CAM-MIM: COHERENCE-AWARE MASKED RECONSTRUCTION FOR SELF-SUPERVISED WIFI HUMAN SENSING

Gang Liu[⋆]*, Yanling Hao*[⋆]*, Yixuan Zou*[⋆]*, Ziyu Zhou*[†]

[⋆] Queen Mary University of London
School of Electronic Engineering and Computer Science
[†] Imperial College London
Department of Electrical and Electronic Engineering

ABSTRACT

WiFi-based human sensing is severely limited by the scarcity of labeled data. Current self-supervised methods often employ random masking, inefficiently learning from predictable background signals instead of crucial motion-induced patterns. We propose CAM-MIM, a framework that introduces a coherence-aware masking strategy. By using 2D autocorrelation as a proxy for signal coherence, our method focuses reconstruction on the most informative low-coherence regions corresponding to human activity. Experiments show that under challenging low-label settings, CAM-MIM achieves 96.92% accuracy, substantially outperforming other self-supervised approaches and nearing fully supervised performance. Our work demonstrates that embedding physical priors into the learning objective can yield more powerful and transferable representations for WiFi sensing. Code will be made publicly available upon acceptance.

***Index Terms*—** WiFi sensing, Channel State Information, self-supervised learning, masked image modeling, coherence-aware masking

1. INTRODUCTION

WiFi-based human sensing has demonstrated extensive value in scenarios such as daily activity recognition, physiological signal monitoring, sleep and health assessment, and security assurance[1, 2, 3, 4]. Commercial WiFi devices can provide Channel State Information (CSI) across subcarriers and transmit/receive antennas. Human motion alters multipath propagation and introduces Doppler shifts, generating discernible patterns in the time-frequency domain of CSI, thereby enabling device-free human activity recognition (WiFi4HAR)[5].

Existing WiFi4HAR methods can be broadly categorized into two types: one involves modeling and feature engineering based on physical and statistical priors (e.g., modeling respiration from periodic phase peaks), combined with traditional classifiers for recognition[6, 7]; the other consists of end-to-end deep learning methods that directly learn discriminative representations from raw or preprocessed CSI, reducing manual feature design and achieving superior performance on various tasks[8]. However, unlike vision or speech, activity labels for CSI can often only be collected in experimental settings with limited coverage. The high cost of data acquisition and the scarcity of labels have become a major bottleneck hindering the development of WiFi4HAR[9].

Self-Supervised Learning (SSL) offers a viable path to alleviate the problem of insufficient labels and has made significant progress in fields like vision[10]. In the domain of WiFi sensing, previous works have explored three main paradigms: contrastive (learning by pulling together representations of augmented views through a consistency constraint)[11], predictive (learning temporal dynamics by predicting future/masked segments)[12], and generative-reconstructive (capturing intrinsic priors by reconstructing the input structure)[13]. Although reconstructive strategies have shown good potential on CSI[14, 15], existing approaches mostly employ random masking, ignoring the physical heterogeneity of CSI: static environments (e.g., reflections from walls, furniture) exhibit slow variations and simple patterns in the local time-frequency domain, presenting a high-coherence background. In contrast, Doppler and multipath effects induced by human motion introduce rapidly changing, non-stationary, and low-coherence components[9, 16]. If masking is applied indiscriminately, the model may expend its capacity on easily predictable high-coherence regions, failing to adequately learn the dynamic signals that are most valuable for sensing.

To address this, this paper proposes a coherence-aware masked modeling framework that incorporates physical priors, named CAM-MIM (Coherence-Aware Masked Modeling). The core idea is to use 2D FFT-based autocorrelation to estimate coherence, construct an unpredictability score per patch, prioritize masking of low-coherence regions, and pre-train under a Masked Image Modeling (MIM) objective[17]. The main contributions are summarized as follows:

1) We propose the CAM-MIM framework, which prioritizes masking low-coherence time-frequency patches during MIM pre-training to enhance the modeling of motion-induced components.
2) We provide a concise and reproducible implementation using FFT/IFFT and patch-wise aggregation, with few tunable parameters and easy integration with convolutional encoders.
3) We demonstrate strong performance with CAM-MIM on public benchmarks under a low-label linear probing setting.

The remainder of this paper is organized as follows. Section II details the proposed CAM-MIM framework and its coherence-aware masking strategy. Section III presents the experimental setup, performance evaluation, and ablation studies. Finally, Section IV concludes the work and outlines future work.

2. METHOD

To overcome the limitations of random masking in prior work, CAM-MIM introduces a physically motivated, coherence-aware masking strategy. Its core novelty lies in utilizing 2D autocorrelation to identify and preferentially mask low-coherence regions that correspond to human motion. This approach compels the model to learn valuable dynamic features from the most informative parts of the CSI signal, rather than redundant background information. The complete pre-training pipeline is illustrated in Fig. 1, and the process unfolds in four key stages that align directly with the figure's

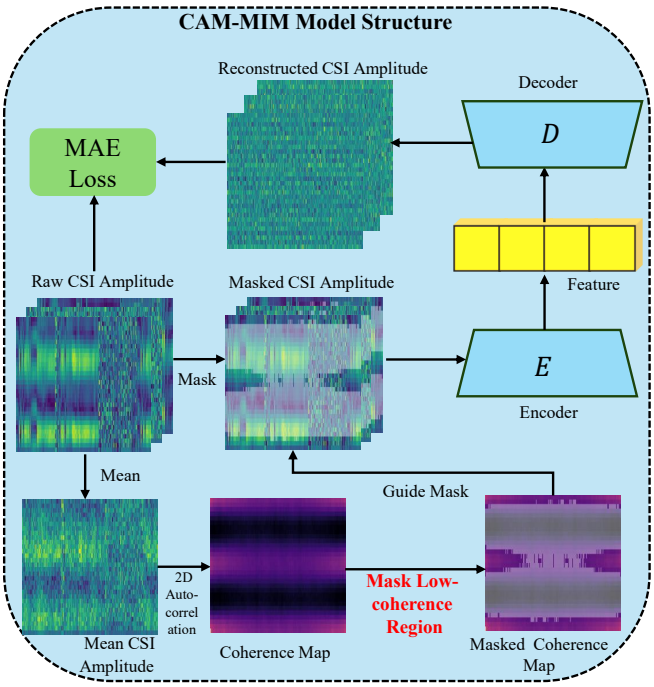


Fig. 1: Overview of CAM-MIM. A global 2D autocorrelation of antenna-averaged CSI provides a coherence proxy; a fixed mapping transfers it to the patch grid to form per-patch scores; low-coherence patches are masked and a convolutional encoder–decoder reconstructs them; the frozen encoder is evaluated by linear probing.

components: (1) computation of the coherence proxy from the mean CSI amplitude, (2) generation of the guide mask and application to low-coherence regions, (3) encoder-decoder reconstruction of masked regions, and (4) downstream evaluation via linear probing. These stages incorporate physical priors into the self-supervised objective, enabling efficient pre-training with minimal tunable parameters. Details are elaborated in the following subsections, with the pre-training procedure summarized in Algorithm 1.

2.1. Coherence Proxy Computation

We consider a WiFi sensing scenario based on MIMO-OFDM [18]. In a single collection period of duration T, the CSI can be represented as a 4D complex tensor $H \in \mathbb{C}^{N\times M\times S\times T}$, where N and M are the number of receive and transmit antennas, respectively, S is the number of subcarriers, and T is the number of time frames. In this work, we use the CSI amplitude tensor, retaining the antenna dimension, as the input:

$$X \in \mathbb{R}^{A\times S\times T}, \qquad X(a,f,t) = |H(a,f,t)|, \tag{1}$$

where A is the number of antenna/link channels.

To obtain a scalar time–frequency map for coherence estimation, we average across antennas:

$$X_{\text{mean}}(f,t) = \tfrac{1}{A}\sum_{a=1}^{A} X(a,f,t). \tag{2}$$

Let $\mathcal{F}\{\cdot\}$ denote the 2D Discrete Fourier Transform. By the Wiener–Khinchin theorem [19], the 2D autocorrelation operator acting on a time–frequency map Z can be written as

$$\mathcal{A}(Z) = \mathcal{F}^{-1}\left(|\mathcal{F}\{Z\}|^2\right), \tag{3}$$

Algorithm 1 CAM-MIM Pre-training

Require: CSI amplitude $X \in \mathbb{R}^{A\times S\times T}$, mask ratio ρ, Encoder f_θ, Decoder g_ϕ
Ensure: Optimized θ, ϕ

$X_{\text{mean}} \leftarrow \text{mean}(X, \text{dim} = A)$
Calculate Coherence Map by Mean Amplitude $\tilde{R} \leftarrow \text{clip}(\mathcal{F}^{-1}(|\mathcal{F}\{X_{\text{mean}}\}|^2)/R(0,0), 0, 1)$
for all patch P_k in X **do**
    Weight $w_k \leftarrow 1 - \text{mean}(\tilde{R}$ over lag window $\Omega_k)$
end for
Masked Patch $\mathcal{M}^{(A)} \leftarrow \text{SampleMask}(\{w_k\}, \rho)$
while not converged **do**
    Apply Mask $X_{\text{in}} \leftarrow (\mathbf{1} - \mathcal{M}^{(A)}) \odot X$
    Encoder and Reconstruct $X_{\text{rec}} \leftarrow g_\phi(f_\theta(X_{\text{in}}))$
    Compute Loss $\mathcal{L} \leftarrow \text{MaskedMAE}(X_{\text{rec}}, X, \mathcal{M}^{(A)})$
    Update parameters θ, ϕ using $\nabla\mathcal{L}$
end while
return Optimized parameters θ

and equivalently in the lag domain $(\Delta f, \Delta t)$ as

$$R_Z(\Delta f, \Delta t) = \sum_{f,t} Z(f,t)\, Z(f+\Delta f,\, t+\Delta t). \tag{4}$$

To align with the encoder's downsampling grid, we divide the (S,T) plane into non-overlapping patches $\mathcal{P} = \{P_k\}$ of size (p_s, p_t).

The coherence proxy (Coherence Map blocks in Fig. 1) is computed on X_{mean} as

$$\begin{aligned} R &= \mathcal{A}(X_{\text{mean}}), \\ \tilde{R}(\Delta f, \Delta t) &= \frac{\max\{0,\, R(\Delta f, \Delta t)\}}{R(0,0)} \in [0,1], \end{aligned} \tag{5}$$

where the nonnegative truncation avoids small negative values caused by numerical error and $\tilde{R}(0,0) = 1$. This proxy quantifies local signal predictability, with high values indicating stationary background and low values signaling motion-induced dynamics.

2.2. Guide Mask Generation and Low-Coherence Masking

Given the patch grid $\mathcal{P}$, we associate each patch P_k with a fixed lag-domain window Ω_k and define its coherence score by averaging $\tilde{R}$ over Ω_k:

$$r_k = \frac{1}{|\Omega_k|}\sum_{(\Delta f,\Delta t)\in\Omega_k} \tilde{R}(\Delta f, \Delta t), \qquad w_k = 1 - r_k. \tag{6}$$

We sample patches without replacement according to

$$p_k = \frac{(w_k + \varepsilon)^\tau}{\sum_{j\in\mathcal{P}}(w_j + \varepsilon)^\tau}, \tag{7}$$

where $\varepsilon > 0$ is a small constant, $\tau \geq 0$ is the sampling exponent (default $\tau = 1$), and the budget selects $m = \lceil \rho|\mathcal{P}| \rceil$ patches with a global mask ratio $\rho \in (0,1)$ (default $\rho = 0.85$). The sampled patch mask is expanded to a pixel-level mask $\mathcal{M} \in \{0,1\}^{S\times T}$ and broadcast across antennas to $\mathcal{M}^{(A)} \in \{0,1\}^{A\times S\times T}$, yielding the masked input ("Mask Low-coherence Region" and "Masked CSI Amplitude" blocks in Fig. 1)

$$X_{\text{in}} = (\mathbf{1} - \mathcal{M}^{(A)}) \odot X. \tag{8}$$

Intuition: Static background is locally smooth, producing high autocorrelation and low intrinsic uncertainty; motion-induced dynamics break local predictability and reduce autocorrelation. Under a local linear–Gaussian approximation, the conditional variance increases as coherence decreases. In 1D, for $X_t = \rho X_{t-1} + \varepsilon_t$ with $\mathrm{Var}(X_t) = \sigma_X^2$,

$$\mathrm{Var}(X_t \mid X_{t-1}) = \sigma_X^2(1-\rho^2), \tag{9}$$

which decreases monotonically in ρ. Mapping ρ to a patch-level r_k gives the analogous 2D behavior: smaller r_k implies larger $\mathrm{Var}(X_{\text{mask}} \mid X_{\text{vis}})$ inside P_k. Therefore, prioritizing low-coherence patches maximizes conditional reconstruction difficulty and yields stronger learning signals for dynamics, while high-coherence patches are often recoverable by simple neighborhood interpolation.

2.3. Encoder-Decoder Reconstruction

A symmetric convolutional encoder–decoder with a 256-dimensional bottleneck reconstructs the masked CSI amplitude X_{rec}. The masked image modeling loss is computed **ONLY on masked elements**:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{\left|\mathcal{M}^{(A)}\right|} \sum_{\substack{(a,f,t):\\ \mathcal{M}^{(A)}(a,f,t)=1}} \left| X_{\text{rec}}(a,f,t) - X(a,f,t) \right|. \tag{10}$$

We do not concatenate the binary mask as an additional input channel to avoid shortcut learning tied to mask boundaries.

2.4. Downstream Evaluation

To evaluate transferability, we adopt k-shot linear probing. The encoder E is frozen and the bottleneck feature $z = E(X) \in \mathbb{R}^{256}$ is fed to a linear head $F(z) = \mathrm{softmax}(Wz + b)$. The loss is

$$\ell_{\text{CE}}(p, y) = -\log p_y, \qquad p = \mathrm{softmax}(Wz + b), \tag{11}$$

where $y \in \{1, \ldots, C\}$, $W \in \mathbb{R}^{C\times 256}$ and $b \in \mathbb{R}^C$. The objective is

$$\min_{W,b} \frac{1}{|\mathcal{S}_k|} \sum_{(X,y)\in\mathcal{S}_k} \ell_{\text{CE}}\big(\mathrm{softmax}(W\,E(X) + b),\, y\big), \tag{12}$$

where $\mathcal{S}_k$ contains k samples per class.

3. EXPERIMENTS

3.1. Dataset

SignFi[20] is a public dataset for WiFi-based hand and sign language recognition, covering both lab and home scenarios with five subjects. It was collected using an 802.11n CSI tool[21], with single sample dimensions of $(3 \times 30 \times 200)$. This study uses the subset collected from User 5 in the home scenario, which comprises 276 classes with 10 samples each, totaling 2760 samples. NLOS [22] records 12 daily activities of 30 users in three environments. Following relevant settings[23], this paper uses all data from environment 1 and consolidates them into 6 activity classes, totaling 3000 samples. The original activity samples have varying durations and are uniformly downsampled to 200 time steps. This dataset has a significant class imbalance.

3.2. Setup

Framework and data protocol. Models are implemented in PyTorch and trained on a single NVIDIA RTX 5090 GPU. Samples are split into training (80%) and testing (20%) sets without data leakage. Self-supervised pre-training uses only the unlabeled training split. Downstream evaluation follows a k-shot linear probing protocol: for the SignFi dataset, we use a challenging 1-shot setting; for the more complex NLOS dataset, a 20-shot setting is used. In each case, k examples per class are randomly sampled from the training split to train the linear head while the encoder remains frozen. The held-out test split is used exclusively for evaluation. We report Accuracy and macro-F1 score as our primary metrics. Input CSI amplitude is Z-score standardized, and no other augmentations or filtering are applied.

Training specifics and fairness. For self-supervised pre-training, we use Adam (learning rate 1×10^{-4}, β_1=0.9, β_2=0.999, no weight decay), batch size 64, and 300 epochs. The time–frequency patch size is $(3, 10)$; the global masking ratio is $\rho = 0.85$; the sampling exponent in the masking distribution is $\tau = 1$; the binary mask is *not* provided as an input channel. The reconstruction network is a symmetric 3-layer convolutional encoder–decoder with a 256-dimensional bottleneck. The loss is MAE computed *only* on masked elements as in Eq. (10). For linear probing, the encoder (including BatchNorm statistics) is frozen and a single linear classifier $F(z) = \mathrm{softmax}(Wz + b)$ is trained with cross-entropy on the k-shot set. Baselines (AutoFi, AutoSen) are re-implemented per their papers under matched I/O (magnitude-only) and matched pre-training budget (300 epochs) for fairness; learning rate and batch size are kept identical unless noted.

3.3. Results Comparison

Table 1: Performance comparison on SignFi (1-shot) and NLOS (20-shot) datasets. Best SSL performance is in bold.

Method	SignFi		NLOS	
	Accuracy	**F1**	**Accuracy**	**F1**
Supervised	99.64	99.61	99.67	99.65
AutoFi	94.75	94.45	55.00	54.80
AutoSen	87.68	86.95	65.67	64.40
CAM-MIM	**96.92**	**96.67**	**67.50**	**67.69**

As shown in Table 1, CAM-MIM consistently outperforms prior self-supervised methods on both benchmarks. On the SignFi dataset under the stringent 1-shot setting, our method achieves 96.92% accuracy, narrowing the gap to the fully supervised model by a significant margin. On the more challenging NLOS dataset (20-shot), CAM-MIM surpasses the contrastive approach AutoFi by over 12 percentage points in accuracy and also outperforms the reconstructive approach AutoSen. These results validate the effectiveness of focusing reconstruction on motion-induced, low-coherence signal components.

The SignFi subset features 276 fine-grained gesture classes from a **single user**, while NLOS includes 6 coarse activity classes with high intra-class variance from **30 different users**. On SignFi, the single-user setting enables both CAM-MIM's local reconstruction and AutoFi's instance-discrimination to capture subtle patterns for class separation; however, AutoSen's phase reconstruction from amplitude struggles with noisy CSI, yielding lower performance. In contrast, on multi-user NLOS, AutoFi collapses due to its contrastive objective conflating inter-class (activity) and inter-user variance, learning user-specific rather than activity-specific features.

Reconstructive methods like CAM-MIM and AutoSen, relying on intra-sample objectives, are inherently robust to such variability. CAM-MIM's top-tier performance across these scenarios highlights its coherence-aware self-reconstruction as a generalizable pre-training paradigm for WiFi human sensing, excelling in fine-grained differentiation and diverse, multi-user environments.

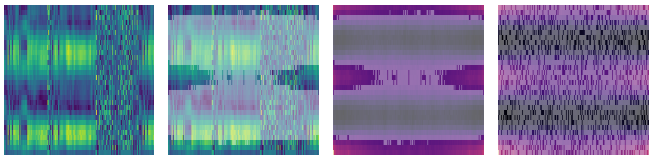
3.4. Ablation Study

Table 2: Ablation experiments on masking strategies.

Method Variant	SignFi		NLOS	
	Accuracy	F1	Accuracy	F1
WithMaskChannel	95.29	94.40	63.50	61.73
Random	95.65	95.48	61.33	59.95
High-CAM-MIM	94.38	93.80	62.50	60.45
CAM-MIM (Ours)	**96.92**	**96.67**	**67.50**	**67.69**

To validate our design choices, we ablate masking strategies while fixing all other hyperparameters (Table 2). Results are consistent across datasets. Replacing low-coherence masking with uniform **random** masking degrades performance (e.g., 1.27% accuracy drop on SignFi, 6.17% on NLOS), confirming that targeted masking better allocates capacity to motion-induced dynamics over predictable backgrounds. The **high-coherence** variant (inverting our strategy) performs worse (e.g., 2.54% and 5.00% drops), as static regions yield trivial reconstructions via interpolation, yielding shallow representations that miss Doppler/multipath cues. Low-coherence prioritization, however, amplifies conditional variance (Eq. (9)), enhancing learning of non-stationary patterns. Adding the binary mask as an input channel (**with mask channel**) also hurts results (e.g., 1.63% and 4.00% drops), promoting shortcut learning on mask boundaries rather than signal context, reducing generalization. These ablations affirm that coherence-aware, low-coherence masking without explicit mask supervision yields robust CSI representations, emphasizing difficulty-aware strategies for heterogeneous signals.

3.5. Visualization of Masking Strategy



(a) Original CSI Amplitude (b) Masked CSI Input (c) CAM-MIM Mask (d) Random Mask

Fig. 2: Visualization of masking strategies. Our coherence-aware mask (c) concentrates on low-coherence regions corresponding to high-amplitude dynamics, unlike a standard random mask (d).

Figure 2 provides an intuitive visual comparison of masking strategies. The original CSI amplitude (a) contains distinct patterns (brighter bands) corresponding to human motion. Our CAM-MIM generated mask (c) strategically targets these information-rich, dynamic regions, as low-coherence areas (derived from 2D autocorrelation) align with Doppler-induced perturbations. In contrast, a standard random mask (d) is applied indiscriminately, often sparing salient motion cues. By forcing reconstruction of these complex parts from limited context (b), CAM-MIM yields representations more attuned to non-stationary signals, enhancing transferability in low-label regimes.

3.6. Mask Ratio Analysis

Table 3: Performance comparison with different mask ratios (ρ).

Mask Ratio	SignFi		NLOS	
	Accuracy	F1	Accuracy	F1
0.75	95.83	95.57	64.67	62.80
0.85	**96.92**	**96.67**	**67.50**	**67.69**
0.95	95.65	95.12	55.67	67.02

We analyze the sensitivity of the mask ratio ρ on downstream performance (Table 3). Across both datasets, results peak at our default $\rho = 0.85$, yielding the highest accuracy and macro-F1 scores, which underscores the importance of striking an optimal balance in masked reconstruction tasks for CSI signals. At a lower ratio ($\rho = 0.75$), the reconstruction becomes overly simplistic: with more visible patches providing abundant context, the model underutilizes its capacity on trivial predictions, particularly for high-coherence backgrounds, leading to suboptimal feature extraction from sparse motion dynamics. This echoes findings in vision-based MIM [17], where insufficient masking dilutes the pretext task's informativeness. Conversely, an aggressive ratio ($\rho = 0.95$) excessively starves contextual information, even when targeting low-coherence regions: while it forces deeper inference on dynamics, the limited visible cues hinder reliable reconstruction of non-stationary patterns, resulting in unstable representations and performance collapse (e.g., sharp drops on NLOS). This highlights a key insight for heterogeneous signals like CSI—extreme masking amplifies noise sensitivity in low-SNR environments, where multipath artifacts demand sufficient surrounding subcarriers for accurate Doppler estimation. The $\rho = 0.85$ sweet spot thus optimally trades off task difficulty against contextual fidelity, enabling the encoder to learn transferable priors on human-induced perturbations without overfitting to isolated fragments. Future extensions could adapt ρ dynamically per sample based on global coherence, further enhancing robustness across varying activity intensities.

4. CONCLUSIONS

We propose CAM-MIM, a coherence-aware MIM framework for SSL in WiFi human sensing, addressing labeled data scarcity via physical priors. Using 2D autocorrelation on mean CSI amplitude as a coherence proxy, we compute per-patch scores to prioritize masking low-coherence regions—targeting motion-induced Doppler/multipath dynamics. Experiments on SignFi (1-shot) and NLOS (20-shot) yield 96.92% and 67.50% accuracy, respectively, outperforming AutoFi/AutoSen by up to 12.5 pp and nearing supervised baselines. Ablations confirm low-coherence masking's superiority over random/high-coherence variants. This physics-informed SSL bridges CSI heterogeneity, yielding transferable representations robust to user/environment variance, with potential in smart homes and health monitoring. Future work will integrate phase cues for finer physiological sensing (e.g., respiration), explore advanced mask guidance like attention-based learnable masking for adaptive difficulty, and enable cross-domain transfer via domain adaptation. Hybrid objectives combining MIM with contrastive/predictive paradigms could further boost generalization in unconstrained, multi-modal WiFi4HAR scenarios.

5. REFERENCES

[1] Q. Pu, S. Gupta, S. Gollakota, and S. Patel, "Whole-home gesture recognition using wireless signals," in *Proc. 19th Annu. Int. Conf. Mobile Comput. Netw. (MobiCom '13)*, 2013, p. 27–38.

[2] F. Adib, H. Mao, Z. Kabelac, D. Katabi, and R. C. Miller, "Smart homes that monitor breathing and heart rate," in *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst. (CHI '15)*, 2015, p. 837–846.

[3] M. Zhao, S. Yue, D. Katabi, T. S. Jaakkola, and M. T. Bianchi, "Learning sleep stages from radio signals: A conditional adversarial architecture," in *Proc. 34th Int. Conf. Machine Learning*, 2017, pp. 4100–4109.

[4] F. Adib and D. Katabi, "See through walls with WiFi!" *ACM SIGCOMM Comput. Commun. Rev.*, vol. 43, no. 4, p. 75–86, 2013.

[5] C. Chen, G. Zhou, and Y. Lin, "Cross-domain WiFi sensing with Channel State Information: A survey," *ACM Comput. Surv.*, vol. 55, no. 11, p. 231, 2023.

[6] K. Ali, M. Alloulah, F. Kawsar, and A. X. Liu, "On goodness of WiFi based monitoring of sleep vital signs in the wild," *IEEE Trans. Mob. Comput.*, vol. 22, no. 1, pp. 341–355, 2023.

[7] S. Yousefi, H. Narui, S. Dayal, S. Ermon, and S. Valaee, "A survey on behavior recognition using WiFi Channel State Information," *IEEE Commun. Mag.*, vol. 55, no. 10, pp. 98–104, 2017.

[8] J. Yang, X. Chen, D. Wang, H. Zou, C. X. Lu, S. Sun, and L. Xie, "SenseFi: A library and benchmark on deep-learning-empowered WiFi human sensing," arXiv preprint arXiv:2207.07859, 2023.

[9] F. Miao, Y. Huang, Z. Lu, T. Ohtsuki, G. Gui, and H. Sari, "Wi-Fi sensing techniques for Human Activity Recognition: Brief survey, potential challenges, and research directions," *ACM Comput. Surv.*, vol. 57, no. 5, p. 107, 2025.

[10] J. Gui, T. Chen, J. Zhang, Q. Cao, Z. Sun, H. Luo, and D. Tao, "A survey on self-supervised learning: Algorithms, applications, and future trends," arXiv preprint arXiv:2301.05712, 2024.

[11] J. Yang, X. Chen, H. Zou, D. Wang, and L. Xie, "AutoFi: Towards automatic WiFi human sensing via geometric self-supervised learning," arXiv preprint arXiv:2205.01629, 2022.

[12] B. Barahimi, H. Tabassum, M. Omer, and O. Waqar, "Context-Aware Predictive Coding: A representation learning framework for WiFi sensing," arXiv preprint arXiv:2410.01825, 2024.

[13] Q. Gao, Y. Hao, and Y. Liu, "Autosen: improving automatic WiFi human sensing through cross-modal autoencoder," arXiv preprint arXiv:2401.05440, 2024.

[14] A. Y. Radwan, M. Yildirim, N. Hasanzadeh, H. Tabassum, and S. Valaee, "A tutorial-cum-survey on self-supervised learning for Wi-Fi sensing: Trends, challenges, and outlook," *IEEE Commun. Surv. Tuts.*, p. 1–1, 2025.

[15] K. Xu, J. Wang, H. Zhu, and D. Zheng, "Evaluating self-supervised learning for WiFi CSI-based Human Activity Recognition," *ACM Trans. Sens. Netw.*, vol. 21, no. 2, p. 21, 2025.

[16] D. M. S. Palipana, "Fall detection using Channel State Information from WiFi devices," Ph.D. dissertation, Cork Institute of Technology, Cork, Ireland, 2019.

[17] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," arXiv preprint arXiv:2111.06377, 2021.

[18] Y. Ma, G. Zhou, and S. Wang, "WiFi sensing with Channel State Information: A survey," *ACM Comput. Surv.*, vol. 52, no. 3, p. 46, 2019.

[19] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 3rd ed. USA: Prentice Hall Press, 2009.

[20] Y. Ma, G. Zhou, S. Wang, H. Zhao, and W. Jung, "SignFi: Sign language recognition using WiFi," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 1, p. 23, 2018.

[21] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Tool release: gathering 802.11n traces with Channel State Information," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 1, p. 53, 2011.

[22] B. A. Alsaify, M. M. Almazari, R. Alazrai, and M. I. Daoud, "A dataset for Wi-Fi-based human activity recognition in line-of-sight and non-line-of-sight indoor environments," *Data Brief*, vol. 33, p. 106534, 2020.

[23] B. A. Alsaify, M. M. Almazari, R. Alazrai, S. Alouneh, and M. I. Daoud, "A CSI-based multi-environment Human Activity Recognition framework," *Appl. Sci.*, vol. 12, no. 2, p. 930, 2022.