Delft University of Technology
Master of Science Thesis in Computer and Embedded Systems Engineering

Leveraging Neural Acoustic Fields for Indoor Localization

Mees Jonker

Embedded Systems

TU Delft Delft University of Technology

Leveraging Neural Acoustic Fields for Indoor Localization

Master of Science Thesis in Computer and Embedded Systems Engineering

Embedded Systems Group
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology
Van Mourik Broekmanweg 6, 2628 XE Delft, The Netherlands

Mees Jonker

October 6th, 2025

Author
Mees Jonker
Title
Leveraging Neural Acoustic Fields for Indoor Localization
MSc Presentation Date
October 20th, 2025

Graduation Committee

Prof. dr. Koen Langendoen	Delft University of Technology
Dr. Guohao Lan	Delft University of Technology
Dr. Nitinder Mohan	Delft University of Technology

Abstract

This thesis presents an analysis-by-synthesis approach for single-microphone indoor localization that inverts Neural Acoustic Fields (NAFs) by comparing synthesized and measured room impulse responses. Inspired by Neural Radiance Fields (NeRFs), NAFs model room impulse responses (RIRs) as continuous functions of spatial coordinates, enabling localization through spectral loss minimization over candidate listener positions. To mitigate computational overhead, we introduce Standard Deviation-Weighted Sampling (SDWS), focusing on informative time-frequency bins. Further, we evaluate regularization effects on loss landscapes. Evaluated on SoundSpaces (simulated, binaural) and RAF (real-world, monaural) datasets, the method shows complementary behavior across datasets. While it outperforms direct regression baselines (ResNet-10, NAF-Direct) in sparse-data regimes on RAF, achieving up to 32% lower mean localization error (on RAF at 10% data), performance is lower on SoundSpaces, likely due to the high acoustic similarity between different locations in the simulated environments. PSO reduces runtime by 75% over grid search while improving accuracy by 14%, and SDWS cuts computation by 40× with only 22% error increase. The approach demonstrates NAF's potential for localization but highlights trade-offs between inference time (5-200 s per query) and performance. Future work could extend the method to jointly estimate listener position and orientation, or to incorporate a hybrid search algorithm for more efficient exploration of the loss space.

"Failure is simply the opportunity to begin again, this time more intelligently."
– Henry Ford

Preface

My fascination for understanding acoustics and its use in localization started with my honours programme project during my bachelor's studies, where I collaborated with researchers from the former ENS group on a robot localization module. This experience left a lasting impression, sparking my curiosity and excitement for the field. As a result, it was an easy choice for me when my supervisor, Guohao Lan, proposed exploring Neural Acoustic Fields. Working on this thesis has been a very rewarding journey. The project has taught me a lot, not only about Neural Acoustic Fields, but also about tackling challenges step by step and finding creative ways forward when things did not work right away.

I would like to express my deep gratitude to my supervisor Dr. Guohao Lan for his patient guidance and advice throughout the project. Your encouragement and insights made the challenges along the way far more manageable. Each time I seemed to get stuck, you helped me see the problem from a fresh perspective and motivated me to keep moving forward. I am also very grateful to my thesis advisor Prof. Koen Langendoen for helping me with graduation matters, providing useful insights, and being on my graduation committee. My thanks also go to Dr. Nitinder Mohan for serving on my committee. Finally, I am especially grateful to my friends, family, and girlfriend for their constant support, not only during the work on this thesis, but throughout my years as a student in Delft.

Mees Jonker

Delft, The Netherlands
6th October 2025

Contents

Chapter 1

Introduction

1.1 Motivation

Indoor localization plays a vital role in supporting a range of location-based services, including indoor navigation, health rehabilitation, and human-computer interaction (HCI) [32]. Classical acoustic techniques, such as time-of-flight (ToF) [49], time-difference-of-arrival (TDoA) [36, 55], or MUSIC [51] often rely on multiple microphones and degrade in performance under multipath effects in complex environments. Methods that rely on a single microphone typically require precise prior knowledge of the room geometry and are only practical in simple and rectangular rooms [44]. Data-driven techniques address these issues by learning patterns from acoustic data, using raw waveforms [59, 46, 37] or features like STFT spectrograms [1, 16] and GCC vectors [63, 20, 60]. However, these methods often require large datasets [16] for model training.

Neural Acoustic Fields (NAFs) [34] are an emerging method for modeling room acoustics by learning a continuous mapping from spatial coordinates to Room Impulse Responses (RIRs). They are inspired by Neural Radiance Fields (NeRFs) [40], which represent 3D scenes as continuous functions mapping spatial coordinates and viewing directions to color and density, enabling photorealistic novel-view synthesis. Similarly, NAFs can synthesize realistic spatial acoustic signals. Most prior work [34, 54, 30] has focused on forward RIR synthesis, leaving their potential for inverse tasks, such as localization, largely unexplored.

Analysis-by-synthesis (AxS) offers a framework to tackle such inverse problems. Rather than predicting positions directly through discriminative regression, AxS generates candidate positions, synthesizes the corresponding signals using a generative model, and then minimizes the discrepancy between these synthesized signals and the actual observations. The candidate with the smallest discrepancy is then selected as the final estimate [18, 42, 4]. This approach has been successfully applied in the field of computer vision. For instance, in [11], an image synthesis network generates object views, which are then compared to the input image to optimize the pose, outperforming direct regression. In the context of visual NeRFs, iNeRF [31] extends this idea by inverting NeRF for camera pose estimation. Similarly, recent acoustic studies [65, 61] demonstrate

the effectiveness of AxS in the acoustic domain, with DAF [65] applied to fallen object localization and DiffRIR [61] to RIR reconstruction. Motivated by these works, this thesis proposes inverting NAFs through an AxS approach to perform single-microphone listener localization based on a single impulse response, given a known source position. Specifically, we design an analysis-by-synthesis pipeline that (i) synthesizes candidate RIRs using a trained NAF, (ii) searches the acoustic loss landscape using grid search or particle swarm optimization, and (iii) reduces computational cost via a Standard Deviation-Weighted Sampling strategy to focus on informative spectrogram regions. We further investigate the effect of data sparsity and regularization on localization accuracy. We evaluate the proposed methods on two public datasets, i.e., the SoundSpaces [9, 10] and RAF [12].

1.2 Contributions

This thesis makes the following key contributions:

- **Inversion of NAF for Localization:** We propose an analysis-by-synthesis method that inverts a Neural Acoustic Field (NAF) to perform listener localization from a single RIR measurement and a known source position.
- **Optimization Strategies:** We design and evaluate both grid-based and particle swarm optimization (PSO) strategies to efficiently navigate the complex acoustic loss landscape. In practice, PSO reduces runtime by up to 75% compared to grid search while simultaneously improving accuracy by 14%.
- **STFT Sub-sampling Improvement:** We introduce and evaluate a Standard Deviation-Weighted Sampling technique, mitigating the significant computational cost of the analysis-by-synthesis loop by focusing on the most informative time-frequency bins in the RIR spectrum. This reduces computation by up to 40× with only a 22% increase in localization error.
- **Evaluation in Sparse-Data Scenarios:** We demonstrate that our method outperforms direct regression baselines on a real-world dataset in scenarios where training data is limited. Specifically, it achieves up to 32% lower mean localization error than the best-performing baseline when trained on only 10% of the RAF dataset.

1.3 Thesis outline

The remainder of this thesis is structured as follows. Chapter 2 reviews the literature related to this work. Chapter 3 presents the proposed approach in detail. The system is evaluated through experiments in Chapter 4. Finally, Chapter 5 concludes the thesis by discussing key findings, limitations, and directions for future work.

Chapter 2

Related work

This chapter reviews related work in acoustic localization, implicit neural representations (INRs), and analysis-by-synthesis (AxS). Specifically, Section 2.1 covers related works in acoustic localization, distinguishing between classical model-based approaches and modern data-driven techniques. Section 2.2 reviews implicit neural representations, with a focus on NeRF and its acoustic counterpart, i.e., NAF. Section 2.3 covers analysis-by-synthesis, from its introduction to recent implementations using NeRF. Section 2.4 reviews public datasets of room impulse responses, including two that serve as the basis for evaluation in this thesis.

2.1 Acoustic localization

Indoor localization is crucial for enabling various location-based services, such as indoor navigation, health rehabilitation, and human-computer interaction (HCI) [32]. A wide variety of signal modalities have been explored for this purpose, such as Wi-Fi [6, 5, 27], Bluetooth [3, 17], visible light [25], and inertial sensors [21]. To improve localization performance, many systems integrate multiple modalities to enhance accuracy and robustness [62].

Among these modalities, acoustic signals stand out for offering high localization accuracy with minimal infrastructure and low latency [32]. The remainder of this review focuses on acoustic localization and distinguishes between two main families of approaches: classical model-based methods [49, 36, 55, 51] and data-driven neural methods [59, 46, 37, 1, 63, 20, 60]. Section 2.1.1 discusses classical, model-driven techniques, such as ToF [49], TDOA [36, 55], and MUSIC [51], that rely on explicit acoustic propagation models. Section 2.1.2 covers data-driven, neural network-based solutions, which learn end-to-end mappings from acoustic inputs to spatial coordinates, often providing greater robustness in reverberant or noisy settings.

2.1.1 Classical techniques

Various physical phenomena can be exploited using model-based approaches for acoustic localization. One of the simplest methods is Time-of-Flight (ToF), which uses the time it takes for a signal to travel from its transmitter to its

receiver. Knowing the speed of sound, one can calculate the distance d between the transceiver pair using a simple formula: $d = c \cdot t$, where c is the speed of sound in the medium, and t is the time delay measured. With multiple fixed-position anchor nodes, these estimated distances define circles in 2D or spheres in 3D around each anchor. The target's position is then determined at the intersection of these loci, typically using trilateration [29]. Despite its simplicity, ToF has several limitations. The speed of sound is temperature-dependent, introducing potential errors in distance estimation if environmental conditions are not accounted for [32]. Moreover, the method typically requires precise synchronization between the transmitter and receiver clocks, which can be challenging to achieve in practice [32].

Another widely used method is Time Difference of Arrival (TDOA) [29], which estimates the origin of a sound by measuring the relative delays between signals received at microphones in known positions. Since it relies only on these relative time differences, it does not require clock synchronization. This technique can be used to estimate the absolute position of a target, as demonstrated in [36], which uses a 2-meter diameter microphone array to localize a sound source in 3D space, achieving a mean accuracy of 4.8 cm. However, when the source is far relative to the array size, the incoming wavefronts appear planar, making it difficult to resolve the source's exact location. In such cases, only the direction-of-arrival (DOA) can be reliably estimated, while range information is lost [55].

To estimate the time delays for TDOA, cross-correlation between microphone signals is commonly used. The generalized cross-correlation with phase transform (GCC-PHAT) is a widely used method that compares only the phase of signals at each microphone to estimate time delays [26].

The mathematical formulations forming the basis for the TDOA algorithm, specifically for a two-dimensional setting, are:

$$\left\{(x_s - x_2)^2 + (y_s - y_2)^2\right\}^{1/2} - \left\{(x_s - x_1)^2 + (y_s - y_1)^2\right\}^{1/2} = T_{21}c, \quad (2.1)$$

$$\left\{(x_s - x_3)^2 + (y_s - y_3)^2\right\}^{1/2} - \left\{(x_s - x_1)^2 + (y_s - y_1)^2\right\}^{1/2} = T_{31}c \quad (2.2)$$

where (x_s, y_s) are the coordinates of the sound source, (x_i, y_i) are the coordinates of the i-th microphone, T_{ij} is the time difference of arrival between microphone i and j, and c is the speed of sound in the medium [33].

A popular method for high-resolution DOA estimation is the Multiple Signal Classification (MUSIC) algorithm, first introduced by Schmidt in 1986 [51]. It decomposes the covariance matrix of a microphone array into "signal" and "noise" subspaces and locates sources by finding steering vectors orthogonal to the noise subspace. Unlike ToF and TDOA methods, which triangulate a source's position from discrete time delays and are constrained by sampling resolution and synchronization, MUSIC operates on narrowband signals, does not need clock alignment, and can resolve sources spaced closer than what conventional array methods allow.

2.1.2 Data-driven techniques

Traditional methods rely on physical models but often struggle with noise, reverberation, and synchronization requirements. Data-driven techniques address these limitations by learning spatial and temporal patterns directly from acoustic data, enabling more robust localization in complex environments. As a result, an increasing number of acoustic localization systems based on deep neural networks (DNNs) have been proposed in recent years [16]. This section briefly reviews such approaches.

Some methods utilize raw audio waveforms directly as inputs [59, 46, 37]. This idea leverages the DNN's ability to learn optimal representations for acoustic localization without hand-crafted features or preprocessing [16]. However, this often leads to more complex networks, as a part of the network needs to be responsible for feature extraction. As a result, many studies use common signal processing representations that emphasize spatial and/or time-frequency characteristics of the signal, such as Short-Time Fourier Transform (STFT) spectrograms [16]. In systems using multiple microphones, STFT spectrograms are typically 3D tensors with dimensions for time, frequency, and channel [16]. In [1], a CNN-based architecture is used for sound event detection (SED) and DOA estimation for a set of sound classes. The model operates directly on STFT spectrograms derived from microphone array recordings. It achieved competitive DOA accuracy (3.4° error for single-source cases) but showed performance degradation in reverberant conditions.

Several learning-based approaches first extract handcrafted acoustic features such as generalized cross-correlation (GCC) vectors before processing them through neural networks for direction-of-arrival estimation. While some methods target only DOA prediction (e.g., achieving 1.37° RMSE in [63] and 4.18° mean error in [20]), others extend to full 2D localization [60]. However, these methods face inherent limitations: they rely on handcrafted features that perform poorly in noisy or reverberant environments, and their fixed preprocessing stages restrict the model's capacity to learn directly from raw input signals.

2.2 Implicit Neural Representations (INR)

Recent studies have shown that fully connected networks can serve as continuous and memory-efficient Implicit Neural Representations (INR) for modeling objects and scenes [52, 50]. One of the key breakthroughs in the field of INR was the introduction of Neural Radiance Fields (NeRF) [40], which model 3D scenes by learning a mapping from spatial coordinates and viewing directions to color and density. This enables photorealistic novel view synthesis from a sparse set of input images.

NeRF inspired a wide range of extensions [38, 66, 47, 64] and applications across modalities, such as LiDAR [22], radio-frequency (RF) [67, 23], and acoustics [34, 54, 30]. This shift toward continuous implicit representations opens new possibilities for efficient and generalizable spatial modeling, especially in scenarios with limited or sparse data [64].

Section 2.2.1 presents a brief overview of NeRF's principles, while Section 2.2.2 introduces Neural Acoustic Fields (NAF), NeRF's acoustic counterpart on which our method builds.

2.2.1 Neural Radiance Fields (NeRF)

Neural Radiance Fields (NeRF), introduced by Mildenhall et al. [40], marked a significant breakthrough in the field of implicit neural representations for 3D scene modeling. NeRF represents a static scene using a fully connected neural network that maps continuous 3D coordinates and viewing directions to volumetric density and emitted radiance. This enables the synthesis of highly realistic images from arbitrary, novel viewpoints given only a sparse set of calibrated input images. Unlike traditional voxel grids or point-based methods, NeRF is both memory-efficient and resolution-independent due to its continuous nature. At the core of NeRF is a neural network F_Θ that approximates the following volumetric scene function:

$$F_\Theta(\mathbf{x}, \mathbf{d}) = (\sigma, \mathbf{c}), \tag{2.3}$$

which maps the 3D location $\mathbf{x} \in \mathbb{R}^3$ and viewing direction $\mathbf{d} = (\theta, \phi)$ to a volume density $\sigma \in \mathbb{R}_{\geq 0}$ at that location and an RGB color $\mathbf{c} \in \mathbb{R}^3$ emitted in that direction. To render a pixel, NeRF casts a ray through the scene and samples a sequence of N points along the ray. The color of the pixel is then computed using volume rendering techniques, integrating the contributions of sampled points based on their predicted density and color [39]:

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^{N} T_i \left(1 - \exp(-\sigma_i \delta_i)\right) \mathbf{c}_i, \quad \text{where} \quad T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right), \tag{2.4}$$

where $\exp(\cdot)$ denotes the exponential function, T_i is the accumulated transmittance up to sample i, and δ_i is the distance between adjacent samples along the ray. This rendering process is differentiable, meaning that NeRF can be trained end-to-end using only image supervision by minimizing the mean squared error between rendered and ground-truth pixels. An overview of the NeRF representation and rendering process is shown in Figure 2.1 [40].

2.2.2 Neural Acoustic Fields (NAF)

Neural Acoustic Fields (NAFs) extend the concept of implicit neural representations to the acoustic domain by learning a continuous mapping from spatial coordinates to room impulse responses (RIRs). Inspired by NeRF [40], the NAF framework was introduced by Luo et al. [34] as a way to model the acoustic behavior of an environment directly from measurements, without relying on explicit room geometry or simulation. A NAF takes a source and receiver coordinate pair as input and predicts the corresponding RIR. Applying this RIR to an anechoic audio signal simulates how the sound would be perceived if emitted and recorded at those specific locations. The applications of NAF include spatial

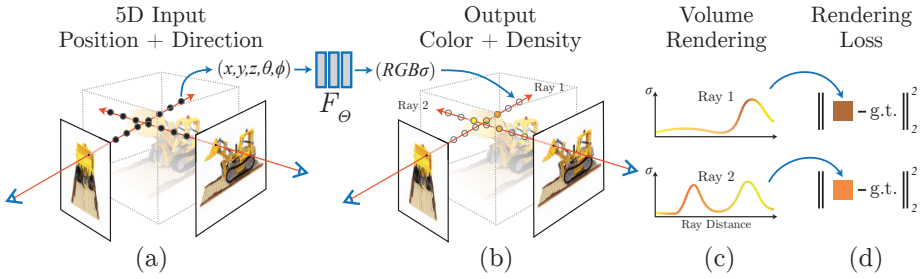

Figure 2.1: **Overview of the NeRF procedure adopted from [40]. Images are synthesized by sampling 5D coordinates (location and viewing direction) along camera rays (a), and these locations are fed into an MLP to produce a color and volume density (b), and volume rendering techniques are used to combine these values into an image (c). This rendering function is differentiable, the scene representation can be optimized by minimizing the difference between synthesized and ground-truth images (d).**

audio rendering for virtual or augmented reality (VR/AR), as well as acoustic scene analysis. For instance, properties like room geometry are implicitly encoded in the NAF and can be leveraged for downstream tasks such as geometry inference, as demonstrated by the authors of NAF [34]. The NAF model can be described using the following function:

$$F_\Theta(\mathbf{q}, \theta, k, \mathbf{q}') = \mathbf{v}, \tag{2.5}$$

where $\mathbf{q} \in \mathbb{R}^3$ represents the listener location, $\mathbf{q}' \in \mathbb{R}^3$ the emitter location, $\theta \in \mathbb{R}^2$ the listener orientation, and $k \in \{0, 1\}$ the ear (binary left/right). Here, the output $\mathbf{v} \in \mathbb{R}^T$ is the time-domain impulse response waveform [34]. This model assumes a directional, binaural listener, similar to how a human listener would perceive sound. In fact, directly outputting the time-domain impulse response is difficult due to its high-dimensional and chaotic nature [34]. Therefore, in practice, the impulse response is represented using the STFT, which is more convenient for neural network prediction due to the smoother nature of the time-frequency space. The final parametrization becomes:

$$F_\Theta(\mathbf{q}, \theta, k, \mathbf{q}', t, f) = [\mathbf{v}_{\text{STFT\_mag}}(t, f), \mathbf{v}_{\text{STFT\_IF}}(t, f)]\,, \tag{2.6}$$

where t and f represent the time and frequency coordinates in the STFT spectrogram, respectively. The outputs $[\mathbf{v}_{\text{STFT\_mag}}, \mathbf{v}_{\text{STFT\_IF}}]$ are the magnitude and phase angle components for that given time and frequency coordinate [34].

To improve convergence and spatial detail at high frequencies, positional encodings are applied to the inputs of NAF, following NeRF's design choices. The reason is that deep networks are biased towards learning lower frequency functions [40]. Having the network operate on raw coordinates results in renderings that perform poorly at representing high-frequency variations in color, in the case of NeRF, and audio, in the case of NAF. To incorporate local geometric detail into the NAF model, the scene is divided into a regular grid of k pixels

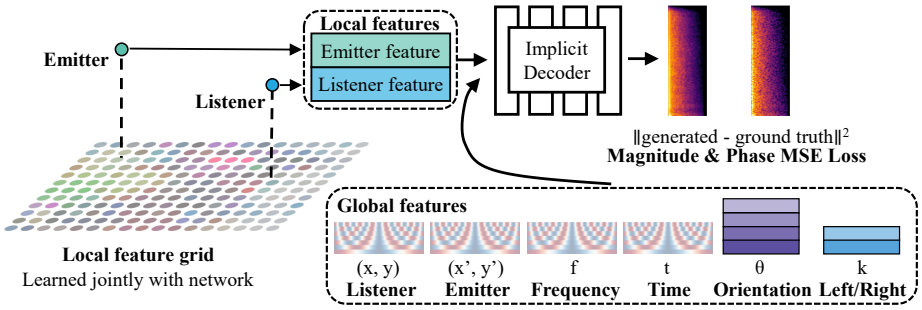

Figure 2.2: **Overview of NAF procedure adopted from [34]. Given emitter and listener coordinates, local features are interpolated from a shared learnable feature grid. These are combined with sinusoidal embeddings of spatial position, frequency, and time, along with discrete embeddings for orientation and ear side. The combined representation is passed through an implicit decoder to predict the room impulse response in both magnitude and phase. The model is trained using an MSE loss between predicted and ground truth spectrograms.**

$P = \{P_1, \ldots, P_k\}$, each storing a learnable feature vector. For any query location (x, y), such as the coordinates of an emitter or receiver, local features are interpolated from the grid using a differentiable weighting function. Specifically, a Nadaraya-Watson estimator [41] with a Gaussian kernel assigns weights to each nearby pixel based on its distance to the query point. The final interpolated feature is a weighted sum of the surrounding grid features. Because the interpolation is fully differentiable, the feature vectors at each grid cell are optimized jointly with the rest of the network during training. These interpolated local features are then combined with sinusoidal positional encodings of the input coordinates and discrete scene embeddings before being passed to the MLP. This combination allows the network to model both global spatial variation and fine-grained local acoustic structure. Figure 3.2 illustrates the NAF method.

2.3 Analysis-by-synthesis

The analysis-by-synthesis (AxS) framework models perception as an active process [18, 42]. Instead of passively interpreting input, the system generates internal hypotheses about the world, synthesizes expected input based on those hypotheses, and compares it to the actual observation. The error is then minimized by adjusting the hypothesis. This approach has a long history in speech perception [18] and has recently gained renewed attention due to its potential for neural modeling.

In language perception, AxS explains how listeners can understand speech even when the input is unclear or noisy. Instead of directly decoding the sound, the brain generates possible interpretations based on prior knowledge of language, simulates what those would sound like, and compares them to the actual input. This process repeats until the internal guess closely matches what was heard. The concept of AxS was further elaborated by Neisser in [42]. Decades later, it was revitalized in [4], in which it is proposed as a unifying framework

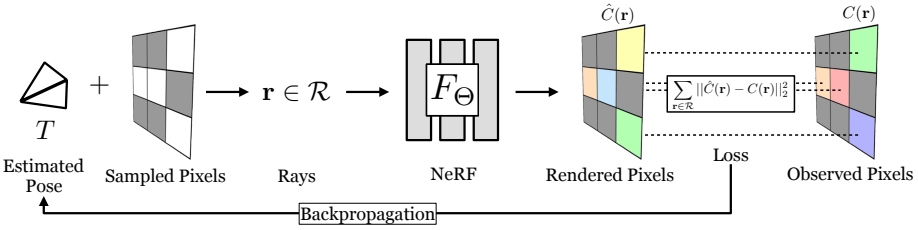

Figure 2.3: **Overview of iNeRF adopted from [31]. Starting from an initial pose estimate, selected rays are rendered using NeRF, and the pose is iteratively refined by minimizing the difference between rendered and observed pixels through end-to-end differentiable optimization.**

for language and vision, grounded in evidence from cognitive neuroscience.

2.3.1 Analysis-by-synthesis in Vision and Acoustics

In computer vision, Brachmann et al. [7] estimate the 6D pose of objects by predicting 3D object coordinates from images and matching them to the observed depth data. While not explicitly framed as AxS, the method fits the same structure: predict, compare, and adjust. More recent work has used neural networks to implement this idea directly. Chen et al. [11] use a network to render synthetic views of objects at different poses, then optimize the pose by comparing these to the input image. This AxS loop outperforms direct pose regression. Building on NeRF, iNeRF [31] estimates pose by rendering images from a neural radiance field and optimizing the pose so that the rendered view matches the input. Starting from an initial pose guess, iNeRF emits a set of rays, samples points along them, and renders pixel values using the trained NeRF model. The pose is then refined by minimizing the difference between the rendered and observed pixels through gradient-based optimization. This AxS framework allows accurate pose estimation. An overview of the pipeline is shown in Figure 2.3 [31].

Disentangled Acoustic Fields (DAFs) [65] apply AxS to acoustic fields by reconstructing a sound's power spectral density (PSD) from disentangled factors: object location, material, type, and a latent scene variable. Unlike STFT representations that struggle with temporal silence gaps, PSD provides a stable frequency-domain target for optimization. The scene latent captures environment-specific acoustics (e.g., reverberation) separately from object properties, enabling generalization across rooms. By comparing synthesized and observed PSDs, DAFs generate uncertainty maps that guide robotic search for fallen objects, outperforming direct regression by 14% in unseen environments. The DiffRIR framework [61] applies analysis-by-synthesis to room acoustics by reconstructing spatial audio from sparse RIR measurements. Unlike DAFs, which model sound through disentangled latent variables, DiffRIR uses interpretable parametric models for sound source directivity and surface reflectivity, optimizing time-domain RIRs rather than PSDs. It combines geometric acoustics (e.g., image-source methods) with a learned residual component, enabling efficient differentiable rendering for immersive audio synthesis. The parameters

of the acoustic model are optimized using analysis-by-synthesis, comparing a synthesized RIR to a ground-truth RIR measured at the same location.

2.4 Datasets

To evaluate the effectiveness of acoustic methods, many acoustic datasets have been introduced over the years. This section will focus specifically on spatial room impulse response (RIR) datasets, since NAF [34] is trained on this type of acoustic data. Section 2.4.1 will cover simulated datasets, while Section 2.4.2 will discuss real-world recorded datasets.

2.4.1 Simulated datasets

A widely used dataset in NAF research is SoundSpaces [9], which provides simulated binaural room impulse responses (RIRs) sampled from emitter-listener pairs on a 0.5 m resolution 2D grid. These simulations cover 85 3D-scanned real-world indoor environments from the Matterport3D dataset [8], as well as 18 high-quality reconstructed scenes from the Replica dataset [53]. NAF and its successors [34, 54, 30] are evaluated on six selected scenes from the Replica dataset. SoundSpaces 2.0 [10] replaces the fixed grid with a geometry-based rendering engine for simulating RIRs at any location or orientation, and adds extensive configuration options. Other simulated datasets include BIRD [15], which contains 100,000 multichannel RIRs generated using the image-source method in randomly sampled, empty rectangular rooms. GWA [58] uses nearly 19,000 professionally designed CAD models of indoor scenes but provides relatively few samples per scene compared to SoundSpaces.

2.4.2 Real-world datasets

Besides these simulated datasets, real-world recorded datasets are essential for evaluating performance of methods like NAF in practical scenarios. Recording a dense dataset of RIRs is a very time-consuming task, which means realism often comes at the expense of data diversity. The recently introduced Real Acoustic Fields (RAF) dataset [12] offers a dense collection of 3D sampled RIRs from two environments, which are furnished and unfurnished versions of the same room. Other real-world datasets, such as MeshRIR [28], BUT Reverb [57], and GTU-RIR [45], were not used in this work since their spatial sampling density is considerably lower compared to RAF. An overview of available datasets is given in Table 2.1.

2.5 Research Gap and Motivation

Existing acoustic localization research, spanning classical methods (e.g., TDOA [36], MUSIC [51], Section 2.1.1) and data-driven approaches (e.g., CNNs on STFT or GCC features [1, 63], Section 2.1.2), primarily focuses on sound source localization using microphone arrays, leaving listener (microphone) localization underexplored despite its potential for localizing household robots, smartphones, or people [44]. Focusing on single-microphone setups, as opposed to arrays,

Dataset	Type	RIRs	Scenes	Scene Types
SoundSpaces [9]	Sim.	17.6M	103	Scanned indoor scenes
SoundSpaces 2.0 [10]	Sim.	-	-	Any input mesh
BIRD [15]	Sim.	100K	100K	Empty shoebox rooms
GWA [58]	Sim.	2M	18.9K	Professional CAD models
MeshRIR [28]	Real	4.4K	2	Acoustic lab
BUT Reverb [57]	Real	1.3K	8	Uni meeting rooms
GTU-RIR [45]	Real	15.2K	11	Uni meeting rooms
RAF [12]	Real	86K	2	Furnished + Unfurnished

Table 2.1: **Overview of popular RIR datasets, both simulated and real. It is clear that simulated datasets offer more samples and scenes, but this comes at the expense of realism.**

enhances compatibility with off-the-shelf devices like smartphones, reducing hardware complexity. Direct regression methods excel in high-data settings but often struggle in low-data or reverberant conditions [1]. At the same time, analysis-by-synthesis (AxS) approaches, as seen in vision (iNeRF [31], Section 2.3) and acoustics (DAF [65], DiffRIR [61], Section 2.3), offer strong performance in sparse-data scenarios and enhanced capabilities such as uncertainty maps. Neural Acoustic Fields (NAF [34], Section 2.2.2) effectively model RIR synthesis, yet their potential for inverse localization tasks remains unexplored. This work addresses these gaps by leveraging a trained NAF to perform listener localization with a known source in low-data, single-microphone setups, enabling accurate localization across varied and acoustically complex spaces.

Chapter 3

Method

This chapter provides a detailed overview of the NAF-based localization method proposed in this thesis. We formulate the problem statement in Section 3.1 and introduce the system design in Section 3.2.

3.1 Problem Statement

Forward problem. Neural Acoustic Fields (NAFs) address the forward problem of acoustic synthesis, i.e., predicting the resulting room impulse response (RIR) from a known configuration of source, listener, and environment. In this setting, the positions are given, and the task is to generate the corresponding acoustic impulse response. We follow the formulation of Luo et al. [34], but we add the emitter orientation term θ_e to the inputs, enabling the model to account for directional emitters. We define the forward model as follows.

Let F_Θ denote a neural acoustic field parameterized by Θ. Given an emitter position $\mathbf{q_e} \in \mathbb{R}^3$ with (optional) orientation θ_e, a listener position $\mathbf{q}_\ell \in \mathbb{R}^3$ with (optional) orientation θ_ℓ, and a channel index $k \in \{1, \ldots, K\}$ (e.g., K=1 for monaural, K=2 for binaural), F_Θ predicts the STFT-domain RIR, denoted by $\mathbf{v}^{(k)}(t, f)$, for each time-frequency bin (t, f):

$$F_\Theta(\mathbf{q}_\ell, \theta_\ell, k, \mathbf{q_e}, \theta_e, t, f) = \mathbf{v}^{(k)}(t, f). \tag{3.1}$$

F_Θ captures multipath reflections and reverberation without explicit geometric modeling. In our work, we consider only the STFT magnitude, omitting the phase component, as opposed to the formulation in Equation (2.6). While the original NAF predicts both magnitude and phase [34], the authors note on their GitHub repository [35] that omitting the phase results in lower spectral error, indicating that NAF is more effective at predicting magnitude. Therefore, our implementation solely predicts the magnitude component.

Inverse problem (listener localization). We now formulate the inverse problem, which is the focus of this thesis: given a trained NAF F_Θ, a known emitter position $\mathbf{q_e}$ (and, if directional, its orientation θ_e), and an observed RIR (STFT spectrum) with K listener channels, $\mathbf{V} = \{\mathbf{v}^{(k)}\}_{k=1}^{K}$, the goal is to estimate the listener position $\mathbf{q}_\ell$. In the most general setting, one could also

estimate θ_ℓ for directional listeners, but in our current implementation, θ_ℓ is assumed to be known in the case of a directional listener.

The reconstruction loss between the ground-truth RIR spectrum $\mathbf{V}$ and the synthesized RIR spectrum $\hat{\mathbf{V}}$ is defined as the mean squared error (MSE) computed over all listener channels, time frames, and frequency bins:

$$\mathcal{L}(\mathbf{V}, \hat{\mathbf{V}}) = \frac{1}{KTF} \sum_{k=1}^{K} \sum_{t=1}^{T} \sum_{f=1}^{F} \left(\mathbf{v}^{(k)}(t,f) - \hat{\mathbf{v}}^{(k)}(t,f) \right)^2, \tag{3.2}$$

where $\hat{\mathbf{v}}^{(k)}$ is synthesized by F_Θ as described in Equation (3.1), K represents the number of channels, and T and F represent the number of time and frequency bins in the STFT spectrum. Inverting the formulation from [34], we can define the following inverse problem statement:

$$\hat{\mathbf{q}}_\ell = \arg\min_{\mathbf{q}_\ell \in \Omega} \mathcal{L}(\mathbf{V},\, F_\Theta(\mathbf{q}_\ell, \theta_\ell, \mathbf{q_e}, \theta_e)), \tag{3.3}$$

where $\hat{\mathbf{q}_\ell}$ is the estimated listener position and Ω denotes the search region, which is limited to the interior of the room. The room geometry is assumed to be known, so Ω is simply taken as the bounding box of the room. When θ_ℓ is unknown and the listener is directional, the search could be extended to $(\mathbf{q}_\ell, \theta_\ell)$, but this is left for future work.

Assumptions (dataset-agnostic). For the scope of this work, we make the following assumptions to ensure the problem remains well-posed. Including additional unknowns, such as the emitter position, would greatly increase complexity and is left beyond the scope of this thesis.

- The environment is static and indoor. F_Θ is trained on RIRs from that space.
- The emitter position $\mathbf{q_e}$ is known; if directional, its orientation θ_e is known.
- No additional sensing modalities (e.g., vision, IMU) are used.
- The room dimensions are known and used to define the search region Ω.

Dataset-specific problem statements. While the general formulation above is dataset-agnostic, each dataset introduces specific constraints on emitter directionality, listener directionality, channel count, and whether orientation is known or estimated. For clarity, we restate the inverse problem for both datasets used in the evaluation of the system, in the notation of Equation (3.3):

- **RAF [12]:** directional emitter with fixed orientation; omnidirectional, single-channel listener ($K = 1$). Listener orientation θ_ℓ is irrelevant. We solve:
$$\hat{\mathbf{q}}_\ell = \arg\min_{\mathbf{q}_\ell \in \Omega} \mathcal{L}\big(\mathbf{V},\, F_\Theta(\mathbf{q}_\ell, -, \mathbf{q_e}, \theta_e)\big).$$
- **SoundSpaces [9]:** omnidirectional emitter; directional, binaural listener (K=2). We *assume known* listener orientation θ_ℓ and optimize only $\mathbf{q}_\ell$:
$$\hat{\mathbf{q}}_\ell = \arg\min_{\mathbf{q}_\ell \in \Omega} \mathcal{L}\big(\mathbf{V},\, F_\Theta(\mathbf{q}_\ell, \theta_\ell, \mathbf{q_e}, -)\big).$$

Estimating listener orientation is feasible by extending the optimization in Equation (3.3) to jointly minimize over $\mathbf{q}_\ell$ and θ_ℓ using grid search or PSO, leveraging binaural cues like inter-channel time/energy differences for uniqueness. This is beyond the scope of this thesis, which focuses on listener position estimation.

3.2 System Design

This section presents the design of the proposed NAF-based listener localization system. We begin with data processing (Section 3.2.1), where the derivation and pre-processing of room impulse responses (RIRs) are described. Next, we detail the NAF architecture and training procedure (Section 3.2.2), which form the core of the system. Building on this, we introduce the localization pipeline (Section 3.2.3), followed by optimizations aimed at improving computational efficiency (Section 3.2.4).

3.2.1 Data Processing

RIR Measurement and Simulation. The room impulse response can be obtained either through direct measurement or simulation. In real-world scenarios, RIRs are typically obtained through direct measurement. Classical measurement techniques include the exponential swept-sine method [14] and maximum-length sequences (MLS) [48], both widely used for acoustic characterization of real spaces. The Real Acoustic Fields (RAF) dataset [12], used in our evaluation, provides RIRs measured in real rooms with exponential sine sweeps, thereby capturing authentic acoustic effects. Alternatively, RIRs can be generated synthetically using acoustic models, such as the image source method [2]. The SoundSpaces dataset [9] is created using this strategy, rendering RIRs via geometric acoustic simulation that combines the image source method with ray tracing.

RIR Pre-processing. The NAF is trained on STFT log-spectrograms of the RIRs (see Figure 3.1 for an example of a time-domain RIR and its STFT representation). Directly training on time-domain RIRs is difficult due to their high-dimensional and chaotic nature [34]. The short-time Fourier transform (STFT) represents a signal in both time and frequency by computing the Fourier transform over short, overlapping windows. For a discrete-time signal $x[n]$ and a window function $w[n]$, the STFT is defined as [43]:

$$X[t, \omega] = \sum_{n=-\infty}^{\infty} x[n]\, w[n-t]\, e^{-j\omega n}, \tag{3.4}$$

where t indexes the time frame and ω is the angular frequency. We use the STFT implementation of the `librosa` Python package [13] with the default Hann window. We use an FFT size of $N_{\text{fft}} = 512$ and a hop length of 128 samples, following the configuration of [34]. The magnitude of the STFT is then converted to the log scale and normalized. The mean $\mu_{(t,f)}$ and standard deviation $\sigma_{(t,f)}$ are computed for each time/frequency index, and each "pixel" of the spectrogram is normalized as follows [34]:

$$\mathbf{v}_{\text{STFT\_mag}}(t, f) = \frac{\mathbf{v}_{\text{STFT\_mag}}(t, f) - \mu_{(t,f)}}{3.0 \times \sigma_{(t,f)}}. \tag{3.5}$$

Although the scaling factor of 3.0 is not explicitly motivated in [34], a reasonable interpretation is that it serves as a practical normalization step, ensuring that most values fall within a bounded range (roughly $[-1, 1]$ under Gaussian assumptions) and thus improving numerical stability during training.

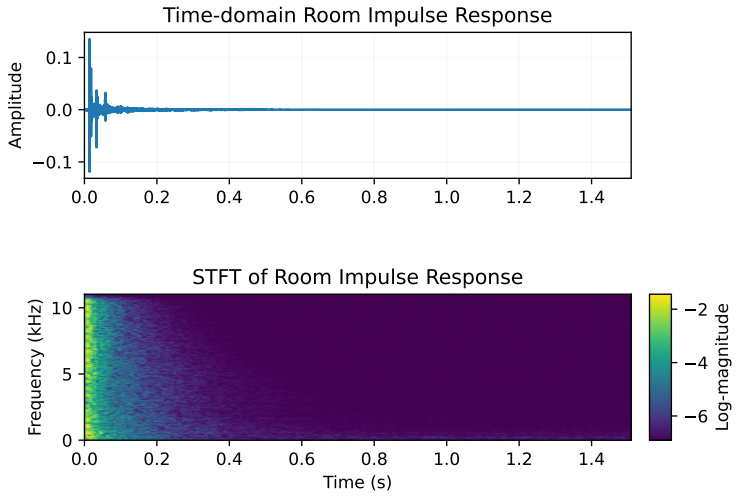

Figure 3.1: **Time-domain RIR (top) and its log-magnitude STFT (bottom) from the RAF dataset.**

3.2.2 NAF Architecture and Training

Neural Acoustic Fields (NAFs) [34] extend the idea of implicit neural representations to the acoustic domain, enabling the synthesis of Room Impulse Responses (RIRs) for arbitrary emitter-listener configurations. Inspired by Neural Radiance Fields (NeRF) [40], NAFs learn a continuous mapping from spatial coordinates (and optional orientations) to a time-frequency representation of the RIR, without requiring explicit room geometry or simulation. As formalized in Section 3.1, a trained NAF model F_Θ takes the positions and orientations of an emitter and listener, as well as a time-frequency index (t, f) as inputs, and predicts the corresponding STFT magnitude. By applying the model over all (t, f), a full RIR can be synthesized and applied to anechoic audio for spatial rendering. The localization system is based on the original NAF [34] rather than newer variations (e.g., INRAS [54], NACF [30]) due to the lack of stable open-source implementations and because our focus lies on the analysis-by-synthesis localization paradigm rather than network architecture innovation. The following four numbered paragraphs (1-4) detail the key components of the NAF and its training process, linking directly to the corresponding numbered parts in Figure 3.2.

1. Global Features and Positional Encoding: Since both the SoundSpaces and RAF datasets lack the full parametrization of an acoustic field as given in Equation (3.1), we train NAF with a dataset-specific restricted parameterization. The core inputs, always present, include the listener position (x_ℓ, y_ℓ, z_ℓ), the emitter position (x_e, y_e, z_e), and the time-frequency indices (t, f). Additional inputs are dataset-dependent: the listener orientation θ_ℓ, the emitter orientation θ_e, and the listener channel index k. These are therefore marked as optional in Figure 3.2.

A distinction can be made between *continuous* and *categorical* variables. Con-

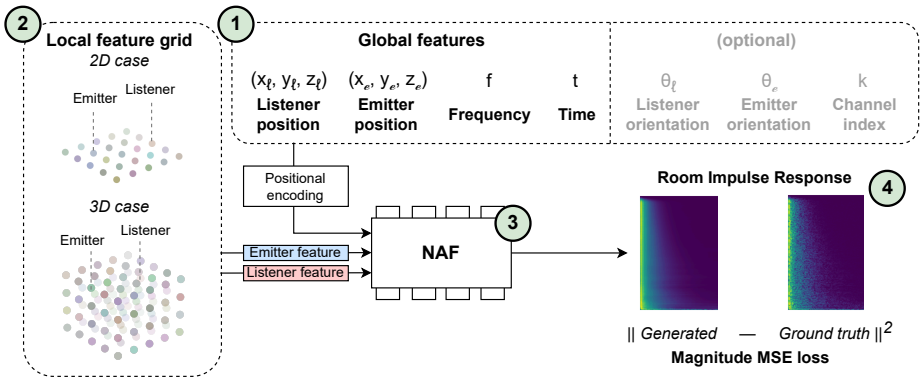

Figure 3.2: **NAF inputs and training setup. Inputs include emitter/listener positions (with optional orientations and channel selection), time, and frequency. Local features are interpolated from a learnable grid and combined with the other inputs. The MLP network predicts STFT magnitudes per time-frequency bin. Training uses the magnitude MSE loss between synthesized and ground-truth spectrograms.**

tinuous variables, including positions and the (t, f) tuple, are scaled to $(-1, 1)$ and encoded with sinusoidal embeddings (as proposed by NeRF [40]) using 10 frequencies of sine and cosine. For positions, the maximum frequency is 2^7, while for time and frequency it is 2^{10}. Categorical variables, such as discrete orientations or the listener channel index, are represented with learned embeddings.

The datasets differ in how positions and orientations are represented. In SoundSpaces, listener and emitter positions are only provided in 2D $(x_\ell, y_\ell), (x_e, y_e) \in \mathbb{R}^2$, with a binaural directional listener. Emitters are omnidirectional, while the listener orientation θ_ℓ is limited to four discrete values $\{0°, 90°, 180°, 270°\}$, embedded via a learned matrix of size $\mathbb{R}^{4\times n}$, from which a single $\mathbb{R}^{1\times n}$ vector is selected. The listener channel index $k \in \{0, 1\}$ is embedded using a matrix of size $\mathbb{R}^{2\times n}$. In RAF, positions are fully 3D $(x_\ell, y_\ell, z_\ell), (x_e, y_e, z_e) \in \mathbb{R}^3$, and listeners are omnidirectional and single-channel. Emitters have a continuous orientation θ_e represented as a quaternion, which is passed through a sinusoidal encoding before feeding it to the network.

2. Local Feature Grid: To incorporate local geometric detail, the acoustic space is discretized into a regular 2D or 3D grid of learnable feature vectors. For any query location (emitter or listener), local features are interpolated from this grid using a differentiable weighting function. Specifically, a Nadaraya-Watson estimator [41] with a Gaussian kernel assigns weights to each nearby grid cell based on Euclidean distance. The interpolated feature vector is a weighted sum of neighboring grid features and is jointly optimized with the rest of the network parameters during training. The inputs and feature grid are illustrated in Figure 3.2, adopted from [34]. The original NAF formulation uses a 2D grid, evaluated on the SoundSpaces dataset with fixed height, but this would impose

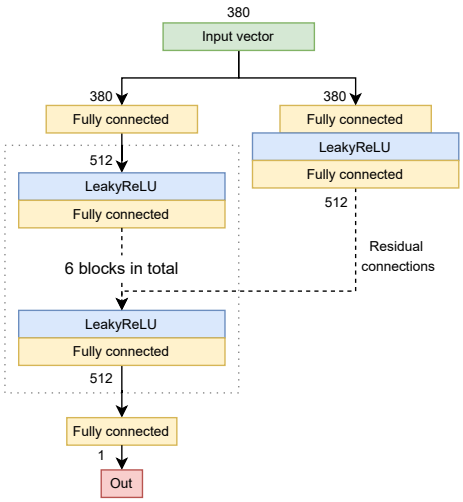

Figure 3.3: **NAF architecture taken from [34]. RIR spectrogram pixels are predicted individually, with (t, f) bins supplied as inputs.**

a height constraint when applying the NAF for localization. Our 3D implementation extends this to (x, y, z) coordinates, capturing height-dependent reflection patterns and enhancing performance in datasets with vertical variation, such as RAF [12].

3. Neural Network Structure: The interpolated local features are concatenated with the positional encodings of spatial coordinates, the time index t, the frequency index f, and any discrete scene or orientation embeddings. This combined representation is passed to a multi-layer perceptron (MLP), illustrated in Figure 3.3, that outputs the STFT magnitude for the given (t, f) bin. As each bin is predicted independently, a complete spectrogram requires evaluating F_Θ over all (t, f) pairs.

4. Model Training and Loss Function: The model is trained using mean squared error (MSE) between predicted and ground-truth spectrograms, computed over all listener channels, time frames, and frequency bins, as defined in Equation (3.2). A small amount of noise sampled from $\mathcal{N}(0, \varepsilon_{\text{reg}})$ is added to both the ground-truth emitter and listener coordinates during training, to prevent degenerate solutions [34]. The regularization parameter ε_{reg} (originally set to 0.1 in [34]) regulates the amount of noise added. In our implementation, ε_{reg} was reduced to 0.01 to improve localization precision (Section 4.3.5).

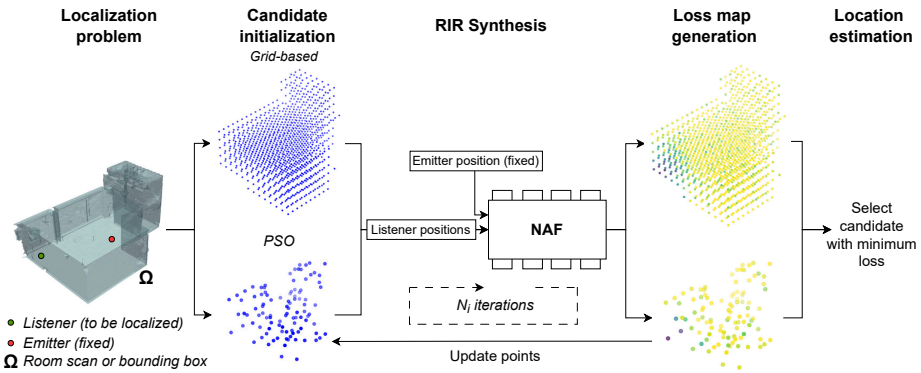

Figure 3.4: **Overview of the acoustic localization pipeline using NAF. The framework estimates the listener's position given a fixed emitter position using a recorded room impulse response (RIR). NAF synthesizes RIRs for candidate listener positions, enabling loss map generation through MSE comparison with the recorded RIR. Two point selection strategies are employed: non-iterative grid-based search and Particle Swarm Optimization (PSO) with updates over N_i iterations. The position with the minimum loss is selected as the final estimate.**

3.2.3 Localization Pipeline

Localization Problem. The inverse localization process follows an *analysis-by-synthesis* strategy inspired by iNeRF [31]. Given a recorded RIR from an unknown listener position, the system synthesizes candidate RIRs using the trained NAF for locations within a bounded search space Ω (defined by the room dimensions), compares them to the target RIR, and selects the position that minimizes the loss in Equation (3.2). The elements of the localization pipeline, as depicted in Figure 3.4, are discussed below.

Candidate Initialization. The localization process begins by defining a set of candidate listener positions within the bounded search space Ω. Depending on the available information, Ω may correspond to a detailed 3D scan of the room geometry, which constrains candidates to physically valid regions, or, in the absence of such data, to a coarse bounding box defined by the room dimensions. Candidate positions are then initialized either uniformly on a predefined grid or sampled from a uniform distribution when using Particle Swarm Optimization (PSO). Each candidate represents a hypothesis of the listener's true position and serves as an input for RIR synthesis. A denser sampling of candidates improves the chances of localizing the true position, though it comes with higher computational cost.

RIR Synthesis. For each candidate position $\mathbf{q}_\ell \in \Omega$, a synthetic RIR is generated using the neural acoustic field F_Θ. This step leverages the NAF as the forward model that links spatial hypotheses to their acoustic signatures. The synthesized responses are then compared against the measured RIR using the mean squared error (MSE) loss, which provides a measure of similarity.

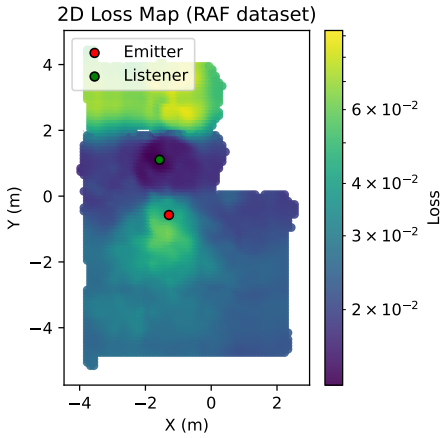

Figure 3.5: **2D slice of a loss map generated at 10 cm resolution. Although the minimum-loss point is close to the true listener location, many local minima exist, complicating gradient descent.**

Loss Map Generation. Once losses have been computed for all candidate positions, they can be aggregated into a spatial *loss map* that represents the error landscape across Ω. This map not only supports localization by identifying the minimum-loss region but also provides a diagnostic view of the search space. Sharp, well-defined minima indicate a clear localization outcome, while flat regions or multiple local minima suggest that localization may be less reliable. In addition to guiding position estimation, loss maps thus offer an interpretable tool for assessing the reliability of the localization process.

Location Estimation. The estimated listener position $\hat{\mathbf{q}}_\ell$ is obtained by selecting the candidate with the minimum loss in the landscape. Other strategies may be explored in future works. For instance, one could estimate the position as a weighted average over the lowest-loss candidates, which might reduce sensitivity to noise or incorrect local minima at the cost of introducing bias. Probabilistic approaches such as maximum-likelihood or Bayesian estimation could also be used to incorporate prior knowledge of likely receiver positions or to explicitly quantify uncertainty in the final estimate. In the context of analysis-by-synthesis, selecting the lowest-loss candidate is the most straightforward strategy, as it reflects the principle that the true location is the one whose synthesized RIR best explains the measurement.

Search Strategies. Several strategies can be used to explore the search space Ω. Inspired by iNeRF [31], gradient descent is a natural choice. In their setting of camera pose estimation, an initial pose is chosen, a view is synthesized with the NeRF, and compared to the real image. The resulting MSE loss, which in training would update the network weights, is instead used to refine the pose estimate. This process is repeated until convergence to the true pose. Directly applying this to acoustics is problematic. A typical acoustic loss landscape (Figure 3.5) is full of local minima, causing gradient descent to get stuck in sub-optimal solutions. We therefore investigate two alternative approaches:

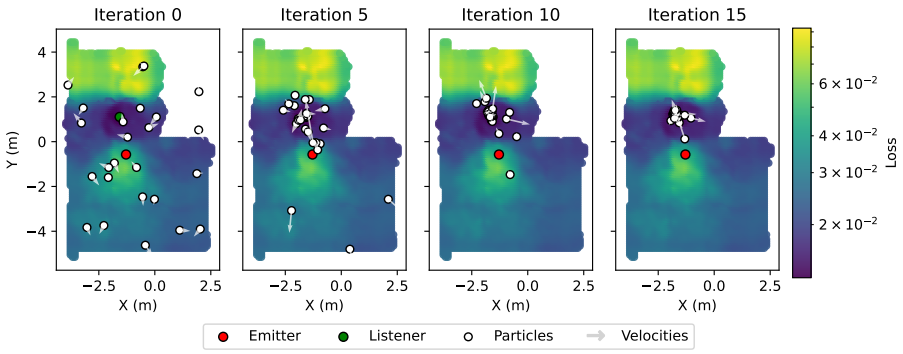

Figure 3.6: **Visualization of the particle swarm optimization (PSO) algorithm applied to explore the search space. The particles quickly converge to the minimum-loss area, providing a much more efficient solution than an exhaustive grid-based search.**

- **Grid-based:** The first is a grid-based search, where Ω is uniformly sampled in either two or three dimensions, depending on the dataset. High-resolution grids are capable of locating the global minimum accurately but incur significant computational costs. For instance, a grid with 10 cm resolution in a $4 \times 5 \times 3$ m environment already requires 60,000 NAF inferences, which may cause problems in practical scenarios where computational power is limited.

- **Particle Swarm Optimization:** The second strategy is Particle Swarm Optimization (PSO) [24], a metaheuristic that maintains a population of particles, each with an individual and group-best location. Initially, N_p particles are distributed randomly within the search space Ω. At every iteration, particles update their velocity and position by balancing inertia, personal experience, and collective information. Specifically, for particle i at iteration t, the velocity and position updates are

$$\mathbf{v}_i^{(t+1)} = w\,\mathbf{v}_i^{(t)} + c_1 r_1 \big(\mathbf{p}_i - \mathbf{x}_i^{(t)}\big) + c_2 r_2 \big(\mathbf{g} - \mathbf{x}_i^{(t)}\big),$$

$$\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t)} + \mathbf{v}_i^{(t+1)},$$

  where w is the inertia weight, c_1 and c_2 are the cognitive and social coefficients, $r_1, r_2 \sim U(0,1)$ are random scalars, $\mathbf{p}_i$ is the personal best position of particle i, and $\mathbf{g}$ is the global best position across the swarm. After each update, positions are clamped to the search space bounds. This process is repeated for N_{iter} iterations, as illustrated in Figure 3.6.

  This method converges to low-loss regions more efficiently than exhaustive grid search, though typically at the expense of some accuracy. By tuning swarm parameters (N_p, N_{iter}, w, c_1, c_2), PSO allows flexible trade-offs between speed and precision.

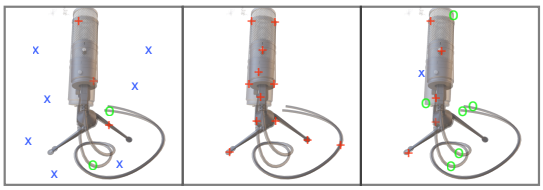

Figure 3.7: **iNeRF sampling strategies (Figure adopted from [31]). '×' denotes pixels on the common background, '+' denotes aligned pixels, and '∘' denotes informative, misaligned pixels. Interest Region sampling focuses on the latter for efficient optimization.**

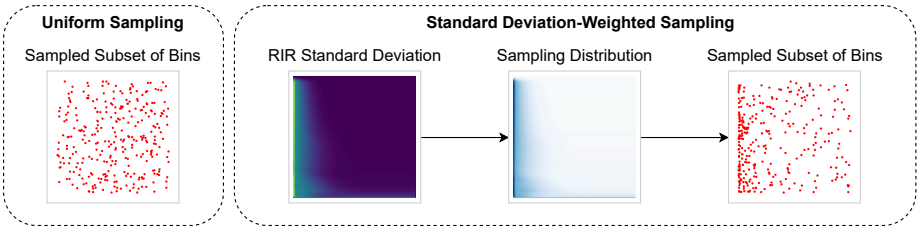

Figure 3.8: **Comparison of pixel selection strategies. In Uniform Sampling, time-frequency bins are selected with equal probability. In Standard Deviation-Weighted Sampling, a probability distribution derived from the per-bin RIR standard deviation guides selection, prioritizing STFT bins with higher variability.**

3.2.4 Optimization

A major drawback of analysis-by-synthesis is its high computational demand. To mitigate this, we take inspiration from iNeRF [31]. The core computational insight from iNeRF is to avoid generating a full synthetic output for every comparison. iNeRF's key contribution is a strategy for selecting a small, informative subset of pixels (rays) to use for gradient calculation in each optimization step, rather than rendering a full image. As illustrated in Figure 3.7, they evaluated three strategies:

- **Random Sampling** is inefficient, as many pixels (marked '×') lie on uninformative regions.
- **Interest Point Sampling** selects feature-rich pixels but often samples points that are already aligned ('+'), providing weak gradients.
- **Interest Region Sampling** proves most effective by sampling from dilated regions around interest points, favoring misaligned pixels ('∘') that provide strong directional gradients for pose correction.

This targeted sampling reduces the number of expensive rendering operations by orders of magnitude. In the acoustic setting, we apply the same principle by selecting only a subset of time-frequency bins from the STFT spectrogram for comparison. Comparable to iNeRF, randomly sampled STFT bins often fall in uninformative regions such as silent RIR segments. To address this,

a Standard Deviation-Weighted Sampling (SDWS) strategy is introduced: we compute the standard deviation for each bin across the training set and construct a sampling distribution that is proportional to these standard deviations. Bins with high variability, such as those containing the direct sound and early reflections, are prioritized, as they carry more discriminative information for localization than low-variance bins (e.g., late reverberation or inactive frequencies). As a baseline, Uniform Sampling (US), which selects bins with equal probability, is also tested. These two strategies are illustrated in Figure 3.8. As demonstrated in Section 4.3, this reduces the number of required NAF inferences per candidate location by one to two orders of magnitude, while only marginally affecting localization accuracy.

Chapter 4

Evaluation

This chapter evaluates the proposed system on two datasets: SoundSpaces [9] (simulated, binaural) and RAF [12] (real-world, monaural), detailed in Section 4.1. Localization performance is evaluated using the full training dataset (Section 4.3.1) and in sparse training data scenarios (Section 4.3.2). The effects of PSO parameters (Section 4.3.3), pixel selection (Section 4.3.4), and regularization (Section 4.3.5) are also studied. Finally, the system's parameter count and computation time are profiled in Section 4.4.

4.1 Datasets

The RAF [12] and SoundSpaces [9] datasets were selected due to their complementary strengths. RAF [12] offers real measurements in 3D spaces, capturing authentic acoustic complexities such as multipath reflections and material properties. At the same time, SoundSpaces [9] provides a larger selection of (simulated) environments. RAF is used to test real-world applicability, while SoundSpaces allows experiments across a wider variety of environments, though at the expense of realism.

For SoundSpaces, we use the provided binaural RIRs (left and right ear channels) with omnidirectional emitters and directional listeners at fixed height, adapting the system to 2D coordinates (x,y) and assuming known listener orientation θ_ℓ. The dataset covers diverse indoor scenes, and our system is evaluated on the same representative subset of six environments, listed in Table 4.1, as the original NAF work [34].

For RAF, we use monaural RIRs (single-channel) with directional emitters (fixed orientation) and omnidirectional listeners. Since the dataset spans two rooms with samples at varying heights, the model is trained with full 3D coordinates. In both cases, datasets are split into training (90%) and test (10%) sets. Positions are normalized to the room bounds, and RIRs are pre-processed into STFT log-spectrograms as described in Section 3.2.2.

4.2 Baselines and evaluation metrics

Baselines. To assess the effectiveness of the proposed NAF-based analysis-by-synthesis approach, we compare it against two direct regression baselines.

Room	Description	Dimensions	Samples
SoundSpaces (2D) [9]			
`frl_apartment_2`	Non-rectangular room	$12.9 \times 7.4\,\mathrm{m}$	240K
`frl_apartment_4`	Non-rectangular room	$7.9 \times 12.8\,\mathrm{m}$	227K
`room_2`	Rectangular room	$6.8 \times 4.9\,\mathrm{m}$	29.5K
`office_4`	Rectangular room	$6.5 \times 6.5\,\mathrm{m}$	67.6K
`apartment_1`	Multi-room apartment	$10.7 \times 7.9\,\mathrm{m}$	264K
`apartment_2`	Multi-room apartment	$9.4 \times 10.2\,\mathrm{m}$	264K
RAF (3D) [12]			
`EmptyRoom`	Empty room	$7.5 \times 9.8 \times 4.1\,\mathrm{m}$	47.5K
`FurnishedRoom`	Furnished room	$7.5 \times 9.8 \times 4.1\,\mathrm{m}$	39.1K

Table 4.1: **Overview of the SoundSpaces [9] and RAF [12] environments used for evaluation, including description, dimensions, and number of samples.**

These baselines are trained to directly predict the listener position from the input STFT spectrogram and emitter position/orientation, using mean squared error (MSE) loss on the position coordinates. We chose ResNet-10 as a widely used, lightweight CNN baseline for spectrogram regression, offering a strong but generic reference point. We designed NAF-Direct to mirror the original NAF architecture while replacing synthesis with direct regression, isolating the effect of analysis-by-synthesis versus direct prediction. Both are covered in more detail below.

- **ResNet-10**: A reduced-complexity variant of ResNet-18 [19], consisting of an initial convolutional layer followed by basic residual blocks (with each block consisting of 2 convolutional layers), global average pooling, and a fully connected output layer. It maintains the same input format as ResNet-18 but with a lower parameter count for fair comparison to NAF. For architecture details, see [19].
- **NAF-Direct**: A custom baseline inspired by the NAF architecture. A convolutional feature extractor is added to compress the STFT spectrogram into a 64-dimensional feature vector, followed by MLP layers that mirror the original NAF structure (Figure 3.2) for direct position regression. This design ensures comparable complexity to the original NAF while enabling direct localization. The NAF-Direct architecture is illustrated in Figure 4.1.

The proposed method uses the original NAF model for RIR synthesis, combined with search strategies (grid-based or PSO) for localization. Unlike the baselines, it does not perform direct regression but optimizes the position via loss minimization in the search space. The proposed method and both baselines are comparable in terms of trainable parameters, with exact numbers provided in Section 4.4.

Evaluation metrics. For evaluating localization performance, we use the 2D Euclidean localization error (in cm) between the estimated and ground-truth listener positions in the (x,y) plane. This metric ensures comparability across

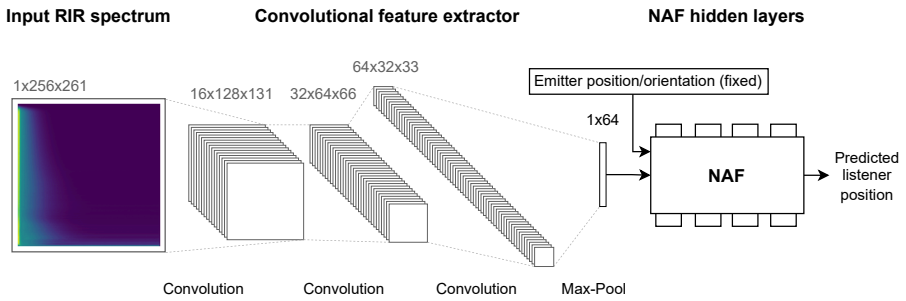

Figure 4.1: **NAF-Direct baseline architecture. A convolutional feature extractor compresses the input STFT spectrogram into a 64-dimensional feature vector, which is passed through MLP layers mirroring the NAF structure, but with a modified output layer that outputs the listener position.**

datasets: SoundSpaces is inherently 2D with fixed height, while for RAF (3D), we project the error by ignoring the z-component, focusing on horizontal accuracy, which is often the primary concern in indoor localization. All results are averaged over 100 random emitter-listener combinations per environment, randomly selected from the test set, with errors reported in centimeters.

Fixed Parameters. Unless specified otherwise, the experiments use:

- NAF as described in Section 3.2.2 (3.6M params, $\varepsilon_{\text{reg}} = 0.01$).
- Grid-based search at 25 cm grid resolution.
- PSO configured with $N_p = 250$ particles, $N_{\text{iter}} = 10$ iterations, inertia $w = 0.7$, and cognitive/social coefficients $C_1 = 1.5$, $C_2 = 1$.
- Standard Deviation-Weighted Sampling (SDWS) at 2.5% STFT pixels.
- Full datasets (100% of training samples).

4.3 Experiments

4.3.1 Localization with models trained on full datasets

Setup. This experiment evaluates the baseline localization performance of the proposed NAF-based analysis-by-synthesis approach using the full training datasets for both SoundSpaces and RAF. We compare grid-based and PSO search strategies against direct regression baselines (ResNet-10, and NAF-Direct) to assess accuracy in high-data regimes. Baselines are trained on flattened STFT spectra plus emitter positions, outputting listener positions via regression with MSE loss.

Results. The mean error of our method (both grid-based and PSO-based) and the two baselines is presented in Figure 4.2. Clearly, both baselines outperform our NAF-based approach in this high-data regime. However, subsequent experiments in Section 4.3.2 demonstrate that the NAF method achieves lower errors when trained on reduced dataset fractions of RAF, indicating its advantage in real-world sparse data conditions. In contrast, its performance on the SoundSpaces dataset is noticeably worse, likely due to differences between simulated and real acoustic environments, which are discussed further in Section 4.3.2.

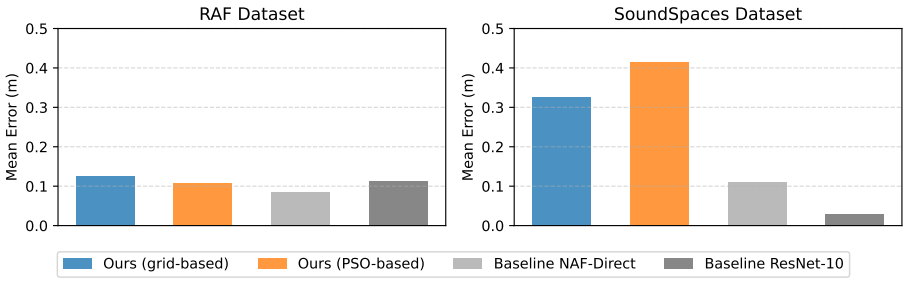

Figure 4.2: **Our method underperforms the baselines when 100% of the training data is available.**

4.3.2 Localization with models trained on sparse data

Setup. Low-data regimes simulate practical constraints, like rapid deployment in new rooms with limited measurements. This experiment tests robustness by training the NAF on randomly sampled subsets (5%, 10%, 25%, 50%, 100%) of the training data for both RAF and SoundSpaces.

Results. Figure 4.3 shows localization error versus training data percentage for the proposed grid-based and PSO-based strategies and the two baselines. In the 5-25% range, our grid-based approach consistently outperforms both baselines for the RAF dataset. Specifically, at 10% data, it achieves 35.1 cm error, and at 25% data 18.6 cm, corresponding to 22% and 32% lower error than NAF-Direct, the strongest baseline in this range. For the SoundSpaces dataset, however, both baselines achieve significantly lower errors across all data percentages. One possible explanation is that the simulated environments have limited acoustic variability and a smooth relationship between position and RIR, allowing the regression models to interpolate well to unseen locations. In contrast, the analysis-by-synthesis approach relies on distinct acoustic cues to identify a unique position, which may be less pronounced in these regular, simulated settings.

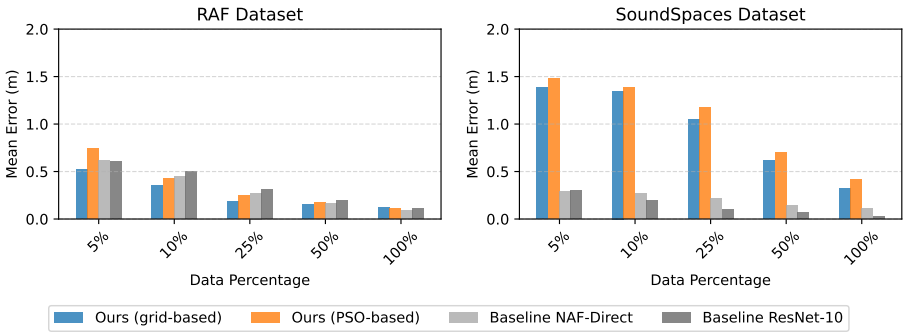

Figure 4.3: **Mean localization error vs. training data percentage used for our method and two baselines. For RAF, in sparse-data scenarios, our grid-based approach outperforms the baselines.**

Acoustic ambiguity analysis. To test this hypothesis, a nearest-neighbor ambiguity analysis was performed. For 100 RIR spectrograms, the five acoustically most similar training samples were found (based on L2 distance), and the spatial distances between their listener positions were measured while keeping the emitter location and listener orientation fixed. Averaged over all environments, the mean top-5 spatial distance was 2.92 m for SoundSpaces and 0.42 m for RAF. The much higher value for SoundSpaces shows that acoustically similar RIRs often come from distant listener positions, revealing strong acoustic ambiguity in the simulated data. RAF, in contrast, shows a tighter link between acoustic and spatial similarity. Because the analysis-by-synthesis method must locate a unique minimum, such ambiguity leads to flatter loss landscapes and less precise localization.

4.3.3 PSO parameters

Setup. This experiment analyzes the impact of PSO hyperparameters on localization performance and efficiency using the full RAF dataset. Grid-based search is computationally expensive, and PSO provides a more efficient exploration of the search space. We vary particle count N_p (50, 100, 250), iterations N_{iter} (10, 25, 50), inertia w (0.5, 0.7, 0.9), and the cognitive and social coefficients C_1/C_2 (1.0, 1.5, 2.0). Grid-based search at various resolutions serves as a baseline.

Results. Figure 4.4 presents a plot of mean localization error versus mean computation time for all PSO parameter combinations, as listed in the previous paragraph, on RAF and SoundSpaces. The plot shows vertical clusters of points, as only N_p and N_{iter} impact computation time, while w, C_1, and C_2 have negligible influence. Hence, configurations with equal N_p and N_{iter} form vertical lines.

For RAF, the 25 cm grid-based method takes 5.46 seconds with a 12.5 cm error, whereas the PSO configuration with ($N_p = 250$, $N_{\text{iter}} = 10$, $C_1 = 1.5$, $C_2 = 1$, $w = 0.7$) achieves a 10.7 cm error in 1.35 seconds, a 75% runtime reduction and a 14% error reduction. There appears to be a clear trade-off between accuracy and computation time.

In fact, the Pareto front of PSO configurations clearly dominates that of the grid-based approach on RAF, yielding solutions with both lower runtime and reduced mean error. For SoundSpaces, this dominance is less clear, suggesting that the relative benefit of PSO is greater in 3D settings. A plausible explanation is that the computational load of grid-based search increases exponentially with dimensionality, rendering it particularly inefficient in 3D spaces, while PSO can exploit this larger search space more effectively, leaving more potential for optimization in the 3D case.

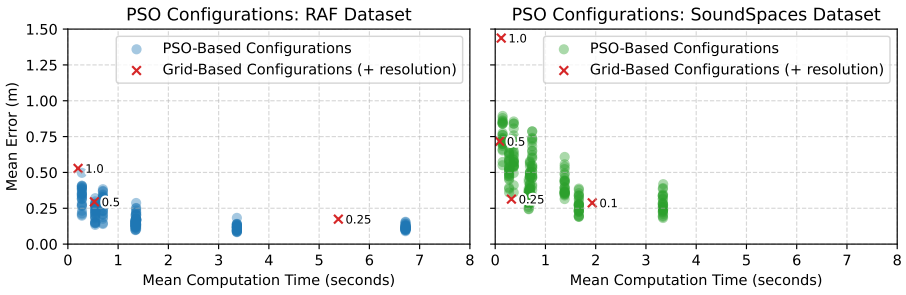

Figure 4.4: **Mean localization error vs. mean computation time for PSO hyperparameter combinations on RAF (blue) and SoundSpaces (green) datasets. Vertical clusters of points arise because only N_p and N_{iter} affect computation time, while w, C_1, and C_2 influence the mean error without affecting runtime. The red crosses show the performance of the grid-based approach at various resolutions. PSO is clearly superior in the case of RAF.**

Individual parameter contributions. Figures 4.5, 4.6, 4.9, 4.7, and 4.8 show individual parameter sweeps for N_p (50-350), N_{iter} (5-35), w (0.1-0.9), C_1 (0.5-2.5), and C_2 (0.5-2.5). In each sweep, all other parameters are fixed to values that showed strong performance across both datasets: $N_p = 100$, $N_{\text{iter}} = 25$, $w = 0.7$, $C_1 = 1.5$, $C_2 = 1.5$. Each figure includes side-by-side plots for RAF and SoundSpaces. Accuracy improves with larger N_p or N_{iter}, though at the cost of longer runtimes. In contrast, adjusting w, C_1, and C_2 shows no clear effect across datasets.

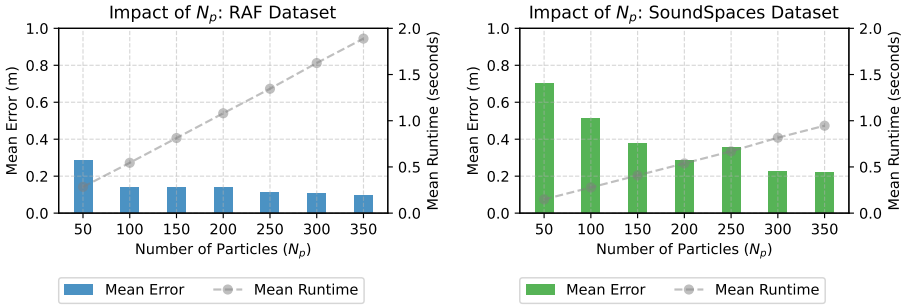

Figure 4.5: **Mean localization error and runtime vs. number of particles N_p for RAF (left) and SoundSpaces (right). Error decreases with increasing N_p, but the improvement diminishes beyond 200 particles.**

Figure 4.5 shows that increasing N_p reduces error but with diminishing returns after 200. There is a clear trade-off between accuracy and runtime. Figure 4.6 similarly shows improvements up to about 25 iterations.

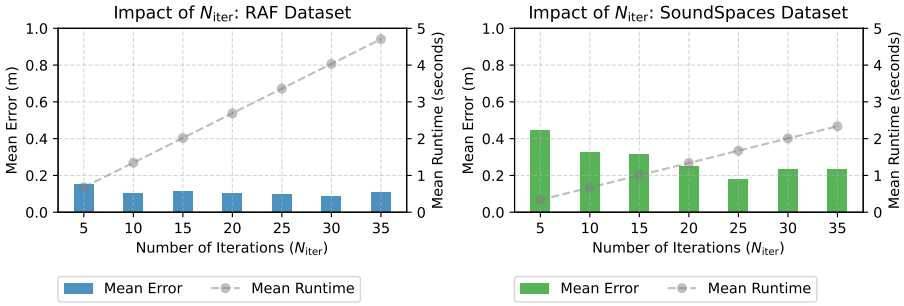

Figure 4.6: **Mean localization error and runtime vs. N_{iter} for RAF (left) and SoundSpaces (right). Error decreases with increasing N_{iter}, but the improvement diminishes beyond 25 iterations.**

Beyond these scaling parameters, C_1 and C_2 control individual and social exploration. As Figure 4.7 and 4.8 show, C_1 and C_2 have no clear relationship to localization accuracy.

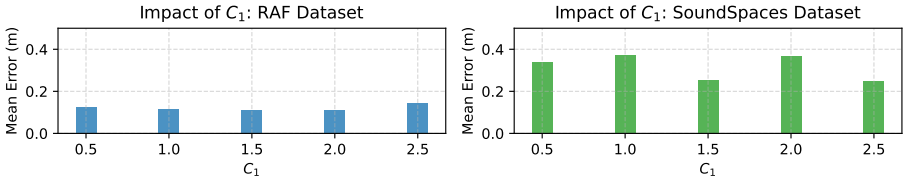

Figure 4.7: **Mean localization error vs. C_1 for RAF (left) and SoundSpaces (right). No clear pattern can be observed across the datasets.**

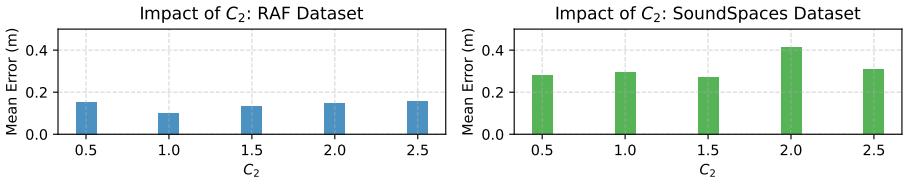

Figure 4.8: **Mean localization error vs. C_2 for RAF (left) and SoundSpaces (right). No clear pattern can be observed across the datasets.**

Finally, the inertia weight w in Figure 4.9 also shows little impact across datasets. Together, these results indicate that N_p and N_{iter} are the key PSO parameters, while w, C_1, and C_2 have minor effects.

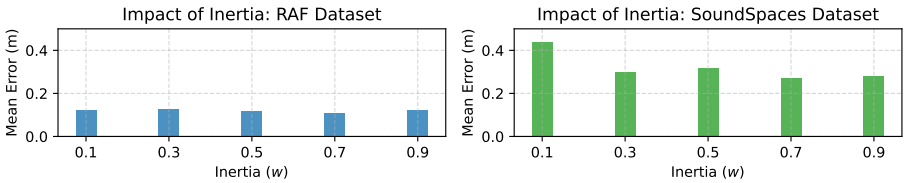

Figure 4.9: **Mean localization error vs. inertia (w) for RAF (left) and SoundSpaces (right). No clear pattern can be observed across the datasets.**

4.3.4 Pixel selection

Setup. As discussed in Section 3.2.4, sampling a subset of STFT pixels to generate and compare spectra accelerates localization. Here, we evaluate this approach on the RAF and SoundSpaces datasets using grid search, comparing Uniform Sampling (US) and Standard Deviation-Weighted Sampling (SDWS) across 1-100% of spectrum pixels.

Results. Figure 4.10 illustrates the trade-off between STFT pixel percentage (proportional to computation time) and localization error for both sampling strategies. SDWS consistently outperforms Uniform Sampling, particularly at low sampling rates, by prioritizing informative high-variance pixels. Using 2.5% of pixels with SDWS (chosen as the default for the remaining experiments) reduces computation time by roughly 40× while increasing error only moderately, from 10.75 cm to 13.16 cm (+22%). This optimization is key to mitigating our method's significant runtime compared to the baselines.

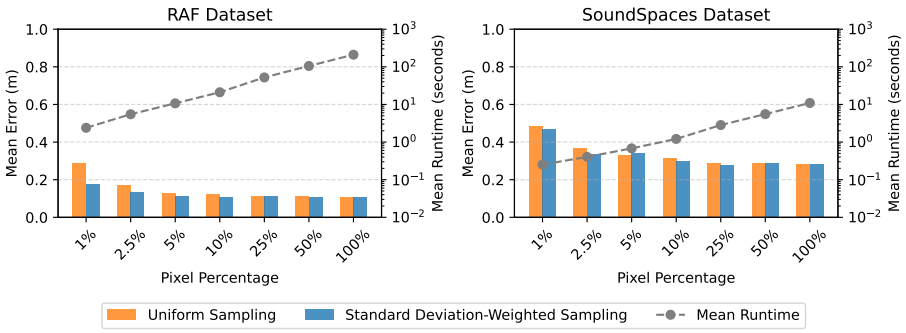

Figure 4.10: **Mean localization error vs. STFT pixel sampling percentage. SDWS clearly outperforms US as the percentage of compared STFT pixels decreases.**

4.3.5 Regularization impact

Setup. This experiment investigates the impact of the regularization parameter ε_{reg} on localization, for both the RAF and SoundSpaces datasets. As discussed in Section 3.2.2, ε_{reg} influences the amount of noise added to the emitter and listener locations during training. We train NAFs using ε_{reg} values (0.01, 0.05, 0.1) and apply grid-based localization to observe the impact of ε_{reg} on localization accuracy.

Results. Figure 4.11 shows the mean localization error for the varying ε_{reg} values over a range of grid resolutions. There is an interesting pattern to observe here. As we decrease the grid resolution, a lower ε_{reg} becomes beneficial, while at higher resolutions, a higher ε_{reg} is better. An explanation for this effect can be found by analyzing the loss maps in Figure 4.12: lower ε_{reg} values produce sharper, more detailed minima (right figure), which improve accuracy on fine grids but can be missed on coarse grids. In contrast, higher ε_{reg} smooths the loss landscape (left panel), making the minima easier to capture with coarse grids, but slightly reducing accuracy on fine grids. This explains why the optimal regularization strength depends on the grid resolution.

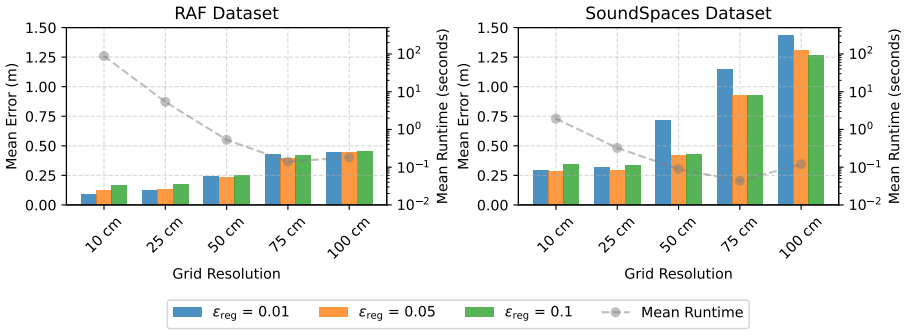

Figure 4.11: **Mean localization error vs. grid resolution for varying ε_{reg} values. Higher ε_{reg} performs better at coarser resolutions (larger step sizes), while lower ε_{reg} is better at finer resolutions.**

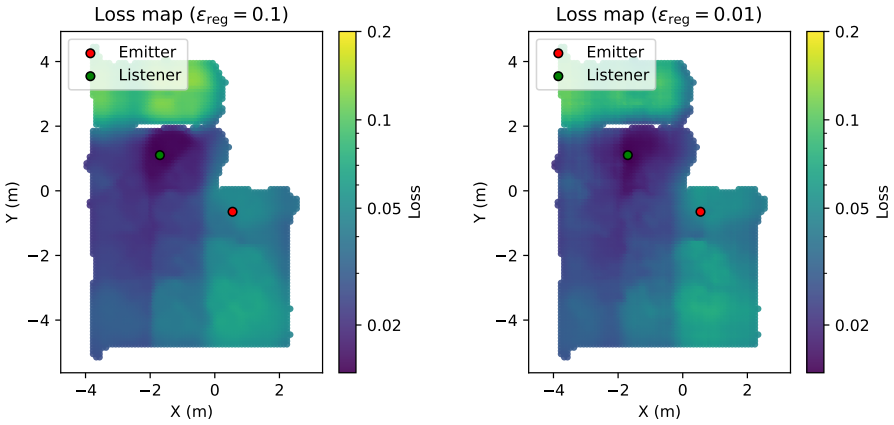

Figure 4.12: **Example loss landscapes from the RAF dataset with $\varepsilon_{\mathbf{reg}} = 0.1$ (left) and $\varepsilon_{\mathbf{reg}} = 0.01$ (right). The right landscape appears slightly more detailed, with the minimum-loss region more strongly concentrated around the ground-truth location, improving localization accuracy with a sufficiently fine search grid.**

4.4 System Profiling

In this section, the computational requirements of the evaluated method and the baselines are covered.

4.4.1 Model Size

Table 4.2 compares our proposed method to the ResNet-10 and NAF-Direct baselines, as introduced in Section 4.2, in terms of parameter count and floating-point operations (FLOPs). For our approach, NAF-Loc, the computational cost depends on the number of STFT pixels synthesized. We therefore provide FLOPs for both the full spectrum (100%) and the reduced setting (2.5%), the latter being the setting used in our experiments. All experiments were performed on Snellius NVIDIA A100 GPU nodes, using 18 of the 72 available CPU cores (Intel Xeon Platinum 8360Y) [56].

Model	**# Params**	**FLOPs**	**Type**
NAF-Loc (100% pixels)	3.6M	148.6G	Analysis-by-synthesis
NAF-Loc (2.5% pixels)		3.72G	Analysis-by-synthesis
ResNet-10 (baseline)	4.9M	112.9G	Direct regression
NAF-Direct (baseline)	2.4M	4.8G	Direct regression

Table 4.2: **Comparison of model complexities for the proposed NAF and direct regression baselines.**

4.4.2 Training Time

Table 4.3 reports the wall-clock training times required to optimize NAF-Loc across the evaluation datasets. Each model was trained for 200 epochs, consistent with the experimental setup used in the original NAF paper [34], and under the same hardware conditions described in Section 4.4.1. For each batch, we sample 20 impulse responses and, for each spectrogram, randomly select 2,000 time-frequency pairs. The observed training times vary considerably between environments, as the room sizes and spatial sample density vary.

Room	**Training Time (hh:mm)**
SoundSpaces [9]	
`frl_apartment_2`	11:07
`frl_apartment_4`	12:02
`room_2`	1:14
`office_4`	2:53
`apartment_1`	12:43
`apartment_2`	12:13
RAF [12]	
`EmptyRoom`	4:40
`FurnishedRoom`	3:51

Table 4.3: **Training times per room in the evaluation datasets. Each network was trained for 200 epochs.**

4.4.3 Inference Time

Table 4.4 reports the time required for a single forward inference by each model. All measurements are obtained on the RAF dataset. For NAF-Loc, we again distinguish between generating all STFT pixels (100%) and generating only 2.5% of the pixels. For the baseline models, a forward inference corresponds to predicting a location directly from a given STFT input. The direct regression baselines are substantially faster, as their inference time and localization time are identical. In contrast, NAF-Loc follows an analysis-by-synthesis strategy in which a single localization requires many STFT inferences, making the total localization time considerably longer. Clearly, NAF-Loc offers a trade-off: higher computational cost in exchange for improved accuracy.

Model	**Inference Time**	**Localization Time**
NAF-Loc (100% of pixels)	19.24 ms	218.4 s
NAF-Loc (2.5% of pixels)	0.481 ms	5.46 s
ResNet-10 (baseline)	0.013 ms	0.058 ms
NAF-Direct (baseline)	0.025 ms	0.058 ms

Table 4.4: **Inference time and total localization time for the proposed NAF-Loc system and direct regression baselines on the RAF dataset.**

Chapter 5

Conclusions and Discussion

This chapter summarizes the key findings of this thesis and reflects on their implications. We first present the main conclusions drawn from the experimental results (Section 5.1). Next, we discuss the limitations of the proposed approach and interpret the broader meaning of the findings (Section 5.2). Finally, we outline promising directions for future research (Section 5.3).

5.1 Conclusions

This thesis has demonstrated that Neural Acoustic Fields can be effectively inverted through an analysis-by-synthesis approach to perform acoustic listener localization. The core finding is that a pre-trained NAF, originally designed for acoustic synthesis, can successfully solve the inverse problem of geometry estimation from a single RIR measurement, and outperform direct regression baselines in sparse-data scenarios.

The proposed method works by synthesizing RIRs for candidate listener positions and identifying the location that minimizes spectral loss when compared to an observed RIR. This approach proved particularly valuable in data-sparse conditions, where it consistently outperformed direct regression baselines. When trained on only 5-25% of the available data, our grid-based search method achieved up to 32% lower error than direct regression approaches, demonstrating its advantage in practical scenarios where extensive data collection may be infeasible.

However, when evaluated on the simulated SoundSpaces dataset, both baselines achieved lower errors across all data percentages. We hypothesize that this difference arises from the regular and low-variability nature of the simulated SoundSpaces environments, where the mapping between position and RIR is smooth and predictable. Under these conditions, direct regression models can interpolate effectively, while the analysis-by-synthesis search is more sensitive to acoustic ambiguities between locations that sound alike.

Analysis-by-synthesis is computationally intensive, as it requires generating and comparing many examples, unlike direct regression, which requires only a

single inference. To address this cost, two key optimizations were developed and evaluated. A Standard Deviation-Weighted Sampling technique reduced the computational cost by focusing on the most informative time-frequency bins, achieving a 40× reduction in required computations with only a minimal impact on the mean error. Additionally, Particle Swarm Optimization proved more efficient than grid search for navigating the 3D loss landscape, with one configuration reducing compute time by 75% while simultaneously reducing the mean error by 14% on the RAF dataset. The clear superiority of PSO over grid search on the 3D RAF dataset stems from the algorithm's ability to efficiently escape local minima in the high-dimensional loss landscape. This demonstrates that metaheuristics such as PSO are not just a faster alternative but a fundamentally more effective search strategy than grid-based search or gradient descent.

The experiments also showed that the combination of model regularization and search resolution affects performance. Lower regularization values create sharper loss landscapes that work well with fine-grid searches, while higher values create smoother landscapes that are better for coarse searches, giving practical insights for system configuration.

In summary, this work shows that analysis-by-synthesis through NAF inversion is a viable approach for acoustic localization, especially when only limited training data is available or when the acoustic environment varies strongly across space. The weaker performance on the simulated SoundSpaces dataset suggests that the method benefits most from the irregularities and richer cues found in real acoustic data, rather than from the smooth, predictable structure of synthetic environments. Therefore, the choice of dataset and the balance between accuracy and computational cost remain important considerations for applying this method effectively.

5.2 Discussion

The computational cost of the analysis-by-synthesis loop remains the most significant limitation of this approach, currently precluding real-time application. The method also currently estimates only the listener position and requires knowing its orientation in the case of a directional listener (as was the case for the SoundSpaces dataset). This represents an important constraint for practical deployment in applications involving directional hearing. Furthermore, the approach assumes a static environment, meaning performance would likely degrade if room acoustics changed after NAF training.

The results highlight a fundamental trade-off between data efficiency and computational cost when choosing between analysis-by-synthesis and direct regression for acoustic localization. The method provides an interpretable output through spatial loss maps, which offer visual representations of localization uncertainty that could be valuable for applications requiring uncertainty quantification. However, for resource-constrained applications, a direct regression approach may still be preferable.

5.3 Future Work

Future work should focus on enhancing the method's efficiency and scope:

- **Joint Position and Orientation Estimation:** A natural extension is to expand the search space to include the listener's orientation (θ_ℓ), leveraging the binaural cues present in multi-channel datasets such as SoundSpaces. If this proves infeasible, an attempt could be made to make localization independent from prior knowledge of θ_ℓ.
- **Multi-Stage Optimization:** Unlike iNeRF [31], our method does not currently exploit the differentiability of NAFs. Gradient descent alone is unsuitable for global search in the complex loss landscape, but combining PSO with local gradient-based refinement may improve both speed and accuracy. A coarse PSO search with few particles (or grid search) could locate the global minimum-loss region, after which gradient descent refines the solution.
- **Architectural Improvements:** This thesis used the original NAF architecture [34], but newer methods [54, 30] achieve better RIR reconstruction results, which may also enhance localization accuracy and efficiency.
- **Practical Implementation:** A key step toward deployment is moving from a research framework to a standalone system on consumer hardware, such as a smartphone or home robot. This requires an offline calibration phase to train a NAF for the environment, and efficient optimization of the analysis-by-synthesis loop. In tracking scenarios, prior position estimates could further narrow the search region and greatly accelerate localization.

Bibliography

[1] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen. Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks. *IEEE J. Sel. Top. Signal Process.*, 13(1):34–48, 2018.

[2] Jont B. Allen and David A. Berkley. Image Method for Efficiently Simulating Small-Room Acoustics. *J. Acoust. Soc. Am.*, 65(4):943–950, 1979.

[3] Marco Altini, Davide Brunelli, Elisabetta Farella, and Luca Benini. Bluetooth Indoor Localization with Multiple Neural Networks. In *Proc. IEEE Int. Symp. Wireless Pervasive Comput. (ISWPC)*, pages 295–300, Modena, Italy, May 2010.

[4] Thomas G. Bever and David Poeppel. Analysis by Synthesis: A (Re-) Emerging Program of Research for Language and Vision. *Biolinguistics*, 4(2–3):174–200, 2010.

[5] Joydeep Biswas and Manuela Veloso. WiFi Localization and Navigation for Autonomous Indoor Mobile Robots. In *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pages 4379–4384, Anchorage, AK, USA, May 2010.

[6] Sujittra Boonsriwai and Anya Apavatjrut. Indoor WiFi Localization on Mobile Devices. In *Proc. Int. Conf. Electr. Eng./Electronics, Comput. Telecommunications and Inf. Technol. (ECTI-CON)*, pages 1–5, Krabi, Thailand, May 2013.

[7] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6D Object Pose Estimation Using 3D Object Coordinates. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 536–551, Zurich, Switzerland, 2014.

[8] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D Data in Indoor Environments. In *Proc. Int. Conf. 3D Vision (3DV)*, pages 667–676, Qingdao, China, 2017.

[9] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. SoundSpaces: Audio-Visual Navigation in 3D Environments. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020.

[10] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W. Robinson, and Kristen Grauman. SoundSpaces 2.0: A Simulation Platform for Visual-Acoustic Learning. In *Proc. Adv. Neural Inf. Process. Syst. Datasets and Benchmarks*, 2022.

[11] Xu Chen, Zijian Dong, Jie Song, Andreas Geiger, and Otmar Hilliges. Category Level Object Pose Estimation via Neural Analysis-by-Synthesis. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 139–156, Glasgow, United Kingdom, 2020.

[12] Ziyang Chen, Israel D. Gebru, Christian Richardt, Anurag Kumar, William Laney, Andrew Owens, and Alexander Richard. Real Acoustic Fields: An Audio-Visual Room Acoustics Dataset and Benchmark. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 21886–21896, Seattle, WA, USA, 2024.

[13] Brian McFee et al. librosa/librosa: 0.10.2.post1, May 2024.

[14] Angelo Farina. Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique. *Preprints Audio Eng. Soc.*, 2000.

[15] François Grondin, Jean-Samuel Lauzon, Simon Michaud, Mirco Ravanelli, and François Michaud. BIRD: Big Impulse Response Dataset. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2020.

[16] Pierre-Amaury Grumiaux, Srđan Kitić, Laurent Girin, and Alexandre Guérin. A Survey of Sound Source Localization with Deep Learning Methods. *J. Acoust. Soc. Am.*, 152(1):107–151, 2022.

[17] Yu Gu and Fuji Ren. Energy-Efficient Indoor Localization of Smart Hand-Held Devices Using Bluetooth. *IEEE Access*, 3:1450–1461, 2015.

[18] M. Halle and K. Stevens. Speech Recognition: A Model and a Program for Research. *IRE Trans. Inf. Theory*, 8(2):155–159, 1962.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 770–778, Las Vegas, NV, USA, 2016.

[20] Weipeng He, Petr Motlicek, and Jean-Marc Odobez. Deep Neural Networks for Multiple Speaker Detection and Localization. In *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pages 74–79, Brisbane, QLD, Australia, May 2018.

[21] Sachini Herath, David Caruso, Chen Liu, Yufan Chen, and Yasutaka Furukawa. Neural Inertial Localization. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 6594–6603, New Orleans, LA, USA, June 2022.

[22] Shengyu Huang, Zan Gojcic, Zian Wang, Francis Williams, Yoni Kasten, Sanja Fidler, Konrad Schindler, and Or Litany. Neural LiDAR Fields for Novel View Synthesis. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 18236–18246, Paris, France, 2023.

[23] Tianshu Huang, John Miller, Akarsh Prabhakara, Tao Jin, Tarana Laroia, Zico Kolter, and Anthony Rowe. DART: Implicit Doppler Tomography for Radar Novel View Synthesis. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 24118–24129, Seattle, WA, USA, 2024.

[24] James Kennedy and Russell Eberhart. Particle Swarm Optimization. In *Proc. IEEE Int. Conf. Neural Netw. (ICNN)*, volume 4, pages 1942–1948, Perth, WA, Australia, 1995.

[25] Musa Furkan Keskin, Ahmet Dundar Sezer, and Sinan Gezici. Localization via Visible Light Systems. *Proc. IEEE*, 106(6):1063–1088, 2018.

[26] C. Knapp and G. Carter. The Generalized Correlation Method for Estimation of Time Delay. *IEEE Trans. Acoust. Speech Signal Process.*, 24(4):320–327, 1976.

[27] Manikanta Kotaru, Kiran Joshi, Dinesh Bharadia, and Sachin Katti. SpotFi: Decimeter Level Localization Using WiFi. *SIGCOMM Comput. Commun. Rev.*, 45(4):269–282, October 2015.

[28] Shoichi Koyama, Tomoya Nishida, Keisuke Kimura, Takumi Abe, Natsuki Ueno, and Jesper Brunnström. MESHRIR: A Dataset of Room Impulse Responses on Meshed Grid Points for Evaluating Sound Field Analysis and Synthesis Methods. In *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, pages 1–5, New Paltz, NY, USA, 2021.

[29] Xinya Li, Zhiqun Daniel Deng, Lynn T. Rauchenstein, and Thomas J. Carlson. Source-Localization Algorithms and Applications Using Time of Arrival and Time Difference of Arrival Measurements. *Rev. Sci. Instrum.*, 87(4):041502, April 2016.

[30] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Neural Acoustic Context Field: Rendering Realistic Room Impulse Response with Neural Fields. *arXiv preprint arXiv:2309.15977*, 2023.

[31] Yen-Chen Lin, Pete Florence, Jonathan T. Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. iNeRF: Inverting Neural Radiance Fields for Pose Estimation. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, pages 1323–1330, Prague, Czech Republic, 2021.

[32] Manni Liu, Linsong Cheng, Kun Qian, Jiliang Wang, Jin Wang, and Yunhao Liu. Indoor Acoustic Localization: A Survey. *Hum.-Centric Comput. Inf. Sci.*, 10(1):24:1–24:24, January 2020.

[33] Tong Liu, Cong Han, Qian Lin Wang, Zhen Quan Li, and Guoan Yang. A Method of Acoustic Emission Source Location for Engine Fault Based on Time Difference Matrix. *Struct. Health Monit.*, 22(1):621–638, 2023.

[34] Andrew Luo, Yilun Du, Michael Tarr, Josh Tenenbaum, Antonio Torralba, and Chuang Gan. Learning Neural Acoustic Fields. *Adv. Neural Inf. Process. Syst.*, 35:3165–3177, 2022.

[35] Andrew Luo, Yilun Du, Michael Tarr, Josh Tenenbaum, Antonio Torralba, and Chuang Gan. Learning Neural Acoustic Fields: Official Code. `https://github.com/aluo-x/Learning_Neural_Acoustic_Fields`, 2022. Accessed: 31 Aug. 2025.

[36] Tim Lübeck, Johannes M. Arend, and Christoph Pörschmann. A Real-Time Application for Sound Source Localization Inside a Spherical Microphone Array. In *Proc. 44th DAGA*, pages 319–322, Munich, Germany, March 2018. Deutsche Gesellschaft für Akustik.

[37] Amjad Yousef Majid, Venkatesha Prasad, Mees Jonker, Casper van der Horst, Lucan de Groot, and Sujay Narayana. AI-Based Simultaneous Audio Localization and Communication for Robots. In *Proc. ACM/IEEE Conf. Internet of Things Design and Implementation (IoTDI)*, IoTDI '23, pages 172–183, San Antonio, TX, USA, May 2023.

[38] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 7206–7215, Virtual, 2021.

[39] Nelson Max. Optical Models for Direct Volume Rendering. *IEEE Trans. Vis. Comput. Graph.*, 1(2):99–108, 1995.

[40] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *Commun. ACM*, 65(1):99–106, December 2021.

[41] E. A. Nadaraya. On Estimating Regression. *Theory Probab. Appl.*, 9(1):141–142, 1964.

[42] Ulric Neisser. *Cognitive Psychology: Classic Edition*. Psychology Press, 1967.

[43] Alan V. Oppenheim and Ronald W. Schafer. *Discrete-Time Signal Processing*. Prentice Hall, Upper Saddle River, NJ, USA, 3rd edition, 2009.

[44] Reza Parhizkar, Ivan Dokmanić, and Martin Vetterli. Single-Channel Indoor Microphone Localization. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pages 1434–1438, Florence, Italy, May 2014.

[45] Mehmet Pekmezci and Yakup Genc. Evaluation of SSIM Loss Function in RIR Generator GANs. *Digit. Signal Process.*, 154:104685, 2024.

[46] Hadrien Pujol, Eric Bavu, and Alexandre Garcia. BeamLearning: An End-to-End Deep Learning Approach for the Angular Localization of Sound Sources Using Raw Multichannel Acoustic Pressure Data. *J. Acoust. Soc. Am.*, 149(6):4248–4263, 2021.

[47] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. KiloNeRF: Speeding up Neural Radiance Fields with Thousands of Tiny MLPs. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 14315–14325, Montreal, QC, Canada, 2021.

[48] Douglas D. Rife and John Vanderkooy. Transfer-Function Measurement with Maximum-Length Sequences. *J. Audio Eng. Soc.*, 37(6):419–444, 1989.

[49] Janos Sallai, György Balogh, Miklos Maroti, Akos Ledeczi, and Branislav Kusy. Acoustic Ranging in Resource-Constrained Sensor Networks. In *Proc. Int. Conf. Wireless Netw. (ICWN)*, pages 467–470, Las Vegas, NV, USA, June 2004.

[50] Luiz Schirmer, Guilherme Schardong, Vinícius da Silva, Hélio Lopes, Tiago Novello, Daniel Yukimura, Thales Magalhaes, Hallison Paz, and Luiz Velho. Neural Networks for Implicit Representations of 3D Scenes. In *Proc. SIBGRAPI*, pages 17–24, Gramado, Brazil, October 2021.

[51] Ralph Schmidt. Multiple Emitter Location and Signal Parameter Estimation. *IEEE Trans. Antennas Propag.*, 34(3):276–280, 1986.

[52] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit Neural Representations with Periodic Activation Functions. In *Adv. Neural Inf. Process. Syst.*, volume 33, pages 7462–7473. Curran Associates, 2020.

[53] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob Engel, Raúl Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, and Richard Newcombe. The Replica Dataset: A Digital Replica of Indoor Spaces. *arXiv preprint arXiv:1906.05797*, 2019.

[54] Kun Su, Mingfei Chen, and Eli Shlizerman. Inras: Implicit Neural Representation for Audio Scenes. *Adv. Neural Inf. Process. Syst.*, 35:8144–8158, 2022.

[55] Yimao Sun, K. C. Ho, and Qun Wan. Solution and Analysis of TDOA Localization of a Near or Distant Source in Closed Form. *IEEE Trans. Signal Process.*, 67(2):320–335, 2019.

[56] SURF User Knowledge Base. Snellius Hardware. `https://servicedesk.surf.nl/wiki/spaces/WIKI/pages/30660208/Snellius+hardware`, 2025. Accessed: 31 Aug. 2025.

[57] Igor Szöke, Miroslav Skácel, Ladislav Mošner, Jakub Paliesek, and Jan Černocký. Building and Evaluation of a Real Room Impulse Response Dataset. *IEEE J. Sel. Top. Signal Process.*, 13(4):863–876, 2019.

[58] Zhenyu Tang, Rohith Aralikatti, Anton Jeran Ratnarajah, and Dinesh Manocha. GWA: A Large High-Quality Acoustic Dataset for Audio Processing. In *Proc. ACM SIGGRAPH Conf. Proc.*, pages 36:1–36:9, Vancouver, BC, Canada, 2022.

[59] Paolo Vecchiotti, Ning Ma, Stefano Squartini, and Guy J. Brown. End-to-End Binaural Sound Localisation from the Raw Waveform. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pages 451–455, Brighton, United Kingdom, May 2019.

[60] Fabio Vesperini, Paolo Vecchiotti, Emanuele Principi, Stefano Squartini, and Francesco Piazza. A Neural Network Based Algorithm for Speaker Localization in a Multi-Room Environment. In *Proc. IEEE Int. Workshop Mach. Learn. Signal Process. (MLSP)*, pages 1–6, Vietri sul Mare, Italy, September 2016.

[61] Mason Wang, Ryosuke Sawata, Samuel Clarke, Ruohan Gao, Shangzhe Wu, and Jiajun Wu. Hearing Anything Anywhere. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

[62] Jiang Xiao, Zimu Zhou, Youwen Yi, and Lionel M. Ni. A Survey on Wireless Indoor Localization from the Device Perspective. *ACM Comput. Surv.*, 49(2):25:1–25:31, June 2016.

[63] Xiong Xiao, Shengkui Zhao, Xionghu Zhong, Douglas L. Jones, Eng Siong Chng, and Haizhou Li. A Learning-Based Approach to Direction of Arrival Estimation in Noisy and Reverberant Environments. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pages 2814–2818, South Brisbane, QLD, Australia, April 2015.

[64] Jiawei Yang, Marco Pavone, and Yue Wang. FreeNeRF: Improving Few-Shot Neural Rendering with Free Frequency Regularization. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 8254–8263, Vancouver, BC, Canada, 2023.

[65] Jie Yin, Andrew Luo, Yilun Du, Anoop Cherian, Tim K. Marks, Jonathan Le Roux, and Chuang Gan. Disentangled Acoustic Fields for Multimodal Physical Scene Understanding. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, pages 557–564, Abu Dhabi, UAE, 2024.

[66] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. NeRF++: Analyzing and Improving Neural Radiance Fields. *arXiv preprint arXiv:2010.07492*, 2020.

[67] Xiaopeng Zhao, Zhenlin An, Qingrui Pan, and Lei Yang. NeRF2: Neural Radio-Frequency Radiance Fields. In *Proc. ACM Int. Conf. Mobile Comput. Netw. (MobiCom)*, pages 1–15, Madrid, Spain, 2023.