Residual Physics-Informed Neural Networks for Robust Indoor Occupancy Estimation using Sonar

Diyar Altinses[1] and Andreas Schwung[2]

***Abstract*— Robust indoor occupancy estimation is critical for smart building safety, especially in emergencies where optical sensors fail due to privacy concerns or smoke. While active acoustic sensing offers a non-intrusive alternative, current black-box models often fail to generalize, while analytical models oversimplify reverberation dynamics. To address this, we introduce a residual physics-informed neural network for acoustic crowd counting using edge devices. Our hybrid architecture fuses Sabine's reverberation theory with a parallel residual branch to correct for non-linear scattering. We validate the approach on a novel real-world sonar dataset collected across three distinct environments. Experimental results demonstrate that our approach outperforms feedforward and recurrent baselines, achieving a detection accuracy of 84.5% and an F1-score of 0.730. Furthermore, the model offers superior interpretability by learning physically plausible absorption parameters and dynamically balancing physical priors with data-driven corrections.**

I. Introduction

In the domain of smart buildings, automated systems are increasingly tasked with ensuring safety, energy efficiency, and security. While modern deep learning algorithms excel at processing high-dimensional sensor data to infer context and activity [1], the robustness of these systems becomes critical in high-stakes environments. A critical example arises in emergency scenarios, such as fires or structural failures. In these situations, real-time knowledge of occupancy, specifically the number and location of trapped individuals, is paramount for first responders to optimize evacuation routes and allocate rescue resources effectively [2]. Accurate crowd density estimation can significantly reduce casualty rates by guiding rescue teams to occupied zones.

Traditionally, occupancy detection relies on optical sensors and surveillance cameras. However, these modalities face severe limitations in emergency contexts. Firstly, visual occlusion caused by smoke or dust renders standard RGB cameras ineffective [3]. Secondly, the deployment of cameras in private spaces, such as bathrooms or changing rooms, is often restricted due to strict privacy regulations [4]. Consequently, there is a need for non-intrusive, privacy-preserving sensing that remains robust under adverse environmental conditions.

Active acoustic probing offers a promising alternative. Unlike light, sound waves permeate smoke and are modulated by the physical presence of bodies, which act as soft absorbers [5]. By analyzing the modulation of the room impulse response and reverberation dynamics, it is possible to infer occupancy counts without capturing personal information.

Despite the potential of acoustics, extracting occupancy counts from raw audio signals remains challenging. State-of-the-art approaches typically employ black box deep neural networks, such as convolutional or recurrent neural networks [6]. While these models achieve high accuracy on training data, they suffer from significant drawbacks: they require massive labeled datasets, lack interpretability, and often fail to extrapolate to scenarios outside their training distribution [7]. Conversely, purely analytical models based on reverberation theories (e.g., Sabine's equation) are often too simplistic to model the complex scattering of real-world furnished rooms.

To bridge the gap between data-driven learning and physical reliability, Physics-Informed Neural Networks (PINNs) have emerged as a powerful paradigm [8]. However, standard PINNs often struggle when the underlying physical equation does not perfectly match the complex reality of acoustic scattering. In this work, we propose a novel, privacy-preserving framework for acoustic crowd estimation using edge devices. We introduce a Residual Physics-Informed Neural Network (Res-PINN) that integrates the Sabine equation of sound absorption while simultaneously learning a residual correction term to account for non-ideal acoustic phenomena. This hybrid architecture allows for robust occupancy estimation even in scenarios with limited training data. The main contributions of this paper are summarized as follows:

(i) We introduce a dataset of active sonar measurements in multiple room environments with varying occupancy levels, designed for emergency sensing research.
(ii) We propose a novel residual physics-informed architecture that combines a hard-constraint layer based on total energy absorption with a learnable residual term.
(iii) We demonstrate through extensive evaluation that our PINN approach offers superior interpretability and extrapolation capabilities compared to black-box models, particularly in unseen occupancy scenarios.

The remainder of this paper is structured as follows: Section II reviews the state-of-the-art in acoustic sensing and physics-informed deep learning. Section III details the data acquisition, introducing our novel real-world dataset and the underlying signal processing pipeline. In Section IV, we present our residual physics-informed neural network using Sabine's reverberation theory. Section V provides an evaluation, benchmarking our approach against other deep learning architectures and analyzing the interpretability. Fi-

[1]Diyar Altinses and [2]Andreas Schwung are with the Department of Automation Technology and Learning Systems, South Westphalia University of Applied Sciences, 59494 Soest, Germany. Email: altinses.diyar@fh-swf.de and schwung.andreas@fh-swf.de

nally, Section VI concludes the study and outlines potential avenues for future research.

II. Related Work

In this section, we review the existing literature regarding non-intrusive occupancy estimation and the recent advancements in scientific machine learning. We categorize the state-of-the-art into two primary domains: (A) device-free acoustic and radio-frequency crowd sensing, and (B) physics-informed neural networks for solving inverse problems.

A. Device-Free Crowd Sensing and Acoustic Monitoring

The demand for privacy-preserving occupancy detection has led to a shift away from optical sensors towards radio-frequency (RF) and acoustic modalities. A significant body of work utilizes Channel State Information (CSI) from Wi-Fi signals to detect human presence. Wang et al. [9] demonstrated that fine-grained CSI amplitude and phase data can capture human motion without wearable devices. Similarly, Xi et al. [10] developed systems to count crowds by analyzing the fading of wireless signals, while Depatla et al. [11] utilized Received Signal Strength variance for occupancy estimation in outdoor environments. However, these RF-based methods often require specialized infrastructure (modified routers) and struggle with multipath interference in complex indoor geometries [12].

In the acoustic domain, research is divided into passive and active sensing. Passive approaches, such as those by Crocco et al. [13], analyze ambient noise (e.g., footsteps, speech) to estimate crowd density. While effective in social settings, passive methods fail in silent scenarios, such as libraries or scenarios involving unconscious individuals during emergencies. Consequently, active acoustic sensing, using ultrasonic or audible chirps, has gained traction. Wang et al. [14] introduced SonarBeat, leveraging phase changes in reflected ultrasonic signals to detect respiration and presence. Nandakumar et al. [15] utilized smartphone speakers and microphones to track finger motion and gestures via Doppler shifts. For room-scale monitoring, Patwal et al. [16] applied machine learning to room impulse responses to classify occupancy levels. Furthermore, advanced deep learning architectures (CNN and LSTM) have been widely adopted to extract features from spectrograms for environmental sound classification [17]. Recent works have also explored using consumer smart speakers for active echolocation to map room boundaries [18].

Despite these advancements, a critical gap remains in the robustness and generalization of current acoustic counting models. Existing methods treat the acoustic signal as a generic high-dimensional feature vector, employing black box deep learning models to learn statistical correlations. These models lack an understanding of the underlying physical laws of sound propagation. Consequently, they require massive labeled datasets for every new room configuration and fail to extrapolate correctly to crowd densities not seen during training. Our work addresses this by integrating the physical Sabine equation directly into the learning process, allowing for robust extrapolation and reduced data dependency.

B. Physics-Informed Learning and Inverse Problems

The paradigm of integrating physical laws into deep learning, known as physics-informed neural networks, was formalized by Raissi et al. [19]. Their seminal work demonstrated that neural networks can solve forward and inverse problems involving partial differential equations by embedding the residuals of the governing equations into the loss function. This approach has revolutionized fields such as fluid dynamics, where Jin et al. [20] successfully applied PINNs to simulate Navier-Stokes equations with limited data. Karniadakis et al. [8] provided a comprehensive review of this field, highlighting the ability of PINNs to act as soft sensors that infer hidden quantities (e.g., coefficients) from sparse measurements.

In the context of wave propagation and acoustics, PINNs have been applied to solve the Helmholtz equation and reconstruct sound speed profiles. Moseley et al. [21] utilized physics-informed networks for seismic wave inversion, while Shukla et al. [22] applied similar concepts to ultrasound non-destructive testing to identify material defects. However, a major challenge in PINNs is the optimization pathology, where the physics-based loss and data-driven loss conflict during training. To mitigate this, Wang et al. [23] proposed adaptive weighting schemes for gradients. Furthermore, hybrid or residual physics learning, as discussed by Willard et al. [24], combines a rigid physics model with a data-driven correction term to account for unmodeled dynamics in complex real-world systems. Zhang et al. [25] applied this hybrid concept to discover unknown terms in differential equations.

While PINNs have proven effective in simulated environments, there is a scarcity of research applying them to noisy, real-world sensor data from edge devices. Most existing acoustic PINNs focus on generating wavefields rather than solving high-level regression tasks like occupancy estimation. Furthermore, pure physics models often fail in real rooms due to non-ideal factors that cannot be captured. We bridge this gap by proposing a residual PINN architecture. Unlike standard approaches, our model uses the physics equation for extrapolation while simultaneously learning a residual correction term to adapt to real-world acoustic environments.

III. Data Acquisition and System Model

To enable the training of the proposed physics-informed neural network and to rigorously benchmark it against classical deep learning architectures, we established a high-fidelity acoustic sensing pipeline. The system operates on the principle of active monostatic sonar, treating the indoor environment as a complex, linear time-variant acoustic channel. An overview of the complete data acquisition and processing chain is depicted in Figure 1. As illustrated, the pipeline transforms raw acoustic reflections captured by a commodity smartphone into a structured dataset through a sequence of

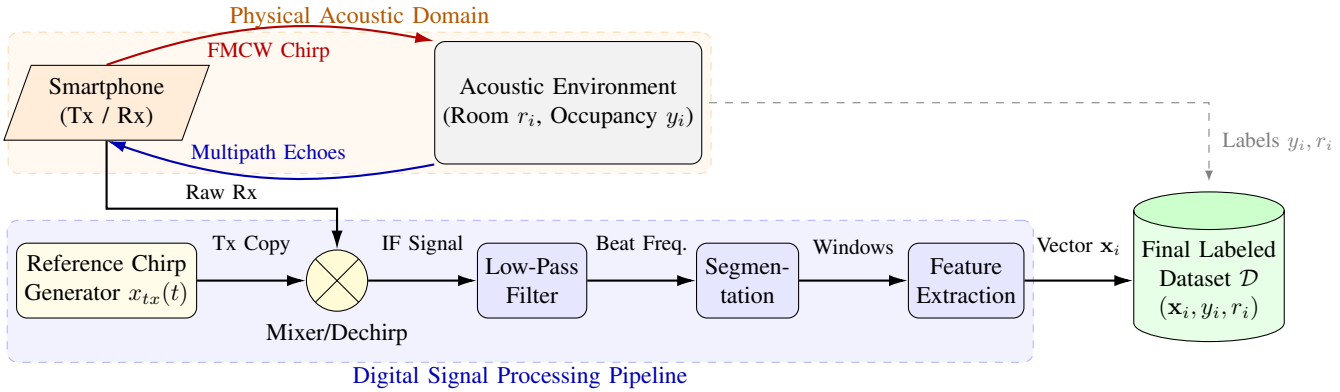


Fig. 1: Overview of the data collection and signal processing pipeline (Compressed View).

analog-to-digital mixing, filtering, and segmentation steps. The following subsections detail the specific components shown in the diagram, including the signal design, the channel physics, the digital signal processing chain, and the experimental geometry.

A. Signal Design: Frequency Modulated Continuous Wave

The choice of the probing signal is critical for determining the range resolution and the signal-to-noise ratio of the system. While simple ultrasonic pulses are computationally cheap, they suffer from a trade-off between peak power and total energy. To overcome the peak-power limitations while maintaining high energy transmission, we utilize a linear frequency-modulated chirp.

The transmitted signal $x_{tx}(t)$ is synthesized digitally and converted to an analog waveform via the device's digital-to-analog converter. The instantaneous frequency $f(t)$ varies linearly over time:

$$f(t) = f_{start} + \frac{B}{T_c}t, \quad 0 \le t \le T_c, \tag{1}$$

where $f_{start} = 17\,\text{kHz}$ denotes the lower cut-off frequency, $B = f_{stop} - f_{start} = 5\,\text{kHz}$ represents the effective bandwidth (sweeping up to 22 kHz), and $T_c = 20\,\text{ms}$ is the chirp duration. The chirp rate μ, which defines the slope of the frequency sweep, is given by $\mu = B/T_c$. The analytic representation of the transmitted pulse can be expressed as:

$$\begin{aligned} x_{tx}(t) = \text{rect}\left(\frac{t - T_c/2}{T_c}\right) \cdot A_{tx} \\ \cdot \exp\left(j2\pi\left(f_{start}t + \frac{\mu}{2}t^2\right)\right), \end{aligned} \tag{2}$$

where A_{tx} is the transmission amplitude and $\text{rect}(\cdot)$ is the rectangular window function.

This signal design offers two decisive advantages for indoor occupancy detection:

(i) Range Resolution (ΔR): The ability to distinguish between two closely spaced reflectors (e.g., a person standing near a wall) is inversely proportional to the bandwidth. Theoretically, $\Delta R = \frac{c}{2B}$, where $c \approx 343\,m/s$ is the speed of sound. By maximizing B within the constraints of the audio hardware (Nyquist limit), we maximize spatial separability.

(ii) Noise Immunity: Since the energy is spread over time T_c, the signal is robust against impulsive noise and narrowband interference, which is filtered out during the pulse compression stage.

Having established the properties of the transmitted waveform, the subsequent analysis focuses on the acoustic channel model.

B. Channel Model and Acoustic Propagation Physics

We model the room as a reverberant acoustic enclosure. In a monostatic configuration, where the transmitter (Tx) and receiver (Rx) are co-located on the same edge device, the received signal $y_{rx}(t)$ is mathematically described as the convolution of the transmitted signal with the room impulse response, $h(t,\tau)$, subject to additive Gaussian noise $\eta(t)$:

$$y_{rx}(t) = \int_{-\infty}^{\infty} h(t,\tau)x_{tx}(t-\tau)d\tau + \eta(t) \tag{3}$$

Here, $h(t,\tau)$ represents the time-variant impulse response. We decompose $h(t,\tau)$ into three physical components:

$$h(t,\tau) = h_{LOS}(\tau) + \sum_{k=1}^{K} h_{wall}^{(k)}(\tau) + \sum_{m=1}^{M} h_{body}^{(m)}(t,\tau). \tag{4}$$

The first term, $h_{LOS}(\tau)$, represents the line-of-sight component traversing the direct path from the loudspeaker to the microphone. The second term, comprising the static multipath reflections $\sum h_{wall}^{(k)}(\tau)$, originates from stationary boundaries such as walls, ceilings, and immobile furniture. The critical term for our analysis is the dynamic body interaction component, $\sum h_{body}^{(m)}(t,\tau)$. Unlike rigid construction materials such as concrete ($\alpha_{concrete} \approx 0.02$), human bodies function as soft porous absorbers with a significantly higher absorption coefficient ($\alpha_{human} \approx 0.3 - 0.5$). Consequently, the presence of a human occupant alters the room impulse response through absorption and scattering.

C. Digital Signal Processing: Matched Filtering

To extract the feature-rich echo profile from the raw continuous recording, we perform pulse compression using a matched filter. This process correlates the received

stream with a time-reversed replica of the transmitted chirp, effectively compressing the long-duration energy of length T_c into a narrow sinc-like pulse. Mathematically, the cross-correlation function $R_{xy}[\tau]$ is computed in the discrete domain. Let $x[n]$ and $y[n]$ be the discrete-time versions of transmitted and received signals, sampled at $F_s = 48\,000$ Hz. The Normalized Cross-Correlation (NCC) is defined as:

$$R_{xy}[\tau] = \frac{\sum_{n=0}^{N-1} x[n] \cdot y[n+\tau]}{\sqrt{\sum_{n=0}^{N-1} |x[n]|^2 \cdot \sum_{n=0}^{N-1} |y[n+\tau]|^2}} \quad (5)$$

The vector $\mathbf{r} = [R_{xy}[0], \ldots, R_{xy}[W]]$ constitutes the echo profile, which serves as the input for the neural networks.

D. Experimental Geometries and Environmental Variance

To prevent the model from overfitting to a single acoustic environment, data acquisition was performed in three distinct rooms ($\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3$). These rooms were selected to maximize variance in the Volume-to-Surface ratio (V/S) and the mean free path ($d_{mfp} = 4V/S$), which are the governing parameters in Sabine's reverberation equation. The first environment, $\mathcal{R}_1$, approximates a typical residential setting with a floor area of $16\,m^2$ and a volume of $\approx 40\,m^3$. It contains soft porous absorbers such as sofas, curtains, and carpets, resulting in a heterogeneous absorption spectrum and frequency-dependent decay times. In contrast, $\mathcal{R}_2$ serves as an acoustic stress test with an area of $12\,m^2$ and a volume of $\approx 30\,m^3$. Characterized by sparse furnishing and hard reflective boundaries like glass windows and plaster walls, this room exhibits a low average absorption coefficient $\bar{\alpha}$, leading to a high reverberation time (RT_{60}) and strong low-frequency standing wave modes. Finally, $\mathcal{R}_3$ represents a highly cluttered environment with a reduced footprint of $8\,m^2$ and a volume of $\approx 20\,m^3$. The high density of scattering objects relative to the room volume significantly reduces the mean free path, creating a dense pattern of early reflections.

E. Protocol and Dataset Construction

The data collection was executed using an edge sensing node, placed on a static surface at a height of $1.5\,m$. To ensure signal fidelity, the application layer utilized the Android AudioSource API to access the raw microphone stream, effectively bypassing system-level noise suppression algorithms that could otherwise distort the room impulse response. We defined a supervised regression task with the target variable $y \in \mathbb{R}_{\geq 0}$ representing the number of occupants.

The collection protocol was divided into two distinct phases for each room to ensure robust generalization. First, a baseline recording was conducted in an empty state for 10 minutes to establish the baseline acoustic signature (h_{empty}). Subsequently, an occupied recording introduced a human subject adhering to a dynamic protocol designed to capture the stochastic nature of human radar cross-sections. This included positioning the subject in non-line-of-sight zones (e.g., corners, behind furniture) to force reliance on diffuse field absorption rather than direct blockage, alongside natural breathing and postural shifts to introduce micro-Doppler effects.

The resulting continuous audio streams were sliced into overlapping windows based on the chirp duration T_c and stride S, yielding a final dataset of $N_{samples} = 4\,140$ labeled tensors. Each sample is stored as a tuple $(\mathbf{x}_i, y_i, r_i)$, containing the correlation history, occupancy label, and room index used to condition the PINN. We have made the raw dataset publicly available at: Repository[1] [26].

F. Exploratory Data Analysis and Feature Separability

Before deploying the deep learning architectures, it is essential to verify the intrinsic discriminative power of the collected acoustic signatures. Specifically, we investigate whether the raw feature vectors $\mathbf{x}_i$ contain sufficient statistical variance to distinguish between occupied and empty states across different room geometries, without relying on the non-linear transformations of a neural network.

To visualize the underlying structure of the high-dimensional data, we employ t-Distributed Stochastic Neighbor Embedding (t-SNE). We project the input feature vectors into a two-dimensional latent space $\mathbf{z}_i \in \mathbb{R}^2$. For this analysis, we utilized a perplexity of 30 and ran the optimization for 1,000 iterations to ensure stable convergence. The resulting projection is visualized in Figure 2.

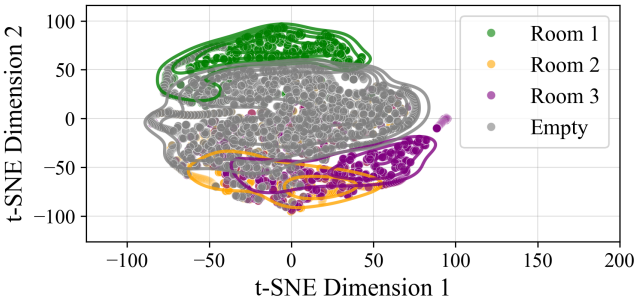


Fig. 2: Visualization (t-SNE) of the learned acoustic feature space. Gray points represent the baseline "Empty" state across all environments, forming a central manifold. Colored clusters correspond to the "Occupied" state in different rooms.

The projection validates our approach, revealing a topology dominated by environmental context that necessitates room-specific embeddings. Crucially, we observe robust class separability: "occupied" states form distinct clusters significantly detached from the central "empty" manifold. This confirms that human presence induces a consistent acoustic shift, proving the raw features possess sufficient signal integrity for the regression task, provided the model conditions on room parameters.

To complement the high-level manifold projection provided by t-SNE and to gain deeper insights into the physical

[1] https://zenodo.org/records/18096214

characteristics of the recorded signals, we visualize in Figure 3 selected raw data using spatio-temporal echograms. This visualization technique stacks consecutive acoustic impulse responses along the vertical axis, representing the temporal evolution of the environment over many chirp cycles. The horizontal axis represents the time delay and the range of reflections within a single measurement window. The color intensity corresponds to the normalized amplitude of the received acoustic return.

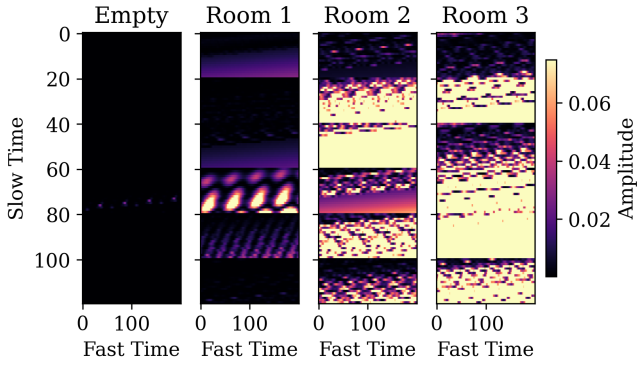


Fig. 3: Spatio-temporal echograms visualizing the normalized acoustic amplitude across "Fast Time" (range samples) and "Slow Time" (sequential acquisition index). The panels show, from left to right: the aggregated empty baseline condition, and occupied conditions in Room 1, Room 2, and Room 3. Brighter colors indicate higher reflection amplitudes.

Figure 3 contrasts the "Empty" baseline with occupied states. While the empty plot shows a low noise floor with minimal features, human presence introduces significant acoustic energy in the selected setting. Room 1 displays clear reflection curves modulating with micro-Doppler effects from respiration. In contrast, Rooms 2 and 3 exhibit complex multipath clutter where human reflections are embedded within strong reverberation tails, particularly in the saturated signal of the confined Room 3. These temporal fluctuations enable the model to disentangle occupancy states from room identities.

IV. Residual Physics-Informed Neural Network

In this section, we formalize the acoustic crowd counting task as a supervised regression problem and systematically construct our Residual Physics-Informed Neural Network (Res-PINN). Our approach moves beyond black-box modeling by explicitly integrating the physical laws of sound absorption while maintaining the flexibility to correct for non-ideal acoustic phenomena via a residual learning branch. The complete architecture, which fuses a physics-based estimation head with a data-driven correction mechanism, is illustrated in Figure 4. In the following subsections, we first define the optimization objective, followed by a step-by-step derivation of the model's components: the shared feature extraction backbone, the physics-compliant estimation branch, and the residual correction mechanism.

A. Problem Definition

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i, r_i)\}_{i=1}^{N}$ denote the dataset consisting of N independent and identically distributed (i.i.d.) samples. Here, $\mathbf{x}_i \in \mathbb{R}^W$ represents the input feature vector (the windowed cross-correlation profile of length W), $y_i \in \mathbb{R}_{\geq 0}$ denotes the ground-truth number of occupants, and $r_i \in \{1, \ldots, K\}$ is a categorical index representing the specific room environment from the set of K available geometries.

The objective is to approximate the unknown mapping function $\Phi : \mathbb{R}^W \times \{1, \ldots, K\} \to \mathbb{R}_{\geq 0}$ such that the expected loss is minimized over the data distribution. We seek the optimal parameters θ^* of a neural network f_θ defined as:

$$\theta^* = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(f_\theta(\mathbf{x}_i, r_i), y_i\right), \tag{6}$$

where $\mathcal{L}(\cdot)$ is the Mean Squared Error (MSE) loss function. Unlike standard regression, where f_θ is an unconstrained function approximator, we impose structural constraints on f_θ derived from architectural acoustics theory.

B. Physics-Informed Learning Model

To effectively leverage the known physical properties of sound propagation, our base architecture is designed to mirror the inverse problem of room acoustics. This model is composed of two sequential stages: a feature extraction backbone and a physics-based estimation head.

1) Feature Extraction Backbone: The high-dimensional input $\mathbf{x}$ is first processed by a shared backbone network $g_\phi(\cdot)$ to extract a compact, noise-robust latent representation. We employ a multi-layer perceptron consisting of dense layers with rectified linear unit activations, defined as:

$$\mathbf{z} = g_\phi(\mathbf{x}) \in \mathbb{R}^D, \tag{7}$$

where $\mathbf{z}$ captures the fundamental decay and reflection patterns of the impulse response, serving as the common input for the subsequent heads.

2) The Physics-Based Estimation Head: This processing head is explicitly designed to satisfy Sabine's reverberation theory. According to Sabine's law, the occupancy N is physically related to the total absorption A_{total} and the room's intrinsic absorption A_{room} via the relation:

$$N_{phy} = \frac{A_{total} - A_{room}}{\alpha_{person}}, \tag{8}$$

where α_{person} is the absorption cross-section of a human ($\approx 0.3 - 0.5$ Sabins). To implement this, the head first estimates the instantaneous total absorption from the latent features using a linear projection parameterized by learnable weights $\mathbf{w}_a \in \mathbb{R}^D$ and bias $b_a \in \mathbb{R}$:

$$\hat{A}_{total} = \text{Softplus}(\mathbf{w}_a^T \mathbf{z} + b_a). \tag{9}$$

The Softplus activation, defined as $\text{Softplus}(x) = \ln(1 + \exp(x))$, ensures strictly positive absorption values. Simultaneously, to account for domain shifts between different environments, we employ a system identification embedding. The baseline absorption of the empty room is learned via a

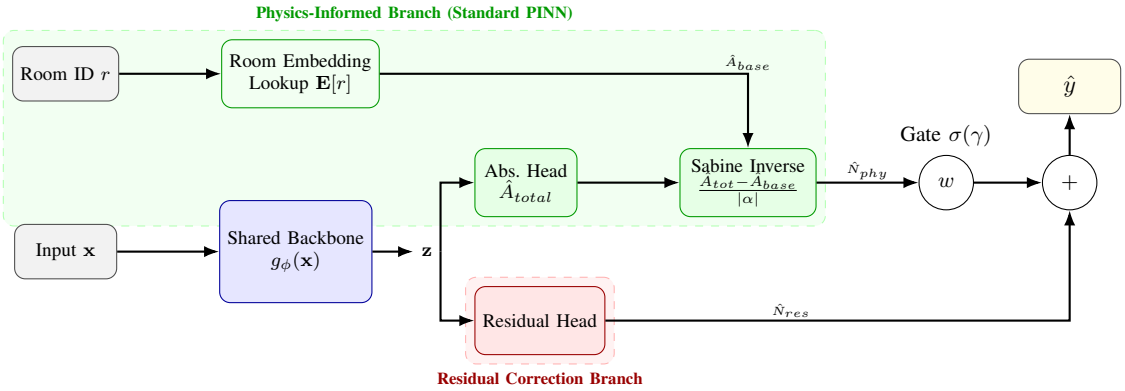


Fig. 4: Architecture of the proposed Res-PINN. The network shares a feature extractor backbone. The upper branch (green) enforces Sabine's physical law, conditioned on a learned room embedding. The lower branch (red) learns a residual correction term. A gating mechanism w dynamically fuses both estimates to produce the final prediction $\hat{y}$.

parameter matrix $\mathbf{E} \in \mathbb{R}^{K\times 1}$, where each row represents a learnable scalar for one of the K rooms:

$$\hat{A}_{room}(r) = \text{Softplus}(\mathbf{E}_r). \tag{10}$$

Here, $\mathbf{E}_r$ denotes the entry corresponding to the room r.

Finally, the physics-compliant occupancy estimate is computed by enforcing the structure of Eq. (8):

$$\hat{N}_{phy} = \frac{\text{ReLU}(\hat{A}_{total} - \hat{A}_{room}(r))}{|\alpha| + \epsilon}. \tag{11}$$

Here, α is a learnable parameter representing the global absorption coefficient per person, and ϵ is a small constant to stabilize the division.

C. Extended Residual Architecture

A purely physics-based model assumes a diffuse sound field, which is often violated in real-world scenarios due to non-linear scattering, standing waves, and non-line-of-sight shadowing. To address this without discarding the physical priors, we extend the model with a residual correction mechanism.

1) The Residual Correction Branch: We introduce a parallel branch designed to capture the physics-defying residuals, i.e., acoustic phenomena that deviate from the linear Sabine approximation. This branch maps the latent features $\mathbf{z}$ to a correction term $\hat{N}_{res}$:

$$\hat{N}_{res} = \text{Linear}(\text{ReLU}(\mathbf{W}_c\mathbf{z} + \mathbf{b}_c)). \tag{12}$$

Unlike the physics head, this branch is unconstrained and allows for negative values, enabling it to correct overestimations caused by constructive interference or underestimations caused by shadowing.

2) Adaptive Gating and Final Prediction: To dynamically balance the reliance on rigid theory versus data-driven correction, we introduce a learnable gating parameter γ. This parameter is passed through a sigmoid function to produce a blending weight $w \in (0, 1)$:

$$w = \sigma(\gamma). \tag{13}$$

The final occupancy prediction $\hat{y}$ is obtained via:

$$\hat{y} = w \cdot \hat{N}_{phy} + (1 - w) \cdot \hat{N}_{res}. \tag{14}$$

This hybrid formulation allows the network to prioritize the robust physical model $\hat{N}_{phy}$ in standard conditions while leveraging the residual term $\hat{N}_{res}$ to compensate for complex scattering effects in challenging environments.

V. Evaluation

In this section, we present a comprehensive empirical validation of the proposed Res-PINN. We benchmark the model against a diverse suite of state-of-the-art deep learning architectures to assess its efficacy in acoustic occupancy estimation. The evaluation is designed to address three primary research objectives: (1) quantifying the performance gain achieved by the hybrid residual architecture over traditional black-box models, (2) assessing the model's convergence stability and robustness to initialization noise, and (3) validating the physical interpretability of the learned parameters.

A. Experimental Setup

The experimental validation was implemented in PyTorch using seven distinct neural architectures: a simple feed-forward neural network (MLP), recurrent neural networks (RNN, GRU, LSTM), a convolutional neural network (CNN) baseline, a physics-informed neural network (see subsection IV-B), and the Res-PINN (see subsection IV-C). We ensure that no model benefits solely from a significantly larger number of parameters. Specifically, the CNN baseline, the PINN, and the Res-PINN share an identical feature extraction backbone (as described in Section IV), differing only in their prediction heads. For the sequence-based models (RNN, GRU, LSTM) and the MLP, the hidden layer dimensions were tuned to yield a total parameter count within the same order of magnitude as the proposed Res-PINN ($\approx$ 12 500 parameters).

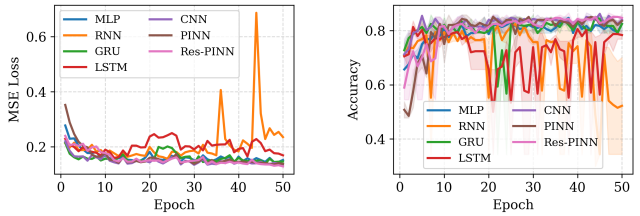
All models shared a consistent data preprocessing pipeline, where raw acoustic signals were windowed with a window size of 50 and a stride of 50, and standardized to zero mean

and unit variance. The dataset was split into training (80%) and validation (20%) sets. All models were trained for 50 epochs using the Adam optimizer with a learning rate of 0.0005 and a batch size of 16, optimizing the MSE loss. To assess statistical robustness, the entire training procedure was repeated over 10 independent runs, with training histories and learnable physical parameters recorded for analysis.

B. Results

In this section, we present the experimental evaluation of the proposed Res-PINN compared to the PINN and various data-driven baselines. We analyze the models' performance in terms of convergence stability, generalization capability, and final classification metrics across 10 independent training runs.

1) Training Dynamics and Convergence: To assess the learning stability and convergence speed, we visualize the validation metrics in Figure 5, which illustrates the MSE loss alongside Accuracy, Recall, and F1-Score. The shaded regions indicate the standard deviation across the 10 runs.



(a) Validation Loss (MSE)

(b) Classification Accuracy

(c) Recall Metric

(d) F1-Score

Fig. 5: Comparative evaluation of the proposed PINN architectures against baseline architectures over training epochs. The plots display the mean performance (solid lines) and standard deviation (shaded regions) across 10 independent runs.

The validation loss (Figure 5a) demonstrates that physics-informed models (PINN and Res-PINN) exhibit significantly smoother convergence compared to pure sequence models. Notably, the RNN suffers from severe instability, characterized by sharp loss spikes and accuracy drops to random guessing (≈50%) in Figure 5b. This suggests that naive recurrent architectures struggle with vanishing gradients on the high-dimensional sonar data. While the standard PINN converges stably, it plateaus at a higher loss and lower F1-score than the Res-PINN. This performance gap validates our hypothesis: the rigid application of Sabine's formula acts as a strong regularizer but introduces a bias when the physical assumptions are violated by local scattering. The Res-PINN overcomes this by leveraging its residual branch to correct these model-plant mismatches, achieving the lowest final MSE and the highest classification accuracy. In terms of F1-Score (Figure 5d), the Res-PINN consistently outperforms the CNN baseline, particularly in the early training phases.

2) Quantitative Performance Comparison: Table I summarizes the final performance metrics averaged over the last 5 epochs of all 10 runs. This comparison highlights the improvements in regression error and classification performance.

TABLE I: Comparison of performances across 10 trials.

Model	Val Loss (MSE)	F1-Score	Accuracy
MLP	0.1494 ± 0.0036	0.647 ± 0.029	0.816 ± 0.010
CNN	0.1420 ± 0.0040	0.706 ± 0.009	0.830 ± 0.003
RNN	0.2333 ± 0.0305	0.388 ± 0.126	0.603 ± 0.124
GRU	0.1493 ± 0.0051	0.652 ± 0.028	0.815 ± 0.010
LSTM	0.1954 ± 0.0371	0.494 ± 0.113	0.749 ± 0.053
PINN	0.1423 ± 0.0058	0.702 ± 0.017	0.834 ± 0.007
Res-PINN	**0.1357 ± 0.0045**	**0.730 ± 0.019**	**0.845 ± 0.009**

The data confirms that the Res-PINN outperforms all baselines across all metrics. Specifically, it achieves the lowest MSE of 0.1357 and the highest F1-Score of 0.730. Comparing the two physics-informed variants, the Res-PINN yields approx. 4% improvement in F1-score over the standard PINN (0.702). This gap validates the necessity of the residual correction branch. Among the black-box models, the CNN performs competitively (F1: 0.706), surpassing the MLP and RNNs, but falls short of the Res-PINN. The poor performance of the simple RNN (0.603 accuracy) and LSTM indicates that these architectures struggle to learn robust features from the raw sonar data. In contrast, the Res-PINN achieves state-of-the-art performance with a compact, interpretable architecture.

3) Interpretability and Physical Consistency: A key advantage of the proposed physics-informed architecture over black-box baselines is its inherent interpretability. Unlike the baselines, the Res-PINN explicitly learns physical parameters that describe the acoustic environment. To validate this property, we analyze the evolution of the learnable absorption coefficient α and the physics-gating weight w during the training process.

Figure 6 visualizes the parameter trajectories averaged across 10 runs. The blue curves represent the estimated acoustic absorption per person (α). The standard PINN exhibits a noisy trajectory that drifts towards a higher value ($\alpha \approx 0.55$). This behavior suggests that the rigid physical model attempts to compensate for unmodeled scattering losses by artificially inflating the absorption coefficient. In contrast, the Res-PINN converges rapidly and smoothly to a lower, stable value of $\alpha \approx 0.42$. This value aligns more closely with empirical Sabine absorption coefficients for human clothing in the ultrasonic range, indicating that the residual branch successfully offloads the non-linear errors. The red line depicts the evolution of the physics gate w in

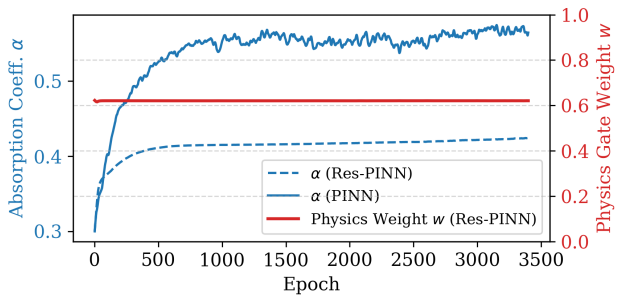


Fig. 6: Evolution of learned physical parameters over the course of training. The left y-axis (blue) tracks the absorption coefficient α (Sabins per person) for both the standard PINN (solid line) and the Res-PINN (dashed line). The right y-axis (red) displays the gating weight w of the Res-PINN.

the Res-PINN. The network attributes roughly 63% of the prediction to the analytical Sabine model and 37% to the data-driven residual correction. This confirms that while the physical law provides a strong and necessary foundation, significant deviations exist in real-world sonar data that require the corrective capacity of the residual network.

VI. Conclusion

In this work, we addressed robust, privacy-preserving occupancy estimation using active sonar. Since purely data-driven models struggle to generalize across varying room geometries, we proposed the Res-PINN, which fuses Sabine's reverberation theory with a residual deep learning branch. Validated on a real-world dataset across three environments, the Res-PINN significantly outperformed baselines, achieving 84.5% accuracy and an F1-score of 0.730 with superior stability. Crucially, the model demonstrated high interpretability, converging to physically plausible absorption coefficients ($\alpha \approx 0.42$) and dynamically balancing the physical prior with data-driven corrections ($w \approx 0.63$). This confirms the residual branch compensates for unmodeled scattering without discarding fundamental acoustic laws. Future work will extend Res-PINN to multi-person crowd counting and dynamic environments with continuous movement. We also aim to deploy the inference pipeline on resource-constrained embedded microcontrollers to demonstrate real-time feasibility.

References

[1] D. Altinses and A. Schwung, "Performance benchmarking of multimodal data-driven approaches in industrial settings," *Machine Learning with Applications*, p. 100691, 2025.

[2] G. Solmaz, F.-J. Wu, F. Cirillo, E. Kovacs, J. R. Santana, L. Sánchez, P. Sotres, and L. Munoz, "Toward understanding crowd mobility in smart cities through the internet of things," *IEEE Communications Magazine*, vol. 57, no. 4, pp. 40–46, 2019.

[3] D. Altinses and A. Schwung, "Multimodal synthetic dataset balancing: a framework for realistic and balanced training data generation in industrial settings," in *IECON 2023-49th Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2023, pp. 1–7.

[4] J. Wang, Q. Gao, M. Pan, and Y. Fang, "Device-free wireless sensing: Challenges, opportunities, and applications," *IEEE network*, vol. 32, no. 2, pp. 132–137, 2018.

[5] W. C. Sabine, *Collected papers on acoustics*. Courier Dover Publications, 2015.

[6] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1, no. 2.

[7] D. Altinses and A. Schwung, "Fault-tolerant multimodal representations learning: Fusion of intermodal and intramodal correlations," *IEEE Transactions on Artificial Intelligence*, pp. 1–13, 2025.

[8] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, "Physics-informed machine learning," *Nature Reviews Physics*, vol. 3, no. 6, pp. 422–440, 2021.

[9] Y. Wang, J. Liu, Y. Chen, M. Gruteser, J. Yang, and H. Liu, "E-eyes: Device-free location-oriented activity identification using fine-grained wifi signatures," in *Proceedings of the 20th annual international conference on Mobile computing and networking*, 2014, pp. 617–628.

[10] W. Xi, J. Zhao, X.-Y. Li, K. Zhao, S. Tang, X. Liu, and Z. Jiang, "Electronic frog eye: Counting crowd using wifi," in *IEEE INFOCOM Conference on Computer Communications*. IEEE, 2014, pp. 361–369.

[11] S. Depatla, A. Muralidharan, and Y. Mostofi, "Occupancy estimation using only wifi power measurements," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 7, pp. 1381–1393, 2015.

[12] F. Adib and D. Katabi, "See through walls with wifi!" in *Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM*, 2013, pp. 75–86.

[13] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," *ACM CSUR*, vol. 48, no. 4, pp. 1–46, 2016.

[14] X. Wang, R. Huang, and S. Mao, "Sonarbeat: Sonar phase for breathing beat monitoring with smartphones," in *26th International Conference on Computer Communication and Networks*. IEEE, 2017, pp. 1–8.

[15] R. Nandakumar, V. Iyer, D. Tan, and S. Gollakota, "Fingerio: Using active sonar for fine-grained finger tracking," in *Conference on Human Factors in Computing Systems*, 2016, pp. 1515–1525.

[16] A. Patwal, M. Diwakar, V. Tripathi, and P. Singh, "Crowd counting analysis using deep learning: A critical review," *Procedia Computer Science*, vol. 218, pp. 2448–2458, 2023.

[17] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.

[18] L. Remaggi, P. J. Jackson, P. Coleman, and W. Wang, "Room boundary estimation from acoustic room impulse responses," in *2014 Sensor Signal Processing for Defence (SSPD)*. IEEE, 2014, pp. 1–5.

[19] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational physics*, vol. 378, pp. 686–707, 2019.

[20] X. Jin, S. Cai, H. Li, and G. E. Karniadakis, "Nsfnets (navier-stokes flow nets): Physics-informed neural networks for the incompressible navier-stokes equations," *JCP*, vol. 426, p. 109951, 2021.

[21] B. Moseley, A. Markham, and T. Nissen-Meyer, "Solving the wave equation with physics-informed deep learning," *arXiv preprint arXiv:2006.11894*, 2020.

[22] K. Shukla, P. C. Di Leoni, J. Blackshire, D. Sparkman, and G. E. Karniadakis, "Physics-informed neural network for ultrasound nondestructive quantification of surface breaking cracks," *Journal of Nondestructive Evaluation*, vol. 39, no. 3, p. 61, 2020.

[23] S. Wang, Y. Teng, and P. Perdikaris, "Understanding and mitigating gradient flow pathologies in physics-informed neural networks," *SIAM Journal on Scientific Computing*, vol. 43, no. 5, pp. A3055–A3081, 2021.

[24] J. Willard, X. Jia, S. Xu, M. Steinbach, and V. Kumar, "Integrating scientific knowledge with machine learning for engineering and environmental systems," *ACM CSUR*, vol. 55, no. 4, pp. 1–37, 2022.

[25] D. Zhang, L. Guo, and G. E. Karniadakis, "Learning in modal space: Solving time-dependent stochastic pdes using physics-informed neural networks," *SISC*, vol. 42, no. 2, pp. 639–665, 2020.

[26] D. Altinses, "Indoor acoustic occupancy dataset: Smartphone- based fmcw sonar in diverse room geometries," Dec. 2025. [Online]. Available: https://doi.org/10.5281/zenodo.18096214