Certified Robustness against Sensor Heterogeneity in Acoustic Sensing

PHUC DUC NGUYEN, Nanyang Technological University, Singapore
YIMIN DAI, Nanyang Technological University, Singapore
XIAO-LI LI[†], Institute for Infocomm Research (I2R), A*STAR, Singapore
RUI TAN, Nanyang Technological University, Singapore

Domain shifts due to microphone hardware heterogeneity pose challenges to machine learning-based acoustic sensing. Existing methods enhance empirical performance but lack theoretical understanding. This paper proposes Certified Adaptive Physics-informed transform (CertiAPT), an approach that provides formal certification on the model accuracy and improves empirical performance against microphone-induced domain shifts. CertiAPT incorporates a novel Adaptive Physics-informed Transform (APT) to create transformations toward the target microphone without requiring application samples collected by the target microphone. It also establishes a theoretical upper bound on accuracy degradation due to microphone characteristic differences on unseen microphones. Furthermore, a robust training method with an APT gradient update scheme leverages APT and certification constraints to tighten the upper bound and improve empirical accuracy across sensor conditions. Extensive experiments on three acoustic sensing tasks, including keyword spotting, room recognition, and automated speech recognition, validate CertiAPT's certified robustness and show accuracy gains, compared with the latest approaches. Our implementation of CertiAPT is available at: https://github.com/bibom108/CertiAPT.

CCS Concepts: • **Computer systems organization** → **Embedded and cyber-physical systems**; • **Computing methodologies** → **Neural networks**; • **Hardware** → *Sensor applications and deployments.*

Additional Key Words and Phrases: Cyber-physical system, Sensor heterogeneity, domain adaptation, certified robustness

ACM Reference Format:
Phuc Duc Nguyen, Yimin Dai, Xiao-Li Li, and Rui Tan. 2025. Certified Robustness against Sensor Heterogeneity in Acoustic Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 3, Article 119 (September 2025), 30 pages. https://doi.org/10.1145/3749481

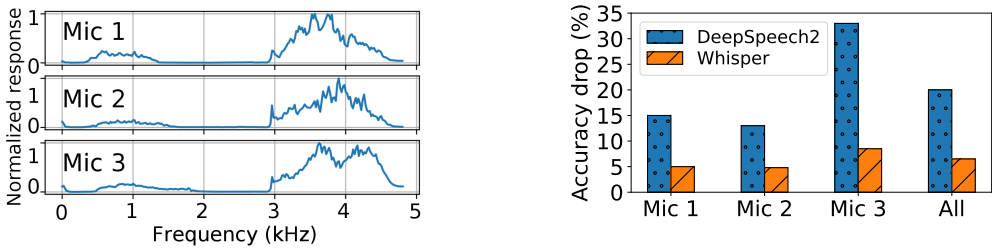
1 INTRODUCTION

Deep learning has brought substantial performance improvements in a range of sensing applications. However, real-world sensing systems often suffer from domain shifts, as they heavily depend on the patterns learned from the training data. Common factors attributed to domain shifts include *environmental variability* and *sensor heterogeneity* that introduce changes to the data distribution. In this paper, we focus on addressing the microphone heterogeneity in acoustic sensing applications. This issue, often overlooked, is in fact critical to sensing performance. Figure 1a shows the frequency response curves (FRCs) of three microphones to the same audio. The differences among the FRCs lead to inconsistencies in the data recorded by the microphones. In particular, the FRC of a user's microphone is generally different from those used to collect the training data for training and validating a deep learning-based acoustic sensing model. The difference leads to a drop in accuracy

[†]Also with Nanyang Technological University, Singapore

Authors' Contact Information: Phuc Duc Nguyen, ducphuc001@e.ntu.edu.sg, Nanyang Technological University, Singapore; Yimin Dai, Nanyang Technological University, Singapore, yimin006@e.ntu.edu.sg; Xiao-Li Li, Institute for Infocomm Research (I2R), A*STAR, Singapore, xlli@i2r.a-star.edu.sg; Rui Tan, Nanyang Technological University, Singapore, tanrui@ntu.edu.sg.

ACM 2474-9567/2025/9-ART119
https://doi.org/10.1145/3749481



(a) The FRCs.

(b) Performance degradation.

Fig. 1. Microphone heterogeneity and its impact. (a) FRCs of three microphones to the same audio. (b) Accuracy drops of two speech recognition models when tested on data collected by various microphones.

from those obtained during the training and validation stages. Figure 1b shows the accuracy drops of two speech recognition models, Baidu DeepSpeech2 [2] and OpenAI Whisper [72], where the latter is trained on a larger dataset and considered a foundation model. Both models exhibit noticeable accuracy drops when tested with data collected by the three microphones mentioned above.

Domain adaptation [51] offers a potential solution to the microphone heterogeneity challenge. It aims to enable models trained on a source domain to generalize effectively to a target domain with a different data distribution. This is achieved by measuring and minimizing the discrepancy or divergence between the two domains [45, 65, 91, 118]. In the context of this paper, the training dataset for building the acoustic sensing model forms the source domain, whereas the inference data collected by the user's microphone forms the target domain. Many domain adaptation techniques require large amounts of target-domain data, making them unpractical in many real-world use scenarios. Few-shot domain adaptation techniques [58, 62, 63, 117] have been applied to reduce the demand on target-domain data, but further reductions while maintaining accuracy are still desirable. To this end, the Physics-Informed Machine Learning (PIML) [29] has received research attention. It integrates physical laws and constraints into the learning process to improve the data efficiency of machine learning. By exploiting the physics governing the domain shifts, PIML has shown promising domain adaptation performance [87], because it can effectively shrink the search space for model optimization [112]. PIML can be also used to generate more realistic data characterizing the target domain for better domain adaptation [55].

Despite the above advances, we still face a major challenge stemming from the uncertainties in the microphone heterogeneity. The existing PIML approaches [55, 87, 112] are based on deterministic parametric physical laws, which limit the learned models' capacity due to nuanced variations and inaccuracy in the parameterization of the underlying physical laws. Moreover, as a general limitation of all existing domain adaptation approaches applied to address sensor heterogeneity, there is a lack of accuracy certification, i.e., the meaningful bounds of accuracy drop are unavailable. As a result, although a certain approach may perform well on the tested microphones, it provides no meaningful information regarding the sensing accuracy of the adapted models on unseen microphones. Under a broader context, we performed a survey on the research papers presented at six sensing-related conferences from 2020 to 2024 on the topics of domain shift, sensor heterogeneity, and cross-platform implementation. Out of 69 papers considered relevant (a majority of them is from IMWUT/UbiComp), only one paper [105], which is not based on deep learning, provides accuracy certification. The survey results can be found in Appendix A.

In this paper, we present Certified Adaptive Physics-informed Transform (CertiAPT), a novel framework that provides accuracy certification for domain shift caused by microphone heterogeneity while empirically improving the accuracy of the adapted model. First, CertiAPT's model adaptation only requires swift profiling of the target microphone, which is agnostic to acoustic sensing applications. Second, CertiAPT provides a meaningful upper

bound of the accuracy drop of the adapted model with respect to the source-domain model's testing accuracy. This accuracy certification process quantifies the model's expected performance degradation in unseen target domains without requiring direct testing on concrete target-domain data. As a result, it offers a guaranteed performance assessment of the model in real-world domain shift scenarios. The core of CertiAPT is a novel Adaptive Physics-informed Transform (APT) with a learnable parameter vector that captures the FRC as a governing characteristic of microphone. Based on a recent accuracy certification theoretical framework developed in [38], we leverage the APT to derive a tighter upper bound on accuracy degradation for target domain data. Built upon APT, CertiAPT applies a proposed robust training method to improve the upper bound and the model's accuracy. The robust training method uses an APT gradient update scheme to complete domain adaptation under the robustness criterion.

We evaluate the effectiveness of CertiAPT on three acoustic sensing applications, which are keyword spotting (KWS), acoustic-based room recognition (ARR), and automated speech recognition (ASR), under open datasets from [55] and our recordings collected from real-world environments. The tightness of the accuracy drop upper bound given by CertiAPT is also evaluated. Empirically, CertiAPT achieves an improvement of 29.58% and 9.98% over PhyAug [55], a physics-informed domain adaptation approach, and 18.66% and 8.01% over CosMix [64], a recent contrastive learning domain adaptation method for limited target domain data, on the KWS and ARR tasks, respectively. For the ASR task, where CosMix is not applicable, CertiAPT achieves an improvement of 10.2% over PhyAug.

Contributions of this paper are summarized as follows:

- We propose CertiAPT, a novel framework that provides a tighter upper bound of drop in accuracy when operating in the source and target domains. CertiAPT is particularly designed to offer guarantee under domain shift induced by microphone heterogeneity.
- We propose APT to enable transforms of source-domain data toward the target domain without requiring target-domain application samples.
- We propose a robust training method that leverages our APT and certification constraints to further tighten the theoretical bound and improve empirical accuracy of the adapted model.
- Our evaluation across three acoustic sensing applications shows that CertiAPT yields significant accuracy improvements compared with the latest relevant domain adaptation approaches and provides formal guarantees on model robustness.

2 RELATED WORK

■ **Few-shot domain adaptation** aims at addressing the domain shift problem with reduced target-domain data, either labeled or unlabeled [83]. Several approaches [58, 66, 113] utilize Generative Adversarial Networks (GANs) to generate additional labeled training data. Other studies focus on aligning the source and target domains through various techniques, such as i) *adversarial learning* [62, 114] that minimizes domain discrepancy by aligning feature representations, ii) *optimal transport* [78] that bridges the source-target gap by modeling domain shifts as a transportation problem, and iii) *test-time adaptation* [95, 111] that adapts the model at inference time to better handle target-domain data. However, to achieve the optimal results, the above approaches still require considerable amounts of target-domain data, e.g., at least 50 to 100 samples in total [78]. Another line of research uses PIML [87] for domain adaptation. PIML leverages physical constraints and abstract mathematical models as prior knowledge to guide the model's learning process. This can be achieved in several methods including i) synthesizing data that follows physical laws [53, 60], ii) embedding physical constraints into the loss function guiding the neural network model training [112], or iii) through a combination of both [34]. This paper presents a new PIML approach following the first method.

■ **Distributionally robust optimization (DRO)** is an approach for minimizing the worst-case expected loss over a set of distributions close to the empirical distribution [74]. This optimization is often challenging, as it requires managing uncertainty across a range of potential data distributions. There are two approaches to address this challenge. The first approach adopts the duality method to relax the problem by converting it into a tractable form in the dual space [7, 9]. The other approaches use the cutting-surface technique, which gradually narrows the feasible region by generating new boundaries called *cutting planes* [6, 73]. In this paper, we adopt the first approach to relax the constraints for our robust training method.

3 PRELIMINARIES

The notations used throughout this paper are summarized in Appendix B.

3.1 Certified Robustness

Certified robustness [46] provides a provable guarantee that a model maintains its accuracy within a bound under a defined perturbation range, ensuring robustness in the presence of adverse or even adversarial conditions [16, 27, 57]. Considering a classifier $h : \mathcal{X} \to \mathcal{Y}$, a given input x, and the classifier output $y = h(x)$, we say h achieves certified robustness at x with a radius ϵ, if $h(x + \delta) = y$ for any perturbation δ satisfying $\|\delta\|_p \leq \epsilon$, where $\|\cdot\|_p$ is the L_p norm. However, the classifier may not achieve certified robustness for every input x. Thus, considering a data distribution $\mathcal{D}$, $B = \inf \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbf{1}_{\|\delta\|_p\leq\epsilon}\{h(x+\delta) = y\}]$ is the tight lower bound of h's accuracy in the presence of input perturbations bounded by ϵ. Note that $\mathbf{1}_{\|\delta\|_p\leq\epsilon}\{h(x+\delta) = y\}$ takes value 0 if there exists a δ such that $h(x + \delta) \neq y$, and takes value 1 otherwise, where $\|\delta\|_p \leq \epsilon$.

However, deriving B analytically is often intractable. To make the analysis tractable, *smoothed classifier* [15] is often used instead of the original classifier h. It is defined as $\tilde{h}(x) = \text{argmax}_{s\in\mathcal{Y}} \mathbb{P}(h(x + n) = s)$, where n follows a zero-mean normal distribution with a standard deviation depending on ϵ. The smoothed classifier can be implemented by yielding the majority of the classifier h's outputs when given many noisified versions of the input x, where the additive noise samples are drawn from the normal distribution. This randomization enables one to derive a certified radius, i.e., a guaranteed neighborhood around any input where $\tilde{h}$ will maintain the same prediction. This technique can generalize across dataset $\mathcal{D}$, without having to solve the generally more difficult problem of computing B directly.

In this paper, we focus on the impact of domain shift on the drop of accuracy from the source domain where the model is trained upon to the target domain. Domain shift can be characterized by bounded distributional perturbation. We consider source- and target-domain data distributions denoted by $\mathcal{D} = (\mathcal{X}, \mathcal{Y})$ and $\tilde{\mathcal{D}} = (\tilde{\mathcal{X}}, \mathcal{Y})$, which are related by $W_1^{\mathbf{c}}(\mathcal{D}, \tilde{\mathcal{D}}) \leq \epsilon$, where $W_1^{\mathbf{c}}(\mathcal{D}, \tilde{\mathcal{D}})$ is the 1-Wasserstein distance between the two distributions. Specifically, $W_1^{\mathbf{c}}(\mathcal{D}, \tilde{\mathcal{D}}) = \inf_{\pi\in\Gamma(\mathcal{D},\tilde{\mathcal{D}})} \mathbb{E}_{(x,x')\sim\pi}\mathbf{c}(x, x')$, where $\mathbf{c} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a distance metric, $\Gamma(\mathcal{D}, \tilde{\mathcal{D}})$ is the set of all couplings of elements in $\mathcal{D}$ and $\tilde{\mathcal{D}}$, π is the joint distribution. In the evaluation experiments conducted in this paper, we set $\mathbf{c}$ as the Euclidean distance between x and x'. The *total variance* of the two distributions is defined as $TV(\mathcal{D}, \tilde{\mathcal{D}}) = \frac{1}{2}\int_\Omega |f_{\mathcal{D}}(x) - f_{\tilde{\mathcal{D}}}(x)|\, dx$, where $f_{\mathcal{D}}(x)$ and $f_{\tilde{\mathcal{D}}}(x)$ are the probability density functions and Ω is the sample space. The following definitions and lemma [38] together describe the impact of domain shift on accuracy drop.

Definition 1 (Parametrizable distribution pair)**.** *$(\mathcal{D}, \tilde{\mathcal{D}})$ is a* parametrizable distribution pair*, if the following three conditions are met:*

(1) $\exists T(\cdot,\cdot)$ such that $\forall x_1 \sim \mathcal{D}$, $\forall x_2 \sim \tilde{\mathcal{D}}$, $\exists\alpha$, $T(x_1, \alpha) = x_2$;
(2) $\exists E(\cdot)$, such that $W_1^c(\mathcal{D}, \tilde{\mathcal{D}}) \leq E(\alpha)$;
(3) If $T(x_1, \alpha_1) = x_2$ and $T(x_2, \alpha_2) = x_3$, $T(x_1, \alpha_1 + \alpha_2) = x_3$.

In Definition 1, α is a constant parameter associated with the transform $T(\cdot, \alpha)$ that maps $\mathcal{D}$ to $\tilde{\mathcal{D}}$.

Definition 2 (Accuracy drop upper bound function)**.** *Given a parametrizable distribution pair $(\mathcal{D}, \tilde{\mathcal{D}})$, any concave function $\psi(\cdot, \cdot)$ meeting the following condition is called an* accuracy drop upper bound function*:*

$$TV(\mathcal{N}(x_1, \eta), \mathcal{N}(x_2, \eta)) \leq \psi(d(x_1, x_2), \eta), \forall x_1 \sim \mathcal{D}, \forall x_2 \sim \tilde{\mathcal{D}}, \tag{1}$$

where $d(\cdot, \cdot)$ is a distance function.

The smoothed classifier is now defined as $\tilde{h}(x) = \text{argmax}_{s \in \mathcal{Y}} \mathbb{P}(h(T(x, \alpha + n)) = s)$. We slightly abuse the notation of $\tilde{h}$ by letting $\mathbb{E}_{(x,y)\sim\mathcal{D}}[\tilde{h}(x, y)]$ and $\mathbb{E}_{(x,y)\sim\tilde{\mathcal{D}}}[\tilde{h}(x, y)]$ denote the accuracy of the smoothed classifier $\tilde{h}$ under $\mathcal{D}$ and $\tilde{\mathcal{D}}$, respectively.

Lemma 1 (Restated from [38])**.** *For a parametrizable distribution pair $(\mathcal{D}, \tilde{\mathcal{D}})$ and any $\psi(\cdot, \cdot)$ given by Definition 2, by denoting $\epsilon = E(\alpha)$ where α is the parameter of the transform $T(\cdot, \alpha)$ mapping $\mathcal{D}$ to $\tilde{\mathcal{D}}$, we have*

$$\left|\mathbb{E}_{(x_1,y_1)\sim\mathcal{D}}[\tilde{h}(x_1, y_1)] - \mathbb{E}_{(x_2,y_2)\sim\tilde{\mathcal{D}}}[\tilde{h}(x_2, y_2)]\right| \leq \psi(\epsilon, \eta). \tag{2}$$

The *certification algorithm*, as detailed in Algorithm 2 of [38], computes the certified accuracy as the right-hand side of the following inequality when given a running parameter ϵ and a fixed η:

$$\mathbb{E}_{(x_2,y_2)\sim\tilde{\mathcal{D}}}[\tilde{h}(x_2, y_2)] \geq \mathbb{E}_{(x_1,y_1)\sim\mathcal{D}}[\tilde{h}(x_1, y_1)] - \psi(\epsilon, \eta). \tag{3}$$

It evaluates the model's certified accuracy on domains that differ from the source domain by at most ϵ.

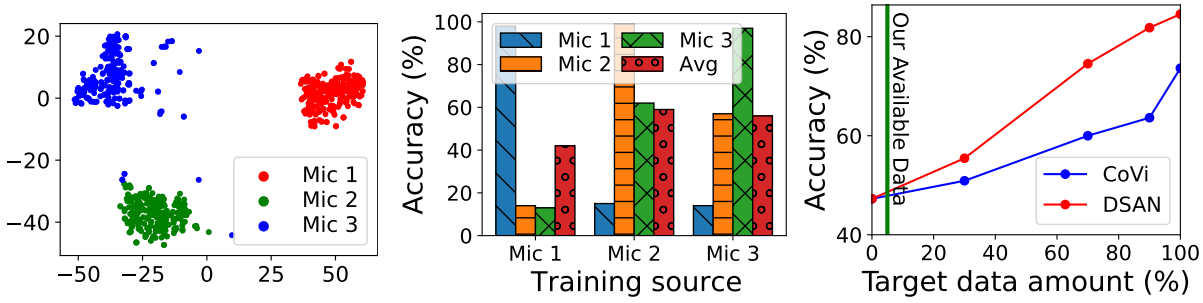
3.2 Robust Training

Robust training is a concept from adversarial learning that involves training the model using samples representing the worst-case scenarios, i.e., samples that violate the certification constraints the most or pose the greatest challenge to the model's robustness. Formally, in the context of domain shift, the objective of robust training [22] is defined as follows:

$$\min_{\theta} \sup_{K \in \mathcal{U}} \mathbb{E}_{(x,y)\sim K}[\ell(\theta, (x, y))], \tag{4}$$

where $\mathcal{U}$ is a set of all the neighbor distributions of the source domain distribution $\mathcal{D}$. The objective is to enhance the model's performance on a distribution K, where the model exhibits its worst-case performance (e.g., the loss function ℓ attains its maximum value on K). The way the set $\mathcal{U}$ is constructed directly affects the optimization process. The common practice of constructing the set $\mathcal{U}$ is to add a small perturbation in the *input space*, where the magnitude of perturbation is controlled by a bound ϵ [15, 75]: $\mathcal{U} = \{K | x' \in \tilde{\mathcal{D}}, x \in \mathcal{D}, ||x' - x||_p \leq \epsilon\}$. Note that, as this paper focuses on acoustic sensing, the *input space* here refers to the space of Mel-Frequency Cepstral Coefficients (MFCC).

However, for acoustic signals, constructing the set $\mathcal{U}$ in the input space using the L_p norm cannot effectively model the effect caused by microphone heterogeneity. Consider a monotonous-frequency sound represented by $p(t) = A_s \sin(2\pi f_s t + \phi_s)$, where $p(t)$ is the sound pressure at time t, A_s is the amplitude of the sound wave, f_s is the frequency, and ϕ_s is the phase shift of the wave. The corresponding voltage induced by the microphone's diaphragm motion can be described by the following expression $V(t) = -Nk \cdot A_s \cdot 2\pi f_s \cos(2\pi f_s t + \phi_s)$, where N is the number of turns in the microphone coil, k is a constant that accounts for both the diaphragm's mechanical sensitivity and the strength of the magnetic field. Due to differences in the coil design and material of various microphones, the conversion equations for transforming sound signals into electrical signals vary among different microphones. This results in a domain shift in the collected data. Therefore, adding a small L_p perturbation in the



(a) (b) (c)

Fig. 2. Impact of microphone heterogeneity. (a) t-SNE visualization of samples in the same KWS class collected by three microphones; (b) Variation in model performance on different training data sources. (c) Domain adaptation techniques and the amount of target domain data required.

input space provides only a partial solution, failing to address the core issue. In Section 5.3, we will propose a more efficient robust training approach.

4 PROBLEM STATEMENT

This paper addresses certification of acoustic sensing robustness against microphone hardware heterogeneity, and how to improve the robustness by adjusting the sensing model.

4.1 A Motivating Example

Figure 2a presents the t-distributed stochastic neighbour embedding (t-SNE) [92] visualization of the data samples in the same KWS class collected from different microphones when the KWS sound is emitted by the same speaker. More details of the data collection process can be found in [55], which provides data as its artifact [1] reused in this work. From the figure, the samples of the same class recorded by the three microphones follow different distributions. These discrepancies among the recorded data cause a model trained on one microphone's data to perform poorly on data from others, as illustrated in Figure 2b. For instance, a model trained on Microphone 1 data can achieve a high testing accuracy of 99% on the same microphone. However, the testing accuracy drops significantly to less than 20% on data from Microphone 2 or Microphone 3. Such drops are also observed when data from Microphone 2 or 3 are used to train the model. One possible solution to address the above issue is to adopt domain adaptation techniques during training. However, these techniques require a large amount of target-domain data, which may not be available in certain applications. Figure 2c illustrates the relationship between the model performance of two domain adaptation approaches, DSAN [117] and CoVi [63], with the amount of target-domain data used. From the figure, when the target-domain data are limited (e.g., less than 5%), the performance of these techniques is comparable to that achieved without using target-domain data. Another limitation of most domain adaptation techniques is the lack of accuracy certification. As a result, the sensing performance on target microphones is not guaranteed.

The above example shows the effect of the domain shift from one specific microphone to another. It provides insights into the more general case considered by this paper, where a large training dataset, possibly collected by many microphones, forms the source domain. As shown in Figure 1, such a general case still suffers from the domain shift problem.

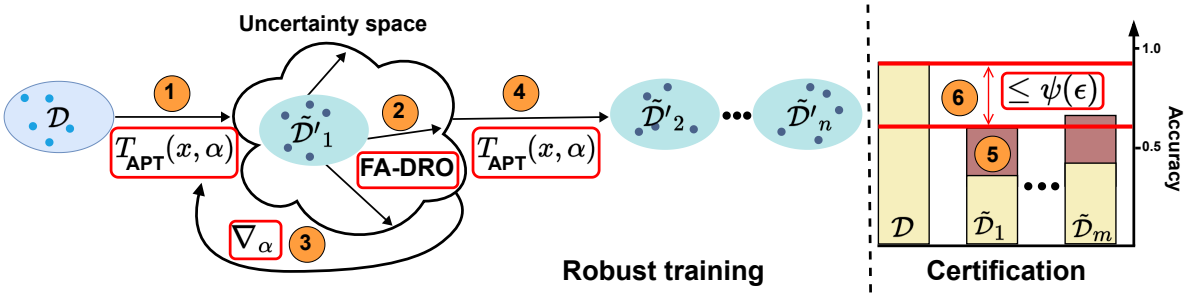


Fig. 3. Overview of the CertiAPT framework. In the robust training procedure, the source-domain is ① transformed via T_{APT}, parameterized by α, forming a domain $\tilde{\mathcal{D}}'_1$. Next, it is ② processed through Frequency-Aware Distributionally Robust Optimization (FA-DRO) to compute the gradient. This gradient is used to ③ update α. This process ④ repeats for n times. The resulting n domains are used to train the classifier, aiming to ⑤ improve empirical accuracy on the target domain $\tilde{\mathcal{D}}$. For certification, the model's accuracy drop between source and *any* target domain $\tilde{\mathcal{D}}_m$ is ⑥ theoretically upper-bounded by $\psi(\epsilon)$, provided that the distance between the source domain and that target domain is ϵ.

4.2 Objectives

The theoretical result in Lemma 1 provides a pathway to certify a sensing model's robustness against domain shift. However, there are three issues in applying Lemma 1 and in improving certified robustness.

■ **Issue 1:** The work [38] shows that given a parametrizable distribution pair $(\mathcal{D}, \tilde{\mathcal{D}})$ and its corresponding transform $T(\cdot, \alpha)$, if c in W_1^c is Euclidean distance and $E(\cdot)$ is L_2 norm, then $W_1^c(\mathcal{D}, \tilde{\mathcal{D}}) \leq E(\alpha)$. Thus, when applying Lemma 1 to a specific sensing application, it is desirable to find a transform $T(\cdot, \alpha)$ to map the data in the source domain $\mathcal{D}$ to the data in a target domain $\tilde{\mathcal{D}}$, such that $T(\cdot, \alpha)$ satisfies the third condition of Lemma 1. This condition ensures that random smoothing in the transform function (i.e. $T(\cdot, \alpha + n)$) remains bounded within the input space, as described in Lemma 1 of [38]. Issue 1 is addressed in Section 5.2.

■ **Issue 2:** The above issue does not involve adjusting the classifier h for improved accuracy. We aim to design a *robust training* approach that is customized for acoustic sensing to fine-tune and optimize the classifier h based on certification criteria and empirical observation. This is the subject of Section 5.3.

■ **Issue 3:** While Definition 2 states the condition that $\psi(\cdot, \cdot)$ must satisfy, it does not provide a concrete function to use. The work [38] has shown that when the distance function used by Definition 2 is Euclidean distance, the error function erf $\left(\frac{\epsilon}{2\sqrt{2}\eta}\right)$ can be a concrete $\psi(\epsilon, \eta)$. However, this only provides a loose upper bound on the accuracy drop, as detailed in Section 5.4. It is desirable to find a tighter bound. Issue 3 is addressed in Section 5.4.

5 CertiAPT

5.1 Overview

This paper presents Certified Adaptive Physics-informed Transform (CertiAPT), a framework designed to mitigate the impact of microphone heterogeneity on model accuracy, while providing accuracy drop guarantee. Figure 3 presents an overview of the CertiAPT framework, which includes the robust training phase and certification phase.

During the robust training phase, unlike conventional methods that require target-domain data to learn distributional shifts, CertiAPT leverages principles of sound acquisition by microphones, avoiding the need for target-domain training data. CertiAPT introduces a novel Adaptive Physics-Informed Transform (APT) based on the microphone's FRC to enable transformations from source-domain data to simulate target-domain conditions.

This approach requires only white noise samples for profiling and does not rely on any application-specific samples from the target microphone during training. The robust training design of CertiAPT is founded upon two key insights. First, since the input space distance between the source- and target-domain samples is often large, the distance between the source-domain and the domain induced by the APT samples should also be increased to better represent potential target-domain conditions. Second, the loss function should directly depend on APT parameters, allowing the transformation to better simulate realistic microphone-induced domain shifts.

Our primary technical contributions for the first phase include the development of APT with learnable parameters, the design of a robust training framework called Frequency-Aware Distributionally Robost Optimization (FA-DRO) that incorporates a novel loss term and facilitates direct optimization on the APT parameters.

The certification phase allows CertiAPT to establish a theoretical upper bound on accuracy degradation in the presence of microphone heterogeneity, based on Lemma 1. The bound is a guarantee of the model accuracy on target domains that differ from the source domain by at most ϵ, without requiring access to application-specific target domain data. Our primary technical contribution for the second phase is the construction of a tighter accuracy drop upper bound function for Lemma 1.

In the following subsections, we detail the proposed APT, the robust training on APT, and the new accuracy drop upper bound. These designs together enable CertiAPT to perform reliably across diverse environments without relying on target-domain data.

5.2 APT

In this section, we present APT to address Issue 1 outlined in Section 4.2. To apply Lemma 1 to microphone heterogeneity, it is essential to define an effective transform function that satisfies the third condition in Definition 1. A potential candidate is the transform function proposed in [55], defined as $T(\mathbf{x}) = \mathbf{F} \otimes \mathbf{x}$, where $\mathbf{x}$ is the input after applying the short-time Fourier transform (STFT), $\mathbf{F}$ represents the target domain microphone's FRC, and $\otimes$ is the Hadamard product. This approach directly addresses the primary source of sensing variability: differences in sensor transfer functions. Note that $\mathbf{x}$ is directly from the source-domain data, eliminating the need for target-domain data. In addition, $\mathbf{F}$ can be obtained by analyzing the microphone's response to white noise.

However, this function suffers from several limitations. First, the transform is static. Once the $\mathbf{F}$ is obtained, the function becomes fixed, failing to capture more nuanced variations within the target domain. Second, the fixed nature of this transform function hinders its integration into the certification process and robust training, both of which require the application of random smoothing to the transform parameters to account for distributional uncertainty. Lastly, the $\mathbf{F}$ is not always stable and accurately measurable. Practical scenarios often involve environmental noise, hardware inconsistencies, and other factors that limit the estimation of $\mathbf{F}$.

To address these limitations, we propose APT defined as $T_{APT}(\mathbf{x}, \alpha) = (e^{\alpha} \otimes \mathbf{F}) \otimes \mathbf{x}$, where α is a learnable vector parameter with the same dimension as $\mathbf{F}$. Here we choose an exponential function e^{α} instead of α to ensure that the conversion equation satisfies the third condition in Definition 1, i.e., $e^{\alpha_1+\alpha_2} = e^{\alpha_1} \otimes e^{\alpha_2}$. This condition ensures that random smoothing in the transform function (i.e. $T(\cdot, \alpha + n)$) remains bounded within the input space, as described in Lemma 1 of [38]. The formulation allows for dynamic adaptability and supports random smoothing, addressing the limitations of the static transform while remaining efficient to profile. We optimize this transform function, along with the model, through our proposed robust training in Section 5.3.

5.3 Robust Training with FA-DRO

We then address Issue 2 outlined in Section 4.2, i.e., training the model to enhance both certified and empirical accuracy, by introducing the robust training framework of CertiAPT. Our main design is the Frequency-Aware Distributionally Robust Optimization (FA-DRO), which is used to learn the worst-case shifts induced by $T_{APT}(x, \alpha)$.

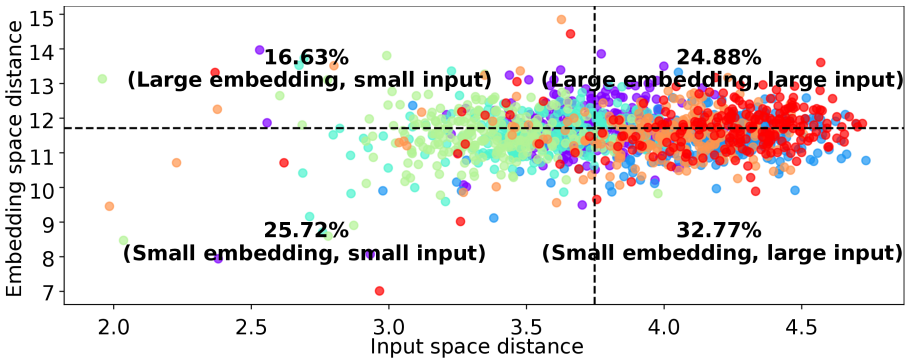


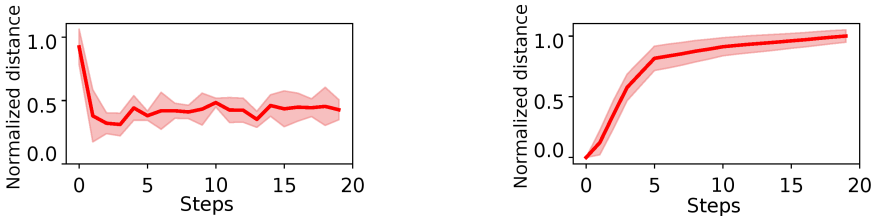
Fig. 4. Distribution of input space and embedding space distances across different classes. Each data point is color-coded by class, representing the distance of the same sample recorded across different microphones. The horizontal and vertical dashed lines are the mean of each distance, divide the distances into large and small regions, with the percentage of data samples in each region annotated.

By optimizing for these extreme scenarios, the model learns to generalize across a range of shifts, thereby reducing sensitivity to both severe and mild variations and improving the certified and empirical accuracy.

5.3.1 FA-DRO. Training the model on worst-case shift data enhances its ability to generalize across a spectrum of shifts, thereby improving performance under both severe and mild shift conditions. To effectively capture the worst-case scenario in robust training, it is necessary to define the distribution set $\mathcal{U}$ in Equation 4. A common approach, beyond the unsuitable definition of $\mathcal{U}$ as a small perturbation in the input space, is to construct $\mathcal{U}$ using a ball of distributions, as proposed in prior works [9, 93]. This is typically expressed as $\mathcal{U} = \{K | W_1^{\mathbf{c_h}}(\mathcal{D}, K) \leq a\}$, where $\mathbf{c_h}$ is the cost function on the *embedding space* (formed by features extracted from an intermediate layer of the model), defined as $\mathbf{c_h}(x', x) = ||h^j(x') - h^j(x)||_2$, where h is a classifier and $h^j(x)$ is the j^{th} layer's output from h with input x. The j is typically the second last layer. This construction is more reasonable for addressing distribution shifts, as it encodes the worst-case constraint (i.e. $W_1^c(\mathcal{D}, \tilde{\mathcal{D}})$) effectively.

However, this approach presents two key challenges. First, we empirically find that samples recorded on different microphones often exhibit similar embedding representations (i.e., they are close in the embedding space) while displaying semantic differences (i.e., they are distant in the input space). Figure 4 illustrates this relationship, showing that most data samples cluster in regions of small embedding distances but large input distances. This observation is consistent with the nature of domain shift, where the underlying data remains fundamentally the same but exhibits value deviations due to sensor-induced variations. The current constraint only addresses the former (embedding space similarity), while neglecting the latter (input space variations). Second, the current definition of $\mathcal{U}$ and its integration into Equation 4 do not account for the specific transformation between domains, as defined in Section 5.2. As a result, the robust training process becomes misaligned with the broader CertiAPT framework, as it does not fully leverage the structured knowledge provided by T_{APT}.

To address the first problem above, we define the set $\mathcal{U} = \{K | K \in \mathcal{D}_{\mathcal{T}}, W_1^{\mathbf{c_h}}(\mathcal{D}, K) \leq a, W_1^{\mathbf{c_X}}(\mathcal{D}, K) \geq b\}$, where $\mathcal{D}_{\mathcal{T}}$ denotes the set of distributions generated from the original distribution $\mathcal{D}$ by all possible parameters α introduced in the T_{APT}. The terms a and b are upper bound and lower bound, respectively, which will be relaxed later. A key contribution here is the incorporation of the term $W_1^{\mathbf{c_X}}$, a constraint derived from our empirical observations, where $\mathbf{c_X}$ denotes the cost function defined on the *input space*. It is intentionally designed in a form similar to $W_1^{\mathbf{c_h}}$ so that it can be relaxed later using the same framework. This constraint is specifically aimed at capturing data points that best represent the target domain. Further details on the construction of $\mathbf{c_X}$ are



(a) Euclidean distance. (b) Proposed distance.

Fig. 5. Comparison of normalized distance values calculated by Euclidean (left) and ours (right) during optimization. As the objective is to maximize the distance, we expect the distance to increase.

deferred to the following section. We then introduce how to optimize the objective defined in Equation 4 with the new $\mathcal{U}$. We adopt the DRO framework in [9, 82], which is commonly used to seek the worst-case expected loss among a ball of distributions, to relax and derive a computable objective for our robust training. This framework requires each cost function to be convex and lower semi-continuous. The detailed derivation steps are provided in Appendix C. The resulting optimization objective is given by:

$$\min_{\theta\in\Theta}\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sup_{x'=T_{APT}(x,\alpha)}\ell(\theta,(x',y))-\gamma\mathbf{c}_{\mathbf{h}}(x',x)+\beta\mathbf{c}_{\mathbf{X}}(x',x)\right]. \tag{5}$$

To address the second problem, we propose to update the gradient computed from the supremum in Equation 5 on the parameter α of the APT function, rather than originally on the input data. Specifically,

$$\alpha_{m+1}\leftarrow\alpha_m+\nabla_\alpha\lambda\Big[\tau\ell\left(\theta,(x'_m,y)\right)-\gamma\mathbf{c}_{\mathbf{h}}(x'_m,x'_{m-1})+\beta\mathbf{c}_{\mathbf{X}}(x'_m,x'_{m-1})\Big], \tag{6}$$

where λ is the learning rate, τ, γ, and β are the weights for the loss terms, and $x'_m=T(x'_{m-1},\alpha_m)$ with $x'_0=x$. Additionally, we introduce a small Gaussian noise to α during training to further enhance the smoothed transform function. The above gradient update aims at leveraging the strengths of PIML and robust training to search for the practical worst-case scenarios induced by T_{APT}, where data samples are prone to misclassification due to their divergence from the original data. Familiarizing the model with these scenarios can effectively address microphone heterogeneity and improve accuracy on unseen microphones.

5.3.2 Frequency-aware Cost Function. In the previous section, the Wasserstein distance $W_1^{\mathbf{c}_\mathbf{X}}$, with the cost function (i.e., distance) $\mathbf{c}_\mathbf{X}$, is introduced to identify samples that are more likely to belong to the target domain. This objective focuses on training the model on samples that are likely to be misclassified due to their significant distance from the original samples. Thus, a straightforward approach to identifying such samples is to look at the cost function calculated by Euclidean distance between two samples. However, we observed that this objective fluctuates significantly, as demonstrated in Figure 5a. This can lead to gradient vanishing or gradient exploding, making the search for worst-case samples ineffective.

To address this problem, we propose a cost function, c_T, that uses an alternative distance that takes into account frequency information. Particularly, time-domain audio data is preprocessed by (DCT $\circ$ log $\circ$ STFT), where STFT is a short-time Fourier transform, log is the logarithm function, DCT is a discrete cosine transform which is linear, and $\circ$ denotes function composition. Thus, we have:

$$\begin{aligned}\mathbf{c}_\mathbf{X}(x',x)&=(\mathrm{DCT}\circ\log)\left((e^\alpha\otimes\mathbf{F})\otimes S_x\right)-(\mathrm{DCT}\circ\log)(S_x)\\&=(\mathrm{DCT}\circ\log)\left((e^\alpha\otimes\mathbf{F})\otimes S_x\oslash S_x\right)\\&=(\mathrm{DCT}\circ\log)\left(e^\alpha\otimes\mathbf{F}\right),\end{aligned}$$

Algorithm 1: Robust training with FA-DRO

Input: number of epoch N, M, source domain dataset $\mathcal{D}$, empty dataset $\hat{\mathcal{D}}$, pre-trained weight θ_0, noise $\epsilon \sim \mathcal{N}(0, \eta)$, epoch list $\mathcal{L}$, learning rate λ, ζ, weight loss τ, γ, β.
Output: learned weight θ_n
Init: $\theta \leftarrow \theta_0$, $\hat{\mathcal{D}} \leftarrow \mathcal{D}$
for $n = 1, \cdots, N$ **do**
  if $n \in \mathcal{L}$ **then**
    for $(x_i, y_i) \sim \mathcal{D}$ **do**
      $\alpha_1 \leftarrow \mathcal{N}(0, \eta)$
      for $m = 1, \cdots, M$ **do**
        $x_i^m \leftarrow T_{APT}(x, \alpha_m + \epsilon)$
        $\alpha_{m+1} \leftarrow \alpha_m + \nabla_\alpha \lambda + \left\{\tau \ell(\theta_n, (x_i^m, y_i)) - \gamma \mathbf{c_h}(x_i^m, x_i^{m-1}) + \frac{\beta}{2}(e^{\alpha_m}\mathbf{F} - 1)^2\right\}$
      end
    end
    Append (x_i^m, y) to $\hat{\mathcal{D}}$
  end
  for $(x_i, y_i) \sim \hat{\mathcal{D}}$ **do**
    $\theta_{n+1} \leftarrow \theta_n - \zeta \nabla_\theta \ell(\theta_n; (x_i, y_i))$
  end
end

where S_x is the vector obtained from the STFT, $x = (\text{DCT} \circ \log)(S_x)$, and $\oslash$ denotes the Hadamard division. To maximize this term, we need to maximize $e^\alpha \otimes \mathbf{F}$. Thus, the objective function is now given by $\sup \mathbf{c_T}(x, x') = \sup e^\alpha \mathbf{F}$. In this paper, we use an alternative formulation $\mathbf{c_T} = \frac{1}{2}(e^\alpha \mathbf{F} - 1)^2$. The cost function $\mathbf{c_T}$ is convex and lower semi-continuous, ensuring sufficient conditions for the DRO. As shown in Figure 5b, our proposed $\mathbf{c_T}$ has a smoother loss value and smaller error bound during optimization, thus mitigating the issues of gradient vanishing and gradient exploding. Overall, the loss function for the α parameter update is presented as follows.

$$\mathcal{L}(\alpha) = -\tau \ell(\theta, (x', y)) + \gamma \mathbf{c_h}(x', x) - \beta \mathbf{c_T}(x', x). \tag{7}$$

5.3.3 Overall Algorithm. The entire robust training pipeline is summarized in Algorithm 1. Our novelty is dashed underlined. Specifically, we train a deep neural network for N epochs on the dataset $\hat{\mathcal{D}}$, which is initialized by source domain dataset $\mathcal{D}$. This training uses standard loss function ℓ, e.g., cross-entropy for classification task. For epochs in the list $\mathcal{L}$, we update the parameter α by Equation 7, augment the samples with T_{APT}, and add those samples into the new training set $\hat{\mathcal{D}}$. To ensure robustness for certification purposes, we employ a randomized smoothing variant of APT by introducing a small Gaussian noise to the parameter α. It is important to note that while our robust training involves two objectives, the overall training time is moderate since M is small, as shown in Section 6.6. Furthermore, as demonstrated in Section 6.5, the training process achieves significantly faster convergence.

5.4 A Tighter Accuracy Drop Upper Bound

Finally, we address Issue 3 outlined in Section 4.2. The previous work [38] has shown that, when $d(\cdot, \cdot)$ in Definition 2 is Euclidean distance, the $\psi(x, \eta) = \text{erf}\left(\frac{x}{2\sqrt{2}\eta}\right)$ is a valid accuracy drop upper bound function.

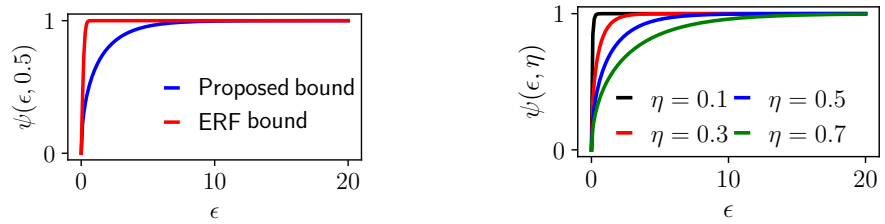


Fig. 6. Our proposed accuracy drop upper bound function. (a) Comparison with the previous erf function. (b) Under various levels of noisification intensity.

However, as shown in Figure 6a, this function with $\eta = 0.5$ quickly rises to 1, even when the distance between the two domains is small. This implies that the error function $\text{erf}(\cdot)$ only provides a loose bound. We propose a tighter bound function as stated by the following theorem.

Theorem 1. *$\psi(d(x_1, x_2); \eta) = \sqrt{1 - e^{\frac{-d(x_1,x_2)}{8\eta^2}}}$ is an accuracy drop upper bound function meeting the conditions set by Definition 2, where $d(\cdot, \cdot)$ is Euclidean distance.*

Proof can be found in the Appendix D. We demonstrate that this new ψ function yields a tighter bound compared with the one based on the error function $\text{erf}(\cdot)$. In Figure 6a, we compare the two functions with a noisification intensity of $\eta = 0.5$. The new ψ function is lower than the previous one, suggesting better tightness. Figure 6b shows the corresponding bounds for various η settings. For a larger η setting, the corresponding bound becomes tighter. This can be explained as follows. When the input values from the two domains cover a wider range, we have higher confidence in believing that the obtained model accuracy and the corresponding difference are reliable. In the rest of this paper, we use the tighter bound given by Theorem 1 for certification algorithm. Note that Theorem 1 is a general result not specific to acoustic sensing.

6 EXPERIMENTS

In this section, we evaluate the effectiveness of CertiAPT in addressing microphone heterogeneity. We select three application scenarios based on acoustic signals: keyword spotting (KWS), automated speech recognition (ASR), and acoustic-based room recognition (ARR), where KWS is widely deployed in real-world applications such as voice assistants, ARR is highly sensitive to environmental conditions and prone to domain shift-induced accuracy drop, and ASR involves larger models and more complex data, thus providing a more comprehensive evaluation of our approach. We use Honk [88], DeepSpeech2 [2], and an MLP as the classifier for KWS, ASR, and ARR, respectively. In the following part, we first compare CertiAPT's empirical accuracy with other methods designed to mitigate domain gaps, including evaluation experiments conducted under two real-world settings with seven different microphones. Then, we compare the proposed theoretical accuracy drop upper bound to assess the bound's tightness. Additionally, we evaluate the effect of incorporating robust training and the APT in the framework. Finally, we conduct a convergence analysis, a measurement of training time, and an ablation study to gain further insight into CertiAPT's accuracy.

■ **Datasets:** The KWS and ASR models are trained on 90% of the Google Speech Commands [101] and LibriSpeech [68], respectively. For testing, we use data collected from five different microphones on the remaining 10% subset of these two datasets [55]. For ARR, we use the acoustic room recognition dataset [55] recorded across 20 rooms using three different smartphones. We pick 80% data from one of the phones for training and test on the remaining 20% data from all the phones. The preprocessing step follows the approach outlined in [55] to ensure a fair

Table 1. Memory and runtime of the models in three tasks, measured on CPU.

Task	Model Size (MB)	Inference Time (ms)
ARR	2.74	0.65
KWS	1.88	38.49
ASR	79.16	645.69

Table 2. Baseline methods for comparision.

Method	Classical DA	Need target domain data	Need physical information	Robust training
SOT [53]	✓	✓		
DSAN [117]		✓		
CoVi [63]		✓		
CosMix [64]		✓		
SG-SCL [37]		✓		
BPA [78]		✓		
W-DRO [93]				✓
PhyAug [55]			✓	

comparison. For consistency, we use accuracy as the evaluation metric for all three tasks. Specifically for ASR, the accuracy is calculated as 100% − word error rate (WER).

■ **Implementations:** We present our hyper-parameter for each task in Appendix E. Our implementation is based on PyTorch and is executed on an Intel Xeon Gold 6246 CPU and two NVIDIA Quadro RTX 8000 GPUs.

■ **Baselines:** In our evaluation, we classify baselines by four criteria: 1) whether they rely on deep learning, 2) whether they require target domain data, 3) whether they are physics-informed, and 4) whether they use robust training techniques. The full list of the baselines is shown in the Table 2. Specifically, DSAN [117] is an unsupervised domain adaptation method, using sub-domain information to align the source and target domain. CoVi [63] is a supervised domain adaptation framework. CosMix [64] employs contrastive learning [11] combined with mixup [107] to address the challenge of limited target domain data. SG-SCL utilizes domain labels to perform contrastive learning for improved domain alignment. BPA [78] is a few-shot learning framework that leverages optimal transport theory to approximate the transform function between different domains. W-DRO [93] is a robust training approach that applies a Wasserstein constraint in the input space to identify worst-case scenarios. PhyAug [55] uses a pure FRC-based transform.

■ **Inference efficiency**: To justify the choices of the backbone models used in the three tasks, we report the inference efficiency on CPU in Table 1. The reported inference time is the average runtime over 10 runs on batches of 64 samples. These models are lightweight and suitable for deployment in real-world applications.

6.1 Effectiveness in Addressing Microphone Heterogeneity

This section presents the quantitative results of the three applications to show the proposed CertiAPT's effectiveness in addressing microphone heterogeneity.

6.1.1 Keyword Spotting. Table 3 shows that CertiAPT outperforms other baseline methods across individual microphones, achieving an average accuracy of 89.96%, which shows an improvement of 4.93% over the PhyAug on average. Note that with the help of robust training, CertiAPT can outperform methods such as DSAN, CoVi, BPA,

Table 3. Quantitative (accuracy) comparison among different testing microphones on KWS, where OD is the accuracy on the source domain testing dataset. The best result is underlined. The final column shows the proportion of target domain data used during training relative to the amount of source domain data.

Method	OD	Mic 1	Mic 2	Mic 3	Mic 4	Mic 5	Average	Target domain data ratio (%)
SOT	30.46	31.78	29.09	32.25	21.16	27.47	28.70	-
Honk	91.91	76.14	72.69	77.35	76.73	74.28	78.18	0.00
W-DRO	93.09	79.60	78.42	80.61	78.50	81.17	81.90	0.00
PhyAug	90.17	85.30	84.45	85.61	81.38	83.28	85.03	0.03
CoVi	91.63	84.77	83.79	85.18	83.87	84.38	85.60	50.0
BPA	89.88	87.59	86.71	88.38	83.94	84.62	86.85	50.0
SG-SCL	91.29	88.84	87.53	88.95	87.97	87.20	86.63	50.0
CosMix	90.62	86.62	85.17	86.95	86.11	85.64	86.85	50.0
DSAN	92.64	87.84	87.89	88.21	88.72	86.71	88.67	50.0
CertiAPT	94.53	90.47	88.43	90.21	87.73	88.39	89.96	0.03

Table 4. Quantitative (average accuracy) comparison among different smartphones on ARR. For each smartphone, the model is trained on the phone's training data and tested on the testing data of all smartphones.

Method	Phone 1	Phone 2	Phone 3	Avg	Target domain data ratio (%)
W-DRO	46.31	63.67	63.46	58.49	0.00
SOT	50.54	69.47	60.10	60.04	-
DSAN	71.35	66.24	69.96	66.85	100
SG-SCL	64.47	71.20	77.82	71.16	100
BPA	90.01	91.52	51.43	77.65	100
CoVi	80.20	77.96	79.14	79.10	100
PhyAug	81.01	78.70	86.98	82.23	0.04
CosMix	77.14	87.09	84.44	82.89	100
CertiAPT	84.67	80.67	91.33	85.56	0.04

SG-SCL, and CosMix, even when these methods are trained with extensive target domain data, while requiring only the FRC and no additional concrete target domain data. While the improvement over the second-best method, DSAN, is approximately 1.3%, CertiAPT demonstrates significantly better coverage performance, as detailed in Section 6.5.

6.1.2 Acoustic-based Room Recognition. Table 4 shows that CertiAPT achieves the highest overall accuracy of 85.56%. Even when additional target domain data comparable to the source domain data is used for training, methods such as DSAN, CoVi, BPA, SG-SCL, and CosMix are unable to surpass CertiAPT's performance. Among these methods, although BPA performs well on two smartphones, it exhibits extremely poor result on the remaining device, highlighting its lack of generalization. In contrast, PIML-based approaches, such as PhyAug and CertiAPT, consistently demonstrate strong accuracy without requiring additional concrete target domain data, as the incorporation of physical information provides greater robustness in this case.

Table 5. Quantitative (accuracy) comparison between different testing microphones on ASR.

Method	OD	Mic 1	Mic 2	Mic 3	Mic 4	Mic 5	Average
W-DRO	93.56	77.69	70.05	79.74	59.96	61.19	73.70
PhyAug	93.06	85.13	80.29	84.86	79.47	69.71	82.09
CertiAPT	93.17	85.61	82.13	86.77	79.81	75.24	83.79

Table 6. Quantitative comparison of microphone characteristics, including noise floor, signal-to-noise ratio (SNR), total harmonic distortion (THD), and dynamic range.

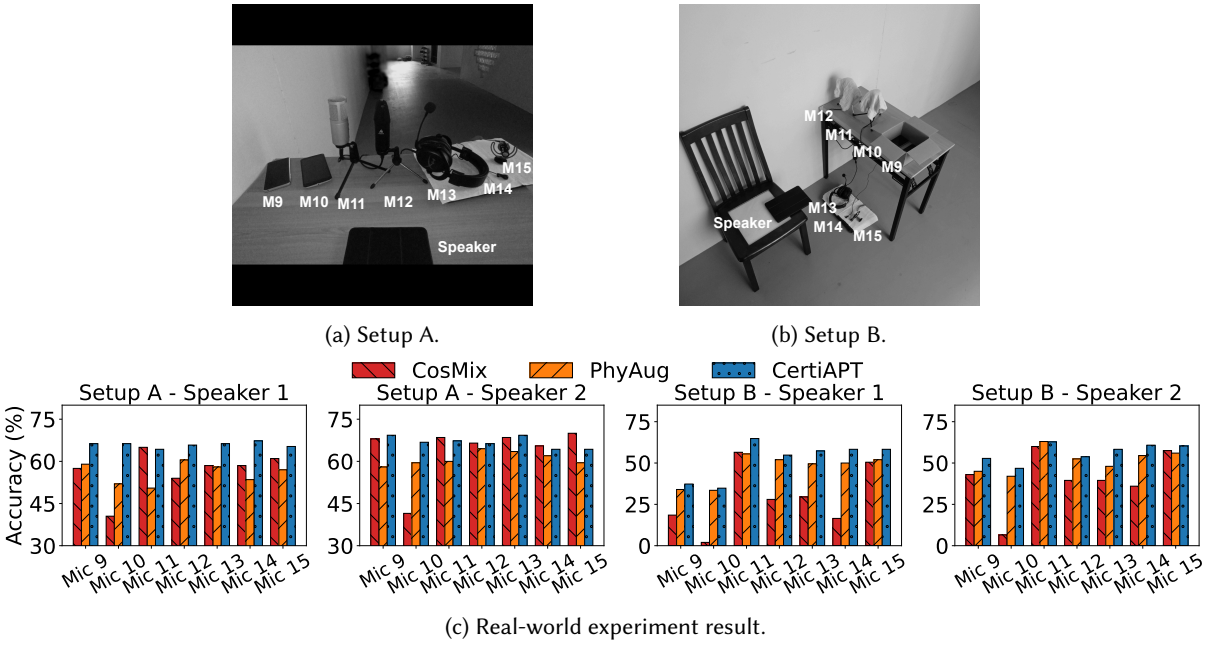
Microphone	Noise Floor (dB)	SNR (dB)	THD (dB)	Dynamic Range (dB)
M9	-44.8	38.2	-21.6	25.9
M10	-65.1	46.0	-11.7	22.3
M11	-72.9	44.8	-24.2	50.3
M12	-61.3	34.8	-9.7	21.1
M13	-59.3	51.1	-13.5	33.5
M14	-65.0	55.4	-25.8	25.5
M15	-70.0	39.4	-28.2	22.9

6.1.3 Automated Speech Recognition. Table 5 presents the accuracy results on the LibriSpeech dataset across various microphones. CertiAPT achieves the highest accuracy across all five test sets, each exhibiting a domain gap from the training data. We exclude the baselines that require the target domain data for training since such data is not available for this task.

6.1.4 Real-world Experiments. We place seven different microphones, denoted as M9 to M15, in two different environments, namely, Setup A and Setup B, for the KWS task. A comparison of the microphone characteristics is summarized in Table 6. Specifically, Setup A, as illustrated in Figure 7a, is positioned in a hallway, with a smartphone used as Speaker 1 and a laptop used as Speaker 2. Setup B, as illustrated in Figure 7b, varies the microphone placement by putting M9 and M10 in a box, covering M11 and M12 with cloth to simulate in-pocket placement, and putting M13 through M15 on the floor. As shown in Figure 7c, CertiAPT consistently outperforms both the primary baseline PhyAug [55] and CosMix [64]. Notably, in Setup B, where microphones are subjected to more severe distortions and obstructions, the performance of CosMix degrades significantly, with accuracy often falling below 40%. In contrast, CertiAPT maintains higher accuracy across all microphones and consistently outperforms CosMix and PhyAug under these challenging conditions. Our approach achieves approximately 60–70% accuracy under these challenging real-world conditions, demonstrating its practicality for deployment in dynamic environments.

6.2 Certified Accuracy against Microphone Heterogeneity

Figure 8 illustrates the certified accuracy (i.e. $\mathbb{E}_{(x_2,y_2)\sim\tilde{\mathcal{D}}}[\tilde{h}(x_2,y_2)] = \mathbb{E}_{(x_1,y_1)\sim\mathcal{D}}[\tilde{h}(x_1,y_1)] - \psi(\epsilon,\eta)$) on unseen target domain data as provided by CertiAPT using Lemma 1. For a given example distance ϵ between the source domain and the target domain, CertiAPT ensures a minimum accuracy with theoretical guarantees when evaluated on target domain data, without requiring direct access to or testing on concrete target domain data. This guarantee applies universally to all target domain data, provided that their distance from the source domain does not exceed ϵ. Under the same ϵ and η, our proposed ψ function achieves significantly higher certified accuracy compared with the erf() function from [38], especially under severe domain shift (i.e., larger ϵ). Notably, our ψ

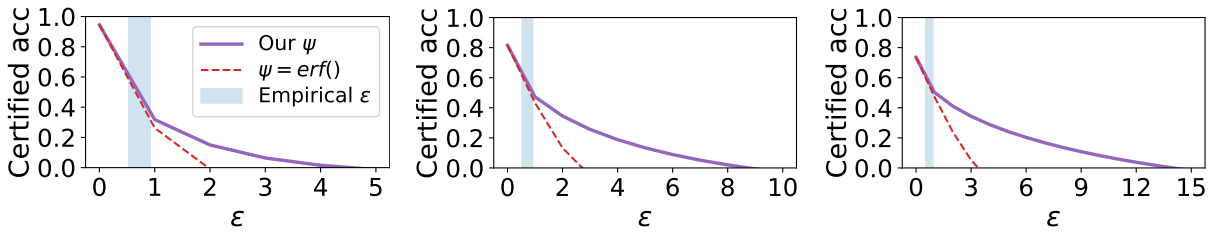


(a) Setup A.

(b) Setup B.

(c) Real-world experiment result.

Fig. 7. Real-world experiments of the KWS task using seven microphones (M9–M15). (a) Setup A, where the microphones are placed in a hallway, with a smartphone serving as Speaker 1 and a laptop as Speaker 2. (b) Setup B, where the placement of the microphones are varied. (c) Model accuracy evaluated on recordings from Setup A and Setup B.



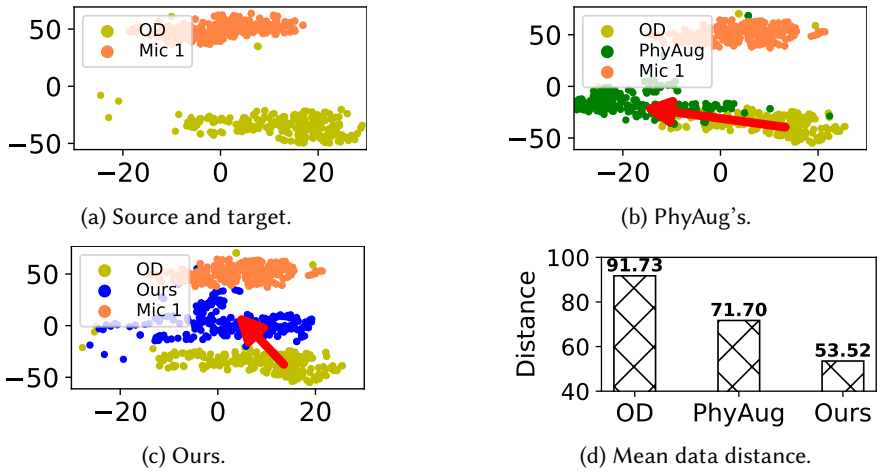
(a) $\eta = 0.5$.

(b) $\eta = 1.0$.

(c) $\eta = 1.5$.

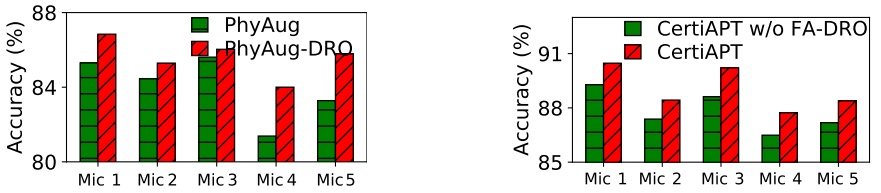
Fig. 8. Certified accuracy comparison between our proposed ψ and $\psi = \text{erf}()$ across various η values, evaluated on the KWS task. The model is trained on $\eta = 0.5$.

provides non-trivial certified accuracy up to $\epsilon = 15$, whereas erf() is limited to $\epsilon = 3$ for $\eta = 1.5$. In addition, we empirically estimate the Wasserstein distance ϵ between the source and target domain data using the available target domain samples, as illustrated by the blue regions in Figure 8. CertiAPT can certify the accuracy within the range of 40% to 60%, demonstrating a significant and meaningful level of certified robustness. Additional results for the ASR and ARR tasks are provided in Appendix F.



(a) Source and target. (b) PhyAug's. (c) Ours. (d) Mean data distance.

Fig. 9. t-SNE visualization of transformed data samples of PhyAug and ours. Figure 9d shows the mean L_2 distance between target data from Mic 1 to original data and to transformed data of PhyAug and ours.



(a) PhyAug (b) CertiAPT

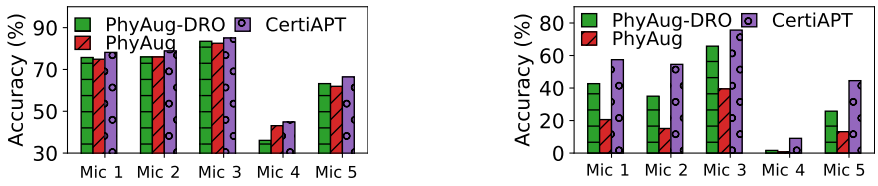
Fig. 10. Evaluate the effect of DRO on PhyAug and CertiAPT framework.

6.3 Effectiveness of APT

To demonstrate the effectiveness of our proposed transform function within the CertiAPT framework in reducing the domain gap, we plot a t-SNE visualization [92] based on the L_2 distance metric in Figure 9. Using a sample from a test set of the KWS task, we translate it to Microphone 1 using our proposed APT and compare it with the same sample collected directly by Microphone 1. Additionally, we apply the transform function from PhyAug for comparison. Our observations indicate that samples from PhyAug (green points) mostly overlap with the original data (yellow points). In contrast, data generated by CertiAPT (blue points) show a much closer alignment with the target domain of Microphone 1 (orange points) than PhyAug's approach. The mean distance across samples is shown in Figure 9d. This indicates that our transform function APT effectively minimizes the domain distance, thus improving the overall accuracy.

6.4 Necessity of Robust Training

■ **Effect of robust training:** In this section, we validate the effectiveness of robust training, specifically the use of DRO. Figure 10a compares the accuracy between PhyAug and PhyAug-DRO, a framework that applies the naive DRO after PhyAug's transform to further optimize the augmented sample. With the add-on robust training framework, PhyAug's accuracy on the KWS task improves by 1.38% on average. With the FA-DRO, CertiAPT's accuracy improves by 1.26% on average as shown in Figure 10b. These results underscore the effectiveness of



(a) $\sigma^2_{\text{noise}} = 0.5$. (b) $\sigma^2_{\text{noise}} = 1.0$.

Fig. 11. Robustness of approaches under noisy input with varying noise standard deviations σ^2_{noise}.

Table 7. Quantitative comparison of uncertainty set approaches on KWS, ASR, and ARR tasks, displayed from top to bottom.

Method	OD	Mic 1	Mic 2	Mic 3	Mic 4	Mic 5	Average
CertiAPT-Rand	93.93	89.28	87.38	88.62	86.49	87.18	88.81
CertiAPT-L_p	94.13	89.48	87.57	89.32	87.34	87.50	89.22
CertiAPT	94.53	90.47	88.43	90.21	87.73	88.39	89.96
CertiAPT-L_p	91.67	80.22	68.76	78.92	78.72	64.53	77.14
CertiAPT-Rand	92.86	83.63	78.17	84.04	74.83	68.83	80.39
CertiAPT	93.17	85.61	82.13	86.77	79.81	75.24	83.79
CertiAPT-Rand	-	80.20	77.42	82.89	-	-	80.17
CertiAPT-L_p	-	76.73	79.24	85.89	-	-	80.62
CertiAPT	-	84.67	80.67	91.33	-	-	85.56

robust training frameworks, such as DRO and FA-DRO, in enhancing model's performance by training the model on worst-case shift data.

Additionally, we evaluate the effect of robust training when dealing with noise. As shown in Figure 11, PhyAug-DRO outperforms PhyAug, particularly at higher noise levels (e.g., 1.0). Moreover, with the proposed FA-DRO, CertiAPT achieves the highest accuracy for all the testing cases on the two noise levels.

■ **Choice of uncertainty set:** Table 7 compares the accuracy of different uncertainty set choices for the DRO. These objectives continue to be employed for updating α in T_{APT}. CertiAPT-Rand is the framework that randomly chooses the uncertainty set from the whole space and CertiAPT-L_p is the framework that chooses the uncertainty set from the neighbor space with an L_p norm constraint. CertiAPT outperforms CertiAPT-L_p across three tasks, notably on ASR with an average improvement of around 5%. These results indicate that using a Wasserstein distance-based uncertainty set, grounded in Lemma 1, effectively address the domain shift problem.

■ **Effect of the cost function $\mathbf{c_T}$:** As discussed in Section 5.3.2, the cost function $\mathbf{c_X}$ in Equation 6 causes unstable loss value, which is later addressed by our proposed cost function $\mathbf{c_T}$. We validate this claim using Figure 12, which illustrates: FFT visualization of the original sample; FFT visualization after including $\mathbf{c_X}$, FFT visualization after including $\mathbf{c_T}$ on Equation 7, respectively; and FFT visualization of reference sample recorded in a real environment. Notably, the sample corresponding to $\mathbf{c_T}$ shows the closest resemblance to the target data, whereas the sample from $\mathbf{c_X}$ lacks several frequency bands, making it less realistic. This is due to $\mathbf{c_X}$ causing unstable gradients during the robust training with APT, as shown in Figure 5a, whereas $\mathbf{c_T}$ provides greater stability and accounts for frequency information during optimization. This frequency awareness is crucial as we aim to calculate the gradient of the APT. Additionally, as shown in Figure 12e, robust training with $\mathbf{c_X}$ leads to a 0.52% and 2.10% reduction in accuracy on KWS and ARR, compared to randomly initialized APT parameters. In

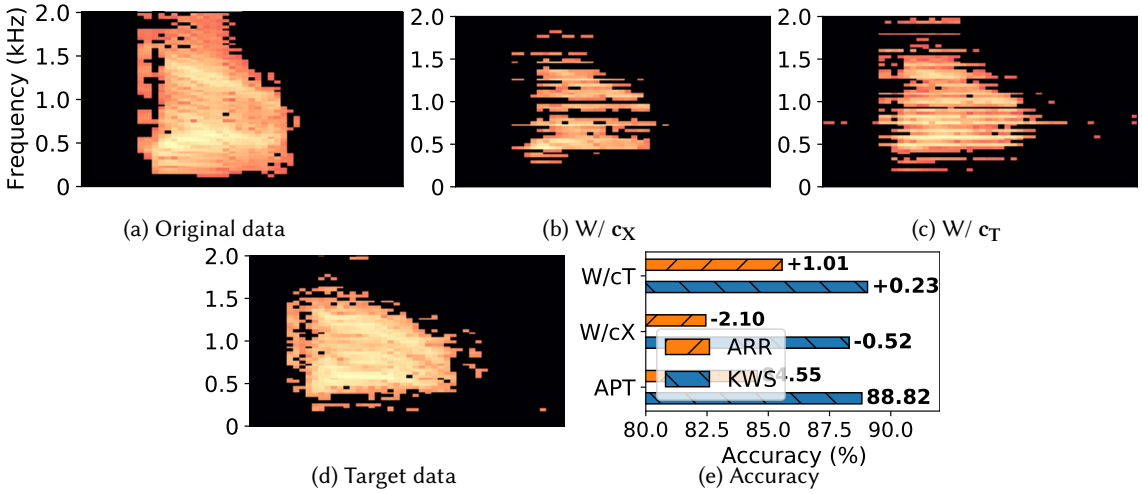


Fig. 12. Impact of the proposed cost function c_T on augmented FFT and its performance over the cost function c_X.

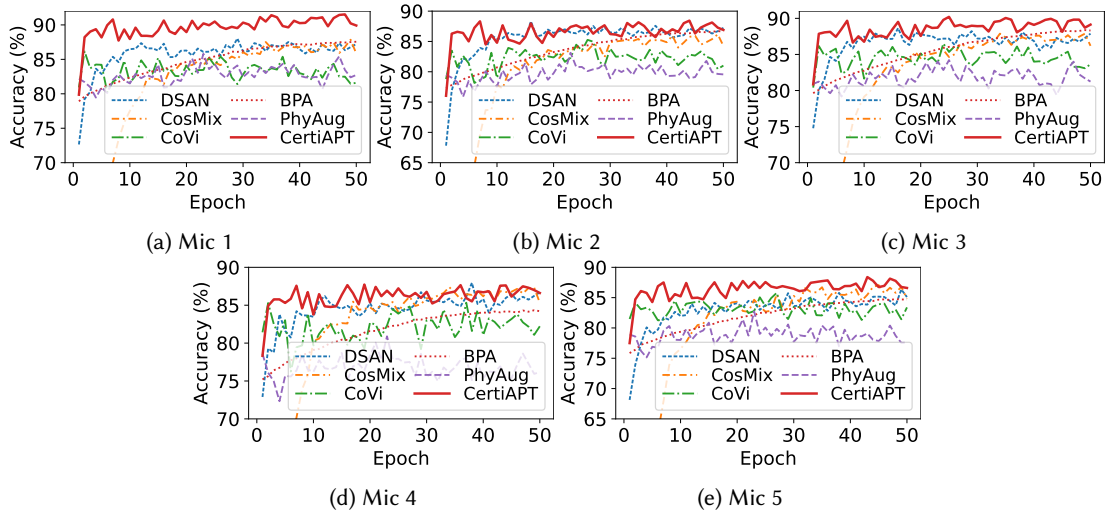
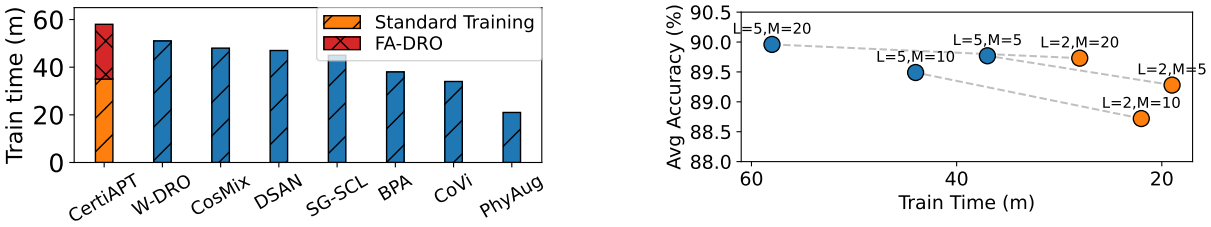


Fig. 13. Comparison of CertiAPT's convergence with other approaches, measured by test accuracy on target domain data collected from five microphones.

contrast, incorporating the stable loss term c_T during robust training results in samples that are more closely aligned with the target domain, thereby enhancing overall performance.

6.5 Convergence Performance

Figure 13 represents the convergence of CertiAPT compared with five other approaches. CertiAPT consistently demonstrates a smooth and rapid convergence in terms of accuracy as the training progresses. By around epoch 20, it achieves near-maximum accuracy, indicating a faster adaptation through robust training with physical



(a) Training time. (b) Impact of M and $\mathcal{L}$.

Fig. 14. Training overhead of CertiAPT on the KWS task. (a) Training times (in minutes) of various methods, presented in descending order. For CertiAPT, the red bar denotes the total time spent in the maximization phase (first stage), and the orange bar corresponds to the minimization phase (second stage) in Algorithm 1. (b) Impact of the number of maximization phases $\mathcal{L}$ and number of steps on each phase M on the training time and average accuracy.

information, even in the absence of target domain data. While PhyAug also integrates physical information, its convergence rate is similar to that of DSAN and CoVi. This highlights the effectiveness of CertiAPT design, founded on robust training with physical transformation and theoretical Wasserstein bound. Additionally, CertiAPT shows a sustained high accuracy over time. In contrast, methods like DSAN, CosMix, and CoVi, which heavily rely on target domain data for learning, require significantly more epochs to reach their peak accuracy levels and exhibit more variability in their accuracy during the early stages of training. Similarly, BPA, which must learn from multiple tasks in the target domain, converges slowly over time despite steady progress.

6.6 Computational Overhead

Figure 14a represents the training time of the KWS task on various methods. Standard training methods, such as PhyAug, complete training in approximately 20 minutes, while domain adaptation methods like CoVi, BPA, SG-SCL, DSAN, and CosMix require around 40 minutes. W-DRO takes slightly longer at approximately 50 minutes. CertiAPT, our proposed method, has the longest training time at 60 minutes, with the FA-DRO phase accounting for 40% of the total time and the remainder dedicated to training the main model. However, as demonstrated in Section 6.5, CertiAPT achieves significantly faster convergence compared to other methods, mitigating the impact of the increased training time.

In addition, the training time can be adjusted by modifying the hyperparameters of the FA-DRO, such as the number of inner maximization steps M, and the number of outer phases $\mathcal{L}$, to achieve a more efficient training process without compromising robustness. As shown in Figure 14b, while CertiAPT achieves its best performance using five maximization phases ($\mathcal{L} = 5$) with 20 steps each ($M = 20$), resulting in a total training time of 58 minutes on the KWS task, this cost can be significantly reduced. For example, by reducing the number of phases and steps, the training time can be brought down to approximately 20 minutes, which is comparable to ordinary training time, while incurring only an accuracy drop of less than 1%. This demonstrates that CertiAPT remains effective without introducing significant training overhead.

6.7 Ablation Study

In Figure 15, we evaluate the contribution of each component within the proposed framework. We begin with the baseline accuracy on the KWS task across five microphones, which achieves an accuracy of 84.00%, and the baseline accuracy for the ARR task is 82.23%. Introducing the simple robust training of DRO boosts the accuracy with a gain of 1.59% on KWS and 2.07% on ARR. Further improvements are observed when performing robust training on the proposed transformation APT, resulting in an accuracy increase of 3.23% on KWS and 0.25% on ARR, suggesting that the domain induced by APT helps the model to generalize better. Additionally, the

(a) KWS task. (b) ARR task.

Fig. 15. Ablation study of CertiAPT framework on KWS and ARR. The metric is average accuracy.

effectiveness of FA-DRO is validated as it improves accuracy by 0.23% and 1.01% on the KWS and ARR tasks, respectively.

7 DISCUSSION

■ **Limitation:** Despite demonstrating improvements in empirical results with theoretical guarantees, CertiAPT has certain limitations. First, the search for worst-case samples during the robust training process can increase the overall training time and impose a substantial memory burden, particularly when handling large datasets such as LibriSpeech [68]. Second, the certified accuracy drop upper bound remains relatively loose compared to the empirical results. In challenging scenarios, such as when the microphone records audio that is significantly different from the original data or is placed further away from the speaker, this upper bound can become trivial, i.e. ≥ 1.

■ **Future work:** Several promising research directions could be explored. First, optimizing the robust training process to reduce computational overhead and memory usage is crucial. Efficient data augmentation strategies [36], parallel processing, and memory-efficient algorithms could be investigated to address these challenges. Second, enhancing theoretical guarantees to provide tighter accuracy bound would be beneficial, especially in challenging conditions. This involves developing more robust certification methods and exploring alternative mathematical frameworks to improve bound tightness.

8 CONCLUSION

This paper presents CertiAPT, a robust framework designed to handle domain shifts from sensor heterogeneity in acoustic sensing applications with accuracy guarantees. Our approach incorporates a novel APT based on the FRC, enabling target-like transformations of source data without target samples. Unlike prior methods, CertiAPT provides certified robustness, reinforced by our robust training on APT. Extensive experiments demonstrate that CertiAPT outperforms in handling domain shifts and noise, while offering non-trivial theoretical accuracy bounds.

Acknowledgment

This research is supported by Singapore Ministry of Education under its AcRF Tier 1 grant RT14/22.

References

[1] [n. d.]. PhyAug Data. https://github.com/jiegev5/PhyAug.

[2] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*. PMLR.

[3] Zhenlin An, Qiongzheng Lin, Ping Li, and Lei Yang. 2020. General-purpose deep tracking platform across protocols for the internet of things. In *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*. 94–106.

[4] Riku Arakawa et al. 2023. Prism-tracker: A framework for multimodal procedure tracking using wearable sensors and state transition information with user-driven handling of errors and uncertainty. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–27.

[5] Abu Bakar, Rishabh Goel, Jasper De Winkel, Jason Huang, Saad Ahmed, Bashima Islam, Przemysław Pawełczak, Kasım Sinan Yıldırım, and Josiah Hester. 2022. Protean: An energy-efficient and heterogeneous platform for adaptive and hardware-accelerated battery-free computing. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*. 207–221.

[6] Manish Bansal, Kuo-Ling Huang, and Sanjay Mehrotra. 2018. Decomposition algorithms for two-stage distributionally robust mixed binary programs. *SIAM Journal on Optimization* (2018).

[7] Dimitris Bertsimas, Xuan Vinh Doan, Karthik Natarajan, and Chung-Piaw Teo. 2010. Models for minimax stochastic linear optimization problems with risk aversion. *Mathematics of Operations Research* (2010).

[8] Sejal Bhalla, Mayank Goel, and Rushil Khurana. 2021. Imu2doppler: Cross-modal domain adaptation for doppler-based activity recognition using imu data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–20.

[9] Jose Blanchet and Karthyek Murthy. 2019. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research* (2019).

[10] Youngjae Chang, Akhil Mathur, Anton Isopoussu, Junehwa Song, and Fahim Kawsar. 2020. A systematic study of unsupervised domain adaptation for robust human-activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–30.

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PmLR, 1597–1607.

[12] Zhe Chen, Tianyue Zheng, Chao Cai, and Jun Luo. 2021. MoVi-Fi: Motion-robust vital signs waveform recovery via deep interpreted RF sensing. In *Proceedings of the 27th annual international conference on mobile computing and networking*. 392–405.

[13] Hyunsung Cho, Akhil Mathur, and Fahim Kawsar. 2022. Flame: Federated learning across multi-device environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–29.

[14] Sungjae Cho, Yoonsu Kim, Jaewoong Jang, and Inseok Hwang. 2023. AI-to-Human Actuation: Boosting Unmodified AI's Robustness by Proactively Inducing Favorable Human Sensing Conditions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 1 (2023), 1–32.

[15] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. 2019. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*. PMLR.

[16] Francesco Croce and Matthias Hein. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*. PMLR.

[17] Shohreh Deldari, Daniel V Smith, Amin Sadri, and Flora Salim. 2020. Espresso: Entropy and shape aware time-series segmentation for processing heterogeneous sensor data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–24.

[18] Yongheng Deng, Weining Chen, Ju Ren, Feng Lyu, Yang Liu, Yunxin Liu, and Yaoxue Zhang. 2022. Tailorfl: Dual-personalized federated learning under system and data heterogeneity. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*. 592–606.

[19] Xiaoran Fan, Longfei Shangguan, Siddharth Rupavatharam, Yanyong Zhang, Jie Xiong, Yunfei Ma, and Richard Howard. 2021. HeadFi: bringing intelligence to all headphones. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 147–159.

[20] Allen-Jasmin Farcas, Myungjin Lee, Ramana Rao Kompella, Hugo Latapie, Gustavo De Veciana, and Radu Marculescu. 2023. MOHAWK: Mobility and Heterogeneity-Aware Dynamic Community Selection for Hierarchical Federated Learning. In *Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation*. 249–261.

[21] Chao Feng, Nan Wang, Yicheng Jiang, Xia Zheng, Kang Li, Zheng Wang, and Xiaojiang Chen. 2022. Wi-learner: Towards one-shot learning for cross-domain wi-fi based gesture recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–27.

[22] Rui Gao, Xi Chen, and Anton J Kleywegt. 2017. Wasserstein Distributionally Robust Optimization and Variation Regularization. *arXiv preprint arXiv:1712.06050* (2017).

[23] Ruiyang Gao, Wenwei Li, Yaxiong Xie, Enze Yi, Leye Wang, Dan Wu, and Daqing Zhang. 2022. Towards robust gesture recognition by characterizing the sensing quality of WiFi signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–26.

[24] Yi Gao, Jiadong Zhang, Gaoyang Guan, and Wei Dong. 2020. LinkLab: A scalable and heterogeneous testbed for remotely developing and experimenting IoT applications. In *2020 IEEE/ACM Fifth International Conference on Internet-of-Things Design and Implementation (IoTDI)*. IEEE, 176–188.

[25] Gines Garcia-Aviles, Andres Garcia-Saavedra, Marco Gramaglia, Xavier Costa-Perez, Pablo Serrano, and Albert Banchs. 2021. Nuberu: Reliable RAN virtualization in shared platforms. In *Proceedings of the 27th Annual International Conference on Mobile Computing and*

Networking. 749–761.

[26] Taesik Gong, Yewon Kim, Adiba Orzikulova, Yunxin Liu, Sung Ju Hwang, Jinwoo Shin, and Sung-Ju Lee. 2023. DAPPER: Label-Free Performance Estimation after Personalization for Heterogeneous Mobile Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 2 (2023), 1–27.

[27] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples.

[28] Yujiao Hao, Boyu Wang, and Rong Zheng. 2023. VALERIAN: Invariant Feature Learning for IMU Sensor-based Human Activity Recognition in the Wild. In *Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation*. 66–78.

[29] Zhongkai Hao, Songming Liu, Yichi Zhang, Chengyang Ying, Yao Feng, Hang Su, and Jun Zhu. 2023. Physics-Informed Machine Learning: A Survey on Problems, Methods and Applications. arXiv:2211.08064 [cs.LG] https://arxiv.org/abs/2211.08064

[30] Alexander Hoelzemann and Kristof Van Laerhoven. 2020. Digging deeper: Towards a better understanding of transfer learning for human activity recognition. In *Proceedings of the 2020 ACM International Symposium on Wearable Computers*. 50–54.

[31] Zhizhang Hu, Yue Zhang, Tong Yu, and Shijia Pan. 2022. VMA: Domain Variance-and Modality-Aware Model Transfer for Fine-Grained Occupant Activity Recognition. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 259–270.

[32] Kai Huang and Wei Gao. 2022. Real-time neural network inference on extremely weak devices: agile offloading with explainable AI. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. 200–213.

[33] Joo Seong Jeong, Jingyu Lee, Donghyun Kim, Changmin Jeon, Changjin Jeong, Youngki Lee, and Byung-Gon Chun. 2022. Band: coordinated multi-dnn inference on heterogeneous mobile processors. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*. 235–247.

[34] Ning Jia, Weiguo Huang, Chuancang Ding, Jun Wang, and Zhongkui Zhu. 2024. Physics-informed unsupervised domain adaptation framework for cross-machine bearing fault diagnosis. *Advanced Engineering Informatics* (2024).

[35] Jinhwan Jung, Jihoon Ryoo, Yung Yi, and Song Min Kim. 2020. Gateway over the air: Towards pervasive internet connectivity for commodity iot. In *Proceedings of the 18th international conference on mobile systems, applications, and services*. 54–66.

[36] Gwantae Kim, David K. Han, and Hanseok Ko. 2021. SpecMix: A Mixed Sample Data Augmentation method for Training withTime-Frequency Domain Features.

[37] June-Woo Kim, Sangmin Bae, Won-Yang Cho, Byungjo Lee, and Ho-Young Jung. 2024. Stethoscope-guided supervised contrastive learning for cross-domain adaptation on respiratory sound classification. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1431–1435.

[38] Aounon Kumar, Alexander Levine, Tom Goldstein, and Soheil Feizi. 2023. Provable Robustness against Wasserstein Distribution Shifts via Input Randomization. In *The Eleventh International Conference on Learning Representations*.

[39] Jaeheon Kwak, Sunjae Lee, Dae R Jeong, Arjun Kumar, Dongjae Shin, Ilju Kim, Donghwa Shin, Kilho Lee, Jinkyu Lee, and Insik Shin. 2023. MixMax: Leveraging Heterogeneous Batteries to Alleviate Low Battery Experience for Mobile Users. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*. 247–260.

[40] Young D Kwon, Jagmohan Chauhan, Hong Jia, Stylianos I Venieris, and Cecilia Mascolo. 2023. LifeLearner: Hardware-Aware Meta Continual Learning System for Embedded Computing Platforms. In *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*. 138–151.

[41] Young D Kwon, Jagmohan Chauhan, and Cecilia Mascolo. 2022. Yono: Modeling multiple heterogeneous neural networks on microcontrollers. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 285–297.

[42] Sunjae Lee, Hayeon Lee, Hoyoung Kim, Sangmin Lee, Jeong Woon Choi, Yuseung Lee, Seono Lee, Ahyeon Kim, Jean Young Song, Sangeun Oh, et al. 2021. FLUID-XP: Flexible user interface distribution for cross-platform experience. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 762–774.

[43] Ang Li, Jingwei Sun, Pengcheng Li, Yu Pu, Hai Li, and Yiran Chen. 2021. Hermes: an efficient federated learning framework for heterogeneous mobile clients. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 420–437.

[44] Ang Li, Jingwei Sun, Xiao Zeng, Mi Zhang, Hai Li, and Yiran Chen. 2021. Fedmask: Joint computation and communication-efficient personalized federated learning via heterogeneous masking. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 42–55.

[45] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. 2018. Domain Generalization with Adversarial Feature Learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[46] Linyi Li, Tao Xie, and Bo Li. 2023. Sok: Certified robustness for deep neural networks. In *2023 IEEE symposium on security and privacy (SP)*. IEEE.

[47] Zimo Liao, Zhicheng Luo, Qianyi Huang, Linfeng Zhang, Fan Wu, Qian Zhang, and Yi Wang. 2021. SMART: screen-based gesture recognition on commodity mobile devices. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 283–295.

[48] Chi Lin, Yongda Yu, Jie Xiong, Yichuan Zhang, Lei Wang, Guowei Wu, and Zhongxuan Luo. 2021. Shrimp: a robust underwater visible light communication system. In *Proceedings of the 27th annual international conference on mobile computing and networking*. 134–146.

[49] Haipeng Liu, Kening Cui, Kaiyuan Hu, Yuheng Wang, Anfu Zhou, Liang Liu, and Huadong Ma. 2022. mTransSee: Enabling environment-independent mmWave sensing based gesture recognition via transfer learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–28.

[50] Tony Liu, Jennifer Nicholas, Max M Theilig, Sharath C Guntuku, Konrad Kording, David C Mohr, and Lyle Ungar. 2019. Machine learning for phone-based relationship estimation: the need to consider population heterogeneity. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 3, 4 (2019), 1–23.

[51] Xiaofeng Liu, Chaehwa Yoo, Fangxu Xing, Hyejin Oh, Georges El Fakhri, Je-Won Kang, Jonghye Woo, et al. 2022. Deep unsupervised domain adaptation: A review of recent advances and perspectives. *APSIPA Transactions on Signal and Information Processing* (2022).

[52] Yilin Liu, Shijia Zhang, and Mahanth Gowda. 2021. When video meets inertial sensors: Zero-shot domain adaptation for finger motion analytics with inertial sensors. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation*. 182–194.

[53] Wang Lu, Yiqiang Chen, Jindong Wang, and Xin Qin. 2021. Cross-domain activity recognition via substructural optimal transport. *Neurocomputing* (2021).

[54] Wang Lu, Jindong Wang, Yiqiang Chen, Sinno Jialin Pan, Chunyu Hu, and Xin Qin. 2022. Semantic-discriminative mixup for generalizable sensor-based cross-domain activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–19.

[55] Wenjie Luo, Zhenyu Yan, Qun Song, and Rui Tan. 2021. PhyAug: Physics-Directed Data Augmentation for Deep Sensing Model Transfer in Cyber-Physical Systems. In *International Conference on Information Processing in Sensor Networks*.

[56] Wenjun Lyu, Guang Wang, Yu Yang, and Desheng Zhang. 2021. Mover: Generalizability Verification of Human Mobility Models via Heterogeneous Use Cases. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–21.

[57] Aleksander Mądry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *stat* (2017).

[58] Akhil Mathur, Anton Isopoussu, Fahim Kawsar, Nadia Berthouze, and Nicholas D. Lane. 2019. Mic2Mic: using cycle-consistent generative adversarial networks to overcome microphone variability in speech systems. In *International Conference on Information Processing in Sensor Networks (IPSN '19)*.

[59] Lakmal Meegahapola, William Droz, Peter Kun, Amalia De Götzen, Chaitanya Nutakki, Shyam Diwakar, Salvador Ruiz Correa, Donglei Song, Hao Xu, Miriam Bidoglia, et al. 2023. Generalization and personalization of mobile sensing-based mood inference models: an analysis of college students in eight countries. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 6, 4 (2023), 1–32.

[60] Apolline Mellot, Antoine Collas, Sylvain Chevallier, Denis Engemann, and Alexandre Gramfort. 2024. Physics-informed and Unsupervised Riemannian Domain Adaptation for Machine Learning on Heterogeneous EEG Datasets. arXiv:2403.15415 [eess.SP] https://arxiv.org/abs/2403.15415

[61] Shenghuan Miao, Ling Chen, Rong Hu, and Yingsong Luo. 2022. Towards a dynamic inter-sensor correlations learning framework for multi-sensor-based wearable human activity recognition. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 6, 3 (2022), 1–25.

[62] Saeid Motiian, Quinn Jones, Seyed Iranmanesh, and Gianfranco Doretto. 2017. Few-shot adversarial domain adaptation. *Advances in neural information processing systems* (2017).

[63] Jaemin Na, Dongyoon Han, Hyung Jin Chang, and Wonjun Hwang. 2022. Contrastive vicinal space for unsupervised domain adaptation. In *European Conference on Computer Vision*. Springer.

[64] Dianwen Ng, Ruixi Zhang, Jia Qi Yip, Chong Zhang, Yukun Ma, Trung Hieu Nguyen, Chongjia Ni, Eng Siong Chng, and Bin Ma. 2023. Contrastive speech mixup for low-resource keyword spotting. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[65] A. Tuan Nguyen, Toan Tran, Yarin Gal, Philip H. S. Torr, and Atilim Günes Baydin. 2021. KL Guided Domain Adaptation. *CoRR* (2021).

[66] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. 2021. Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

[67] Xiaomin Ouyang, Zhiyuan Xie, Heming Fu, Sitong Cheng, Li Pan, Neiwen Ling, Guoliang Xing, Jiayu Zhou, and Jianwei Huang. 2023. Harmony: Heterogeneous multi-modal federated learning through disentangled model training. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*. 530–543.

[68] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE.

[69] Weiwu Pang, Chunyu Xia, Branden Leong, Fawad Ahmad, Jeongyeup Paek, and Ramesh Govindan. 2023. UbiPose: Towards Ubiquitous Outdoor AR Pose Tracking using Aerial Meshes. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. 1–16.

[70] Alejandro Blanco Pizarro, Joan Palacios Beltrán, Marco Cominelli, Francesco Gringoli, and Joerg Widmer. 2021. Accurate ubiquitous localization with off-the-shelf IEEE 802.11 ac devices. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. 241–254.

[71] Xin Qin, Yiqiang Chen, Jindong Wang, and Chaohui Yu. 2019. Cross-dataset activity recognition via adaptive spatial-temporal transfer learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (2019), 1–25.
[72] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org.
[73] Hamed Rahimian, Güzin Bayraksan, and Tito Homem-de Mello. 2019. Identifying effective scenarios in distributionally robust stochastic programs with total variation distance. *Mathematical Programming* (2019).
[74] Hamed Rahimian and Sanjay Mehrotra. 2022. Frameworks and Results in Distributionally Robust Optimization. *Open Journal of Mathematical Optimization* (2022).
[75] Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. 2019. Provably robust deep learning via adversarially trained smoothed classifiers. *Advances in neural information processing systems* (2019).
[76] Sandeep Singh Sandha, Joseph Noor, Fatima M Anwar, and Mani Srivastava. 2020. Time awareness in deep learning-based multimodal fusion across smartphone platforms. In *2020 IEEE/ACM Fifth International Conference on Internet-of-Things Design and Implementation (IoTDI)*. IEEE, 149–156.
[77] Panneer Selvam Santhalingam, Al Amin Hosain, Ding Zhang, Parth Pathak, Huzefa Rangwala, and Raja Kushalnagar. 2020. mmASL: Environment-independent ASL gesture recognition using 60 GHz millimeter-wave signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–30.
[78] Daniel Shalam and Simon Korman. 2024. The Balanced-Pairwise-Affinities Feature Transform. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria*. OpenReview.net.
[79] Leming Shen, Qiang Yang, Kaiyan Cui, Yuanqing Zheng, Xiao-Yong Wei, Jianwei Liu, and Jinsong Han. 2024. FedConv: A Learning-on-Model Paradigm for Heterogeneous Federated Clients. In *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services*. 398–411.
[80] Taoran Sheng and Manfred Huber. 2020. Weakly supervised multi-task representation learning for human activity analysis using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (2020), 1–18.
[81] Jaemin Shin, Yuanchun Li, Yunxin Liu, and Sung-Ju Lee. 2022. Fedbalancer: Data and pace control for efficient federated learning on heterogeneous clients. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*. 436–449.
[82] Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. 2020. Certifying Some Distributional Robustness with Principled Adversarial Training. arXiv:1710.10571 [stat.ML] https://arxiv.org/abs/1710.10571
[83] Yisheng Song, Ting Wang, Puyu Cai, Subrota K Mondal, and Jyoti Prakash Sahoo. 2023. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *Comput. Surveys* (2023).
[84] Jannis Strecker, Khakim Akhunov, Federico Carbone, Kimberly García, Kenan Bektaş, Andres Gomez, Simon Mayer, and Kasim Sinan Yildirim. 2023. MR Object Identification and Interaction: Fusing Object Situation Information from Heterogeneous Sources. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–26.
[85] Jingwei Sun, Ang Li, Lin Duan, Samiul Alam, Xuliang Deng, Xin Guo, Haiming Wang, Maria Gorlatova, Mi Zhang, Hai Li, et al. 2022. FedSEA: A semi-asynchronous federated learning framework for extremely heterogeneous devices. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*. 106–119.
[86] Mahan Tabatabaie, Suining He, and Kang G Shin. 2023. Cross-Modality Graph-Based Language and Sensor Data Co-Learning of Human-Mobility Interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–25.
[87] Rui Tan and Wenjie Luo. 2023. Physics-Informed Machine Learning Model Generalization in AIoT: Opportunites and Challenges. In *Proceedings of Cyber-Physical Systems and Internet of Things Week 2023*. Association for Computing Machinery.
[88] Raphael Tang and Jimmy Lin. 2017. Honk: A PyTorch Reimplementation of Convolutional Neural Networks for Keyword Spotting.
[89] Catherine Tong, Jinchen Ge, and Nicholas D Lane. 2021. Zero-shot learning for imu-based activity recognition using video embeddings. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–23.
[90] Vu Tran, Gihan Jayatilaka, Ashwin Ashok, and Archan Misra. 2021. DeepLight: Robust & unobtrusive real-time screen-camera communication for real-world displays. In *Proceedings of the 20th International Conference on Information Processing in Sensor Networks (co-located with CPS-IoT Week 2021)*. 238–253.
[91] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. 2014. Deep Domain Confusion: Maximizing for Domain Invariance. *CoRR* (2014).
[92] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
[93] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. 2018. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems* (2018).
[94] Chao Wang, Christopher Gill, and Chenyang Lu. 2020. Adaptive data replication in real-time reliable edge computing for internet of things. In *2020 IEEE/ACM fifth international conference on internet-of-things design and implementation (IoTDI)*. IEEE, 128–134.
[95] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *International Conference on Learning Representations*.

[96] Haoyu Wang, Jiazhao Wang, Demin Gao, and Wenchao Jiang. 2024. NNCTC: Physical Layer Cross-Technology Communication via Neural Networks. *arXiv preprint arXiv:2403.10014* (2024).
[97] Jike Wang, Shanmu Wang, Yasha Iravantchi, Mingke Wang, Alanson Sample, Kang G Shin, Xinbing Wang, Chenghu Zhou, and Dongyao Chen. 2023. METRO: Magnetic Road Markings for All-weather, Smart Roads. In *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*. 280–293.
[98] Jiankun Wang, Zenghua Zhao, Mengling Ou, Jiayang Cui, and Bin Wu. 2023. Automatic update for wi-fi fingerprinting indoor localization via multi-target domain adaptation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 2 (2023), 1–27.
[99] Mingke Wang, Qing Luo, Yasha Iravantchi, Xiaomeng Chen, Alanson Sample, Kang G Shin, Xiaohua Tian, Xinbing Wang, and Dongyao Chen. 2022. Automatic calibration of magnetic tracking. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. 391–404.
[100] Qianru Wang, Bin Guo, Lu Cheng, and Zhiwen Yu. 2023. sUrban: Stable Prediction for Unseen Urban Data from Location-based Sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–20.
[101] Pete Warden. 2018. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition.
[102] Jianyu Wei, Ting Cao, Shijie Cao, Shiqi Jiang, Shaowei Fu, Mao Yang, Yanyong Zhang, and Yunxin Liu. 2023. Nn-stretch: Automatic neural network branching for parallel inference on heterogeneous multi-processors. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*. 70–83.
[103] Hao Wen, Yuanchun Li, Zunshuai Zhang, Shiqi Jiang, Xiaozhou Ye, Ye Ouyang, Yaqin Zhang, and Yunxin Liu. 2023. Adaptivenet: Post-deployment neural architecture adaptation for diverse edge environments. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. 1–17.
[104] Huatao Xu, Pengfei Zhou, Rui Tan, and Mo Li. 2023. Practically Adopting Human Activity Recognition. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*.
[105] Jialiang Yan, Siyao Cheng, Zhijun Li, and Jie Liu. 2022. PCTC: Parallel cross technology communication in heterogeneous wireless systems. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE.
[106] Shuangjiao Zhai, Zhanyong Tang, Petteri Nurmi, Dingyi Fang, Xiaojiang Chen, and Zheng Wang. 2021. RISE: Robust wireless sensing using probabilistic and statistical assessments. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 309–322.
[107] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* (2017).
[108] Jin Zhang, Zhuangzhuang Chen, Chengwen Luo, Bo Wei, Salil S Kanhere, and Jianqiang Li. 2022. MetaGanFi: Cross-domain unseen individual identification using WiFi signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–21.
[109] Li Lyna Zhang, Shihao Han, Jianyu Wei, Ningxin Zheng, Ting Cao, Yuqing Yang, and Yunxin Liu. 2021. Nn-meter: Towards accurate latency prediction of deep-learning model inference on diverse edge devices. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. 81–93.
[110] Qingzhao Zhang, Xumiao Zhang, Ruiyang Zhu, Fan Bai, Mohammad Naserian, and Z Morley Mao. 2023. Robust real-time multi-vehicle collaboration on asynchronous sensors. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. 1–15.
[111] Wenyu Zhang, Li Shen, Wanyue Zhang, and Chuan-Sheng Foo. 2022. Few-Shot Adaptation of Pre-Trained Networks for Domain Shift. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*. International Joint Conferences on Artificial Intelligence Organization.
[112] Xiyuan Zhang, Xiaohan Fu, Diyan Teng, Chengyu Dong, Keerthivasan Vijayakumar, Jiayun Zhang, Ranak Roy Chowdhury, Junsheng Han, Dezhi Hong, Rashmi Kulkarni, et al. 2023. Physics-Informed Data Denoising for Real-Life Sensing Systems. In *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*.
[113] Zicheng Zhang, Yinglu Liu, Congying Han, Tiande Guo, Ting Yao, and Tao Mei. 2022. Generalized one-shot domain adaptation of generative adversarial networks. *Advances in Neural Information processing systems* (2022).
[114] An Zhao, Mingyu Ding, Zhiwu Lu, Tao Xiang, Yulei Niu, Jiechao Guan, and Ji-Rong Wen. 2021. Domain-Adaptive Few-Shot Learning. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*.
[115] Tianyue Zheng, Ang Li, Zhe Chen, Hongbo Wang, and Jun Luo. 2023. Autofed: Heterogeneity-aware federated multimodal learning for robust autonomous driving. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. 1–15.
[116] Zihao Zhou, Aihua Ran, Shuxiao Chen, Guodan Wei, Hongbin Sun, Xuan Zhang, and Yang Li. 2021. Few-shot cross domain battery capacity estimation. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*. 703–711.
[117] Yongchun Zhu, Fuzhen Zhuang, Jindong Wang, Guolin Ke, Jingwu Chen, Jiang Bian, Hui Xiong, and Qing He. 2021. Deep Subdomain Adaptation Network for Image Classification. *IEEE Transactions on Neural Networks and Learning Systems* (2021).

Table 8. Heterogeneity papers presented at sensing-related conferences in 2020 to 2024.

Venue	Papers addressing heterogeneity related topics	Papers providing certification
IMWUT/UbiComp	[8, 10, 13, 17, 23, 30, 50, 56, 71, 77, 80, 89, 116] [4, 14, 21, 26, 49, 54, 59, 61, 84, 86, 98, 100, 108]	-
SenSys	[5, 18, 40, 44, 85, 97]	-
IPSN	[31, 41, 55, 90, 96, 105]	[105]
IoTDI	[20, 24, 28, 52, 76, 94]	-
MobiCom	[12, 19, 25, 32, 42, 43, 47, 48, 99, 106, 115] [69, 103, 104, 110]	-
MobiSys	[3, 33, 35, 39, 67, 70, 79, 81, 102, 109]	-
Total	69	1

Table 9. Summary of Notations

Notation	Description
$\mathcal{D}$	Source domain distribution
$\tilde{\mathcal{D}}$	Target domain distribution
$T(\cdot, \alpha)$	Transformation function with parameter α
ψ	Accuracy drop upper bound function
η	Standard deviation of noise in certified robustness
$\mathcal{U}$	Set of distributions in robust training
h	Classifier
$\tilde{h}$	Smoothed classifier
c_h	Cost function in the embedding space
c_X	Cost function in the input space
θ	The parameters of the classifier
ℓ	Loss function of the classifier
$\mathcal{L}(\alpha)$	Loss function of the transformation function
$\mathcal{N}(x, \eta)$	Normal distribution with mean x and standard deviation η
$\mathbf{F}$	Frequency response curve (FRC)

[118] Fuzhen Zhuang, Xiaohu Cheng, Ping Luo, Sinno Jialin Pan, and Qing He. 2015. Supervised Representation Learning: Transfer Learning with Deep Autoencoders. In *International Joint Conference on Artificial Intelligence, IJCAI.*

A Survey of Papers on Domain Shifts

We surveyed papers presented at IMWUT/UbiComp, SenSys, IPSN, IoTDI, MobiCom, and MobiSys from 2020 to 2024, focusing on topics such as domain shift, sensor heterogeneity, and cross-platform implementation. The survey results are presented in Table 8. Out of 69 relevant papers, only one paper [105] includes performance certification.

B Notations

The notations used throughout this paper are summarized in Table 9.

C Distributionally Robust Optimization (DRO)

This framework includes two steps. First, given $\mathcal{U} = \{K | K \in \mathcal{D}_{\mathcal{T}}, W_1^{\mathbf{c_h}}(\mathcal{D}, K) \leq a, W_1^{\mathbf{c_X}}(\mathcal{D}, K) \geq b\}$, it uses Lagrangian relaxation with two Lagrange multipliers γ, β to make Equation 4 more tractable, since the upper bound a and lower bound b of the distribution balls in $\mathcal{U}$ are arbitrary.

$$\min_{\theta \in \Theta} \sup_{K} \mathbb{E}_{(x,y)\sim K}[\ell(\theta, (x, y))] - \gamma W_1^{\mathbf{c_h}}(\mathcal{D}, K) + \beta W_1^{\mathbf{c_X}}(\mathcal{D}, K). \tag{8}$$

The problem still involves a maximization over distributions K, requiring sampling lots of K until converging, which is challenging due to the slow convergence and complexity of directly optimizing over an infinite-dimensional space of distributions. Thus, the second step of the DRO is to simplify the maximization over distributions K with a more tractable optimization problem over finite-dimensional Lagrange multipliers, which ensures the problem remains computationally feasible. This can be done through Theorem 1 in [9], given that the cost functions $\mathbf{c_h}$ and $\mathbf{c_X}$ are non-negative, lower semi-continuous, and satisfy $\mathbf{c_h}(x', x) = 0$ and $\mathbf{c_X}(x', x) = 0$ when $x = x'$. Proposition 1 in [82] gives similar result given $\mathbf{c_h}$ and $\mathbf{c_X}$ are convex and continuous. The compact notation of Equation 8 is expressed as follows:

$$\min_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim \mathcal{D}} \left[\sup_{x' = T_{APT}(x,\alpha)} \ell(\theta, (x', y)) - \gamma \mathbf{c_h}(x', x) + \beta \mathbf{c_X}(x', x) \right]. \tag{9}$$

D Proof of Theorem 1

Proof. Given two distributions $P \sim \mathcal{N}(x_1, \eta^2)$ and $Q \sim \mathcal{N}(x_2, \eta^2)$, we denote $f(x)$, $g(x)$ as the probability density functions of P and Q. First, we prove $\int \sqrt{f(x)g(x)}dx = e^{\frac{-d(x_1,x_2)}{8\eta^2}}$.

$$\begin{aligned}
\int \sqrt{f(x)g(x)}dx &= \frac{1}{\sqrt{2\pi\eta^2}} \int e^{\frac{(x-x_1)^2+(x-x_2)^2}{-4\eta^2}} dx \\
&= \frac{1}{\sqrt{2\pi\eta^2}} \int e^{(\frac{(x-\frac{x_1+x_2}{2})^2}{-2\eta^2} - \frac{(x_1-x_2)^2}{8\eta^2})} dx \\
&= \frac{1}{\sqrt{2\pi\eta^2}} e^{-\frac{(x_1-x_2)^2}{8\eta^2}} \int e^{\frac{(x-\frac{x_1+x_2}{2})^2}{-2\eta^2}} dx \\
&= \frac{1}{\sqrt{2\pi\eta^2}} e^{-\frac{(x_1-x_2)^2}{8\eta^2}} \sqrt{2\pi\eta^2}.
\end{aligned}$$

Given $d(x_1, x_2) = (x_1 - x_2)^2$, we obtain:

$$\int \sqrt{f(x)g(x)}dx = e^{\frac{-d(x_1,x_2)}{8\eta^2}}. \tag{10}$$

Next, we define the total variation distance as $TV(P, Q) = \frac{1}{2}\int |f(x) - g(x)|dx$. Using the Cauchz inequality, we have: $|f(x) - g(x)| \leq |\sqrt{f(x)} - \sqrt{g(x)}||\sqrt{f(x)} + \sqrt{g(x)}| \leq |\sqrt{f(x)} - \sqrt{g(x)}|\sqrt{2}\sqrt{f(x) + g(x)}$. This leads to the

Table 10. Hyper-parameters for KWS, ARR, and ASR.

Task	Hyper-parameters
KWS	batch size = 64, $N = 50$, $M = 20$, $\mathcal{L} = \{0, 10, 20, 30, 40\}$, $\sigma^2 = 0.5$, optimizer = SGD $\zeta = 0.001$, $\lambda = 0.1$, $\tau = -0.1$, $\gamma = 1000$, $\beta = -1.0$.
ARR	batch size = 64, $N = 20$, $M = 20$, $\mathcal{L} = \{0, 10, 20, 30, 40\}$, $\sigma^2 = 0.05$, optimizer = Adam $\zeta = 0.001$, $\lambda = \{1.0, 0.1, 0.1\}$, $\tau = \{-0.1, -0.01, 0.01\}$, $\gamma = \{10, 1.0, 1.0\}$, $\beta = \{-1.0, -0.001, -0.01\}$.
ASR	batch size = 32, $N = 40$, $M = 20$, $\sigma^2 = 0.5$, optimizer = SGD, $\zeta = 0.0003$, $\lambda = 0.1$ $\tau = -10^{-5}$, $\gamma = 1.0$, $\beta = -0.1$.

following:

$$\begin{aligned} TV(P,Q) &\leq \frac{\sqrt{2}}{2}\int |\sqrt{f(x)} - \sqrt{g(x)}|dx \int \sqrt{f(x)+g(x)}dx \\ &\leq \frac{\sqrt{2}}{2}\int |\sqrt{f(x)} - \sqrt{g(x)}|dx \sqrt{\int \frac{f(x)+g(x)}{2}dx} \\ &= \frac{\sqrt{2}}{2}\int |\sqrt{f(x)} - \sqrt{g(x)}|dx \\ &\leq \frac{\sqrt{2}}{2}\sqrt{\int (\sqrt{f(x)} - \sqrt{g(x)})^2 dx}\sqrt{\int 1dx} \\ &= \sqrt{1 - \int \sqrt{f(x)g(x)}dx}. \end{aligned} \tag{11}$$

By combining (10) and (11), we have:

$$TV(P,Q) \leq \sqrt{1 - e^{\frac{-d(x_1,x_2)}{8\eta^2}}}. \tag{12}$$

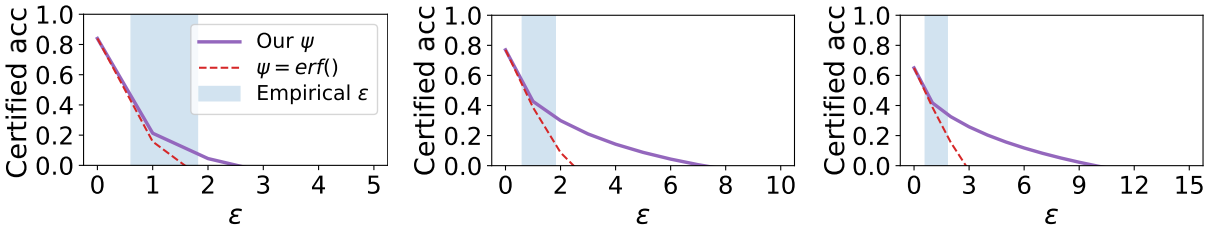
That concludes the proof. □

E Hyper-parameters

We present our hyper-parameters for each task in Table 10. Specifically, N and ζ represent the number of epochs and learning rate for model training, while M and λ pertain to the FA-DRO process. $\mathcal{L}$ denotes the epoch at which FA-DRO is applied, and σ^2 represents the standard deviation of the noise used for randomized smoothing. The parameters τ, γ, and β are the loss weights for Equation 7.

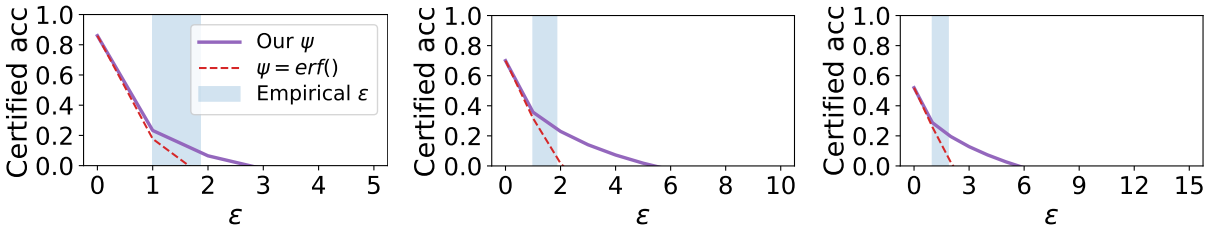
F Certified accuracy for ASR and ARR tasks

Certified accuracy for ASR and ARR tasks are presented in Figure 16 and Figure 17, respectively.



(a) $\eta = 0.5$. (b) $\eta = 1.0$. (c) $\eta = 1.5$.

Fig. 16. Certified accuracy comparison between our proposed ψ and $\psi = \text{erf}()$ across various η values, evaluated on the ASR task.



(a) $\eta = 0.5$. (b) $\eta = 1.0$. (c) $\eta = 1.5$.

Fig. 17. Certified accuracy comparison between our proposed ψ and $\psi = \text{erf}()$ across various η values, evaluated on the ARR task.