AdaSleep: A Robust Source-Free Domain Adaptation Framework for Unobtrusive Sleep Staging via Belt and Radar

XUAN LUO, University of Science and Technology of China, China
ZHI LU, University of Science and Technology of China, China
FANG ZHOU, University of Science and Technology of China, China
YU PU, University of Science and Technology of China, China
DONGHENG ZHANG, University of Science and Technology of China, China
YAN CHEN*, University of Science and Technology of China, China

Unobtrusive sleep staging based on respiratory signals enables long-term sleep monitoring, yet real-world deployment is often hindered by domain shifts caused by demographic heterogeneity, environmental noise, and differences in sensing modality. Privacy constraints can make source data inaccessible, motivating Source-Free Domain Adaptation (SFDA), which uses only a source-pretrained model and unlabeled target data. However, existing self-training based SFDA methods are vulnerable to confirmation bias: signal artifacts can induce erroneous pseudo-labels that are subsequently reinforced during adaptation. To address these challenges, we propose AdaSleep, a robust framework for adapting a source-pretrained model using unlabeled target data. AdaSleep introduces a consistency-based self-training strategy built on a Teacher-Student architecture. By aggregating predictions across stochastically augmented views, AdaSleep reduces sensitivity to random perturbations and produces more stable pseudo-labels. It further applies a Quality Gate to restrict model updates with reliable pseudo-labels, thereby reducing error propagation. Extensive evaluations across four real-world datasets spanning abdominal belt and radar sensing in hospital and home environments show that AdaSleep outperforms the compared SFDA baselines, with average improvements of 3.67% in accuracy and 3.90% in Macro-F1 over the source-only baseline. These results support AdaSleep as a reliable adaptation strategy for privacy-sensitive unobtrusive sleep monitoring.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

Additional Key Words and Phrases: Sleep staging, contactless sensing, source-free domain adaptation

ACM Reference Format:
Xuan Luo, Zhi Lu, Fang Zhou, Yu Pu, Dongheng Zhang, and Yan Chen. 2026. AdaSleep: A Robust Source-Free Domain Adaptation Framework for Unobtrusive Sleep Staging via Belt and Radar. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 10, 3, Article 136 (September 2026), 28 pages. https://doi.org/10.1145/3832040

1 INTRODUCTION

Sleep staging plays a critical role in diagnosing sleep disorders and assessing sleep quality. Polysomnography (PSG) remains the clinical gold standard, with sleep stages scored according to the guidelines of the American Academy

*Corresponding author

This work was supported by Space Medical Experiment Project of CMSP under grant HYZHXMX01005.
Authors' Contact Information: Xuan Luo, University of Science and Technology of China, Hefei, China, lxxyz@mail.ustc.edu.cn; Zhi Lu, University of Science and Technology of China, Hefei, China, zhilu@ustc.edu.cn; Fang Zhou, University of Science and Technology of China, Hefei, China, fangzhou1958@mail.ustc.edu.cn; Yu Pu, University of Science and Technology of China, Hefei, China, puyu@mail.ustc.edu.cn; Dongheng Zhang, University of Science and Technology of China, Hefei, China, dongheng@ustc.edu.cn; Yan Chen (corresponding author), University of Science and Technology of China, Hefei, China, eecyan@ustc.edu.cn.

ACM 2474-9567/2026/9-ART136
https://doi.org/10.1145/3832040

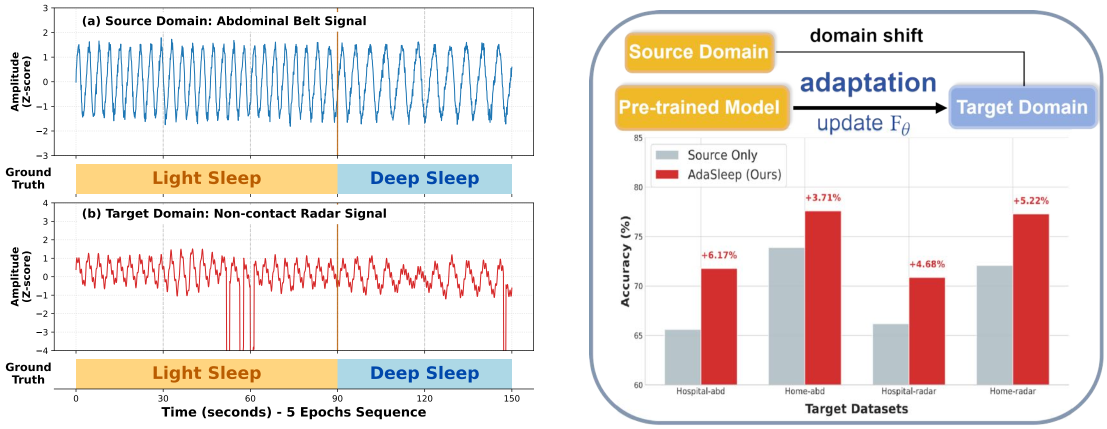
of Sleep Medicine (AASM) [3, 23]. However, PSG requires multiple on-body sensors, specialized equipment, and expert annotation, which limits its practicality for long-term unobtrusive monitoring outside controlled sleep-laboratory settings [3, 10]. To enable more convenient sleep assessment, recent studies have explored respiratory-signal-based sleep staging using both contact-based respiratory belts and contactless radio-frequency (RF) sensing systems [10, 15, 36, 40, 43, 46]. Recent evidence shows that useful sleep-stage information can be extracted from both belt-derived and RF-derived respiratory signals, making respiration a promising modality for continuous and unobtrusive sleep monitoring [10, 46].

However, applying these models in real-world deployment remains difficult because of domain shift [31, 45]. Models pretrained on a source dataset often struggle to generalize to an unseen target dataset when the underlying data distributions differ. In sleep staging with respiratory signals, this shift can arise from inter-subject physiological variability, changes in recording environments, and differences in sensing modality. In contrast to PSG scoring, which directly exploits multiple electrophysiological channels defined by the AASM standard, respiratory-based staging must infer sleep states from subtler surrogate patterns, making it more sensitive to distortion and noise [3, 10]. The challenge becomes even greater when transferring from contact-based abdominal belts to contactless radar systems. Radar-based vital-sign sensing is particularly susceptible to multipath/clutter, non-stationary interference, and motion-induced waveform corruption [1, 8, 41]. Fig. 1a illustrates this signal-level discrepancy between belt-derived and radar-derived respiration. Such sensing- and environment-induced distortions substantially reduce the transferability of models trained in one domain, limiting reliable deployment across diverse populations and environments.

Current approaches, such as fine-tuning models with labeled target data or aligning features between modalities, often rely on the availability of target labels, access to source data, or data from multiple source domains. These requirements are impractical in real-world scenarios where target annotations are scarce or unavailable and privacy concerns often restrict access to source data. Furthermore, in cross-sensor scenarios, the differences between modalities are not limited to variations in data distributions, but also involve fundamental disparities in how each sensor captures and processes respiratory signals. This makes feature alignment particularly challenging. As a result, methods that require extensive data, complex alignment, or source-domain access are difficult to apply in practice, leaving only pretrained models and unlabeled target data available for adaptation.

To overcome these limitations, we adopt the Source-Free Domain Adaptation (SFDA) paradigm [13, 16, 17], which adapts a pretrained model to the target domain without requiring source data or labeled target samples. SFDA methods commonly rely on self-training or related entropy minimization adaptation objectives to refine the model on target data [11, 30]. However, applying this strategy to sleep staging is complicated by the nature of respiratory signals, which are highly susceptible to non-stationary noise, multipath/clutter, and motion artifacts [1, 41, 43]. These issues can lead to overconfident but incorrect pseudo-labels, causing confirmation bias and error propagation during adaptation [2, 11].

To address this challenge, we propose AdaSleep, a robust SFDA framework designed for sleep staging using respiratory signals from abdominal belt and radar sensing modalities. To mitigate confirmation bias in self-training [2], AdaSleep goes beyond simple confidence measures by introducing **consistency-based pseudo-labeling** combined with a rigorous **Quality Gate**. This design is related to consistency regularization and confidence-based pseudo-label selection [26, 28], but instantiates them through Teacher-side multi-view aggregation without adopting a weak-to-strong student augmentation protocol. Robust pseudo-labels are generated by aggregating Teacher predictions across multiple augmented views of the respiratory signal, in line with augmentation-based consistency principles [42], using consistency to filter random noise. The Quality Gate ensures that only pseudo-labels exhibiting high consistency across these views are selected, thereby updating the model with reliable supervision. This approach helps reduce error accumulation during adaptation, enabling the model to learn from high-quality and consistent signals. As shown in Fig. 1, AdaSleep improves performance compared with the source-only baseline, indicating its effectiveness in real-world settings.



(a) Signal-Level Discrepancy: Contrast between source belt and target radar signals.

(b) Performance Gain: AdaSleep boosts accuracy.

Fig. 1. **The Challenge and Our Effective Solution.** (a) The signal-level discrepancy between belt-derived and radar-derived respiration. Unlike the rhythmic patterns of the source domain (abdominal belt), the target radar-derived respiratory waveform is more vulnerable to non-linear distortions and environmental artifacts. (b) AdaSleep (red) robustly adapts the model, achieving substantial accuracy improvements over the source-only baseline (grey) across all transfer scenarios, despite the challenge.

We evaluate AdaSleep on real-world datasets collected from diverse clinical and home environments. To assess the robustness of the model, we categorize experiments into two scenarios that represent varying levels of difficulty: **Intra-Modality Adaptation** where the domain shift primarily arises from differences in subject demographics and environmental noise, and **Cross-Modality Adaptation** which introduces additional challenges due to differences in sensor modalities. The key contributions of our work are outlined as follows:

- We propose AdaSleep, a robust Source-Free Domain Adaptation framework for unobtrusive sleep staging. It targets the challenging cross-modality shift, relying solely on the pretrained model and unlabeled target data without accessing source data.
- We introduce a consistency-based self-training strategy integrated with a rigorous Quality Gate within a Teacher-Student architecture. By aggregating predictions across augmented views of the respiratory signal, the method filters unreliable pseudo-labels and reduces error propagation during adaptation.
- We conduct extensive evaluations across four real-world datasets, including both intra-modality and cross-modality scenarios, using three distinct backbone architectures. The results demonstrate that AdaSleep consistently outperforms the compared source-free domain adaptation methods, with an average accuracy improvement of **3.67%** and a Macro-F1 boost of **3.90%**.

2 PROBLEM FORMULATION

2.1 The Domain Shift: Physiology, Modality, and Environment

Standard deep learning models rely on the assumption that train and test data are identically distributed. However, a model is often pretrained on a large source domain, and then deployed to a different target domain in the real world. The distribution gap between the source and target domain leads to substantial performance degradation.

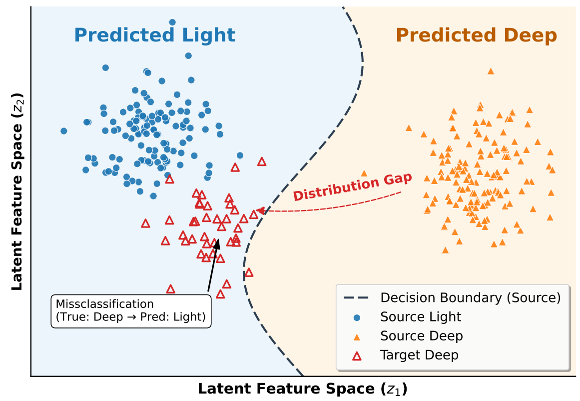


Fig. 2. **Feature-Level Misalignment.** The t-SNE visualization reveals a severe distribution shift. Target Deep sleep samples (red triangles) drift across the source-learned decision boundary (dashed line) into the Source Light region (blue area). This misalignment causes the pretrained model to erroneously classify Deep sleep as Light sleep, underscoring the necessity for domain adaptation.

In this work, a model is pretrained on a large-scale source dataset, abdominal belt recordings from SHHS1, and then deployed to a different target domain. Sleep Heart Health Study (SHHS) is a multi-center cohort study [25]. This target domain may involve abdominal belt data or radar data collected in a different environment. We examine the discrepancy at two connected levels: the respiratory waveform itself and the learned feature representation. We use the transfer from abdominal belt data to radar data as a representative cross-modality example.

Signal-Level Discrepancy. As shown in Fig. 1a, the source and target respiratory waveforms can differ substantially at the signal level. The source signal is an abdominal belt recording from SHHS1, whereas the target signal is a radar recording from the Home-radar domain. In the cross-modality setting, abdominal belt signals directly reflect abdominal motion and usually exhibit relatively stable respiratory rhythms. Meanwhile, radar sensing captures respiratory micromotion through phase modulation, which makes the observed signal more sensitive to attenuation and environmental interference. As a result, radar waveforms are more easily affected by distortions and irregular fluctuations that do not correspond to true physiological activity. These perturbations can alter the observed signal pattern and may mislead a pretrained model. For example, the model may incorrectly interpret sensor artifacts as rapid or irregular breathing, which can lead to confusion with REM or Wake stages. In intra-modality scenarios, such as abd-to-abd adaptation, the sensing modality remains unchanged. There are still differences in subject demographics and background noise across collection environments, causing noticeable signal-level discrepancies.

Feature-Level Misalignment. The signal-level discrepancy further propagates into the latent feature space, so the problem occurs in both cross-modality and intra-modality scenarios. We visualize the high-level features of a model pretrained on SHHS1 and then directly tested on the Hospital-radar domain using t-SNE, as shown in Fig. 2. Although the model learns a well-defined decision boundary on the source domain that separates Light from Deep sleep, it fails to align with the target distribution. As a result, part of the target Deep sleep samples drift toward the source Light sleep region and cross the source-learned decision boundary. Without proper adaptation, this feature-level misalignment leads to systematic projection errors in the latent space and eventually causes substantial degradation in staging performance.

To formulate the distribution problem, we define the input space $\mathcal{X}$ of respiratory signals and the label space $\mathcal{Y}$ of sleep stages. A domain $\mathcal{D}$ is represented by a joint distribution $P(X, Y)$. We denote the labeled Source Domain as $\mathcal{D}_S = \{(x_s^i, y_s^i)\}_{i=1}^{N_S} \sim P_S(X, Y)$, which consists of standard abdominal belt recordings. In contrast, the unlabeled Target Domain is denoted as $\mathcal{D}_T = \{x_t^j\}_{j=1}^{N_T} \sim P_T(X, Y)$, where labels are strictly inaccessible during adaptation. Crucially, the target domain is heterogeneous, potentially involving distinct sensing modalities such as non-contact radar or contact-based belts deployed in diverse clinical or home environments. The fundamental challenge stems from the joint distribution discrepancy where $P_S(X, Y) \neq P_T(X, Y)$.

Unlike general computer vision tasks, this discrepancy in sleep staging is closely tied to physical sensing principles and physiological population differences. We formulate this gap by modeling the underlying signal generation process. Let $z(t)$ denote the latent physiological respiratory dynamics that drive the observed breathing motion at time t. The observed signal $x_d(t)$ in a specific domain $d \in \{S, T\}$ is generated as:

$$x_d(t) = \mathcal{T}_d(z(t)) + \eta_d(t) \tag{1}$$

Here, $\mathcal{T}_d(\cdot)$ represents the domain-specific sensor transfer function that maps latent respiratory dynamics to observed sensor readings, and $\eta_d(t)$ denotes the additive environmental noise. Based on Eq. 1, we interpret the domain shift from three perspectives, progressing from physiological priors to sensing principles.

Physiological Shift. The transition from a large-scale community cohort (SHHS1) to specific target populations can introduce label distribution shift. Unlike the source domain, which covers a broad spread of the general population (including both healthy individuals and those with chronic sleep disordered breathing), target domains often represent specific subsets with skewed physiological characteristics. For instance, in our clinical scenario (Hospital dataset), inpatients typically suffer from acute pathologies or severe sleep disorders. These conditions manifest as highly fragmented sleep architectures, characterized by distinct physiological feature modes compared to the general community baseline. Mathematically, this heterogeneity can be expressed as a divergence in the prior probabilities of sleep stages ($P_S(Y) \neq P_T(Y)$). In clinical targets, this may appear as higher proportions of Wake and Light sleep and reduced Deep sleep frequencies. This distributional discrepancy can be quantified by the Kullback-Leibler divergence:

$$D_{KL}(P_T(Y) || P_S(Y)) = \sum_{k \in \mathcal{Y}} P_T(y = k) \log \frac{P_T(y = k)}{P_S(y = k)} > 0 \tag{2}$$

This metric indicates that the label distribution in the target domain differs from the source training data.

Environmental Interference. Environmental conditions introduce unpredictable noise distributions that vary substantially between domains. Let $\eta_d(t)$ denote the additive noise term in Eq. 1. Its probability density function, denoted as $P(\eta)$, exhibits distinct statistical properties across domains. In the source domain (SHHS1), measurements were collected in residential settings using contact-based belts. The interference is primarily dominated by stationary instrumental noise and minor motion artifacts, which can be approximated by a Gaussian distribution: $\eta_S \sim \mathcal{N}(0, \sigma_S^2)$. In contrast, the target domain involves complex scenarios, such as non-contact radar sensing in multi-patient clinical wards. This environment is subject to distinct physical interference, including radar multi-path reflections (clutter) and sporadic, high amplitude body movements from patients or medical staff [1, 8, 41].

To capture this heterogeneity, we use a representative mixture view of the target noise distribution $P_T(\eta)$, combining stationary background noise and sporadic artifacts:

$$P_T(\eta) = (1 - \gamma) \cdot P_{\text{bg}}(\eta) + \gamma \cdot P_{\text{art}}(\eta) \tag{3}$$

where $\gamma \in [0, 1]$ represents the occurrence probability of impulsive artifacts. Here, the two components are defined as follows. $P_{\text{bg}}(\eta) = \mathcal{N}(0, \sigma_T^2)$ denotes the continuous background thermal noise. $P_{\text{art}}(\eta) = \mathcal{U}(-B, B)$

represents the sporadic artifacts modeled by a uniform distribution bounded by amplitude B, accounting for unpredictable, high-magnitude disturbances.

This mixture view indicates that the target domain can contain heavier-tailed outliers relative to the cleaner source measurements.

Non-Linear Sensing Modality Distortion. In cross-modality transfer, a major gap arises from the distinct sensing modalities, specifically, the transfer from contact-based belts to contactless radar. This shift changes the physical representation of respiratory signals, manifesting as a disparity in the sensor transfer function $\mathcal{T}_d$.

Source Domain (Approximately Linear): The RIP belt measures abdominal tension. Its transfer function can be approximated as linear to the respiratory volume:

$$x_S(t) \approx \alpha \cdot z(t) + \epsilon_{inst} \tag{4}$$

where α is a scaling factor.

Target Domain (Non-Linear): Radar detects chest displacement via phase modulation. A simplified phase-based observation can be written as a non-linear sinusoidal transformation:

$$x_T(t) \approx A(d) \cdot \cos\left(\frac{4\pi}{\lambda} z(t) + \phi_0\right) + \eta_{clutter} \tag{5}$$

where λ is the radar wavelength. This illustrates how respiratory motion can enter the radar observation through a non-linear phase relationship.

This functional divergence between $\mathcal{T}_T$ and $\mathcal{T}_S$ leads to a generalized covariate shift. Consequently, the marginal distributions can differ not merely in statistical moments (e.g., mean and variance) but also in their underlying manifold structure: $P_S(\mathcal{T}_S(z)) \neq P_T(\mathcal{T}_T(z))$.

2.2 Source-Free Domain Adaptation Setting

To bridge the aforementioned domain shifts without source data, we turn to a robust solution, the Source-Free Domain Adaptation (SFDA) framework. Unlike conventional adaptation methods that require access to source and target data, SFDA adapts the model using only a pretrained source model and unlabeled target data.

In the first phase (pretraining), we train the model F_θ on the labeled source domain $\mathcal{D}_S = (\mathcal{X}_S, \mathcal{Y}_S)$, where $\mathcal{X}_S$ represents the input respiratory signals and $\mathcal{Y}_S$ denotes the corresponding sleep stage labels. The objective is to minimize the standard supervised cross-entropy loss, learning general sleep staging representations. Upon completion, we retain only the model parameters, denoted as the Source Model. In the second phase (domain adaptation), the model is deployed to a target domain where only the input sequences $\mathcal{X}_T$ are observable. Here, $\mathcal{X}_T$ consists of unlabeled respiratory sequences collected from diverse scenarios like contactless radar in hospital wards or home environments. Since target labels $\mathcal{Y}_T$ are unavailable, the optimization goal shifts from supervised learning to unsupervised adaptation. The challenge is therefore to adapt the source model using only $\mathcal{X}_T$, without supervised feedback from target labels.

Conventional SFDA methods typically rely on self-training, where the model iteratively generates pseudo-labels for unlabeled target samples based on reliability criteria such as confidence or stability. However, applying this strategy to cross-modality sleep staging presents a severe risk. As analyzed in Section 2.1, target signals, especially radar, are corrupted by non-stationary noise and non-linear distortions. A model pretrained on belt data is prone to misinterpreting these sensor artifacts as physiological features, yielding overconfident but incorrect predictions. Using such noisy pseudo-labels for self-training creates a confirmation-bias loop that reinforces errors caused by artifacts rather than authentic physiological patterns, ultimately degrading adaptation.

Addressing this challenge requires a mechanism that extracts trustworthy learning signals from noisy target data. Given the non-linear sensing distortion discussed in Section 2.1, standard alignment strategies may inadvertently align feature representations to artifacts rather than true physiological structures. This motivates two design requirements for AdaSleep: consistency for invariance and gating for reliability. Consistency across valid augmented views helps capture structural patterns that persist despite random sensor noise, while the Quality Gate filters unstable or ambiguous pseudo-labels so that model updates focus on selected reliable supervision.

3 METHODOLOGY

This section details our proposed source-free domain adaptation algorithm **AdaSleep** for the sleep staging task. The core objective of AdaSleep is to robustly adapt a pretrained model F_{θ_S} to unlabeled target data $\mathcal{D}_T$ drawn from a distribution substantially different from that of the source domain. As illustrated in Fig. 3, the framework integrates two synergistic strategies: (1) Entropy Minimization, which sharpens predictive distributions and reduces uncertainty, and (2) consistency-based pseudo-labeling with a Quality Gate. Together, they provide stable supervision while reducing the influence of artifacts.

3.1 Preliminary

Following the formulation in Section 2, we focus on the source-free domain adaptation (SFDA) task. Our objective is to adapt a pretrained model F_{θ_S} for unlabeled target data $\mathcal{D}_T$. The adaptation relies solely on the pretrained model parameters θ_S and the unlabeled target sequences, without access to source data and target labels.

Sleep staging is fundamentally a sequence learning problem, requiring the analysis of temporal correlations across consecutive sleep epochs. We adopt a standard sequence-to-sequence architecture as the backbone model F, which maps an input sequence of sleep signals to a sequence of sleep stage probability vectors. Here, a sleep epoch denotes a standard 30-second physiological segment used for sleep staging. Formally, a single input sample is denoted as $\mathbf{x} = (x_1, x_2, \ldots, x_L)$, representing a sequence of L sleep epochs, where x_t is the signal from the t-th sleep epoch. To facilitate adaptation, we structurally decompose the model F into three functional modules, where $F = g \circ s \circ f$. The epoch feature extractor f extracts discriminative local features from each sleep epoch independently. This module captures morphological characteristics of the physiological signals and maps each epoch to $z_t = f(x_t)$, producing the feature sequence $\mathbf{z} = (z_1, z_2, \ldots, z_L)$. The sequence encoder s aggregates the sequence of local features $(z_1, \ldots, z_L)$ to model long-term temporal dependencies and stage transitions. It produces a contextualized latent representation $(h_1, h_2, \ldots, h_L) = s(\mathbf{z})$ for the whole sequence. The classifier g projects the contextual features into the label space and produces the probability distribution sequence $\mathbf{p}$:

$$\mathbf{p} = (p_1, \ldots, p_L) = (g(h_1), \ldots, g(h_L)) \tag{6}$$

where $p_t \in \mathbb{R}^K$ denotes the predicted probability distribution for the t-th sleep epoch over K sleep stages, and K is the number of sleep stages.

The formulated mapping $F = g \circ s \circ f$ serves as a generic abstraction for sequence-to-sequence sleep staging. To adapt the model to the target domain, we initialize all modules with the pretrained weights θ_S. Crucially, to preserve the label semantics, we freeze the parameters of the classifier g and update only the feature extractor f and sequence encoder s. This strategy ensures that the target feature representations are aligned to the fixed source hypothesis space, preventing semantic drift during the unsupervised adaptation process. Let θ denote the trainable parameters during adaptation, i.e., parameters of f and s. Therefore, our goal is to optimize θ to minimize the unsupervised loss, which we specify in Sections 3.2 and 3.3.

To demonstrate the robustness of our AdaSleep framework, we use three distinct backbone architectures that represent mainstream methodologies in time-series modeling. ResTransnet serves as a primary baseline, combining a 1D-ResNet as the extractor f to capture morphological waveforms and a Transformer as the encoder s to model long-term temporal dependencies via self-attention. NSSPS [40] is a representative RNN-based model that utilizes a multi-scale CNN for feature extraction and Simple Recurrent Units (SRU) for efficient temporal sequence learning. RadioSleep [10] is a hybrid architecture tailored for radio-based sleep monitoring, employing

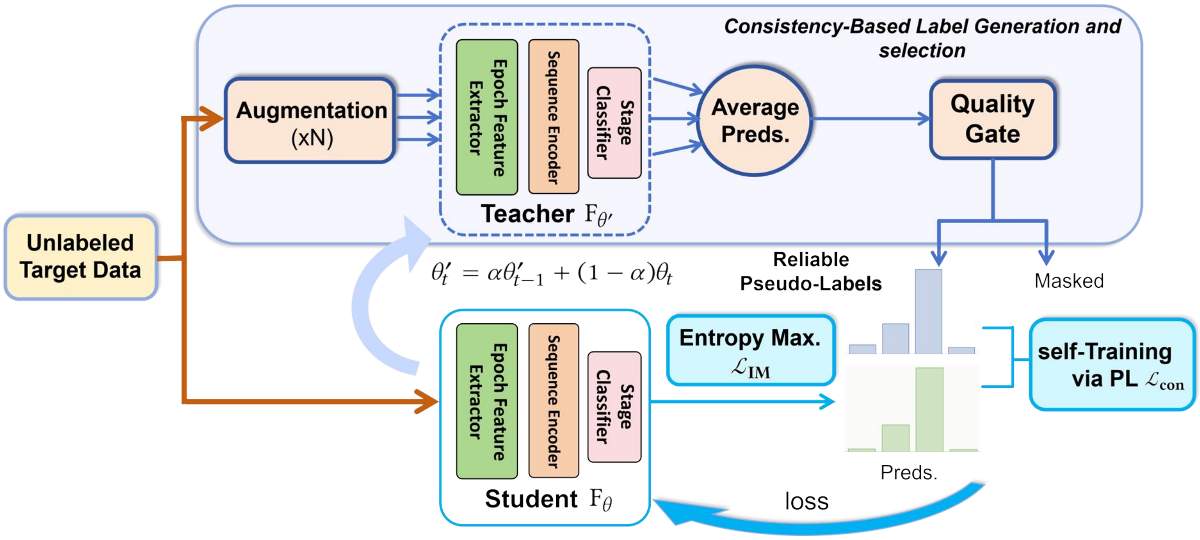


Fig. 3. Overview of AdaSleep. **Consistency-Based Label Generation** aggregates predictions across stochastically augmented views to reduce sensitivity to random noise. A **Quality Gate** selects high-reliability pseudo-labels before Student updates, limiting error propagation. The Student is optimized via **robust self-training** ($\mathcal{L}_{con}$) and **information maximization** ($\mathcal{L}_{IM}$).

deep residual blocks followed by an LSTM combined with an Attention mechanism to capture complex temporal dynamics. In the subsequent adaptation phase, we apply our proposed strategies across these backbones.

3.2 Entropy Minimization

To reduce predictive uncertainty, we adopt Information Maximization (IM) [16] as one of our adaptation objectives. IM combines entropy minimization [9] with a diversity constraint to mitigate mode collapse under sleep stage class imbalance.

Entropy Minimization. Different sleep stages correspond to distinct physiological states and respiratory patterns. A robust model should therefore make confident predictions for target epochs with recognizable patterns. However, due to domain shift, the pretrained model may produce high-entropy predictions on unseen target data. We therefore minimize the conditional entropy of the model's prediction to sharpen its predictions. Let $\mathbf{p}_{i,t} \in \mathbb{R}^K$ denote the predicted probability distribution for the t-th sleep epoch of the i-th sequence in a batch of size B, and let $\epsilon = 10^{-6}$ be a small constant for numerical stability. The entropy minimization loss is formulated as:

$$\mathcal{L}_{ent} = \frac{1}{BL} \sum_{i=1}^{B} \sum_{t=1}^{L} \left(-\sum_{k=1}^{K} p_{i,t}^{(k)} \log \left(p_{i,t}^{(k)} + \epsilon \right) \right) \tag{7}$$

By minimizing $\mathcal{L}_{ent}$, we encourage target representations to move away from decision boundaries defined by the fixed classifier. However, relying solely on entropy minimization entails a substantial risk, particularly given the inherent class imbalance in sleep data. In standard PSG-based staging, Wake, N1, N2, N3, REM, the N2 stage typically dominates the night. In our 4-class setting based on respiratory signals, we aggregate N1 and N2 into Light sleep, while retaining Wake, Deep, and REM. Consequently, the Light sleep class often constitutes the absolute majority, while Deep and REM appear as minority classes.

Diversity Preservation via Distribution Regularization. To prevent the model from falling into a trivial solution, known as mode collapse, where the network erroneously assigns all target samples to the dominant Light sleep class to minimize uncertainty, we introduce a distributional regularization term. We leverage the Kullback-Leibler (KL) divergence to measure the discrepancy between the empirical marginal distribution $\hat{\mathbf{p}}$ and a prior discrete uniform distribution $\mathbf{u}$, where $\mathbf{u} = [1/K, \ldots, 1/K]^\top$. Importantly, this uniform prior is not intended to model the true sleep-stage distribution in the target domain, which is naturally non-uniform and usually dominated by Light sleep. Instead, it serves only as a soft regularizer on the model's empirical marginal predictions during adaptation, discouraging the degenerate solution in which ambiguous target epochs collapse excessively into the majority class. The empirical marginal distribution $\hat{\mathbf{p}} = [\hat{p}_1, \ldots, \hat{p}_K]^\top$ and the diversity loss are defined as follows.

$$\hat{p}_k = \frac{1}{BL} \sum_{i=1}^{B} \sum_{t=1}^{L} p_{i,t}^{(k)} \tag{8}$$

$$\mathcal{L}_{div} = D_{KL}(\hat{\mathbf{p}} \parallel \mathbf{u}) = \sum_{k=1}^{K} \hat{p}_k \log \frac{\hat{p}_k}{1/K} = \sum_{k=1}^{K} \hat{p}_k \log(\hat{p}_k) + \log K, \tag{9}$$

This formulation effectively normalizes the loss value within the range $[0, \log K]$. The minimum value of 0 is achieved when the predicted marginal distribution is perfectly uniform ($\hat{p}_k = 1/K$ for all k), representing the maximum predictive diversity. The maximum value of $\log K$ occurs during complete mode collapse, where the model assigns all sleep epochs to a single category. Therefore, minimizing $\mathcal{L}_{div}$ does not force the adapted model to match a physiologically unrealistic uniform label prior. It simply counterbalances the collapse tendency induced by entropy minimization, while the final predictions are still determined jointly by the target data, the fixed source hypothesis, and the other adaptation losses. By minimizing $\mathcal{L}_{div}$, we reduce the risk that the model excessively favors majority stages and suppresses clinically relevant minority stages, such as Deep sleep and REM, during source-free adaptation.

We combine entropy minimization to reduce prediction uncertainty with diversity regularization to discourage prediction collapse to a single class.

$$\mathcal{L}_{IM} = \mathcal{L}_{ent} + \lambda_{div} \mathcal{L}_{div} \tag{10}$$

where λ_{div} is a trade-off that controls the contribution of diversity regularization relative to the entropy term. We set λ_{div} to 0.001 in all experiments. This loss serves as the fundamental objective, allowing the model to leverage the intrinsic statistical structure of the target data.

3.3 Robust Self-Training with Reliable Pseudo-Labeling

Domain shift between source and target domain often leads to unreliable predictions on the target domain. While standard Information Maximization encourages the model to output confident predictions, it lacks a mechanism to verify their semantic correctness. The pretrained model may yield overconfident but incorrect predictions on the target domain. This creates a confirmation-bias risk, especially on ambiguous sleep stages distorted by signal artifacts. To address this issue, we propose a robust self-training framework designed to extract reliable supervision by distinguishing authentic physiological features from environmental artifacts. Unlike naive pseudo-labeling strategies that utilize all predictions, our approach integrates a Teacher-Student architecture combined with two cooperative mechanisms: (1) **Consistency-Based Label Generation**, which aggregates predictions across valid stochastically augmented views to filter out random noise; and (2) **a rigorous Quality Gate**, which restricts model updates using selected high-reliability pseudo-labels. This design guides domain adaptation with pseudo-labels that exhibit both high certainty and structural stability, thereby reducing noise accumulation.

3.3.1 Stabilizing Supervision via Mean Teacher. In the absence of ground-truth labels, ensuring reliable supervision is crucial. Physiological signals in the target domain, particularly those from non-contact sensors, are inherently unstable and contain substantial noise. Relying solely on the current predictions of the learning model to generate pseudo-labels is risky. This is because stochastic gradient updates can cause instability in decision boundaries, leading to inconsistent guidance across training steps.

To mitigate this instability, we employ the Mean Teacher framework [28]. At the start of adaptation, we initialize two identical networks with the pretrained model weights. For clarity and simplicity in notation, we denote the Student model as F_θ and the Teacher model as $F_{\theta'}$, omitting the target domain subscript T. During training, the Student model F_θ is updated via standard backpropagation to minimize the loss. In contrast, the Teacher model $F_{\theta'}$ serves as a stable reference for pseudo-label generation and does not receive gradient updates. Instead, its parameters are updated as the Exponential Moving Average (EMA) of the Student's weights.

$$\theta'_t = \alpha\theta'_{t-1} + (1-\alpha)\theta_t \tag{11}$$

where t denotes the current training step, and α is a momentum coefficient, which is typically set to 0.999.

This EMA mechanism effectively functions as a temporal low-pass filter on the model parameters. By smoothing out the rapid changes in the Student model, the Teacher model $F_{\theta'}$ yields more robust and temporally stable representations. It ensures stable predictions, which is essential for generating high-quality pseudo-labels in the next steps.

3.3.2 Consistency-Based Label Generation. To extract reliable supervision signals from the noisy target domain and reduce the risk of error propagation during self-training, we propose a Consistency-Based Label Generation strategy. This approach is grounded in the physiological invariance principle, authentic sleep patterns, such as stable Deep sleep respiratory rhythms, should remain structurally stable across valid signal perturbations, whereas predictions originating from random artifacts are typically fragile and inconsistent. Therefore, instead of relying on a single prediction, we generate robust pseudo-labels by aggregating predictions from multiple stochastically augmented views of the same input.

For an input sequence $\mathbf{x}$, we generate N augmented views by applying a composite augmentation function $\mathcal{A}(\cdot)$. This consistency-based aggregation is computed by averaging the probability distributions across all views. To maximize diversity while maintaining semantic integrity, each view is created by sequentially applying three distinct stochastic perturbations. Formally, the augmentation can be represented as:

$$\mathcal{A}(\mathbf{x}) = \mathcal{T}_{warp}(\mathcal{T}_{slice}(\mathcal{T}_{jitter}(\mathbf{x}))) \tag{12}$$

We use transformations as follows [12].

First, to simulate high-frequency signal fluctuations and environmental interference, we employ Jittering, which injects random Gaussian noise into the raw sequence. This operation acts as a regularization mechanism against signal artifacts, preventing the model from overfitting to micro texture noise and enforcing it to focus on the fundamental structural morphology of the respiratory waveforms.

Next, to address the inherent physiological variability across different subjects, particularly the large differences in baseline respiratory rates ranging from 12 to 20 breaths per minute, we apply Window Slicing. This transformation involves randomly cropping a continuous segment, typically retaining 90% of the original length, and linearly interpolating it back to the input resolution. By simulating variations in global signal frequency, this transformation ensures that the learned representations are robust to the inherent temporal variability of physiological events.

Finally, we utilize Window Warping to capture the non-stationary dynamics inherent in complex sleep stages. This is particularly crucial for modeling the irregular rhythms observed during REM. Unlike the global operation of slicing, warping selectively compresses or stretches randomly chosen sub-windows, typically covering 10% of the sequence, while preserving the overall sequence duration. This simulates local temporal distortions, enabling the model to be robust against short-term irregularities while preserving the global temporal context.

The Teacher model $F_{\theta'}$ processes these N stochastically perturbed views. To filter out inconsistent predictions originating from random artifacts, we aggregate the outputs to derive a robust pseudo-label, denoted as $\tilde{y}_{i,t}$. The consistency-based aggregation is computed by averaging the probability distributions across all views.

$$\tilde{y}_{i,t} = \frac{1}{N} \sum_{n=1}^{N} \sigma\left(F_{\theta'}(\mathcal{A}_n(\mathbf{x}_{i,t}))\right) \tag{13}$$

where $\mathcal{A}_n$ represents the n-th stochastic realization of the composite augmentation, and $\sigma(\cdot)$ denotes the softmax function. Note that $\tilde{y}_{i,t} \in \mathbb{R}^K$ represents a refined probability distribution rather than a hard class index. In all main experiments, we use $N = 8$ augmented views by default. This setting provides a practical trade-off between pseudo-label stability and adaptation efficiency, and its effect is analyzed separately in Section 4.6. By enforcing output consistency across these perturbations, we effectively distill reliable supervision.

3.3.3 Reliable Supervision via Quality Gate. Through the consistency-based strategy, we derive robust soft pseudo-labels for the unlabeled target domain. However, it is crucial to recognize that stability under augmentation is a necessary but not sufficient condition for semantic correctness. While averaging multiple views filters out random fluctuations, it does not eliminate structural ambiguities. For instance, samples situated near decision boundaries, such as the subtle transition between Light sleep and Wakefulness, often yield consistent yet uncertain predictions that consistently hover around 0.55 probability. Forcing the Student model to align its predictive distribution with these ambiguous targets introduces a risk of confirmation bias, causing the network to mistakenly reinforce its own uncertainty rather than learning distinct decision boundaries.

To address this challenge, we implement a strict **Quality Gate** to filter the supervision signals. We use the confidence of the Teacher's aggregated soft prediction as an operational indicator of semantic reliability. Consequently, we define a binary Quality Indicator $Q_{i,t}$ to select the supervision signals that satisfy this confidence criterion.

This indicator evaluates whether the generated pseudo-label distribution for a specific sleep epoch (i, t) is sufficiently reliable to guide the model update. Formally, based on the Teacher's robust soft pseudo-label $\tilde{y}_{i,t}$ derived in Eq. 13, the gate is defined as:

$$Q_{i,t} = \mathbb{I}(\max(\tilde{y}_{i,t}) \geq \tau), \tag{14}$$

where $\max(\tilde{y}_{i,t})$ extracts the maximum probability confidence, $\mathbb{I}(\cdot)$ is the indicator function, and τ is the confidence threshold. If $Q_{i,t} = 1$, the supervision signal is accepted as sufficiently reliable under this criterion. If $Q_{i,t} = 0$, it is discarded to reduce the influence of uncertain supervision. In practice, τ determines how conservatively unlabeled epochs are converted into supervision. A lower threshold admits more epochs but increases the chance of including uncertain pseudo-labels, whereas a higher threshold retains only highly confident pseudo-labels but provides less supervision. Confidence is therefore used as a consistent filtering signal rather than a guarantee of pseudo-label correctness after distribution shift.

The reliable consistency loss is computed over the selected high-confidence supervision signals. Distinct from standard cross-entropy which relies on hard labels, we employ the Kullback-Leibler (KL) Divergence to align the Student's predictive distribution $F_\theta(\mathbf{x}_{i,t})$ with the robust soft pseudo-label $\tilde{y}_{i,t}$. This allows the Student to inherit not only the dominant class prediction but also the structural knowledge regarding inter-class relationships embedded in the Teacher's soft targets.

$F_\theta(\mathbf{x}_{i,t})$ represents the predicted probability distribution of the Student model. The consistency loss is defined as:

$$\mathcal{L}_{con} = \frac{1}{\sum_{i=1}^{B}\sum_{t=1}^{L} Q_{i,t}} \sum_{i=1}^{B}\sum_{t=1}^{L} Q_{i,t} \cdot D_{KL}(\tilde{y}_{i,t} \parallel F_\theta(\mathbf{x}_{i,t})) \tag{15}$$

where $D_{KL}(\mathbf{p} \parallel \mathbf{q}) = \sum_k p_k \log \frac{p_k}{q_k}$ measures the distributional discrepancy. In this formulation, the Quality Indicator $Q_{i,t}$ strictly filters the supervision signals, ensuring that the model is updated based on reliable pseudo-labels. The denominator $\sum Q_{i,t}$ computes the average loss over these selected samples rather than the entire batch. This dynamic normalization is crucial for stability, as it maintains a consistent gradient magnitude even when the number of selected samples fluctuates considerably during the early stages of training.

Although the threshold τ remains fixed, the set of **selected instances evolves dynamically throughout the training process**. Initially, the Teacher model exhibits high confidence mainly on samples with distinct, prototypical respiratory patterns, such as stable breathing rhythms often associated with Deep sleep. As the model's representation capability improves via these reliable gradients, its certainty regarding more subtle or complex patterns naturally increases. Consequently, harder samples, such as those situated at the ambiguous boundaries between Light sleep and Wakefulness, progressively satisfy the threshold criteria and are integrated into the training set later in the process. This dynamic selection strategy allows the model to refine its decision boundaries in a stable, progressive manner without being misled by early-stage confusion.

The final training objective integrates Information Maximization with robust self-training:

$$\mathcal{L}_{total} = \mathcal{L}_{IM} + \lambda_{con}\mathcal{L}_{con} \tag{16}$$

Here, Information Maximization $\mathcal{L}_{IM}$ facilitates global distribution alignment by encouraging diverse and confident predictions, while $\mathcal{L}_{con}$ provides explicit semantic guidance through reliable pseudo-labels. λ_{con} balances the contribution between Information Maximization and supervised consistency, enabling AdaSleep to progressively adapt to the target domain while effectively preventing error propagation.

4 EXPERIMENT

In this section, we evaluate the effectiveness of the proposed AdaSleep framework. We utilize the large-scale SHHS1-abd dataset as the source domain for pretraining. SHHS1-abd is derived from the Sleep Heart Health Study (SHHS), a multi-center cohort study [25]. To thoroughly assess the model's adaptability, we employ four unlabeled target datasets derived from distinct real-world collection scenarios. They represent a challenging clinical setting and a relatively stable home setting, respectively, allowing us to benchmark performance under different levels of signal interference and domain shift.

4.1 Datasets

A summary of the datasets is provided in Table 1. The source domain consists of abdominal respiratory signals extracted from PSG recordings, while the target domains include both abdominal signals and simultaneously collected non-contact radar signals.

Source Domain Dataset. The source dataset used for pretraining is **SHHS1-abd**, a subset derived from Sleep Heart Health Study Visit 1 (SHHS-1) of the Sleep Heart Health Study (SHHS), a multi-center cohort study [25]. This landmark study investigated the relationship between sleep-disordered breathing and cardiovascular disease. We used abdominal respiratory signals acquired by respiratory inductance plethysmography from 5,451 overnight recordings. These signals provide a high quality labeled reference for learning generalized sleep features.

Target Domain Datasets. We use datasets collected from two distinct environments: hospital wards and private homes. These datasets allow us to construct scenarios with varying degrees of domain shift difficulty. All experiments involving human participants were approved by the Institutional Review Board (IRB).

Table 1. Summary of the datasets. SHHS1-abd serves as the labeled source domain. The target domains include Hospital-abd, Hospital-radar, Home-abd, and Home-radar. They cover diverse environments, including hospital wards and private homes. All datasets are standardized to a 10 Hz sampling rate.

Dataset	Recordings	Population	Environment	Modality
Source Domain				
SHHS1-abd	5,451	Community Cohort	Residential	Abd Belt
Target Domains				
Hospital-abd	72	Inpatients	Hospital Ward	Abd Belt
Hospital-radar	72	Inpatients	Hospital Ward	Radar
Home-abd	35	Healthy Volunteers	Residential	Abd Belt
Home-radar	35	Healthy Volunteers	Residential	Radar

As shown in Fig. 4, for both target domain collection scenarios, the PSG reference signals were acquired with a Philips Alice PDX system, while the radar signals were recorded simultaneously using a 3-Tx/4-Rx MIMO FMCW radar operating at a 60 GHz center frequency with a bandwidth of 4 GHz. The abdominal belt was worn as part of the PSG setup, and the radar was mounted near the head of the bed at an approximate distance of 1 m and tilted downward by about 45° toward the chest area to maintain a consistent sensing geometry.

Clinical Environment represents a challenging adaptation scenario involving 72 overnight recordings from patients with varying degrees of sleep disorders. In this setting, contact-based PSG and non-contact radar signals were collected simultaneously in multi-bed hospital wards. The patient population exhibits complex and irregular sleep architectures that differ from those in the general cohort, introducing physiological distribution shifts. In addition, the multi-bed ward setting introduces unpredictable background noise and movements from medical staff or other patients at night. From these temporally aligned recordings, we derive **Hospital-abd** (using PSG abdominal belts) and **Hospital-radar** datasets. This configuration allows us to evaluate the model's robustness to clinical noise and pathological variability.

Private Home Environment represents a home-based sleep monitoring scenario, consisting of 35 overnight recordings from primarily healthy individuals collected in their private bedrooms. Similar to the hospital scenario, radar and PSG signals were recorded synchronously to ensure strict temporal alignment. Compared with the hospital setting, the home setting generally involves fewer external disturbances during sleep, and the sleep patterns of these participants are typically less heterogeneous. From these synchronized recordings, we derive **Home-abd** and **Home-radar** datasets. Because the abdominal belt and radar signals were recorded simultaneously in each environment, Hospital-abd and Hospital-radar share the same 72 nights, and Home-abd and Home-radar share the same 35 nights.

4.2 Implementation Details

4.2.1 Signal Preprocessing. To ensure a unified input representation across various domains, we apply a standardized preprocessing pipeline. For radar-based datasets, the raw intermediate frequency (IF) signals acquired by the TI AWR6843AOP 60 GHz MIMO FMCW radar are first processed with beamforming on the virtual antenna array to improve spatial selectivity. We then compute the Breathing-to-Noise Ratio (BNR) for each range bin following prior RF respiration work [38], identify the bins that contain the clearest breathing information, and select the five bins with the highest BNR values. The selected bins are coherently aggregated to obtain the final respiratory waveform, rather than directly using the raw phase signal from a single range bin. All respiratory signals (both abdominal belt and radar) are downsampled from 100 Hz to 10 Hz. We then map the original sleep

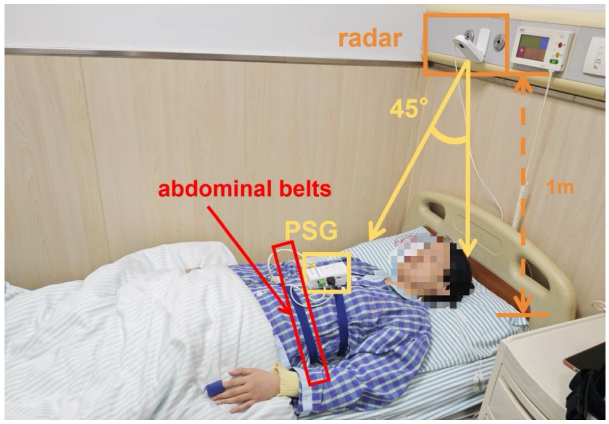


Fig. 4. **Experimental setup for target domain data collection.** The figure shows the placement of the radar, the PSG system, and the abdominal belt, and illustrates the radar position relative to the subject, including its approximate distance and orientation.

stages into four classes: Wake, Light (merge N1 and N2), Deep (N3), and REM. To mitigate amplitude variations across subjects and sensors, we apply subject-specific Z-score normalization. For each overnight recording x, we compute the global mean μ and standard deviation σ across the entire night. Each sleep epoch x_i is normalized as:

$$x_i' = \frac{x_i - \mu}{\sigma + \epsilon} \tag{17}$$

where $\epsilon = 10^{-6}$ is a small constant for numerical stability. This process reduces scale differences and centers each overnight recording.

The ResTransnet model takes the whole-night sequence as a single input sample to capture long-term temporal dependencies. Since sleep duration varies across subjects (typically ranging from 1,100 to 1,500 sleep epochs), we employ a padding-and-masking strategy for batch processing. All input sequences are zero-padded to a fixed maximum length of $L_{max} = 1600$ sleep epochs. During both training and inference, a binary mask is generated to exclude these padded regions from computation, ensuring that the model's attention and loss calculations focus solely on valid sleep epochs.

4.2.2 Source Model Pre-training. To evaluate the effectiveness of AdaSleep, we pretrain three distinct backbones ResTransnet, NSSPS [40], and RadioSleep [10], on the source domain SHHS1-abd. The model is pretrained on the source SHHS1-abd dataset in a supervised manner using SGD optimizer. Due to the memory intensity of whole night sequence modeling, the batch size is set to 8. The initial learning rate is set to 2.0×10^{-3} and follows a cosine annealing schedule to stabilize convergence. The pretraining process runs for 200 training epochs. We split the SHHS1 source subjects into 70% for training and 30% for validation, saving the checkpoint with the highest Macro-F1 score on the validation set.

4.2.3 Domain Adaptation Protocol. During adaptation, only the pretrained model and unlabeled target recordings are used. We leverage the complete unlabeled dataset, allowing multiple adaptation epochs to capture its global distribution before inference. After adaptation, the optimized model is used to perform final sleep staging predictions. During self-training, we freeze the classifier g and update only the feature extractor f and sequence encoder s. This strategy aligns the target features to the fixed source hypothesis space, preventing semantic drift while allowing the encoder to adapt to signal-level and temporal-level shifts.

We use 0.80 as the reference value for the Quality Gate threshold because, in an analysis of the pretrained model on the SHHS validation set, accuracy for epochs whose prediction confidence was no lower than 0.80 reached 90.95%. For Hospital-abd, we set $\tau = 0.81$ and use the Adam optimizer with a learning rate of 1.0×10^{-4}, a weight decay of 1.0×10^{-4}, and 6 adaptation epochs. For Hospital-radar, we set $\tau = 0.80$ and use a learning rate of 1.0×10^{-4} for 10 adaptation epochs. For Home-abd, we set $\tau = 0.78$ and use a learning rate of 6.0×10^{-4} for 5 adaptation epochs. For Home-radar, we set $\tau = 0.82$ and use a learning rate of 2.0×10^{-4} for 10 adaptation epochs. Across all experiments, we use $N = 8$ augmented Teacher views, and the Teacher model's Exponential Moving Average momentum coefficient α is fixed at 0.999.

4.3 Main Results: Adaptation Performance

Compared Methods. To comprehensively evaluate the effectiveness of AdaSleep, we benchmark it against several representative source-free domain adaptation methods. Source Only denotes the source-only baseline, i.e., direct deployment of the pretrained model on the target domain without adaptation. We compare our method with SHOT [16], which aligns feature distributions via information maximization and self-supervised pseudo-labeling. We also compare with C-SFDA [11], which employs a curriculum learning strategy to progressively leverage target samples based on uncertainty estimation. Additionally, we include Tent [30], an online test-time adaptation method that updates model parameters by minimizing the entropy of the incoming testing stream in real time. Finally, we introduce AdaSleep Online as a variant that adapts the model in an online streaming manner rather than using the complete unlabeled target set. In addition, we report DANN [7] and CDAN [19] as standard UDA references. These two methods are listed separately from the SFDA baselines because they need access to labeled source data during adaptation.

We pretrain three distinct backbones ResTransnet, NSSPS, and RadioSleep, on the source domain SHHS1-abd. These pretrained models are then adapted to the target domains using the compared adaptation methods.

Intra-Modality Domain Adaptation. We first evaluate AdaSleep in intra-modality scenarios where both source and target domains utilize abdominal respiratory belts. Despite the shared sensor modality, the source-only baseline exhibits performance degradation when deployed to target domains, particularly on Hospital-abd. For instance, the accuracy of ResTransnet drops to 66.19% in the hospital setting. This decline is consistent with differences in inpatient breathing patterns and with background interference in clinical wards. AdaSleep improves performance under these conditions. Using the ResTransnet backbone, it recovers the accuracy to 71.78% on Hospital-abd and achieves 77.60% on Home-abd, with Macro-F1 scores reaching 67.25% and 76.64%, respectively. The same improvement trend is observed across the three evaluated backbones.

For NSSPS, which starts with a higher source baseline (68.91% accuracy on Hospital-abd), AdaSleep further improves the accuracy to 72.21% and the Macro-F1 to 69.33%, outperforming the competing source-free adaptation methods. The higher Macro-F1 is consistent with better class-balanced recognition, including for the underrepresented REM and Deep-sleep stages. A similar trend is evident with the RadioSleep backbone, where AdaSleep achieves the strongest Home-abd performance among the SFDA methods.

Cross-Modality Domain Adaptation. We further evaluate AdaSleep in the challenging cross-modality scenario, adapting models pretrained on contact-based belt signals to non-contact radar data. As illustrated in Table 2, the source-only baseline experiences substantial performance degradation. Radar-derived respiratory signals can be affected by non-linear distortions and environmental clutter.

In the noise-intensive Hospital-radar setting, where physiological signals can be affected by motion artifacts, the source-only ResTransnet yields a Macro-F1 of 61.18%. AdaSleep improves this result, achieving an accuracy of 70.87% and a Macro-F1 of 66.38%. The same improvement trend is observed with the NSSPS and RadioSleep backbones.

Table 2. Performance comparison of different domain adaptation methods on the ResTransnet, NSSPS, and RadioSleep backbones. Methods marked with † are standard UDA methods (DANN, CDAN). SHOT, Tent, C-SFDA, AdaSleep Online, and AdaSleep are SFDA methods. Accuracy and Macro-F1 are reported in percentage, and Cohen's kappa (κ) is also reported. The best results among SFDA methods are highlighted in **bold**.

Backbone	Method	Hospital-abd			Home-abd			Hospital-radar			Home-radar		
		Acc	MF1	κ	Acc	MF1	κ	Acc	MF1	κ	Acc	MF1	κ
ResTransnet	Source Only	66.19	63.43	0.515	73.89	73.02	0.612	65.61	61.18	0.468	72.07	71.11	0.621
	DANN†	72.85	69.32	0.558	78.27	77.04	0.694	71.41	68.25	0.547	77.41	76.82	0.683
	CDAN†	73.55	69.85	0.572	78.88	77.84	0.702	71.24	68.16	0.539	79.51	77.38	0.700
	SHOT	68.56	65.78	0.534	75.14	74.32	0.635	66.15	62.18	0.475	73.26	72.94	0.646
	Tent	66.10	63.56	0.512	74.80	73.95	0.628	66.67	62.36	0.470	73.16	72.46	0.637
	C-SFDA	66.95	64.10	0.518	76.28	75.57	0.655	68.18	65.34	0.511	69.33	66.66	0.552
	AdaSleep Online	66.97	64.08	0.520	74.69	74.16	0.625	67.45	62.13	0.477	72.32	71.41	0.628
	AdaSleep (Ours)	**71.78**	**67.25**	**0.537**	**77.60**	**76.64**	**0.675**	**70.87**	**66.38**	**0.526**	**77.29**	**76.26**	**0.680**
NSSPS	Source Only	68.91	65.43	0.518	78.49	77.70	0.685	67.13	62.70	0.480	76.48	75.87	0.664
	DANN†	72.83	69.54	0.571	81.44	80.20	0.737	71.65	68.31	0.552	80.16	79.43	0.719
	CDAN†	73.46	70.24	0.589	81.97	80.41	0.738	72.96	68.99	0.570	81.19	80.21	0.736
	SHOT	69.14	65.96	0.522	80.75	79.44	0.713	68.02	64.86	0.504	78.54	76.27	0.687
	Tent	69.11	65.64	0.517	80.68	79.23	0.708	66.90	63.57	0.479	79.71	78.39	0.704
	C-SFDA	71.55	68.76	0.548	79.34	78.41	0.696	70.16	67.21	0.525	78.84	78.01	0.693
	AdaSleep Online	70.44	67.74	0.541	78.67	77.89	0.688	69.34	64.40	0.513	77.74	76.20	0.686
	AdaSleep (Ours)	**72.21**	**69.33**	**0.562**	**80.98**	**79.56**	**0.715**	**71.08**	**67.75**	**0.538**	**79.96**	**79.23**	**0.708**
RadioSleep	Source Only	69.05	65.66	0.521	78.40	77.39	0.682	68.82	64.09	0.513	77.94	76.38	0.677
	DANN†	72.63	69.52	0.568	81.27	80.44	0.724	71.82	68.57	0.556	80.48	79.83	0.719
	CDAN†	73.34	70.41	0.582	81.86	81.03	0.737	72.58	69.21	0.569	81.04	80.32	0.728
	SHOT	71.99	67.79	0.546	80.46	78.82	0.706	71.27	67.22	0.548	79.58	78.92	0.702
	Tent	70.21	66.43	0.533	79.60	78.97	0.693	69.21	64.76	0.516	79.38	78.24	0.696
	C-SFDA	71.45	67.32	0.542	79.03	78.20	0.688	68.92	64.01	0.511	78.90	77.39	0.686
	AdaSleep Online	70.26	66.70	0.537	79.20	78.06	0.691	69.87	64.88	0.519	78.41	77.06	0.681
	AdaSleep (Ours)	**72.49**	**69.86**	**0.574**	**81.11**	**80.37**	**0.721**	**71.46**	**68.40**	**0.552**	**80.15**	**79.71**	**0.723**

In the more stable Home-radar setting, AdaSleep also achieves substantial gains. Specifically, ResTransnet, NSSPS, and RadioSleep achieve accuracies of 77.29%, 79.96%, and 80.15% on the Home-radar dataset, respectively.

C-SFDA decreases performance on Home-radar with ResTransnet from 72.07% Acc and 71.11% Macro-F1 to 69.33% and 66.66%, respectively, although its accuracy increases in the other settings. C-SFDA uses a curriculum self-training strategy that progressively incorporates pseudo-labels according to batch-relative confidence and uncertainty. Under belt-to-radar transfer, errors caused by radar distortions or ambiguous transitions may still appear relatively confident and stable, and thus enter the curriculum early. If selected, such noisy pseudo-labels may be reinforced in subsequent adaptation steps, providing one plausible explanation for the negative transfer in this setting. This observation motivates AdaSleep's multi-view aggregation and explicit confidence threshold.

These results suggest that AdaSleep narrows the observed performance gap between the radar and belt target domains. Finally, our offline framework consistently outperforms the AdaSleep Online variant. This comparison supports the benefit of using the complete unlabeled target set for offline pseudo-label selection. AdaSleep adapts on the complete unlabeled target set, which allows the Teacher to refine pseudo-label selection using the global target distribution. By contrast, AdaSleep Online updates the model batch by batch and can only rely on the

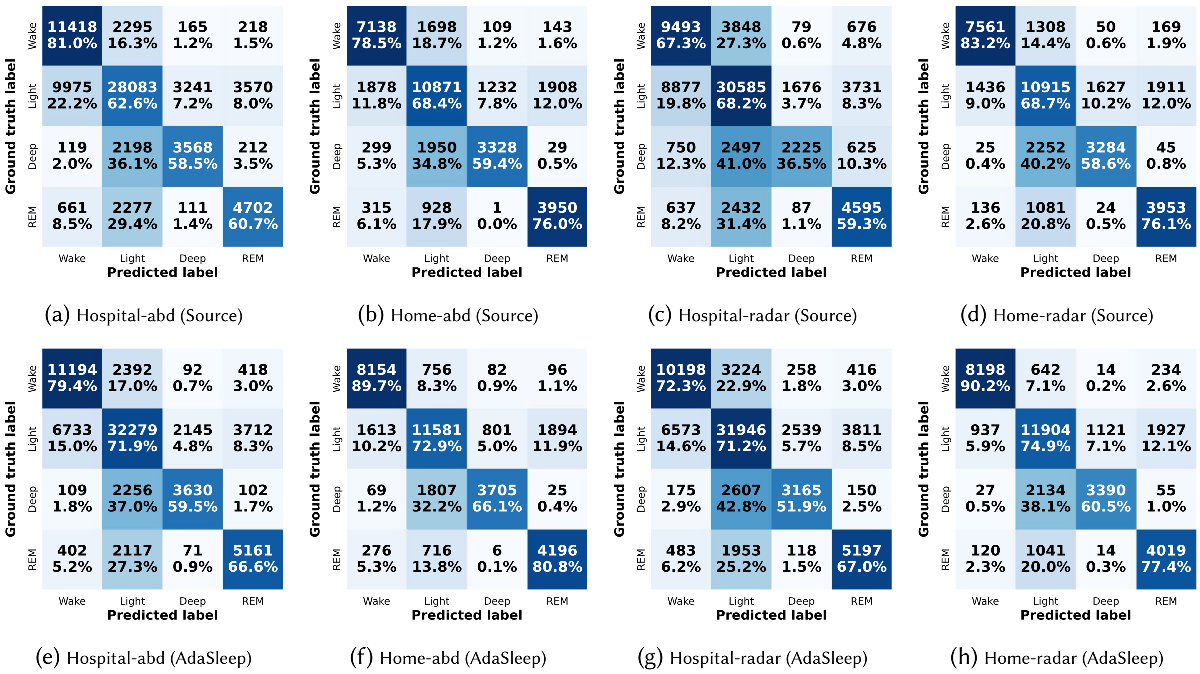


Fig. 5. Comparison of confusion matrices between the source-only baseline (top row) and AdaSleep (bottom row) across intra-modality (a, b, e, f) and cross-modality (c, d, g, h) scenarios. Each row of the matrix represents the ground truth labels, while each column corresponds to the predicted stages, with values normalized by the total number of samples per class. The displayed matrices show reduced confusion of Deep sleep and REM with Light sleep for AdaSleep in several domains, particularly radar domains.

current incoming target batch. Nevertheless, online adaptation remains practically relevant in home monitoring, where full target data are unavailable before prediction.

To further investigate class-wise performance, Fig. 5 illustrates the confusion matrices of the ResTransnet backbone under source-only and AdaSleep configurations across all target domains. The source-only model exhibits substantial inter-class confusion, particularly in cross-modality scenarios. On Hospital-radar, 41.0% of Deep-sleep epochs are misclassified as Light sleep. With AdaSleep, the observed recall of Deep sleep increases from 36.5% to 51.9%. The matrices also show reduced confusion between REM and Light sleep in several domains. These class-wise patterns complement the aggregate metrics but do not establish a causal mechanism.

Comparison with Standard UDA References. The added DANN and CDAN results provide a reference for understanding the performance cost of removing source-data access. As shown in Table 2, standard UDA methods, especially CDAN, usually achieve the strongest results because they can use labeled source data together with unlabeled target data during adaptation. AdaSleep is not intended to replace standard UDA under this less restricted setting. Instead, it targets the stricter SFDA setting, where only the pretrained model and unlabeled target data are available. Under this source-free setting, AdaSleep remains competitive. Across the three backbones and four target domains, it consistently improves over the source-only baseline and outperforms the other SFDA baselines in the evaluated settings.

In summary, AdaSleep improves over the source-only baseline in both intra-modality and cross-modality scenarios. This pattern is consistent across the three evaluated backbones and four target domains. The results support AdaSleep as a practical adaptation approach when source recordings are unavailable.

4.4 Statistical Reliability on Home Benchmarks

Because Home-abd and Home-radar each contain 35 overnight recordings, we further report recording level uncertainty estimates to avoid over-interpreting small margins on these two benchmarks. We use the whole overnight recording as the resampling unit and report the mean together with the half-width of the 95% bootstrap confidence interval.

Table 3. Recording-level mean performance with 95% bootstrap confidence intervals (CI half-widths) on the Home benchmarks (35 recordings each) under the NSSPS backbone. Metrics are reported as mean ± CI half-width.

Dataset	Method	Acc (%)	MF1 (%)	κ
Home-abd	SHOT	79.81 ± 2.69	78.08 ± 3.35	0.693 ± 0.042
Home-abd	Tent	79.48 ± 2.12	77.60 ± 2.64	0.689 ± 0.032
Home-abd	C-SFDA	77.34 ± 3.09	76.67 ± 2.99	0.662 ± 0.037
Home-abd	AdaSleep	79.94 ± 2.64	78.47 ± 2.71	0.702 ± 0.039
Home-radar	SHOT	77.65 ± 2.24	75.43 ± 2.17	0.657 ± 0.027
Home-radar	Tent	78.93 ± 1.73	77.84 ± 1.91	0.689 ± 0.026
Home-radar	C-SFDA	77.97 ± 3.13	77.47 ± 2.48	0.677 ± 0.034
Home-radar	AdaSleep	79.08 ± 1.83	78.16 ± 2.02	0.694 ± 0.028

Table 3 provides a recording-level uncertainty view of the Home results. On Home-radar, AdaSleep has the highest point estimates under NSSPS, but its intervals overlap with those of competing methods. On Home-abd, its margin over SHOT is also small. We therefore treat the Home results as supportive evidence and avoid drawing strong conclusions from these narrow differences alone.

4.5 Ablation Study: Component Contribution

To investigate the incremental contribution of each component in AdaSleep, we conduct an ablation study on the challenging Home-radar target domain. Table 4 reports the performance of the ResTransnet backbone under four progressive configurations, alongside results for NSSPS and RadioSleep to assess whether the same pattern holds across backbones. We explicitly define the ablation variants constructed by progressively integrating specific loss components. The source-only baseline is the pretrained model deployed directly on the target domain without adaptation.

Impact of Entropy Minimization and Diversity Preservation. We first examine the two information maximization components. Model (a) uses Entropy Minimization $\mathcal{L}_{ent}$ to suppress predictive uncertainty. As shown in Table 4, model (a) improves the overall accuracy to 75.67%. However, the class-wise results show an imbalance: the F1 score for Light sleep reaches 73.44%, while the F1 score for Deep sleep is 62.52%. This disparity is consistent with the risk of using $\mathcal{L}_{ent}$ alone on imbalanced data. When driven only by uncertainty reduction, ambiguous samples can be assigned to the dominant class, producing a low-entropy but class-imbalanced solution.

To address this issue, model (b) builds upon model (a) by introducing Diversity Preservation ($\mathcal{L}_{div}$). This variant acts as a global regularizer, enforcing a constraint on the marginal label distribution to prevent the model from ignoring underrepresented classes. Consequently, model (b) partially mitigates the bias. The F1 score for Deep sleep improves to 64.22%, and the overall accuracy increases to 76.17%. This comparison supports the value

Table 4. Ablation study quantifying the incremental contribution of each component on the Home-radar dataset. $\mathcal{L}_{ent}$ denotes entropy minimization, $\mathcal{L}_{div}$ denotes diversity regularization, $\mathcal{L}_{con}$ denotes consistency-based self-training, and Q_τ denotes the Quality Gate. The best results are highlighted in bold.

Model	Components				Overall		Class-wise F1			
	$\mathcal{L}_{ent}$	$\mathcal{L}_{div}$	$\mathcal{L}_{con}$	Q_τ	Acc	MF1	Wake	Light	Deep	REM
ResTransnet										
Source Only					72.07	71.11	82.88	69.42	62.01	70.14
(a)	✓				75.67	72.42	84.17	73.44	62.52	69.53
(b)	✓	✓			76.17	73.77	85.54	73.92	64.22	71.38
(c)	✓	✓	✓		76.61	74.39	86.67	74.11	65.70	71.09
(d)	✓	✓	✓	✓	**77.29**	**76.26**	**89.25**	**75.32**	**68.83**	**71.63**
NSSPS										
Source Only					76.48	75.87	86.50	74.29	68.12	74.57
(e)	✓				77.50	76.51	88.16	76.87	66.53	74.47
(f)	✓	✓			78.62	77.76	88.66	77.12	69.34	75.91
(g)	✓	✓	✓		79.18	78.41	89.19	77.58	70.84	76.04
(h)	✓	✓	✓	✓	**79.96**	**79.23**	**90.53**	**78.12**	**72.46**	**75.82**
RadioSleep										
Source Only					77.94	76.38	87.09	75.51	67.07	75.86
(i)	✓				78.58	76.90	88.23	77.54	65.59	76.25
(j)	✓	✓			79.22	78.19	88.87	77.81	69.58	76.50
(k)	✓	✓	✓		79.50	79.00	89.26	78.36	71.21	77.16
(l)	✓	✓	✓	✓	**80.15**	**79.71**	**89.65**	**79.13**	**72.23**	**77.84**

of maintaining distributional diversity for unsupervised adaptation, particularly for identifying minority sleep stages.

Impact of Consistency Loss. Compared with model (b), adding Consistency Loss ($\mathcal{L}_{con}$) in model (c) increases Accuracy only from 76.17% to 76.61% and Macro-F1 from 73.77% to 74.39%. This limited gain indicates that prediction stability alone does not ensure correctness: consistency training can reinforce noisy Teacher predictions and increase the risk of confirmation bias.

The Role of the Quality Gate (Q_τ). Model (d) adds the Quality Gate (Q_τ). As shown in Table 4, the Quality Gate raises ResTransnet's accuracy from 76.61% to 77.29% and Macro-F1 from 74.39% to 76.26%. By excluding predictions below the threshold, it limits consistency updates to pseudo-labels that satisfy the confidence criterion and reduces the risk of reinforcing low-confidence predictions.

The incremental improvement trend is also observed across the evaluated architectures. As shown in Table 4, in models (e)–(h) for NSSPS and (i)–(l) for RadioSleep, the full AdaSleep framework outperforms the corresponding partial variants. Specifically, for the RadioSleep backbone, the inclusion of the Quality Gate (model l) boosts the Macro-F1 from 79.00% to 79.71%. These results support the applicability of AdaSleep across the three evaluated backbones.

4.6 Sensitivity Analysis of Adaptation Hyperparameters

The ablation results above support the role of consistency-based self-training, but they do not show how the main hyperparameters affect the adaptation process. We therefore analyze two key hyperparameters in AdaSleep: the number of augmented Teacher views N and the Quality Gate threshold τ. The former controls the stability of multi-view pseudo-label generation, while the latter controls how conservatively target epochs are selected for consistency-based self-training. To keep the analysis aligned with the main mechanism study, we use the ResTransnet backbone and conduct the experiments on Home-radar. For τ, we further examine the same trend on the more challenging Hospital-radar dataset. Unless otherwise stated, all other parameters are kept unchanged.

Table 5. Sensitivity analysis of the number of augmented views N on the Home-radar dataset using the ResTransnet backbone. The default setting used in the main experiments is marked by †.

N	Acc (%)	MF1 (%)	κ	t_{epoch} (s)	t_{total} (s)
Source Only	72.07	71.11	0.621	–	–
1	74.17	72.95	0.640	6.7	67.2
2	76.68	75.36	0.653	11.0	110.1
4	76.74	75.62	0.661	19.7	196.7
6	77.08	75.88	0.676	28.2	281.7
8†	77.29	76.26	0.680	36.7	366.6
16	77.26	76.21	0.691	71.3	712.8

We first analyze the effect of N. This parameter determines how many stochastic Teacher predictions are aggregated to form each soft pseudo-label. For each setting, we report Accuracy (Acc), Macro-F1 (MF1), Cohen's kappa (κ), the average adaptation time per training epoch (t_{epoch}), and the total adaptation time (t_{total}). Table 5 shows that increasing N from 1 to 8 consistently improves adaptation performance. Compared with $N = 1$, the default setting $N = 8$ improves Acc from 74.17 to 77.29, MF1 from 72.95 to 76.26, and κ from 0.640 to 0.680. This trend is consistent with aggregating multiple augmented Teacher views to reduce sensitivity to perturbation noise and stabilize pseudo-labels for self-training.

The improvement from larger N comes with additional computational cost, because the Teacher must process more augmented views for each input. This trend is reflected in both t_{epoch} and t_{total}. For example, increasing N from 8 to 16 nearly doubles the total adaptation time from 366.6 to 712.8 seconds, while bringing almost no additional gain in Acc or MF1. These results show that $N = 8$ provides a practical balance between pseudo-label reliability and adaptation efficiency, which is why it is used as the default setting in the main experiments.

We then analyze the effect of the Quality Gate threshold τ. Before adaptation, we conducted an analysis of the pretrained ResTransnet on the labeled SHHS validation set; the model was not adapted on SHHS. For epochs whose prediction confidence was no lower than 0.80, accuracy reached 90.95%. We therefore use 0.80 as the reference value and evaluate representative thresholds within $[0.70, 0.90]$.

The threshold determines how conservatively unlabeled epochs are admitted into consistency-based self-training. For each evaluated threshold, we report Acc, MF1, κ, and the pseudo-label selection ratio:

$$\text{Selection ratio} = 100\% \times \frac{\sum_{i=1}^{N_T} \sum_{t=1}^{L_i} Q_{i,t}}{\sum_{i=1}^{N_T} L_i}, \tag{18}$$

where N_T is the number of target sequences and L_i is the number of valid sleep epochs in the i-th sequence. This ratio measures the proportion of epochs that pass the Quality Gate and are used for consistency-based self-training.

Tables 6 and 7 show that increasing τ reduces the selection ratio, while Acc, Macro-F1, and κ remain within a narrow range for thresholds near 0.80. The reported target performance metrics are used only for post-hoc sensitivity evaluation, not for threshold selection. During adaptation, τ is fixed in advance and applied only to unlabeled Teacher confidence and consistency signals.

Table 6. Sensitivity analysis of the Quality Gate threshold τ on Home-radar using the ResTransnet backbone. The default setting in the main experiments is marked by †.

τ	Acc (%)	MF1 (%)	κ	Selection ratio (%)
Source Only	72.07	71.11	0.621	–
0.76	75.68	73.24	0.649	76.53
0.78	76.16	75.01	0.655	73.63
0.80	76.66	75.54	0.668	71.64
0.82†	77.29	76.26	0.680	69.63
0.84	76.72	75.84	0.671	65.41
0.86	77.04	75.77	0.672	61.70

Table 7. Sensitivity analysis of the Quality Gate threshold τ on Hospital-radar using the ResTransnet backbone. The default setting in the main experiments is marked by †.

τ	Acc (%)	MF1 (%)	κ	Selection ratio (%)
Source Only	65.61	61.18	0.468	–
0.76	69.94	64.55	0.514	66.41
0.78	70.11	65.99	0.521	60.82
0.80†	70.87	66.38	0.526	57.42
0.82	70.14	65.71	0.524	53.66
0.84	70.46	65.20	0.517	49.14
0.88	69.03	64.12	0.501	30.71

4.7 Computational Cost Analysis

To assess the practical cost of AdaSleep, we report a representative runtime and memory comparison on Home-radar using the ResTransnet backbone. Runtime and peak memory were measured on an NVIDIA GeForce RTX 4090 with batch size 1 over 10 adaptation passes. Timing covers adaptation updates and excludes model loading, data loading, and post-processing. Since AdaSleep relies on multi-view Teacher inference for pseudo-label generation, this analysis focuses on adaptation cost rather than deployment-time inference cost.

Table 8. Computational cost comparison on Home-radar with ResTransnet. Acc and MF1 repeat the main-result values in Table 2 for context, and runtime and peak GPU memory are measured in the separate RTX 4090 benchmark described in the text.

Method	Acc	MF1	t_{epoch} (s)	t_{total} (s)	Peak GPU memory (GB)	Trainable params
Source Only	72.07	71.11	–	–	–	–
SHOT	73.26	72.94	2.47	24.67	5.853	16.68M
Tent	73.16	72.46	1.36	13.56	5.193	15.1K
C-SFDA	69.33	66.66	14.17	141.72	8.442	16.82M
AdaSleep Online	72.32	71.41	34.80	348.01	3.315	16.68M
AdaSleep	77.29	76.26	34.48	344.80	3.315	16.68M

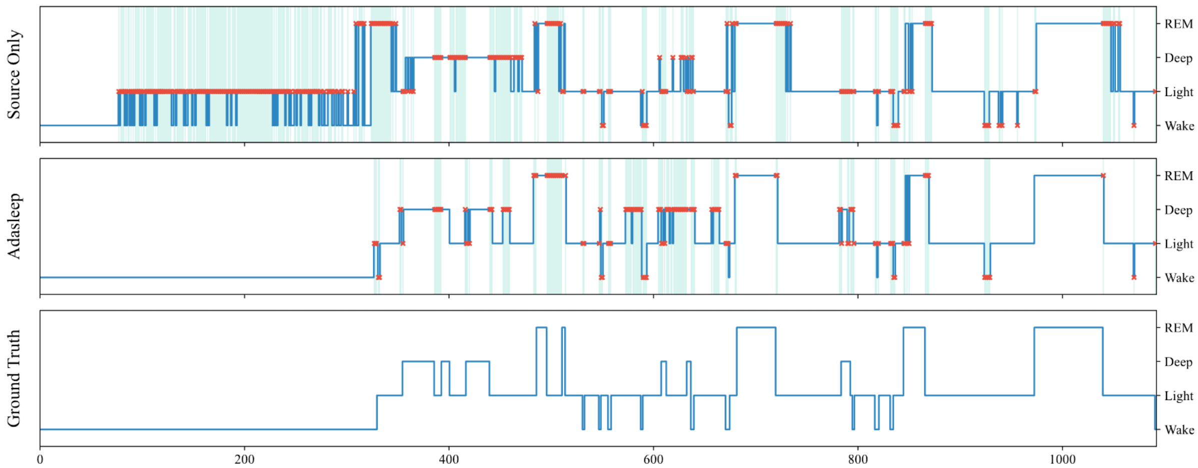


Fig. 6. Qualitative visualization of domain adaptation performance. The figure compares the hypnograms generated by the source only baseline (top) and AdaSleep (middle) against the Ground Truth (bottom) for a representative subject. Red crosses denote misclassified sleep epochs. The comparison highlights two key improvements: (1) Elimination of Instability. AdaSleep effectively filters out noise-induced artifacts during the prolonged Wake stage (sleep epochs 0–320), whereas the baseline exhibits severe high-frequency flickering; (2) Recovery of Deep sleep. AdaSleep correctly identifies Deep sleep segments at sleep epochs 350–450 and 600–700 that are consistently misclassified as Light sleep by the baseline due to signal attenuation.

Table 8 shows that AdaSleep requires more adaptation time than Tent and SHOT, mainly because the Teacher processes $N = 8$ augmented views for each input before applying the Quality Gate. Compared with C-SFDA, AdaSleep has a higher per-epoch and total runtime (34.48 vs. 14.17 seconds per epoch; 344.80 vs. 141.72 seconds in total), but lower peak GPU memory (3.315 GB vs. 8.442 GB). This indicates that the multi-view design increases runtime without a proportional memory increase because the Teacher is evaluated without gradients. After adaptation, inference uses only the adapted Student model, so AdaSleep does not introduce additional inference cost.

4.8 Training Stability and Noise Robustness

To validate the efficiency and robustness of AdaSleep, we show the process throughout the adaptation phase on the Home-radar dataset. Fig.7 analyzes the training behavior from two complementary perspectives, training stability and internal learning trends.

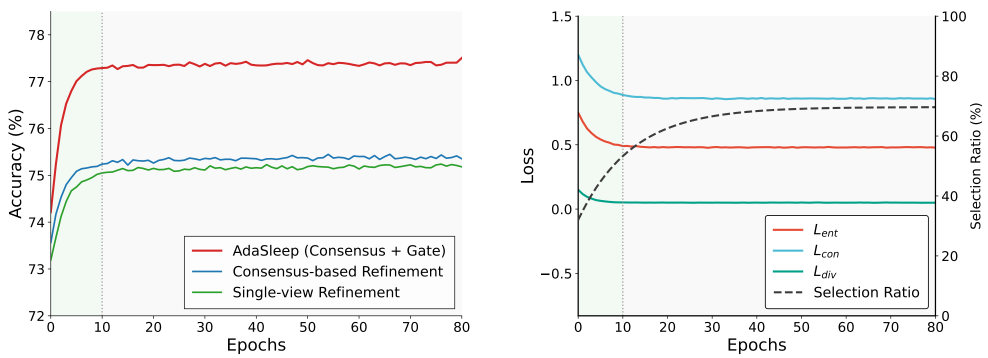
Training Stability and Convergence. Fig. 7a reports the test accuracy across adaptation training epochs. The results reveal a distinct two-phase behavior: a rapid adaptation phase during training epochs 0-10, followed by a stable plateau. The full AdaSleep framework (red line) not only exhibits the steepest initial ascent—indicating high data efficiency—but also maintains a steady performance ceiling up to 80 training epochs. In contrast to the baselines, AdaSleep shows no signs of performance degradation or overfitting despite the prolonged training. This confirms that the proposed Quality Gate effectively filters out noise, preventing the catastrophic forgetting often seen in source-free adaptation when noisy pseudo-labels accumulate over time.

Noise Robustness. Fig.7b investigates the internal state of the model and reveals why AdaSleep converges robustly. We observe the three loss components $\mathcal{L}_{ent}, \mathcal{L}_{con}, \mathcal{L}_{div}$ decrease, indicating that the model progressively gains confidence and aligns with the teacher's knowledge. Meanwhile, the Selection Ratio (dashed line) displays an adaptive increase. Crucially, this ratio does not grow linearly to 100%, instead, it saturates at approximately

69% around training epoch 10. This implies an emergent noise rejection capability, rather than forcing the model to fit every target sample (which would risk overfitting to environment noise or sensor artifacts), AdaSleep naturally establishes an balance and prevent error propagation. It selectively learns from reliable patterns while excluding the remaining uncertainty samples, thereby ensuring robust convergence in the source-free setting.

4.9 Case Study: Qualitative Visualization

To intuitively evaluate the model's performance in real-world scenarios, we present a comprehensive analysis that integrates qualitative hypnogram visualization, quantitative misclassification statistics, and t-SNE feature representation for a representative recording from Home-radar dataset.



(a) Adaptation Accuracy Evolution

(b) Loss Convergence and Selection Ratio

Fig. 7. **Analysis of adaptation dynamics on the Home-radar dataset.** (a) AdaSleep achieves rapid convergence within the first 10 training epochs and maintains a high accuracy plateau without degradation, verifying its **training stability** and **robustness against overfitting**. (b) The losses decrease while the selection ratio increases and plateaus near ~69%, and a subset of target epochs remains excluded from consistency updates.

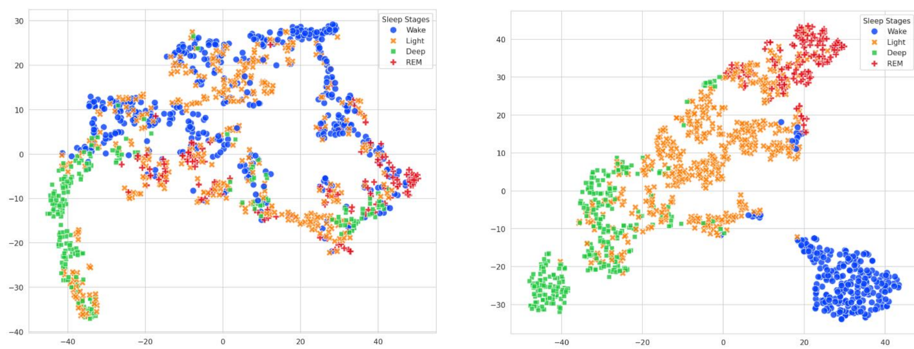
Elimination of Noise-Induced Instability in Wake Stages. The significant improvement is observed during the prolonged wakefulness period (Epochs 0–320) in Fig. 6. As depicted in the top panel, the source only baseline exhibits severe frequent transitions, by misclassifying unstable radar signals as Light sleep. In contrast, AdaSleep (middle panel) generates a spatially consistent prediction that aligns with the Ground Truth. Corroborating this visual observation, the quantitative results in Table 9 reveal a drop in the Wake misclassification rate from 58.71% to 5.90%, with an improvement of 52.81%. This drastic reduction confirms that our consistency-based strategy effectively filters out noise artifacts, as micro-movements that plague the baseline, preventing the model from overfitting to sensor fluctuations.

Rectification of Stage Confusion in Deep sleep. AdaSleep also rectifies the critical stage confusion in Deep sleep. As shown in Fig. 6, the baseline fails to recognize the consolidated Deep sleep blocks in intervals 350–450 and 600–700, erroneously mapping them to Light sleep due to the amplitude attenuation of radar signals. AdaSleep successfully recovers these segments. Quantitatively, this corresponds to a 7.50% reduction in the Deep class misclassification rate in Table 9. While the numerical margin appears smaller than that of Wake, its clinical significance is profound, as accurate Deep sleep quantification is vital for assessing sleep quality.

Feature Space Alignment. To investigate the underlying mechanism driving these performance gains, we visualize the learned feature representations using t-SNE in Fig. 8. As shown in Fig. 8(a), the distributions of Wake (blue) and Light sleep (orange) are heavily entangled with the source only. This feature space confusion directly

Table 9. Quantitative analysis of class-wise misclassification rates on the Home-radar target domain with ResTransnet. The table reports the percentage of erroneous predictions for each sleep stage before and after adaptation.

Sleep Stage	Source Only	AdaSleep	Reduction
Wake	58.71	5.90	52.81
Light	27.47	23.40	4.07
Deep	28.75	21.25	7.50
REM	7.19	5.76	1.44



(a) Source Only (b) AdaSleep

Fig. 8. t-SNE visualization of feature representations. (a) The **source-only** features show overlap, particularly between Wake (blue) and Light sleep (orange). (b) **AdaSleep** shows more separated clusters for Wake and Deep sleep (green) in this visualization.

explains the "flickering" predictions observed in the hypnogram, as the model lacks a clear decision boundary to distinguish resting wakefulness from shallow sleep. Furthermore, the Deep sleep samples (green) are scattered and mixed with Light sleep, leading to the underestimation of deep sleep duration. Conversely, AdaSleep (Fig. 8(b)) achieves clear class separability. The Wake samples form a tight, distinct cluster well-separated from sleep stages, explaining the 52.81% error reduction. Similarly, the Deep sleep features are aggregated into a cohesive manifold. This distinct feature alignment demonstrates that AdaSleep has successfully projected the distorted radar signals into a domain-invariant space, preserving the semantic structure of sleep stages despite the sensing modality shift.

5 Related Work

5.1 Sleep Staging

Sleep staging has traditionally relied on manual scoring of polysomnography (PSG) recordings according to the guidelines of the American Academy of Sleep Medicine (AASM) [3]. With the development of deep learning, automated sleep staging based on physiological signals such as EEG and EOG has achieved substantial progress [23]. Representative early frameworks such as DeepSleepNet [27] combined convolutional feature extraction with recurrent modeling of sleep transitions. Subsequent sequence-to-sequence models, including SeqSleepNet [20] and XSleepNet [21], further modeled temporal dependencies across successive sleep epochs.

More recent architectures have extended this direction toward longer temporal contexts and whole-night modeling [22, 44], while attention-based designs have been explored to emphasize salient patterns in single-channel EEG recordings [6]. These studies establish strong supervised sleep staging backbones, but most still depend on contact-based neurophysiological sensing or assume that training and deployment data are drawn from similar recording conditions.

Respiratory signals provide a complementary route toward less obtrusive sleep monitoring. Large-scale studies have shown that nocturnal respiratory signals contain clinically meaningful information for sleep analysis and related health assessment [34, 46]. At the same time, radio-frequency sensing has enabled contactless monitoring of vital signs and nocturnal physiological or body signals [1, 5, 8, 37, 39, 41]. For sleep sensing, recent systems have used millimeter wave radar [40], commodity Wi-Fi [36], UWB sensing [15], and RF-based neural architectures [43] to infer sleep stages or related nocturnal states without attaching electrodes. In particular, He et al. [10] showed that sleep staging can be performed using either abdominal belt or radar-derived respiratory signals, supporting the feasibility of both contact-based and contactless respiratory modalities.

However, demonstrating that respiratory or RF signals contain sleep-stage information does not by itself solve the deployment problem. Respiratory waveforms vary across subjects, environments, and sensing modalities; radar-derived signals are additionally affected by propagation, body motion, and background clutter. As a result, a model trained on one respiratory domain can degrade when transferred to another domain, especially in abdominal-belt-to-radar settings where both environment and modality change. Recent work has begun to study source-free adaptation for personalized EEG-based sleep staging [45], but respiratory-signal-based sleep staging under cross-environment and cross-modality shifts remains insufficiently explored. This gap motivates an adaptation framework that can use a pretrained sleep staging model and unlabeled target recordings without requiring source data or target labels during adaptation.

5.2 Source-Free Domain Adaptation

Unsupervised domain adaptation (UDA) aims to reduce distribution mismatch between source and target domains, but many conventional UDA methods require access to source data during adaptation [7, 18]. This requirement is often impractical in privacy-sensitive healthcare scenarios, where source recordings may be restricted by institutional policy or patient privacy. Source-Free Domain Adaptation (SFDA) addresses this constraint by adapting a pretrained model using unlabeled target data at adaptation time [13, 16]. This setting matches practical sleep monitoring deployments more closely: a pretrained model can be distributed, while raw source data remain inaccessible.

Existing SFDA methods can be broadly divided into data-based and model-based strategies. Data-based methods approximate source-domain structure through source-like data generation, virtual domain modeling, or source-free domain construction [14, 29, 35], but they may introduce additional modeling assumptions and computational overhead. Model-based methods have therefore become a major line of SFDA research. SHOT [16] freezes the source hypothesis and adapts target representations through information maximization and pseudo-labeling. Later methods exploit target structure through neighborhood consistency, prototype refinement, hypothesis or labeling transfer, and contrastive objectives [4, 17, 24, 32, 33]. Related test-time adaptation methods such as Tent [30] minimize prediction entropy on unlabeled target data, while C-SFDA [11] uses curriculum self-training to prioritize more reliable pseudo-labels. Together, these works show that entropy-based objectives, self-training, pseudo-label selection, and target-structure refinement are effective tools for source-free adaptation.

The challenge for respiratory sleep staging is that these generic adaptation signals can become unreliable under noisy physiological sensing. Pseudo-labeling is known to suffer from confirmation bias when incorrect model predictions are reused as training targets [2]. In respiratory recordings, this risk is amplified by subject-specific breathing patterns, sensor noise, radar artifacts, and ambiguous transitions among sleep stages. A confident prediction may therefore reflect a stable artifact or a boundary-stage ambiguity rather than a reliable physiological pattern. Existing SFDA and test-time adaptation methods provide useful general mechanisms, but they do not explicitly distinguish reliable respiratory supervision from noisy pseudo-labels under stochastic signal perturbations. AdaSleep is designed for this specific setting: it follows the offline SFDA assumption of adapting only from a pretrained model and unlabeled target data, and introduces a reliable supervision framework for respiratory-signal-based sleep staging under severe domain shift. Rather than treating every target prediction as supervision, AdaSleep aggregates Teacher predictions from multiple valid augmented views of the same respiratory sequence to form a soft pseudo-label, and then applies a Quality Gate so that Student updates are driven only by target epochs accepted as sufficiently reliable under this operational criterion. This positions AdaSleep as a domain-specific SFDA framework for robust respiratory sleep staging, with the goal of reducing unreliable pseudo-label supervision rather than proposing a new general-purpose SFDA primitive.

6 DISCUSSION

AdaSleep targets strict SFDA for respiratory sleep staging across populations, environments, and sensing modalities, using only a source-pretrained model and unlabeled target data. This setting covers population and environmental shifts within the same modality, as well as additional sensing differences and motion artifacts when transferring from abdominal belt signals to radar. AdaSleep uses multiple Teacher views to stabilize pseudo-labels and a Quality Gate to reduce error propagation and confirmation bias. However, the current datasets remain limited in scale and coverage. Future work will expand data collection across devices, centers, and populations to improve generalization.

7 CONCLUSION

In this paper, we presented **AdaSleep**, a source-free domain adaptation framework for respiratory-signal-based sleep staging that adapts a source-pretrained model using only unlabeled target data. AdaSleep combines information maximization with Teacher–Student multi-view consistency and a Quality Gate to reduce the influence of unreliable pseudo-label supervision under respiratory signal shift. Across four target domains, three backbone architectures, and both intra- and cross-modality scenarios, it improves Accuracy by 3.67 percentage points and Macro-F1 by 3.90 percentage points on average relative to the source-only baseline. These findings support reliable pseudo-label selection as a useful direction for source-free adaptation in privacy-sensitive unobtrusive sleep monitoring.

Acknowledgment

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

References

[1] Mostafa Alizadeh, George Shaker, João Carlos Martins De Almeida, Plinio Pelegrini Morita, and Safeddin Safavi-Naeini. 2019. Remote monitoring of human vital signs using mm-wave FMCW radar. *IEEE Access* 7 (2019), 54958–54968.

[2] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International joint conference on neural networks (IJCNN)*. IEEE, 1–8.

[3] Richard B Berry, Rita Brooks, Charlene E Gamaldo, Susan M Harding, Carole Marcus, Bradley V Vaughn, et al. 2012. 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 (2012), 7.

[4] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. 2022. Contrastive Test-Time Adaptation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 295–305. https://doi.org/10.1109/CVPR52688.2022.00039

[5] Sergei Churkin and Lesya Anishchenko. 2015. Millimeter-wave radar for vital signs monitoring. In *2015 IEEE International Conference on Microwaves, Communications, Antennas and Electronic Systems (COMCAS)*. IEEE, 1–4.

[6] Emadeldeen Eldele, Zhenghua Chen, Chengyu Liu, et al. 2021. An attention-based deep learning approach for sleep stage classification with single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 29 (2021), 809–818.

[7] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research* 17, 59 (2016), 1–35. http://jmlr.org/papers/v17/15-239.html

[8] Hanqin Gong, Jinbo Chen, Guixin Xu, Jianwen Tong, Yadong Li, Dongheng Zhang, Yang Hu, and Yan Chen. 2025. OSense: Omni-Directional Heartbeat Sensing With Radio Signal. *IEEE Internet of Things Journal* 12, 13 (2025), 25269–25285. https://doi.org/10.1109/JIOT.2025.3558733

[9] Yves Grandvalet and Yoshua Bengio. 2004. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, Vol. 17.

[10] Hao He, Chao Li, Wolfgang Ganglberger, et al. 2025. What Radio Waves Tell Us about Sleep! *SLEEP* 48, 1 (2025), zsae187.

[11] Nazmul Karim, Niluthpol Chowdhury Mithun, Abhinav Rajvanshi, Han-pang Chiu, Supun Samarasekera, and Nazanin Rahnavard. 2023. C-SFDA: A Curriculum Learning Aided Self-Training Framework for Efficient Source Free Domain Adaptation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 24120–24131. https://doi.org/10.1109/CVPR52729.2023.02310

[12] Arthur Le Guennec, Simon Malinowski, and Romain Tavenard. 2016. Data augmentation for time series classification using convolutional neural networks. In *ECML/PKDD workshop on advanced analytics and learning on temporal data*.

[13] Jingjing Li, Zhiqi Yu, Zhekai Du, Lei Zhu, and Heng Tao Shen. 2024. A Comprehensive Survey on Source-Free Domain Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 8 (2024), 5743–5762. https://doi.org/10.1109/TPAMI.2024.3370978

[14] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. 2020. Model Adaptation: Unsupervised Domain Adaptation Without Source Data. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9638–9647. https://doi.org/10.1109/CVPR42600.2020.00966

[15] Song Li, Biao Jin, Fan Zhang, Zhaohong Wang, Jian Ma, Xiaojie Ren, and Hongbo Liu. 2024. Hypnos: A Contactless Sleep Stage Monitoring System Using UWB Signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–27.

[16] Jian Liang, Dapeng Hu, and Jiashi Feng. 2020. Do We Really Need to Access the Source Data? Source Hypothesis Transfer for Unsupervised Domain Adaptation. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 6028–6039. https://proceedings.mlr.press/v119/liang20a.html

[17] Jian Liang, Dapeng Hu, Yunbo Wang, Ran He, and Jiashi Feng. 2022. Source Data-Absent Unsupervised Domain Adaptation Through Hypothesis Transfer and Labeling Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 11 (2022), 8602–8617. https://doi.org/10.1109/TPAMI.2021.3103390

[18] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. 2015. Learning Transferable Features with Deep Adaptation Networks. In *International Conference on Machine Learning*.

[19] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. 2018. Conditional Adversarial Domain Adaptation. In *Advances in Neural Information Processing Systems*.

[20] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chén, and Maarten De Vos. 2019. SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 27, 3 (2019), 400–410.

[21] Huy Phan, Oliver Y Chen, Minh C Tran, Philipp Koch, Alfred Mertins, and Maarten De Vos. 2021. XSleepNet: Multi-view sequential model for automatic sleep staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[22] Huy Phan, Kristian P Lorenzen, Elisabeth Heremans, et al. 2023. L-SeqSleepNet: Whole-cycle long sequence modelling for automatic sleep staging. *arXiv preprint arXiv:2301.03441* (2023).

[23] Huy Phan and Kaare Mikkelsen. 2022. Automatic sleep staging of EEG signals: recent development, challenges, and future directions. *Physiological Measurement* 43, 4 (2022), 04TR01.

[24] Sanqing Qu, Guang Chen, Jing Zhang, Zhijun Li, Wei He, and Dacheng Tao. 2022. BMD: A General Class-Balanced Multicentric Dynamic Prototype Strategy for Source-Free Domain Adaptation. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV* (Tel Aviv, Israel). Springer-Verlag, Berlin, Heidelberg, 165–182. https://doi.org/10.1007/978-3-031-19830-4_10

[25] 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. 1997. The sleep heart health study: design, rationale, and methods. *Sleep* 20, 12 (1997), 1077–1085.

[26] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. 2020. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In *Advances in Neural Information Processing Systems*, Vol. 33. 596–608.

[27] Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. 2017. DeepSleepNet: A Model for Automatic Sleep Stage Scoring Based on Raw Single-Channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25, 11 (2017), 1998–2008. https://doi.org/10.1109/TNSRE.2017.2721116

[28] Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, Vol. 30.

[29] Jiayi Tian, Jing Zhang, Wen Li, and Dong Xu. 2022. VDM-DA: Virtual Domain Modeling for Source Data-Free Domain Adaptation. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 6 (2022), 3749–3760. https://doi.org/10.1109/TCSVT.2021.3111034

[30] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *International Conference on Learning Representations*. https://api.semanticscholar.org/CorpusID:232278031

[31] Garrett Wilson and Diane J Cook. 2020. A Survey of Unsupervised Deep Domain Adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)* 11, 5 (2020), 1–46.

[32] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. 2021. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS '21)*. Curran Associates Inc., Red Hook, NY, USA, Article 2250, 13 pages.

[33] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. 2021. Generalized Source-free Domain Adaptation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 8958–8967. https://doi.org/10.1109/ICCV48922.2021.00885

[34] Yuzhe Yang, Yuan Yuan, Guo Zhang, et al. 2022. Artificial intelligence-enabled detection and assessment of Parkinson's disease using nocturnal breathing signals. *Nature Medicine* 28, 10 (2022), 2207–2215.

[35] Hao-Wei Yeh, Baoyao Yang, Pong C. Yuen, and Tatsuya Harada. 2021. SoFA: Source-data-free Feature Alignment for Unsupervised Domain Adaptation. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 474–483. https://doi.org/10.1109/WACV48630.2021.00052

[36] Bohan Yu, Yuxiang Wang, Kai Niu, et al. 2021. WiFi-Sleep: Sleep stage monitoring using commodity Wi-Fi devices. *IEEE Internet of Things Journal* 8, 18 (2021), 13900–13913.

[37] Yuqin Yuan, Jinbo Chen, Dongheng Zhang, Ruixu Geng, Hanqin Gong, Guixin Xu, Yu Pu, Zhi Lu, Yang Hu, Dong Zhang, et al. 2025. Atrial fibrillation detection via contactless radio monitoring and knowledge transfer. *Nature Communications* 16, 1 (2025), 4317.

[38] Shichao Yue, Hao He, Hao Wang, Hariharan Rahul, and Dina Katabi. 2018. Extracting Multi-Person Respiration from Entangled RF Signals. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 2, Article 86 (July 2018), 22 pages. https://doi.org/10.1145/3214289

[39] Shichao Yue, Yuzhe Yang, Hao Wang, Hariharan Rahul, and Dina Katabi. 2020. BodyCompass: Monitoring sleep posture with wireless signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (2020), 1–25.

[40] Qiang Zhai, Tao Tang, Xin Lu, Xinyu Zhou, Changzhi Li, Juan Yi, and Ting Liu. 2022. Machine Learning-Enabled Noncontact Sleep Structure Prediction. *Advanced Intelligent Systems* 4 (2022), 2100227.

[41] Dongheng Zhang, Yang Hu, and Yan Chen. 2021. MTrack: Tracking Multiperson Moving Trajectories and Vital Signs With Radio Signals. *IEEE Internet of Things Journal* 8, 5 (2021), 3904–3914. https://doi.org/10.1109/JIOT.2020.3025820

[42] Marvin Zhang, Sergey Levine, and Chelsea Finn. 2022. MEMO: Test Time Robustness via Adaptation and Augmentation. *Advances in neural information processing systems* 35 (2022), 38629–38642.

[43] Mingmin Zhao, Shichao Yue, Dina Katabi, Tommi S Jaakkola, and Matt T Bianchi. 2017. Learning sleep stages from radio signals: A conditional adversarial architecture. *International Conference on Machine Learning (ICML)* (2017).

[44] Fang Zhou, Zhi Lu, Zhi Wu, Gaohan Ye, Lingjie Shu, Yu Pu, Beilei Wang, Dong Zhang, Dongheng Zhang, Yang Hu, et al. 2025. WN-Sleep: Modeling Whole-Night Data for Improved Sleep Staging Classification. *IEEE Journal of Biomedical and Health Informatics* (2025).

[45] Yangxuan Zhou, Sha Zhao, Jiquan Wang, Haiteng Jiang, Shijian Li, Benyan Luo, Tao Li, and Gang Pan. 2025. Personalized Sleep Staging Leveraging Source-free Unsupervised Domain Adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence* (2025), 14529–14537.

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