Dual-Timescale Embeddings from CSI Amplitude for Cross-Domain WiFi Activity Recognition

Nguyen Tien Dung[1,a)] Kota Tsubouchi[2] Nobuhiko Nishio[3]
Masamichi Shimosaka[1]

Abstract: Cross-domain WiFi activity recognition requires models to remain accurate across changes in location, orientation, and environment. Existing methods often use engineered motion representations, such as Doppler frequency shift (DFS) profiles or body-coordinate velocity profiles (BVPs), to reduce domain variation. Whether direct channel state information (CSI) can remain useful when its temporal dynamics are modeled explicitly is less well understood. To examine this question, we develop a dual-timescale model that learns complementary short- and long-term embeddings from normalized CSI amplitude. The model requires no phase processing, time–frequency transformation, BVP reconstruction, or adaptation using recordings from the evaluation domain. On Widar3.0, it achieves 95.20% mean accuracy across cross-location, cross-orientation, and cross-environment evaluations, compared with 87.59% for the BVP-based pipeline, and approaches methods based on phase and DFS. Accuracy also improves as observations from additional receivers are integrated. These results show that direct temporal modeling of CSI amplitude can provide a compact alternative to engineered motion representations.

Keywords: WiFi sensing, channel state information, gesture recognition, domain generalization, temporal modeling, Widar3.0

1. Introduction

The central challenge in WiFi-based human activity recognition is maintaining accuracy when the sensing domain changes. Channel state information (CSI) captures motion-induced changes in wireless propagation [1], but it also varies with the room, the person's location and orientation, and the transmitter–receiver geometry. Consequently, a model trained under one set of sensing conditions may fail under another [2], [3].

Most cross-domain approaches attempt to suppress domain variation before or during recognition. Widar3.0 compared denoised CSI, Doppler frequency shift (DFS) profiles, and body-coordinate velocity profiles (BVPs) using the same CNN–GRU model family; direct CSI performed substantially worse, whereas BVP improved accuracy by about 52 points [3]. Representation-based methods have since constructed phase-derived images [4] or combined phase with DFS [5], while feature-level approaches have used augmentation and style randomization [6], [7]. Collectively, these studies suggest that conventional direct-CSI models do not transfer reliably without either representation engineering or feature-level regularization.

The remaining question is whether the temporal organization of amplitude can provide transferable activity cues despite the domain sensitivity of direct CSI. Human motion produces rapid channel variations within a slower signal evolution [8]. We hypothesize that relationships between these time scales retain activity-specific structure across sensing conditions. This hypothesis does not require amplitude itself to be domain-invariant; instead, it asks whether explicit temporal modeling can extract information that remains useful for recognition.

Guided by this hypothesis, we make temporal scale, rather than motion- representation engineering, the central design choice. One path emphasizes short-term differences, while the other preserves slowly varying context. The model compresses both paths into temporal embedding sequences, exchanges information across time, and integrates corresponding embeddings from multiple receivers. All feature extraction operates on normalized amplitude, avoiding phase processing, time–frequency transformation, and BVP reconstruction. Figure 1 summarizes this design.

Our contributions are threefold. First, we introduce a compact dual-timescale architecture that learns directly from normalized CSI amplitude. Second, we compare the proposed model with representative methods based on amplitude, phase, DFS, and BVP across changes in location, orientation, and environment. Third, we examine how input representation, the number of receiver observations, and individual model components affect recognition. Across the

[1] Institute of Science Tokyo
[2] LY Corporation
[3] Ritsumeikan University
[a)] nguyen@miubiq.cs.titech.ac.jp

© 1992 Information Processing Society of Japan

evaluated domain shifts, the proposed model surpasses the BVP-based pipeline and approaches the phase–DFS Wi-CBR system, positioning direct temporal modeling of amplitude as a competitive alternative to engineered motion representations.

2. Related Work

2.1 Cross-Domain WiFi Recognition

Cross-domain WiFi recognition methods differ primarily in what data they assume are available from a new sensing domain. EI learns features that suppress environment information [2], while AirFi and WiSR pursue generalization without new-domain recordings through feature augmentation and style randomization, respectively [6], [7]. WiGr uses labeled support samples from each new domain [9], whereas AaD and Wi-SFDAGR adapt using unlabeled recordings from that domain [10], [11]. These settings represent distinct recognition scenarios and are not directly comparable without aligning their data assumptions.

A complementary line of work focuses on the recognition architecture itself. Early model-based work related temporal CSI variation to human motion [8], and later methods combined convolutional feature extraction with long-range sequence modeling. THAT uses a two-stream temporal architecture, while Recurrent ConFormer combines convolution and recurrent processing [12], [13]. These methods provide amplitude-based architectural contrasts to systems built around engineered motion representations.

2.2 Engineered Motion Representations

Widar3.0 introduced a cross-domain recognition pipeline that estimates DFS from multiple wireless links and reconstructs a body-coordinate velocity profile (BVP) [3]. Expressing motion in body coordinates aims to reduce the dependence of link-level measurements on the user's location and orientation. BVP therefore provides a representative engineered motion description against which direct temporal modeling of CSI amplitude can be contrasted. WiGRUNT instead constructs phase-derived images and processes them with a dual-attention network [4]. Wi-CBR processes phase and DFS images in separate self-attention branches and combines their outputs through a Saliency Guidance Module [5]. Because these systems change both the input representation and the recognition network, we treat them as complete pipelines rather than as representation-only variants.

3. Proposed Method

The proposed model converts amplitude sequences from multiple receivers into a compact activity representation through three stages: dual-timescale encoding, receiver aggregation, and temporal mixing followed by classification. Each recording is resampled to 2,000 time steps and contains 90 antenna–subcarrier features per receiver.

3.1 Dual-Timescale Encoding

For each receiver, the 90 amplitude features are standardized independently at every time step. The resulting sequence is processed along two paths. The temporal-difference path emphasizes rapid changes between adjacent samples, whereas the pooled-context path smooths local variation to preserve slower signal evolution. Each path then applies a compact dilated temporal convolutional encoder and compresses the complete recording into 12 embeddings of 32 dimensions. Thus, the two paths share a common output form while emphasizing different temporal characteristics.

3.2 Receiver Aggregation and Classification

The encoder weights are shared across receivers. Let $Z_{p,r} \in \mathbb{R}^{L\times D}$ denote the $L = 12$ embeddings of dimension $D = 32$ produced by path p for receiver r. Corresponding embeddings are averaged across the R available receivers:

$$\bar{Z}_p = \frac{1}{R}\sum_{r=1}^{R} Z_{p,r}. \tag{1}$$

The resulting representation is independent of receiver order. Two gated temporal-mixer blocks then exchange information among the 12 embedding positions. An attention-pooling layer converts each path into a summary vector, and a learned gate combines the two summaries before six-class classification. Because the encoder is shared and receiver outputs are averaged, the number of trainable parameters does not depend on the number of receiver observations.

4. Evaluation

4.1 Widar3.0 Cross-Domain Protocols

Widar3.0 contains six gestures. We use five cross-location (CL) folds, five cross-orientation (CO) folds, and three cross-environment (CE) folds. Each fold uses one domain of the corresponding factor for evaluation and the remaining domains for training. Within the training domains, recordings are stratified by domain and gesture. Each complete recording belongs to only one partition, so the training, validation, and evaluation sets contain no overlapping temporal windows. Table 1 summarizes the resulting folds.

Table 1 Widar3.0 cross-domain protocols used in this study.

Protocol	Folds	Total recordings	Evaluation/fold
CL	5	6,750	1,350
CO	5	6,750	1,350
CE	3	12,745	2,988–6,750

4.2 Comparison and Reporting Protocol

For every fold, each method is trained on recordings from the training domains and evaluated on the remaining domain after every epoch. We report the highest accuracy observed on that evaluation domain during training. All comparison methods are implemented as complete pipelines and use the same Widar3.0 fold construction; the reported values are obtained from our experiments rather than copied

© 1992 Information Processing Society of Japan

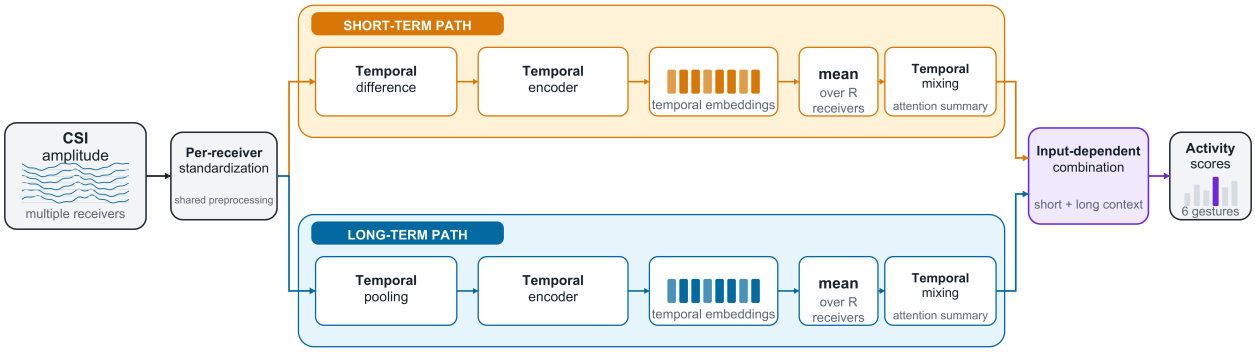


Fig. 1 Overview of the proposed model. A shared encoder processes each receiver along a temporal-difference path and a pooled-context path. Corresponding receiver embeddings are averaged within each path, mixed across time, summarized, and combined by a learned input-dependent weight.

from the original papers.

The proposed model and Wi-CBR are each trained three times to characterize variation across training runs. The receiver-count and input-representation analyses average two runs over all 13 folds, while the component analysis uses one complete run over the same folds. In the receiver analysis, the first receiver, the first three receivers, or all six receivers are supplied to the same shared encoder; all other model settings remain unchanged.

For a given protocol, accuracy is averaged over its constituent folds. The overall mean gives equal weight to the CL, CO, and CE protocol means rather than weighting protocols by their numbers of folds.

5. Results

5.1 Comparison with Existing Methods

Table 2 compares the proposed model with representative methods that use amplitude or engineered CSI representations.

Table 2 Cross-domain accuracy (%) on Widar3.0. Proposed and Wi-CBR values average three training runs; the other rows use one run.

Method	Input	CL	CO	CE	Mean
THAT [12]	amplitude	62.00	48.68	28.24	46.31
Recurrent ConFormer [13]	amplitude	65.90	77.23	31.13	58.09
Widar3.0 BVP [3]	BVP	89.36	83.04	90.36	87.59
WiGRUNT [4]	phase image	96.64	92.18	93.97	94.26
Proposed model	amplitude	95.77	**95.52**	94.31	95.20
Wi-CBR [5]	phase + DFS	**98.02**	94.73	**96.45**	**96.40**

Among the amplitude-based methods, the proposed model achieves the highest accuracy in CL, CO, and CE. Its 95.20% overall mean also exceeds the BVP-based pipeline and the phase-image WiGRUNT system. The phase–DFS Wi-CBR system has the highest overall mean, while the proposed model is higher in CO.

The proposed model contains 59,501 trainable parameters, compared with 145,062 for the BVP-based pipeline and 22,494,574 for Wi-CBR. These counts describe the recognition networks and exclude representation preprocessing and system-level costs.

Table 3 summarizes variation across the three training runs. The proposed model ranges from 95.05% to 95.40%, while Wi-CBR ranges from 96.28% to 96.52%. Their standard deviations across run-level protocol means are 0.18 and 0.12 points, respectively.

Table 3 Variation in overall accuracy (%) across three training runs.

Method	Repetitions	Mean ± SD	Range
Proposed model	3	95.20 ± 0.18	95.05–95.40
Wi-CBR	3	96.40 ± 0.12	96.28–96.52

Table 4 reports the proposed model's accuracy on each fold, averaged over three training runs. The fold means range from 91.85% on CE1 to 97.44% on CE2.

Table 4 Proposed-model accuracy (%) on each fold, averaged over three training runs.

Protocol	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean
CL	95.16	96.25	95.14	95.46	96.86	95.77
CO	93.38	96.89	95.73	96.57	95.01	95.52
CE	91.85	97.44	93.64	–	–	94.31

5.2 Influence of Receiver Count

The number of receiver observations has a consistent effect across the three cross-domain protocols. Table 5 compares the proposed model when supplied with one, three, or six receivers.

Table 5 Accuracy (%) by number of receiver observations, averaged over two training runs.

Receivers	CL	CO	CE	Mean
One	87.75	83.73	78.45	83.31
Three	93.35	91.47	91.39	92.07
Six	**95.84**	**95.59**	**94.25**	**95.22**

Accuracy increases with each additional receiver setting in CL, CO, and CE. The overall mean rises by 8.76 points from one to three receivers and by a further 3.15 points from three to six receivers.

© 1992 Information Processing Society of Japan

5.3 Input Representation Analysis

Input representation substantially affects the performance of the dual-timescale model. Table 6 compares amplitude, phase, and DFS using the same model family, six-receiver setting, and evaluation folds; only the input projection changes to match the dimensionality of each representation.

Table 6 Accuracy (%) by input representation, averaged over two training runs.

Input	CL	CO	CE	Mean
Amplitude	**95.84**	**95.59**	**94.25**	**95.22**
DFS	93.47	90.43	93.09	92.33
Phase	86.10	91.70	86.94	88.25

Amplitude achieves the highest mean in all three protocols, followed by DFS and phase. Its overall mean exceeds DFS by 2.89 points and phase by 6.97 points. Phase is comparatively stronger in CO than in CL or CE.

5.4 Component Ablation

Table 7 compares the proposed model with variants that use the same input in both paths or omit temporal mixing. All variants use CSI amplitude, six receiver observations, and mean aggregation across receivers.

Table 7 Component analysis across the three Widar3.0 protocols. Values are accuracy (%) from one complete training run.

Configuration	CL	CO	CE	Mean
Difference + pooled context (proposed)	96.00	**95.59**	**94.60**	**95.40**
Raw amplitude in both paths	**96.19**	95.10	94.10	95.13
Temporal difference in both paths	94.64	93.66	93.35	93.88
Pooled context in both paths	93.70	93.20	88.38	91.76
Without cross-temporal mixing	85.84	85.85	82.11	84.60

The proposed combination of temporal difference and pooled context obtains the highest overall mean, followed closely by the two-path raw-amplitude variant. Using temporal differences in both paths reduces the mean by 1.52 points, and using pooled context in both paths reduces it by 3.64 points. Omitting temporal mixing produces the largest decrease, at 10.80 points.

6. Discussion

6.1 Interpretation of Results

The central finding is that direct temporal modeling makes normalized CSI amplitude competitive with pipelines built around engineered motion representations. The proposed model outperforms the amplitude-based comparison methods and the BVP-based pipeline, while remaining close to the phase–DFS Wi-CBR system. It does so with a compact recognition network, although the parameter comparison does not include representation preprocessing or system-level latency. Together, these observations show that improving how amplitude is modeled can provide an alternative to replacing it with an engineered motion representation.

The receiver analysis highlights the value of complementary spatial observations. Different receivers observe a gesture through different propagation paths, so motion that is weak or distorted on one link may remain visible on another. Averaging their encoded features can preserve patterns shared across links while reducing receiver-specific variation. Because the encoder is shared, additional receivers increase computation but not the number of trainable parameters. The smaller increase from three to six receivers suggests that the first additional observations provide much of the available complementary information.

The input analysis further shows that the proposed model is particularly well matched to amplitude. Under the same model family, amplitude outperforms both phase and DFS across the three protocols. This pattern indicates that the model can extract useful motion information without first converting CSI into an engineered representation.

The component analysis distinguishes the roles of path-specific preprocessing and temporal interaction. The two raw-amplitude paths remain close to the proposed configuration, whereas removing the temporal mixer produces the largest decrease. The different path inputs therefore provide a useful temporal bias, while interaction across the compressed embedding sequence has the larger observed effect.

6.2 Limitations

The main comparison evaluates each method as a complete pipeline. The BVP and Wi-CBR systems differ from the proposed model in their representations, preprocessing, recognition networks, objectives, and computational costs. Table 2 therefore compares complete approaches rather than isolating the effect of a single representation or architectural choice.

The repeated training runs characterize sensitivity to initialization but do not capture variation across users, recording sessions, or independent datasets. In addition, the reported maximum for each fold is selected using accuracy on its evaluation domain. The results therefore describe the best observed epoch under this reporting rule rather than performance selected without access to evaluation-domain labels.

7. Conclusion

We investigated whether explicit temporal modeling can support cross-domain activity recognition directly from normalized CSI amplitude. The proposed model combines short-term differences with slower context, exchanges information across their compact embedding sequences, and integrates observations from multiple receivers. Across the Widar3.0 protocols, it outperforms the BVP-based pipeline and approaches the phase–DFS Wi-CBR system; the accompanying analyses also highlight the benefits of additional receiver observations and temporal interaction. These results establish amplitude-based temporal modeling as a compact and competitive direction for cross-domain WiFi activity recognition.

© 1992 Information Processing Society of Japan

References

[1] Yousefi, S., Narui, H., Dayal, S., Ermon, S. and Valaee, S.: A Survey on Behavior Recognition Using WiFi Channel State Information, *IEEE Communications Magazine*, Vol. 55, No. 10, pp. 98–104 (online), DOI: 10.1109/MCOM.2017.1700082 (2017).

[2] Jiang, W., Miao, C., Ma, F., Yao, S., Wang, Y., Yuan, Y., Xue, H., Song, C., Ma, X., Koutsonikolas, D., Xu, W. and Su, L.: Towards Environment Independent Device Free Human Activity Recognition, *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, pp. 289–304 (online), DOI: 10.1145/3241539.3241548 (2018).

[3] Zhang, Y., Zheng, Y., Qian, K., Zhang, G., Liu, Y., Wu, C. and Yang, Z.: Widar3.0: Zero-Effort Cross-Domain Gesture Recognition With Wi-Fi, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 44, No. 11, pp. 8671–8688 (online), DOI: 10.1109/TPAMI.2021.3105387 (2022).

[4] Gu, Y., Zhang, X., Wang, Y., Wang, M., Yan, H., Ji, Y., Liu, Z., Li, J. and Dong, M.: WiGRUNT: WiFi-Enabled Gesture Recognition Using Dual-Attention Network, *IEEE Transactions on Human-Machine Systems*, Vol. 52, No. 4, pp. 736–746 (online), DOI: 10.1109/THMS.2022.3163189 (2022).

[5] Zhang, R., Tang, S., Yan, H., Zhang, X. and Guo, J.: Wi-CBR: Salient-Aware Adaptive WiFi Sensing for Cross-Domain Behavior Recognition, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40, No. 2, pp. 1552–1560 (online), DOI: 10.1609/aaai.v40i2.37131 (2026).

[6] Wang, D., Yang, J., Cui, W., Xie, L. and Sun, S.: AirFi: Empowering WiFi-Based Passive Human Gesture Recognition to Unseen Environment via Domain Generalization, *IEEE Transactions on Mobile Computing*, Vol. 23, No. 2, pp. 1156–1168 (online), DOI: 10.1109/TMC.2022.3230665 (2024).

[7] Liu, S., Chen, Z., Wu, M., Liu, C. and Chen, L.: WiSR: Wireless Domain Generalization Based on Style Randomization, *IEEE Transactions on Mobile Computing*, Vol. 23, No. 5, pp. 4520–4532 (online), DOI: 10.1109/TMC.2023.3292229 (2024).

[8] Wang, W., Liu, A. X., Shahzad, M., Ling, K. and Lu, S.: Understanding and Modeling of WiFi Signal Based Human Activity Recognition, *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, pp. 65–76 (online), DOI: 10.1145/2789168.2790093 (2015).

[9] Zhang, X., Tang, C., Yin, K. and Ni, Q.: WiFi-Based Cross-Domain Gesture Recognition via Modified Prototypical Networks, *IEEE Internet of Things Journal*, Vol. 9, No. 11, pp. 8584–8596 (online), DOI: 10.1109/JIOT.2021.3114309 (2022).

[10] Yang, S., Wang, Y., Wang, K., Jui, S. and van de Weijer, J.: Attracting and Dispersing: A Simple Approach for Source-Free Domain Adaptation, *Advances in Neural Information Processing Systems*, Vol. 35, pp. 5802–5815 (online), DOI: 10.52202/068431-0420 (2022).

[11] Yan, H., Zhang, X., Huang, J., Feng, Y., Li, M., Wang, A., Ou, W., Wang, H. and Liu, Z.: Wi-SFDAGR: WiFi-Based Cross-Domain Gesture Recognition via Source-Free Domain Adaptation, *IEEE Internet of Things Journal*, Vol. 12, No. 13, pp. 24159–24173 (online), DOI: 10.1109/JIOT.2025.3554228 (2025).

[12] Li, B., Cui, W., Wang, W., Zhang, L., Chen, Z. and Wu, M.: Two-Stream Convolution Augmented Transformer for Human Activity Recognition, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, No. 1, pp. 286–293 (online), DOI: 10.1609/aaai.v35i1.16103 (2021).

[13] Shang, M. and Hong, X.: Recurrent ConFormer for WiFi Activity Recognition, *IEEE/CAA Journal of Automatica Sinica*, Vol. 10, No. 6, pp. 1491–1493 (online), DOI: 10.1109/JAS.2023.123291 (2023).

© 1992 Information Processing Society of Japan