Context-Aware Predictive Coding: A Representation Learning Framework for WiFi Sensing

Borna Barahimi[1, 2*] **Hina Tabassum**[1] **Mohammad Omer**[2] **Omer Waqar**[3]
[1]York University
[2]Cognitive Systems Corp.
[3]University of the Fraser Valley
bornab@yorku.ca

Abstract

WiFi sensing is an emerging technology that uses wireless signals for sensing applications like human activity recognitions, but challenges like limited labeled data and complexity of channel state information (CSI) data, hinder model performance and generalization. We propose Context-Aware Predictive Coding (CAPC), a novel self-supervised learning (SSL) framework for CSI-based WiFi sensing. CAPC integrates temporal contrastive prediction with an augmentation-based contextual approach, which captures temporal dependencies while eliminating CSI estimation and transceiver errors through a proposed dual view augmentation method. The combined strategy ensures that CAPC learns robust and informative representations while maintaining temporal coherence and contextual consistency. CAPC outperforms supervised and SSL baselines, especially with limited labeled data from unseen environments and in transfer learning to new datasets and tasks, making it a strong alternative for WiFi sensing applications.

1 Introduction

The widespread availability of the internet has resulted in a proliferation of WiFi-enabled devices across various settings, from commercial to residential areas. WiFi-enabled devices have evolved beyond basic connections, now functioning as sensors for human sensing applications through the analysis of wireless signal characteristics such as channel state information (CSI). CSI captures detailed wireless channel properties, including amplitude and phase, allowing for the detection of environmental changes caused by human movements or objects [1]. WiFi sensing offers a cost-efficient, privacy-preserving, and non-invasive solution that leverages existing infrastructure and functions effectively even in non-line-of-sight scenarios. This technology supports diverse applications, including localization [2], human activity recognition (HAR) [3], gesture recognition [4, 5, 6], and respiration detection [7].

Deep learning models have achieved state-of-the-art performance in extracting CSI patterns for complex activities without requiring extensive preprocessing [8]. However, their success is limited by the scale and diversity of labeled data, as CSI data is challenging to collect and annotate. Recent methods focus on data-efficient approaches, particularly self-supervised learning (SSL), to leverage unlabeled data for training. SSL allows models to learn general features from large unlabeled datasets and fine-tune with minimal labeled data, addressing the scarcity of CSI labels and improving performance in real-world environments [9, 10, 11, 12].

*This research was supported by an Alliance Grant funded by the Natural Sciences and Engineering Research Council of Canada (NSERC).

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

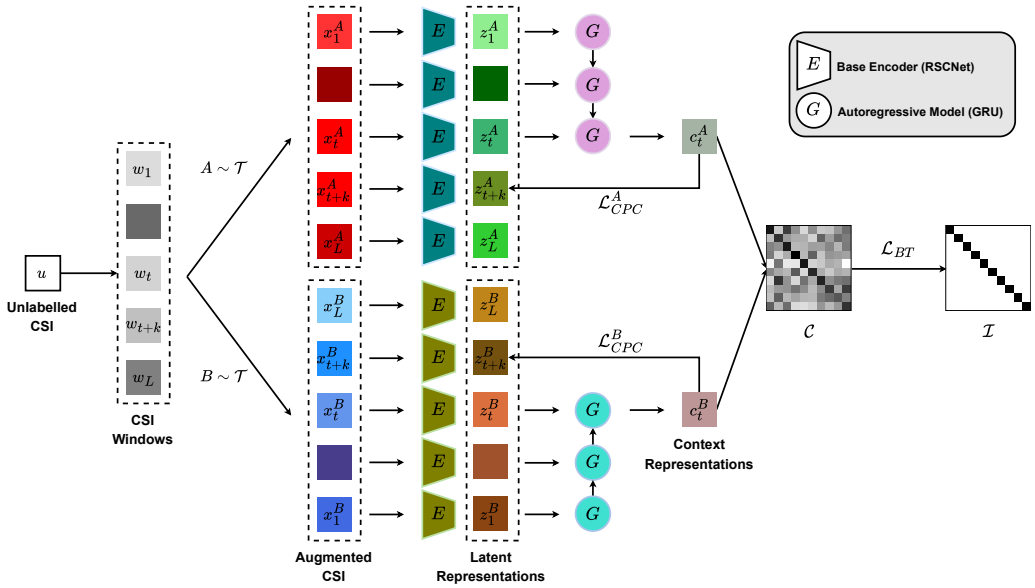


Figure 1: Overview of the CAPC's architecture. Here, $\boldsymbol{w_t}$ denotes a window of sample $\boldsymbol{u}$. The symbols $\boldsymbol{x_t}$, $\boldsymbol{Z_t}$, and $\boldsymbol{c_t}$ represent the augmented CSI for window t, the latent representation of this window, and the accumulated context embedding up to window t, respectively. Different colours signify distinction in the windows, their representations, and model parameters between branches A and B.

To date, various SSL methods have been considered for WiFi sensing [13, 14, 15, 16, 9] that utilize contrastive learning. However, they do not consider the temporal aspects of CSI, wireless propagation channel, or transceiver characteristics. In contrast to images that primarily exhibit spatial features, CSI, a time-series data, is predominantly defined by its temporal relationships [17, 18]. Moreover, common augmentation methods designed for computer vision tasks, like colour distortion or rotation, typically do not align well with the nature of CSI data [19]. Finally, the recent developments in non-contrastive methods, which do not rely on negative samples [20, 21], have yet to be explored in the WiFi sensing domain.

In this study, we propose a novel SSL framework for CSI WiFi sensing, named ***C**ontext-**A**ware **P**redictive **C**oding*[2] (**CAPC**). CAPC combines the ability of Contrastive Predictive Coding (CPC) [22] to capture temporal patterns in the CSI time-series with the Barlow Twins (BT) [20] objective, which encourages the creation of representations invariant to augmentations while minimizing redundancy. To fully leverage the BT framework and better separate environment-related information from distortions, we introduce a novel augmentation strategy tailored to CSI-based sensing. This strategy utilizes uplink and downlink CSI views to maximize the mutual information between their representations. This approach effectively isolates free-space propagation effects while reducing electronic distortions and estimation errors. Our results demonstrate that CAPC significantly improves performance in unseen environments and enhances transfer learning capabilities, outperforming existing SSL baselines, especially when limited labeled data is available.

2 Methodology

Given a CSI sample $\boldsymbol{u_i}$ in the unlabeled batch U with dimensions $N_a \times N_s \times N_t$, where N_a denotes the number of transmitter-receiver antenna links, N_s denotes the number of subcarriers, and N_t denotes the number of packets or timestamps, we segment this a sample into multiple windows $\{\boldsymbol{w_1^i}, \ldots, \boldsymbol{w_L^i}\}$ with $N_f \leq N_t$ number of CSI frames along the time dimension.

[2]The source code of CAPC is available on GitHub at https://github.com/bornabr/CAPC.

CAPC employs a multi-task objective: (1) predicting future windows $\{\boldsymbol{w_{t+1}^i}, \ldots, \boldsymbol{w_{t+T}^i}\}$ based on prior windows $\{\boldsymbol{w_1^i}, \ldots, \boldsymbol{w_t^i}\}$, where t is a random window index and T is the number of future windows for prediction. (2) generating two augmented views, $\boldsymbol{x_t^A}$ and $\boldsymbol{x_t^B}$, and maximizing their similarity.

We use two base encoders with distinct parameters, E_{θ^A} and E_{θ^B}, to generate latent representations for the two views, $\boldsymbol{z_t^A} = E_{\theta^A}(\boldsymbol{x_t^A})$ and $\boldsymbol{z_t^B} = E_{\theta^B}(\boldsymbol{x_t^B})$, where $\boldsymbol{z_t^A}, \boldsymbol{z_t^B} \in \mathbb{R}^D$. The goal is to make the latent representations $\boldsymbol{z}$ feature-rich and robust for downstream tasks. Using an autoregressive model G, CAPC summarizes the latent representations from $\boldsymbol{z_1}$ up to window $\boldsymbol{z_t}$, creating $\boldsymbol{c_t^A} = G_{\gamma^A}(\boldsymbol{z_{\leq t}^A})$ and $\boldsymbol{c_t^B} = G_{\gamma^B}(\boldsymbol{z_{\leq t}^B})$, where $\boldsymbol{c_t^A}, \boldsymbol{c_t^B} \in \mathbb{R}^H$.

To capture the contextual and temporal information from the time-series CSI, we leverage a temporal contrastive predictive loss inspired by CPC [22] and the BT [20] loss for augmentation-based contextual consistency.

Temporal Contrastive Prediction Loss This loss ensures the model learns representations that capture mutual information and temporal continuity between the windows by predicting near-future windows. The context embedding $\boldsymbol{c_t^i}$ of sample i predicts future windows $t+k$, for each $k \leq T$. Instead of explicit predictions, the representation $\boldsymbol{z_{t+k}^i}$ serves as a positive sample, while representations from other instances, $\boldsymbol{z_{t+k}^j}$, are negative samples. The loss for sample i and prediction of window $t+k$ is:

$$\ell_{temp}^{i,k} = -\log \frac{\exp((\boldsymbol{z_{t+k}^i})^{\boldsymbol{T}} \boldsymbol{W_k} \boldsymbol{c_t^i})}{\sum_{j=1}^{N} \exp((\boldsymbol{z_{t+k}^j})^{\boldsymbol{T}} \boldsymbol{W_k} \boldsymbol{c_t^i})}, \tag{1}$$

where N is the batch size and $\boldsymbol{W_k} \in \mathbb{R}^{D \times H}$ is a linear layer used for predicting window $t+k$. Note that this loss is applied to each view of samples independently but using the same timestamp t and transformations $\boldsymbol{W}$.

Augmentation-based Contextual Consistency Loss To enhance robustness against augmentations, we use the covariance loss of BT [20]. The covariance matrix between context embeddings of views $\boldsymbol{c}^A$ and $\boldsymbol{c}^B$ is:

$$\boldsymbol{\mathcal{C}}_{ij} = \frac{\sum_{b=1}^{N} \boldsymbol{c}_i^{b,A} \boldsymbol{c}_j^{b,B}}{\sqrt{\sum_{b=1}^{N} (\boldsymbol{c}_i^{b,A})^2 \sum_b (\boldsymbol{c}_j^{b,B})^2}}, \tag{2}$$

where b is the index of batch sample, and i, j indexes the features of the context latent representations. The contextual loss ensures invariance to augmentations by setting the diagonal of the cross-correlation matrix to 1 and minimizing redundancy by setting off-diagonal elements to 0. The overall contextual loss is controlled by the hyperparameter λ and is defined as:

$$\mathcal{L}_{context} = \underbrace{\sum_{i=1}^{H} (\boldsymbol{\mathcal{C}}_{ii} - 1)^2}_{\text{invariance term}} + \underbrace{\lambda \sum_{i,j=1}^{H} \boldsymbol{\mathcal{C}}_{ij}^2}_{\text{redundancy reduction term}}, \tag{3}$$

The two losses are complementary: temporal loss extracts dynamic time-dependent trends, while contextual loss is timeless and enforces augmentation invariance. The total loss is controlled by the hyperparameter β, balancing the two losses, and is defined as:

$$\mathcal{L} = \mathcal{L}_{context} + \beta \left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{k=1}^{T} (\ell_{temp}^{i,k,A} + \ell_{temp}^{i,k,B}) \right). \tag{4}$$

Type of augmentations plays a crucial role in SSL by helping models learn robust features for downstream tasks by ensuring invariance to distortions. We propose a novel augmentation, called *dual view*, for wireless sensing, which leverages the reciprocity of wireless channels ([23]) to isolate the free space propagation effects from electronic distortions. By assigning uplink and downlink CSI as $\boldsymbol{x}^A$ and $\boldsymbol{x}^B$ views, using these two enforces the contextual loss to extract the mutual information between them which is the core characteristics of the channel that are essential for sensing applications.

3 Results and Discussions

Method	SignFi Home				UT HAR		
	2	4	6	Avg.	10	20	Avg.
Supervised	-	57.97	78.99	45.65	8.0	56.0	32.0
SimCLR	59.6	82.25	92.57	78.14	52.6	57.2	54.9
CPC	55.89	75.82	86.23	72.65	52.2	54.4	53.3
BT	54.08	76.00	92.66	74.25	52.8	56.0	54.4
AutoFi	59.15	79.62	92.84	77.20	51.2	55.0	53.1
CAPC‡	63.41	85.51	92.48	80.47	49.8	57.4	53.6
CAPC	**65.67**	**88.50**	**93.84**	**82.67**	**54.2**	**59.2**	**56.7**

Table 1: **Evaluations results** of linear classification from frozen representations of pre-trained encoders with CAPC and baseline SSL methods on SignFi lab data.

Method	**Dual View**	Time Mask	Subcarrier Mask	Time Flip	Noise	Avg.
AutoFi	49.98	80.16	79.64	80.67	82.97	74.68
BT	79.22	87.73	86.76	84.00	87.77	88.10
SimCLR	86.50	**89.33**	88.60	89.04	88.95	88.60
CAPC	**91.16**	87.41	**89.67**	**89.11**	**91.89**	**89.85**
Avg.	76.72	86.16	86.17	85.71	**87.89**	84.53

Table 2: Summary of the average accuracies achieved using various augmentations across different methods.

We evaluate the proposed CAPC model and other SSL baselines using linear classification on the SignFi home [24] and UT HAR datasets. The SignFi dataset is used for pre-training in a lab setting and evaluated in a home environment for sign language detection, while the UT HAR [25] dataset is employed for transfer learning to HAR task. In both cases, the pre-trained encoder is fixed and a linear classifier on top of it is trained using a limited number of labeled samples or *shots*, testing the effectiveness of learned representations. Table 1 shows that CAPC consistently outperforms other methods across both datasets. Notably, using dual view augmentation in CAPC leads to a significant performance boost compared to using other mask augmentations. This improvement is especially prominent in transfer learning, demonstrating CAPC's ability to capture more generalizable and task-agnostic features. This evaluation underscores the superiority of CAPC with and without dual view for time-series CSI data. It also shows that leveraging wireless channel characteristics, particularly through dual view augmentation, offers a notable performance boost for WiFi sensing tasks.

We investigated five augmentations, applied individually and in pairs, across our CAPC model and various baselines to identify the best combination for each method. The average performance of each augmentation is summarized in Table 2, with full details of the experiments and augmentation combinations provided in Appendix E. Across all SSL methods, noise augmentation consistently enhanced model generalization, yielding the highest average performance. Following noise, CAPC demonstrated a significant advantage with dual view augmentation, while the baselines struggled to fully utilize it. This highlights CAPC's unique ability to handle stronger augmentations. SimCLR outperformed CAPC in the time mask scenario, likely because CAPC's temporal prediction component may be affected by masking. Despite this, CAPC excelled in most augmentation types, showcasing its robustness and adaptability to various transformations.

4 Conclusion

In this paper, we proposed CAPC, a representation learning framework for CSI data and WiFi sensing, featuring a time-series-specific architecture. Specifically, we employed a loss function that combines future prediction and embedding consistency pretext tasks for SSL, ensuring the generated representations are both temporally informative and robust to data distortions inherent in downstream tasks. We also developed a novel augmentation technique to reduce electronic distortions from transceivers and isolate free space propagation effects on the channel. Extensive cross-domain experiments demonstrated that CAPC, with and without dual view augmentation, surpasses baseline SSL methods in downstream HAR and gesture recognition tasks, particularly in few-shot scenarios. Furthermore, using the novel dual view augmentation significantly boosts CAPC's performance. Future works may explore other architectural designs such as the attention mechanism [26] and Fourier Transform based STFNet blocks [27], integrating other modalities like vision, [28, 29, 30] or CSI streams from multiple WiFi devices [31], as well as considering more complex tasks such as multi-user sensing [32].

‡Refers to the CAPC model being trained with the second best combination of augmentations, noise and subcarrier mask, instead of noise and dual view.

References

[1] Steven M. Hernandez and Eyuphan Bulut. Wifi sensing on the edge: Signal processing techniques and challenges for real-world systems. *IEEE Commun. Surveys & Tutorials*, 25(1): 46–76, 2023. doi: 10.1109/COMST.2022.3209144.

[2] Rui Zhou, Meng Hao, Xiang Lu, Mingjie Tang, and Yang Fu. Device-free localization based on CSI fingerprints and deep neural networks. In *2018 15th Annual IEEE Intl. Conf. on Sensing, Commun., and Networking (SECON)*, pages 1–9, 2018. doi: 10.1109/SAHCN.2018.8397121.

[3] Zhenguo Shi, Qingqing Cheng, J Andrew Zhang, and Richard Yi Da Xu. Environment-robust wifi-based human activity recognition using enhanced CSI and deep learning. *IEEE Internet of Things Jrnl.*, 9(24):24643–24654, 2022.

[4] Yanchao Zhao, Ran Gao, Shangqing Liu, Lei Xie, Jie Wu, Huawei Tu, and Bing Chen. Device-free secure interaction with hand gestures in wifi-enabled iot environment. *IEEE Internet of Things Jrnl.*, 8(7):5619–5631, 2021. doi: 10.1109/JIOT.2020.3032623.

[5] Qifan Pu, Sidhant Gupta, Shyamnath Gollakota, and Shwetak Patel. Whole-home gesture recognition using wireless signals. In *Proc. of the 19th Annual Intl. Conf. on Mobile Comput. & Networking*, pages 27–38, 2013.

[6] Kai Niu, Fusang Zhang, Xuanzhi Wang, Qin Lv, Haitong Luo, and Daqing Zhang. Understanding wifi signal frequency features for position-independent gesture sensing. *IEEE Trans. on Mobile Comput.*, 21(11):4156–4171, 2022. doi: 10.1109/TMC.2021.3063135.

[7] Youwei Zeng, Dan Wu, Jie Xiong, Enze Yi, Ruiyang Gao, and Daqing Zhang. Farsense: Pushing the range limit of wifi-based respiration sensing with CSI ratio of two antennas. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 3(3), sep 2019. doi: 10.1145/3351279. URL `https://doi.org/10.1145/3351279`.

[8] Jie Zhang, Zhanyong Tang, Meng Li, Dingyi Fang, Petteri Nurmi, and Zheng Wang. Crosssense: Towards cross-site and large-scale wifi sensing. In *Proc. of the 24th Annual Intl. Conf. on Mobile Comput. and Networking*, MobiCom '18, page 305–320, New York, NY, USA, Oct 2018. Association for Comput. Machinery. ISBN 978-1-4503-5903-0. doi: 10.1145/3241539.3241570. URL `https://doi.org/10.1145/3241539.3241570`.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Intl. Conf. on Machine Learning*, pages 1597–1607. PMLR, 2020.

[10] Jianfei Yang, Xinyan Chen, Han Zou, Dazhuo Wang, and Lihua Xie. Autofi: Towards automatic wifi human sensing via geometric self-supervised learning. *IEEE Internet of Things Jrnl.*, 2022.

[11] Li-Hsiang Shen, Kai-Jui Chen, An-Hung Hsiao, and Kai-Ten Feng. Bts: Bifold teacher-student in semi-supervised learning for indoor two-room presence detection under time-varying CSI, 2023.

[12] Sijie Ji, Yaxiong Xie, and Mo Li. Sifall: Practical online fall detection with rf sensing. In *Proc. of the 20th ACM Conf. on Embedded Networked Sensor Systems*, pages 563–577, 2022.

[13] Ruiyuan Song, Dongheng Zhang, Zhi Wu, Cong Yu, Chunyang Xie, Shuai Yang, Yang Hu, and Yan Chen. Rf-url: Unsupervised representation learning for rf sensing. In *Proc. of the 28th Annual Intl. Conf. on Mobile Comput. And Networking*, MobiCom '22, page 282–295, New York, NY, USA, 2022. Association for Comput. Machinery. ISBN 9781450391818. doi: 10.1145/3495243.3560529. URL `https://doi.org/10.1145/3495243.3560529`.

[14] Dongxin Liu, Tianshi Wang, Shengzhong Liu, Ruijie Wang, Shuochao Yao, and Tarek Abdelzaher. Contrastive self-supervised representation learning for sensing signals from the time-frequency perspective. In *2021 Intl. Conf. on Computer Commun. and Networks (ICCCN)*, pages 1–10, 2021. doi: 10.1109/ICCCN52240.2021.9522151.

[15] Ke Xu, Jiangtao Wang, Le Zhang, Hongyuan Zhu, and Dingchang Zheng. Dual-stream contrastive learning for channel state information based human activity recognition. *IEEE Jrnl of Biomedical and Health Informatics*, 27(1):329–338, 2023. doi: 10.1109/JBHI.2022.3219640.

[16] Mohammud J. Bocus, Hok-Shing Lau, Ryan McConville, Robert J. Piechocki, and Raul Santos-Rodriguez. Self-supervised wifi-based activity recognition. In *2022 IEEE Globecom Workshops (GC Wkshps)*, pages 552–557, 2022. doi: 10.1109/GCWkshps56602.2022.10008537.

[17] Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi. Unsupervised scalable representation learning for multivariate time series, 2020.

[18] Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. Self-supervised contrastive representation learning for semi-supervised time-series classification. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 45(12): 15604–15618, 2023. doi: 10.1109/TPAMI.2023.3308189.

[19] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[20] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction, 2021.

[21] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning, 2022.

[22] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[23] François Rottenberg, Rui Wang, Jianzhong Zhang, and Andreas F Molisch. Channel extrapolation in fdd massive mimo: Theoretical analysis and numerical validation. In *2019 IEEE Global Commun. Conf. (GLOBECOM)*, pages 1–7. IEEE, 2019.

[24] Yongsen Ma, Gang Zhou, Shuangquan Wang, Hongyang Zhao, and Woosub Jung. Signfi: Sign language recognition using wifi. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(1), mar 2018. doi: 10.1145/3191755. URL `https://doi.org/10.1145/3191755`.

[25] Siamak Yousefi, Hirokazu Narui, Sankalp Dayal, Stefano Ermon, and Shahrokh Valaee. A survey on behavior recognition using wifi channel state information. *IEEE Commun. Magazine*, 55(10):98–104, 2017. doi: 10.1109/MCOM.2017.1700082.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[27] Shuochao Yao, Ailing Piao, Wenjun Jiang, Yiran Zhao, Huajie Shao, Shengzhong Liu, Dongxin Liu, Jinyang Li, Tianshi Wang, Shaohan Hu, Lu Su, Jiawei Han, and Tarek Abdelzaher. Stfnets: Learning sensing signals from the time-frequency perspective with short-time fourier neural networks. In *The World Wide Web Conf.*, WWW '19, page 2192–2202, New York, NY, USA, 2019. Association for Comput. Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313426. URL `https://doi.org/10.1145/3308558.3313426`.

[28] Jianfei Yang, Shijie Tang, Yuecong Xu, Yunjiao Zhou, and Lihua Xie. Maskfi: Unsupervised learning of wifi and vision representations for multimodal human activity recognition. *arXiv preprint arXiv:2402.19258*, 2024.

[29] Lang Deng, Jianfei Yang, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. Gaitfi: Robust device-free human identification via wifi and vision multimodal learning. *IEEE Internet of Things Jrnl.*, 10(1):625–636, 2022.

[30] Jianfei Yang, He Huang, Yunjiao Zhou, Xinyan Chen, Yuecong Xu, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing. *Advances in Neural Information Processing Systems*, 36, 2024.

[31] Zheng Yang, Yi Zhang, Guidong Zhang, and Yue Zheng. Widar 3.0: Wifi-based activity recognition dataset, 2020. URL `https://dx.doi.org/10.21227/7znf-qp86`.

[32] Shuokang Huang, Kaihan Li, Di You, Yichong Chen, Arvin Lin, Siying Liu, Xiaohui Li, and Julie A McCann. Wimans: A benchmark dataset for wifi-based multi-user activity sensing. *arXiv preprint arXiv:2402.09430*, 2024.

[33] Chunjing Xiao, Daojun Han, Yongsen Ma, and Zhiguang Qin. CSIgan: Robust channel state information-based activity recognition with gans. *IEEE Internet of Things Jrnl.*, 6(6): 10191–10204, 2019.

[34] Yongsen Ma, Gang Zhou, and Shuangquan Wang. Wifi sensing with channel state information: A survey. *ACM Comput. Surv.*, 52(3), jun 2019. ISSN 0360-0300. doi: 10.1145/3310194. URL `https://doi.org/10.1145/3310194`.

[35] Andrii Zhuravchak, Oleg Kapshii, and Evangelos Pournaras. Human activity recognition based on WiFi CSI data - a deep neural network approach. *Procedia Computer Science*, 198:59–66, 2022.

[36] Zhenghua Chen, Le Zhang, Chaoyang Jiang, Zhiguang Cao, and Wei Cui. Wifi CSI based passive human activity recognition using attention based blstm. *IEEE Trans. on Mobile Comput.*, 18(11):2714–2724, 2019. doi: 10.1109/TMC.2018.2878233.

[37] Chunjing Xiao, Yue Lei, Yongsen Ma, Fan Zhou, and Zhiguang Qin. Deepseg: Deep-learning-based activity segmentation framework for activity recognition using wifi. *IEEE Internet of Things Jrnl.*, 8(7):5669–5681, 2021. doi: 10.1109/JIOT.2020.3033173.

[38] Borna Barahimi, Hakam Singh, Hina Tabassum, Omer Waqar, and Mohammad Omer. Rscnet: Dynamic CSI compression for cloud-based wifi sensing. In *ICC 2024 - IEEE International Conf. on Commun.*, pages 4179–4184. IEEE, 2024.

[39] Bing Li, Wei Cui, Wei Wang, Le Zhang, Zhenghua Chen, and Min Wu. Two-stream convolution augmented transformer for human activity recognition. In *Proc. of the AAAI Conf. on Artificial Intelligence*, volume 35, pages 286–293, 2021.

[40] Yu Gu, Xiang Zhang, Yantong Wang, Meng Wang, Huan Yan, Yusheng Ji, Zhi Liu, Jianhua Li, and Mianxiong Dong. WiGRUNT: WiFi-enabled gesture recognition using dual-attention network. *IEEE Trans. on Human-Machine Systems*, 52(4):736–746, 2022. doi: 10.1109/THMS.2022.3163189.

[41] Yi Zhang, Yue Zheng, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. Widar3.0: Zero-effort cross-domain gesture recognition with wi-fi. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 44(11):8671–8688, 2022. doi: 10.1109/TPAMI.2021.3105387.

[42] Han Zou, Jianfei Yang, Yuxun Zhou, Lihua Xie, and Costas J. Spanos. Robust wifi-enabled device-free gesture recognition via unsupervised adversarial domain adaptation. In *2018 27th Intl. Conf. on Computer Commun. and Networks (ICCCN)*, pages 1–8, 2018. doi: 10.1109/ICCCN.2018.8487345.

[43] Xi Chen, Hang Li, Chenyi Zhou, Xue Liu, Di Wu, and Gregory Dudek. Fidora: Robust wifi-based indoor localization via unsupervised domain adaptation. *IEEE Internet of Things Jrnl.*, 9 (12):9872–9888, 2022. doi: 10.1109/JIOT.2022.3163391.

[44] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[45] Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. *Advances in Neural Inf. Processing Systems*, 35:26671–26685, 2022.

[46] Chunjing Xiao, Yanhui Han, Wei Yang, Yane Hou, Fangzhan Shi, and Kevin Chetty. Diffusion model-based contrastive learning for human activity recognition. *IEEE Internet of Things Jrnl.*, pages 1–1, 2024. doi: 10.1109/JIOT.2024.3429245.

[47] Daniel Halperin, Wenjun Hu, Anmol Sheth, and David Wetherall. Tool release: Gathering 802.11n traces with channel state information. *SIGCOMM Comput. Commun. Rev.*, 41(1):53, jan 2011. ISSN 0146-4833. doi: 10.1145/1925861.1925870. URL `https://doi.org/10.1145/1925861.1925870`.

[48] Yang You, Igor Gitman, and Boris Ginsburg. Scaling SGD batch size to 32k for imagenet training. *CoRR*, abs/1708.03888, 2017. URL `http://arxiv.org/abs/1708.03888`.

[49] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with restarts. *CoRR*, abs/1608.03983, 2016. URL `http://arxiv.org/abs/1608.03983`.

[50] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[51] Mostafa Neo Mohsenvand, Mohammad Rasool Izadi, and Pattie Maes. Contrastive representation learning for electroencephalogram classification. In Emily Alsentzer, Matthew B. A. McDermott, Fabian Falck, Suproteem K. Sarkar, Subhrajit Roy, and Stephanie L. Hyland, editors, *Proc. of the Machine Learning for Health NeurIPS Workshop*, volume

136 of *Proc. of Machine Learning Research*, pages 238–253. PMLR, 11 Dec 2020. URL https://proceedings.mlr.press/v136/mohsenvand20a.html.

[52] Kemal Davaslioglu, Serdar Boztaş, Mehmet Can Ertem, Yalin E. Sagduyu, and Ender Ayanoglu. Self-supervised rf signal representation learning for nextg signal classification with deep learning. *IEEE Wireless Commun. Letters*, 12(1):65–69, 2023. doi: 10.1109/LWC.2022.3217292.

[53] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

[54] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Jrnl. of Machine Learning Research*, 9(11), 2008.

A Related Work

A.1 Supervised Learning for WiFi Sensing

Recent advances in deep learning enabled sensing have demonstrated remarkable potential in extracting and modeling complex patterns from CSI without intensive feature engineering [33, 34]. Methods based on Long Short-Term Memory (LSTM) networks were used to extract temporal activity-related information from CSI data [35], [36] and achieved state-of-the-art accuracy at the time. Convolutional Neural Networks (CNNs), commonly used to extract spatial features, have been used to segment CSI based on activities [37] and compress CSI data for cloud-based sensing [38]. Additionally, THAT [39] proposed a two-stream Transformer-based model to utilize both time-over-channel and channel-over-time features and out-performed LSTM and CNN based methods. ResNet architectures with attention mechanisms for gesture recognition [40] were proposed as well for cross-domain HAR. Notably, some studies have also addressed the environmental dependencies of deep learning methods. Widar3.0 [41] used a hand-crafted environment-independent feature called body-coordinate velocity profile (BVP) as the input of their deep learning model and AFEE-MatNet [3] applied a matching network to learn common features among environments.

A.2 Semi-supervised learning for WiFi Sensing

Nevertheless, the aforementioned research works follow a supervised approach, thus necessitating large volumes of labelled CSI data. A fundamental challenge is to train the model without labels, while enabling generalization across different environment settings. Recently, a handful of research works considered semi-supervised learning to reduce the reliance on labelled datasets. For instance, in WiADG [42], where a pre-trained encoder, initially trained in a supervised manner, is fine-tuned for new environments using adversarial networks. This semi-supervised approach can adapt to new settings without the additional labelled data in the target environment, though it is contingent on having pre-labelled data for the source environment. Fidora [43] used a semi-supervised approach by employing Variational Auto Encoders (VAEs) to augment the labelled dataset. Subsequently, a feature extractor is applied to generate a representation from labelled and unlabelled samples. These representations are then processed through a decoder and classifier for CSI reconstruction and classification, respectively. SiFall [12] also leveraged VAEs in conjunction with anomaly detection techniques for fall detection. BTS [11] proposed a semi-supervised teacher-student learning approach inspired by BYOL [44] to solve time-varying effects induced by environment changes.

A.3 Self-supervised learning for WiFi Sensing

SSL is emerging as a powerful technique that leverages unlabelled data to train deep learning models, generating effective representations without the need for explicit labels. However, it requires knowledge of what makes some samples semantically close to others [45]. In the initial phase, the model undergoes pre-training using unlabelled dataset, thereby generating useful representations from the data without relying on explicit labels. In the subsequent stage, the model shifts to supervised learning using a limited labelled target dataset. During this phase, fine-tuning takes place, exploiting the previously learned representations to improve the performance.

Recent SSL-based WiFi sensing methods, such as RF-URL[13], Lau et al. [16], STF-CSL[14], DualConFi[15], and CLAR[46], learn invariant representations by contrasting or aligning different views of the input samples, which are created through various augmentation techniques. These

methods focus on creating representations that minimize the distance between similar instances in an embedding space. RF-URL used Doppler-frequency-spectrum, Angle of Arrival, and Time of Flight augmentations as well as InfoNCE contrastive loss function. STF-CSL integrated Short-Time-Fourier-Neural Networks (STFNets) in their encoders, with a variety of frequency and time-domain augmentations. In [16], different views of CSI corresponding to a specific activity (or sample) captured by several receivers placed at various locations are considered positive samples, while views of other unrelated samples are treated as negative samples. CLAR, a more recent work, utilizes diffusion models to generate augmented samples. Except for RF-URL, these methods adopt a contrastive learning strategy akin to SimCLR [9], using the NT-Xent loss function. More recently, AutoFi[10] has emerged with a non-contrastive geometric SSL method and few-shot learning for WiFi sensing.

However, the aforementioned research overlooks the temporal aspects of CSI, as well as wireless propagation channel, and transceiver characteristics. Additionally, the considered augmentation techniques often do not align well with the inherent nature of CSI data. Furthermore, recent advancements in non-contrastive methods [20, 21], which eliminate the need for negative samples, have not been explored in the WiFi sensing domain.

B Datasets

B.1 SignFi

We employed the SignFi gesture recognition dataset [24], specifically designed for sign language gesture recognition tasks. This dataset features a significant volume of data instances but has a limited number of samples per class due to its extensive variety of sign language words (classes), totalling 276. The SignFi dataset, acquired through the Intel 5300 NIC [47] in a single-transmitter single-receiver setup, consists of 3 antennas at the receiver, a single antenna for the transmitter, 30 subcarriers, and 200 packets per data instance, resulting in each sample possessing $3 \times 30 \times 200$ dimensions. Its key attributes include:
(1) Multiple environment setups: SignFi dataset comes from two different environments: a home and a lab. This variety helps us test how well our method adapts to new environments. We used the lab dataset as the unlabelled dataset for the SSL pre-training as it contained more samples and the home dataset with labels for the supervised evaluation.
(2) Dual view CSI: SignFi dataset includes synchronized uplink and downlink CSI for each sample in the dataset allowing us to utilize them for dual view augmentation in SSL pre-training. To the best of our knowledge, SignFi is the only database offering synchronized uplink and downlink CSI.
(3) Substantial Sample Volume: SignFi dataset comprises 5520 instances from the lab environment (20 samples per class) and 2760 instances from the home environment (10 samples per class), making it a notably large dataset.
(4) User Diversity: Lab environment samples are collected independently for each of the 5 users participating in the experiments, meaning that at any given time, only one person is performing an activity in the room. In contrast, home environment samples are collected from a single user.

B.2 UT HAR

We also employed the UT HAR dataset [25] for evaluating transfer learning. Specifically, we tested the RSCNet backbone encoders, which were pre-trained using unlabelled CSI data in the SignFi lab, on a subset of labelled samples from UT HAR. These experiments aimed to assess the effectiveness of the CAPC representations for adapting to new tasks and environments.

The UT HAR data was collected using the Intel 5300 NIC [47], similar to the setup used in SignFi. The data samples have dimensions of 3 antenna links, 30 subcarriers, and 250 packets. Given that the antenna and subcarrier dimensions align with those of SignFi, and considering our use of a windowing size $N_f = 10$ in the encoders for both CAPC and all baseline models, we were able to use the same pre-trained encoders from SignFi without requiring additional preprocessing.

Additionally, this dataset, which consists of samples all collected from one user, categorizes human activities into seven types: lying down, falling, walking, running, sitting down, standing up, and empty room. The choice to use this dataset solely for fine-tuning and transfer learning purposes is due to its relatively small size, with only 3,977 samples in the training set.

B.3 Evaluation Criteria

Our evaluation criteria shown in Figure 2 are designed to assess how effectively a pre-trained encoder with SSL adapts to downstream WiFi sensing tasks. Specifically, we aim to understand the applicability of the encoder's representations when confronted with a limited number of labelled samples from the downstream task.

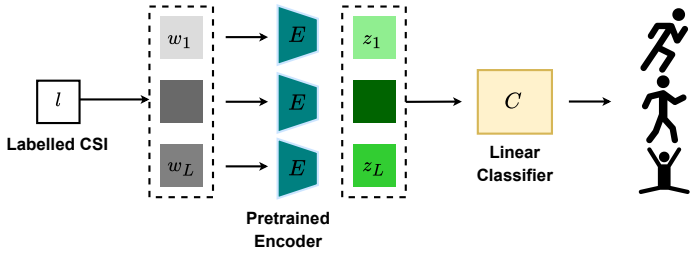


Figure 2: **Supervised evaluation:** A linear classifier $C_{\boldsymbol{\phi}}$ is fine-tuned with labelled CSI based on the concatenated representations from all windows generated by the pre-trained encoder $E_{\boldsymbol{\theta^A}}$. The pre-trained encoder's weights $\boldsymbol{\theta^A}$ are frozen in linear classification.

We analyze if the representations produced by the encoder are sufficient and generalizable for the downstream tasks without additional data-specific training. The encoder's weights, denoted as $\boldsymbol{\theta^A}$, are frozen to prevent the infusion of task-specific information. The latent representations, generated by the encoder for each data window, are concatenated and used as input for a linear classifier $C_{\boldsymbol{\phi}}$. This classifier is trained using the cross-entropy loss function, defined by:

$$\mathcal{L}_c(\boldsymbol{x}, \boldsymbol{y}) = -\mathbb{E}_{(\boldsymbol{x},y)} \sum_{i=1}^{n} \left[\mathbb{I}[y=i] \log \left(C_{\boldsymbol{\phi}} \left(E_{\boldsymbol{\theta^A}}(\boldsymbol{w_{1,\ldots,L}})\right)\right)\right], \tag{5}$$

Here, n is the number of class labels, and $E_{\boldsymbol{\theta^A}}(\boldsymbol{w_{1,\ldots,L}})$ represents the merged embeddings generated by the encoder across all CSI windows of each sample i.e. $\{\boldsymbol{w_t}\}_1^L$. This setup tests the hypothesis of how a well-trained encoder can enable a linear classifier to perform effectively on downstream tasks without the additional information of the task in the representations.

C Baselines

To thoroughly evaluate the effectiveness of our proposed method, CAPC, we conducted comparisons with four established SSL methods. Specifically, we compared CAPC against: *(1) SimCLR* [9], which serves as a standard contrastive SSL benchmark; *(2) Barlow Twins* [20] and *(3) CPC* [22], both included as ablation studies due to their integration within CAPC's framework; *(4) AutoFi* [10], a recent SSL method tailored for CSI WiFi sensing. Furthermore, we compare CAPC against a fully supervised training model where the encoder $E_{\boldsymbol{\theta}}$ is randomly initialized and trained with the labelled data from scratch. This comparison is intended to highlight the performance gains of SSL even with limited training labels.

Remark: To ensure a fair comparison, all methods utilized the same backbone encoder architecture, namely the RSCNet encoder [38]. Additionally, unlike CAPC and CPC, which employ an autoregressive model during pre-training, SimCLR, Barlow Twins, and AutoFi use a projector to independently map each window's representation into an embedding space for applying their respective loss functions. Specifically, we adopted the Barlow Twins' projector configuration, which consists of three fully connected layers paired with ReLU activation functions. By using the same backbone encoder across all methods, we maintained consistent model complexity, as detailed by the FLOP counts in Figure 5c. This consistency ensures that any differences in representation quality are solely attributable to SSL methodology and augmentations rather than model complexity, thus guaranteeing a fair and balanced comparison among CAPC and the baseline methods.

D Training Configuration

Our model, developed using PyTorch, employs the LARS optimizer during the SSL phase [48]. The training lasts for 300 epochs with a batch size of 128. We initialize the learning rate for weights at 0.2 and for biases and batch normalization parameters at 0.0048. The initial 10 epochs act as a warm-up period for the learning rate, which is then reduced following a cosine decay schedule [49]. The weight decay is configured at 1.5×10^{-6}. The trade-off parameter of Barlow Twins loss is set to $\lambda = 0.002$ and the trade-off parameter of CAPC β is set to 50 to scale the loss terms. These configurations largely follow those reported in [20]. In our model, unlike [20], the weights between the twin networks are not shared to enhance performance. Regarding the number of future windows prediction T in CAPC and CPC, we chose 9 for CAPC and 2 for CPC, as these values show better performance for each method as shown in Figure 5c.

During the evaluation phase, we switch to the Adam optimizer [50], maintaining a batch size of 512. For linear evaluation, the learning rate starts at 10^{-2} and follows a cosine decay schedule over 100 epochs, training only the linear classifier.

For the model architecture, we use a window size $N_f = 10$, an encoder embedding size $D = 128$, and 128 nodes in the GRU autoregressive model's hidden layer, the projection hidden layers and the embedding size h. The number of nodes in the hidden layer of the linear classifier C_ϕ is half of its input size, which equals the embedding size D multiplied by the sequence length L.

E Augmentations

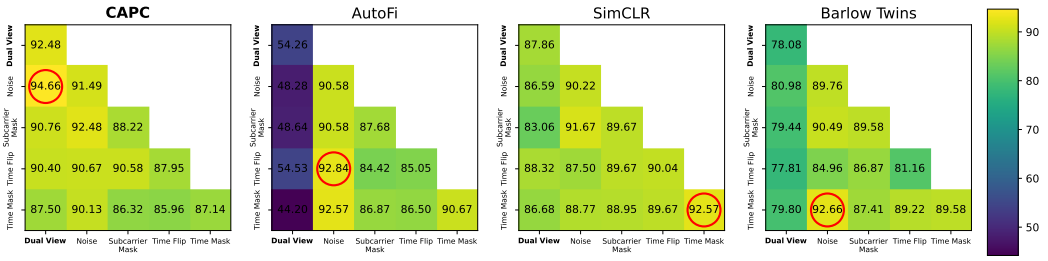


Figure 3: Linear evaluation of individual and compositional data augmentations. Each diagonal element represents the effect of a single transformation, while off-diagonal elements illustrate the combined impact of two sequentially applied transformations. We report the accuracies (in %) with 6 shots in the labelled dataset. Red circles indicate the best combination of augmentations.

In addition to the proposed dual view augmentation, we evaluated several common augmentations used in the context of time-series [51, 52] and WiFi sensing [14]:

- **Gaussian Noise:** introduces random Gaussian noise with zero mean and 0.1 standard deviation into data, simulating inherent noise in CSI, and enhancing model resilience.
- **Time Flip:** flips CSI samples along the time axis to accommodate palindromic activities, where time-reversed patterns represent the same activity.
- **Time Mask:** randomly masks a segment of the time dimension in each CSI window, varying its location. This teaches the model to predict missing temporal information, simulating scenarios with temporal disruptions or signal losses.
- **Subcarrier Mask:** masks a set of subcarriers in each CSI sample, forcing the model to infer activities using the remaining subcarriers. This augmentation improves the handling of frequency-selective fading or interference and underscores different subcarriers' significance in activity recognition.

Choosing best augmentations for each method: We conduct a comprehensive analysis of the aforementioned augmentations, both individually and in combination, on the performance of our method and the baselines. This allowed us to identify the optimal augmentation mix for each method and to demonstrate the robustness of each method in response to the various augmentations. The augmentations chosen for each method are in Figure 3. For CAPC, the chosen augmentations are dual view and noise. For SimCLR, it is time mask. The Barlow Twins use noise and time mask, while AutoFi employs noise and time flip. CPC relies solely on prediction for its SSL task, thus not using any augmentations.

Remarks: In experiments where the dual view augmentation was not applied, we treated the uplink and downlink CSI as separate samples, effectively doubling the dataset size during the SSL phase. This approach highlights the flexibility of our framework, showing that it is not limited to using dual view augmentation and can still operate effectively with either uplink or downlink CSI samples. Unlike dual view augmentation, which requires both CSI samples, CAPC can function effectively without needing both.

F Comparative analysis of contextual loss

One of the key novelties of our CAPC method is the integration of hybrid temporal and contextual losses. The use of prediction and autoregressive models to capture temporal dependencies is well-recognized. However, the rationale behind choosing a specific type of contextual loss might not be immediately apparent as various contrastive [9] and non-contrastive methods [20], [21], [10] are available. Our investigation shown in Figure 4 included three distinct loss functions: Barlow Twins, as utilized in our framework, along with SimCLR, and AutoFi. Our findings indicate that the non-contrastive loss from Barlow Twins consistently outperforms both the contrastive loss used in CAPC with SimCLR and the loss of AutoFi across various training scenarios, often by a significant margin. For example, in a scenario with four shots, CAPC equipped with Barlow Twins achieved an accuracy of 88.50%, markedly higher than the 79.35% and 75.63% seen with AutoFi and SimCLR, respectively. Additionally, AutoFi consistently outperforms SimCLR by an average of 2%. These results suggest that non-contrastive loss functions serve as more effective contextual losses for our proposed CAPC method.

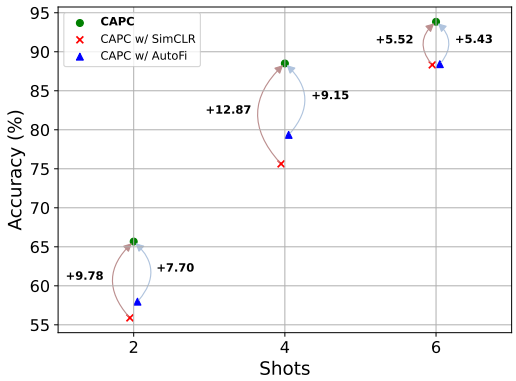
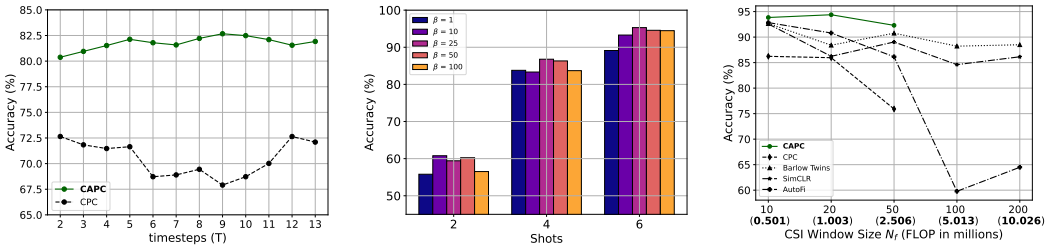


Figure 4: A comparative study of the proposed CAPC method. The CAPC w/ SimCLR and AutoFi mean that we have replaced the Barlow Twins loss in our design with SimCLR and AutoFi, respectively. Showcasing that Barlow Twins has superior performance for enforcing context embedding consistency. We report the experiments under linear evaluation of SignFi Home dataset with 2, 4, and 6 shots.

G Hyperparameter selection and sensitivity analysis

We conducted sensitivity analyses on three primary hyperparameters within the CAPC framework: (1) we illustrate the impact of the number of future windows or timesteps, T, predicted by CAPC during SSL, on linear evaluation; (2) we demonstrate the effect of the trade-off parameter, β, which represents the weight of $\mathcal{L}_{CPC}$ in the loss function; (3) we depict the influence of the number of frames per CSI window, N_f, and encoder complexity comparing CAPC to baseline SSL methods.

The analysis in Figure 5a reveals the impact of T on the performance of CAPC and CPC. For both frameworks, we present the linear evaluation performance averaged across 2, 4, and 6 shot scenarios in SignFi Home. Both methods exhibit fluctuations; however, CAPC demonstrates greater robustness and fewer fluctuations across different T values, outperforming CPC by approximately 11.3% on average. Generally, CAPC achieved better results with higher T values, particularly at $T = 9$. Conversely, CPC's performance initially declined with increasing T values but improved



(a) (b) (c)

Figure 5: Sensitivity analysis experiments on SignFi dataset: (a) Illustrates how the accuracy of linear evaluation is affected by varying the number of predicted future windows (T) for 2, 4, and 6 samples per class, highlighting that CAPC is significantly more stable across different values of T compared to CPC; (b) Depicts the influence of the coefficient β on CAPC's performance under linear evaluation; (c) Examines the impact of different window sizes, N_f, on the accuracy of CAPC and baseline methods during linear evaluation for 6 samples per class. It also shows the computational complexity of the encoder with varying window sizes. Here, $T = 2$ for both CAPC and CPC due to constraints imposed by the limited number of windows at higher window sizes ($T \leq L - 2$).

upon reaching $T = 12$ and $T = 13$, with $T = 2$ being the optimal value. This implies that CAPC may be more effective at capturing common information over an extended number of windows.

We further investigated the sensitivity of CAPC to the hyperparameter β, which balances the significance of temporal and contextual consistency in the embeddings. We found that CAPC is relatively insensitive to variations in β, with values of 25 and 50 both demonstrating slightly improved performance overall, as illustrated in Figure 5b.

We extensively examined SSL with different window sizes, N_f, for CAPC and all baselines within the flexible RSCNet framework, as shown in Figure 5c. We specifically evaluated window sizes of 10, 20, and 50 for CAPC and CPC because larger N_f values reduce the sequence length L, and both CPC and CAPC require $L \geq 3$ to predict future windows effectively. For other methods, we extended our examination to include window sizes of 100 and 200, corresponding to the complete sequence of the CSI in SignFi. Our results indicate that our proposed CAPC, along with Barlow Twins and SimCLR, demonstrated robust performance across different N_f values, with CAPC showing superior accuracy in all tested cases. In contrast, AutoFi and CPC experienced significant performance declines as the number of windows increased.

This suggests that CSI segmentation is a viable approach for SSL in WiFi sensing, benefiting methods beyond our proposed model. Not only did all methods generally perform better with lower N_f values, but, as Figure 5c illustrates, these lower values also simplify the complexity of the encoder E due to smaller input sizes. This reduction enhances efficiency and reduces computational demands, making it particularly suitable for resource-limited edge devices where these models are deployed. Based on these findings, we selected $N_f = 10$ for subsequent experiments, as it generally yielded strong performance across all methods.

H Collapse analysis

To demonstrate that our method does not suffer from dimensional collapse, we present the singular value spectrum of the embedding space (Figure 6) of the pre-trained encoder of CAPC and other baselines. Complete collapse occurs when all singular values fall to zero, indicating that the representation has become constant and carry no useful information. If only some singular values drop to zero, it suggests partial dimensional collapse, where the encoder has not fully utilized the embedding space [53]. As depicted in Figure 6, no singular values fall to zero for any of the methods, confirming that none, including ours, undergo dimensional collapse.

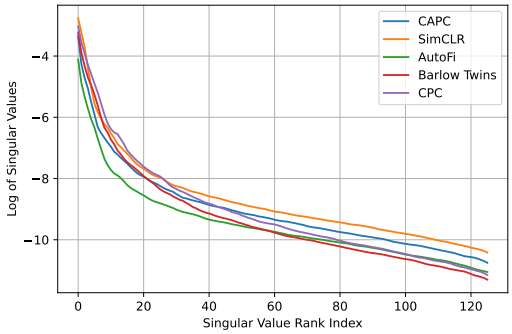


Figure 6: Singular value spectrum of the representation space (z) of CAPC compared to baselines on the SignFi Home dataset validation set. Each embedding vector is of size 128. The spectrum displays the singular values of the covariance matrix of these embedding vectors, sorted and plotted on a logarithmic scale. No singular values drop to zero, indicating that none of the methods, including ours, experience dimensional collapse.

I t-SNE visualization

In Figure 7, we present a t-SNE visualization [54] of the encoded CSI with the proposed CAPC pre-trained encoder, marked by the colored labels in the SignFi Home dataset. This visualization demonstrates the discriminative power of the embeddings for sign language recognition, effectively clustering similar gestures together. Despite the large number of labels and the lack of supervision or predefined labels during training, the encoder effectively segregates the data into distinct clusters. Moreover, the input data originates from an unseen environment, further underscoring the capability of CAPC to not only extract relevant discriminative features for downstream tasks but also to generalize effectively in new environments.

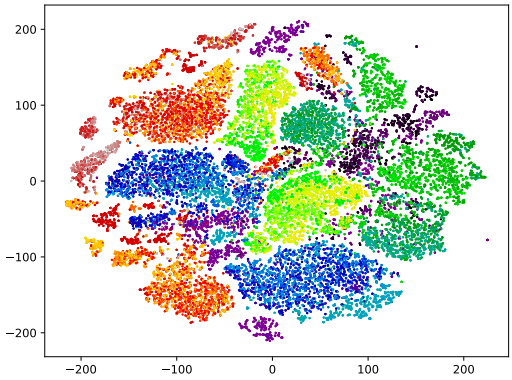


Figure 7: t-SNE visualization of SignFi Home dataset representations, trained using the CAPC SSL method on the SignFi Lab dataset. Each color corresponds to a distinct sign language label.