CIG-MAE: Cross-Modal Information-Guided Masked Autoencoder for Self-Supervised WiFi Sensing

Gang Liu, Yanling Hao, and Yixuan Zou

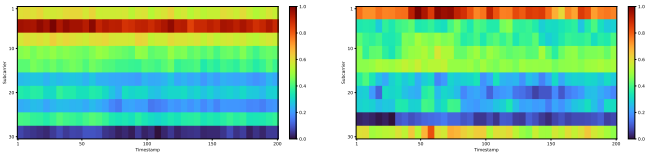
***Abstract*—Human Action Recognition using WiFi Channel State Information (CSI) has emerged as an attractive alternative to vision-based methods due to its ubiquity, device-agnostic nature, and inherent privacy-preserving capabilities. However, the high cost of manual annotation and the limited scale of publicly available CSI datasets restrict the performance of supervised approaches. Self-supervised learning (SSL) offers a promising avenue, but existing contrastive paradigms rely on data augmentations that conflict with the physical semantics of radio signals and require large-batch training, making them poorly suited for CSI. To overcome these challenges, we introduce CIG-MAE—a Cross-modal Information-Guided Masked Autoencoder—that reconstructs both the amplitude and phase of CSI using a symmetric dual-stream architecture with a high masking ratio. Specifically, we propose an Adaptive Information-Guided Masking strategy that dynamically allocates attention to time–frequency regions with high information density to improve learning efficiency, and incorporate a Barlow Twins regularizer to align cross-modal representations without negative samples. Experiments on three public datasets show that CIG-MAE consistently outperforms SOTA SSL methods and even surpasses a fully supervised baseline, demonstrating superior data efficiency, robustness, and representation generalization.**

***Index Terms*—Human activity recognition, WiFi sensing, self-supervised learning, masked autoencoder, cross-modal learning, adaptive information-guided masking.**

I. Introduction

Human Action Recognition (HAR) has garnered significant attention for its broad applications in health monitoring, elderly care, and smart homes [1]–[5]. Among various sensing modalities, HAR based on WiFi Channel State Information (CSI-HAR) has emerged as a particularly promising approach due to its ubiquity, device-agnostic nature, privacy-preserving characteristics, and capability to operate in non-line-of-sight (NLOS) environments [6], [7].

Despite its potential, the practical deployment of CSI-HAR is significantly hampered by the high cost and difficulty of acquiring large-scale, accurately labeled datasets. The performance of conventional supervised learning models is fundamentally constrained by data scarcity and the challenge of covering diverse subjects and environments, which can lead to poor generalization and domain shift issues [8], [9]. To mitigate this dependency on labeled data, Self-Supervised Learning (SSL) has been introduced as a powerful paradigm capable of leveraging large volumes of unlabeled CSI to learn transferable representations and enhance robustness in unseen scenarios [10], [11].



(a) Amplitude Reconstruction Heatmap (b) Phase Reconstruction Heatmap

Fig. 1. Reconstruction error heatmaps for amplitude and phase on the subcarrier-time plane. A larger reconstruction error (warmer color) is interpreted as higher information density. Both modalities show high difficulty on low-frequency subcarriers, but their high-difficulty regions do not fully overlap, reflecting modal complementarity and non-uniform information distribution.

However, mainstream SSL paradigms, predominantly developed for computer vision and natural language processing, often prove suboptimal when directly applied to WiFi CSI data. In particular, contrastive learning depends heavily on data augmentations that lack consistency with radio propagation physics and typically requires large-batch training, which is incompatible with the small-dataset reality of WiFi sensing. These mismatches can corrupt semantic structure and undermine the ability to learn invariant features.

A key factor that has been largely overlooked is the intrinsic complementary nature of the amplitude and phase components of CSI. While much existing SSL work relies solely on amplitude [12], [13], phase has been shown to be highly sensitive to micro-motion and propagation geometry. Yet, even when phase is used, it is often treated as a unidirectional prediction target rather than a jointly learned modality, failing to capture their inherent coupling [14]. Moreover, as illustrated in Fig. 1, the information density—interpreted as reconstruction difficulty—exhibits strong non-uniformity across the time–frequency axis for both modalities, and critically, their high-density regions only partially overlap. This indicates that amplitude and phase provide comple-

(Corresponding author: Yanling Hao.)

Gang Liu and Yixuan Zou are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K. (e-mail: gangliu2004@outlook.com; yixuan.zou@qmul.ac.uk).

Yanling Hao is with the School of Computing and Engineering, University of West London, London W5 5RF, U.K. (e-mail: yanling.hao@uwl.ac.uk).

mentary physical information and that a well-designed SSL framework must explicitly model and leverage this complementarity to obtain discriminative and physically meaningful representations.

To address these challenges, we propose the Cross-modal Information-guided Masked Autoencoder (CIG-MAE), a self-supervised framework tailored to the physical structure and sensing properties of CSI. CIG-MAE adopts a symmetric dual-stream architecture that jointly reconstructs amplitude and phase under a high masking ratio, eliminating the need for problematic augmentations and large batch sizes. Furthermore, we introduce an Adaptive Information-guided Masking (AIM) mechanism that allocates visibility to time–frequency regions with higher information density, thereby enhancing representation learning efficiency and boosting model performance. The main contributions of this work are summarized as follows:

- We propose a symmetric dual-stream self-supervised reconstruction framework, CIG-MAE, tailored to the physical properties of CSI. This framework reconstructs the amplitude and phase components of CSI with a high masking ratio and introduces a non-contrastive Barlow Twins (BT) regularizer to align the dual-stream representations and eliminate feature redundancy. This design explicitly models the complementarity between amplitude and phase, effectively avoiding the reliance of existing methods on data augmentations that are inconsistent with physical semantics and on large-batch configurations.
- We design an AIM strategy to enhance data efficiency and representation quality. This strategy introduces a learnable policy network that can evaluate the information value of each time-frequency region, thereby dynamically focusing the model's "visibility budget" on regions with higher information density that are more critical for learning discriminative features, addressing the inefficiency of random masking from a mechanistic perspective.
- We conducted systematic evaluations on three public datasets. The experimental results show that under the standard linear probing protocol, the performance of the proposed CIG-MAE not only significantly surpasses various advanced self-supervised baselines but even exceeds a fully supervised model in some particular scenarios. Furthermore, exhaustive ablation studies and parameter analyses have verified the effectiveness of our model's components and its overall robustness.

The remainder of this paper is organized as follows: Section II reviews related work. Section III elaborates on our proposed CIG-MAE framework. Section IV reports detailed experimental settings and results. Finally, Section V concludes the paper and discusses future research directions.

II. Related Work

SSL has achieved remarkable success across various domains [15]–[18] and is gradually being adopted for signal processing tasks [19]–[21]. Early research on SSL for HAR explored various pretext tasks. Saeed et al. [20] systematically designed eight different pretext tasks such as feature prediction and transformation recognition for diverse signals. Subsequently, they extended SSL to federated settings [22], proposing scalogram-signal correspondence learning via wavelet transforms. As research deepened, methods specifically tailored to the physical structure of WiFi CSI have emerged, broadly categorized into discriminative (contrastive/predictive) and generative paradigms.

A. Contrastive and Predictive Paradigms

The fundamental idea of contrastive learning is to maximize representation consistency between different views of the same sample while minimizing consistency between different samples [23]–[26]. In WiFi sensing, current works optimize this paradigm primarily through view construction and objective function design.

Regarding *view construction*, Lau et al. [27] treated CSI from spatially separated receivers as positive pairs. Liu et al. [28] introduced STFNets to mine time-frequency features, demonstrating the importance of frequency domain augmentations. Song et al. [21] focused on multi-representation consistency, utilizing a translator-predictor structure to map distinct radio frequency (RF) features like Angle of Arrival (AoA) and Doppler Frequency Shift (DFS) into a unified space. Xu et al. [29] employed a dual-stream architecture to collaboratively model spatial and channel features. Lyons et al. [30] proposed WiFiAct, which combines an environment-invariant preprocessing pipeline with a Bayesian CNN for generalized learning, though its partial reliance on labels distinguishes it from pure SSL.

Regarding *objective functions*, Yang et al. [31] proposed AutoFi, which enhances consistency by minimizing the Kullback-Leibler (KL) divergence of probability distributions alongside mutual information and geometric structure constraints. Chen et al. [13] addressed the instability of KL divergence by maximizing Jensen-Shannon (JS) divergence and introducing Gaussian regularization to improve generalization. Recently, Xiao et al. [12] incorporated diffusion models, specifically Denoising Diffusion Probabilistic Models (DDPM) to generate high-quality, temporally specific augmentations and designed adaptive weights based on activity information.

In predictive learning, the model learns representations that capture the dynamics and latent structures of the data by predicting future segments, missing context, or certain transformation properties [32], [33]. Haresamudram et al. [34] applied Contrastive Predictive Coding (CPC) to wearable sensors. In CSI-HAR, Barahimi et al. [35] proposed CAPC,

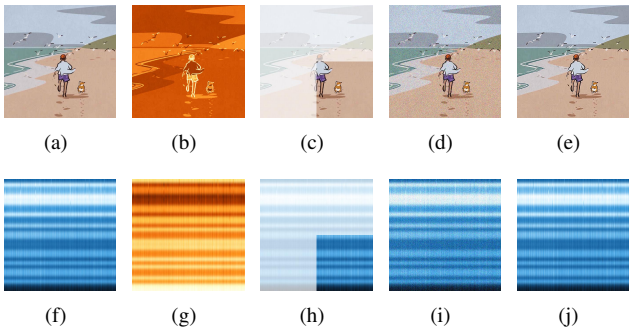


Fig. 2. Illustration of five visualization/augmentation operators applied to a natural image (top row) and to CSI data (bottom row). From left to right: Original, Color Mapping, Crop, Gaussian Noise, and Unsharp Mask. Directly transferring vision augmentations to CSI can misalign with RF physical semantics.

combining CPC with BT to leverage uplink/downlink views for capturing temporal context and view consistency.

However, these discriminative paradigms face two inherent limitations in the CSI context. First, they rely heavily on view construction via augmentations (e.g., cropping, noise injection) that are often semantically inconsistent with the physical properties of radio signals [36], [37], as illustrated in Fig. 2. While solutions like using multiple receivers or specialized augmentations exist [21], [27], [35], [38], they often introduce extra hardware costs or complexity. Second, contrastive learning typically requires large batches and numerous negative samples to perform well [23], [26], a condition that large-scale image datasets like ImageNet can satisfy [39]. This is fundamentally at odds with the CSI domain, where public datasets are typically small, containing only hundreds to thousands of samples [40]. Predictive methods, which learn by predicting future or masked segments [35], often still rely on contrastive objectives and are sensitive to hyperparameters, magnifying training instability under small-sample, high-noise conditions.

B. Generative and Reconstructive Paradigms

Generative learning, which forces models to learn intrinsic structures by reconstructing masked inputs, offers a solution more suitable for small-sample scenarios with specific physical structures [41]. In general sensor domains, Cheng et al. proposed MaskCAE [19], utilizing sparse convolutions to process masked sensor data without explicit augmentation. Miao et al. [42] proposed STMAE for wearable devices, employing spatial-temporal masking and an asymmetric autoencoder to model device correlations. In WiFi-specific contexts, Yang et al. [43] proposed MaskFi, which transforms multi-modal (WiFi-Vision) data into discrete tokens via a VQ-VAE and reconstructs them using a shared Transformer. Ji et al. [44] proposed SiFall, framing fall detection as anomaly detection by using a VAE to model non-fall activities and utilizing reconstruction error as the anomaly signal. Focusing on modal correlation, Gao et al. [14] proposed AutoSen, a cross-modal autoencoder that reconstructs phase from amplitude to capture intrinsic CSI semantics.

Despite avoiding complex augmentations, existing generative methods exhibit structural shortcomings. Approaches like MaskFi rely on heavy backbones (e.g., Transformers) that are computationally prohibitive and unstable on small, noisy CSI datasets [45]. Asymmetric designs, such as AutoSen's unidirectional prediction, create an information bottleneck by assuming amplitude fully specifies phase. Moreover, standard uniform random masking ignores the non-uniform information density of CSI, potentially wasting learning budgets on redundant regions [46].

III. Methodology

To address the challenges of data scarcity and non-uniform information density in WiFi sensing, we propose CIG-MAE, a self-supervised framework tailored to the physical representation of CSI. As illustrated in Fig. 3, the proposed framework processes synchronized dual-stream CSI through a *Filter-then-Reconstruct* paradigm. The pipeline is composed of three integral modules: an *AIM* mechanism that actively identifies information-dense signal patches; a *Dual-Stream Convolutional Backbone* that reconstructs the signal from these focused regions; and a *Cross-Modal Representation Alignment* module that enforces semantic consistency between modalities. The specific design motivations and details are described below.

A. Preliminaries: Signal Formulation and Properties

WiFi CSI characterizes the wireless channel at a fine-grained level using Orthogonal Frequency Division Multiplexing (OFDM). Mathematically, the CSI for subcarrier k at time t is represented as a complex value:

$$H_k(t) = A_k(t)e^{j\phi_k(t)} \tag{1}$$

where $A_k(t)$ is the amplitude and $\phi_k(t)$ is the phase. Consequently, the input to our model is defined as a synchronized dual-stream tensor $(\mathbf{X}^A, \mathbf{X}^P)$, where each $\mathbf{X}^m \in \mathbb{R}^{N\times S\times T}$ for modality $m \in \{A, P\}$, with N, S, and T denoting antennas, subcarriers, and time steps, respectively.

This formulation underscores two critical physical properties that inform our architectural design. First, the coupling of $A_k(t)$ and $\phi_k(t)$ implies that single-stream models inherently discard complementary modal information, necessitating a **symmetric dual-stream architecture**. Second, human activities typically manifest as locally structured patterns on the time-frequency plane ($S \times T$). Unlike vision tasks, CSI is heavily contaminated by background. This renders standard random masking inefficient, motivating the introduction of an adaptive masking strategy.

B. Adaptive Information-Guided Masking (AIM)

Standard masked autoencoders typically employ uniform random masking. However, given the high noise floor in CSI

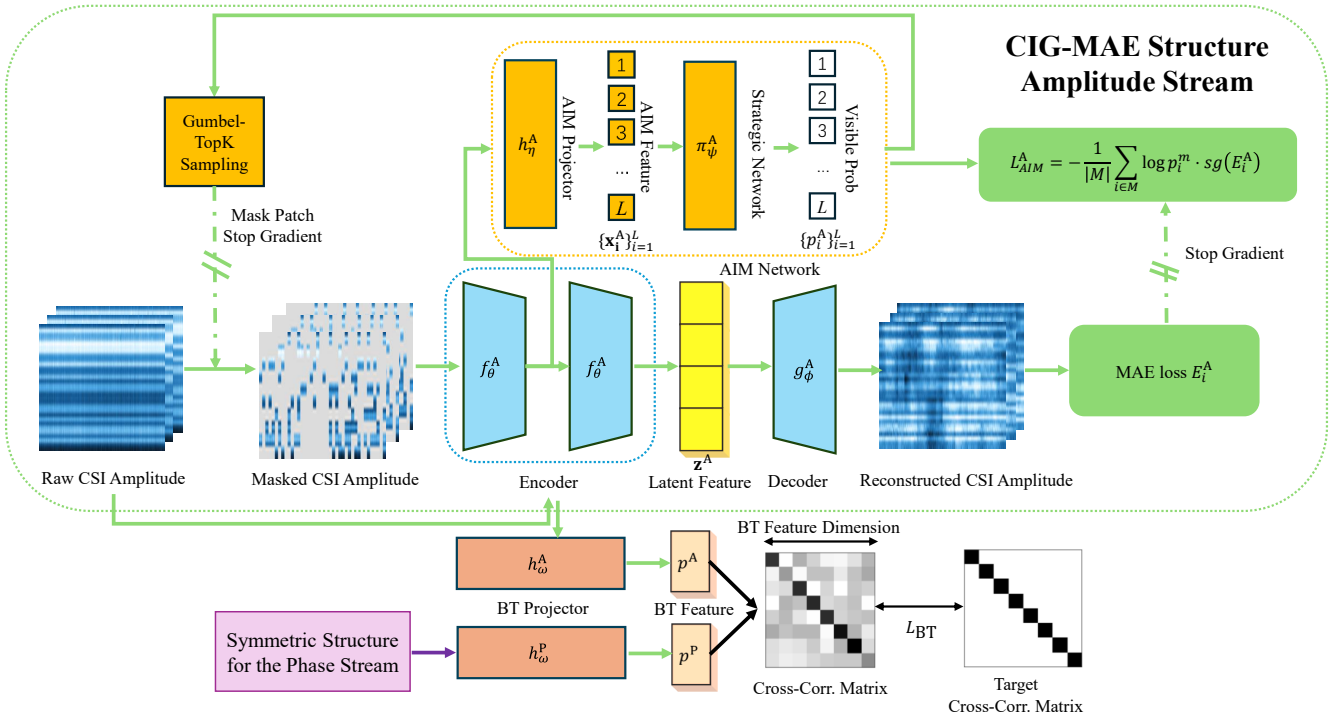


Fig. 3. Schematic overview of the CIG-MAE framework. The architecture integrates three core modules corresponding to Section III: (1) **Adaptive Information-Guided Masking (Sec. III-B)**: The policy network evaluates information density to generate non-uniform masks on the $S \times T$ plane; (2) **Dual-Stream Convolutional Backbone (Sec. III-C)**: Symmetric CNN encoder-decoder networks reconstruct the signal from masked amplitude and phase streams; (3) **Cross-Modal Representation Alignment (Sec. III-D)**: A Barlow Twins regularizer aligns unmasked representations to ensure semantic consistency.

data, such a strategy usually fails to distinguish between information-dense and redundant regions, leading to a significant waste of computational resources on meaningless reconstruction tasks [47], [48]. To mitigate this, we incorporate the AIM strategy [49] to dynamically allocate visibility to regions with high information density.

Implicit Patch Embedding. We first discretize the input $\mathbf{X}^m$ into a grid of patches. We define the patch indices as $\mathcal{L} = \{1, \ldots, L\}$, where each patch has dimensions (s_p, t_p). To avoid the computational overhead of a separate tokenizer, we strictly align the kernel size and stride of the encoder's first convolutional layer with (s_p, t_p). This yields native patch embeddings, which are projected to a policy feature space $\{\mathbf{x}_i^m\}_{i=1}^L$ via a linear layer h_η^m.

Policy Network and Probabilistic Sampling. A lightweight Policy Network π_ψ^m, implemented as a single-layer Transformer Encoder, evaluates the contextual importance of each patch. It aggregates global information via multi-head self-attention (MHA) and produces hidden states $\mathbf{U}$ via an MLP:

$$\mathbf{Z} = \{\mathbf{x}_i^m\} + \mathrm{MHA}(\mathrm{LN}(\{\mathbf{x}_i^m\})), \quad \mathbf{U} = \mathbf{Z} + \mathrm{MLP}(\mathrm{LN}(\mathbf{Z})). \tag{2}$$

The importance probability p_i^m (i.e., the probability of being **visible**) for the i-th patch is then derived via a softmax projection:

$$p_i^m = \frac{\exp(\mathbf{w}^\top \mathbf{U}_i)}{\sum_{j=1}^{L} \exp(\mathbf{w}^\top \mathbf{U}_j)}. \tag{3}$$

Based on $\{p_i^m\}$, we employ Gumbel-TopK sampling to partition the patch set $\mathcal{L}$ into a masked set $\mathcal{M}$ (where $|\mathcal{M}| = \rho L$, mask ratio $\rho \in (0, 1)$) and a visible set $\mathcal{V}$. This generates the binary mask Π applied to the input: $\tilde{\mathbf{X}}^m = \mathbf{X}^m \odot \Pi$.

Policy Optimization via Reconstruction Reward. The core innovation of AIM lies in how we define *valuable* patches without ground-truth labels. We propose a hypothesis rooted in information theory: *patches that are difficult to reconstruct from their context contain higher information density (e.g., human activity), whereas patches that are easily inferable correspond to redundant background.*

Guided by this intuition, we formulate the masking process as a reinforcement learning problem [50], [51]. We treat the mean absolute reconstruction error E_i^m of a patch i as a **reward signal**. Unlike adversarial approaches that simply mask salient regions, our goal is to identify high-error regions and increase their likelihood of being preserved in future iterations. Given the high masking ratio (e.g., 95%) in WiFi sensing, random masking risks removing the sparse activity signal entirely. By explicitly encouraging the network to select difficult-to-reconstruct regions into the visible set $\mathcal{V}$,

we ensure the model focuses its limited visibility budget on capturing the discriminative "skeleton" of the signal.

Formally, we define the policy gradient loss to maximize the expected reconstruction error of the sampled masked set $\mathcal{M}$ with respect to the visibility probabilities. By minimizing the following negative objective:

$$\mathcal{L}_{\text{AIM}}^{m} = -\frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \log p_i^m \cdot \text{stopgrad}(E_i^m), \tag{4}$$

we effectively encourage the policy network to assign higher visibility probabilities p_i^m to patches that yield high reconstruction errors E_i^m when masked. Note that we apply a stopgrad operator to the reward E_i^m. This decouples the optimization: it prevents the backbone from artificially increasing reconstruction error to satisfy the policy's objective, ensuring that the encoder focuses solely on minimizing reconstruction loss while the policy independently learns to identify and preserve information-dense examples.

C. Dual-Stream Convolutional Reconstruction Backbone

The core of CIG-MAE is the reconstruction of the original signal from the partial input filtered by AIM. We deliberately employ a pure Convolutional Neural Network (CNN) encoder-decoder architecture. This choice is substantiated by both physical characteristics and practical constraints:

1) **Inductive Bias Match:** CNNs capture local structural and translation-equivariant features, which aligns with CSI patterns [52]. In contrast, Transformer-based backbones (ViTs) model global dependencies and require massive datasets to learn such local priors [53]. However, existing CSI datasets are limited in scale, often ranging from hundreds to thousands of samples [54]–[57], with only a few reaching tens of thousands [58]. This scarcity is insufficient to train data-hungry Transformers effectively.
2) **Deployment Efficiency:** WiFi sensing applications are typically deployed on resource-constrained edge devices [59]. As empirically validated in Section IV-G, our CNN backbone requires significantly fewer FLOPs and parameters compared to ViTs, making real-time inference feasible without sacrificing accuracy.

Reconstruction Process. The masked input $\tilde{\mathbf{X}}^m$ is mapped by the encoder f_θ^m to a latent representation $\mathbf{z}^m$, and subsequently reconstructed by the decoder g_ϕ^m to yield $\hat{\mathbf{X}}^m$. To quantify quality, we compute the Mean Absolute Error (MAE) exclusively on the masked pixels $\Omega(\Pi) = \{(n, s, t) \mid \Pi(s, t) = 0\}$:

$$\mathcal{L}_{\text{rec}}^{m} = \frac{1}{|\Omega(\Pi)|} \sum_{(n,s,t) \in \Omega(\Pi)} \left| \hat{\mathbf{X}}_{n,s,t}^m - \mathbf{X}_{n,s,t}^m \right|. \tag{5}$$

The choice of an unnormalized reconstruction target combined with the MAE function is deliberately tailored to the physical properties of CSI. First, regarding the loss metric, we employ MAE over the standard Mean Squared Error (MSE). CSI signals are prone to sporadic, high-amplitude noise spikes due to hardware imperfections and multipath interference. MSE imposes a quadratic penalty on such outliers, which can dominate the gradient and destabilize training. In contrast, MAE's linear penalty provides robustness, ensuring the model focuses on learning the underlying activity structure rather than fitting these transient artifacts. Second, regarding the target, we avoid per-sample normalization during loss computation. Normalization tends to compress the signal's dynamic range, potentially obscuring the semantic distinction between high-magnitude activity bursts and low-magnitude background fluctuations. By preserving the full dynamic range, we force the model to prioritize the reconstruction of significant, activity-induced signal variations. The empirical superiority of this design is validated in Section IV-E.

D. Cross-Modal Representation Alignment

While the dual streams independently reconstruct their respective modalities, it is crucial that their learned latent representations are semantically consistent. To achieve this without relying on negative samples, we introduce a BT regularizer [24].

To ensure the statistical stability of the correlation estimates, this alignment process is performed on the original, *unmasked* input $\mathbf{X}^m$, paralleling the reconstruction path. For each modality $m \in \{A, P\}$, the projection vector $\mathbf{p}^m$ is obtained through the cascaded mapping of the encoder f_θ^m and the BT projection head h_ω^m:

$$\mathbf{p}^m = h_\omega^m(f_\theta^m(\mathbf{X}^m)). \tag{6}$$

Within a data batch of size B, we assess the alignment between the two streams by computing the cross-correlation matrix $\mathbf{C}$:

$$\mathbf{C}_{ij} = \frac{1}{B} \sum_{b=1}^{B} \tilde{\mathbf{p}}_{b,i}^A \, \tilde{\mathbf{p}}_{b,j}^P, \tag{7}$$

where $\tilde{\mathbf{p}}$ denotes the representations after batch normalization. The BT loss is then defined to strictly penalize deviations from the identity matrix:

$$\mathcal{L}_{\text{BT}} = \sum_i (1 - \mathbf{C}_{ii})^2 + \lambda \sum_{i \neq j} \mathbf{C}_{ij}^2. \tag{8}$$

The first term (invariance) forces the diagonal elements to 1, aligning the representations of amplitude and phase. The second term (redundancy reduction) forces off-diagonal elements to 0, decorrelating the feature dimensions to maximize information content.

E. Overall Training Objective

The total loss is a weighted sum of the three objectives:

$$\mathcal{L}_{\text{CIG-MAE}} = w_{\text{rec}} \mathcal{L}_{\text{rec}} + w_{\text{aim}} \mathcal{L}_{\text{AIM}} + w_{\text{bt}} \mathcal{L}_{\text{BT}}. \tag{9}$$

Algorithm 1 CIG-MAE Pre-training Procedure

Require: Data $(\mathbf{X}^A, \mathbf{X}^P)$, Backbones $\Theta = \{f_\theta, g_\phi, h_\omega\}$, Policy Nets $\Psi = \{\pi_\psi, h_\eta\}$
Ensure: Optimized Encoders f_θ^A, f_θ^P

while not converged **do**
  Sample batch $(\mathbf{X}^A, \mathbf{X}^P)$
  // Phase 1: Adaptive Masking (Sec. III-B)
  // Policy network π_ψ estimates importance; $\mathcal{M}$ is sampled via Gumbel-TopK
  $\mathcal{M}^A \leftarrow \text{AIM}(\mathbf{X}^A, \Psi^A)$; $\mathcal{M}^P \leftarrow \text{AIM}(\mathbf{X}^P, \Psi^P)$
  // Phase 2: Dual-Stream Reconstruction (Sec. III-C)
  // Apply mask and reconstruct using Encoder f_θ and Decoder g_ϕ
  $\tilde{\mathbf{X}}^m \leftarrow \mathbf{X}^m \odot (\mathbf{1} - \mathcal{M}^m)$ for $m \in \{A, P\}$
  $\hat{\mathbf{X}}^m \leftarrow g_\phi^m(f_\theta^m(\tilde{\mathbf{X}}^m))$ for $m \in \{A, P\}$
  $\mathcal{L}_{\text{rec}} \leftarrow \text{MaskedMAE}(\hat{\mathbf{X}}^{A,P}, \mathbf{X}^{A,P}, \mathcal{M}^{A,P})$
  // Phase 3: Cross-Modal Alignment (Sec. III-D)
  // Project unmasked features via h_ω for Barlow Twins
  $\mathbf{p}^A \leftarrow h_\omega^A(f_\theta^A(\mathbf{X}^A))$; $\mathbf{p}^P \leftarrow h_\omega^P(f_\theta^P(\mathbf{X}^P))$
  $\mathcal{L}_{\text{BT}} \leftarrow \text{BTLoss}(\mathbf{p}^A, \mathbf{p}^P)$
  // Phase 4: Decoupled Parameter Optimization
  // 1. Update Policy Ψ: maximize reconstruction error of masked patches
  $E^{A,P} \leftarrow \text{PixelWiseError}(\hat{\mathbf{X}}^{A,P}, \mathbf{X}^{A,P})$
  $\mathcal{L}_{\text{AIM}} \leftarrow \text{PolicyGradient}(E^{A,P}, \mathcal{M}^{A,P}, \Psi)$
  $\Psi \leftarrow \text{Optimizer}(\nabla_\Psi(w_{\text{aim}}\mathcal{L}_{\text{AIM}}))$
  // 2. Update Backbone Θ: minimize reconstruction and redundancy
  $\Theta \leftarrow \text{Optimizer}(\nabla_\Theta(\mathcal{L}_{\text{rec}} + w_{\text{bt}}\mathcal{L}_{\text{BT}}))$
end while
return Optimized encoders f_θ^A, f_θ^P

We employ a **decoupled update strategy** to distinguish the optimization targets of the generative and adversarial components. Specifically, the backbone parameters $\Theta = \{f_\theta, g_\phi, h_\omega\}$ are updated by minimizing the reconstruction and alignment losses ($\mathcal{L}_{\text{rec}} + w_{\text{bt}}\mathcal{L}_{\text{BT}}$), while the policy parameters $\Psi = \{\pi_\psi, h_\eta\}$ are updated solely by the policy gradient loss $\mathcal{L}_{\text{AIM}}$. This separation ensures that the policy network independently learns to identify information-dense regions without interference from the reconstruction objective. The complete pre-training procedure is summarized in Algorithm 1.

IV. Experiments

This section systematically evaluates the proposed CIG-MAE framework. First, its effectiveness is validated by comparing it with several SOTA self-supervised learning methods on three public CSI datasets. Subsequently, the model's behavior and performance are analyzed through ablation studies, hyperparameter sensitivity analysis, and settings with varying amounts of labeled data.

A. Experimental Setup

Datasets. We evaluate CIG-MAE on three datasets representing diverse sensing scenarios:

- **SignFi** [55] (Sign Language): A subset from User 5 (home environment), comprising 2,760 samples across 276 classes, collected using an 802.11n CSI tool [60]. The input tensor shape is $3 \times 30 \times 200$.
- **NLOS** [54] (Daily Activities): Data from Environment 1, consolidated into 6 classes (3,000 samples) following [61], [62]. Variable durations $[754, 1601]$ were downsampled to 200.
- **HTHI** [63] (Human-Interaction): Contains 4,800 trials of 12 interaction types from 40 pairs. Durations $[1040, 2249]$ were downsampled to 200.

Protocol & Implementation. We implement the model in PyTorch on an NVIDIA RTX 5090. Datasets are stratified (8:2 train/test split). Raw CSI phase is linearly calibrated [64], and both modalities are z-score normalized. We adhere to the standard k-shot linear probing protocol: pre-trained encoders are frozen, and concatenated representations are fed to a linear classifier. **Pre-training** runs for 300 epochs (batch size 256) using AdamW ($lr = 1 \times 10^{-4}$, weight decay 0.01, $\beta = (0.9, 0.95)$). We set the mask ratio $\rho = 0.95$, patch size $(3, 5)$, and loss weights $w_{\text{rec}} = 1, w_{\text{bt}} = 0.2, w_{\text{aim}} = 1 \times 10^{-4}$. The AIM policy features $\mathbf{x}^m$ and encoder latent features $\mathbf{z}^m$ are both 256. Critically, the BT projection head is a 3-layer MLP with width 1024 (SignFi/HTHI) or 5096 (NLOS). **Downstream** evaluation uses 1-shot (SignFi) or 10-shot (NLOS/HTHI) learning for 100 epochs (batch size 32, $lr = 1 \times 10^{-3}$). Architecture details are provided in Table I.

B. Baselines

We compare CIG-MAE against a fully supervised upper bound and three SOTA self-supervised methods:

- **Supervised Model**: Trained end-to-end on the full labeled dataset using the exact same backbone as CIG-MAE, serving as the performance upper bound.
- **AutoFi** [31]: A contrastive method that jointly optimizes probabilistic consistency, mutual information maximization, and geometric structure consistency.
- **AutoSen** [14]: A cross-modal autoencoder that learns joint representations through the asymmetric proxy task of reconstructing phase from amplitude.
- **CAPC** [35]: Combines CPC with BT. It predicts future latent representations to capture temporal context while using BT to reduce feature redundancy.

C. Performance Comparison

Table II compares CIG-MAE against SOTA SSL methods and a supervised baseline. CIG-MAE consistently outperforms all counterparts. Notably, on SignFi, it achieves 98.55% accuracy, surpassing even the supervised baseline (96.38%), indicating superior generalization. This advantage

TABLE I
NETWORK ARCHITECTURE USED IN CIG-MAE. E, D, AND F REPRESENT ENCODER, DECODER, AND CLASSIFIER, RESPECTIVELY.

Layer	E **(Encoder)**	D **(Decoder)**	F **(Classifier)**
Input CSI: $3(6) \times 30 \times 200$ (antenna $\times$ subcarrier $\times$ timestamp)			
1	Conv, $C = 128$, $K = (3, 5)$, $S = (3, 5)$	MLP: $256 \to 512 \times 5 \times 10$	FC: $512 \to C$, Softmax
2	Conv, $C = 256$, $K = (2, 2)$, $S = (2, 2)$	DeConv, $C = 512$, $K = (1, 2)$, $S = (1, 2)$	—
3	Conv, $C = 512$, $K = (1, 2)$, $S = (1, 2)$	DeConv, $C = 256$, $K = (2, 2)$, $S = (2, 2)$	—
4	MLP: $512 \times 5 \times 10 \to 256$	DeConv, $C = 128$, $K = (3, 5)$, $S = (3, 5)$	—

widens in complex scenarios (NLOS and HTHI), where CIG-MAE maintains robustness with accuracies of 63.50% and 40.10% respectively, establishing a clear margin over all SSL baselines.

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT METHODS ON SIGNFI, NLOS, AND HTHI DATASETS WITH A LINEAR CLASSIFIER.

Method	SignFi		NLOS		HTHI	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
Supervised	96.38	95.75	99.83	99.86	97.50	97.50
AutoFi	94.75	94.37	48.00	46.67	32.08	31.43
AutoSen	89.49	88.75	51.83	50.71	32.60	31.25
CAPC	94.20	94.60	47.49	47.58	35.72	36.33
CIG-MAE	98.55	98.37	63.50	62.35	40.10	40.05

The performance gap stems from the misalignment of baselines with CSI characteristics:

- **AutoFi** falters in multi-user settings (NLOS, HTHI) as its contrastive paradigm relies on simplistic augmentations insufficient to handle high intra-class variance.
- **CAPC** imposes a restrictive temporal continuity bias, causing it to neglect self-contained spatio-temporal patterns critical for fine-grained actions.
- **AutoSen**'s asymmetric design creates an information bottleneck by assuming amplitude fully specifies phase, thereby losing unique phase-dependent information.

In contrast, CIG-MAE overcomes these limitations through a synergistic design. Its masked autoencoding paradigm captures holistic intrinsic structures rather than brittle inter-sample relations. The symmetric, BT-regularized architecture ensures complete, decorrelated fusion of amplitude and phase. Crucially, AIM dynamically focuses the learning budget on discriminative regions, ensuring robustness against noise where other models fail.

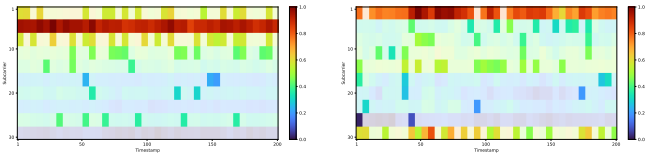
D. Ablation Study

We validate the contribution of each component by systematically dismantling the full model (Table III).

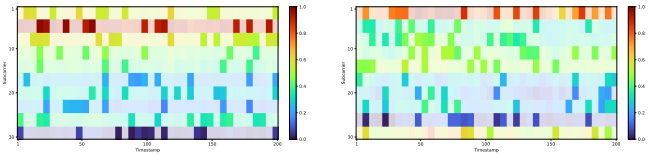
Impact of AIM. Replacing AIM with random masking (*w/o AIM*) degrades performance across all datasets, most notably dropping 5.42% accuracy on HTHI. This confirms that focusing on information-dense regions is critical. Fig. 4 visualizes this mechanism: AIM preserves activity-induced signal changes while masking redundant background. Interestingly, on HTHI, *w/o AIM* (which retains BT) performs worse (34.68%) than the unconstrained *Dual-Stream MAE* (37.60%). This indicates that enforcing statistical decorrelation without guiding the model to salient features can be counterproductive.

TABLE III
THE PERFORMANCE WITH DIFFERENT DESIGN CHOICES FOR SIGNFI, NLOS, AND HTHI DATA.

Method	SignFi		NLOS		HTHI	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
Single-Stream (Amp.)	95.65	95.48	60.33	59.57	37.08	37.73
Dual-Stream MAE	96.37	96.16	55.50	55.37	37.60	38.33
w/o BT	97.10	96.73	50.83	53.99	37.70	38.17
w/o AIM	96.92	96.72	61.17	62.27	34.68	35.59
CIG-MAE	98.55	98.37	63.50	62.35	40.10	40.05



(a) AIM-guided Amplitude Masking (b) AIM-guided Phase Masking

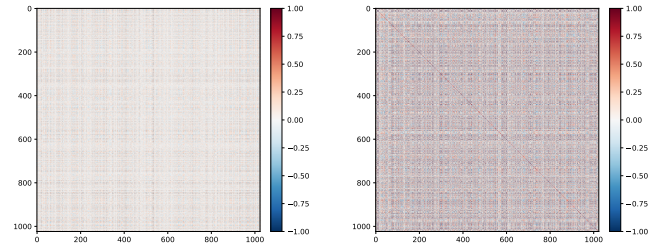


(c) Random Amplitude Masking (d) Random Phase Masking

Fig. 4. Demonstration of AIM's visibility allocation. Compared to random masking, AIM suppresses low-information regions and increases the visibility of high-information regions, concentrating the learning budget on critical time-frequency blocks related to the activity, thereby obtaining more discriminative representations.

Impact of BT Regularizer. BT is essential for coordinating the dual streams. Removing it (*w/o BT*) causes a catastrophic 12.67% accuracy drop on NLOS. Furthermore, *w/o BT* underperforms the basic *Dual-Stream MAE* (50.83% vs 55.50%), suggesting that in noisy conditions, AIM without alignment constraints may guide streams toward divergent, modality-specific artifacts. Fig. 5 demonstrates how BT ef-

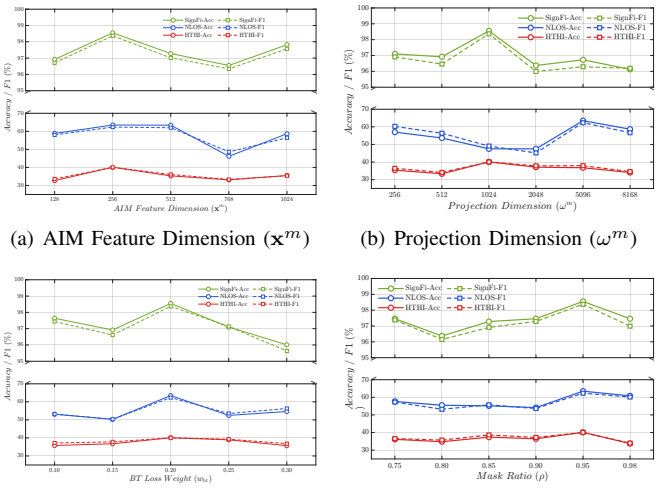
fectively aligns representations and reduces redundancy.



(a) Amplitude-Phase Correlation Matrix without BT Constraint (b) Amplitude-Phase Correlation Matrix with BT Constraint

Fig. 5. Demonstration of the alignment and decorrelation effect of BT. After introducing BT, the diagonal elements of the correlation matrix approach 1, and the off-diagonal elements approach 0, reflecting enhanced consistency and reduced redundancy in the corresponding dimensions of the two streams.

Dual-Stream Architecture. Comparison between *Single-Stream* and *Dual-Stream MAE* reveals that naively combining amplitude and phase improves performance on SignFi/HTHI but degrades it on NLOS. This conflict underscores the necessity of the BT regularizer to resolve feature redundancy, validating the synergistic design of the full CIG-MAE model.



(a) AIM Feature Dimension ($\mathbf{x}^m$) (b) Projection Dimension (ω^m)

(c) BT Loss Weight (w_{bt}) (d) Mask Ratio (ρ)

Fig. 6. Sensitivity analysis of CIG-MAE's key hyperparameters across the SignFi, NLOS, and HTHI datasets. The results show that while some parameters are stable, others adapt to data characteristics, demonstrating the model's robustness and flexibility. (a) AIM feature dimension; (b) BT projection head dimension; (c) BT loss weight; (d) Mask ratio.

E. Sensitivity Analysis

We analyze key hyperparameters in Fig. 6 and reconstruction objectives in Table IV.

Hyperparameters. The AIM feature dimension ($\mathbf{x}^m$) consistently peaks at 256; lower dimensions lack expressive capacity, while higher ones risk overfitting to spurious correlations. The BT loss weight (w_{bt}) stabilizes at 0.2, striking a critical balance: lower weights provide insufficient regularization, while higher weights over-optimize decorrelation at the expense of preserving fine-grained signal details. In contrast, the projection head dimension (ω^m) adapts to dataset complexity. While 1024 suffices for SignFi/HTHI, the noisier NLOS dataset requires a larger capacity (5096) to effectively disentangle features. Finally, a high mask ratio ($\rho = 0.95$) is consistently optimal, confirming that forcing reconstruction from sparse inputs (5%) compels the model to learn deep intrinsic structures rather than memorizing superficial patterns.

TABLE IV
IMPACT OF DIFFERENT RECONSTRUCTION LOSS FUNCTIONS ON CIG-MAE PERFORMANCE FOR SIGNFI, NLOS, AND HTHI DATA.

Method	SignFi		NLOS		HTHI	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
MAE (w/o Norm.)	98.55	98.37	63.50	62.35	40.10	40.05
MAE (w/ Norm.)	96.37	96.30	54.66	54.33	35.10	35.09
MSE (w/o Norm.)	95.10	94.75	46.66	48.45	38.64	39.96
MSE (w/ Norm.)	95.11	94.75	57.66	56.72	33.85	34.90

Loss Function. Table IV shows that *MAE (w/o Norm.)* consistently yields the best performance. MAE is preferred over MSE because its linear penalty is robust to sporadic high-amplitude noise spikes (outliers), avoiding the gradient dominance caused by quadratic penalties. Furthermore, omitting normalization preserves the original dynamic range, which is crucial for distinguishing high-magnitude activity bursts from low-magnitude background fluctuations.

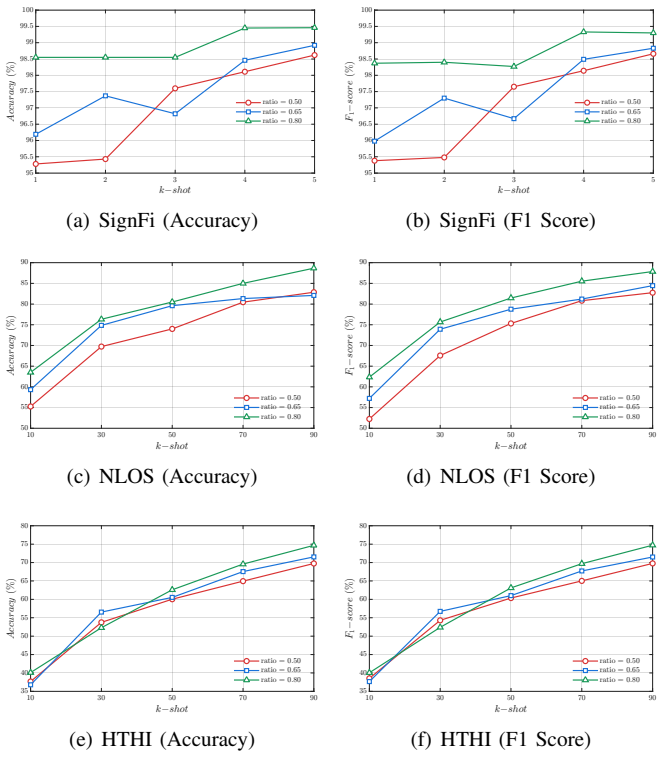
F. Data Efficiency Analysis

We analyze the impact of labeled fine-tuning data size (k) and unlabeled pre-training data scale (50%, 65%, 80%) in Fig. 7.

First, performance monotonically improves with labeled samples (k) across all settings. This confirms that CIG-MAE learns a meaningful, linearly separable feature space where the classifier can effectively leverage additional supervision to refine decision boundaries.

Second, the impact of pre-training data volume depends heavily on dataset complexity. On the clean SignFi dataset, the gain from 50% to 80% data is minimal, suggesting diminishing returns. Conversely, on the noisy NLOS and HTHI datasets, the gap is substantial; for instance, in the NLOS 10-shot scenario, increasing data from 50% to 80% boosts the F1-score from 52.24% to 62.35%. This highlights that diverse unlabeled examples are essential for learning invariant features in complex scenarios.

Crucially, on NLOS and HTHI, the performance advantage of larger pre-training data *widens* as k increases. This indicates that a superior pre-training foundation creates a better representation space with a higher performance ceiling, allowing the model to continuously capitalize on labeled data without the early plateauing observed in models trained on smaller subsets.



(a) SignFi (Accuracy) (b) SignFi (F1 Score)

(c) NLOS (Accuracy) (d) NLOS (F1 Score)

(e) HTHI (Accuracy) (f) HTHI (F1 Score)

Fig. 7. Impact of unlabeled pre-training data size and labeled fine-tuning data size on model performance. The x-axis represents the number of labeled samples per class used in k-shot learning, and different curves correspond to different pre-training data ratios (50%, 65%, 80%). (a) SignFi (Accuracy); (b) SignFi (F1 Score); (c) NLOS (Accuracy); (d) NLOS (F1 Score); (e) HTHI (Accuracy); (f) HTHI (F1 Score).

G. Backbone Comparison

To validate the suitability of our design for practical IoT deployment, we benchmark the proposed CNN backbone against a ViT-based counterpart. Specifically, we replaced the CNN encoder-decoder with a ViT equivalent following the standard Masked Autoencoder design [16], [53], while keeping other components unchanged. The ViT baseline consists of an encoder with 6 blocks (1024 embedding dim, 16 heads) and a decoder with 4 blocks (512 dim, 8 heads). The comprehensive comparison is detailed in Table V.

TABLE V
PERFORMANCE AND EFFICIENCY COMPARISON BETWEEN THE PROPOSED CNN BACKBONE AND THE VIT BASELINE.

Method	SignFi		NLOS		HTHI	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
CNN (Ours)	**98.55**	**98.37**	**63.50**	**62.35**	**40.10**	**40.05**
ViT	89.85	89.06	34.16	32.89	34.27	32.88
Params	**CNN: 53.46MB**		ViT: 608MB			
FLOPs	**CNN: 0.14G**		ViT: 129.98G			
Memory	**CNN: 131.97MB**		ViT: 580.13MB			

Performance Superiority. The CNN-based model substantially outperforms the ViT counterpart across all datasets (e.g., +29.34% accuracy on NLOS). This gap stems from the inherent inductive biases of CNNs, such as locality and translation equivariance, which are highly compatible with the structured features of CSI time-frequency maps. In contrast, ViT lacks these priors and struggles to learn effectively from the limited scale of CSI datasets.

Edge Deployment Feasibility. Critically, the CNN architecture demonstrates superior efficiency, utilizing approximately $10\times$ fewer parameters and $1000\times$ fewer FLOPs than the ViT. This extreme computational economy is decisive for WiFi sensing: it enables the framework to be deployed directly on resource-constrained edge devices (e.g., commercial WiFi access points), facilitating ubiquitous, real-time sensing without expensive hardware accelerators.

V. CONCLUSION

This paper proposes CIG-MAE, a generative self-supervised framework tailored for CSI-HAR. By integrating dual-stream masked reconstruction, AIM, and BT regularization, CIG-MAE effectively leverages the complementarity of amplitude and phase. This design eliminates the reliance on problematic data augmentations and negative samples while capturing intrinsic signal structures. Experiments on three public datasets demonstrate that CIG-MAE consistently achieves SOTA performance, surpassing supervised baselines in generalization. Critically, our comparison with ViT backbones confirms that the proposed CNN-based architecture offers superior accuracy with orders of magnitude lower computational cost ($1000\times$ fewer FLOPs), establishing a feasible path for deploying advanced sensing algorithms on resource-constrained edge devices. Future work will focus on two directions: evaluating scalability on larger, diverse datasets to further test generalization, and extending the framework to complex multi-person scenarios. Additionally, we plan to explore model compression techniques to further optimize real-time inference on commercial WiFi hardware.

REFERENCES

[1] Z. Wang, J. A. Zhang, M. Xu, and Y. J. Guo, "Single-Target Real-Time Passive WiFi Tracking," *IEEE Trans. Mob. Comput.*, vol. 22, no. 6, pp. 3724–3742, Jun. 2023.

[2] Y. Gu, X. Zhang, Z. Liu, and F. Ren, "WiFi-Based Real-Time Breathing and Heart Rate Monitoring During Sleep," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[3] S. Yue, Y. Yang, H. Wang, H. Rahul, and D. Katabi, "BodyCompass: Monitoring sleep posture with wireless signals," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 2, p. 66, Jun. 2020.

[4] H. Wang, D. Zhang, Y. Wang, J. Ma, Y. Wang, and S. Li, "RT-Fall: A real-time and contactless fall detection system with commodity WiFi devices," *IEEE Trans. Mob. Comput.*, vol. 16, no. 2, pp. 511–526, Feb. 2017.

[5] L. Guo, Z. Lu, S. Zhou, X. Wen, and Z. He, "Emergency Semantic Feature Vector Extraction From WiFi Signals for In-Home Monitoring of Elderly," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 6, pp. 1423–1438, Nov. 2021.

[6] C. Wu, B. Wang, O. C. Au, and K. J. R. Liu, "Wi-fi can do more: Toward ubiquitous wireless sensing," *IEEE Communications Standards Magazine*, vol. 6, no. 2, pp. 42–49, Jun. 2022.

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

[8] S. G. Dhekane and T. Ploetz, "Transfer Learning in Human Activity Recognition: A Survey," 2024, arXiv preprint arXiv:2401.10185.

[9] W. Jiang, C. Miao, F. Ma, S. Yao, Y. Wang, Y. Yuan, H. Xue, C. Song, X. Ma, D. Koutsonikolas, W. Xu, and L. Su, "Towards Environment Independent Device Free Human Activity Recognition," in *Proc. 24th Annu. Int. Conf. Mobile Comput. Netw.*, 2018, pp. 289–304.

[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," 2024, arXiv preprint arXiv:2301.05712.

[11] K. Xu, J. Wang, H. Zhu, and D. Zheng, "Self-Supervised Learning for WiFi CSI-Based Human Activity Recognition: A Systematic Study," 2023, arXiv preprint arXiv:2308.02412.

[12] C. Xiao, Y. Han, W. Yang, Y. Hou, F. Shi, and K. Chetty, "Diffusion-Model-Based Contrastive Learning for Human Activity Recognition," *IEEE Internet Things J.*, vol. 11, no. 20, pp. 33 525–33 536, Oct. 2024.

[13] B. Chen, J. Wang, Y. Lv, Q. Gao, M. Pan, and Y. Fang, "Device-Free Wireless Sensing With Few Labels Through Mutual Information Maximization," *IEEE Internet Things J.*, vol. 11, no. 6, pp. 10 513–10 524, Mar. 2024.

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

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2019, arXiv preprint arXiv:1810.04805.

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

[17] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," 2022, arXiv preprint arXiv:2106.08254.

[18] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, T. N. Sainath, and S. Watanabe, "Self-Supervised Speech Representation Learning: A Review," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1179–1210, Oct. 2022.

[19] D. Cheng, L. Zhang, L. Qin, S. Wang, H. Wu, and A. Song, "MaskCAE: Masked convolutional autoencoder via sensor data reconstruction for self-supervised human activity recognition," *IEEE J. Biomed. Health Inform.*, vol. 28, no. 5, pp. 2687–2698, May 2024.

[20] A. Saeed, V. Ungureanu, and B. Gfeller, "Sense and Learn: Self-supervision for omnipresent sensors," *Mach. Learn. Appl.*, vol. 6, p. 100152, 2021.

[21] R. Song, D. Zhang, Z. Wu, C. Yu, C. Xie, S. Yang, Y. Hu, and Y. Chen, "RF-URL: unsupervised representation learning for RF sensing," in *Proc. 28th Annu. Int. Conf. Mobile Comput. Netw.*, 2022, pp. 282–295.

[22] A. Saeed, F. D. Salim, T. Ozcelebi, and J. Lukkien, "Federated Self-Supervised Learning of Multisensor Representations for Embedded Intelligence," *IEEE Internet Things J.*, vol. 8, no. 2, pp. 1030–1040, Jan. 2021.

[23] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," 2020, arXiv preprint arXiv:1911.05722.

[24] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," *arXiv preprint arXiv:2103.03230*, 2021.

[25] A. Bardes, J. Ponce, and Y. LeCun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," *arXiv preprint arXiv:2105.04906*, 2022.

[26] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," 2020, arXiv preprint arXiv:2002.05709.

[27] M. J. Bocus, H.-S. Lau, R. McConville, R. J. Piechocki, and R. Santos-Rodriguez, "Self-Supervised WiFi-Based Activity Recognition," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2022, pp. 552–557.

[28] D. Liu, T. Wang, S. Liu, R. Wang, S. Yao, and T. Abdelzaher, "Contrastive Self-Supervised Representation Learning for Sensing Signals from the Time-Frequency Perspective," in *Proc. Int. Conf. Comput. Commun. Netw. (ICCCN)*, Jul. 2021, pp. 1–10.

[29] K. Xu, J. Wang, L. Zhang, H. Zhu, and D. Zheng, "Dual-Stream Contrastive Learning for Channel State Information Based Human Activity Recognition," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 1, pp. 329–338, Jan. 2023.

[30] N. Lyons, A. Santra, V. K. Ramanna, K. Uln, R. Taori, and A. Pandey, "WiFiAct: Enhancing Human Sensing Through Environment Robust Preprocessing and Bayesian Self-Supervised Learning," in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, April 2024, pp. 13 391–13 395.

[31] J. Yang, X. Chen, H. Zou, D. Wang, and L. Xie, "AutoFi: Toward automatic Wi-Fi human sensing via geometric self-supervised learning," *IEEE Internet Things J.*, vol. 10, no. 8, pp. 7416–7425, Apr. 2023.

[32] A. van den Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," 2019, arXiv preprint arXiv:1807.03748.

[33] K. Zhang, Q. Wen, C. Zhang, R. Cai, M. Jin, Y. Liu, J. Y. Zhang, Y. Liang, G. Pang, D. Song, and S. Pan, "Self-supervised learning for time series analysis: Taxonomy, progress, and prospects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 10, pp. 6775–6794, 2024.

[34] H. Haresamudram, I. Essa, and T. Ploetz, "Contrastive Predictive Coding for Human Activity Recognition," 2020, arXiv preprint arXiv:2012.05333.

[35] B. Barahimi, H. Tabassum, M. Omer, and O. Waqar, "Context-Aware Predictive Coding: A Representation Learning Framework for WiFi Sensing," 2024, arXiv preprint arXiv:2410.01825.

[36] J. Strohmayer and M. Kampel, *Data Augmentation Techniques for Cross-Domain WiFi CSI-Based Human Activity Recognition*, 2024, pp. 42–56.

[37] O. G. Serbetci, J.-H. Lee, D. Burghal, and A. F. Molisch, "Simple and Effective Augmentation Methods for CSI Based Indoor Localization," 2023, arXiv preprint arXiv:2211.10790.

[38] W. Hou and C. Wu, "RFBoost: Understanding and boosting deep WiFi sensing via physical data augmentation," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 8, no. 2, pp. 1–26, May 2024.

[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[40] 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.*, pp. 1–1, 2025.

[41] X. Kong and X. Zhang, "Understanding Masked Image Modeling via Learning Occlusion Invariant Feature," 2022, arXiv preprint arXiv:2208.04164.

[42] S. Miao, L. Chen, and R. Hu, "Spatial-Temporal Masked Autoencoder for Multi-Device Wearable Human Activity Recognition," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 7, no. 4, p. 172, Jan. 2024.

[43] J. Yang, S. Tang, Y. Xu, Y. Zhou, and L. Xie, "MaskFi: Unsupervised learning of WiFi and vision representations for multimodal human activity recognition," 2024, arXiv preprint arXiv:2402.19258.

[44] S. Ji, Y. Xie, and M. Li, "SiFall: Practical online fall detection with RF sensing," in *Proc. 20th ACM Conf. Embedded Netw. Sensor Syst.*, Nov. 2022, pp. 563–577.

[45] Y. Liu, E. Sangineto, W. Bi, N. Sebe, B. Lepri, and M. D. Nadai, "Efficient Training of Visual Transformers with Small Datasets," 2021, arXiv preprint arXiv:2106.03746.

[46] I. Kakogeorgiou, S. Gidaris, B. Psomas, Y. Avrithis, A. Bursuc, K. Karantzalos, and N. Komodakis, *What to Hide from Your Students: Attention-Guided Masked Image Modeling*, 2022, pp. 300–318.

[47] Z. Yang, Y. Zhang, K. Qian, and C. Wu, "SLNet: A spectrogram learning neural network for deep wireless sensing," in *Proc. 20th USENIX Symp. Netw. Syst. Des. Implementation (NSDI 23)*, Apr. 2023, pp. 1221–1236.

[48] Y. Taso, S.-C. Yeh, Y.-Y. Liang, C.-H. Wang, and S.-H. Fang, "Subcarrier selection for efficient CSI-based indoor localization," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 383, no. 1, p. 012017, Jul. 2018.

[49] W. G. C. Bandara, N. Patel, A. Gholami, M. Nikkhah, M. Agrawal, and V. M. Patel, "AdaMAE: Adaptive masking for efficient spatiotemporal learning with masked autoencoders," 2022, arXiv preprint arXiv:2211.09120.
[50] Z. Zheng, J. Oh, and S. Singh, "On Learning Intrinsic Rewards for Policy Gradient Methods," 2018, arXiv preprint arXiv:1804.06459.
[51] M. J. Tyszkiewicz, P. Fua, and E. Trulls, "DISK: Learning local features with policy gradient," 2020, arXiv preprint arXiv:2006.13566.
[52] 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," 2023, arXiv preprint arXiv:2207.07859.
[53] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2021, arXiv preprint arXiv:2010.11929.
[54] 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 in Brief*, vol. 33, p. 106534, 2020.
[55] 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, Mar. 2018.
[56] 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, Oct. 2017.
[57] S. Huang, K. Li, D. You, Y. Chen, A. Lin, S. Liu, X. Li, and J. A. McCann, "WiMANS: A benchmark dataset for WiFi-based multi-user activity sensing," 2024, arXiv preprint arXiv:2402.09430.
[58] Z. Yang, Y. Zhang, G. Zhang, Y. Zheng, and G. Chi, "Widar 3.0: WiFi-based activity recognition dataset," IEEE Dataport, 2020.
[59] O. Jouini, K. Sethom, A. Namoun, N. Aljohani, M. H. Alanazi, and M. N. Alanazi, "A Survey of Machine Learning in Edge Computing: Techniques, Frameworks, Applications, Issues, and Research Directions," *Technologies*, vol. 12, no. 6, p. 81, 2024.
[60] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Tool release: Gathering 802.11n traces with channel state information," *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 1, p. 53, Jan. 2011.
[61] 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.
[62] B. A. Alsaify, M. M. Almazari, R. Alazrai, and M. I. Daoud, "Exploiting Wi-Fi Signals for Human Activity Recognition," in *Proc. 12th Int. Conf. Inf. Commun. Syst. (ICICS)*, May 2021, pp. 245–250.
[63] R. Alazrai, A. Awad, B. Alsaify, M. Hababeh, and M. I. Daoud, "A dataset for wi-fi-based human-to-human interaction recognition," *Data in Brief*, vol. 31, p. 105668, 2020.
[64] S. Sen, B. Radunovic, R. R. Choudhury, and T. Minka, "You are facing the Mona Lisa: Spot localization using PHY layer information," in *Proc. 10th Int. Conf. Mobile Syst., Appl., Services*, 2012, pp. 183–196.