RF-Motion-MAE: Uncertainty-Aware Multimodal Masked Autoencoder for Smart-Home Occupancy

Sahar Rezagholi
Rutgers University
Multimodal Machine Learning for Sensing Systems
Saharrgl@Winlab.rutgers.edu

Abstract

Accurate occupancy detection is a key enabler for smart buildings, enabling energy efficiency, security, and intelligent automation. WiFi Channel State Information (CSI) provides a privacy-preserving sensing modality but is highly sensitive to noise and static conditions. This work proposes a self-supervised Masked Autoencoder (MAE) framework that jointly learns from raw CSI and motion-derived features. By reconstructing masked signal segments during pretraining, the model learns robust temporal representations. Experimental results on the WiSA dataset demonstrate improved accuracy and confidence-aware prediction compared to supervised baselines.

Keywords

WiFi CSI, Occupancy Detection, Masked Autoencoder, Self-Supervised Learning, Smart Buildings

ACM Reference Format:
Sahar Rezagholi. 2025. RF-Motion-MAE: Uncertainty-Aware Multimodal Masked Autoencoder for Smart-Home Occupancy. In . ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Smart homes aim to improve energy efficiency, safety, and quality of life by enabling environments that can sense, interpret, and respond to human presence and activity. A fundamental building block of such systems is *occupancy detection*, which determines whether a space is empty or occupied. Accurate occupancy information enables a wide range of applications, including adaptive heating and cooling, intelligent lighting control, security monitoring, and activity-aware automation.

Traditional occupancy detection approaches rely on dedicated sensors such as passive infrared (PIR), cameras, ultrasonic sensors, or wearable devices [11], [13], . While effective in controlled settings, these solutions suffer from practical limitations. Vision-based systems raise privacy concerns, require line-of-sight, and are sensitive to lighting conditions. PIR sensors fail to detect stationary occupants, while wearable-based approaches depend on user compliance and device availability. These challenges motivate the exploration of *device-free* sensing techniques that can operate unobtrusively using existing infrastructure.

Recent research has demonstrated that radio frequency (RF) signals, particularly WiFi Channel State Information (CSI), can serve as a powerful sensing modality for human presence and activity recognition [6], [12], [10]. Human motion and body dynamics perturb wireless propagation paths, inducing measurable changes in CSI amplitude and phase. RF-based sensing preserves privacy, works in non-line-of-sight conditions, and leverages ubiquitous WiFi deployments, making it attractive for smart home environments.

Prior RF-based occupancy detection methods can be broadly categorized into *RF-only* and *motion-only* approaches. RF-only methods directly learn from raw CSI measurements using handcrafted features or supervised deep learning models. While effective, these methods are sensitive to environmental noise, hardware variability, and static multipath effects. Motion-only approaches extract temporal dynamics such as variance, short-time energy, or gradients to capture human movement patterns. However, motion-based methods struggle to detect stationary occupants and often fail in low-activity scenarios. Hybrid RF + motion systems have shown improved robustness but typically rely on fully supervised training, which requires extensive labeled data and limits generalization.

Motivated by these limitations, this project proposes a *self-supervised, multimodal occupancy detection framework* based on Masked Autoencoders (MAE). The key idea is to jointly model raw CSI signals and derived motion features, while leveraging self-supervised pretraining to learn robust temporal representations. By masking a portion of the input sequence and reconstructing missing segments, the MAE encoder is forced to capture global temporal structure rather than memorizing local patterns. The pretrained encoder is then fine-tuned with a lightweight classifier to perform binary occupancy detection (Empty vs. Occupied).

The proposed approach is evaluated using the WiSA dataset, which contains CSI recordings collected from real indoor environments with multiple subjects performing diverse activities. Experiments demonstrate that the MAE-based model consistently outperforms RF-only, motion-only, and RF+motion baselines. In particular, the model achieves its best performance with a moderate masking ratio, highlighting the importance of balanced self-supervised learning for noisy RF signals.

The remainder of this paper is organized as follows. Section 2 reviews related work on RF-based sensing and self-supervised learning. Section 3 describes the proposed MAE-based multimodal framework in detail. Section 4 presents the experimental setup and dataset. Section 5 reports quantitative results and analysis. Finally, Section 6 concludes the paper and outlines future research directions.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 Related Work and Motivation

2.1 RF-based and Motion-based Occupancy Detection

Occupancy detection is a fundamental component of smart home and smart building systems, enabling applications such as energy-efficient HVAC control, safety monitoring, and context-aware automation. Traditional approaches rely on physical sensors such as PIR, cameras, or wearable devices; however, these solutions suffer from privacy concerns, limited coverage, or high deployment cost. As a result, device-free sensing using WiFi Channel State Information (CSI) has emerged as a promising alternative.

RF-only methods leverage variations in CSI amplitude or phase caused by human presence and movement. These approaches are effective for detecting motion and coarse occupancy patterns, but they exhibit notable limitations when occupants remain stationary or move slowly. In such cases, CSI variations become weak and ambiguous, leading to degraded detection accuracy. Furthermore, RF-only methods are highly sensitive to environmental noise, multipath fading, and hardware variations across different deployment settings.

Motion-only methods attempt to address these limitations by explicitly extracting temporal motion features, such as signal variance, short-time energy, or temporal gradients. While motion features are more robust to environmental changes, they inherently fail when occupants are static (e.g., sitting, reading, or lying down), which is a common scenario in residential environments. As a result, motion-only models suffer from low recall in real-world occupancy detection.

RF + Motion fusion methods combine raw CSI measurements with handcrafted motion features to improve robustness. These hybrid approaches have demonstrated improved performance compared to unimodal models by mitigating the weaknesses of RF-only and motion-only methods. However, existing fusion techniques typically rely on fully supervised learning and handcrafted features, making them sensitive to missing data, noise, and distribution shifts across environments. Moreover, they often require large amounts of labeled data, which is expensive and difficult to collect in practice.

2.2 Motivation for Masked Autoencoder-based Learning

A key challenge across existing RF-based occupancy detection methods is their limited ability to learn robust representations under noisy, incomplete, or low-motion conditions. WiFi CSI data is inherently imperfect: packets may be dropped, measurements may be corrupted by interference, and human motion patterns vary significantly across users and environments. Supervised learning alone is insufficient to address these challenges, as it encourages models to overfit to specific motion patterns or environmental conditions.

Recent advances in **Masked Autoencoders (MAEs)** have demonstrated that self-supervised masked reconstruction is an effective strategy for learning robust and generalizable representations from high-dimensional and noisy data [5], [7],[8]. He et al. [4] showed that sparsely masking and reconstructing visual tokens enables models to learn global structure while being resilient to missing information. This paradigm was later extended to spatiotemporal data, including videos [2], where MAEs were shown to capture long-term temporal dependencies under partial observation.

More recently, Social-MAE [1] applied masked autoencoder learning to human motion trajectories, demonstrating that reconstructing masked motion tokens enables robust representation learning under noisy and incomplete observations. This work highlights the suitability of MAE-based learning for temporal motion data, particularly in scenarios with limited supervision.

3 Proposed Approach: RF-Motion-MAE

3.1 Overview

We propose **RF-Motion-MAE**, a self-supervised and uncertainty-aware framework for **binary occupancy detection** (Empty vs. Occupied) in smart-home environments. The method fuses (i) **RF modality** from WiFi CSI windows and (ii) a **motion modality** *derived from CSI* (not a separate PIR sensor) to improve robustness under noise, multipath, and low-motion occupants.

The pipeline has two stages: (i) **MAE pretraining** to reconstruct masked RF+motion patches using MSE, and (ii) **fine-tuning** where we discard the decoder and train an encoder + classifier head for occupancy labels.

3.2 Data Representation and Tokenization

Each sample is a time window of length $T = 500$ frames. The RF input is a CSI-derived feature matrix $\mathbf{X}_{\mathrm{RF}} \in \mathbb{R}^{T\times F}$ (e.g., $F = 90$ subcarrier links), and the motion stream $\mathbf{X}_{\mathrm{M}} \in \mathbb{R}^{T\times 4}$ contains four motion-energy features computed from CSI. We concatenate them along feature dimension:

$$\mathbf{X} = [\mathbf{X}_{\mathrm{RF}} \parallel \mathbf{X}_{\mathrm{M}}] \in \mathbb{R}^{T\times(F+4)}. \tag{1}$$

We split time into non-overlapping patches of length $p = 10$ frames, producing $P = T/p = 50$ tokens. Each token is a flattened patch:

$$\mathbf{x}_i \in \mathbb{R}^{p(F+4)} \quad i = 1, \ldots, P, \tag{2}$$

then projected to an embedding dimension D (e.g., $D = 128$) by a linear layer (PatchEmbed). This fixes concrete shapes and patching as requested in feedback.

3.3 Motion Feature Formulation (Derived Modality)

Let $\mathbf{R} \in \mathbb{R}^{T\times F}$ be the RF window (CSI amplitude features). We compute motion features per time step t using statistics over RF channels [14]:

(1) Mean amplitude.

$$\mu(t) = \frac{1}{F}\sum_{f=1}^{F} R(t,f). \tag{3}$$

(2) Amplitude variance.

$$\mathrm{Var}(t) = \frac{1}{F}\sum_{f=1}^{F} \big(R(t,f) - \mu(t)\big)^2. \tag{4}$$

(3) Short-time energy (STE).

$$\mathrm{STE}(t) = \frac{1}{F}\sum_{f=1}^{F} R(t,f)^2. \tag{5}$$

(4) Temporal gradients. We use absolute first-order differences:

$$g_\mu(t) = |\mu(t) - \mu(t-1)|, \quad g_{\mathrm{STE}}(t) = |\mathrm{STE}(t) - \mathrm{STE}(t-1)|. \tag{6}$$

Finally, motion at time t is $\mathbf{m}(t) = [\mathrm{Var}(t), \mathrm{STE}(t), g_\mu(t), g_{\mathrm{STE}}(t)]$, giving $\mathbf{X}_{\mathrm{M}} \in \mathbb{R}^{T\times 4}$. This provides explicit, reproducible motion definitions as requested.

3.4 Masked Autoencoder (MAE) Pretraining

Given token embeddings $\{\mathbf{z}_i\}_{i=1}^{P}$, we randomly mask a fraction r (e.g., $r \in \{0.2, 0.35, 0.4, 0.45, 0.55, 0.6\}$) of tokens and keep the rest:

$$\mathcal{I}_{\mathrm{keep}} \cup \mathcal{I}_{\mathrm{mask}} = \{1,\ldots,P\}, \quad |\mathcal{I}_{\mathrm{mask}}| = \lfloor rP \rfloor. \tag{7}$$

Encoder: a TransformerEncoder processes only visible tokens to learn a robust latent representation.

Decoder: we insert learned `[MASK]` tokens back into the *original patch positions* and decode to predict the missing patches. In practice this requires: (i) restoring token order using $\mathcal{I}_{\mathrm{keep}}/\mathcal{I}_{\mathrm{mask}}$ and (ii) adding positional information (time). This fixes the "no ordering / no position" issue.

Reconstruction loss (masked-only). Let $\hat{\mathbf{X}}$ be reconstructed patches and $\mathbf{X}$ be the true patches. We compute MSE only over masked indices:

$$\mathcal{L}_{\mathrm{MAE}} = \frac{1}{|\mathcal{I}_{\mathrm{mask}}|}\sum_{i\in\mathcal{I}_{\mathrm{mask}}} \|\hat{\mathbf{x}}_i - \mathbf{x}_i\|_2^2. \tag{8}$$

This matches the MAE training objective described in the feedback.

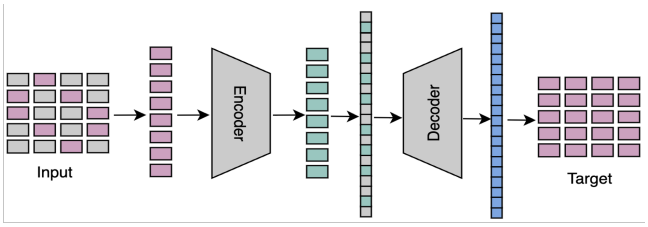

Figure 1: Masked Autoencoder architecture for multimodal CSI representation learning.

3.5 Occupancy Label Definition and Decision Process

Ground-Truth Occupancy Labels from the Dataset. In this project, occupancy labels are derived directly from the **WiSA CSI dataset annotations**. Each CSI recording corresponds to a *human activity session* performed by a volunteer in a room. Activities such as *walking, cleaning, reading, writing, squats, running, jumping, arm training*, and *lying down* all imply the **presence of at least one person**.

Therefore, we define occupancy as a **binary label**:

$$y = \begin{cases} 1 & \text{if any human activity is present (Occupied)} \\ 0 & \text{if no human activity is present (Empty)} \end{cases} \tag{9}$$

In the WiSA dataset, files labeled as `Clean` or background CSI recordings correspond to **empty-room conditions** and are assigned $y = 0$. All other activity recordings are assigned $y = 1$. This labeling strategy matches the goal of smart-home occupancy detection rather than fine-grained activity recognition.

Input to the Classifier. After MAE pretraining, each input sample is a time window $\mathbf{X} \in \mathbb{R}^{500\times 94}$, consisting of:

- 90 RF (CSI-based) features
- 4 motion features derived from RF (variance, energy, temporal gradients)

The MAE encoder transforms the input into a sequence of latent tokens. These tokens are aggregated using temporal pooling (mean pooling) to obtain a fixed-length representation[9]:

$$\mathbf{h} = \frac{1}{P}\sum_{i=1}^{P} \mathbf{z}_i \tag{10}$$

where $\mathbf{z}_i$ are encoder outputs for each patch.

Classifier Decision Rule. A lightweight classifier head is trained on top of the frozen or fine-tuned encoder:

$$\mathbf{s} = \mathbf{W}\mathbf{h} + \mathbf{b} \tag{11}$$

where $\mathbf{s} \in \mathbb{R}^2$ are logits corresponding to *Empty* and *Occupied*.

The predicted occupancy label is obtained using softmax:

$$\hat{y} = \arg\max_{c\in\{\mathrm{Empty,Occupied}\}} \mathrm{softmax}(\mathbf{s})_c \tag{12}$$

Interpretation.

- The model does **not** explicitly detect people.
- Instead, it learns **RF and motion patterns** that statistically differ between empty environments and environments with human presence.
- Even low-motion activities (e.g., reading or lying down) alter CSI statistics through multipath and micro-movements, which are captured by the MAE-learned representation.

3.6 Uncertainty via MC Dropout (Selective Classification)

To estimate epistemic uncertainty, dropout remains ON at inference and we run K stochastic forward passes. We average softmax probabilities:

$$\bar{\mathbf{p}} = \frac{1}{K}\sum_{k=1}^{K} \mathbf{p}^{(k)}. \tag{13}$$

Uncertainty is computed using predictive entropy:

$$H(\bar{\mathbf{p}}) = -\sum_c \bar{p}_c \log \bar{p}_c, \tag{14}$$

or alternatively variance across $\mathbf{p}^{(k)}$. We then **reject** samples with $H(\bar{\mathbf{p}}) > \tau$ to trade coverage for higher accuracy.

This allows the system to output *"I am not confident"* instead of a wrong occupancy decision, which is critical for safety-aware smart-home applications.

Algorithm 1: RF-Motion-MAE Training and Uncertainty

Input: RF windows $\mathbf{R} \in \mathbb{R}^{T \times F}$, occupancy labels $y \in \{0, 1\}$ (for FT)
Output: occupancy predictor with uncertainty-aware rejection

Stage 1: Self-supervised MAE pretraining
Compute motion stream $\mathbf{X}_{\mathrm{M}}$ from $\mathbf{R}$ using Var/STE/gradients.
Concatenate $\mathbf{X} = [\mathbf{R} \| \mathbf{X}_{\mathrm{M}}] \in \mathbb{R}^{T \times (F+4)}$.
Patchify time into $P = T/p$ tokens; embed tokens to dimension D.
Sample mask set $\mathcal{I}_{\mathrm{mask}}$ (ratio r), keep $\mathcal{I}_{\mathrm{keep}}$.
Encode visible tokens: $\mathbf{H} = \mathrm{Enc}(\mathbf{Z}_{\mathcal{I}_{\mathrm{keep}}})$.
Reinsert `[MASK]` tokens into original positions + positional embeddings; decode.
Minimize $\mathcal{L}_{\mathrm{MAE}}$ (MSE on masked tokens only).

Stage 2: Supervised fine-tuning for occupancy
Discard decoder; attach classifier head to encoder output.
Train with cross-entropy on labeled occupancy windows.

uncertainty-aware
Run MC dropout K times to get $\{\mathbf{p}^{(k)}\}_{k=1}^{K}$, compute $\bar{\mathbf{p}}$.
Compute uncertainty (entropy or variance). Reject if $H(\bar{\mathbf{p}}) > \tau$; else predict $\arg\max \bar{\mathbf{p}}$.

4 Dataset and Experimental Setup

4.1 WiSA Dataset Description

This work uses the **WiSA (WiFi Sensing for Activities) dataset**, a publicly available CSI-based human activity dataset collected in indoor environments for RF sensing research [3].

Environment and Hardware. CSI data were collected in two residential indoor spaces:

- Bedroom: 3.8 m × 2.4 m
- Living room: 3.2 m × 4.4 m

Two Ubuntu 14.04 computers equipped with Intel 5300 network interface cards and three external antennas were used. One device acted as a transmitter with a single active antenna, while the receiver used all three antennas, resulting in a 3×1 antenna configuration.

The 802.11n CSI Tool captured CSI measurements at:

- 30 subcarriers per antenna link
- 20 MHz bandwidth in the 5 GHz band
- Sampling rate of 1000 Hz

All devices were placed at a height of 1.3 m, with transmitter–receiver separation of 2.2 m or 3.2 m depending on room layout.

Participants and Activities. The dataset includes recordings from 15 volunteers (9 females and 6 males), aged between 23 and 26. Each participant performed a subset of nine daily activities:

- **Light**: lying down, reading, writing
- **Moderate**: walking, cleaning, arm training
- **Intense**: running, squats, jumping

Each recording session lasted approximately 20 minutes, during which participants performed activities naturally without external interference.

Occupancy Labels. For the purpose of occupancy detection, activities implying human presence were labeled as **Occupied**. Clean or background CSI recordings without human motion were labeled as **Empty**. This converts the original multi-class activity dataset into a binary occupancy detection problem.

—

4.2 Data Preprocessing and Feature Construction

Sliding Window Segmentation. Raw CSI amplitude data were segmented using a sliding window of 500 time steps. Each window represents a short temporal snapshot of RF dynamics and serves as one input sample.

Formally, each sample is represented as:

$$\mathbf{X} \in \mathbb{R}^{500 \times 90}$$

where 90 corresponds to CSI features extracted from subcarriers and antennas.

Motion Feature Extraction. To enhance robustness against noise and improve sensitivity to human presence, four motion-related features were computed directly from RF amplitude signals:

- Variance of amplitude
- Short-time energy (STE)
- Temporal gradient of mean amplitude
- Temporal gradient of energy

These features capture both signal intensity and temporal change caused by human movement.

The motion features are concatenated with RF features to form a multimodal input:

$$\mathbf{X}_{\mathrm{RF+Motion}} \in \mathbb{R}^{500 \times 94}$$

Normalization and Tensor Construction. All features were converted to floating-point tensors and normalized implicitly through MAE pretraining. Each sample is paired with a binary occupancy label and loaded using PyTorch DataLoaders.

—

4.3 Experimental Setup

Data Splitting. The dataset was randomly split into:

- 80% training data
- 20% validation data

The training set was used for MAE pretraining and classifier fine-tuning, while the validation set was used exclusively for performance evaluation.

MAE Pretraining Configuration. A Masked Autoencoder (MAE) was trained in a self-supervised manner using only input data, without labels. Key configuration parameters include:

- Patch size: 10 time steps
- Encoder depth: 4 Transformer layers
- Number of attention heads: 4
- Mask ratios: 20%–60%

The MAE was optimized using mean squared error loss between reconstructed and ground-truth masked patches.

Fine-Tuning for Occupancy Classification. After pretraining, the decoder was discarded. The encoder was fine-tuned together with a lightweight classifier head to predict Empty vs. Occupied using cross-entropy loss.

Evaluation Metrics. Model performance was evaluated using:

- Classification accuracy
- F1-score
- Confusion matrix analysis
- Selective classification (accuracy vs. coverage)

This setup allows analysis of both predictive performance and uncertainty handling, which is essential for smart-home deployment.

5 Results and Discussion

This section evaluates the proposed MAE-based RF and motion fusion framework for binary occupancy detection. We compare against RF-only, motion-only, and supervised fusion baselines, analyze the impact of MAE masking ratios, and study classification reliability through confusion matrices and selective classification.

5.1 Baseline Model Comparison

Table 1 compares the proposed MAE fine-tuned model against three baselines: RF-only with dropout, motion-only with dropout, and supervised RF+motion fusion.

Table 1: Model comparison for occupancy detection.

Model	Input	Best Val Acc	Best Val F1
RF–Dropout	RF only	0.860	0.880
Motion-only	Motion only	0.778	0.832
RF+Motion–Dropout	RF + Motion	0.895	0.913
MAE–FT (RF+Motion)	**RF + Motion**	**0.913**	**0.929**

The motion-only model achieves the lowest performance, as handcrafted motion features fail to reliably capture low-motion or stationary occupants (e.g., sitting, reading). The RF-only model performs better by leveraging CSI variations caused by human presence, but remains sensitive to environmental noise and multipath effects.

Supervised fusion of RF and motion improves performance, indicating complementary information between modalities. The proposed MAE-pretrained model further improves validation accuracy from 0.895 to 0.913 and F1-score from 0.913 to 0.929. This gain highlights the effectiveness of self-supervised MAE pretraining in learning robust temporal representations under missing and noisy RF conditions.

5.2 Effect of MAE Masking Ratio

To analyze the impact of masking strength during MAE pretraining, we evaluate multiple mask ratios, as shown in Table 2.

Performance peaks at a masking ratio of 35%. Lower masking (20%) makes reconstruction too easy, limiting the encoder's ability to learn robust representations. Excessively high masking (55–60%) forces the model to hallucinate large portions of the signal, degrading reconstruction quality and downstream performance. These results indicate that moderate masking provides the optimal balance between learning temporal structure and maintaining reconstruction stability.

Table 2: Effect of MAE masking ratio on fine-tuned performance.

Mask Ratio	Best Val Acc	Best Val F1
MAE–FT (20%)	0.896	0.914
MAE–FT (35%)	**0.913**	**0.929**
MAE–FT (40%)	0.907	0.923
MAE–FT (45%)	0.906	0.922
MAE–FT (55%)	0.902	0.919
MAE–FT (60%)	0.900	0.917

5.3 Confusion Matrix Analysis

Figure 2 presents the normalized confusion matrix for the MAE–FT model.

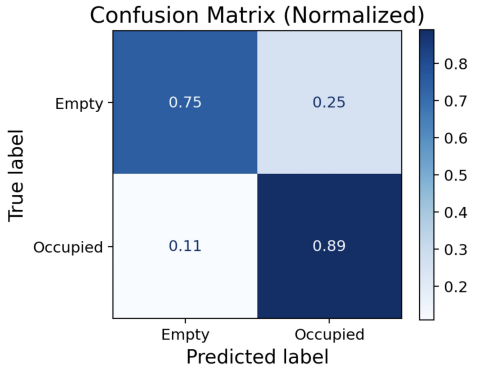

Figure 2: Normalized confusion matrix for MAE–FT (RF+Motion).

The model achieves a high recall for the *Occupied* class (0.89), indicating strong sensitivity to human presence. The recall for the *Empty* class is lower (0.75), with 25% of empty samples misclassified as occupied. These false positives are primarily caused by environmental RF fluctuations, multipath dynamics, or background motion-like noise.

Missed detections of occupied scenes (11%) are mainly associated with low-motion activities, such as sitting or reading, where both RF and motion signals exhibit limited variation.

5.4 Selective Classification and Confidence Analysis

Figure 3 illustrates accuracy versus coverage under selective classification. As the confidence threshold increases, the model accepts fewer samples but achieves significantly higher accuracy.

The model achieves near-perfect accuracy (approximately 0.98–0.99) when operating on its most confident 64–65% of samples. Performance degrades as coverage approaches 100%, where ambiguous cases such as low-motion occupancy and noisy RF segments are included.

This behavior suggests a practical deployment strategy in smart home systems: high-confidence predictions can be accepted directly, while low-confidence samples can be deferred for temporal smoothing or multimodal fusion with additional sensors.

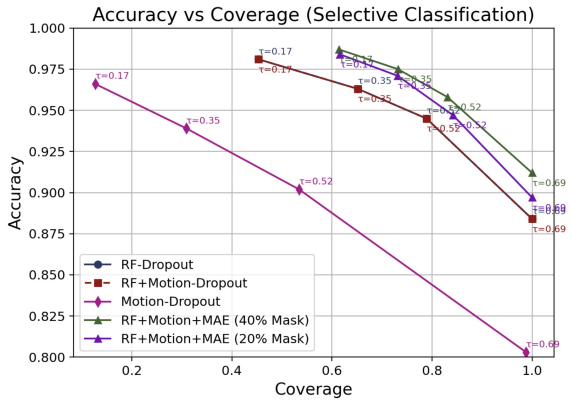

Figure 3: Accuracy vs. coverage under selective classification.

Overall, the results demonstrate that MAE-based self-supervised pretraining significantly enhances occupancy detection performance by improving robustness to missing data, noise, and environmental variability. The proposed approach consistently outperforms RF-only, motion-only, and supervised fusion baselines, achieving strong accuracy, balanced class performance, and reliable confidence estimation.

6 Conclusion

This work presented a self-supervised multimodal framework for binary occupancy detection using WiFi Channel State Information (CSI) and motion-derived features. By leveraging a Masked Autoencoder (MAE) pretraining strategy, the proposed approach learns robust temporal representations from partially observed RF signals before fine-tuning for occupancy classification. This design addresses key challenges in RF-based sensing, including environmental noise, multipath effects, missing data, and low-motion occupancy scenarios.

Experimental results on the WiSA dataset demonstrate that MAE-based pretraining consistently improves performance over RF-only, motion-only, and supervised RF+motion baselines. The proposed MAE–FT model achieves a best validation accuracy of 0.913 and an F1-score of 0.929, with optimal performance observed at a moderate masking ratio of 35%. These results confirm that self-supervised reconstruction of masked RF and motion patches enables the encoder to capture meaningful temporal structure that generalizes well to downstream occupancy detection.

Further analysis using confusion matrices shows strong sensitivity to occupied scenes, while selective classification experiments reveal that the model can achieve near-perfect accuracy on its most confident predictions. This reliability-aware behavior is particularly valuable for real-world smart building applications, where high-confidence decisions can be prioritized and ambiguous cases can be deferred or fused with additional sensing modalities.

Overall, this study demonstrates that MAE-based self-supervised learning is a promising direction for robust, privacy-preserving occupancy detection in smart environments. Future work will explore multi-person occupancy estimation, cross-domain generalization to unseen environments and hardware, and integration with additional sensing modalities to further enhance system robustness and scalability.

Acknowledgment

This work was supported by WINLAB at Rutgers University. The author thanks her advisor and colleagues for their guidance and feedback.

References

[1] Mahsa Ehsanpour, Ian Reid, and Hamid Rezatofighi. 2025. Social-MAE: Social masked autoencoder for multi-person motion representation learning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 13913–13919.

[2] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. 2022. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems* 35 (2022), 35946–35958.

[3] WiSA Research Group. 2024. *WiSA: Environment and CSI Collection Dataset*. doi:10.6084/m9.figshare.24939765.v1

[4] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16000–16009.

[5] Junkun Jiang, Jie Chen, and Yike Guo. 2022. A dual-masked auto-encoder for robust motion capture with spatial-temporal skeletal token completion. In *Proceedings of the 30th ACM International Conference on Multimedia*. 5123–5131.

[6] Minyoung Jung, Joosang Lee, Donghyun Kim, Dongjun Park, and Taeyeon Kim. 2025. Implementing Wi-Fi CSI-Based Room-Level Occupancy Estimation: an Experimental Study in Multi-zone Residential Environments. *Journal of Building Engineering* (2025), 113155.

[7] Jintang Li, Ruofan Wu, Wangbin Sun, Liang Chen, Sheng Tian, Liang Zhu, Changhua Meng, Zibin Zheng, and Weiqiang Wang. 2023. What's behind the mask: Understanding masked graph modeling for graph autoencoders. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1268–1279.

[8] Yaqian Liang, Shanshan Zhao, Baosheng Yu, Jing Zhang, and Fazhi He. 2022. Meshmae: Masked autoencoders for 3d mesh data analysis. In *European conference on computer vision*. Springer, 37–54.

[9] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130* (2017).

[10] Muhammad Salman, Lismer Andres Caceres-Najarro, Young-Duk Seo, and Youngtae Noh. 2024. WiSOM: WiFi-enabled self-adaptive system for monitoring the occupancy in smart buildings. *Energy* 294 (2024), 130420.

[11] Azad Shokrollahi, Jan Persson, Reza Malekian, Arezoo Sarkheyli-haegele, and Fredrik Karlsson. 2024. PIR Sensor-Based Occupancy Monitoring in Smart Buildings: A Review of Methodologies and Machine Learning Approaches. (2024).

[12] Nan Zhang, Jianchao Zhang, Baotian Chang, Junzhu Duan, Boni Su, Jingchao Xie, Ying Ji, and Fangyu Li. 2025. A Bi-LSTM-based Wi-Fi CSI approach for non-intrusive human behavior recognition in smart buildings. *Energy and Buildings* (2025), 116059.

[13] Wuxia Zhang, John Calautit, Paige Wenbin Tien, Yupeng Wu, and Shuangyu Wei. 2024. Deep learning models for vision-based occupancy detection in high occupancy buildings. *Journal of Building Engineering* 98 (2024), 111355.

[14] Mingmin Zhao, Yonglong Tian, Hang Zhao, Mohammad Abu Alsheikh, Tianhong Li, Rumen Hristov, Zachary Kabelac, Dina Katabi, and Antonio Torralba. 2018. RF-based 3D skeletons. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*. 267–281.