Cross-Sensor Domain Generalization for Non-Contact Sleep Staging

Jie Deng[1], **Zhi Lu**[1], **Fang Zhou**[1], **Yu Pu**[1], **Zhi Wu**[1], **Beilei Wang**[2], **Yang Hu**[1] and **Yan Chen**[1*]
[1] School of Cyber Science and Technology, University of Science and Technology of China
[2] Department of Cardiovascular,The First Affiliated Hospital of USTC , Division of Life Sciences and Medicine , University of Science and Technology of China
gutendzx@mail.ustc.edu.cn , zhilu@ustc.edu.cn , {fangzhou1958,puyu,wzwyyx}@mail.ustc.edu.cn,
{blwangxnk,eeyhu,eecyan}@ustc.edu.cn

Abstract

Sleep staging is pivotal for assessing sleep quality and clinical diagnostics. While contactless radio-frequency (RF) sensing offers an unobtrusive alternative to polysomnography (PSG), existing methods often struggle to generalize across diverse devices and environments due to the scarcity of annotated RF data. To address this limitation, we propose XSensorSleep, a cross-sensor domain generalization framework trained exclusively on large-scale respiratory datasets collected from chest and abdominal belts and directly transferable to RF-derived respiratory signals. To bridge the cross-sensor domain gap, the training data are partitioned into multiple pseudo-domains based on data sources and sensor types, encouraging the model to learn domain-invariant representations from heterogeneous respiratory patterns. Furthermore, we introduce a hierarchical alignment strategy in which Epoch-Level Feature Alignment (ELFA) mitigates sensor-specific morphological discrepancies, while Spectral Temporal Alignment (STA) captures invariant global sleep structure by aligning whole-night sleep transition dynamics through rotation-invariant spectral representations, enabling the model to emphasize stable physiological rhythms over device-dependent signal variations. Extensive evaluations on diverse real-world RF datasets from both home and clinical environments demonstrate that XSensorSleep achieves strong cross-sensor generalization, supporting practical, automated, and label-free non-contact sleep staging.

1 Introduction

Accurate sleep staging is fundamental for assessing sleep quality and diagnosing sleep disorders. According to the American Academy of Sleep Medicine (AASM) standard [Berry *et al.*, 2012], sleep is categorized into Wake, N1, N2, N3, and REM stages using 30-second epochs. Polysomnography (PSG) remains the clinical gold standard [Rundo and Downey III, 2019], but its reliance on multiple contact-based sensors makes it intrusive in practice and susceptible to the first-night effect [Tamaki *et al.*, 2005; Byun *et al.*, 2019; Ding *et al.*, 2022]. Despite recent advances in automated sleep staging [Thapa *et al.*, 2024; Lee *et al.*, 2025], these practical limitations continue to motivate the development of non-contact sleep monitoring approaches.

Among emerging paradigms, radio-frequency (RF) sensing has attracted increasing attention for non-contact sleep staging [Zhao *et al.*, 2017a; Zhai *et al.*, 2022; Wang *et al.*, 2025; Zhang *et al.*, 2019; Yu *et al.*, 2021]. RF signals reflected from the human body are modulated by subtle physiological motions, with respiration providing the most informative cues for sleep staging [Li *et al.*, 2024; Dong *et al.*, 2024; Duan *et al.*, 2025; Carley and Farabi, 2016]. This enables unobtrusive, long-term monitoring without body-worn sensors. However, practical adoption of RF-based sleep staging remains limited by the lack of large-scale, publicly available labeled RF datasets, which hinders robust model training and generalization.

In light of this limitation, the similarity between RF-derived respiratory signals and chest and abdominal belt recordings [Yue *et al.*, 2018; Zhang *et al.*, 2020; Hao *et al.*, 2024] enables the transfer of models trained on large-scale belt-based respiration datasets to RF-derived signals. However, such transfer is non-trivial due to fundamental differences in measurement characteristics across sensors. Belt sensors are tightly coupled to the body surface and provide high signal-to-noise respiratory waveforms, whereas RF sensing is highly sensitive to global body motion and environmental multi-path effects. In particular, non-respiratory movements during Wake periods can dominate RF reflections, intermittently corrupting or fragmenting respiratory signals and leading to lower and more variable signal quality than belt recordings. This results in pronounced cross-sensor distribution shifts in the absence of labeled RF data during training.

Most existing domain generalization methods promote robustness by enforcing similarity or consistency in learned representations across source domains, typically via feature alignment or invariance constraints [Zhou *et al.*, 2023; Wang *et al.*, 2022]. Such methods implicitly assume that tar-

[*]The corresponding author is Yan Chen

get domain shifts are moderate and interpolatable from the source domains. In cross-sensor sleep staging, this assumption may be violated, as RF sensing induces large and sensor-specific transformations beyond simple distributional variations [Zhao *et al.*, 2017b]. Consequently, strong similarity constraints can become overly restrictive [Jin *et al.*, 2020], and cross-sensor generalization may benefit from representations that preserve stable structural properties of sleep dynamics under nuisance transformations.

In this paper, we propose XSensorSleep, a cross-sensor domain generalization framework for non-contact sleep staging that leverages stable properties of sleep dynamics under sensor-induced variability. The central idea is that although RF-based respiratory signals are subject to substantial noise and distortion at fine temporal scales, the large-scale organization of sleep, namely sleep architecture, is governed by relatively stable physiological dynamics across sensing modalities. XSensorSleep incorporates stage-aware data modeling to reflect sensing-dependent variability across wake and sleep states, and adopts a hierarchical alignment strategy to capture invariant structure at multiple temporal scales. Local alignment promotes robustness to sensor- and placement-related variations in short respiratory segments, while global spectral alignment focuses on the dominant dynamical structure of whole-night sleep sequences, emphasizing principal modes of temporal organization and reducing sensitivity to sensor-amplified noise.

We collect RF-based datasets from both home and clinical environments and evaluate XSensorSleep on these datasets without using labeled RF data during training or fine-tuning. The framework is integrated into existing processing pipelines to perform fully automated, non-contact sleep staging.

Our contributions are threefold:

- We identify a fundamental limitation of existing sequence modeling approaches in cross-sensor sleep staging, namely that the assumption of preserved fine-grained temporal correspondence across sensing modalities often does not hold in non-contact RF settings.
- We propose XSensorSleep, a cross-sensor domain generalization framework that leverages stable sleep dynamics beyond sensor-induced temporal distortions through joint modeling of local respiratory representations and global whole-night sleep structure.
- We validate the proposed framework through extensive evaluations on real-world RF datasets collected from both home and clinical environments, where no labeled RF data is used for training or fine-tuning.

2 Related Work

2.1 Non-contact Automatic Sleep Staging

Recent research has explored non-contact sleep staging using radio frequency signals as an alternative to traditional PSG-based approaches. By leveraging the ability of RF signals to capture subtle physiological activities during sleep, these methods aim to enable unobtrusive and automatic sleep staging.

Zhao et al. proposed a Conditional Adversarial Network (CAN) that learns sleep stage representations directly from RF signals, demonstrating improved accuracy even when labeled data is limited [Zhao *et al.*, 2017a]. Kwon et al. introduced a framework based on impulse radio ultra-wideband (IR-UWB) radar, employing an attention-based LSTM model to capture long-range dependencies and salient temporal features in sleep patterns [Kwon *et al.*, 2021]. Zhai et al. developed a method that combines a convolutional recurrent neural network (CRNN) with neural conditional random fields (CRF), effectively modeling both the temporal dynamics and transition structures of sleep stages from RF data [Zhai *et al.*, 2022]. Li et al. designed a system using a CNN-based encoder together with a Transformer sequence model and contrastive learning, enabling accurate sleep staging from UWB signals and jointly performing sleep apnea detection as an auxiliary task [Li *et al.*, 2024]. Most recently, He et al. proposed a deep neural architecture comprising multiple convolutional residual blocks for local feature extraction, followed by an LSTM layer and attention mechanism to model global temporal dependencies in RF-reflected signals [He *et al.*, 2025].

Despite these advances, most existing RF-based sleep staging methods are often limited by small dataset sizes, lack of generalizability across environments and devices, and dependence on substantial labeled RF data, which motivates the need for robust, domain-generalizable approaches.

2.2 Domain Generalization

Domain generalization (DG) aims to train models to perform reliably on unseen domains [Blanchard *et al.*, 2011; Muandet *et al.*, 2013]. A widely adopted strategy for DG is to encourage domain-invariant feature learning by reducing the divergence between feature distributions across domains. Several classic methods have been developed to minimize feature distribution discrepancies across domains, such as maximum mean discrepancy (MMD) [Tzeng *et al.*, 2014], second-order correlation alignment (CORAL) [Sun and Saenko, 2016], and similarity metric-based approaches [Dou *et al.*, 2019].In addition, adversarial frameworks, such as domain adversarial neural networks (DANN) [Li *et al.*, 2018], have become standard techniques in domain generalization research.

In sleep staging, recent efforts like SleepDG [Wang *et al.*, 2024] employ multi-level alignment to bridge gaps between heterogeneous PSG datasets. However, while SleepDG focuses on *intra-modality* shifts, **XSensorSleep** addresses the more challenging *cross-sensor* scenario where physical measurement mechanisms differ fundamentally (e.g., from belts to non-contact RF).

Unlike SleepDG's reliance on noise-sensitive point-to-point alignment, we propose **Spectral Temporal Alignment (STA)**. Recognizing that sleep architecture dynamics remain physiologically invariant despite varying signal morphologies, STA aligns the **eigenvalue spectra** of temporal correlation matrices. This spectral approach captures the universal rhythm of sleep transitions, enabling robust zero-shot generalization to unseen sensing modalities by focusing on intrinsic dynamical complexity rather than device-dependent fluctuations.

3 Method

3.1 Problem Formulation

In this work, we study contactless sleep staging under a cross-sensor domain generalization setting, where models are trained on contact-based respiratory signals (e.g., chest and abdominal belts) and evaluated on RF-based respiratory measurements, which serve as the target domain. This setting is challenging due to substantial domain shifts in signal morphology and temporal structure induced by differences in sensing mechanisms.

We consider M labeled source domains. Each source domain m is associated with a joint distribution $p_m(x, y)$ and a labeled dataset $D_m = \{(x_n^m, y_n^m)\}_{n=1}^{N_m}$. The RF-based target domain follows an unknown distribution $p_{\text{target}}(x, y)$ that differs from all source distributions $p_m(x, y)$.

Each sample (x_n^m, y_n^m) from the m-th domain represents a whole-night respiratory recording annotated with sleep stage labels from the set $\mathcal{Y} = \{1, \ldots, C\}$, corresponding to Wake, Light (N1+N2), Deep (N3), and REM. Specifically,

$$x_n^m = \{x_{n,1}^m, \ldots, x_{n,L}^m\},\ y_n^m = \{y_{n,1}^m, \ldots, y_{n,L}^m\}, \tag{1}$$

where each respiratory epoch $x_{n,\ell}^m \in \mathbb{R}^d$ is associated with a label $y_{n,\ell}^m \in \mathcal{Y}$.

The objective is to learn a predictive function that generalizes from belt-based respiratory recordings to RF-based respiratory signals.

3.2 Framework Overview

As illustrated in Fig. 1, our framework addresses the cross-domain gap between belt-based and RF-derived respiratory signals through a single-stage domain generalization approach. Labeled belt data are partitioned into multiple pseudo-domains based on dataset source, sensor type, and subject pool to expose the model to diverse respiratory patterns. To enhance generalization, we incorporate stage-aware data augmentation and enforce both epoch-level feature alignment and whole-night sequence alignment, encouraging domain-invariant representations. The trained model is then directly applied to RF-derived respiratory signals for sleep stage prediction without accessing the target domain.

3.3 Model Architecture

We adopt a hierarchical architecture $\mathcal{F} = f \circ h \circ g$ to model full-night respiratory sequences, where the epoch encoder g maps each respiratory epoch to a latent representation, the sequence encoder h captures long-range temporal dependencies and sleep-stage transition dynamics across the night, and the classifier f outputs sleep stage probabilities for each epoch.

Specifically, for a whole-night respiratory sequence x_n^m, the epoch encoder g is applied independently to each respiratory epoch:

$$\mathbf{f}_{n,\ell}^m = g(x_{n,\ell}^m),\ \ell = 1, \ldots, L. \tag{2}$$

The resulting sequence of epoch-level embeddings $\mathbf{F}_n^m = \{\mathbf{f}_{n,\ell}^m\}_{\ell=1}^L$ is then processed by the sequence encoder h to model global temporal context:

$$\mathbf{H}_n^m = h(\mathbf{F}_n^m) = \{\mathbf{h}_{n,\ell}^m\}_{\ell=1}^L. \tag{3}$$

Finally, the classifier f predicts sleep stage labels for each epoch based on the temporally enhanced representations:

$$\hat{y}_{n,\ell}^m = f(\mathbf{h}_{n,\ell}^m),\ \ell = 1, \ldots, L. \tag{4}$$

In our implementation, the epoch encoder g is instantiated as a ResNet34 backbone [He *et al.*, 2016] to extract local, epoch-level representations from respiratory waveforms. The sequence encoder h is implemented using a bidirectional Structured State Space Sequence (BiS4) model [Gu *et al.*, 2022], which captures long-range temporal dependencies by incorporating both past and future contextual information. The classifier f is realized as a lightweight multilayer perceptron that outputs sleep stage probabilities for each epoch.

3.4 Stage-Aware Data Augmentation

Directly optimizing the prediction model $\mathcal{F}$ via empirical risk minimization (ERM) on contact-based source domains often leads to suboptimal generalization to RF-based respiratory signals. This is primarily due to substantial sensing-induced domain shifts, where different sensing modalities exhibit stage-dependent response characteristics. In particular, while contact-based respiratory belts predominantly capture respiration-induced chest or abdominal expansion, non-contact radar systems are sensitive to all surface displacements. As a result, during the Wake stage, large-amplitude body movements can dominate radar measurements and overwhelm respiration-related micro-motions, yielding signal patterns that differ qualitatively from belt-based recordings.

Motivated by this sensing asymmetry, we propose Stage-Aware Data Augmentation (SADA), which applies stage-dependent perturbations to better reflect distinct radar sensing regimes during training. For wake epochs, an aggressive augmentation $\mathcal{A}_{\text{wake}}$ introduces high-variance noise, amplitude scaling, and soft temporal masking to simulate motion-dominated radar signals. For sleep epochs, a mild augmentation $\mathcal{A}_{\text{sleep}}$ applies attenuated perturbations to preserve respiration-dominant morphology.

Formally, for an input respiratory epoch $x_{n,\ell}^m$ with label $y_{n,\ell}^m$, the augmented signal $\tilde{x}_{n,\ell}^m$ is defined as

$$\tilde{x}_{n,\ell}^m = \begin{cases} \mathcal{A}_{\text{wake}}(x_{n,\ell}^m; \theta_{\text{wake}}), & y_{n,\ell}^m = \text{Wake}, \\ \mathcal{A}_{\text{sleep}}(x_{n,\ell}^m; \theta_{\text{sleep}}), & \text{otherwise}, \end{cases} \tag{5}$$

where θ_{wake} and θ_{sleep} denote stage-specific augmentation parameters.

3.5 Multi-source Domain Generalization

We partition the N public belt-based datasets into $2N$ pseudo-domains based on dataset origin and sensor placement (chest or abdomen). To prevent subject-level information leakage, chest- and abdomen-belt signals are drawn from disjoint subject pools, and subjects appearing in multiple datasets are excluded.

Epoch-level Feature Alignment (ELFA). To reduce distributional discrepancies in local respiratory patterns across source domains, we introduce an epoch-level alignment loss. Given the sets of epoch-level features $\mathbf{F}_n^p$ and $\mathbf{F}_n^q$ extracted by the epoch encoder g in Eq. 2 from different source domains,

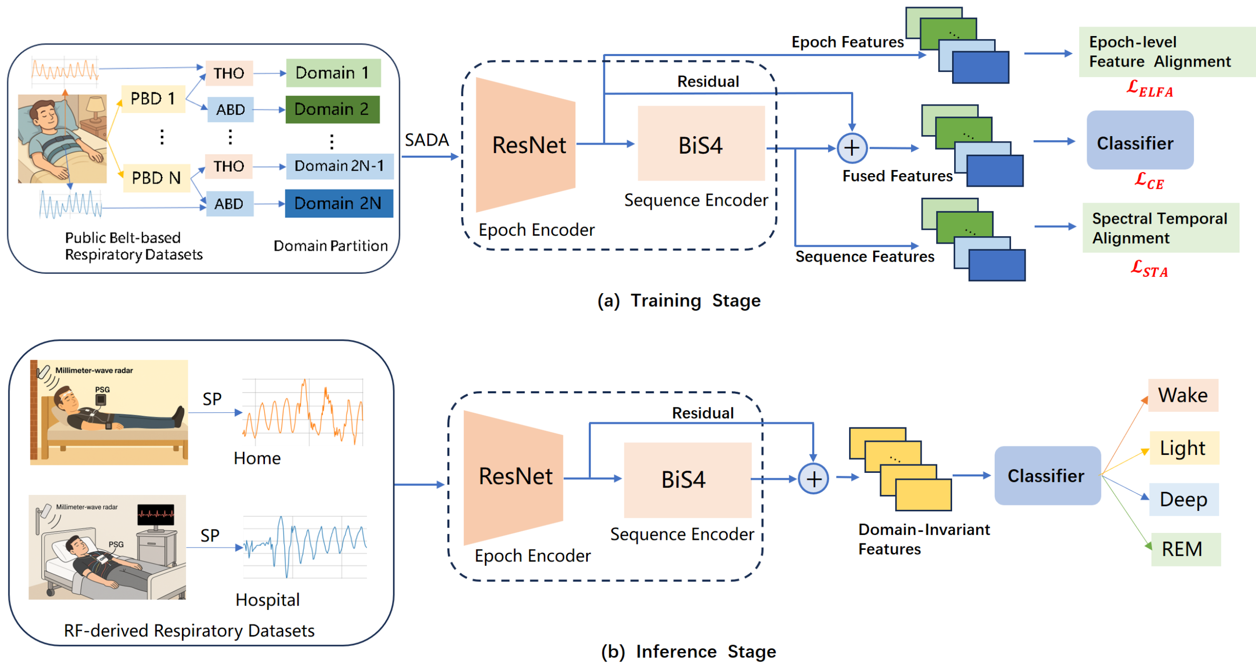

Figure 1: Overview of XSensorSleep. PBD refers to public belt-based respiratory datasets, ABD refers to abdominal belt respiratory signals, THO refers to thoracic (chest) belt respiratory signals, and SP refers to signal processing, SADA represents stage-aware data augmentation.

we align their mean and covariance following CORAL [Sun and Saenko, 2016]:

$$\mathcal{L}_{\text{ELFA}} = \sum_{p<q} \left(\|\boldsymbol{\mu}^p - \boldsymbol{\mu}^q\|_2^2 + \frac{1}{4d^2} \|\boldsymbol{\Sigma}^p - \boldsymbol{\Sigma}^q\|_F^2 \right), \quad (6)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ denote the corresponding mean vector and covariance matrix, and d is the feature dimensionality. By minimizing pairwise statistical discrepancies across pseudo-domains, ELFA encourages the encoder to learn domain-invariant epoch-level representations that suppress sensor- and dataset-specific variations in respiratory morphology.

Spectral Temporal Alignment (STA). While ELFA aligns local epoch-level representations, respiratory-based sleep staging additionally depends on long-range transition dynamics that define the global sleep architecture. However, existing sequence-level alignment methods typically assume direct temporal correspondence across domains, an assumption often violated in cross-sensor settings due to sensor-specific distortions and motion-induced shifts. We therefore propose *Spectral Temporal Alignment* (STA) to align dominant temporal modes of whole-night dynamics.

For a whole-night recording, given the sets of in-context features $\mathbf{H}_n^m$ extracted by the sequence encoder in Eq. 3. We first construct a temporal correlation matrix $\mathbf{R}_n^m \in \mathbb{R}^{L\times L}$ using cosine similarity:

$$(R_n^m)_{i,j} = \frac{\langle \mathbf{h}_{n,i}^m, \mathbf{h}_{n,j}^m \rangle}{\|\mathbf{h}_{n,i}^m\| \cdot \|\mathbf{h}_{n,j}^m\|}, \; i,j \in \{1,\ldots,L\}. \quad (7)$$

Then the domain-level average correlation matrix is computed as:

$$\bar{R}^m = \frac{1}{N_m} \sum_{n=1}^{N_m} R_n^m, \quad (8)$$

where N_m is the number of recordings.

Directly aligning correlation matrices via element-wise distances may impose unnecessarily strict correspondence at the level of individual temporal interactions, which is less desirable under sensor-induced distortions. Instead, STA aligns the eigenvalue spectra of $\bar{R}^m$, which summarize the dominant temporal modes of whole-night dynamics and are invariant to orthogonal transformations. The STA loss is defined as:

$$\mathcal{L}_{\text{STA}} = \sum_{p<q} \|\boldsymbol{\lambda}_K^p - \boldsymbol{\lambda}_K^q\|_2^2 \quad (9)$$

where $\boldsymbol{\lambda}_K$ denotes the top-K eigenvalues of corresponding correlation matrix. Aligning the leading K eigenvalues promotes agreement in large-scale temporal structure across domains, while reducing sensitivity to fine-grained variations that are susceptible to sensor-induced distortions.

3.6 Training and Inference

The total loss function during training is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{ELFA} + \lambda_2 \mathcal{L}_{STA}, \quad (10)$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss for classification, and λ_1, λ_2 are balancing hyperparameters.

The model is trained exclusively on the multi-source, labeled belt-based respiratory data using the above objective. At test time, $\mathcal{F} = f \circ h \circ g$ is directly deployed on RF-derived respiratory signals, achieving zero-shot generalization to unseen target domains.

Dataset	Modality	Total	Used	Usage
Source Domain (Contact-based Sensors)				
SHHS1	Chest/Abdo Belt	5793	3083	Train / Val.
SHHS2	Chest/Abdo Belt	2651	2554	Train / Val.
Target Domain (RF-based Sensors)				
Home	FMCW Radar	35	35	Test
Hospital	FMCW Radar	57	57	Test

Table 1: Datasets used in our experiments. Belt-based datasets serve as source domains for training and validation, while radar-based datasets are used exclusively for evaluation.

4 Experiments

4.1 Datasets and Preprocessing

Datasets

We used two datasets: (1) multi-source belt-based respiratory signals from public sleep databases, and (2) a self-collected mmWave radar respiratory dataset (Table 1).

Belt-based Respiratory Datasets. We used the Sleep Heart Health Study (SHHS), including SHHS1 and SHHS2 cohorts [Zhang *et al.*, 2018; Quan *et al.*, 1997]. SHHS1 contains over 5,000 subjects, while SHHS2 is a subset re-collected approximately three years later. To avoid subject leakage in domain generalization experiments, overlapping subjects between cohorts were removed. Each overnight recording provides synchronized chest and abdominal belt signals annotated in 30-second epochs following AASM standards.

Radar-based Respiratory Datasets. Our target domain is a self-collected mmWave radar dataset with 92 overnight recordings from 71 subjects in hospital and home environments. Specifically, 57 recordings were collected from 57 subjects in hospital settings and 35 from 14 subjects at home. Data acquisition used 77GHZ and 60GHZ FMCW radar systems positioned obliquely at the bedside. All procedures were IRB approved.

Preprocessing

Pseudo-domain Partitioning. For both SHHS1 and SHHS2, subjects were randomly divided into two disjoint groups of comparable size, assigned to chest-belt and abdomen-belt domains, respectively. Each subject appears in only one domain, with no overlap within or across datasets. This results in four balanced pseudo-domains (SHHS1-chest, SHHS1-abdomen, SHHS2-chest, SHHS2-abdomen) with unique subjects.

Signal Processing. Radar signals were beamformed to extract respiratory waveforms by selecting points with the maximum Breath-to-Noise Ratio (BNR) . This dynamic selection enables physical adaptation to different sleeping postures (e.g., supine, side, or stomach) by focusing on the body region with the strongest respiratory modulation. All belt-based and radar-derived signals were resampled to 10 Hz and bandpass filtered (0.1–2 Hz) to preserve respiratory components and suppress artifacts. Each epoch was normalized using min-max scaling. Non-overlapping 30-second epochs were then concatenated to form full-night time-series inputs for the model.

4.2 Experiment Settings

XSensorSleep is implemented in PyTorch and optimized using AdamW with a cosine learning rate scheduler, an initial learning rate of 4×10^{-4}, and a weight decay of 1×10^{-4}. The model is trained for 50 epochs with a batch size of 6, a dropout rate of 0.1, and regularization coefficients $\lambda_1 = \lambda_2 = 0.1$. Each overnight respiratory recording is represented as a fixed-length sequence of 1500 epochs, corresponding to 12.5 hours of data. Shorter recordings are zero-padded with masks, while longer ones are truncated, with padded epochs excluded during temporal correlation computation.

The feature dimension is set to 512, and the number of leading eigenvalues for whole-night structure alignment is $K = 20$. Model performance is evaluated using Accuracy (ACC), Macro-F1 (MF1), and Cohen's Kappa (κ).

4.3 Experiment Results

Comparison with Existing Non-Contact Sleep Staging Methods Using RF Signals

We compare XSensorSleep with representative RF-based sleep staging methods spanning domain generalization, temporal modeling, and modern deep architectures, including Zhao et al. (2017) [Zhao *et al.*, 2017a], Kwon et al. (2021) [Kwon *et al.*, 2021], Zhai et al. (2022) [Zhai *et al.*, 2022], He et al. (2025) [He *et al.*, 2025], and Zhuang et al. (2025) [Zhuang *et al.*, 2025], where the first four models were reproduced and trained on the same large-scale multi-source belt datasets as XSensorSleep and evaluated on the RF-derived test sets, and the fifth method was evaluated using its publicly available pre-trained weights.

Results in Table 2 show that XSensorSleep consistently outperforms all prior RF-based methods in accuracy, macro-F1, and Cohen's Kappa across both home and hospital settings, demonstrating strong generalization enabled by multi-source partitioning and whole-night alignment. Performance on the hospital dataset is approximately 5 points lower than in the home setting, primarily due to increased RF noise from higher personnel movement and a greater prevalence of severe sleep apnea, consistent with prior studies [Bianchi *et al.*, 2010; Korkalainen *et al.*, 2019].

Comparison with Existing DG Methods

We further evaluate XSensorSleep against state-of-the-art domain generalization methods, including ERM, MMD [Tzeng *et al.*, 2014], CORAL [Sun and Saenko, 2016], DANN [Ganin and Lempitsky, 2015], IRM [Arjovsky *et al.*, 2019], and the sleep-specific SleepDG [Wang *et al.*, 2024]. All methods use the same backbone and are evaluated under identical experimental settings.

As shown in Table 3, our method consistently outperforms all baselines on the target domain across all metrics. The ERM baseline exhibits a notable performance drop, particularly in the Hospital environment, suggesting a pronounced modality gap between contact-based belt signals and non-contact radar measurements, where direct model transfer may be insufficient. General DG methods (e.g., CORAL and DANN) yield only marginal improvements over this baseline, possibly because sample-level distribution alignment

Method	Home			Hospital			Overall Performance		
	ACC	MF1	κ	ACC	MF1	κ	ACC	MF1	κ
RF-Specific Baselines (Trained on Belt Data)									
Zhao et al. (2017)	79.35	78.99	0.708	70.83	65.35	0.516	74.12	71.56	0.607
Kwon et al. (2021)	70.93	65.26	0.565	63.80	55.92	0.378	66.56	60.31	0.461
Zhai et al. (2022)	72.19	67.64	0.586	65.24	60.51	0.424	67.91	63.18	0.588
He et al. (2025)	76.41	71.56	0.655	68.84	64.41	0.496	71.75	67.23	0.588
Large-scale Foundation Model (Zero-shot Transfer)									
Zhuang et al. (2025)	80.70	79.49	0.723	73.21	67.63	0.568	76.10	73.12	0.625
XSensorSleep (Ours)	**81.67**	**80.43**	**0.735**	**76.30**	**70.63**	**0.592**	**78.38**	**75.23**	**0.657**

Table 2: Performance comparison on Home and Hospital RF datasets. We categorize baselines into RF-specific models (re-trained) and pre-trained foundation models (directly evaluated). The best results are highlighted in **bold**.

Method	Home (RF)			Hospital (RF)			Overall (RF)			Validation (Belt)			Perf. Gap (Δ)		
	ACC	MF1	κ	ACC	MF1	κ	ACC	MF1	κ	ACC	MF1	κ	ACC↓	MF1↓	κ↓
ERM	78.76	76.88	0.683	73.62	67.18	0.552	75.61	71.54	0.611	79.77	76.40	0.702	4.16	4.86	0.091
MMD	79.04	77.67	0.695	74.21	68.23	0.563	76.08	72.66	0.625	78.63	74.95	0.685	2.55	2.29	0.060
CORAL	79.41	78.07	0.701	75.15	69.10	0.573	76.81	73.81	0.631	79.46	75.64	0.696	2.65	1.83	0.065
DANN	80.21	79.02	0.702	74.29	68.17	0.571	76.54	73.33	0.632	79.45	75.11	0.692	2.91	1.78	0.060
IRM	80.03	79.23	0.709	74.75	68.28	0.559	76.79	73.56	0.629	79.50	75.24	0.694	2.71	1.68	0.065
SleepDG	80.56	79.30	0.713	74.64	69.07	0.564	76.93	74.17	0.636	80.50	77.20	0.713	3.57	3.03	0.077
XSensorSleep	**81.67**	**80.43**	**0.735**	**76.30**	**70.63**	**0.592**	**78.38**	**75.23**	**0.657**	80.01	75.77	0.703	**1.63**	**0.54**	**0.046**

Table 3: Detailed performance across testing domains (Home and Hospital RF), overall performance, and validation stage (Belt). Perf. Gap (Δ) denotes the absolute drop between Validation and Overall Testing. Best results in **bold**.

Variant	Modules			Home			Hospital			Overall Performance		
	SADA	ELFA	STA	ACC	MF1	κ	ACC	MF1	κ	ACC	MF1	κ
ERM				78.76	76.88	0.683	73.62	67.18	0.552	75.61	71.54	0.611
(a)	✓			80.02	78.53	0.710	75.59	68.39	0.566	77.31 (+1.70)	73.11 (+1.57)	0.637 (+0.026)
(b)		✓		79.84	78.81	0.707	74.62	68.75	0.563	76.64 (+1.03)	73.46 (+1.92)	0.628 (+0.017)
(c)			✓	80.26	79.42	0.705	74.86	68.46	0.570	76.95(+1.34)	73.72 (+2.21)	0.633 (+0.022)
(d)	✓	✓		80.80	79.33	0.719	75.44	68.86	0.576	77.52 (+1.91)	73.81 (+2.27)	0.645 (+0.034)
(e)	✓		✓	80.94	79.45	0.723	76.07	69.45	0.591	77.97 (+2.36)	74.63 (+3.09)	0.651 (+0.040)
(f)	✓	✓	✓	**81.67**	**80.43**	**0.735**	**76.30**	**70.63**	**0.592**	**78.38 (+2.77)**	**75.23 (+3.69)**	**0.657 (+0.046)**

Table 4: Ablation study of different components on Home and Hospital datasets. ✓ indicates the inclusion of a module. **Source-only** denotes the model trained on source belt data and directly tested on radar data without alignment. All metrics are in percentage (%) except Cohen's κ (decimal). Values in parentheses denote improvements over Source-only.

does not fully capture the periodic morphology of respiratory signals. In contrast, XSensorSleep achieves an overall κ improvement of 0.021 over SleepDG, with the largest gain of 0.028 observed in the Hospital dataset, suggesting that STA contributes to improved robustness under clinical noise and pathological breathing conditions. Notably, XSensorSleep also exhibits the smallest performance drop from in-domain belt validation to cross-domain radar testing, indicating improved robustness to modality shift.

4.4 Ablation Experiment

To assess the contribution of each component, we conduct ablation studies summarized in Table 4, where (f) denotes the full model and the ERM model serves as a lower-bound baseline.

Individual Impact of Modules. As shown in Table 4, each individual module consistently improves performance over the ERM baseline. SADA (a) introduces stage-aware perturbations that enhance robustness to varying radar signal-to-noise ratios, resulting in a 0.026 increase in overall κ. ELFA (b) performs epoch-level alignment to mitigate morphological distortions arising from the transition from mechanical pressure sensors to RF-based sensing. Notably, STA (c) independently yields a 2.21% improvement in Macro-F1, suggesting that preserving the global "physiological skeleton" provides an effective constraint for cross-modal transfer.

Synergistic Effects of Combined Modules. Combining modules leads to further performance gains, highlighting

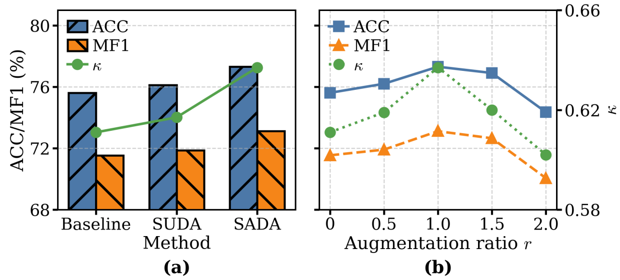

Figure 2: Evaluation of Stage-Aware Data Augmentation. (a) Comparison between different augmentation methods. (b) Sensitivity analysis of the augmentation ratio in SADA.

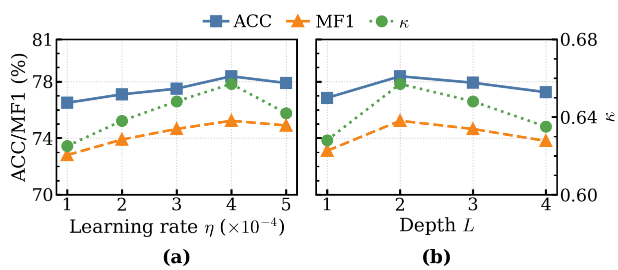

Figure 3: Ablation studies on training hyperparameters. (a) Sensitivity of model performance to the learning rate. (b) Performance comparison of BiS4 model with different hierarchical depths.

their complementary roles. The combination of SADA and ELFA (d) achieves an overall κ of 0.645, suggesting that data-level augmentation from SADA enriches the feature space and facilitates more robust local invariant alignment by ELFA between the belt and radar domains. The SADA and STA combination (e) further improves κ to 0.651, where SADA stabilizes epoch-level representations and enables STA to more accurately capture global spectral-temporal transition dynamics across a full night without being dominated by noisy epochs.

4.5 Analysis of Stage-Aware Data Augmentation

We compare our proposed SADA with two settings: (1) *Baseline* (no augmentation) and (2) *Stage-Unaware Data Augmentation (SUDA)*, which applies uniform transformations across all sleep stages. As illustrated in Fig. 2(a), SADA consistently outperforms both counterparts. This confirms that tailoring augmentation to stage-specific radar motion sensitivity effectively enhances feature representation. Furthermore, Fig. 2(b) investigates the sensitivity to augmentation ratios, revealing that doubling the training data (ratio = 1.0) achieves the optimal balance between data diversity and performance.

4.6 Hyperparameter Analysis

We analyze the sensitivity of XSensorSleep to the learning rate η and the number of BiS4 layers L_{s4}. As shown in Fig. 3(a), varying $\eta \in \{1 \times 10^{-4}, 2 \times 10^{-4}, 3 \times 10^{-4}, 4 \times 10^{-4}, 5 \times 10^{-4}\}$ reveals that performance peaks at $\eta = 4 \times 10^{-4}$, with smaller values leading to insufficient convergence and larger ones causing training instability. Fig. 3(b) further shows that increasing L_{s4} from 1 to 2 improves MF1, while performance saturates beyond $L_{s4} = 2$ and slightly degrades at larger depths due to overfitting. We therefore set $\eta = 4 \times 10^{-4}$ and $L_{s4} = 2$.

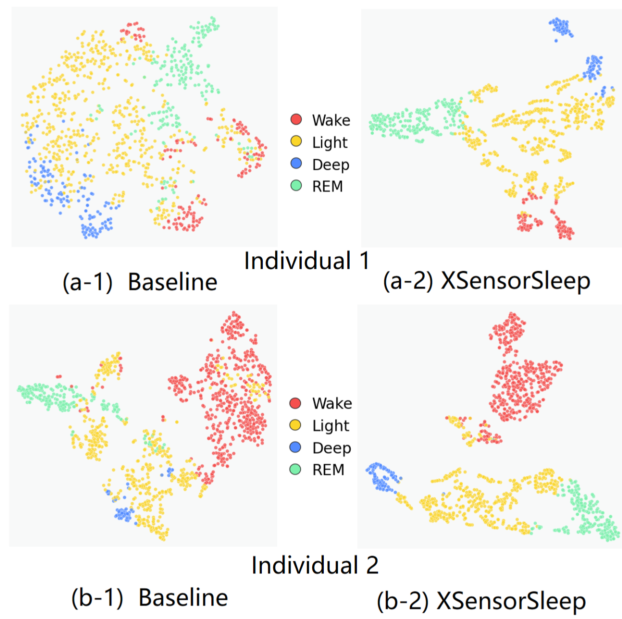

Figure 4: t-SNE visualization of sleep features. Individual 1 is from the Home dataset, and Individual 2 is from the Hospital dataset.

4.7 Feature Visualization

To further evaluate feature quality, we visualize intermediate fused representations using t-SNE [Maaten and Hinton, 2008] for two individuals from distinct environments: one from the Home setting and one from the Hospital setting. Fig. 4(a-1) and (a-2) show the feature distributions for the Home individual under the Baseline and XSensorSleep, respectively, while Fig. 4(b-1) and (b-2) present the corresponding results for the Hospital individual. Compared to the Baseline, XSensorSleep yields more compact clusters for the same sleep stage in both environments, indicating improved feature separability and more consistent representations across domains.

5 Conclusion

We propose XSensorSleep, which formulates RF-based contactless sleep staging as a cross-sensor domain generalization problem and enables zero-shot transfer from large-scale belt-based respiratory datasets to unseen radar signals. The framework combines multi-source pseudo-domain partitioning with stage-aware data augmentation and hierarchical alignment at both the epoch and whole-night levels, explicitly addressing sensing-induced distortions and temporal mismatch across modalities. Experiments on home and hospital RF datasets show that these design choices consistently improve feature separability and cross-sensor robustness compared to existing methods.

Acknowledgments

This work was supported by Space Medical Experiment Project of CMSP under grant HYZHXMX01005.

References

[Arjovsky *et al.*, 2019] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[Berry *et al.*, 2012] Richard B Berry, Rita Brooks, Charlene E Gamaldo, Susan M Harding, Carole Marcus, Bradley V Vaughn, et al. The aasm manual for the scoring of sleep and associated events. *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine*, 176(2012):7, 2012.

[Bianchi *et al.*, 2010] Matt T Bianchi, Sydney S Cash, Joseph Mietus, Chung-Kang Peng, and Robert Thomas. Obstructive sleep apnea alters sleep stage transition dynamics. *PloS one*, 5(6):e11356, 2010.

[Blanchard *et al.*, 2011] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24, 2011.

[Byun *et al.*, 2019] Jong-Ho Byun, Keun Tae Kim, Hye-jin Moon, Gholam K Motamedi, and Yong Won Cho. The first night effect during polysomnography, and patients' estimates of sleep quality. *Psychiatry research*, 274:27–29, 2019.

[Carley and Farabi, 2016] David W Carley and Sarah S Farabi. Physiology of sleep. *Diabetes spectrum: a publication of the American Diabetes Association*, 29(1):5, 2016.

[Ding *et al.*, 2022] Lei Ding, Baixin Chen, Yanyuan Dai, and Yun Li. A meta-analysis of the first-night effect in healthy individuals for the full age spectrum. *Sleep Medicine*, 89:159–165, 2022.

[Dong *et al.*, 2024] Shuqin Dong, Li Wen, Yuchen Li, Jingyun Lu, Zhi Zhang, Chengmei Yuan, and Changzhan Gu. Remote respiratory variables tracking with biomedical radar-based iot system during sleep. *IEEE Internet of Things Journal*, 11(11):19937–19948, 2024.

[Dou *et al.*, 2019] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. *Advances in neural information processing systems*, 32, 2019.

[Duan *et al.*, 2025] Xiang Duan, Kunpeng Li, Xuan Liu, Yang Li, Yuqing Hou, Xiaowei He, and Sheng Tang. A non-contact sleep monitoring method based on millimeter-wave radar. *IEEE Transactions on Instrumentation and Measurement*, 2025.

[Ganin and Lempitsky, 2015] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[Gu *et al.*, 2022] Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

[Hao *et al.*, 2024] Zhanjun Hao, Yue Wang, Fenfang Li, Guozhen Ding, and Yifei Gao. Mmwave-rm: A respiration monitoring and pattern classification system based on mmwave radar. *Sensors (Basel, Switzerland)*, 24(13):4315, 2024.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[He *et al.*, 2025] Hao He, Chao Li, Wolfgang Ganglberger, Kaileigh Gallagher, Rumen Hristov, Michail Ouroutzoglou, Haoqi Sun, Jimeng Sun, M Brandon Westover, and Dina Katabi. What radio waves tell us about sleep! *Sleep*, 48(1):zsae187, 2025.

[Jin *et al.*, 2020] Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Feature Alignment and Restoration for Domain Generalization and Adaptation, 2020.

[Korkalainen *et al.*, 2019] Henri Korkalainen, Juhani Aakko, Sami Nikkonen, Samu Kainulainen, Akseli Leino, Brett Duce, Isaac O Afara, Sami Myllymaa, Juha Töyräs, and Timo Leppänen. Accurate deep learning-based sleep staging in a clinical population with suspected obstructive sleep apnea. *IEEE journal of biomedical and health informatics*, 24(7):2073–2081, 2019.

[Kwon *et al.*, 2021] Hyun Bin Kwon, Sang Ho Choi, Dongseok Lee, Dongyeon Son, Heenam Yoon, Mi Hyun Lee, Yu Jin Lee, and Kwang Suk Park. Attention-based lstm for non-contact sleep stage classification using ir-uwb radar. *IEEE Journal of Biomedical and Health Informatics*, 25(10):3844–3853, 2021.

[Lee *et al.*, 2025] Hyojin Lee, You Rim Choi, Hyun Kyung Lee, Jaemin Jeong, Joopyo Hong, Hyun-Woo Shin, and Hyung-Sin Kim. Explainable vision transformer for automatic visual sleep staging on multimodal psg signals. *npj Digital Medicine*, 8(1):55, 2025.

[Li *et al.*, 2018] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 624–639, 2018.

[Li *et al.*, 2024] Siheng Li, Beihong Jin, Fusang Zhang, Zhi Wang, Junqi Ma, Xiaoyong Ren, and Haiqin Liu. Hypnos: A contactless sleep stage monitoring system using uwb signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(3):1–27, 2024.

[Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[Muandet *et al.*, 2013] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International conference on machine learning*, pages 10–18. PMLR, 2013.

[Quan *et al.*, 1997] Stuart F Quan, Barbara V Howard, Conrad Iber, James P Kiley, F Javier Nieto, George T O'Connor, David M Rapoport, Susan Redline, John Robbins, Jonathan M Samet, et al. The sleep heart health study: design, rationale, and methods. *Sleep*, 20(12):1077–1085, 1997.

[Rundo and Downey III, 2019] Jessica Vensel Rundo and Ralph Downey III. Polysomnography. *Handbook of clinical neurology*, 160:381–392, 2019.

[Sun and Saenko, 2016] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer vision–ECCV 2016 workshops: Amsterdam, the Netherlands, October 8-10 and 15-16, 2016, proceedings, part III 14*, pages 443–450. Springer, 2016.

[Tamaki *et al.*, 2005] Masako Tamaki, Hiroshi Nittono, Mitsuo Hayashi, and Tadao Hori. Examination of the first-night effect during the sleep-onset period. *Sleep*, 28(2):195–202, 2005.

[Thapa *et al.*, 2024] Rahul Thapa, Bryan He, Magnus Ruud Kjaer, Hyatt Moore IV, Gauri Ganjoo, Emmanuel Mignot, and James Y Zou. Sleepfm: Multi-modal representation learning for sleep across ecg, eeg and respiratory signals. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024.

[Tzeng *et al.*, 2014] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[Wang *et al.*, 2022] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 35:1–1, 2022.

[Wang *et al.*, 2024] Jiquan Wang, Sha Zhao, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Generalizable sleep staging via multi-level domain alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 265–273, 2024.

[Wang *et al.*, 2025] Wei Wang, Zhaoxi Chen, Wenyu Zhang, Zetao Wang, Xiang Zhao, Chenyang Li, Jian Guan, Shankai Yin, and Gang Li. Fusion of millimeter-wave radar and pulse oximeter data for low-burden diagnosis of obstructive sleep apnea-hypopnea syndrome. *arXiv preprint arXiv:2501.15264*, 2025.

[Yu *et al.*, 2021] Bohan Yu, Yuxiang Wang, Kai Niu, Youwei Zeng, Tao Gu, Leye Wang, Cuntai Guan, and Daqing Zhang. Wifi-sleep: Sleep stage monitoring using commodity wi-fi devices. *IEEE internet of things journal*, 8(18):13900–13913, 2021.

[Yue *et al.*, 2018] Shichao Yue, Hao He, Hao Wang, Hariharan Rahul, and Dina Katabi. Extracting multi-person respiration from entangled rf signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(2):1–22, 2018.

[Zhai *et al.*, 2022] Qian Zhai, Tingyu Tang, Xiaoling Lu, Xiaoxi Zhou, Chunguang Li, Jingang Yi, and Tao Liu. Machine learning-enabled noncontact sleep structure prediction. *Advanced Intelligent Systems*, 4(5):2100227, 2022.

[Zhang *et al.*, 2018] Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The national sleep research resource: towards a sleep data commons. *Journal of the American Medical Informatics Association*, 25(10):1351–1358, 2018.

[Zhang *et al.*, 2019] Feng Zhang, Chenshu Wu, Beibei Wang, Min Wu, Daniel Bugos, Hangfang Zhang, and KJ Ray Liu. Smars: Sleep monitoring via ambient radio signals. *IEEE Transactions on Mobile Computing*, 20(1):217–231, 2019.

[Zhang *et al.*, 2020] Dongheng Zhang, Yang Hu, and Yan Chen. Mtrack: Tracking multiperson moving trajectories and vital signs with radio signals. *IEEE Internet of Things Journal*, 8(5):3904–3914, 2020.

[Zhao *et al.*, 2017a] Mingmin Zhao, Shichao Yue, Dina Katabi, Tommi S Jaakkola, and Matt T Bianchi. Learning sleep stages from radio signals: A conditional adversarial architecture. In *International conference on machine learning*, pages 4100–4109. PMLR, 2017.

[Zhao *et al.*, 2017b] Mingmin Zhao, Shichao Yue, Dina Katabi, Tommi S Jaakkola, and Matt T Bianchi. Learning sleep stages from radio signals: A conditional adversarial architecture. In *Proceedings of the 34th International Conference on Machine Learning*, pages 4100–4109. PMLR, 2017.

[Zhou *et al.*, 2023] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2023.

[Zhuang *et al.*, 2025] Zhongxu Zhuang, Biao Xue, Qiang An, Hui Chu, Yue Zhang, Rui Chen, Jing Xu, Ning Ding, Xiaochuan Cui, E Wang, et al. Advancing sleep health equity through deep learning on large-scale nocturnal respiratory signals. *Nature Communications*, 16(1):9334, 2025.