as a standard baseline rather than an optional enhancement.

1 Introduction

Human activity recognition (HAR) using smartphone sensors has advanced with deep learning, yet most models are still evaluated within a single dataset using random splits or cross-validation, assuming that train and test samples share the same distribution (*i.i.d.* assumption). In real-world scenarios, however, models frequently encounter distribution shifts arising from changes in sensor placement, device type, user behavior, and environmental conditions [1]. Figure 1 illustrates these shifts with t-SNE projections from the DAGHAR benchmark [2], which contains six datasets from different domains (collected with different devices, body positions, and protocols). The plots highlight cross-position and cross-device differences, as well as the more severe cross-dataset shift combining both factors and additional variability in collection protocols and demography.

Domain Generalization (DG) methods aim to improve robustness to such shifts by training on multiple source domains and evaluating on unseen targets [3]. DG techniques can be grouped into three families: (a) objective-level methods, which modify the optimization objective, while treating the model as

*This project was supported by the Ministry of Science, Technology, and Innovation of Brazil, with resources granted by the Federal Law 8.248 of October 23, 1991, under the PPI-Softex. The project was coordinated by Softex and published as Intelligent agents for mobile platforms based on Cognitive Architecture technology [01245.003479/2024-10]. The authors also thank CNPq (315399/2023-6) and Fapesp (2013/08293-7) for their financial support, and Discovery Laboratory for their computational resources.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

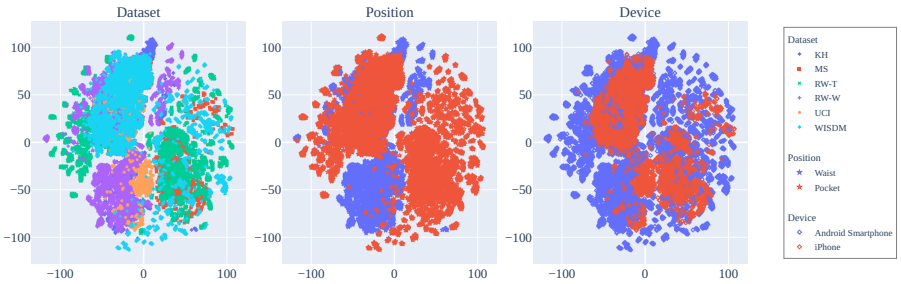

Fig. 1: t-SNE projections of DAGHAR samples in the frequency domain. The same projection is shown across panels, colored by dataset, body position, and device type to highlight different sources of domain shift. Clear cluster separation indicates strong distribution differences.

a black box; (b) architecture modifiers, which insert lightweight modules into models to perturb intermediate representations; and (c) initialization strategies, typically based on Self-Supervised Learning (SSL) pretraining, which provide more robust starting points. Although complementary, these components are often evaluated in isolation in prior HAR work [1, 4], leading to conclusions drawn from suboptimal baselines [5].

In this work, we systematically assess the contribution of each DG component to cross-dataset generalization across multiple HAR architectures, from CNNs to Transformers. Using the DAGHAR benchmark under a leave-one-dataset-out setup, we show that SSL-based initialization is one of the most important component in DG performance and should be treated as a standard component of evaluation. Feature-level augmentations provide complementary improvements, whereas objective-level methods alone offer limited benefits.

2 Domain Generalization Problem

DG aims to learn predictors that remain reliable under distribution shifts. Let $\mathcal{X}$ and $\mathcal{Y}$ denote the input and output spaces, and let a domain be a dataset $S = \{(x_i, y_i)\}_{i=1}^{n}$ drawn from a joint distribution $P_{\mathcal{XY}}$. In DG, the training set $\mathcal{S}_{\text{train}}$ contains M source domains $\{S^i\}_{i=1}^{M}$, each drawn from a distinct distribution $P^i_{\mathcal{XY}}$. The goal is to learn a predictor $h : \mathcal{X} \to \mathcal{Y}$ that generalizes to an unseen target domain $S_{\text{test}} \sim P^u_{\mathcal{XY}}$, whose distribution is unavailable during training [1].

A model (predictor) consists of a feature extractor f_θ and a classifier g_ϕ, combined as $h(x) = g_\phi(f_\theta(x))$. Training minimizes a prediction loss $\mathcal{L}$. The standard baseline, Empirical Risk Minimization (ERM), optimizes this loss over all source domains. Although ERM does not explicitly enforce domain invariance, it remains a central baseline in DG studies [3], as training with data from

[1] The label space $\mathcal{Y}$ is assumed to be shared across domains.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

diverse domains can implicitly encourage representations that generalize to unseen distributions, depending on data diversity and model capacity.

2.1 Domain Generalization Methods

DG methods aim to extract invariances shared across domains while suppressing domain-specific variation.[2]

Domain-Invariant Feature Exploration (DIFEX) *[4]* combines frequency-based guidance with cross-domain alignment. A teacher network processes the phase-only FFT of x to obtain invariant features z_T. The student representation $z = f_\theta(x)$ is split into (z_1, z_2), where z_1 is guided toward z_T and z_2 is aligned across domains using CORAL [6]. The objective is

$$\mathcal{L}_{\text{DIFEX}} = \mathcal{L}_{\text{cls}} + \alpha\|z_1 - z_T\|_2^2 + \beta(-\|z_1 - z_2\|_2^2) + \lambda\mathcal{L}_{\text{CORAL}}(z_2), \qquad (1)$$

encouraging frequency invariance, complementary representations, and second-order alignment.

MixStyle *[7]* perturbs intermediate features by mixing channel-wise statistics. For a feature map f with per-channel mean μ and std σ, and another sample (μ', σ'), a mixing coefficient λ sampled from a Beta distribution produces mixed statistics $\tilde{\mu} = \lambda\mu + (1-\lambda)\mu'$ and $\tilde{\sigma} = \lambda\sigma + (1-\lambda)\sigma'$, yielding the stylized feature

$$f_{\text{mix}} = \tilde{\sigma}\frac{f - \mu}{\sigma} + \tilde{\mu}.$$

During training, the non-trainable MixStyle layer is applied with probability p to features from selected layers, exposing the model to a broader range of domain styles and improving its robustness.

Time-Frequency Consistency (TF-C) *[8]* is an SSL method originally designed for representation learning, but also evaluated under DG conditions. TF-C enforces consistency between temporal and spectral views of the same input. Given $x \in \mathbb{R}^{C\times T}$ (C sensor axes and T time steps), augmentations generate $x_t = \mathcal{T}_{\text{time}}(x)$ and $x_f = \mathcal{T}_{\text{freq}}(\text{FFT}(x))$, where $\mathcal{T}$ is a set of stochastic transformations. Two encoders G_t and G_f produce latent features $h_t = G_t(x_t)$ and $h_f = G_f(x_f)$, which are projected to embeddings $z_t = R_t(h_t)$ and $z_f = R_f(h_f)$.

TF-C combines three contrastive objectives: a time-view loss $\mathcal{L}_t$, a frequency-view loss $\mathcal{L}_f$, and a consistency loss aligning the two views,

$$\mathcal{L}_{\text{TF-C}} = \lambda(\mathcal{L}_t + \mathcal{L}_f) + (1-\lambda)\mathcal{L}_{\text{consistency}}(z_t, z_f), \qquad (2)$$

typically implemented with NT-Xent loss. The λ parameter balances view-specific and cross-view alignment. By aligning temporal and spectral structure, TF-C yields encoders that capture domain-stable information and substantially improve downstream HAR performance after fine-tuning.

[2] We assume the target distribution shares some structure with at least one source domain so that transferable features exist.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

2.2 Limitations of Prior HAR DG Studies

Although DIFEX, MixStyle, and TF-C target complementary aspects of DG (objective-level, architecture-level, and initialization strategies, respectively), prior HAR studies typically evaluate them in isolation [1, 4], leading to incomplete conclusions and weak baselines [5]. To address this, we systematically assess each DG component, individually and in combination, across multiple HAR architectures using the DAGHAR benchmark, clarifying their contributions to domain generalization across HAR datasets.

3 Methodology

We follow a standardized evaluation protocol inspired by DomainBed [3], ensuring consistent batching, training, and model selection across all configurations.

Datasets Experiments use the DAGHAR benchmark [2], which unifies six public HAR datasets under a common format and split structure. We adopt a leave-one-dataset-out (LODO) strategy: in each run, one dataset is held out as the target domain, while the remaining five serve as sources. The train and validation splits of all source datasets form the training set, and their test splits are used for model selection (our validation set); the target dataset remains unseen until evaluation. We report performance across all target splits rather than only the test split as in prior work [2].

Training and Optimization Models are trained with domain-balanced mini-batches of 60 samples (12 per source domain). Training lasts 5000 steps using the ADAM optimizer with learning rates 10^{-3} and 10^{-4}, reporting the best result. The model with the highest validation accuracy is selected for evaluation. Each configuration is run with three random seeds. Unlike prior work [1, 2], which uses early stopping or domain truncation, we follow DomainBed's sampling and model-selection procedures.

Models and DG Techniques We evaluate four representative HAR backbones [1, 2]: CNN-PFF, ResNet-SE-5, TS2Vec, and IMUTransformer, covering convolutional and transformer-based architectures. Each backbone is combined with DG methods spanning objective-level strategies (ERM, DIFEX), architecture modifiers (MixStyle), and initialization strategies (Random or TF-C). For DIFEX, we set $\alpha = 1$, $\beta = 1$, and $\lambda = 1$, following the authors' reference implementation. For MixStyle, we apply the module after the first three convolutional layers (CNNs) or transformer blocks (Transformers) using a mixing probability of $p = 0.5$. For TF-C, each backbone is pretrained for 10000 steps on the combined source datasets, and the weights of the time-view encoder G_t is used to initialize the backbone of the target model. Pretraining excludes the held-out dataset, requiring six independent runs to maintain the LODO protocol.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

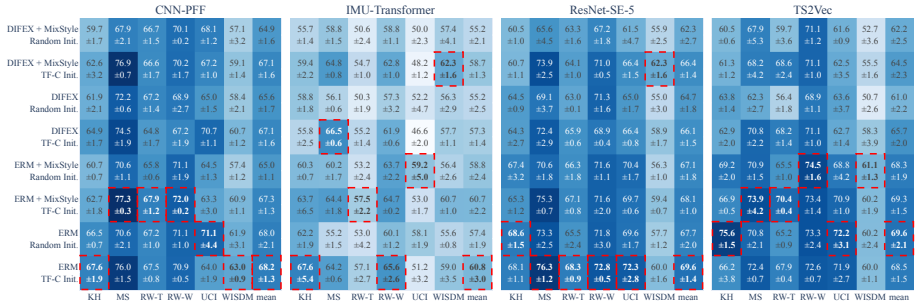

Fig. 2: Average accuracy (%) for each backbone (subplots), DG technique, and initialization strategy (rows) across all target datasets (columns) in the DAGHAR benchmark under a LODO protocol. Best results for each model-dataset pair are highlighted with a red dashed outline.

4 Experimental Results

Figure 2 summarizes the results across all backbones, DG techniques, and target domains. Some observations emerge: (i) TF-C pretraining provides the largest and most consistent gains, improving mean accuracy by almost 2% across all backbones. (ii) MixStyle yields some pontual improvements on CNN-PF and TS2Vec but has limited overall effect, with a mean decrease of about 0.39%. (iii) DIFEX does not consistently outperform simple ERM, even when combined with TF-C or MixStyle, and on average decreases performance by roughly 3.14%.

These observations were further validated through Wilcoxon paired tests, comparing all combinations of DG techniques independent of backbone and dataset. Figure 3 shows these results, in which arrows denote statistically significant differences ($p < 0.05$), pointing from the higher to the lower-performing combination. The analysis confirms that TF-C pretraining yields significant improvements, that MixStyle is an optional complementary component and that DIFEX shows no significant advantage under any condition.

These findings suggest that robust SSL initialization is a primary component of DG in HAR and should be treated as a standard evaluation component rather than an optional enhancement, contrary to previous assumptions [1, 4].

5 Conclusion

This work presents a systematic evaluation of DG techniques for smartphone-based HAR across multiple backbones and datasets in the DAGHAR benchmark. Using a leave-one-dataset-out setup, we show that SSL initialization consistently provides the largest gains in cross-dataset generalization, outperforming objective-level methods such as DIFEX. Feature-level augmentations like MixStyle offer modest complementary improvements but cannot substitute for strong initialization. Thus, robust SSL initialization is a key component for

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964.
Available from http://www.i6doc.com/en/.

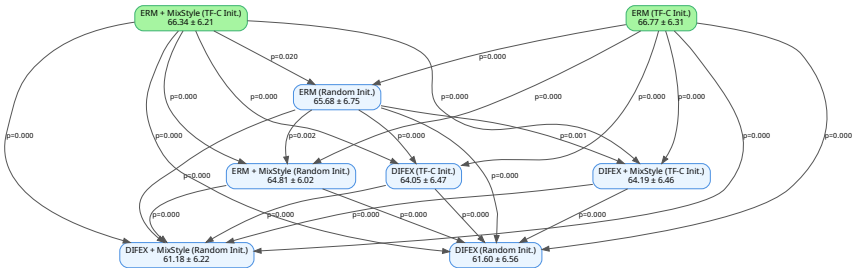

Fig. 3: Wilcoxon paired comparisons between DG technique combinations (nodes), independent of model and dataset. Arrows indicate significant differences ($p < 0.05$), pointing from higher- to lower-performing combinations. Green nodes indicate combinations with no incoming arrows (best ones).

DG in HAR and should be treated as a standard evaluation element rather than an optional enhancement [1, 4].

Although our work evaluates a selected set of DG methods, it spans a wider range of architectures than prior studies (which is typically limited to a single model or architecture type) and shows that these conclusions hold for both convolutional and transformer-based backbones. This foundation can guide future research on more effective DG strategies and support more reliable HAR systems in real-world applications such as healthcare and fitness monitoring.

References

[1] O. Napoli and E. Borin. On domain generalization for human activity recognition with mix-based methods. In *Proceedings of 33th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 491–496, 2025.

[2] O. Napoli, D. Duarte, P. Alves, D. H. P. Soto, H. E. de Oliveira, A. Rocha, L. Boccato, and E. Borin. A benchmark for domain adaptation and generalization in smartphone-based human activity recognition. *Scientific Data*, 11(1):1192, 2024.

[3] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

[4] Wang Lu, Jindong Wang, Haoliang Li, Yiqiang Chen, and Xing Xie. Domain-invariant feature exploration for domain generalization. *Transactions on Machine Learning Research*, 2022, 2022.

[5] P. Teterwak, K. Saito, T. Tsiligkaridis, K. Saenko, and B. A Plummer. Erm++: An improved baseline for domain generalization. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8525–8535. IEEE, 2025.

[6] B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.

[7] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021.

[8] X. Zhang, Z. Zhao, T. Tsiligkaridis, and M. Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in neural information processing systems*, 35:3988–4003, 2022.