Towards Generalizable Wireless Sensing Models via Pre-training on Multi-Source Datasets

Chen Gong
School of Computer Science
Peking University
Beijing, China
gongchen17@pku.edu.cn

Bo Liang
School of Computer Science
Peking University
Beijing, China
rambo@pku.edu.cn

Qihao Zhu
School of Computer Science
Peking University
Beijing, China
zhuqh@pku.edu.cn

Wei Gao
University of Pittsburgh
Pittsburgh, USA
weigao@pitt.edu

Yin Chen
Reitaku University
Kashiwa, Japan
ychen@reitaku-u.ac.jp

Jin Nakazawa
Keio University
Tokyo, Japan
jin@sfc.keio.ac.jp

Chenren Xu[*†]
School of Computer Science
Peking University
Beijing, China
chenren@pku.edu.cn

Abstract

The prevailing single-source paradigm in wireless sensing produces specialized models that are unscalable and generalize poorly to new tasks. Multi-source pre-training offers a path toward a generalist backbone but poses challenges including task heterogeneity, data redundancy, structural incompatibility, and the lack of a general-purpose pre-training objective. To address these issues, we propose WiSwiss, a comprehensive self-supervised multi-source pre-training framework that learns a general-purpose backbone for each modality. WiSwiss integrates semantic deduplication for dataset curation and a transformation-invariant pre-training objective. Experiments show that WiSwiss outperforms models trained from scratch, improving WiFi and mmWave performance by 4.5% and 10.3%, respectively, while reducing fine-tuning data requirements by 22.2% and 28.6%. We also present a qualitative study of scaling laws, showing that gains are task-dependent and that larger models require sufficiently large and diverse pre-training corpora to achieve substantial improvements.

CCS Concepts

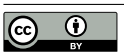
• **Computing methodologies** → **Artificial intelligence**.

Keywords

Pre-training, Wireless Sensing

ACM Reference Format:
Chen Gong, Bo Liang, Qihao Zhu, Wei Gao, Yin Chen, Jin Nakazawa, and Chenren Xu. 2026. Towards Generalizable Wireless Sensing Models via Pre-training on Multi-Source Datasets. In *ACM/IEEE International Conference on Embedded Artificial Intelligence and Sensing Systems (SenSys '26), May 11–14, 2026, Saint Malo, France.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3774906.3802759

[*]Also with Key Laboratory of High Confidence Software Technologies, Ministry of Education (PKU).
[†]Corresponding author: chenren@pku.edu.cn



SenSys '26, Saint Malo, France

ACM ISBN 979-8-4007-2309-4/26/05
https://doi.org/10.1145/3774906.3802759

1 Introduction

Deep learning-enhanced wireless sensing has enabled a diverse spectrum of applications, ranging from human activity recognition [83] to indoor localization [2]. However, the prevailing research paradigm remains fundamentally constrained to a *single-source* setup. Models are typically trained on task-specific datasets in a single domain to become *specialists* for that one task [43]. This reliance on a single-source paradigm results in a bottleneck for real-world deployment: It necessitates that for each new sensing task, environment, or hardware configuration, a large-scale labeled dataset will be collected from scratch [16]. Such a process is costly, time-consuming and unscalable. Consequently, models developed under this paradigm exhibit poor performance in low-data regimes [81].

In recent years, a major paradigm shift has emerged, driven by the success of large-scale models in natural language processing [15] and computer vision [11]. This shift moves from specialist models toward a *generalist pre-trained backbone* that learns general-purpose representations from broad and diverse data [7, 79]. Such a backbone can serve as a universal feature extractor and can be efficiently adapted (*e.g.,* fine-tuned) to diverse downstream tasks, including tasks that are unseen during pre-training. Consequently, it alleviates the data collection bottleneck and supports effective deployment in new scenarios where labeled data is limited [49, 57].

Translating this generalist ambition to wireless sensing introduces a set of distinct challenges. First, training a generalist model requires a large and diverse corpus, which raises the challenge of **heterogeneity across sensing domains and tasks**. Unlike modalities with relatively uniform representations (*e.g.,* language), wireless signals are highly sensitive to the scenario and hardware configuration. Furthermore, typical applications range from classification tasks like gesture/action recognition to regression tasks like pose estimation and object detection, and different tasks require

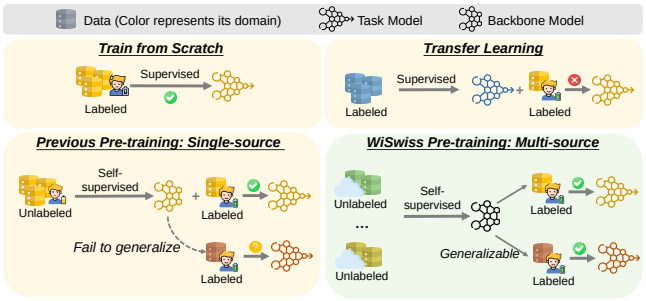


Figure 1: Comparison of Wireless Sensing Paradigms: Specialist Models vs. the WiSwiss Generalist Backbone Model.

features of varying granularity [48]. For instance, gesture recognition requires fine-grained human motion, while indoor localization focuses on coarse-grained spatial distribution [91]. This divergence makes defining a unified supervised optimization target difficult. As our preliminary experiments confirm (§2.1), transfer learning across disparate tasks frequently fails, resulting in negative transfer where source-domain knowledge harms target-domain performance, as illustrated in the transfer learning part of Fig. 1.

This fundamental limitation of supervised learning necessitates a shift to *self-supervised pre-training on multi-source data*. However, this shift introduces additional challenges at both data and model levels. At the data level, one major challenge is **data redundancy**. Aggregating multiple datasets inevitably over-represents common but uninformative samples (*e.g.,* static 'no activity' signal segments) and introduces large volumes of semantically repetitive data (*e.g.,* instances of the same gesture). Such redundancy can bias model optimization, causing overfitting to frequent patterns, and waste computational resources [46]. Therefore, data deduplication is a necessary prerequisite for efficient and unbiased pre-training.

Another critical data-level challenge is **structural incompatibility**. Data shapes vary even within the same modality. For example, the time-frequency dimensions of WiFi Channel State Information (CSI) signals differ across datasets collected with different hardware platforms or sampling durations [95], and mmWave radars exhibit heterogeneous range and angle resolutions. A generalist backbone should be flexible enough to process variable-shaped inputs natively, without relying on distortive resizing or padding that may destroy critical signal structures [50].

Beyond these data-level issues, there is also a model-level challenge: the **absence of a general-purpose pre-training objective**. The key question is which objective can learn representations that generalize beyond the training domain. Standard masked reconstruction is effective at capturing local signal details, but it may overfit to low-level domain-specific artifacts (*e.g.,* the unique multipath pattern of a room) rather than high-level transferable semantics (*e.g.,* the human motion) [68, 69]. By contrast, contrastive learning is prone to learning shortcuts from the data augmentations rather than the meaningful signal structures [51]. Existing pre-training strategies are fragmented, typically emphasizing either local detail or global semantics in isolation, and do not provide a unified objective for learning a general-purpose backbone.

Although pre-training has been explored in wireless sensing [42, 61, 67–69], existing methods are not designed for the generalist objective in our work. As illustrated in Fig. 1, these approaches typically remain within a single-source setting, using data collected with uniform hardware and in limited scenarios. As a result, their downstream applications are narrow and restricted to one or two similar tasks. These methods neither address the combined challenges of multi-source heterogeneity and cross-task generalization, nor provide mechanisms to effectively handle them.

In this paper, we introduce WiSwiss, the first systematic framework for self-supervised multi-source pre-training in wireless sensing. It provides a holistic data-to-model pipeline that addresses the full set of challenges involved in learning a general-purpose backbone for diverse tasks within each wireless modality.

To address task heterogeneity, we adopt self-supervised learning to avoid the negative transfer associated with supervised cross-task transfer. To handle data-level challenges, we first develop a data deduplication pipeline. It uses an encoder model trained under perturbation invariance, thereby identifying repetitive data without relying on pre-existing universal feature extractor [1]. To adapt to varying input formats, WiSwiss is built on a flexible transformer architecture with Multi-Dimensional Rotary Position Embedding (RoPE) [32]. This design enables the model to process variable-shaped inputs natively while preserving their intrinsic structure.

To address the model-level challenge, we further introduce a transformation-invariant pre-training objective. This objective enables the model to learn generalizable features by complementing local masked reconstruction with a global consistency loss. The consistency loss is defined with physically meaningful transformation [68, 69], which encourages the model to capture underlying semantics while remaining invariant to scenario-specific artifacts. By jointly optimizing local fine-grained reconstruction and global *transformation invariance*, our pre-training strategy balances essential signal detail and transferable semantics, leading to better generalization and higher accuracy on downstream tasks in the WiFi and mmWave modalities. We validate WiSwiss through extensive experiments on six open-source datasets across the WiFi and mmWave modalities, covering both classification and regression tasks. We use CSI for WiFi and time-range-angle cubes for mmWave. Overall, WiSwiss improves performance on WiFi and mmWave tasks by 4.5% and 10.3%, respectively, while reducing fine-tuning data requirements by 22.2% and 28.6%.

Our contributions in this paper are summarized as follows:

- We introduce WiSwiss, the first systematic framework for multi-source self-supervised pre-training in wireless sensing, which learns generalizable representations for diverse tasks.
- We develop a holistic pipeline that includes semantic deduplication to handle data redundancy and a transformation-invariant objective for learning robust and generalizable features.
- We validate WiSwiss through extensive experiments and demonstrate its superior performance and data efficiency. We further present a pioneering analysis of scaling laws in wireless sensing, offering insights into the roles of model size and data volume during pre-training and fine-tuning.

Our code is available at https://github.com/gcc17/WiSwiss.

2 Background & Motivation

To motivate this work, we first present empirical evidence of negative transfer in wireless sensing. This phenomenon, demonstrated in §2.1, reveals a critical limitation of supervised transfer learning and motivates the adoption of a self-supervised approach. We then analyze the principles underlying large-scale pre-training in language and vision, and use their established workflows as a lens to identify the gaps in current wireless pre-training practices.

2.1 The Bottleneck of Negative Transfer

To demonstrate the limitations of supervised transfer learning, we conduct a preliminary study. Specifically, a model is first trained with labeled supervision on source datasets (Widar [92] and/or SignFi [55]), and is then fine-tuned on the XRF55 dataset [74] for two downstream tasks: action recognition and human identification.

As shown in Tab. 2, for both tasks, models pre-trained with labeled supervision on the source datasets perform *worse* than a baseline model trained from scratch on the target dataset. This negative transfer effect highlights a fundamental limitation of supervised learning: it encourages the model to learn representations that are overly specialized to the source domains and labels, thereby limiting generalization to unseen target domains.

The observation of negative transfer effect motivates the adoption of a self-supervised pre-training framework. By learning directly from data without reliance on task-specific labels, such a framework enables the development of a backbone model [7] that captures universal characteristics of wireless signals, and supports effective adaptation to diverse downstream tasks.

2.2 LLM Lessons: Systematic Pipeline

Having established that self-supervised learning is a more viable direction than supervised transfer, we now examine how to construct such a framework. The success of self-supervised pre-trained models in language and vision is built on a systematic pipeline of three stages: (1) aggregating diverse data, (2) curating the aggregated data, and (3) applying an effective model optimization strategy.

Diverse Data Sources. Foundation models such as BERT [15] and GPT-3 [8] demonstrate that training on large-scale and diverse datasets is critical for learning robust and generalizable representations. The scale and heterogeneity of the training corpus enable models to capture a broad range of patterns, which support strong performance across diverse tasks.

Efficient Data Curation. Another critical component of model pre-training is the curation of large-scale corpora. For example, the C4 dataset [65] applies heuristic filtering and deduplication to reduce bias from over-represented text patterns [46], and shrinks the dataset from 6.21 TB to 806.87 GB [71] to improve training efficiency. In general, existing deduplication pipelines remove exact duplicates through token hashing and identify semantic duplicates by using existing encoder models to retrieve and cluster similar embeddings [1]. However, these methods are either designed for discrete tokenized data or depend on powerful encoders.

Effective Model Optimization. Beyond the dominant paradigm of masked modeling, which is effective at capturing fine-grained local details [30], contrastive learning provides a flexible framework for learning a globally structured embedding space. It can promote class separation (supervised) [44], inter-modality consistency [64], and instance discrimination (self-supervised) [12, 31].

2.3 Wireless Gap: Incomplete Approach

Existing efforts in wireless pre-training lack a unified framework, which results in several key limitations that we aim to address.

Single-source Bias. Current approaches perform both pre-training and fine-tuning on data collected from identical devices and environments [22, 42, 61, 67–69]. Such a approach fails to produce models that are robust to domain shifts inherent in real-world deployments, thereby undermining the primary benefit of pre-training.

Uncurated Data. The challenge of data redundancy in wireless signals remains largely unexplored. Existing methods utilize all available data without filtering or deduplication [96], leading to models that are biased toward frequent but uninformative signal patterns and result in inefficient use of computational resources.

Under-explored Objectives. Although prior work in wireless sensing has demonstrated the potential of self-supervised learning, they have largely diverged into two separate directions. The first direction focuses on improving masked signal modeling, such as keeping high-energy regions during masking to emphasize informative content [42]. The second direction adopts contrastive learning, where signal processing techniques are used to construct positive samples [67–69]. However, these strategies have been explored independently. The absence of a unified objective that jointly captures local details and global semantics leaves a critical research gap: the design of a pre-training strategy that is well aligned with the characteristics of wireless signals remains under-explored.

To clearly position WiSwiss within the landscape of recent advancements, we compare it with representative wireless pre-training frameworks in Tab. 1. Existing methods rely on fixed input dimensions and uncurated datasets, which inherently limit their applicability to similar downstream tasks. In contrast, WiSwiss is designed to address the heterogeneity of multi-source data. By integrating data curation, shape adaptation, and a unified pre-training objective, our framework enables robust cross-task generalization.

3 Overview

Building a multi-source generalist backbone for wireless sensing requires addressing a set of systemic challenges. At the data level, the framework must handle data redundancy arising from large-scale aggregation and the structural incompatibility of signals collected from heterogeneous hardware. At the model level, it requires a general-purpose pre-training objective that captures transferable semantics, rather than overfitting to domain-specific artifacts or augmentation-induced shortcuts.

To address this complete set of challenges, WiSwiss is designed around a unifying principle: *learning through invariance*. Instead of learning task-specific details, such as the exact activity represented by a signal, the framework focuses on learning features that remain invariant under changing conditions. Compared with task-specific details that are entangled with environmental variations [56, 89], such invariant representations are more broadly applicable, and

Table 1: Comparison of Representative Wireless Sensing Pre-training Frameworks.

Framework	Pre-training Objective	Modality Coverage	Input Shape	Data Source	Tasks	Data Curation
RF-URL [68, 69]	Contrastive learning	WiFi, mmWave	Fixed	Single	1 WiFi task, 2 mmWave tasks	None
PhyMask [42]	Masked recon.	Acoustic signals	Fixed	Single	1 task per model	None
COSMO [61]	Contrastive fusion	Multimodal (mmWave+depth+IMU)	Fixed	Single*	1 task	None
AM-FM [96]	Contrastive + masked recon. + energy	WiFi	Fixed	Multiple (proprietary)	3 tasks**	None
WiSwiss (Ours)	**Masked recon. + trans-invariance**	**WiFi, mmWave**	**Dynamic**	**Multiple (public)**	**3 WiFi tasks, 3 mmWave tasks**	**Semantic deduplication**

* While COSMO handles multi-modality, the model is trained per-dataset, making it a single-source paradigm.
** We treat the CSI-Bench dataset with 7 sub-tasks as a single task. Both AM-FM and our framework include this dataset for downstream evaluation.

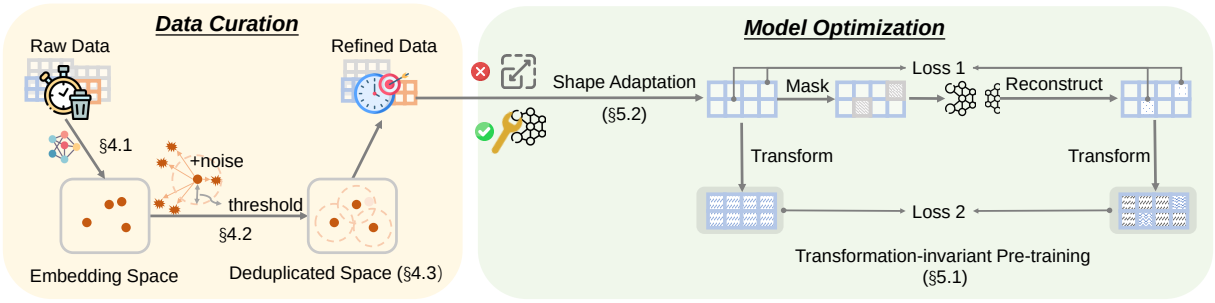


Figure 2: WiSwiss Workflow.

Table 2: Transfer from Widar/SignFi to XRF55.

Method	Action Recognition	Human Identification
Train from Scratch	**62.66%**	**86.96%**
Transfer from Widar+SignFi	57.42%	83.37%
Transfer from Widar	61.02%	84.79%
Transfer from SignFi	61.00%	83.27%

provide useful initialization for downstream models. This principle is systematically applied at both the data and model levels to improve data efficiency and generalization, as illustrated in Fig. 2.

4 Data Curation

Effective deduplication requires an encoder that maps input data to semantically meaningful representations. However, in our setting, such an encoder is itself the intended output of the pre-training pipeline, which creates a circular dependency between data curation and model training. To resolve this issue, we introduce a semantic deduplication approach that operates from scratch, without task-specific labels or any pre-existing feature encoder. Specifically, the proposed method identifies and removes semantic redundancies that are not restricted to particular tasks, thereby enabling efficient learning of generalizable representations.

The proposed data curation pipeline is built on the principle of perturbation invariance. The key assumption is that the semantic content of a sample remains unchanged under moderate semantics-preserving perturbations. This property allows us to construct positive pairs by treating perturbed versions of the same sample as semantically equivalent. We then learn an embedding space in which such pairs are pulled closer together, so that proximity in the feature space reflects semantic similarity.

The deuplication pipeline consists of three stages: (1) extracting perturbation-invariant embeddings, (2) identifying duplicates according to latent space similarity, and (3) removing redundant samples. The pipeline has two key components. First, we introduce a tailored contrastive learning framework to train an embedding model that is inherently invariant to perturbations, without relying on external supervision or pre-trained models. Second, we develop a dynamic data-driven algorithm to determine the similarity threshold for duplicate detection automatically, which improves the robustness and adaptability of the deduplication process.

4.1 Feature Encoder Training

To enforce perturbation invariance, we design a contrastive learning framework that trains a feature encoder to map an input signal and its perturbed version to nearby points in the latent space, while separating representations of semantically distinct signals.

Although the feature encoder is built on a standard contrastive learning backbone, our use of contrastive learning differs from prior work. Existing methods, including self-supervised instance discrimination [12, 31] and supervised class separation [44, 94], are typically developed to learn generalized feature extractors from relatively uniform datasets. In contrast, we re-purpose contrastive learning for semantic deduplication over heterogeneous multi-source data. Moreover, we construct negative pairs locally within each dataset by using its labels, which avoids the cross-dataset label misalignment that does not arise in standard modalities such as language.

Positive and Negative Sample Construction. The choice of perturbation is critical for learning meaningful invariance. We use Additive White Gaussian Noise (AWGN) to construct positive pairs, as it is a fundamental channel model that captures thermal noise and uncontrolled interference in wireless systems [63, 72]. Its statistical simplicity and physical relevance have made it an effective augmentation for robust deep learning in both computer vision [73] and wireless sensing [14].

A key reason for this choice is that AWGN serves as a minimal and unstructured perturbation. Although real wireless signals are also affected by structured distortions such as carrier frequency offset (CFO) [75] and sampling frequency offset (SFO) [45], our goal here is not to model the full space of possible impairments. It is important to clarify that the deduplication encoder is not designed to learn a fully general representation for downstream tasks. Instead, its purpose is to provide a conservative similarity metric for near-duplicate detection. More complex augmentations may be beneficial

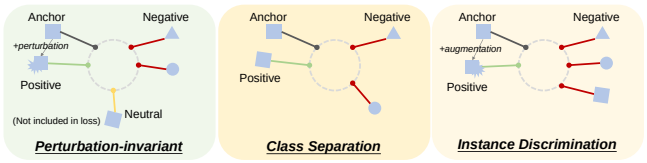


Figure 3: Comparison between Our Contrastive Learning for Deduplication and Existing Ones.

in specific applications, but they are not universally semantics-preserving and are often strongly task-dependent [24]. By contrast, AWGN is a minimal modality-agnostic perturbation that preserves signal semantics under moderate variance. Enforcing invariance to this noise helps the encoder avoid task-specific augmentation shortcuts and extract stable features for deduplication.

For each original signal x_i in a mini-batch, we treat it as an *anchor* and construct its *positive counterpart* x_i^+ with Gaussian noise:

$$x_i^+ = x_i + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2)$$

where σ^2 is a small variance hyperparameter.

Negative samples are defined to encourage separation between signals from different task-specific classes. For an anchor x_i with class label y_i, its negatives are all other samples x_j in the mini-batch with $y_j \neq y_i$. Samples that share the same class label as the anchor are excluded from the negative set. This design prevents the encoder from penalizing intra-class similarity and instead focuses optimization on inter-class separation. As a result, the learned embedding space is better suited to near-duplicate detection and less prone to overfitting to source-specific intra-class distributions. A conceptual comparison between our perturbation-invariant contrastive design and prior methods for supervised class separation [44] and self-supervised instance discrimination [31] is shown in Fig. 3.

The feature encoder used for deduplication is adapted from a pre-trained ResNet-50 and projects input signals into a 128-dimensional latent space. We choose the ResNet family because of its stable optimization and reliable convergence behavior, which are desirable for learning effective similarity representations [38].

Standard self-supervised contrastive learning, such as instance discrimination, is a natural alternative, but it treats all other samples in the batch as negatives. This can push apart semantically identical samples and produce a latent space that is unsuitable for near-duplicate detection. In contrast, we use local task-specific labels to define negatives, which separates distinct classes without penalizing intra-class similarity. Because these labels are used within individual dataset batches, the method also avoids the cross-dataset label misalignment.

Loss Function. The feature encoder is trained to maximize the similarity between an anchor and its positive counterpart while minimizing similarity to negative samples in the same mini-batch. To achieve this, we adopt the InfoNCE loss, a standard objective in contrastive learning for representation [12, 59].

Given a feature encoder $\mathcal{M}$, an anchor embedding $z_i = \mathcal{M}(x_i)$, and its positive embedding $z_i^+ = \mathcal{M}(x_i^+)$, the loss treats positive-pair identification as a classification problem over one positive and $N-1$ negatives $\{z_k\}_{k=1}^{N-1}$ from the same mini-batch. The loss for the i-th anchor is:

$$\mathcal{L}_i = -\log \frac{\exp(\text{sim}(z_i, z_i^+)/\tau)}{\exp(\text{sim}(z_i, z_i^+)/\tau) + \sum_{k=1}^{N-1} \exp(\text{sim}(z_i, z_k)/\tau)}$$

where $\text{sim}(\cdot,\cdot)$ measures the similarity between two embeddings, and τ is a temperature hyperparameter that controls the sharpness of the distribution. A smaller τ places greater emphasis on hard negatives, which are close to the anchor in the latent space but belong to different classes. In our implementation, we use cosine similarity, defined as $\frac{\boldsymbol{u}^\top \boldsymbol{v}}{\|\boldsymbol{u}\|\|\boldsymbol{v}\|}$. The temperature is set to $\tau = 0.07$, following common practice in prior work [5].

4.2 Threshold-based Duplicate Detection

After converting signals into embedding vectors, the next step is to identify semantically redundant samples. Although clustering methods such as k-means [4] are often used to group data, they are not well suited to large-scale deduplication. A fundamental limitation is that they require the number of clusters, K, to be specified in advance. In our setting, however, the number of unique informative samples is unknown and is precisely what the curation process aims to uncover. Moreover, heuristic methods for selecting K, such as the elbow method, are computationally expensive because they require repeated clustering over a range of K values [33]. This cost is prohibitive for large-scale datasets. Therefore, we adopt a simpler and more efficient threshold-based strategy: two signals x_i and x_j are regarded as semantic duplicates if the similarity between their embedding vectors z_i and z_j exceeds a threshold δ. Formally, if $\text{sim}(z_i, z_j) > \delta$, then (x_i, x_j) is a duplicate pair.

Dynamic Threshold Determination. A fixed similarity threshold is suboptimal because the density and geometry of the latent space can vary substantially across datasets. We therefore determine δ dynamically for each dataset by leveraging the properties of embedding distribution. Because the encoder is explicitly trained to map a signal x_i and its noise-augmented counterpart x_i^+ to nearby locations in the latent space, the similarity distribution of these positive pairs, $\text{sim}(z_i, z_i^+)$, provides a natural data-driven reference for semantic duplication. In practice, we set δ to the 10th percentile of the positive-pair similarity distribution.

This dynamic threshold is intentionally conservative. By choosing a lower percentile, we reduce the risk of incorrectly removing unique samples and thus better preserve the diversity of the curated dataset. This design accepts the trade-off that a small amount of redundancy may remain. By contrast, a more aggressive threshold (*e.g.,* a higher percentile) would remove more samples but would also increase the risk of discarding semantically unique instances near the boundary of the positive-pair distribution, thereby reducing the coverage of the dataset for pre-training.

4.3 Efficient Duplicate Removal

A brute-force pairwise comparison of all N embeddings requires $O(N^2)$ computation and is infeasible at large scale. To address this issue, we design an efficient multi-stage duplicate removal pipeline based on the Facebook AI Similarity Search (FAISS) library [19, 40], which supports accelerated approximate nearest neighbor search.

We first build a FAISS index to organize the embedding space for efficient similarity retrieval. For each sample, we then search

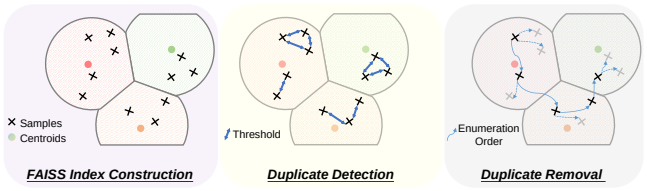


Figure 4: Deduplication Process.

for its k nearest neighbors and use these candidates to construct a duplicate graph. In this graph, each node represents a sample, and an undirected edge is added between x_i and x_j if their similarity $\text{sim}(z_i, z_j)$ exceeds the threshold δ.

To obtain a non-redundant subset from this graph, we apply a greedy traversal algorithm. The algorithm scans the nodes sequentially. When it encounters a node that has not been marked for removal, that node is added to the curated subset. All of its neighbors in the duplicate graph are then marked for removal, since they are considered duplicates of the retained sample. This process continues until every node has been either retained or removed. The overall pipeline, including FAISS indexing, neighbor retrieval, and graph-based pruning, is illustrated in Fig. 4.

5 Dynamic Shape Adaptation

Another challenge in learning a generalist backbone is the variation in input dimensions across datasets. These differences arise from variations in subcarrier counts, FFT resolutions, and signal collection durations. A common solution is to resample all inputs to a fixed size [60]. However, this strategy is unsuitable for wireless signals because it can distort critical temporal and spectral structures. To avoid this issue, we enable native support for variable-shaped inputs through two architectural modifications to the standard Vision Transformer (ViT) [18]: a flexible convolutional patch embedding layer and a multi-dimensional rotary position embedding.

Flexible Convolutional Patch Embedding. Our architecture is based on the ViT, which divides an input into a grid of non-overlapping patches. Each patch is projected into a fixed-dimensional embedding through a convolutional layer, and the resulting embeddings are flattened into a 1D sequence for subsequent transformer blocks. Unlike the original ViT, where patches contain only spatial pixel dimensions [18], our patches cover either 2D time-frequency inputs or 3D time-frequency-spatial inputs. As a result, the output sequence length scales naturally with the input dimensions, which preserves the native signal resolution without distortion.

Multi-Dimensional Rotary Position Embedding. The variable-length sequences produced by patch embedding require a positional encoding scheme that is not tied to a fixed sequence length. We therefore adopt rotary position embedding (RoPE) [70], which injects relative positional information by applying rotational transformations to the query and key vectors in self-attention according to token positions. In the original 1D formulation, the rotation for embedding channel k at token position n is defined as:

$$\mathbf{R}_{\theta_k}(n) = e^{i\theta_k n}, \quad \text{where } \theta_k = 10000^{-2k/d}$$

where d is the embedding dimension.

However, this 1D formulation is insufficient for multi-dimensional wireless signals. After flattening, the sequence no longer preserves

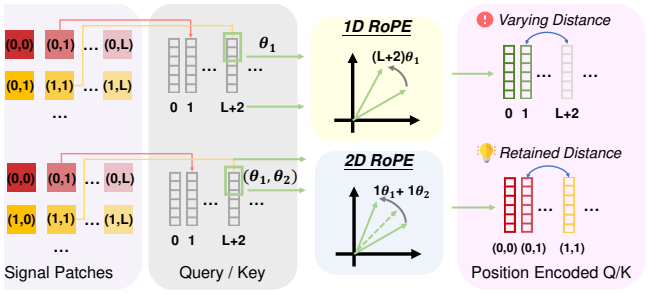


Figure 5: Comparison between 1D and 2D RoPE.

the original structural relationships among patches. For example, as illustrated in the upper part of Fig. 5, two patches such as (0,1) and (1,1), which are adjacent in time, may be separated in the flattened sequence by a distance that depends on the size of other dimensions. Consequently, the model cannot accurately capture the intrinsic temporal and spectral topology of the signal.

To preserve this structure, we extend RoPE to multi-dimensional inputs [32]. For each patch at flattened position n, we recover its original coordinates in each dimension (*e.g.,* frequency p_n^f and time p_n^t), and compute a composite rotation by combining dimension-specific rotary angles, as illustrated in the lower part of Fig. 5.

$$\mathbf{R}_{\theta_k}(p_n^f, p_n^t) = e^{i(\theta_k^f p_n^f + \theta_k^t p_n^t)}$$

This composite transformation makes the self-attention scores responsive to the true dimension arrangement of signal patches.

6 Model Optimization

Given curated multi-source datasets, the next challenge is to design a pre-training objective that learns generalizable representations. Existing self-supervised objectives have respective limitations. Masked modeling can overfit to low-level features, which weakens generalization across tasks and datasets [82, 86]. For contrastive learning, supervised variants are affected by inconsistent label distributions across heterogeneous wireless datasets [56], while self-supervised variants are prone to learning trivial shortcuts from data augmentations [51]. To address these issues, we introduce a transformation-invariant pre-training objective that balances fine-grained local detail with semantically meaningful global structure.

Although prior studies have incorporated signal processing into representation learning [68, 69], they enforce consistency between 'views' derived from specialized hardware configurations, such as multi-antenna arrays. This dependence limits their applicability to specific datasets and collection setups. In contrast, our framework is built on hardware-agnostic transformations, which makes it applicable to a broader range of wireless sensing datasets.

Our objective combines a standard reconstruction loss with a transformation consistency loss. The overall loss is defined as:

$$\mathcal{L} = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2 + \lambda \|\varphi(\boldsymbol{x}) - \varphi(\hat{\boldsymbol{x}})\|_2^2$$

Here, $\boldsymbol{x}$ denotes the original signal and $\hat{\boldsymbol{x}}$ denotes the reconstruction from the masked input. The function $\varphi(\cdot)$ denotes a modality-specific transformation. The reconstruction term encourages accurate recovery of masked signal content. The transformation term

serves as a regularizer that encourages the model to preserve task-relevant signal characteristics, as defined by φ, during reconstruction. The hyperparameter λ controls the trade-off between reconstruction fidelity and transformation invariance.

Target Signal Transformations. The choice of the transformation function $\varphi(\cdot)$ is critical. We select transformations that are well established in their respective domains for extracting robust high-level features, so that the pre-training objective is guided toward semantically meaningful invariance.

For the WiFi modality, we define φ as the transformation from raw CSI to the doppler frequency spectrum (DFS). DFS is effective at capturing motion-induced frequency shifts while suppressing static environmental clutter, which makes it a robust representation for activity recognition [26, 36, 75]. By enforcing consistency in the DFS domain, the model is encouraged to focus on primary motion components that are important across a wide range of sensing tasks.

For the mmWave modality, we use the constant false alarm rate (CFAR) algorithm as φ. CFAR is a standard adaptive thresholding method in radar signal processing for distinguishing salient targets from background noise and clutter [66, 77]. Incorporating this transformation encourages the model to learn features that support robust separation of salient objects from background noise.

Although extremely low-SNR inputs can occasionally make transformed representations less informative, the transformation consistency term acts only as a regularizer and is jointly optimized with the primary masked reconstruction objective.

Differentiable and Accelerated Transformations. To integrate these signal processing transformations into a GPU-accelerated pre-training pipeline, they should be reformulated as fully differentiable and parallelizable PyTorch operations. This is challenging because traditional signal processing algorithms involve sequential and sample-specific logic that is incompatible with batched GPU operation. Therefore, we develop differentiable and vectorized implementations for each modality-specific transformation.

For the CSI-to-DFS transformation, the main computational bottlenecks are the infinite impulse response (IIR) filtering and PCA stages. The original IIR filter uses a sequential recurrence and is therefore not suitable for parallel execution. We replace it with an efficient frequency-domain bandpass filter. For PCA-based denoising, which extracts the first principal component, direct eigendecomposition using `torch.linalg.eigh` is numerically unstable during backpropagation and can cause vanishing or exploding gradients. To ensure stable end-to-end training, we instead use the *power method* to iteratively approximate the first eigenvector of the covariance matrix [25]. This method relies on repeated matrix-vector multiplication and normalization, and provides an efficient differentiable alternative with stable gradient flow.

After accelerating these core operations, we further adapt the pipeline to batched inputs. One challenge is reference antenna selection, as the optimal antenna differs across samples. A per-sample loop would eliminate the benefit of parallelization. We therefore use a vectorized masking strategy: selection criteria are computed for all antennas in parallel, a dynamic one-hot mask is generated from the resulting indices, and the mask is applied in a batched manner. These optimizations reduce the batched CSI-to-DFS transformation time from 0.5 seconds to 0.008 seconds, yielding a 68× speedup and making the transformation practical for large-scale pre-training.

For the mmWave transformation, we implement a differentiable version of standard cell-averaging CFAR (CA-CFAR). Conventional CFAR has two steps: estimating the local noise power of each cell from its surrounding guard and reference cells, and applying a hard threshold for target detection. We reformulate both steps as differentiable operations. Noise estimation is implemented efficiently as a 2D convolution with a fixed kernel that averages the required neighboring cells. The main difficulty lies in thresholding, because a hard threshold is non-differentiable and yields zero gradients. We therefore replace it with a differentiable soft threshold using a temperature-scaled sigmoid function:

$$\hat{S} = \sigma(\tau \cdot (\mathrm{SNR} - \delta)) = \frac{1}{1 + e^{-(\mathrm{SNR}-\delta)/\tau}}$$

Here, SNR denotes the signal-to-noise ratio of a cell, δ denotes the detection threshold, and τ is a temperature parameter that controls the sharpness of the soft decision boundary. This formulation allows gradients to propagate through the full detection map.

Although the transformation function $\varphi(\cdot)$ is modality-specific, it serves as a flexible plug-in regularizer within the pre-training objective. Extending WiSwiss to a new wireless modality only requires the design of a differentiable vectorized transformation for that signal type, while the core pipeline of semantic deduplication, shape adaptation, and pre-training objective remains unchanged.

7 Implementation

Data Preprocessing. To explicitly account for variations in signal intensity, we apply standard amplitude normalization. Following RF-Diffusion [14], each input signal sequence is normalized by its average signal power, which stabilizes training by reducing the impact of energy shifts while preserving the relative variations.

Training Setup. Pre-training and fine-tuning are conducted on a workstation equipped with Genuine Intel (R) CPU and NVIDIA 3090 GPUs. We use the *Accelerate* library [29] for efficient multi-GPU training. For pre-training, models are trained for 10 epochs with a learning rate of 5×10^{-4} with cosine decay and a batch size of 32. For downstream fine-tuning, we attach a lightweight task-specific layer to the pre-trained backbones, which are then fine-tuned for 100 epochs with a learning rate of 2×10^{-4} and a batch size of 32. For comparison, baseline models are trained from scratch for 100 epochs with a learning rate of 3×10^{-4} and a batch size of 32. A smaller learning rate for fine-tuning aims to avoid much deviation from the pre-trained representations [62].

8 Evaluation

8.1 Experiment Setup

8.1.1 Datasets and Tasks. Our goal is to develop a generalist backbone for each wireless sensing modality. We conduct experiments on WiFi and mmWave datasets, with details in Table 3. Dataset selection is guided by three principles for rigorous evaluation:

- Disjoint Pre-training and Fine-tuning Data. To emulate realistic deployment scenarios, we use separate datasets for pre-training and fine-tuning. This setup evaluates the model's ability to generalize to new tasks or scenarios using limited labeled data.

- Diverse Downstream Tasks. We fine-tune and evaluate the pre-trained model on both classification and regression tasks.
- Standardized and Robust Benchmarks. For fair comparison and reproducibility, we select datasets widely used as benchmarks. We also prioritize datasets with a large number of samples to ensure statistical robustness and reduce the impact of randomness.

Table 3: Datasets for Pre-training and Fine-tuning.

Modality	Stage	Dataset	Task	# Train	# Test/Valid	Shape
WiFi	Pre-training	SignFi [55]	Gesture	7500	-	(90,200)
		Widar [92]	Gesture	66384	-	(90,L)
		MM-Fi [85]	Pose	320760	-	(342,10)
	Fine-tuning	XRF55 [74]	Action	23100	9900	(90,1000)
		CSI-Bench [95]	Classification	19129	27399	(58/64/..., L)
		Wi-Pose [21]	Pose	19927	33753	(342,5)
mmWave	Pre-training	MM-DCDR [23]	Action	6832	-	(64,128,128)
		MCD [52, 53]	Gesture	8040	-	(32,128,32)
		RT-Pose [34]	Pose	253333	-	(16,256,16)
	Fine-tuning	XRF55 [74]	Action	23100	9900	(16,256,64)
		RADDet [88]	Detection	8126	2032	(16,256,256)
		HuPR [47]	Pose	30000	6000	(8,64,64)

L represents variable time length.

8.1.2 Baselines and Ablations. To demonstrate the effectiveness of WiSwiss, we compare it against two primary baselines and conduct ablation studies evaluating five model variants.

- **Baselines:** We compare with (1) PhyMask [42][1], an adaptive masking method for self-supervised pre-training that keeps information-rich and temporally coherent regions, and (2) a model trained from scratch in a supervised manner for each individual task, representing the approach without pre-training.
- **Ablation Studies:** To validate our design choices, we evaluate five variants of WiSwiss along two dimensions. The first dimension investigates the training procedure: (1) without data deduplication and (2) without the transformation-invariant loss. The second dimension examines the architecture by replacing dynamic RoPE with fixed positional embeddings (PE) using three approaches: (3) interpolated PE, (4) resizing all inputs to a small fixed shape, and (5) resizing all inputs to a large fixed shape.

8.1.3 Task Signal Format. We evaluate our framework on two primary wireless modalities. For WiFi tasks, the input is a 2D time-frequency representation of CSI, while for mmWave tasks, it is a 3D time-range-angle data cube. This format is directly derived from raw ADC data via a near-lossless FFT and provides richer information than point clouds [54]. For each modality, we pre-train a single general-purpose backbone that is shared and fine-tuned across all downstream datasets. The backbone is modality-specific but universal within its domain.

8.1.4 Pre-training Data Scaling Strategy. To study the impact of both data quantity and diversity, we progressively increase the pre-training corpus from 10k to 160k samples. Rather than sampling randomly from a mixed pool, we add entire datasets sequentially following the order in Tab. 3. This per-dataset inclusion strategy enables analysis of how performance scales not only with total data volume but also with the number of distinct data sources.

8.1.5 Model Architecture. Our model architecture is based on Vision Transformer [18], which partitions the continuous input into non-overlapping patches. Similar to ViT for 2D images [78] and Video ViT for 3D videos [6], we use a patch size of 16×16 for 2D WiFi inputs and $2 \times 16 \times 16$ for 3D mmWave inputs. To study the effect of model scale, we consider three configurations: tiny, small, and base. These variants differ in embedding dimension, number of attention heads, and number of layers, as summarized in Tab. 4.

Table 4: Model Architecture Configurations.

Scale	Embedding Size	# Atten. Heads	# Atten. Layers	# Parameters
Tiny	192	3	12	0.7 M
Small	384	6	12	2.7 M
Base	768	12	12	10.8 M

8.1.6 Evaluation Metrics. We evaluate model performance using standard task-specific metrics. For classification tasks, we report accuracy. In addition, since the CSI-Bench dataset [95] includes multiple classification tasks, we report the average accuracy over five tasks absent from other datasets (breathing detection, fall detection, human identification, room-level localization, and motion source recognition). For pose estimation, we use Mean Per Joint Position Error (MPJPE) [58]. For object detection, we report mean Average Precision (mAP) [39]. Higher accuracy and mAP indicate better performance, whereas lower MPJPE indicates better performance.

8.1.7 Statistical Robustness. To account for the randomness in the fine-tuning process (*e.g.,* data shuffling), we repeat each experiment five times with different random seeds. In our result figures, the mean performance is depicted by a solid line, and the full data range is visualized as a surrounding shaded error band.

8.2 Pre-training Scaling

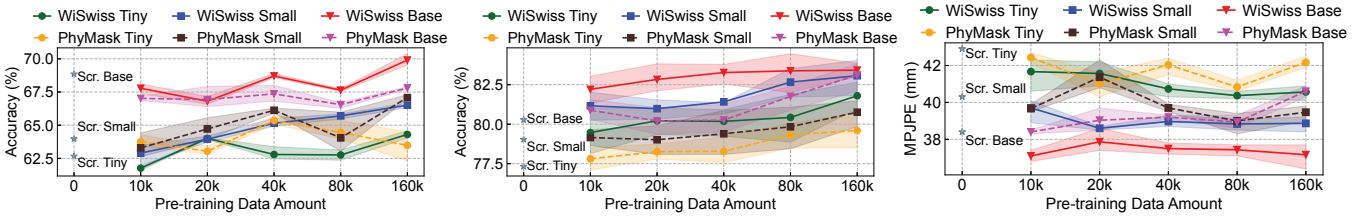
We evaluate the performance of WiSwiss on both WiFi and mmWave modalities, with detailed results shown in Fig. 6 and Fig. 7. In these figures, solid lines represent the performance of WiSwiss under different model scales, dotted lines indicate PhyMask, and star markers denote the from-scratch baseline. We analyze the results from three perspectives: (1) the advantage of our method over baselines, and (2) the effects of pre-training data volume and model scale.

Pre-training Advantage. Across all tasks and model sizes, our pre-training consistently improves performance compared with training from scratch. These improvements are detailed below:

- For the WiFi modality, pre-training boosts accuracy on the XRF55 dataset from 62.7% to 64.3% (a 2.6% relative gain) and improves the average accuracy on CSI-Bench from 77.3% to 81.8% (a 5.7% relative gain). For the Wi-Pose regression task, it reduces the pose estimation error by 5.3%, from 42.9 mm to 40.6 mm.
- The benefits are more pronounced for the mmWave modality. Accuracy on the XRF55 dataset increases from 76.3% to 80.0% with a 4.8% relative gain. On the HuPR dataset, pre-training achieves a 14.6% reduction in pose estimation error, lowering the MPJPE from 56.3 mm to 48.1 mm. mAP on the RADDet dataset increases from 0.43 to 0.48 with a 11.6% relative gain.
- On average, pre-training with WiSwiss improves performance by 4.5% on WiFi tasks and 10.3% on mmWave tasks. One explanation for the greater gains on mmWave is that its higher operating frequency can provide less environmentally sensitive representations [90] compared to WiFi signals.

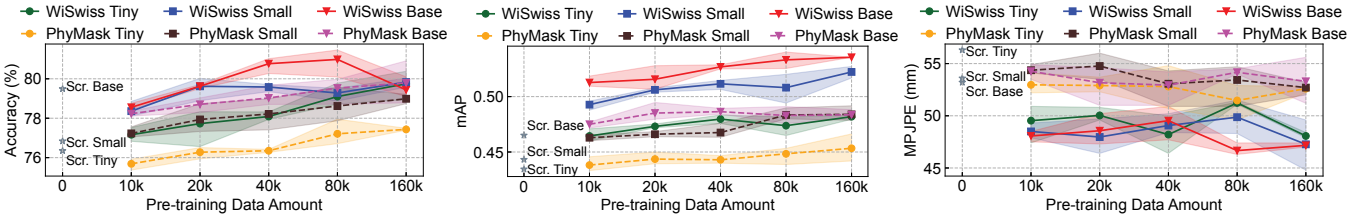
Overall, WiSwiss demonstrates higher accuracy and lower error compared to PhyMask across the majority of benchmarks, with only isolated exceptions, such as the small model on the WiFi XRF55

[1]We adopt the weighted adaptive mask Generation (W-PhyMask) variant, assigning equal weights of 0.5 to the energy distribution and coherence masking components.



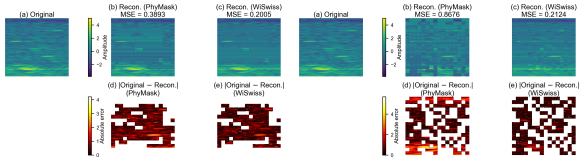
(a) XRF55 (↑) (b) CSI-Bench (↑) (c) Wi-Pose (↓)

Figure 6: Scaling of Pre-training Data Amount and Model Size in WiFi Modality.



(a) XRF55 (↑) (b) RADDet (↑) (c) HuPR (↓)

Figure 7: Scaling of Pre-training Data Amount and Model Size in mmWave Modality.



(a) PhyMask Mask Scheme (b) Random Mask Scheme

Figure 8: Reconstruction Comparison.

dataset. Specifically, the average performance improvements for the WiFi modality are -0.05% on XRF55, 2.4% on CSI-Bench, and 2.8% on Wi-Pose. For the mmWave modality, WiSwiss achieves gains of 1.6% on XRF55, 7.6% on RADDet, and 8.6% on HuPR.

To further analyze the gains of WiSwiss over PhyMask, particularly for mmWave, we conduct an intermediate reconstruction study on the RADDet dataset, which is used only for fine-tuning and is unseen during pre-training for both methods. As shown in Fig. 8, we evaluate reconstruction under two masking strategies: PhyMask masking and random masking. In both cases, WiSwiss achieves lower Mean Squared Error (MSE) than PhyMask, indicating more accurate reconstructions. We attribute this improvement to the sparsity of informative regions in mmWave signals. Masking strategies that prioritize specific regions can lead to overfitting, causing the model to memorize background noise rather than learn meaningful foreground structure. In contrast, our approach combines random masking with a foreground CFAR loss, which encourages the model to learn representations better suited to mmWave sensing tasks.

Pre-training Data Volume Impact. Increasing the volume of pre-training data generally improves downstream performance, confirming the benefit of the multi-source strategy in providing sufficient data diversity and scale for learning generalizable representations. However, these gains exhibit diminishing returns as the corpus grows, consistent with established scaling laws in large-scale machine learning [35, 41]. We further observe that saturation occurs earlier for less complex tasks, as indicated by higher baseline performance. For example, the performance of CSI-Bench and mmWave XRF55 classification tasks plateaus earlier than the more challenging WiFi XRF55 task.

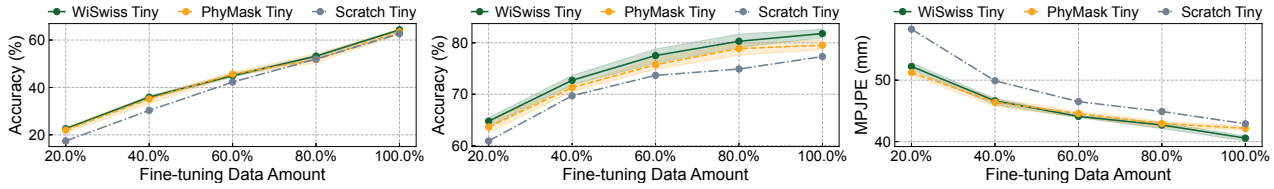
Data Source Impact. Sequentially incorporating data from distinct sources consistently improves downstream performance. This trend highlights the importance of data diversity for generalization, as a single-source dataset is often limited in scale and fails to capture the full range of real-world variations. By aggregating multiple datasets, the proposed approach not only increases data volume but also captures a richer set of variations in signal characteristics.

Model Size Impact. Larger models consistently achieve higher peak performance on downstream tasks. We observe that larger models benefit more from increased data and require larger corpora to reach their full capacity. For example, on the WiFi XRF55 dataset, the base model continues to improve up to 80k samples, whereas the tiny model saturates at around 20k. A similar trend is observed for mmWave: on HuPR, the base model requires 80k samples compared with 40k for the tiny model, and on mmWave XRF55 dataset, it requires 80k samples compared with 20k for the small model.

Summary of Scaling Insights. In summary, the experiments reveal three key scaling dynamics for wireless pre-training. First, the proposed pre-training approach consistently outperforms training from scratch. Second, performance improves with both the size and diversity of the pre-training corpus but exhibits diminishing returns.Finally, model and data scale must be matched: larger models achieve higher performance but require diverse datasets.

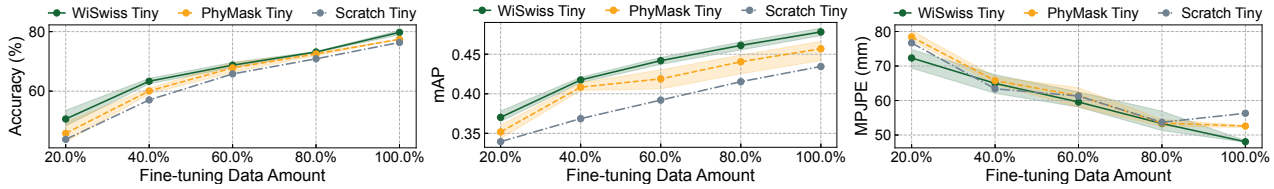
8.3 Fine-tuning Scaling

We next analyze the impact of fine-tuning dataset size on downstream performance. Detailed results are shown in Fig. 9 and Fig. 10, where the solid line represents the performance of WiSwiss with 160k pre-training samples, and the dotted lines show PhyMask and training from scratch using the tiny model. All methods improve as more fine-tuning data becomes available. Our analysis focuses on



(a) XRF55 (↑) (b) CSI-Bench (↑) (c) Wi-Pose (↓)

Figure 9: Scaling of Fine-tuning Data Amount in WiFi Modality.



(a) XRF55 (↑) (b) RADDet (↑) (c) HuPR (↓)

Figure 10: Scaling of Fine-tuning Data Amount in mmWave Modality.

two main benefits of pre-training: (1) performance amplification, especially in low-data regimes, and (2) improved data efficiency.

Performance Amplification in Low-Data Regimes. A key advantage of pre-training is that its gains are most pronounced when labeled fine-tuning data is scarce. This result indicates that the learned representations provide a strong prior and act as an effective regularizer against overfitting on small datasets [8, 15]. This effect is clear on the WiFi XRF55 dataset. With only 20% of the training data, the model trained from scratch achieves 17.4% accuracy. In contrast, the pre-trained WiSwiss model reaches 22.6% accuracy with a 29.9% relative gain. With the full training set, the from-scratch baseline reaches 62.7% and WiSwiss achieves 64.3%, a smaller relative gain of 2.6%. This pattern of diminishing relative improvement is observed across nearly all datasets.

Pre-training for Data Efficiency. We further quantify the data efficiency of pre-training by estimating how much less fine-tuning data WiSwiss requires to match the performance of a baseline model trained on full dataset. As an exhaustive sweep is not performed, these estimates are obtained via linear interpolation of the performance curves. The pre-trained model consistently requires less data to reach the baseline peak performance. For example, on CSI-Bench, the from-scratch model achieves 77.3% accuracy, while WiSwiss reaches 72.7% with 40% data and 77.5% with 60% data, corresponding to an estimated 40.8% reduction in required data. The estimated data savings are 3.0% on WiFi XRF55 and 22.9% on Wi-Pose, yielding an average of 22.2% for WiFi datasets. For mmWave, the savings are 10.3%, 46.1%, and 29.5% on XRF55, RADDet, and HuPR, respectively, with an average of 28.6%.

Advantage over PhyMask. Overall, WiSwiss outperforms PhyMask across most experimental settings. In the WiFi modality, WiSwiss achieves average improvements of 1.5% on XRF55, 2.2% on CSI-Bench, and 0.6% on Wi-Pose. In the mmWave modality, the gains are larger, with improvements of 4.2% on XRF55, 4.5% on RADDet, and 4.2% on HuPR.

Summary of Fine-tuning Insights. The analysis highlights two primary benefits of pre-training with WiSwiss. First, it improves performance in low-data regimes, which makes it suited for scenarios where labeled data is limited or costly to obtain. Second, it improves data efficiency consistently. These results demonstrate the practical value of the framework for real-world deployment.

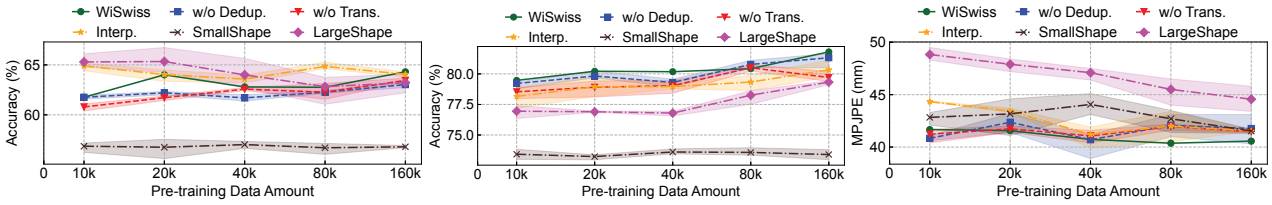
8.4 Ablation Studies

To quantify the contribution of each design component, we compare the full WiSwiss framework with five ablated variants. As described in §8.1.2, these variants examine the effects of data curation, training objectives, and positional encoding. For the positional embedding variants, the *Small Shape* setting resizes all inputs to the minimum dimensions observed in the fine-tuning datasets, while the *Large Shape* setting resizes all inputs to the maximum dimensions. For WiFi, these correspond to (342, 5) from Wi-Pose and (96, 1000) from XRF55, respectively. For mmWave, the corresponding shapes are (8, 64, 64) from HuPR and (16, 256, 256) from RADDet. In the *Interpolation* variant, the initial positional embedding size is set to match that of the large-shape configuration. The results are shown in Fig. 11 and Fig. 12, where solid lines denote the full WiSwiss and dotted lines denote the ablated variants. For a direct quantitative comparison, Tab. 5 reports the average performance degradation of each component across the full pre-training corpus.

Necessity of Data Curation and Better Model Optimization. Across all datasets, removing the transformation-invariant loss results in a performance drop, which suggests that our pre-training objective is important to learn generalizable representations. Besides, training on uncurated data consistently degrades performance. This aligns with findings from the language domain, where data repetition is known to cause models to memorize over-represented patterns, which harms generalization and overall performance [46, 65].

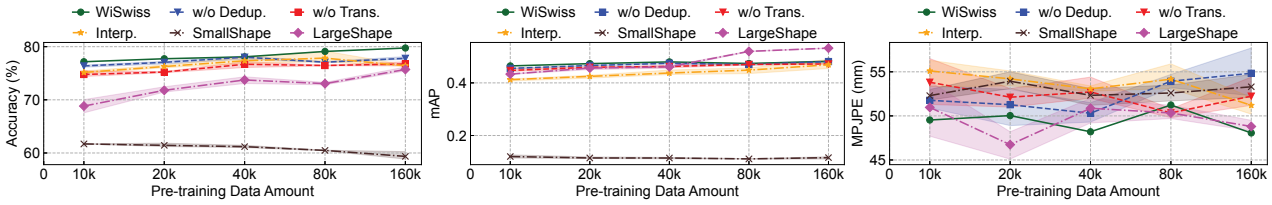
Necessity of Shape Adaptation. As shown in Tab. 5, relying on fixed position embeddings generally leads to performance degradation across most experimental setups.

- **Position Interpolation:** The interpolation variant shows performance degradation in nearly all cases, with WiFi XRF55 as the only exception, showing a slight improvement. However, as the pre-training data volume increases (Fig. 11a), it eventually



(a) XRF55 (↑) (b) CSI-Bench (↑) (c) Wi-Pose (↓)

Figure 11: Ablation Studies in WiFi Modality



(a) XRF55 (↑) (b) RADDet (↑) (c) HuPR (↓)

Figure 12: Ablation Studies in mmWave Modality

underperforms our RoPE method. This behavior is likely caused by accumulated shape distortions.

- **Small Shape:** Resizing all inputs to the smallest fine-tuning shape leads to substantial performance degradation. In particular, half of the benchmarks exhibit drops exceeding 10%.
- **Large Shape:** Resizing all inputs to the maximum shape yields mixed results. Datasets that naturally match this resolution and avoid interpolation distortion show marginal improvements over RoPE However, forcing all datasets to this size increases input dimensions by more than 15×. Overall, the RoPE approach achieves a better balance, providing strong average performance without unnecessary computational cost.

Summary of Ablation Insights. In summary, our ablation studies confirm the necessity of each proposed design component. Systematic data deduplication remains a critical step for high-quality representation learning, while our transformation-invariant loss enhances the model's capacity to extract robust features. Besides, the model's ability to dynamically adapt to diverse data shapes avoids the distortions associated with fixed-dimension scaling, proving essential for broad generalization.

Table 5: Average Performance Degradation.

Variants	WiFi			mmWave		
	XRF55 (↑)	CSI-Bench (↑)	Wi-Pose (↓)	XRF55 (↑)	RADDet (↑)	HuPR (↓)
w/o Deduplication	1.5%	0.4%	1.4%	1.4%	1.1%	6.1%
w/o Transformation	1.5%	1.3%	1.3%	3.0%	2.6%	5.8%
Interpolation	-1.9%	1.6%	3.7%	2.2%	7.7%	8.4%
Small Shape	**10.0%**	8.7%	4.6%	**22.3%**	**75.7%**	7.1%
Large Shape	-1.6%	3.4%	**14.1%**	7.3%	-1.2%	0.3%

8.5 Sensitivity Study

To study the effect of deduplication parameters while controlling for total data volume, we conduct a sensitivity analysis on a fixed pool of 40,000 un-deduplicated WiFi samples. We vary two key parameters independently: the additive Gaussian noise variance (σ^2) and the distribution percentile (p) used to determine the similarity threshold. The resulting data removal ratios and downstream performance are reported in Tab. 6, where the best result for each dataset is shown in bold and the worst is underlined.

Across all datasets, the best performance is achieved when the data removal ratio remains below 10%. By contrast, aggressive deduplication, such as setting σ^2 to 5.0^2 and removing more than 40% of the data, substantially degrades downstream performance on XRF55 and CSI-Bench. This result suggests that overly strict similarity thresholds remove informative patterns and reduce the coverage and diversity of the pre-training corpus.

Table 6: Sensitivity of Deduplication Parameters.

Setting	σ^2	p (%)	Removed (%)	XRF55 ↑	CSI-Bench ↑	Wi-Pose ↓
Default	0.5^2	10	9.7	66.0	**82.1**	41.0
Vary AWGN variance σ^2 (fix p = 10%)						
σ^2 low	0.1^2	10	6.7	**67.5**	81.9	41.5
σ^2 high	2.0^2	10	21.9	65.2	81.7	41.1
σ^2 very high	5.0^2	10	40.2	62.0	79.8	41.0
Vary percentile p (fix $\sigma^2 = 0.5^2$)						
p low	0.5^2	1	5.1	65.4	81.8	**39.8**
p high	0.5^2	40	14.1	65.8	81.0	41.2
p very high	0.5^2	80	18.4	65.4	81.5	41.0

We also examine the sensitivity of our framework to the CFAR threshold, which controls the required prominence of foreground objects relative to the background. As shown in Tab. 7, an excessively high threshold suppresses valid foreground detections, causing them to be excluded from the loss computation and resulting in degraded performance. Conversely, a too low threshold causes the model to attend to background noise, which obscures the structural representations of true objects and impairs performance.

Table 7: Sensitivity of CFAR Threshold.

Setting	CFAR threshold	XRF55 ↑	RADDet ↑	HuPR ↓
Default	5.0	**78.1**	**48.0**	48.2
low	2.0	76.1	44.5	52.4
high	10.0	76.8	46.3	**47.1**
very high	20.0	72.8	44.1	53.2

8.6 Robustness to Data Configuration

To verify that the observed gains are not tied to a particular data configuration, we evaluate an alternative setup by reversing the roles of the pre-training and fine-tuning datasets. We replicate the training pipeline using the tiny model, with 40k samples for pre-training and 20k samples for fine-tuning. To ensure statistical reliability,

all fine-tuning experiments use 5-fold cross-validation, and we report the average performance. As shown in Tab. 8, WiSwiss still consistently outperforms the baseline trained from scratch. The proposed approach achieves an average improvement of 9.9% for WiFi and 2.8% for mmWave. The lower absolute performance on Widar, MM-Fi, and RT-Pose, is due to the smaller fine-tuning budget.

Table 8: Results under Another Data Configuration.

Modality	Dataset	Metric	Scratch	WiSwiss	Improvement
WiFi	SignFi	Acc (%) ↑	57.6	59.8	3.7%
	Widar	Acc (%) ↑	40.1	44.3	10.5%
	MM-Fi	MPJPE (mm) ↓	92.6	78.3	15.4%
mmWave	MM-DCDR	Acc (%) ↑	94.7	98.2	3.7%
	MCD	Acc (%) ↑	92.4	94.0	1.7%
	RT-Pose	MPJPE (mm) ↓	280.3	271.3	3.1%

8.7 Training Efficiency

We analyze the computational efficiency of the data curation and pre-training pipeline in Tab. 9. For the WiFi modality, training the deduplication encoder incurs a modest one-time cost of 0.16 GPU hours. The inference stage requires 20 seconds for 10k samples. In addition to improving performance, deduplication reduces overall pre-training time by removing redundant data: When pre-training the base model on 160k samples with a duplicate rate of 9.7%, the reduction in training time (0.33 hours) exceeds the deduplication overhead (0.25 hours). Moreover, since the deduplication encoder is trained only once, this cost can be amortized across experiments. Similar trends are also observed for the mmWave modality.

Table 9: Computational Cost Analysis.

Modality	**Pre-training** (10k samples)			**Deduplication**	
	Tiny	Small	Base	Training (total)	Inference (10k samples)
WiFi	0.15 hours	0.17 hours	0.21 hours	0.16 hours	20 seconds
mmWave	0.32 hours	0.44 hours	0.80 hours	0.32 hours	76 seconds

9 Discussion

Role of Synthetic Data. An alternative approach to improving generalization in wireless sensing is to use generative models, which addresses data scarcity by increasing the quantity and diversity of training samples [14, 26, 80, 93]. In contrast, our work focuses on learning general-purpose representations from large-scale heterogeneous real-world datasets. The two paradigms are complementary: synthetic data that emulate unseen scenarios can serve as an additional source for pre-training, potentially improving robustness and domain coverage. A systematic study of this integration remains an important direction for future work.

Multi-modal Generalist Models. A natural consideration is to develop a single backbone that can process multiple wireless modalities. However, due to differences in physical properties and data structures, this work focuses on modality-specific generalist models as a first step. A unified multi-modal backbone would require addressing fundamental challenges, including large-scale paired data collection and cross-modal representation alignment [64, 87]. Although recent studies have explored alignment between wireless signals and vision [76] or language [9], they are limited to specific tasks and do not provide a general-purpose backbone.

Device Heterogeneity across Datasets. While the proposed framework natively supports variable input shapes, it does not explicitly model hardware-specific characteristics such as bandwidth or center frequency. We hypothesize that pre-training on a diverse corpus enables the model to learn device-invariant features, a capability encouraged by our objective. The generalization performance across datasets collected with different hardware provides empirical support for this hypothesis. Explicit modeling of device-specific factors, for example through an adaptive input layer, is left for future work.

Better Fine-tuning Method. We employed a full-model fine-tuning approach, as the primary focus of this paper is to evaluate the generalization of pre-trained representations with multi-source corpus. Possible optimization in the fine-tuning phase includes LoRA for efficiency [37] or multi-layer probing for performance [20]. We consider their exploration an important but orthogonal direction.

More Rigorous Scaling Law. Our experiments provide the first empirical evidence for scaling laws in the wireless sensing domain. We acknowledge that our analysis is qualitative compared to the rigorous power-law curves established in the language domain [35, 41]. However, current scarcity of large public datasets makes a similarly precise study impractical at this time. Instead, our goal is to demonstrate the existence and direction of these trends.

Magnitude of Performance Gains. The average performance gains of WiSwiss are 4.5% for WiFi and 10.3% for mmWave. Although these improvements may appear modest, they are consistent with prior pre-training results in wireless sensing [69, 84] and are comparable to gains reported for seminal models such as BERT on language benchmarks using a 3.3-billion-word pre-training corpus [15]. This comparison suggests a clear direction: achieving transformative performance, as observed in models such as GPT-4, requires further scaling both data and model capacity [3].

Compatibility with Other Data Formats. The proposed framework standardizes inputs to 2D WiFi CSI arrays and 3D mmWave data cubes, which introduces compatibility challenge for other task-specific formats. For example, the unidirectional CSI-to-DFS objective does not support models that take DFS as input. Similarly, the grid-based ViT architecture is not directly applicable to sparse point clouds commonly used in mmWave applications [10, 85]. Future work may address these limitations by exploring bi-directional transformations or pre-trained models for other formats.

10 Related Work

Self-Supervised Learning for Time-Frequency Data. The challenge of learning from 2D time-frequency data is not unique to wireless sensing. The spectrograms of audio with time-frequency structure [17] has inspired many specialized pre-training objectives. These methods often augment standard masked reconstruction with a complementary task to learn richer features. For instance, some models incorporate a patch-level instance discrimination task to improve local feature quality [27, 28], while others focus on learning multi-scale features to capture both fine-grained and global context [13]. Inspired by this method of enhancing reconstruction with domain-specific objectives, our work uses signal transformations to encourage better semantic understanding of wireless data.

Self-Supervised Learning for Wireless Data. Adapting strategies from other domains, SSL for wireless signals has largely diverged into two primary paradigms. The first masked modeling focuses on learning fine-grained local features. Existing efforts [22, 42] advance it by keeping informative signal regions, such as those with higher energy. The second paradigm learns global semantic representations with contrastive learning, optimizing the model to produce similar representations for signal augmentations [68, 69]. However, these approaches are explored in isolation, and suffer from missing global semantics and augmentation-dependent shortcut. Our work is the first to bridge this gap by introducing a unified objective, and learns a robust representation that captures essential features of local and global scales.

11 Conclusion

We present WiSwiss, a pre-training framework for generalizable wireless representations by integrating careful multi-source data curation with a transformation-invariant pre-training objective. The robust performance of these representation across various downstream tasks demonstrates the practical viability of a unified wireless generalist backbone.s Furthermore, our qualitative scaling law analysis presents a call to action: Achieving the next level of performance will necessitate a concerted community effort focused on both large-scale data collection and the co-design of scalable models and objectives, a direction for which our work establishes the essential principles.

Acknowledgments

We sincerely thank the anonymous shepherd and reviewers for their insightful critique and constructive feedback, which have significantly improved the quality of this paper. This work is supported by National Natural Science Foundation of China (Grant No. U25B2040 and 62272010). Chenren Xu (chenren@pku.edu.cn) is the corresponding author.

References

[1] Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. 2023. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540* (2023).
[2] Moustafa Abbas, Moustafa Elhamshary, Hamada Rizk, Marwan Torki, and Moustafa Youssef. 2019. WiDeep: WiFi-based accurate and robust indoor localization system using deep learning. In *IEEE PerCom*.
[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv:2303.08774* (2023).
[4] Mohiuddin Ahmed, Raihan Seraj, and Syed Mohammed Shamsul Islam. 2020. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics* (2020).
[5] Arash Khoeini. 2025. InfoNCE Loss - PyTorch Implementation. https://github.com/arashkhoeini/infonce [Online; accessed 23-August-2025].
[6] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. ViViT: A Video Vision Transformer. arXiv:arXiv 2103.15691
[7] Rishi Bommasani. 2021. On the opportunities and risks of foundation models. *arXiv:2108.07258* (2021).
[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *NeurIPS* (2020).
[9] Qiming Cao, Hongfei Xue, Tianci Liu, Xingchen Wang, Haoyu Wang, Xincheng Zhang, and Lu Su. 2024. mmclip: Boosting mmwave-based zero-shot har via signal-text alignment. In *ACM SenSys*.
[10] Anjun Chen, Xiangyu Wang, Shaohao Zhu, Yanxu Li, Jiming Chen, and Qi Ye. 2022. mmbody benchmark: 3d body reconstruction dataset and analysis for millimeter wave radar. In *ACM MM*.
[11] Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. 2023. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research* (2023).
[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *ICML*.
[13] Wenxi Chen, Yuzhe Liang, Ziyang Ma, Zhisheng Zheng, and Xie Chen. 2024. EAT: Self-Supervised Pre-Training with Efficient Audio Transformer. In *IJCAI*.
[14] Guoxuan Chi, Zheng Yang, Chenshu Wu, Jingao Xu, Yuchong Gao, Yunhao Liu, and Tony Xiao Han. 2024. RF-Diffusion: Radio Signal Generation via Time-Frequency Diffusion. *ACM MobiCom* (2024).
[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *ACL*.
[16] Shuya Ding, Zhe Chen, Tianyue Zheng, and Jun Luo. 2020. RF-net: A unified meta-learning framework for RF-enabled one-shot human activity recognition. In *ACM SenSys*.
[17] Xuefu Dong, Liqiang Xu, Lixing He, Zengyi Han, Ken Christofferson, Yifei Chen, Akihito Taya, Yuuki Nishiyama, and Kaoru Sezaki. 2025. Poster: Recognizing Hidden-in-the-Ear Private Key for Reliable Silent Speech Interface Using Multi-Task Learning. *ACM UbiComp* (2025).
[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929* (2020).
[19] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. *arXiv:2401.08281* (2024).
[20] Utku Evci, Vincent Dumoulin, Hugo Larochelle, and Michael C Mozer. 2022. Head2toe: Utilizing intermediate representations for better transfer learning. In *ICML*.
[21] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2017. RMPE: Regional Multi-person Pose Estimation. In *IEEE ICCV*.
[22] Liang Fang, Ruiyuan Song, Zhi Lu, Dongheng Zhang, Yang Hu, Qibin Sun, and Yan Chen. 2024. Prism: Pre-training rf signals in sparsity-aware masked autoencoders. In *IEEE INFOCOM*.
[23] Yating Gao, Ruixu Geng, Dongheng Zhang, Yang Hu, Hui Lin, and Yan Chen. 2024. MM-DCDR: A Benchmark of Device Configuration and Data Representation for mmWave-Based Human Sensing. In *IEEE WCSP*.
[24] Zijun Gao, Haibao Liu, and Lingbo Li. 2025. Data augmentation for time-series classification: An extensive empirical study and comprehensive survey. *JAIR* 83 (2025).
[25] Gene H Golub and Charles F Van Loan. 2013. *Matrix computations*. JHU press.
[26] Chen Gong, Bo Liang, Wei Gao, and Chenren Xu. 2025. Data Can Speak for Itself: Quality-guided Utilization of Wireless Synthetic Data. *ACM MobiSys* (2025).
[27] Yuan Gong, Yu-An Chung, and James Glass. 2021. AST: Audio Spectrogram Transformer. In *Interspeech*.
[28] Yuan Gong, Cheng-I Lai, Yu-An Chung, and James Glass. 2022. SSAST: Self-Supervised Audio Spectrogram Transformer. In *AAAI*.
[29] Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. 2022. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate.
[30] Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. 2024. A Survey on Self-Supervised Learning: Algorithms, Applications, and Future Trends. *IEEE TPAMI* (2024).
[31] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF CVPR*.
[32] Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. 2024. Rotary position embedding for vision transformer. In *Springer ECCV*.
[33] Indra Herdiana, M Alfin Kamal, Mutia Nur Estri, et al. 2025. A More Precise Elbow Method for Optimum K-means Clustering. *arXiv:2502.00851* (2025).
[34] Yuan-Hao Ho, Jen-Hao Cheng, Sheng Yao Kuan, Zhongyu Jiang, Wenhao Chai, Hsiang-Wei Huang, Chih-Lung Lin, and Jenq-Neng Hwang. 2024. RT-Pose: A 4D Radar Tensor-based 3D Human Pose Estimation and Localization Benchmark. *arXiv:2407.13930* (2024).
[35] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv:2203.15556* (2022).
[36] Weiying Hou and Chenshu Wu. 2024. RFBoost: Understanding and Boosting Deep WiFi Sensing via Physical Data Augmentation. *ACM IMWUT* (2024).
[37] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. arXiv 2021. *arXiv:2106.09685* (2021).
[38] Kaixuan Huang, Yuqing Wang, Molei Tao, and Tuo Zhao. 2020. Why Do Deep Residual Networks Generalize Better than Deep Feedforward Networks?—A Neural Tangent Kernel Perspective. *NeurIPS* (2020).

[39] Jonathan Hui. [n. d.]. mAP (mean Average Precision) for Object Detection. https://jonathan-hui.medium.com/map-mean-average-precision-for-object-detection-45c121a31173. [Online; accessed 31-August-2025].

[40] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* (2019).

[41] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv:2001.08361* (2020).

[42] Denizhan Kara, Tomoyoshi Kimura, Yatong Chen, Jinyang Li, Ruijie Wang, Yizhuo Chen, Tianshi Wang, Shengzhong Liu, and Tarek Abdelzaher. 2024. PhyMask: An Adaptive Masking Paradigm for Efficient Self-Supervised Learning in IoT. In *ACM SenSys*.

[43] Saurabh Kataria, Yi Wu, Zhaoliang Chen, Hyunjung Gloria Kwak, Yuhao Xu, Lovely Yeswanth Panchumarthi, Ran Xiao, Jiaying Lu, Ayca Ermis, Anni Zhao, Runze Yan, Alex Federov, Zewen Liu, Xu Wu, Wei Jin, Carl Yang, Jocelyn Grunwell, Stephanie R. Brown, Amit Shah, Craig S. Jabaley, Tim Buchman, Sivasubramanium V. Bhavani, Randall J. Lee, and Xiao Hu. 2025. Generalist vs Specialist Time Series Foundation Models: Investigating Potential Emergent Behaviors in Assessing Human Health Using PPG Signals.

[44] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *NeurIPS* (2020).

[45] Ruiqi Kong and He Chen. 2025. CIRSense: Rethinking WiFi Sensing with Channel Impulse Response. *arXiv preprint arXiv:2510.11374* (2025).

[46] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2021. Deduplicating training data makes language models better. *arXiv:2107.06499* (2021).

[47] Shih-Po Lee, Niraj Prakash Kini, Wen-Hsiao Peng, Ching-Wen Ma, and Jenq-Neng Hwang. 2023. HuPR: A Benchmark for Human Pose Estimation Using Millimeter Wave Radar. In *IEEE/CVF WACV*.

[48] Chenning Li, Zhichao Cao, and Yunhao Liu. 2021. Deep AI Enabled Ubiquitous Wireless Sensing. *ACM CSUR* (2021).

[49] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. 2023. Multimodal Foundation Models: From Specialists to General-Purpose Assistants. *Found. Trends Comput. Graph. Vis.* (2023).

[50] Shengqiang Li, Menglong Xu, and Xiao-Lei Zhang. 2021. Conformer-based End-to-end Speech Recognition With Rotary Position Embedding. *APSIPA ASC* (2021).

[51] Tianhong Li, Lijie Fan, Yuan Yuan, and Dina Katabi. 2022. Unsupervised learning for human sensing using radio signals. In *IEEE/CVF WACV*.

[52] Yadong Li, Dongheng Zhang, Jinbo Chen, Jinwei Wan, Dong Zhang, Yang Hu, Qibin Sun, and Yan Chen. 2022. DI-Gesture: Domain-Independent and Real-Time Gesture Recognition with Millimeter-Wave Signals. In *GLOBECOM*.

[53] Yadong Li, Dongheng Zhang, Jinbo Chen, Jinwei Wan, Dong Zhang, Yang Hu, Qibin Sun, and Yan Chen. 2022. Towards Domain-Independent and Real-Time Gesture Recognition Using Mmwave Signal. *IEEE Transactions on Mobile Computing* (2022).

[54] Yang Liu, Feng Wang, Naiyan Wang, and Zhao-Xiang Zhang. 2023. Echoes beyond points: Unleashing the power of raw radar data in multi-modality fusion. *NeurIPS* (2023).

[55] Yongsen Ma, Gang Zhou, Shuangquan Wang, Hongyang Zhao, and Woosub Jung. 2018. SignFi: Sign language recognition using WiFi. *ACM IMWUT* (2018).

[56] Yimin Mao, Zhengxin Guo, Biyun Sheng, Linqing Gui, and Fu Xiao. 2024. Wi-Cro: WiFi-Based Cross Domain Activity Recognition via Modified GAN. *IEEE TVT* (2024).

[57] Sai Munikoti, Ian Stewart, Sameera Horawalavithana, Henry Kvinge, Tegan H. Emerson, Sandra E Thompson, and Karl Pazdernik. 2024. Generalist Multimodal AI: A Review of Architectures, Challenges and Opportunities. *ArXiv* (2024).

[58] OECD.AI. 2025. Mean Per Joint Position Error (MPJPE). https://oecd.ai/en/catalogue/metrics/mean-per-joint-position-error-mpjpe [Online; accessed 20-August-2025].

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

[60] OpenAI. 2025. CLIP input image size. https://github.com/openai/CLIP/issues/379 [Online; accessed 24-August-2025].

[61] Xiaomin Ouyang, Xian Shuai, Jiayu Zhou, Ivy Wang Shi, Zhiyuan Xie, Guoliang Xing, and Jianwei Huang. 2022. Cosmo: contrastive fusion learning with small data for multimodal human activity recognition. In *ACM MobiCom*.

[62] Aldo Pareja, Nikhil Shivakumar Nayak, Hao Wang, Krishnateja Killamsetty, Shivchander Sudalairaj, Wenlong Zhao, Seungwook Han, Abhishek Bhandwaldar, Guangxuan Xu, Kai Xu, et al. 2024. Unveiling the secret recipe: A guide for supervised fine-tuning small llms. *arXiv:2412.13337* (2024).

[63] John G Proakis and Masoud Salehi. 2001. *Digital communications*. McGraw-hill New York.

[64] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

[65] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR* (2020).

[66] Karthik Ramasubramanian. 2017. mmWave radar for automotive and industrial applications. *Texas Instruments* (2017).

[67] Ruiyuan Song, Zhi Lu, Dongheng Zhang, Liang Fang, Zhi Wu, Yang Hu, Qibin Sun, and Yan Chen. 2025. Unleashing the Potential of Self-Supervised RF Learning With Group Shuffle . *IEEE TMC* (2025).

[68] Ruiyuan Song, Dongheng Zhang, Zhi Wu, Cong Yu, Chunyang Xie, Shuai Yang, Yang Hu, and Yan Chen. 2022. RF-URL: unsupervised representation learning for RF sensing. In *ACM MobiCom*.

[69] Ruiyuan Song, Dongheng Zhang, Zhi Wu, Cong Yu, Chunyang Xie, Shuai Yang, Yang Hu, and Yan Chen. 2025. RF-URL 2.0: A General Unsupervised Representation Learning Method for RF Sensing. *IEEE TPAMI* (2025).

[70] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Elsevier Neurocomputing* (2024).

[71] TensorFlow. 2025. C4 Dataset. https://www.tensorflow.org/datasets/catalog/c4 [Online; accessed 22-August-2025].

[72] David Tse and Pramod Viswanath. 2005. *Fundamentals of wireless communication*. Cambridge university press.

[73] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *ICML*.

[74] Fei Wang, Yizhe Lv, Mengdie Zhu, Han Ding, and Jinsong Han. 2024. XRF55: A Radio Frequency Dataset for Human Indoor Action Analysis. *ACM IMWUT* (2024).

[75] Wei Wang, Alex X Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. 2015. Understanding and modeling of wifi signal based human activity recognition. In *ACM MobiCom*.

[76] Yuxuan Weng, Guoquan Wu, Tianyue Zheng, Yanbing Yang, and Jun Luo. 2024. Large model for small data: Foundation model for cross-modal rf human activity recognition. In *ACM SenSys*.

[77] Wikipedia Contributors. 2025. Constant false alarm rate. https://en.wikipedia.org/wiki/Constant_false_alarm_rate [Online; accessed 23-August-2025].

[78] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. 2020. Visual Transformers: Token-based Image Representation and Processing for Computer Vision. arXiv:arXiv 2006.03677

[79] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. 2025. Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. *Nature Communications* (2025).

[80] Chunjing Xiao, Daojun Han, Yongsen Ma, and Zhiguang Qin. 2019. CsiGAN: Robust channel state information-based activity recognition with GANs. *IEEE IoTJ* (2019).

[81] Rui Xiao, Jianwei Liu, Jinsong Han, and Kui Ren. 2021. OneFi: One-Shot Recognition for Unseen Gesture via COTS WiFi. *ACM SenSys* (2021).

[82] Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. 2022. Revealing the Dark Secrets of Masked Image Modeling. *IEEE/CVF CVPR* (2022).

[83] Jianfei Yang, Xinyan Chen, Han Zou, Chris Xiaoxuan Lu, Dazhuo Wang, Sumei Sun, and Lihua Xie. 2023. SenseFi: A library and benchmark on deep-learning-empowered WiFi human sensing. *Patterns* (2023).

[84] Jianfei Yang, Xinyan Chen, Han Zou, Dazhuo Wang, and Lihua Xie. 2022. Autofi: Toward automatic wi-fi human sensing via geometric self-supervised learning. *IEEE Internet of Things Journal* 10, 8 (2022), 7416–7425.

[85] Jianfei Yang, He Huang, Yunjiao Zhou, Xinyan Chen, Yuecong Xu, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. 2023. Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing. *NeurIPS* (2023).

[86] Yuchong Yao, Nandakishor Desai, and Marimuthu Palaniswami. 2022. Masked contrastive representation learning. *arXiv:2211.06012* (2022).

[87] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review* (2024).

[88] Ao Zhang, Farzan Erlik Nowruzi, and Robert Laganiere. 2021. RADDet: Range-Azimuth-Doppler based Radar Object Detection for Dynamic Road Users. In *CRV*.

[89] Jie Zhang, Zhanyong Tang, Meng Li, Dingyi Fang, Petteri Nurmi, and Zheng Wang. 2018. CrossSense: Towards Cross-Site and Large-Scale WiFi Sensing. In *ACM MobiCom*.

[90] Jia Zhang, Rui Xi, Yuan He, Yimiao Sun, Xiuzhen Guo, Weiguo Wang, Xin Na, Yunhao Liu, Zhenguo Shi, and Tao Gu. 2023. A survey of mmWave-based human sensing: Technology, platforms and applications. *IEEE Communications Surveys & Tutorials* (2023).

[91] Xie Zhang, Chengpei Tang, Yasong An, and Kang Yin. 2021. WiFi-based Multi-task Sensing. *ArXiv* (2021).

[92] Yi Zhang, Yue Zheng, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. 2021. Widar3. 0: Zero-effort cross-domain gesture recognition with

Wi-Fi. *IEEE TPAMI* (2021).
[93] Xiaopeng Zhao, Zhenlin An, Qingrui Pan, and Lei Yang. 2023. Nerf2: Neural radio-frequency radiance fields. In *ACM MobiCom*.
[94] Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. 2021. Neighborhood Contrastive Learning for Novel Class Discovery. In *IEEE/CVF CVPR*.
[95] Guozhen Zhu, Yuqian Hu, Weihang Gao, Wei-Hsiang Wang, Beibei Wang, and KJ Liu. 2025. CSI-Bench: A Large-Scale In-the-Wild Dataset for Multi-task WiFi Sensing. *arXiv:2505.21866* (2025).
[96] Guozhen Zhu, Yuqian Hu, Sakila Jayaweera, Weihang Gao, Wei-Hsiang Wang, Jiaxuan Zhang, Beibei Wang, Chenshu Wu, and KJ Liu. 2026. AM-FM: A Foundation Model for Ambient Intelligence Through WiFi. *arXiv:2602.11200* (2026).