

**FROM ECHOES TO SHAPES: DOLPHIN-INSPIRED UNDERWATER
TARGET RECONSTRUCTION USING DEEP LEARNING**

by

HOU PENGKUN

(B.Eng., Beijing University of Posts and Communications)

A DISSERTATION SUBMITTED FOR THE DEGREE OF

MASTER OF COMPUTING

in

COMPUTER SCIENCE

in

SCHOOL OF COMPUTING

of the

NATIONAL UNIVERSITY OF SINGAPORE

2026

Advisors:

Associate Professor Martin Henz
Associate Professor Mandar Chitre
Dr. Matthias Hoffmann-Kuhnt

Examiner:

Associate Professor Wang Ye

Declaration

I hereby declare that this dissertation is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the dissertation.

This dissertation has also not been submitted for any degree in any university previously.



HOU Pengkun

1 April 2026

Acknowledgments

Looking back on my short but fulfilling one-year Master's journey, I would like to express my sincere gratitude towards all the kindness and support I received along the way.

Firstly, I would like to thank my advisor, Prof. Martin Henz. Both my FYP and Master's projects would not have been possible without his support throughout this process. I would also like to express my gratitude to Prof. Mandar Chitre and Dr. Matthias Hoffmann-Kuhnt for all the insightful discussions and continuous support during the project's development. They guided me through the early stages of research and helped me discover the joy of it.

I would like to thank all my friends and seniors in ARL, who brightened a big part of my daily life during my Master's study. Special thanks go to Dr. Peng Luyuan, who provided many valuable insights into this project. I would also like to extend my gratitude to Ginsan, whose echolocation clicks made this work possible in the first place.

Last but not least, I would like to thank my boyfriend Longan, my dear friends, and my family for their unconditional support and belief in me. Without them, I would not be who I am today.

Contents

Acknowledgments	i
Abstract	v
List of Figures	vi
List of Tables	viii
1 Introduction	1
1.1 Background and Motivation	1
1.2 Problem Statement	2
1.3 Contributions	4
1.4 Dissertation Synopsis	5
2 Literature Review	7
2.1 Dolphin Echolocation and Shape Perception	7
2.2 From Echoes to Target Geometry	9
2.3 Deep Learning for Acoustic Inverse Problems	10
3 Methodology	12
3.1 Problem Formulation	12
3.1.1 Forward Problem	12
3.1.2 Inverse Problem	13
3.1.3 Reconstruction Strategy	13
3.2 Data Preparation	14
3.2.1 Synthetic Data Generation	14
3.2.2 Real Data Processing	17
3.3 Network Architecture	21

3.3.1	Per-Sensor Echo Encoding	21
3.3.2	Cross-Sensor Compression	22
3.3.3	Multi-View Aggregation	22
3.3.4	Occupancy Prediction Head	22
3.4	Training	22
3.4.1	Loss Function	22
3.4.2	Training Procedure	24
4	Experiments and Results	25
4.1	Evaluation Protocol	25
4.1.1	Pixel-wise Metrics	25
4.1.2	Multiple-Choice Matching Protocol	26
4.2	Synthetic Data Experiment	27
4.2.1	Simulator Configuration	27
4.2.2	Target Patterns	28
4.2.3	Training Configuration	28
4.2.4	Results	29
4.3	Real Data Experiment	31
4.3.1	EV-MTS Experiment and Recording Setup	32
4.3.2	Data Selection and Processing	35
4.3.3	Ground-Truth Occupancy Maps	36
4.3.4	Training Configuration	37
4.3.5	Results	39
5	Discussion	41
5.1	Shape Information in Echo Waveforms	41
5.2	Real Data: Per-Pixel Discrimination and Implications	43
5.3	Comparison with Prior Work	45
5.3.1	Comparison with Prior Approaches	45
5.3.2	Comparison with Dolphin Behavioral Performance	47
5.4	Limitations	48
5.5	Future Directions	50
6	Conclusion	52

Abstract

From Echoes to Shapes: Dolphin-Inspired Underwater Target Reconstruction using
Deep Learning

by

HOU Pengkun

Master of Computing in Computer Science

National University of Singapore

Dolphins can identify the shape of underwater objects using echolocation alone. This implies that echoes carry sufficient geometric information to support spatial shape inference. Inspired by this finding, this dissertation investigates whether a deep learning model can reconstruct the two-dimensional shape of an underwater target directly from echo waveforms.

We develop an end-to-end pipeline comprising three components: a synthetic data generator that simulates echo waveforms from a given 2D occupancy map; a CNN+MLP network that reconstructs the target shape from multi-channel waveforms; and a multiple-choice evaluation protocol, modeled after the dolphin behavioral paradigm, in which the model selects the correct target from a set of candidates based on its reconstruction.

The system is evaluated in two settings. On synthetic data with unseen target shapes, the model achieves 93% accuracy on 2-choice and 81% on 4-choice matching trials, demonstrating that the network learns a generalizable inverse mapping from echoes to spatial structure. On real data from dolphin echolocation recordings, the model attains 84% accuracy on a 2-choice task with two target types, providing evidence that the echo waveforms carry spatially localized target information that survives the complex underwater conditions. Together, these results establish that learned shape reconstruction from echo waveforms alone is feasible and that the required acoustic information is preserved under realistic experimental conditions, providing initial insights toward deep learning based biomimetic sonar systems capable of shape-level perception.

List of Figures

1.1	Problem setup. (a) An acoustic source emits a broadband click that propagates through the hydrophone array, reflects off occupied cells in the target grid, and returns to the receivers. The pink lines illustrate the two-leg propagation path (source–target–hydrophone) for a single scatterer; each hydrophone records a superposition of such reflections from all occupied cells. (b) The inverse problem: given multi-channel echo waveforms recorded at hydrophones $H_1, \dots, H_n$, infer the unknown target occupancy map.	3
3.1	Mnist digit with different thresholds.	16
3.2	Model Architecture.	21
4.1	Synthetic data geometry layout.	27
4.2	Reconstruction results on two held-out synthetic test samples. In each panel, the left subplot shows the ground-truth binary occupancy map and the right subplot shows the model’s predicted probability map. . .	30
4.3	Representative correct trials from the 2-choice (top) and 4-choice (bottom) multiple-choice matching experiments on synthetic data. In each panel, the top row shows the query target and the model’s reconstructed probability map. The bottom row shows the candidate occupancy maps, each annotated with its Dice similarity to the reconstruction. The correct candidate is marked with a star.	32
4.4	The EV-MTS experimental setup, reproduced from [34]. (A) Layout of the sample box (right, underwater) and display boxes (left, in air). (B) The dolphin echolocating on the concealed object underwater. (C) The dolphin visually selecting the matching alternative in air.	33

4.5	Layout of the 32-channel hydrophone array used in the real data recordings.	34
4.6	Target SQ (left) and FF (right)	34
4.7	10 echo clusters in TDOA space, t-SNE applied.	35
4.8	Ground-truth occupancy maps for the SQ (left) and FF (right) targets. The red shade shows the physical extent of each PVC pipe target. . . .	36
4.9	Reconstruction results on two held-out real test samples.	38
4.10	Representative correct trials from the 2-choice multiple-choice matching experiments on real data.	39
5.1	A correct 2-choice matching trial illustrating the effect of reconstruction blurriness on discriminative power. Although the two candidates differ visibly in their fine structure, the model’s diffuse probability map yields nearly identical Dice scores (0.757 vs. 0.758), resulting in a correct selection by a margin of only 0.001. This example shows that even in successful trials, the reconstruction may lack sufficient sharpness to reliably distinguish a candidates.	42
5.2	Sparsity-aware (SA) processor visualizations of dolphin echolocation data for the SQ and FF targets, reproduced from [33]. (a, b) Target objects. (c, d) SA processor output. (e, f) SA output with shading masks overlaid to highlight matching features.	46

List of Tables

4.1	Simulator configuration parameters.	28
4.2	Training hyperparameters (final configuration after grid search).	29

Chapter 1

Introduction

1.1 Background and Motivation

Underwater remote sensing is a fundamental task in applications such as marine exploration, environmental monitoring, and autonomous navigation. Compared with terrestrial environments, underwater environments are more complex, posing challenges to conventional sensing modalities. Specifically, electromagnetic waves attenuate rapidly in water, while optical imaging is severely constrained by light absorption and turbidity. Acoustic sensing, particularly sonar, has therefore become the primary modality for underwater perception [30, 31, 22, 35]. By transmitting acoustic pulses and analyzing the returning echoes, active sonar systems can infer spatial information about submerged objects. However, the spatial resolution of such systems is fundamentally constrained by the physical aperture, that is, the effective spatial extent of the sensor array, and the operating frequency [14]. In practice, imaging quality is further degraded by reverberation, multipath propagation, and low signal-to-noise ratios underwater [22]. These constraints make fine-grained, shape-level reconstruction a persistent challenge for engineered sonar systems of compact size.

A remarkable example of active sonar exists in nature. The biological sonar of dolphins achieves remarkably superior perceptual acuity under comparable frequency and aperture constraints, consistently outperforming engineered sonars of similar size [1, 28]. Decades of behavioral research have demonstrated that dolphins can discriminate object material composition, detect sub-millimeter differences in wall

thickness, and resolve angular separations below one degree in highly dynamic underwater environments [3, 2, 16]. These capabilities far exceed what current engineered systems of equivalent scale can achieve.

Dolphins accomplish this through echolocation, a process governed by the same transmit-receive principle as active sonar: they emit broadband clicks through the melon, a lipid-filled structure in the forehead that focuses the outgoing acoustic energy into a directional beam, and receive the returning echoes via the acoustic fat channels in the lower jaw [1]. Among the most remarkable abilities enabled by echolocation is shape recognition. In the echoic-to-visual matching-to-sample (EV-MTS) paradigm, a dolphin first inspects a sample object solely through echolocation, and must then identify the matching object from a set of visual alternatives presented in air. Early work showed that dolphins could rapidly and accurately match acoustically perceived objects to visually seen targets, performing at near-errorless levels [27, 15, 26]. This cross-modal equivalence implies that the acoustic echo field carries sufficient geometric information to support detailed spatial inference, which is a capability we seek to replicate in engineered systems.

Inspired by these biological findings, prior work has demonstrated that array-processing techniques, when applied to dolphin-like click signals, can extract coarse shape features from the echo field [33]. However, such method produces spatial reflectivity maps that indicate where scattering energy is concentrated, rather than explicit predictions of target shape. Given the demonstrated success of deep learning in solving inverse problems across wave-based imaging, acoustic sensing, and inverse scattering [25, 23, 20, 24], it is timely to revisit this acoustic-to-spatial reconstruction problem with data-driven methods. This naturally leads to the central research question of the present work: can a learned model reconstruct target shape directly from multi-channel echo waveforms, achieving the kind of acoustic-to-spatial transformation that dolphin neural processing evidently performs?

1.2 Problem Statement

To address the research question, we can formulate the problem as an acoustic inverse problem. The physical setup consists of a stationary planar hydrophone array, a fixed target at a known distance in front of the array, and a mobile acoustic source

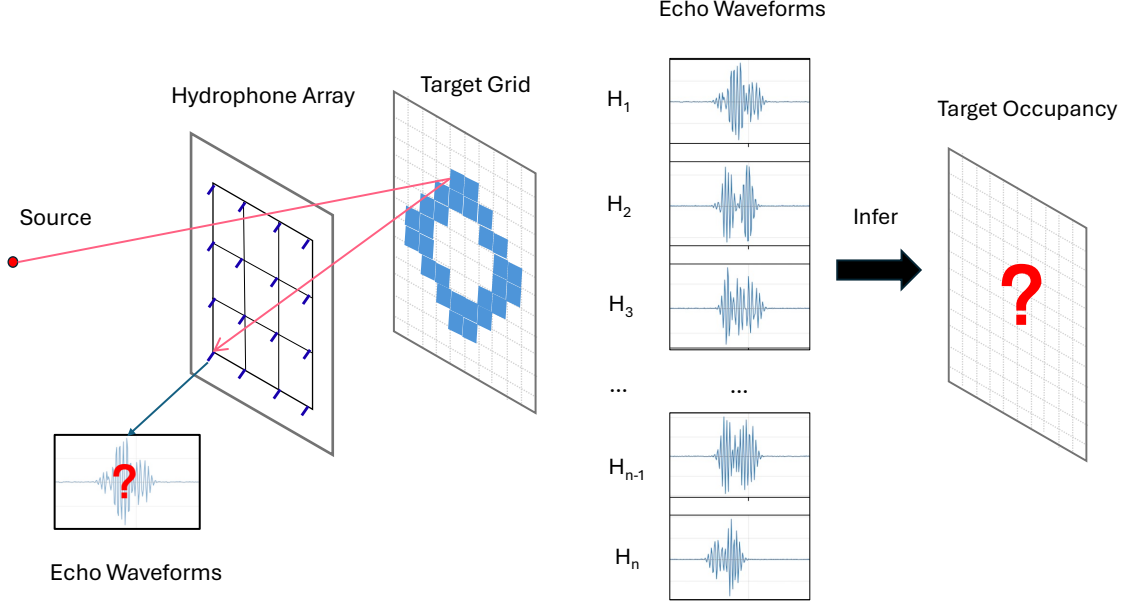


Figure 1.1: Problem setup. (a) An acoustic source emits a broadband click that propagates through the hydrophone array, reflects off occupied cells in the target grid, and returns to the receivers. The pink lines illustrate the two-leg propagation path (source–target–hydrophone) for a single scatterer; each hydrophone records a superposition of such reflections from all occupied cells. (b) The inverse problem: given multi-channel echo waveforms recorded at hydrophones $H_1, \dots, H_n$, infer the unknown target occupancy map.

that emits broadband clicks from varying positions behind the array (Fig. 1.1a). In the forward direction, the physics is simple: given a source position, a target occupancy map, and the propagation environment, one can compute the echo waveforms arriving at each hydrophone through standard acoustic modeling. The inverse direction asks us to recover the target shape from the recorded echoes alone (Fig. 1.1b). This is an ill-posed inverse problem: the received echo at each hydrophone is a superposition of reflections from all scattering elements of the target, and recovering individual contributions from this summed signal constitutes a deconvolution problem whose stability degrades with increasing target complexity.

We start by considering a two-dimensional formulation, where the target is reconstructed as a pixel-wise binary occupancy map on a 2D plane. Extending this formulation to 3D volumetric reconstruction is conceptually straightforward, as depth information can be directly inferred from travel time measurements. In contrast, recovering the lateral spatial structure of a target from echo waveforms

requires extracting fine-scale geometric information that is implicitly encoded and cannot be obtained from simple timing cues. We therefore begin with the 2D shape reconstruction, fixing depth and focusing on lateral structure.

Conventional array processing techniques such as beamforming are sensitive to environmental variability in underwater settings and can be computationally expensive, particularly when high spatial resolution is required. Deep learning provides an alternative strategy. Rather than relying on an explicit forward model, a neural network can learn the underlying relationship between echo waveforms and spatial occupancy from training data. Once trained, the network applies this learned mapping to reconstruct targets it has never encountered, without requiring any prior association between specific echo patterns and specific shapes.

We evaluate this approach in two settings. First, a synthetic dataset generated with a physics-based acoustic simulator provides diverse target shapes, enabling evaluation of true reconstruction on targets never seen during training. Second, real echo data collected from dolphin echolocation experiments test whether the approach transfers to physical acoustic conditions, where the model must generalize across different echo realizations of a limited set of known target geometries.

1.3 Contributions

The main contributions of this work are as follows.

First, we formulate underwater acoustic target shape reconstruction as a learned inverse problem, in which a neural network maps raw multi-channel echo waveforms directly to a pixel-wise 2D binary occupancy map of the target. This data-driven formulation offers an alternative to conventional array processing approaches such as beamforming, which require explicit forward models and are sensitive to environmental variability. By learning the inverse mapping end-to-end from training data, the framework sidesteps these constraints and can, in principle, recover spatial structure that is only implicitly encoded in the echo signals.

Second, we develop two complementary data pipelines that together make the learned approach feasible. A physics-based synthetic data generation pipeline pairs an acoustic propagation simulator with a large pool of diverse target shapes, producing unlimited echo-shape training pairs with known ground truth and enabling

evaluation of true reconstruction on targets never seen during training. A real data processing pipeline extracts a structured echo database from dolphin echolocation recordings collected during EV-MTS experiments, incorporating click detection, channel quality assessment, and source position clustering across a multi-channel hydrophone array.

Third, we adopt a multiple-choice evaluation protocol inspired by EV-MTS paradigm used in dolphin sensory research. This protocol evaluates whether the model’s reconstruction carries enough shape information to select the correct target from a set of alternatives, directly mirroring the task that dolphins perform in cross-modal matching experiments. The shared evaluation framework enables a meaningful comparison between the computational model and the biological system it seeks to emulate.

Finally, through experiments on both synthetic and real data, we provide empirical evidence that multi-channel echo waveforms encode sufficient geometric information to support shape-level spatial inference. On synthetic data, the model reconstructs previously unseen target shapes well above chance in both 2-choice and 4-choice matching protocols, demonstrating that the network learns a generalizable inverse mapping. On real dolphin echolocation recordings, the model reliably discriminates between two physically distinct target shapes while producing spatially structured per-pixel predictions under the complex underwater environment. These findings lend computational support to the hypothesis that echo waveforms carry rich spatial information, and may inform the design of future biomimetic sonar systems inspired by dolphin echolocation.

1.4 Dissertation Synopsis

The rest of this dissertation is organized as follows: Chapter 2 reviews prior work on dolphin echolocation, sonar-based target recognition, and deep learning for acoustic inverse problems, situating the present study in the context of both the biological and computational literature. Chapter 3 formalizes the reconstruction task as an acoustic inverse problem and presents the methodology, covering the synthetic and real data pipelines, the network architecture, and the training procedure. Chapter 4 describes the experimental setup for both synthetic and real data settings,

CHAPTER 1. INTRODUCTION

and reports reconstruction and multiple-choice matching results. Chapter 5 interprets the results, compares the model's performance with dolphin behavioral data, and discusses limitations and future directions. Chapter 6 concludes the dissertation.

Chapter 2

Literature Review

This chapter reviews the three bodies of literature that converge in the present work. Section 2.1 summarizes behavioral evidence that dolphins perceive object shape through echolocation, establishing the biological capability that motivates this dissertation. Section 2.2 frames the recovery of target geometry from echo waveforms as an acoustic inverse problem and reviews the conventional signal processing approaches that have been applied to it. Section 2.3 surveys recent advances in deep learning for wave-based inverse problems, with particular attention to systems that recover spatial structure from acoustic echoes, and identifies the gap that the present work addresses.

2.1 Dolphin Echolocation and Shape Perception

Dolphins possess one of the most sophisticated biological sonar systems known in nature. Through their specialized melon organ, a lipid-filled structure in the forehead that focuses the emitted sound into a directional beam, dolphins emit broadband click signals in the frequency range of approximately 30 to 150 kHz, and receive returning echoes primarily through the lower jaw and surrounding fatty tissues [1]. In the echolocation process, dolphins integrate information across multiple echo features to identify targets [10]. Subsequent decades of behavioral research have established that dolphins can discriminate targets differing in quantity, size, shape, material composition, and wall thickness with remarkable precision [3, 11, 29]. For instance, Au and Pawloski (1992) demonstrated that a bottlenose dolphin could discriminate wall thickness differences as small as 0.23 mm in hollow aluminium cylinders at a range of 8 m, implying sensitivity to echo time-of-arrival differences

CHAPTER 2. LITERATURE REVIEW

on the order of hundreds of nanoseconds [2].

One of the capabilities that motivates this dissertation is the perception of shape through echolocation. The most compelling evidence that dolphins form genuine spatial representations from echoes comes from cross-modal echoic-to-visual(E-V) matching-to-sample (MTS) experiments, in which a dolphin first inspects a sample object using echolocation alone underwater, and must subsequently identify the matching object from visually presented alternatives in air [27]. The critical observation is that dolphins perform at near-errorless levels on the very first trial with novel objects they have never encountered before [15]. Since there has been no opportunity for associative learning between echoic and visual experiences of these objects, the dolphin must be constructing some form of shape representation directly from the echo returns, one that can be directly matched against its visual representation of the same object.

A natural concern is whether this cross-modal ability could be explained by the dolphin attending to a single acoustic feature rather than perceiving the object's overall geometry. Subsequent experiments addressed this by introducing alternative objects that deliberately shared class features with the sample, such as symmetry, closure, or linear orientation, and by including a "no-match" option that the dolphin could select when no alternative matched the sample [26]. If the dolphin were simply matching on a local feature, it would frequently confuse alternatives sharing that feature, or default to selecting the closest match rather than correctly choosing the no-match option when no true match was present. Instead, matching accuracy remained high and no obvious vacillation was observed, indicating that the dolphin perceived the global shape of objects rather than relying on isolated features. Furthermore, performance was symmetric across modality directions: accuracy in the visual-to-echoic (V-E) condition was comparable to echoic-to-visual (E-V), and the same object pairs that proved difficult in one direction also proved difficult in the other [27]. This symmetry suggests that the representations arising from echolocation and from vision are functionally equivalent, that is, echolocation provides the dolphin with a spatial description of object structure that is comparable in nature to what vision provides. It is this remarkable capacity for acoustic shape perception that the present dissertation seeks to replicate computationally.

The EV-MTS paradigm is also central to the present work. Wei et al. combined

behavioral EV-MTS experiments with finite-element acoustic simulations to investigate the potential limitations of dolphin echolocation for shape discrimination [34]. Their study tested dolphins across objects made of materials producing varying echo strengths, including air-filled PVC, water-filled PVC, and foam, and found that matching performance depended strongly on the acoustic contrast of the target material. This work provided both the behavioral data and the experimental setup from which the present dissertation draws its real-world acoustic recordings.

2.2 From Echoes to Target Geometry

Reconstructing the spatial structure of an underwater target from its acoustic echo returns is fundamentally an inverse problem. In the forward process, a known source emits an acoustic pulse that propagates through water, interacts with a target whose geometry and material properties shape the scattered field, and produces echo waveforms recorded at an array of receivers. The inverse problem seeks to recover the target’s spatial structure from these received signals. This problem is ill-posed in the classical sense [9]: the continuous scattered field does not uniquely determine the scatterer geometry from finitely many receivers, and the received echoes represent a superposition of reflections from all occupied regions of the target, so that recovering individual scatterer contributions requires solving a deconvolution problem whose stability degrades rapidly with increasing target complexity. Our discrete formulation on a pixel grid (Section 3.2.1.3) represents a deliberate regularization of this continuous problem, reducing the unknowns to a finite and tractable number.

Traditional approaches to underwater acoustic imaging address this inverse problem through various forms of beamforming. Conventional delay-and-sum beamforming and its adaptive variants estimate the spatial distribution of acoustic scatterers by exploiting time delays across array elements. While effective for source localization, these methods face fundamental resolution limits governed by the array aperture and the operating frequency [32, 35]. For compact arrays such as those relevant to dolphin-scale or portable biomimetic systems, the aperture is inherently small relative to the wavelength, leading to wide mainlobes and prominent grating lobes that severely degrade spatial resolution [33]. Synthetic aperture sonar can overcome aperture limitations by combining measurements from multiple platform

positions, but it requires precise motion estimation [14]. More advanced methods, including time-reversal techniques and matched field processing, exploit full wavefield information but depend on accurate environmental knowledge and are computationally expensive, making them unsuitable for real-time deployment [4, 12].

A critical observation, and one that motivates the present work, is that none of these conventional signal processing methods directly reconstructs target shape. Beamforming produces spatial power maps that indicate where scattering is concentrated, but these maps do not correspond to object boundaries or occupancy. Matched field processing estimates source or scatterer locations rather than geometric structure. Sparsity-aware approaches [33] produce improved spatial reflectivity maps rather than explicit shape predictions. The gap between scattering-intensity maps and actual target geometry arises because the relationship between echo structure and object boundary is inherently nonlinear, particularly when targets have fine structural features or when sparse sampling introduces ambiguity in the reconstructed map. This fundamental limitation of model-based signal processing approaches motivates the exploration of data-driven methods that can learn the nonlinear mapping from echo waveforms directly to spatial target representations.

2.3 Deep Learning for Acoustic Inverse Problems

Deep learning has emerged as a powerful framework for solving ill-posed inverse problems across a wide range of domains. The core insight is that neural networks can learn implicit regularization from training data, effectively encoding prior knowledge about the solution space without requiring explicit mathematical formulations of the regularizer [18]. This end-to-end learning paradigm has demonstrated success across a wide range of wave-based inverse problems, including medical image reconstruction from sparse measurements, seismic waveform inversion for subsurface velocity estimation, and electromagnetic inverse scattering for shape recovery, establishing that neural networks can learn to invert complex wave-propagation physics when provided with sufficient training data [25, 23, 20, 24]. These applications share a structural similarity with the present work: in each case, the input consists of indirect physical measurements from which a spatial map must be recovered, the

CHAPTER 2. LITERATURE REVIEW

inverse mapping is ill-posed, and deep learning provides the implicit regularization needed to produce plausible reconstructions.

Closer to the present domain, several recent works have applied deep learning to the specific problem of recovering spatial information from acoustic echoes. Christensen et al. developed BatVision, a bat-inspired system that uses a binaural microphone pair to record echoes from emitted chirp signals [8, 7, 5]. Their encoder-decoder architecture processes raw waveforms through temporal convolutions to predict depth maps and grayscale images of indoor environments, demonstrating that meaningful scene geometry can be learned from echo signals alone; the system operates from a single static emitter-receiver position without exploiting movement. Gao et al. proposed VisualEchoes, a framework that learns spatial image representations through echolocation in simulated 3D environments, where the agent moves through the scene and accumulates observations across multiple spatial positions [13]. Lindell et al. demonstrated acoustic non-line-of-sight imaging from a fixed loudspeaker-microphone configuration, reconstructing hidden scene geometry through learned processing of echo signals without relying on emitter or receiver motion [21]. These works collectively establish that deep learning can bridge the gap between 1D acoustic time-domain signals and 3D spatial representations, providing the conceptual foundation for the present dissertation. A common thread across these works is that they operate at room or scene scale, recovering environmental layout such as depth maps and floor plans from echolocation signals. The problem of reconstructing the shape of a discrete object at close range from multi-channel echo waveforms remains largely unexplored.

Having established the gap in echo-to-shape reconstruction, we turn to the choice of spatial output representation for the reconstructed target. In computer vision and robotics, occupancy-based representations have become a standard approach for encoding spatial geometry. Our output representation is a 2D pixel-wise binary occupancy map, in which each cell of a discretized spatial grid is assigned an occupancy probability. This can be viewed as a deliberate simplification of the volumetric occupancy grids used in 3D reconstruction [6]: by restricting prediction to a single 2D plane, we reduce the dimensionality of the output space and the data requirements for training, while still capturing the essential spatial structure of planar targets.

Chapter 3

Methodology

This chapter presents the methodology for reconstructing underwater target shapes from acoustic echo signals. The work comprises two main components: (1) data generation and preparation, which covers both the physics-based synthetic data pipeline and the real dolphin echolocation data processing pipeline; and (2) a deep learning network that learns the inverse mapping from echo waveforms to target occupancy maps. Section 3.1 formalizes the reconstruction task as an acoustic inverse problem. Section 3.2 describes the data preparation for synthetic and real data pipelines. Section 3.3 details the network architecture, and Section 3.4 presents the loss function and the training procedure.

3.1 Problem Formulation

This work addresses the problem of reconstructing the shape of an underwater target from acoustic echo signals, framed as an acoustic inverse problem. The formulation is motivated by the EV-MTS experiments, which demonstrated that dolphins can infer object shape from echolocation alone, suggesting that spatial representations can be derived purely from returned acoustic signals.

3.1.1 Forward Problem

In the forward acoustic process, a sound source (dolphin) at a known position $\mathbf{s} \in \mathbb{R}^3$ emits a click that propagates through water, interacts with the target, and produces echo waveforms received by an array of C hydrophones. The target is represented as a binary occupancy map $\mathbf{y} \in [0, 1]^{P \times P}$ defined on a discretized 2D pixel grid, where each pixel indicates whether the corresponding spatial region is

occupied ($y_{ij} = 1$) or empty ($y_{ij} = 0$). The forward mapping can be expressed as:

$$\mathbf{x} = \mathcal{F}(\mathbf{y}, \mathbf{s}) + \boldsymbol{\eta} \quad (3.1)$$

where $\mathbf{x} \in \mathbb{R}^{C \times L}$ denotes the multi-channel echo waveforms of length L , $\mathcal{F}$ represents the acoustic propagation operator determined by the physical environment, and $\boldsymbol{\eta}$ accounts for signal perturbations such as ambient noise. The grid resolution P is chosen as a trade-off between shape distinguishability and the solvability of the inverse problem: a coarser grid loses meaningful shape detail, while a finer grid increases the number of unknowns relative to the information content of the echo signals, making the reconstruction more severely underdetermined.

3.1.2 Inverse Problem

The reconstruction task is to invert this forward process: given observed echo waveforms $\mathbf{x}$, recover the target occupancy map $\mathbf{y}$. We approach this inverse problem using a data-driven strategy. Rather than attempting to analytically invert the forward operator $\mathcal{F}$, we train a neural network f_θ to learn the inverse mapping directly from data:

$$\hat{\mathbf{y}} = f_\theta(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M) \quad (3.2)$$

where $\mathbf{x}_m = \mathcal{F}(\mathbf{y}, \mathbf{s}_m) + \boldsymbol{\eta}_m$ for $m = 1, \dots, M$ are echo observations of the same target from M different source positions $\{\mathbf{s}_1, \dots, \mathbf{s}_M\}$, and $\hat{\mathbf{y}} \in [0, 1]^{P \times P}$ is the predicted occupancy probability map. The use of multiple source positions mirrors the biological observation that dolphins emit sequences of clicks from slightly varying locations and beam positions during echolocation, providing the model with complementary acoustic perspectives of the target.

3.1.3 Reconstruction Strategy

The target space is discretized into a $P \times P$ pixel grid, where each pixel independently receives a predicted occupancy probability. The model outputs a probability map $\hat{\mathbf{y}} \in [0, 1]^{P \times P}$ via a sigmoid activation, where values close to 1 indicate high confidence that the corresponding spatial region is occupied, and values close to 0 indicate likely absence. The collection of pixel-wise predictions forms the reconstructed target shape. This pixel-wise formulation allows the use

of overlap-based loss functions that directly measure the similarity between the predicted and ground-truth shapes (detailed in Section 3.4).

3.2 Data Preparation

This section describes how the echo data used for training and evaluation are produced. The system operates on two data sources: physics-based synthetic data generated by an acoustic simulator (Section 3.2.1), and real dolphin echolocation recordings processed from EV-MTS experiments (Section 3.2.2).

3.2.1 Synthetic Data Generation

To study the inverse reconstruction problem in a controlled setting where ground-truth occupancy maps are available, we generate synthetic echo data using a physics-based acoustic simulator. The simulator models the essential physics of the real experimental setup: a sound source emits a wideband pulse, which propagates through water, reflects off occupied target pixels, and arrives at an array of hydrophones. The following subsections describe the source signal model, the propagation and receiver geometry, the target representation, the data augmentation strategy, and the online data generation procedure.

3.2.1.1 Source Signal Model

The emitted pulse is modelled as a damped sinusoidal waveform, taking the form:

$$s(t) = A e^{-\frac{1}{2}\left(\frac{t-t_0}{\sigma}\right)^2} e^{-\alpha \frac{|t-t_0|}{\Delta\tau}} \cos(2\pi f_c (t - t_0)) w(t) \quad (3.3)$$

where a Gaussian envelope and an exponential decay term jointly shape the pulse amplitude, f_c is the centre frequency, t_0 is the time offset of the click peak, σ controls the Gaussian envelope width, $\Delta\tau$ is the click duration, α is the exponential damping factor, and $w(t)$ is a rectangular window that truncates the signal to $|t - t_0| \leq 2\Delta\tau$. The waveform formulation is chosen to produce a broadband pulse consistent with the characteristics reported for bottlenose dolphin clicks [1].

3.2.1.2 Acoustic Propagation Model

The simulator models acoustic propagation in a simplified environment with constant sound speed c , where only direct-path transmission and single-bounce target reflections are considered. Boundary effects such as tank wall reflections and surface or bottom interactions are not modeled. The geometry consists of three elements: a point source at position $\mathbf{s}$, a planar hydrophone array of C elements at known positions, and a target plane discretized into a $P \times P$ pixel grid at a fixed distance from the array. Each occupied pixel acts as an independent point scatterer that reflects the incident pulse back towards the receivers.

The received signal at hydrophone m is computed as a combination of a direct-path arrival and reflections from all occupied pixels:

$$x_m(t) = \frac{s(t - d_{sm}/c)}{d_{sm}} + \sum_{(i,j) \in \mathcal{O}} \frac{s(t - (d_{sp_{ij}} + d_{p_{ij}m})/c)}{d_{sp_{ij}} \cdot d_{p_{ij}m}} + \eta_m(t) \quad (3.4)$$

where d_{sm} is the source-to-hydrophone distance (direct path), $d_{sp_{ij}}$ and $d_{p_{ij}m}$ are the source-to-pixel and pixel-to-hydrophone distances respectively, $\mathcal{O} = \{(i, j) : y_{ij} = 1\}$ is the set of occupied pixels, and $\eta_m(t)$ is additive Gaussian noise at a controlled signal-to-noise ratio. The amplitude of each arrival decays inversely with propagation distance, modelling geometric spreading loss.

3.2.1.3 Target Representation

The choice of target shapes is critical for training a general-purpose shape reconstruction model. Early experiments with randomly scattered pixels produced patterns that lacked structural coherence, making it difficult for the network to learn meaningful shape features. Conversely, using a small set of predefined geometric shapes led to poor generalization, as the model tended to memorize the limited set of training patterns. These observations motivated the need for a target source that provides both large-scale diversity and structurally coherent, contiguous shapes.

We adopt the MNIST handwritten digit dataset [19] as the source of target patterns. Each digit image is downsampled and binarized by thresholding, then pasted onto the $P \times P$ target grid at a random position. MNIST is well-suited for this purpose: its large pool of training images provides ample shape diversity, the digit strokes are naturally contiguous, and the patterns span a range of structural complexities.

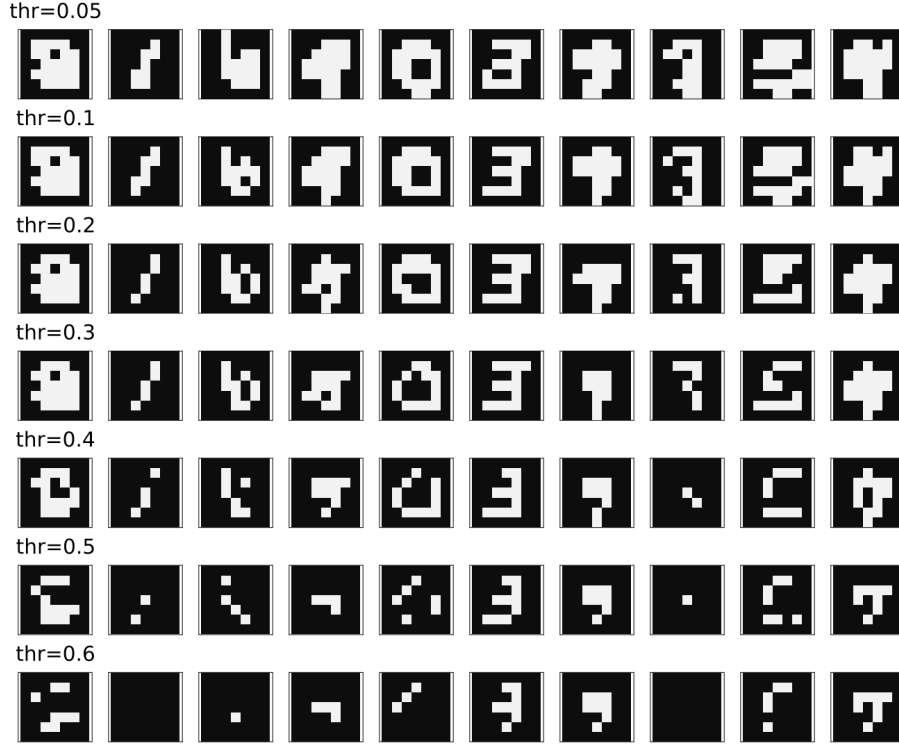


Figure 3.1: Mnist digit with different thresholds.

To further increase diversity, three forms of data augmentation are applied during occupancy generation:

1. **Random binarization threshold.** The threshold for converting the greyscale digit to a binary mask is sampled uniformly from a predefined range for each sample. Lower thresholds produce thicker, more filled shapes, while higher thresholds yield thinner, sparser patterns, as shown in Figure 3.1.
2. **Random rotation.** Each digit pattern is rotated by a random multiple of 90° before placement.
3. **Random placement.** The digit pattern is placed at a uniformly random position within the target grid, varying the spatial location of the target.

These augmentations ensure that, in combination with the large MNIST pool, the model encounters a virtually unlimited set of unique target configurations during training.

3.2.1.4 Source Position Variation

For each target occupancy pattern, multiple echo observations are generated by varying the source position within a bounded 3D region in front of the hydrophone array. This produces the set of M multi-view observations described in Section 3.1.

During training, source positions are drawn uniformly at random from this region for each sample. During evaluation, a fixed set of M positions is generated using a deterministic procedure to ensure reproducibility: candidate positions are constructed on a 3D grid within the source region, and M well-spread positions are selected via farthest-point sampling.

3.2.1.5 Online Data Generation

Rather than pre-generating a fixed dataset, training samples are synthesized on the fly. In each training iteration, a fresh occupancy pattern is generated by sampling an MNIST digit and applying the augmentation pipeline described in Section 3.2.1.3. For this pattern, the simulator is run M times with different source positions, each producing a C -channel echo observation, forming one training group. A training batch consists of N_{batch}/M such groups, where N_{batch} is the total batch size. This online generation strategy can effectively provide an unbounded training set, which greatly reduces the risk of overfitting.

3.2.2 Real Data Processing

The real echo data are derived from controlled EV-MTS experiments in which a trained bottlenose dolphin (*Tursiops aduncus*) actively echolocated on target objects while a multi-channel hydrophone array recorded the emitted clicks and returning echoes. The raw recordings contain a mixture of acoustic events that the algorithm needs to deal with. When the dolphin emits a click, the hydrophone array receives a direct-path arrival followed by target echoes, which are reflections from the sample object that carry shape information and is needed for reconstruction. However, the recordings also contain tank wall echoes, which are secondary reflections from the enclosure boundaries that arrive after the target echoes and act as noise. In addition, ambient noise and dolphin whistles are present throughout. The processing pipeline must therefore (1) detect genuine dolphin clicks while suppressing tank wall echoes that can be falsely detected as clicks, (2) extract the target echo component from

each detected click, and (3) estimate the dolphin’s source position for multi-view grouping. These three stages are described below.

3.2.2.1 Click Detection

The first step is isolating individual click events from continuous multi-channel recordings, so the target echoes could be extracted subsequently. Our detection pipeline builds upon the algorithm described in our earlier work [17], with several refinements to improve robustness on the multi-channel data used in the present study.

Signal conditioning. The raw recordings are first bandpass-filtered to retain the frequency range of dolphin clicks while attenuating low-frequency noise. The Hilbert transform is then applied to each channel independently to obtain the instantaneous amplitude envelope, which serves as the basis for all subsequent peak detection.

Channel quality assessment. Not all hydrophone channels produce usable signals across all recordings. Some channels exhibit persistently elevated baseline noise that would contaminate the detection envelope. We identify and exclude such channels: for each channel, the median envelope amplitude is computed, and channels whose median deviates from the cross-channel median by more than a threshold (measured in units of the median absolute deviation) are flagged as bad and excluded from further processing.

Detection envelope construction. Rather than summing channel envelopes (which amplifies noise) or relying on a single fixed reference channel (which may miss clicks arriving predominantly at other sensors), we construct the detection envelope as the sample-wise maximum across all good channels. At each time sample, the loudest channel is selected, producing an adaptive envelope that tracks whichever sensor is closest to the source at each moment.

Tonal masking. Dolphins occasionally produce tonal vocalizations (whistles) alongside their echolocation clicks. Because whistles also have high amplitude, they must be suppressed before peak detection to avoid being detected as clicks. We detect tonal regions using short-time spectral analysis on the summed filtered signal

across good channels: frames in which a small number of frequency bins concentrate a disproportionate fraction of the total spectral energy are classified as tonal and zeroed out in the detection envelope.

Global peak detection and filtering. Candidate click events are identified as local maxima of the detection envelope, subject to a minimum inter-peak distance. An adaptive amplitude threshold, derived from the median of the non-zero envelope values, separates genuine clicks from the noise floor. The remaining candidates are then passed through two filtering stages inherited from our earlier work to suppress false positives from tank echoes [17]. The first filter suppresses peaks whose amplitude falls below a dynamic fraction of the preceding strong peak, based on the observation that tank echoes exhibit sharp amplitude drops relative to the originating click. The second filter is applied twice to remove any peak whose amplitude is less than 90% of both its immediate neighbors, targeting residual tank echo artifacts.

Per-channel refinement. After global detection on the composite envelope, each detected click event is refined on every individual channel. Within a coarse window around the global peak, each channel’s envelope is searched for its own local maximum, yielding per-channel peak times, amplitudes, and frequencies. A fine window centered on each per-channel peak is used to extract the click waveform segment for feature computation. The dominant channel for each click is defined as the good channel with the highest peak amplitude.

3.2.2.2 Echo Extraction

For the reconstruction task, we require the click echo component of each click event, i.e., the portion of the received signal that contains reflections from the target, with the direct-path click removed.

The echo window boundaries are determined from the per-channel peak times computed during refinement. The echo start is defined as the latest peak time across channels with a sufficiently strong signal (those exceeding a fraction of the dominant channel’s amplitude). This corresponds to the last direct-path arrival, after which the signal transitions from direct click to target reflections. The echo end is defined

as the earliest peak time plus a fixed delay corresponding to the back-wall reflection distance, ensuring that the full set of target echoes is captured while excluding later multipath arrivals from the tank walls.

The echo waveform for each channel is then extracted from the bandpass-filtered signal within this window and stored alongside metadata including per-channel peak times, the dominant channel index, and echo window boundaries.

3.2.2.3 Source Position Clustering

In the synthetic data pipeline, each echo observation is generated with a known source position. For real data, the dolphin’s position at each click emission is unknown and must be estimated. Rather than recovering absolute 3D coordinates, we cluster clicks into groups that originate from approximately the same position, using time-difference-of-arrival (TDOA) features as a proxy for spatial location.

For each detected click, a TDOA feature vector is computed as the difference in peak arrival times between each feature channel and a stable reference channel. The reference and feature channels are chosen as the ones that are consistently marked as good across all recordings.

The TDOA feature vectors are then clustered using K-Means to identify the K most frequent dolphin positions. Outlier clicks, whose TDOA vectors fall far from their assigned cluster centroid (measured in units of the median absolute deviation), are discarded. Additionally, clusters that span fewer than a minimum number of distinct recording sessions are removed to guard against session-specific artifacts. The surviving clusters define the set of source positions used for multi-view grouping during training.

3.2.2.4 Echo Database Construction

The outputs of the preceding stages are assembled into a structured echo database. For each click that belongs to a retained cluster, its echo waveform is read from the stored per-channel segments, then normalized using z-score standardization. Only channels that are consistently good across all recordings are included, yielding a fixed channel subset.

The resulting database is indexed by target type and cluster identity, with each entry containing an array of multi-channel echo waveforms. This structure mirrors

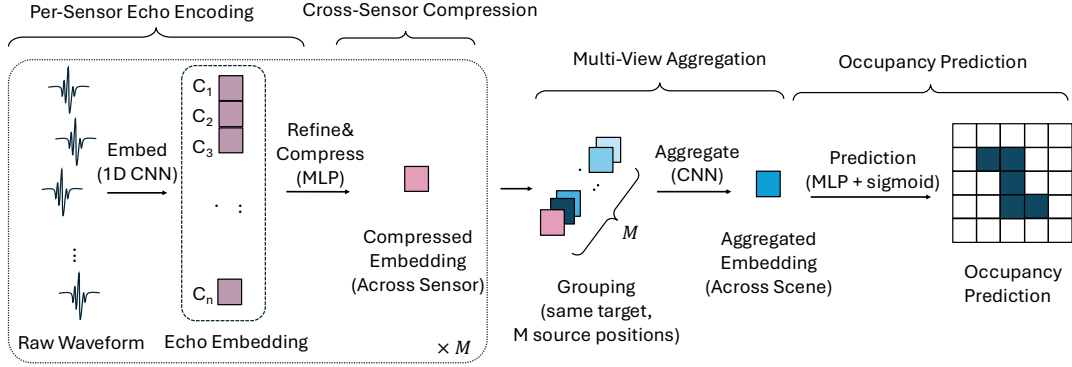


Figure 3.2: Model Architecture.

the synthetic data format: each cluster corresponds to a source position, and the echoes within a cluster form a group analogous to the M multi-view observations in the synthetic pipeline. The database is stored in HDF5 format for efficient random access during training and evaluation.

3.3 Network Architecture

This section describes the neural network f_θ introduced in Section 3.1. The network takes as input a group of M multi-channel echo waveforms $\{\mathbf{x}_1, \dots, \mathbf{x}_M\}$, each of shape $C \times L$, and outputs a pixel-wise occupancy probability map $\hat{\mathbf{y}} \in [0, 1]^{P \times P}$. The architecture consists of four stages, illustrated in Fig. 3.2.

3.3.1 Per-Sensor Echo Encoding

Each of the C hydrophone channels carries a raw echo waveform of length L . These waveforms are processed independently by a shared one-dimensional convolutional neural network (1D CNN) that takes a single-channel input and produces a feature map of fixed temporal length. Because the same network weights are applied to every channel, the encoder learns a universal echo representation while preserving the per-channel signal variations that encode spatial information about the target geometry. The per-sensor feature vectors are then refined by a shared MLP encoder applied independently to each sensor. This encoder operates per token to enrich each sensor’s representation before the cross-sensor stage.

3.3.2 Cross-Sensor Compression

The C refined per-sensor feature vectors are then fused into a single representation that captures cross-channel relationships such as inter-sensor amplitude and timing differences. We employ a learned MLP that projects across the sensor dimension, compressing C tokens into a single click-level embedding. This learned compression allows the network to weight sensors adaptively, attending more to channels that carry stronger or more informative echo signals for a given source-target configuration.

3.3.3 Multi-View Aggregation

As described in Section 3.1, the model receives M echo observations of the same target from different source positions. Each observation produces a click-level embedding through the encoding and compression stages above. The M embeddings are grouped and aggregated into a single scene-level representation via a learned 1D convolution over the group dimension. This aggregation allows the network to combine the complementary spatial information provided by different source positions, analogous to how a dolphin integrates information across a train of echolocation clicks emitted from varying positions and beam locations.

3.3.4 Occupancy Prediction Head

The scene-level embedding is mapped to the output occupancy map through a two-layer MLP followed by a sigmoid activation. Each output value represents the predicted probability that the corresponding pixel is occupied. To ensure numerical stability during training, the raw logits are passed through a soft clamping function before the sigmoid, bounding them to a fixed range and preventing gradient saturation from extreme predictions.

3.4 Training

3.4.1 Loss Function

The reconstruction task is a pixel-wise binary prediction problem with a class imbalance: in most target configurations, the majority of pixels are unoccupied. To

address this, we employ a composite loss function that combines three complementary terms.

Weighted binary cross-entropy. The base loss is the standard binary cross-entropy (BCE), augmented with a positive-class weight w_{pos} that up-weights the loss contribution of occupied pixels:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{P^2} \sum_{i,j} [w_{\text{pos}} y_{ij} \log \hat{y}_{ij} + (1 - y_{ij}) \log(1 - \hat{y}_{ij})] \quad (3.5)$$

where $y_{ij} \in [0, 1]$ is the ground-truth occupancy label and $\hat{y}_{ij} \in (0, 1)$ is the predicted probability. In typical target configurations, the majority of pixels are unoccupied, causing a standard BCE loss to be dominated by the empty-pixel class. The weight w_{pos} compensates for this imbalance by increasing the penalty for misclassifying occupied pixels, and is computed from the class ratio in each batch and clamped to a predefined range to prevent extreme weighting.

Dice loss. While BCE optimizes per-pixel accuracy, it does not directly encourage spatial overlap between the predicted and ground-truth shapes. We supplement it with the Dice loss, which quantifies the overlap between the predicted and ground-truth occupancy maps:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \sum_{i,j} \hat{y}_{ij} y_{ij} + \epsilon}{\sum_{i,j} \hat{y}_{ij} + \sum_{i,j} y_{ij} + \epsilon} \quad (3.6)$$

where ϵ is a smoothing constant for numerical stability. This term directly constrains the global shape similarity between the prediction and the target, complementing the pixel-level supervision provided by BCE.

L1 sparsity regularization. Inspired by sparsity-aware processing strategies used in previous work [33], we add an L1 penalty on the predicted probabilities to encourage sparse predictions:

$$\mathcal{L}_{\text{L1}} = \frac{1}{P^2} \sum_{i,j} |\hat{y}_{ij}| \quad (3.7)$$

This term encourages sparse predictions, pushing the probability map towards zero for pixels that do not belong to the target.

Total loss. The three terms are combined as a weighted sum:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \lambda_{\text{dice}} \mathcal{L}_{\text{Dice}} + \lambda_{\text{L1}} \mathcal{L}_{\text{L1}} \quad (3.8)$$

where λ_{dice} and λ_{L1} are scalar coefficients that control the relative contribution of each auxiliary loss. These coefficients, together with other training hyperparameters, are selected via grid search as described in Section 4.2.3.

3.4.2 Training Procedure

The network is trained end-to-end using the Adam optimizer. To account for the different roles of the network components, per-layer learning rates are used: the CNN encoder, the MLP encoder, and the prediction head each receive separate learning rates, which are jointly selected during hyperparameter search.

At each training iteration, a batch of training groups is generated on the fly by the synthetic data pipeline (Section 3.2.1.5). Each group consists of M echo waveforms from different source positions paired with a shared ground-truth occupancy map. The model processes the full batch, and the total loss is backpropagated through all stages. Optionally, a pixel masking strategy can be applied during training: instead of computing the loss over the full P^2 pixel grid, the loss is evaluated on a randomly sampled subset of pixels per training sample. This variant is described in the context of the real data experiment (Section 4.3), where it is used to mitigate overfitting to a small number of target shapes.

Early stopping based on the validation loss is used to prevent overfitting, with the best checkpoint retained for evaluation. The validation set consists of MNIST test split which is not seen during training, ensuring that the model is evaluated on genuinely novel target shapes.

During inference on real data, the forward pass is identical: echo waveforms extracted from dolphin recordings (Section 3.2.2) are fed through the same network to produce occupancy predictions, with clicks grouped by their TDOA-derived position cluster serving as the M multi-view observations.

Chapter 4

Experiments and Results

This chapter presents the experimental evaluation of the proposed reconstruction system on both synthetic and real data. Section 4.1 defines the evaluation metrics and the multiple-choice matching protocol shared by both experiments. Section 4.2 describes the synthetic data experiment, including the simulator configuration, training setup, and reconstruction results. Section 4.3 describes the real data experiment conducted on dolphin echolocation recordings, including the behavioral experiment setup, the data processing and the reconstruction results.

4.1 Evaluation Protocol

4.1.1 Pixel-wise Metrics

Reconstruction quality is assessed using per-pixel accuracy, computed between the predicted occupancy map $\hat{\mathbf{y}}$ and the ground-truth map $\mathbf{y}$ after binarizing the predictions at a threshold of 0.5:

$$\text{Accuracy} = \frac{1}{P^2} \sum_{i,j} \mathbf{1}[\hat{y}_{ij} = y_{ij}] \quad (4.1)$$

Because the occupancy maps are dominated by unoccupied pixels, a trivial all-zero prediction already achieves high accuracy. The multiple-choice matching protocol (Section 4.1.2) therefore serves as the primary evaluation of reconstruction quality, as it tests whether the model captures sufficient shape information to distinguish the correct target from distractors.

4.1.2 Multiple-Choice Matching Protocol

In addition to pixel-wise metrics, we evaluate the model using a multiple-choice matching protocol that is directly inspired by the EV-MTS behavioral paradigm used in dolphin sensory research [34]. In an EV-MTS experiment, a dolphin is presented with a target via echolocation and must subsequently select the matching object from a set of visual alternatives. We replicate this paradigm computationally to enable a direct comparison between the model’s shape discrimination ability and the dolphin’s behavioral performance.

In each trial, the model is presented with echo waveforms corresponding to a query target and asked to select the correct target shape from a set of N_c candidate occupancy maps (where $N_c = 2$ or $N_c = 4$). The model first produces a reconstructed occupancy map $\hat{\mathbf{y}}$ from the input echoes. Each candidate occupancy map $\mathbf{y}^{(k)}$ ($k = 1, \dots, N_c$) is then compared to $\hat{\mathbf{y}}$ using the Dice coefficient as a similarity measure, consistent with model loss:

$$\text{Dice}(\hat{\mathbf{y}}, \mathbf{y}^{(k)}) = \frac{2 \sum_{i,j} \hat{y}_{ij} y_{ij}^{(k)} + \epsilon}{\sum_{i,j} \hat{y}_{ij} + \sum_{i,j} y_{ij}^{(k)} + \epsilon} \quad (4.2)$$

The candidate with the highest Dice similarity is selected as the model’s answer. A trial is scored as correct if the selected candidate matches the true query target. The overall accuracy across all trials is reported. Each condition is evaluated over 100 independent trials. In each trial, N_c candidate digit patterns are sampled without replacement from the MNIST test set and placed at the center of the target grid. One candidate is randomly designated as the query target, and the model receives simulated echo waveforms for that target from $M = 10$ fixed source positions.

This evaluation protocol serves two purposes. First, it provides an intuitive measure of reconstruction quality that goes beyond pixel-wise statistics: the model must produce a prediction that is globally more similar to the correct target than to any distractor. Second, it establishes a direct parallel with the dolphin behavioral experiments, enabling a meaningful comparison between the model’s performance and that of the biological system it seeks to emulate. The use of both 2-choice and 4-choice variants allows the difficulty of the matching task to be controlled, with chance-level accuracy at 50% and 25% respectively.

3D Geometry layout for synthetic data generation

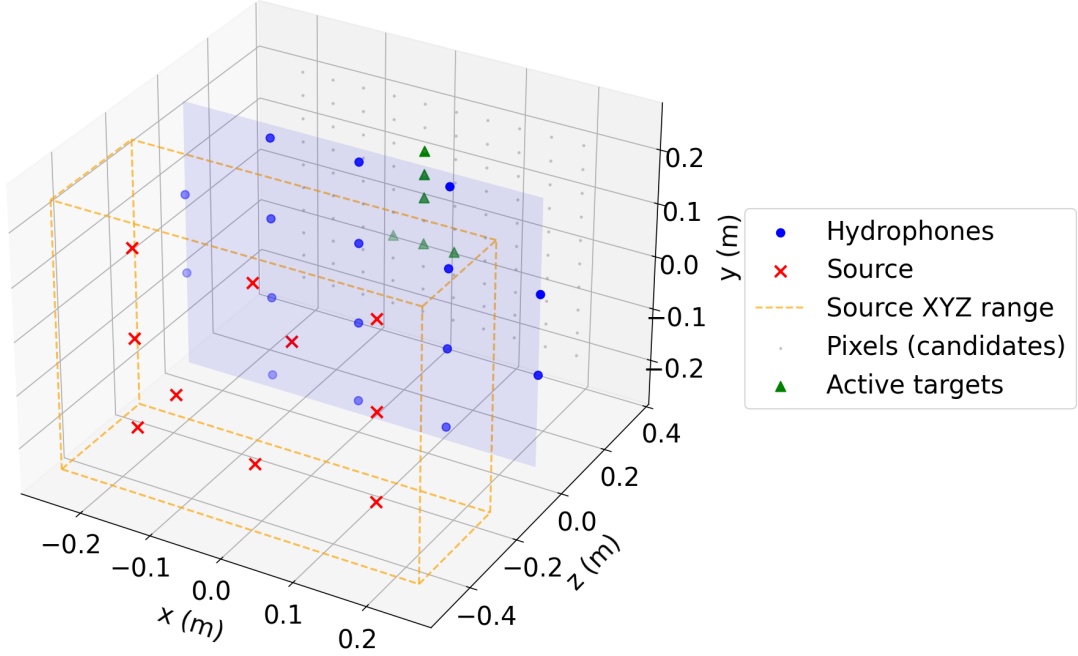


Figure 4.1: Synthetic data geometry layout.

4.2 Synthetic Data Experiment

4.2.1 Simulator Configuration

The acoustic simulator is configured to model a controlled underwater environment matching the physical dimensions of the real experimental setup. The geometry is defined in a coordinate system centered on the hydrophone array (Fig. 4.1), with the x - and y -axes spanning the lateral and vertical extents of the array, respectively. The array lies in the $z = 0$ plane, and the target is placed at $z = 0.35$ m in the forward direction.

The hydrophone array consists of $C = 16$ sensors arranged in a planar 3–5–5–3 layout, spanning approximately 0.50×0.45 m in the x - y plane, with inter spacings of 0.125 m horizontally and 0.15 m vertically. The acoustic source is sampled uniformly from a bounded 3D region behind the array: $x \in [-0.25, 0.25]$ m, $y \in [-0.25, 0.25]$ m, and $z \in [-0.5, -0.2]$ m. The target plane is a $P \times P = 10 \times 10$ pixel grid covering a

CHAPTER 4. EXPERIMENTS AND RESULTS

Table 4.1: Simulator configuration parameters.

Parameter	Symbol	Value
Number of hydrophones	C	16
Grid resolution	P	10×10 pixels
Sound speed	c	1500 m/s
Centre frequency	f_c	120 kHz
Signal-to-noise ratio	SNR	20 dB
Sampling rate	f_s	500 kHz
Group size (views per target)	M	10

0.4×0.4 m area, centered at the origin of the x - y plane.

The sound speed is set to $c = 1500$ m/s, consistent with typical values for seawater. The source signal is a damped Gabor pulse (Equation 3.3) with a center frequency of $f_c = 120$ kHz, a click duration parameter of $\Delta\tau = 50$ μ s, a Gaussian envelope width of $\sigma = \Delta\tau/4 = 12.5$ μ s, and an exponential damping factor of $\alpha = 2.0$. Additive white Gaussian noise is applied at a signal-to-noise ratio of 20 dB.

During evaluation, $M = 10$ source positions are selected using farthest-point sampling on a regular grid within the source region, ensuring a well-spread set of acoustic perspectives. Table 4.1 summarizes the key simulator parameters. The acoustic simulator is implemented in Python.

4.2.2 Target Patterns

Target patterns are derived from the MNIST handwritten digit dataset [19]. Each 28×28 digit image is downsampled to 7×7 pixels and binarized at a random threshold sampled uniformly from $[0.05, 0.5]$, then pasted at a random position within the 10×10 target grid. Random 90° rotations are applied to further increase diversity. The MNIST training set (60000 images) is used for training data generation, and the test set (10000 images) is reserved for evaluation. Because training samples are generated online with random augmentations and source positions, the effective training set is unbounded despite using a finite digit pool.

4.2.3 Training Configuration

All training experiments are conducted on a single NVIDIA A100 GPU. The model is implemented in PyTorch. The network is trained using the Adam optimizer

CHAPTER 4. EXPERIMENTS AND RESULTS

Table 4.2: Training hyperparameters (final configuration after grid search).

Hyperparameter	Value
Optimizer	Adam
Learning rate (CNN encoder)	0.001
Learning rate (MLP encoder)	0.003
Learning rate (prediction head)	0.005
Batch size (N_{batch})	100
Group size (M)	10
Dice loss weight (λ_{dice})	0.3
L1 loss weight (λ_{l1})	0.01
Dropout rate	0.1

with per-layer learning rates selected via grid search. Table 4.2 lists the final hyperparameter configuration. The loss function coefficients λ_{dice} and λ_{l1} (Equation 3.8) and the per-layer learning rates are jointly selected by grid search. The positive-class weight w_{pos} in the BCE loss is computed dynamically from the batch class ratio and clamped to $[1.0, 2.0]$. Early stopping based on validation loss is used to select the final checkpoint.

4.2.4 Results

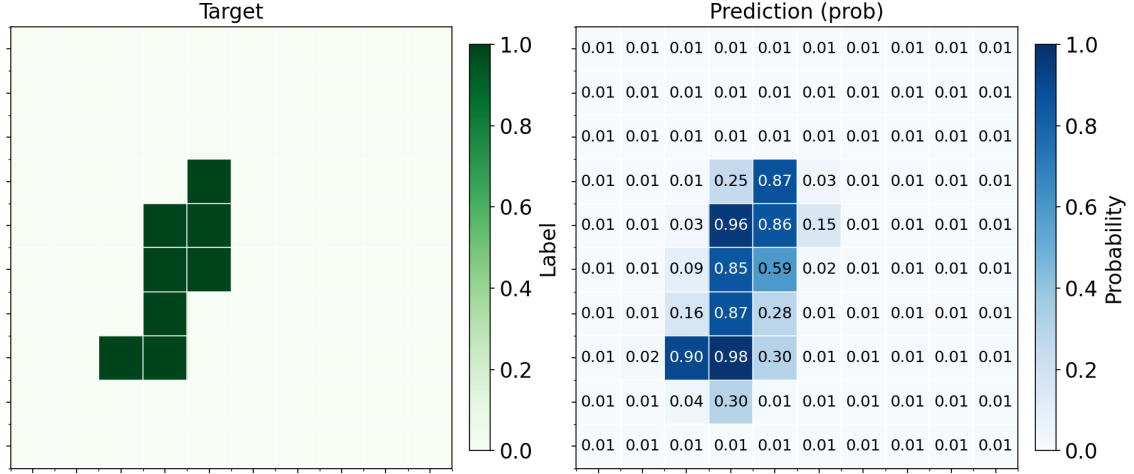
4.2.4.1 Reconstruction Performance

On the held-out MNIST test set, the model achieves a per-pixel accuracy of 95.2% after training for 75 epochs with online data generation. Figure 4.2 presents qualitative reconstruction results on held-out test samples. In each panel, the left subplot shows the ground-truth binary occupancy map, where each green cell indicates an occupied pixel in the target grid. The right subplot shows the corresponding model output as a continuous probability map, where each cell value represents the predicted occupancy probability prior to thresholding. Higher values are rendered in darker blue, allowing the spatial distribution of the model’s confidence to be read directly from the figure.

Across a diverse set of digit-shaped targets, the model consistently recovers the coarse spatial structure of each occupancy pattern: the predicted maps correctly localize the target within the grid and capture the overall silhouette, even when the target is pasted at varying positions within the 10×10 grid. The predictions tend to

CHAPTER 4. EXPERIMENTS AND RESULTS

Sample 175 | test split



Sample 73 | test split

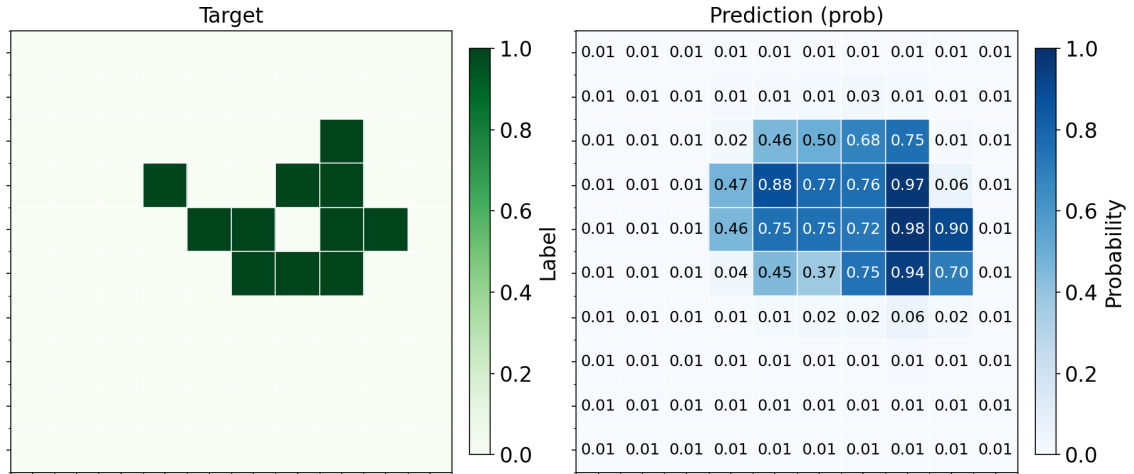


Figure 4.2: Reconstruction results on two held-out synthetic test samples. In each panel, the left subplot shows the ground-truth binary occupancy map and the right subplot shows the model's predicted probability map.

assign high probabilities (> 0.8) to pixels near the center of occupied regions, with values tapering off toward the boundaries. This soft falloff means that while the dominant structural features including rough aspect ratio and overall spatial extent are faithfully reproduced, fine-grained boundary details are occasionally smoothed, and thin strokes or small protrusions may be attenuated. One notable limitation is that hollow or concave structures are not well resolved: when the target contains an interior gap surrounded by occupied pixels, the model tends to fill in the gap with intermediate probabilities rather than predicting a clear void, as visible in the lower panel of Figure 4.2.

4.2.4.2 Multiple-Choice Matching Results

The multiple-choice matching accuracy was reported on the synthetic test set, evaluated over 100 independent trials per condition. The model achieves 93% accuracy on 2-choice trials and 81% on 4-choice trials, both substantially above the chance levels of 50% and 25% respectively.

Figure 4.3 shows representative trials from both conditions. Each panel displays a single trial. The top row shows the query target (left) and the model’s reconstructed probability map (right). The bottom row shows the N_c candidate occupancy maps, each annotated with its Dice similarity to the reconstruction. The correct candidate is marked with a star; the candidate selected by the model (i.e., the one with the highest Dice score) is highlighted. A trial is correct when these two coincide. In correct trials, the reconstructed occupancy map exhibits a visibly higher Dice similarity to the ground-truth candidate than to any distractor. In incorrect trials, errors often occur when two candidates share a similar global structure, such as shapes with comparable width and vertical extent. In these cases, the reconstruction’s soft probability map may match the wrong candidate.

4.3 Real Data Experiment

The synthetic experiment demonstrates that the proposed architecture can reconstruct novel target shapes from simulated echo waveforms. The real data experiment tests whether the same architecture, retrained on physically recorded dolphin echolocation signals, can learn to extract shape information from real acoustic

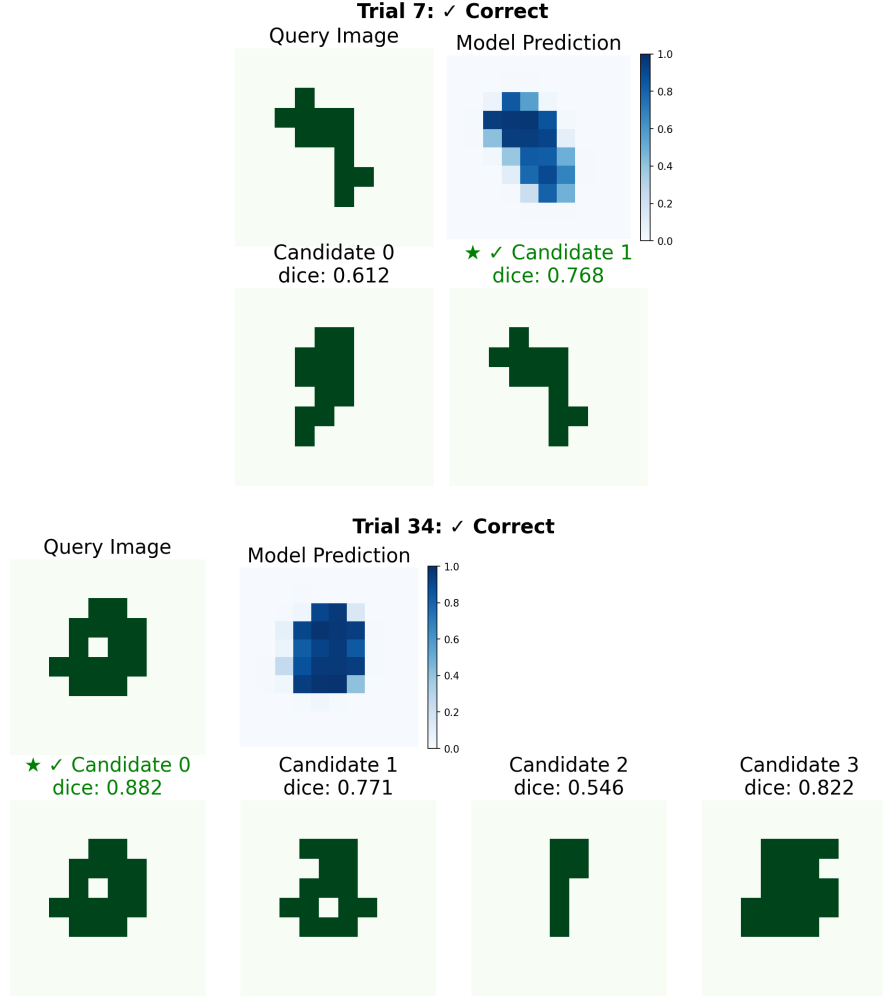


Figure 4.3: Representative correct trials from the 2-choice (top) and 4-choice (bottom) multiple-choice matching experiments on synthetic data. In each panel, the top row shows the query target and the model’s reconstructed probability map. The bottom row shows the candidate occupancy maps, each annotated with its Dice similarity to the reconstruction. The correct candidate is marked with a star.

data. The model structure, loss function, and training procedure are identical to those used for the synthetic experiment; only the data source changes.

4.3.1 EV-MTS Experiment and Recording Setup

The real echolocation data were collected during EV-MTS experiments conducted at the Marine Mammal Breeding and Research Centre (MMBRC) of Ocean Park Hong Kong (OPHK). The subject was an adult male Indo-Pacific bottlenose dolphin (*Tursiops aduncus*) named Ginsan.

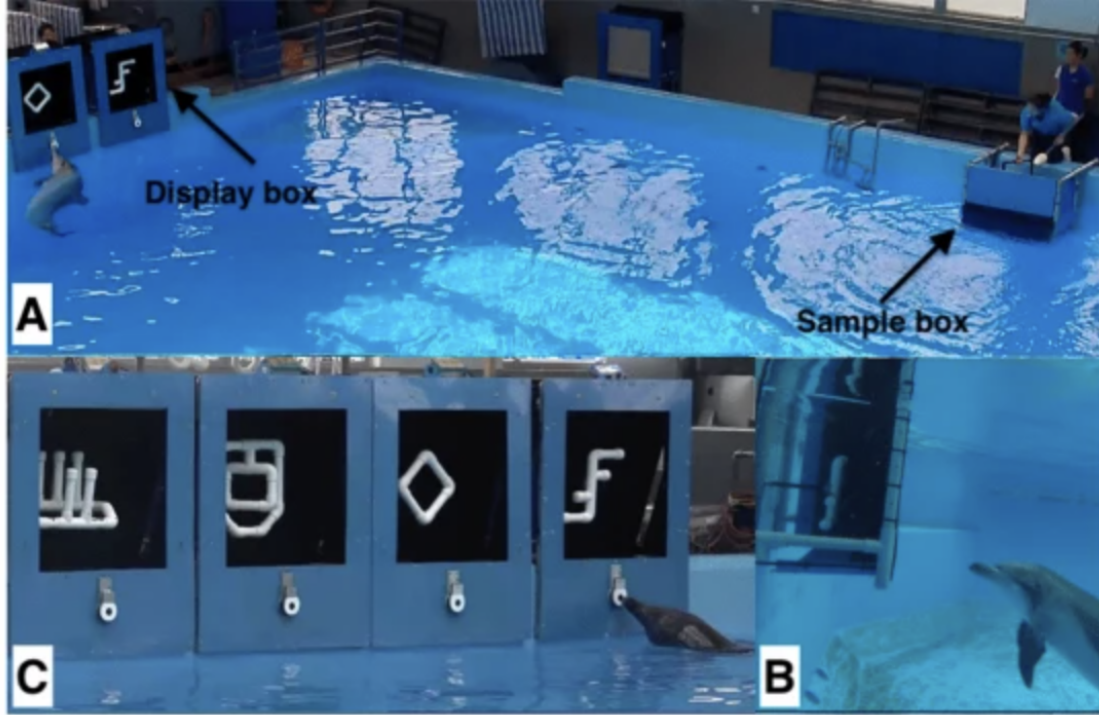


Figure 4.4: The EV-MTS experimental setup, reproduced from [34]. (A) Layout of the sample box (right, underwater) and display boxes (left, in air). (B) The dolphin echolocating on the concealed object underwater. (C) The dolphin visually selecting the matching alternative in air.

The experimental setup follows the protocol described by Wei et al. [34] and Vishnu et al. [33], as shown in Figure 4.4. In each trial, a sample object constructed from air-filled PVC pipes is placed inside an underwater sample box (0.9×0.9 m). A 3 mm opaque plexiglass screen, acoustically transparent in water, is mounted on the front of the box to block the dolphin’s vision while allowing echolocation signals to pass through. The dolphin freely positions itself in front of the box and echolocates on the concealed object. After the echoic inspection, the dolphin swims to the opposite side of the pool, where up to four alternative objects are revealed in air for visual matching. A planar hydrophone array is placed behind the plexiglass screen to record the acoustic signals at a sampling rate of 500 kHz. The recording hardware was upgraded during the course of the project from a single hydrophone to 16 channels and eventually to 32 channels. The array layout is shown in Figure 4.5.

Two target shapes from the EV-MTS paradigm are used in this study: SQ (square) and FF (double-F), both constructed from 32 mm diameter air-filled PVC

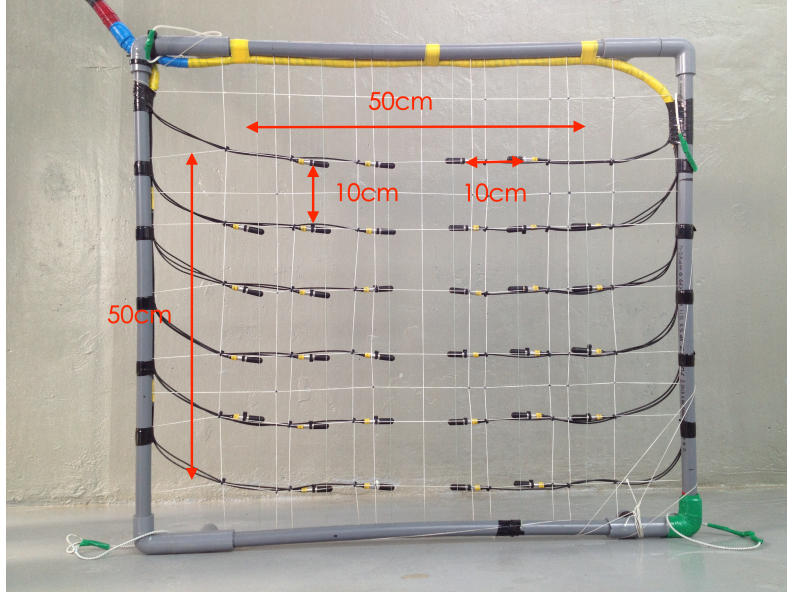


Figure 4.5: Layout of the 32-channel hydrophone array used in the real data recordings.

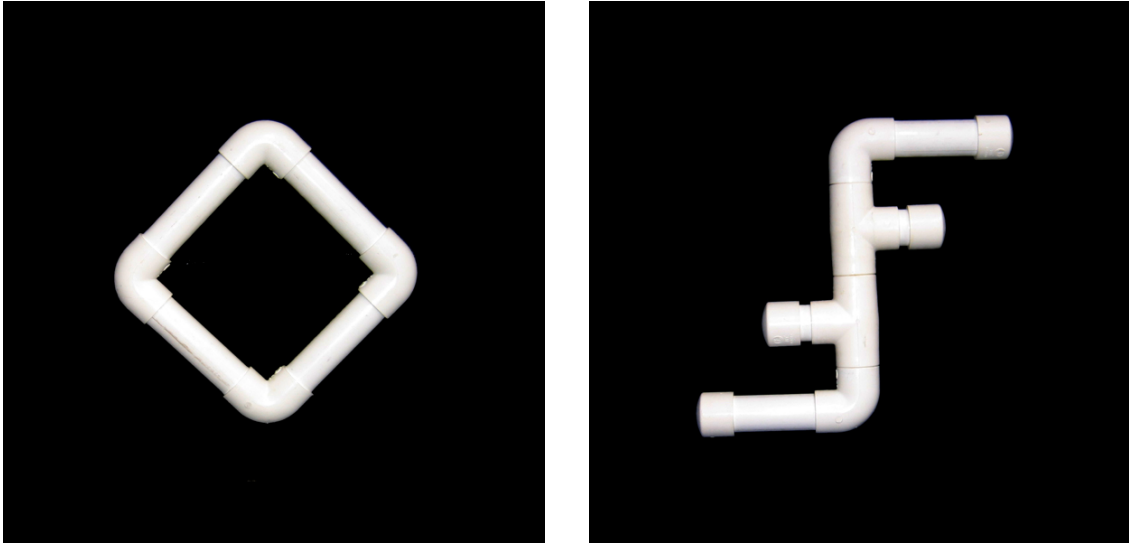


Figure 4.6: Target SQ (left) and FF (right)

pipes, as shown in Figure 4.6. These two shapes are equated for reflective surface area to ensure comparable overall acoustic reflectivity. From the full recording archive, we select 51 trials recorded with the 32-channel acquisition system in which the target is either SQ or FF.

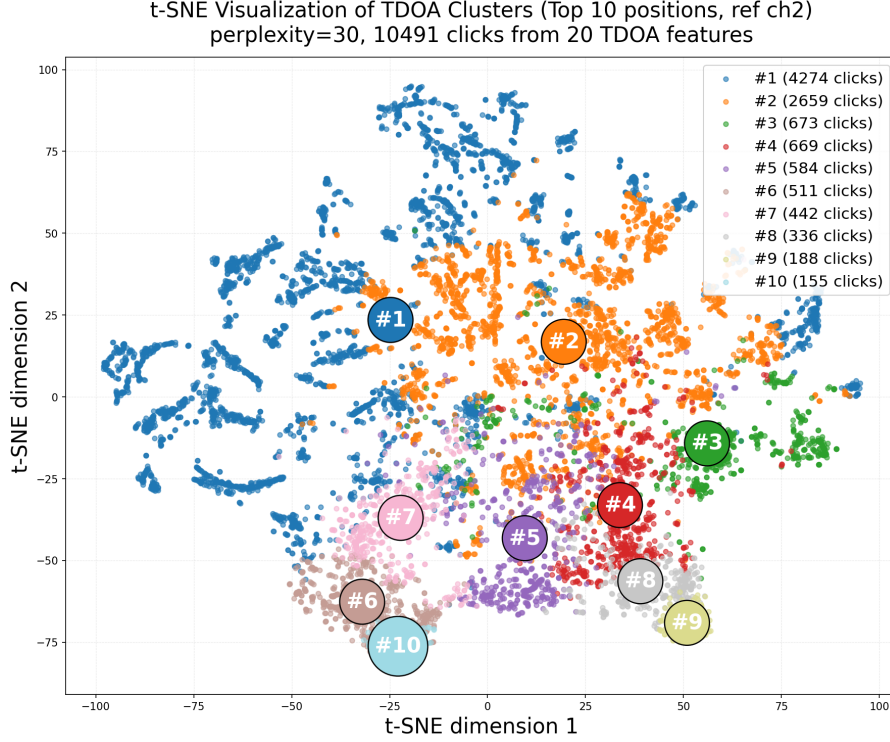


Figure 4.7: 10 echo clusters in TDOA space, t-SNE applied.

4.3.2 Data Selection and Processing

The click detection pipeline (Section 3.2.2.1) is implemented in Julia and applied to each trial independently. The number of detected clicks per trial varies widely, reflecting the dolphin’s variable click emission rate across sessions. In total, 10491 valid clicks are extracted across all 51 trials.

After click detection and echo extraction (Section 3.2.2.2), TDOA-based source position clustering (Section 3.2.2.3) is applied to group clicks by approximate dolphin position. K-Means clustering with $K = 10$ is used, with an outlier rejection threshold of 3σ ($867.5\mu s$) from each cluster centroid in TDOA space. 481 outliers are rejected at this threshold. The resulting clusters range in size from 155 to 4274 clicks. Figure 4.7 visualizes the 10 clusters in two dimensions using t-SNE applied to the TDOA feature vectors, confirming that the clusters correspond to well-separated regions in TDOA space. The construction of the echo database is implemented in Python.

Of the 32 recording channels, 20 are identified as consistently good across all

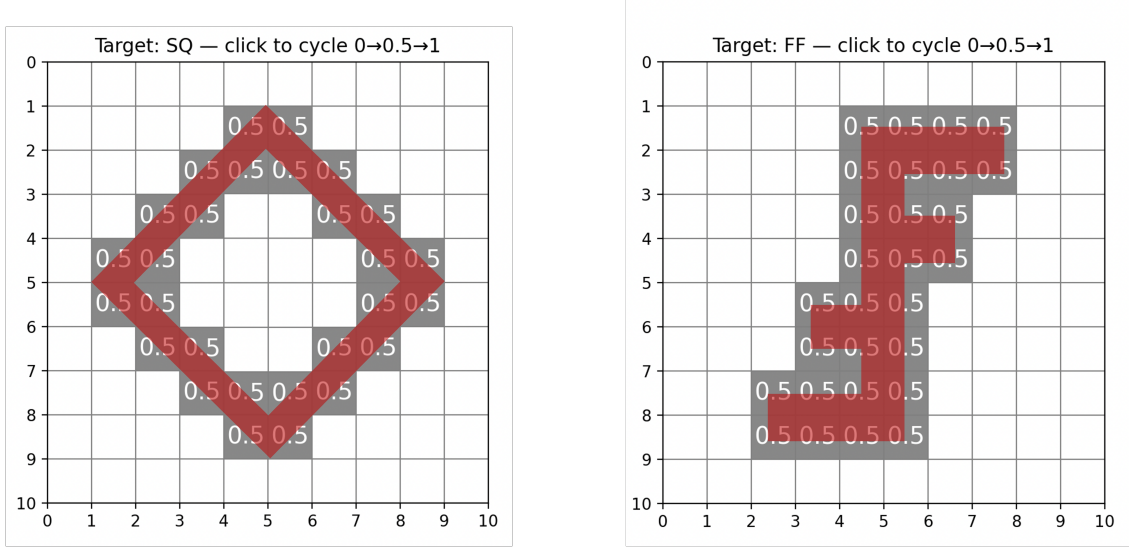


Figure 4.8: Ground-truth occupancy maps for the SQ (left) and FF (right) targets. The red shade shows the physical extent of each PVC pipe target.

sessions based on the channel quality assessment procedure (Section 3.2.2.1), using one channel (ch2) as the reference for TDOA computation and 19 channels as feature channels.

4.3.3 Ground-Truth Occupancy Maps

To train and evaluate the model on real data, a ground-truth occupancy map must be defined for each target shape. Because the model outputs a $P \times P = 10 \times 10$ pixel grid using the same representation as the synthetic experiment, the physical target shapes are manually mapped onto this grid. Figure 4.8 illustrates this mapping: the red shade shows the physical extent of each PVC pipe target overlaid on the discretized grid. Each pixel that intersects any part of a pipe segment is assigned a value of 0.5, preserving the target shape continuity; pixels that are entirely empty are assigned a value of 0. A uniform nonzero value is used for all occupied pixels, along with the same number of occupied pixels, to prevent the model from developing a bias toward one target shape due to differences in label.

Unlike the synthetic experiment, where each evaluation trial uses a distinct MNIST digit pattern, the real data experiment uses only two fixed ground-truth maps. The purpose of this mapping is not to evaluate fine-grained reconstruction accuracy on diverse shapes, but rather to provide a consistent supervisory signal for

training.

4.3.4 Training Configuration

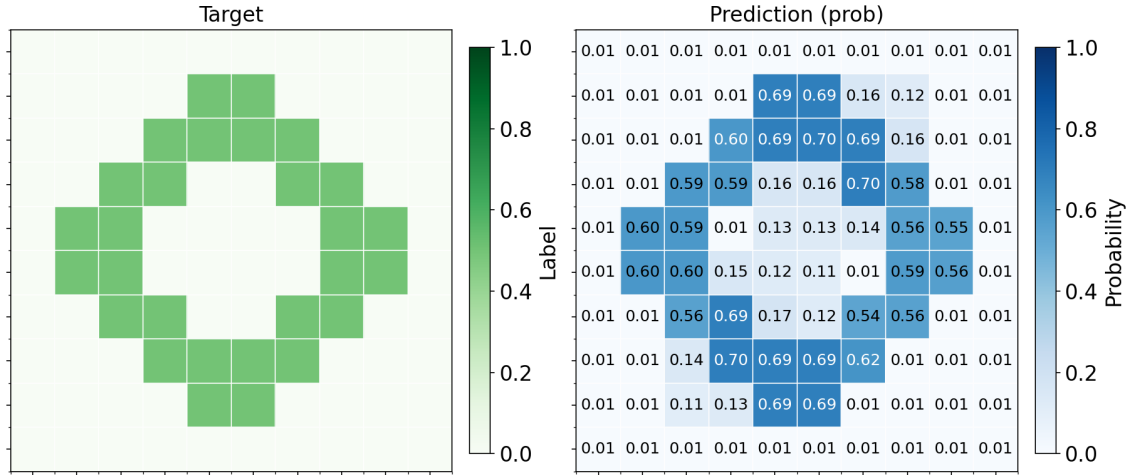
The echo database is split into training and test sets at the per-cluster level: within each cluster, 80% of the echoes are randomly assigned to training and the remaining 20% to testing, using a fixed random seed for reproducibility. During training, each batch is constructed by randomly selecting a target type (SQ or FF), sampling one echo from each of the $K = 10$ position clusters for that target, and grouping them as the M multi-view observations. The corresponding ground-truth occupancy map is the 10×10 grid described in Section 4.3.3. Training uses the same composite loss function (Equation 3.8) and Adam optimizer as the synthetic experiment, with hyperparameters carried over from the synthetic grid search.

When trained with the standard loss computed over all $P^2 = 100$ pixels, the model rapidly converges to near-binary predictions that closely replicate the ground-truth template for each target type, achieving near-perfect matching accuracy but producing no meaningful per-pixel variation. To encourage per-pixel discriminative predictions, a pixel masking strategy is applied: in each training iteration, the loss is computed on a single randomly selected pixel per sample rather than the full grid. This forces the network to produce a meaningful prediction at every spatial location independently. Under single-pixel masking, the Dice and L1 loss components degenerate to single-pixel operations and lose their intended roles as shape-overlap and sparsity measures respectively; the effective training signal is therefore dominated by the per-pixel binary cross-entropy term. The composite loss formulation is retained for consistency with the synthetic training pipeline, though the auxiliary terms contribute negligibly under this setting.

For real data, we set the model’s input channel C to 20 to accept all the good channels from the real array setup. The echo waveforms are normalized using z-score standardization, matching the preprocessing applied during synthetic training.

CHAPTER 4. EXPERIMENTS AND RESULTS

Sample 7 | test split



Sample 10 | test split

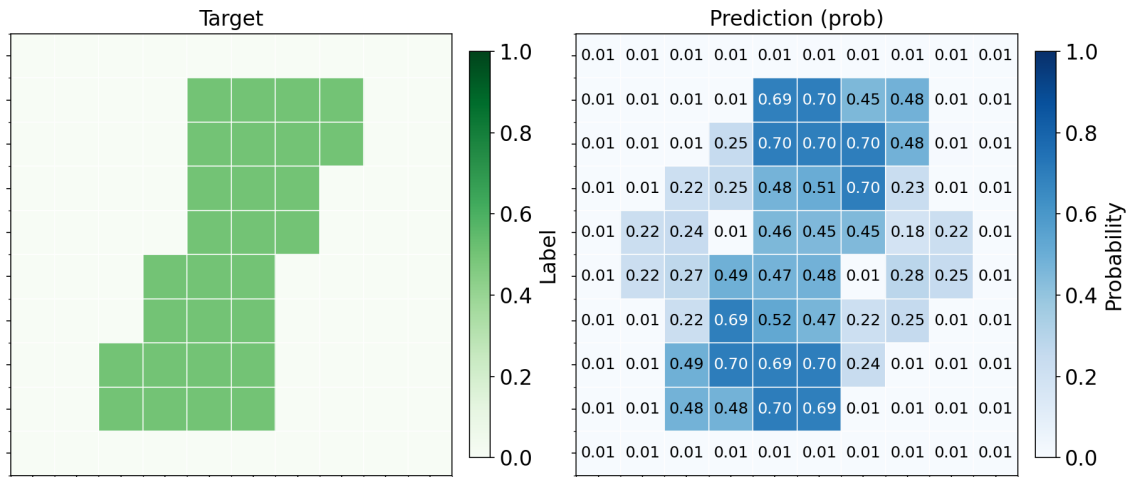


Figure 4.9: Reconstruction results on two held-out real test samples.

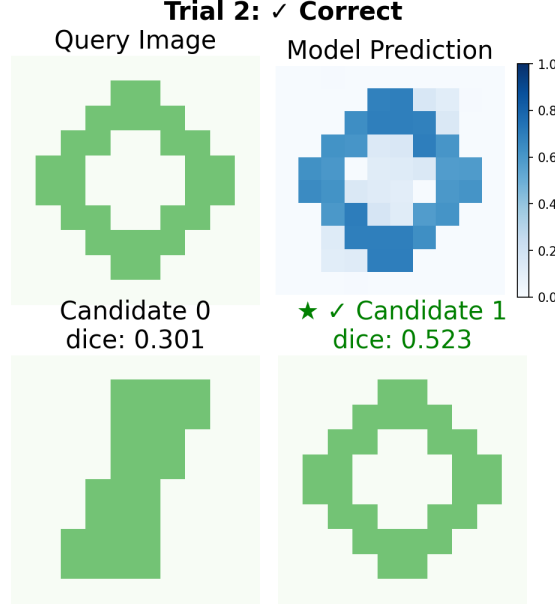


Figure 4.10: Representative correct trials from the 2-choice multiple-choice matching experiments on real data.

4.3.5 Results

4.3.5.1 Reconstruction Performance

On the echo database test split, the model achieves a per-pixel accuracy of 91.34% after training for 86 epochs. Figure 4.9 presents representative reconstruction results on held-out test echoes for the FF target. The pixel-masked model outputs spatially varying probability maps with intermediate confidence values, typically peaking around 0.7 for occupied regions. This variation across pixels indicates that the model has learned per-pixel discriminative behavior rather than memorizing a fixed output template for each target type. The predicted probability maps show higher confidence in regions where the two target shapes overlap spatially, and lower confidence in regions that are unique to one target.

4.3.5.2 Multiple-Choice Matching Results

Because only two target shapes are available in the real dataset, the multiple-choice matching evaluation reduces to a 2-choice protocol: given a set of echo waveforms, the model reconstructs an occupancy map and selects whichever of the two ground-truth maps (SQ or FF) yields a higher Dice similarity. The model

CHAPTER 4. EXPERIMENTS AND RESULTS

achieves 84% accuracy over 100 test trials, with a chance level of 50%. One example of the multiple choice experiment is shown in Figure 4.10. This accuracy is well above chance, confirming that the echo waveforms carry target-discriminative information that the model can extract.

Chapter 5

Discussion

This chapter interprets the experimental results presented in Chapter 4 and situates them in the broader context of acoustic shape reconstruction and dolphin echolocation research. Section 5.1 discusses what the synthetic data results reveal about the information content of echo waveforms. Section 5.2 examines the real data results and the distinction between reconstruction and classification. Section 5.3 compares the proposed approach with prior work and with dolphin behavioral performance. Section 5.4 identifies the limitations of the current work, and Section 5.5 outlines directions for future research.

5.1 Shape Information in Echo Waveforms

The synthetic data results provide direct evidence that multi-channel echo waveforms encode sufficient geometric information to support pixel-wise shape reconstruction of previously unseen targets. The model achieves 93% accuracy on 2-choice matching trials and 81% on 4-choice trials, both substantially above the chance levels of 50% and 25% respectively. These results are obtained on targets drawn from the MNIST test set, which the model has never encountered during training, confirming that the network has learned a generalizable inverse mapping from echo waveforms to occupancy maps rather than memorizing target-specific echo associations.

It is worth emphasizing that the MNIST digits, after downsampling to 7×7 pixels, binarization at a random threshold, and random 90° rotation, no longer resemble recognizable digit classes. The resulting patterns function as a diverse pool of arbitrary binary shapes with varying spatial extent, connectivity, and orientation. The choice of MNIST is therefore motivated not by digit recognition but by the

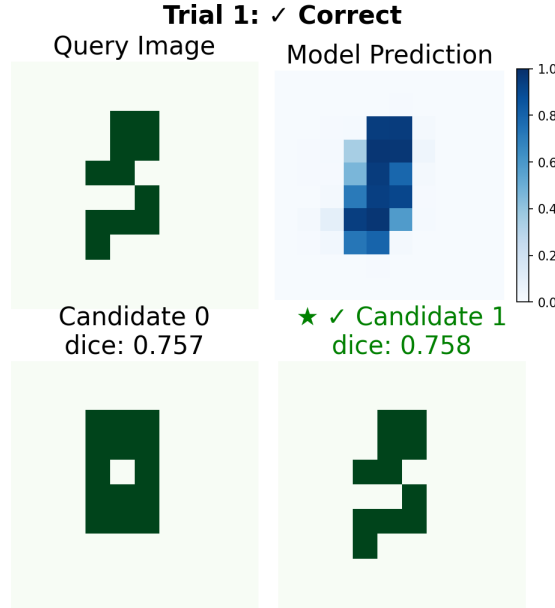


Figure 5.1: A correct 2-choice matching trial illustrating the effect of reconstruction blurriness on discriminative power. Although the two candidates differ visibly in their fine structure, the model’s diffuse probability map yields nearly identical Dice scores (0.757 vs. 0.758), resulting in a correct selection by a margin of only 0.001. This example shows that even in successful trials, the reconstruction may lack sufficient sharpness to reliably distinguish a candidates.

need for a large and structurally diverse set of contiguous binary patterns suitable for training and evaluation.

As a first attempt at learned acoustic shape reconstruction, the 10×10 pixel resolution is deliberately coarse, yet the model demonstrates effective recovery of the dominant spatial structure of each target: the overall silhouette and spatial extent are faithfully reproduced in the predicted probability maps. The reconstruction quality is sufficient to discriminate the correct target from distractors in the majority of matching trials, which is the functional criterion that matters for the shape recognition task.

The error cases reveal a characteristic failure mode that is informative about the difficulty of the inverse problem. In incorrect matching trials, the model typically produces a reconstruction that captures the coarse spatial layout of the target but fails to resolve fine-grained details such as small interior gaps, thin strokes, or precise boundary placement. The predicted probability maps in these cases

tend to be diffuse, assigning intermediate confidence values across a broad region rather than producing sharp, well-localized occupancy predictions. As a result, the reconstruction may be roughly consistent with multiple candidate shapes, and the Dice similarity scores between the prediction and different candidates can differ by only a narrow margin. For example, figure 5.1 illustrates that even in correct trials, the margin can be extremely narrow when the reconstruction is diffuse.

This blurriness reflects the inherent difficulty of the acoustic inverse problem: extracting fine spatial detail from echo waveforms requires the model to reliably decode subtle differences in arrival times and amplitude patterns across channels, and for some target configurations the echo signals may simply not carry enough discriminative information to support a sharp reconstruction. The issue is not primarily one of grid resolution, as even at 10×10 pixels many targets are structurally distinct; rather, the model’s learned inverse mapping does not always produce reconstructions that are precise enough to distinguish between candidates that differ in their fine details.

The gap between 2-choice and 4-choice accuracy (93% vs. 81%) is consistent with this interpretation. In the 2-choice setting, the reconstruction need only be globally more similar to the correct target than to a single distractor, a relatively lenient requirement. In the 4-choice setting, the probability that at least one distractor is structurally similar to the correct target increases, demanding finer shape fidelity from the reconstruction to maintain discrimination. Despite this increased difficulty, the 81% accuracy on 4-choice trials remains well above chance and indicates that the model captures shape information beyond what would be needed for coarse category-level discrimination.

5.2 Real Data: Per-Pixel Discrimination and Implications

The real data experiment serves a fundamentally different purpose from the synthetic experiment. While the synthetic setting evaluates true reconstruction, where the model must recover the shape of a target it has never seen, the real data setting tests whether the same architecture can learn to extract shape-relevant information from physically recorded dolphin echolocation signals. The distinction

is important because the real dataset contains only two target shapes (SQ and FF), and this constraint shapes both the training strategy and the interpretation of the results.

With only two targets in the training set, a model trained with the standard full-grid loss rapidly converges to template memorization: it learns to associate the acoustic signature of each target type with a near-binary replica of the corresponding ground-truth map, achieving near-perfect matching accuracy but producing no meaningful spatial variation in its predictions. This classification-like behavior is expected, as the network has sufficient capacity to memorize two fixed output templates.

The pixel masking training strategy addresses this issue by forcing the model to predict each pixel independently, with the loss evaluated on a single randomly selected pixel per sample. Under this regime, the model can no longer rely on reproducing a global template and must instead learn to make locally grounded predictions at every spatial location. The resulting probability maps exhibit structured spatial variation: predicted confidence is higher in regions where the SQ and FF targets overlap on the grid, and lower in regions occupied by only one of the two shapes. This pattern suggests that the model has learned which spatial locations are acoustically distinguishable between the two targets, rather than simply identifying target identity and recalling a template.

This result carries implications beyond what the earlier template-memorizing model could demonstrate. The spatially varying predictions provide evidence that the echo waveforms contain not only global target-identity information but also spatially localized information that supports per-pixel occupancy estimation. The real echolocation recordings are subject to physical complexities absent from the synthetic data, including multipath reflections from the tank walls and water surface, variable dolphin position and beam orientation across trials, ambient noise from the marine environment, and interactions between the acoustic field and the plexiglass screen. That the model produces spatially structured per-pixel predictions under these conditions demonstrates that fine-grained acoustic information survives physical propagation in the real experimental environment.

However, an important interpretive caveat remains. With only two targets, the model’s confidence map may reflect a spatial discriminability map, encoding which

pixel locations are most informative for distinguishing SQ from FF, rather than a true per-pixel occupancy estimate. These two interpretations are indistinguishable in a two-target setting. Resolving this ambiguity requires training and evaluating on a larger set of target shapes, where the model must predict per-pixel occupancy for targets it has not encountered during training. The synthetic experiment already demonstrates that this generalization is achievable in a controlled setting; the open question is whether it transfers to real acoustic data with sufficient target diversity.

5.3 Comparison with Prior Work

5.3.1 Comparison with Prior Approaches

The present work relates to two bodies of prior research: physics-based array processing methods for acoustic shape visualization, and deep learning systems that recover spatial information from echo signals in other domains.

The most directly comparable prior work is the biomimetic sonar system of Vishnu et al. [33], which applied sparsity-aware (SA) signal processing to dolphin echolocation recordings and biomimetic sonar signals collected in the same experimental setup used in this dissertation. Their SA processor, based on compressed sensing principles, produces spatial reflectivity maps that reveal the distribution of scattering energy across the target plane, and achieves substantially improved spatial resolution compared to conventional Bartlett beamforming.

A key difference between the two approaches lies in the nature of the output. The SA processor produces a continuous-valued spatial power map that indicates where acoustic scattering is concentrated. Shape features are visible in these maps, particularly when combined with knowledge of the target geometry, but the output does not constitute an explicit shape prediction; interpreting the map in terms of target boundaries requires either human judgment or additional post-processing. Figure 5.2 shows the SA processing output applied to dolphin echolocation recordings of the same SQ and FF targets used in the present study. The present work, by contrast, produces a direct pixel-wise occupancy prediction that constitutes an explicit shape estimate. The model outputs a probability for each grid cell indicating whether that spatial region is occupied, yielding a discrete reconstruction that can

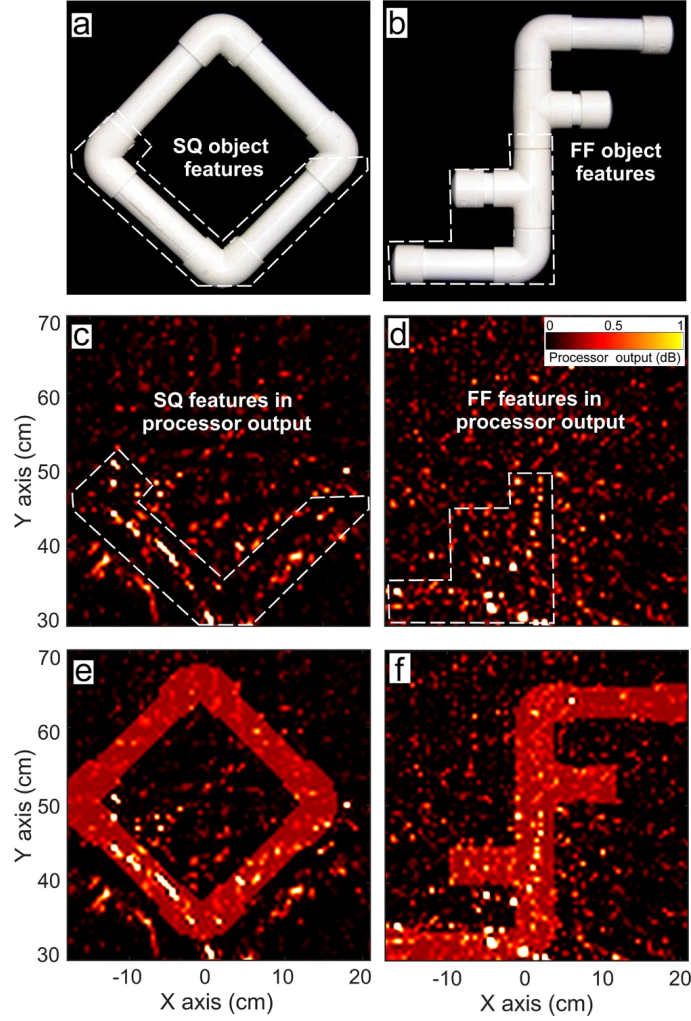


Figure 5.2: Sparsity-aware (SA) processor visualizations of dolphin echolocation data for the SQ and FF targets, reproduced from [33]. (a, b) Target objects. (c, d) SA processor output. (e, f) SA output with shading masks overlaid to highlight matching features.

be evaluated automatically against ground-truth occupancy maps without further interpretation.

A second distinction concerns generalization. The SA processing pipeline operates on physical principles and does not require training data, making it applicable to any target without prior exposure. However, as a signal processing method, it does not learn from data and cannot improve its output by leveraging patterns across many target-echo pairs. The deep learning approach requires a training phase with ground-truth labels, but in return it learns an inverse mapping that generalizes to novel target shapes, as demonstrated by the synthetic data results on unseen MNIST

patterns.

These differences make the two approaches complementary rather than competitive. The SA processing provides a principled, training-free baseline grounded in acoustic physics, while the deep learning approach offers an explicit shape prediction with the capacity to learn from data. A natural direction for future work is to combine the strengths of both paradigms, for example by using physics-based processing to provide spatial priors or feature representations as input to a learned reconstruction network.

Beyond this specific experimental context, a growing body of work has demonstrated that deep learning can bridge the gap between acoustic echo signals and spatial representations. Systems such as BatVision [8], which predicts depth maps and grayscale images from binaural echoes of chirp signals in indoor environments, and VisualEchoes [13], which learns spatial image representations through echolocation in simulated 3D scenes, establish the broader feasibility of learned echo-to-spatial mappings. These works operate in different domains (in-air, room-scale, single or binaural receivers) and address different tasks (scene-level depth estimation rather than object-level shape reconstruction), but they share the core principle that neural networks can learn to extract geometric information from echo waveforms. The present work extends this principle to the underwater domain with a multi-channel hydrophone array, targeting explicit object shape reconstruction at the scale relevant to dolphin echolocation research.

5.3.2 Comparison with Dolphin Behavioral Performance

The multiple-choice evaluation protocol adopted in this work was designed to enable a comparison with the dolphin behavioral experiments of Wei et al. [34], in which the dolphin subject Ginsan performed EV-MTS trials with the same target types (SQ and FF) used in the real data experiment. Because only two target shapes are available in the real dataset, the model’s evaluation is limited to a 2-choice protocol, whereas the dolphin performed 4-alternative matching trials, achieving 92.5% accuracy with air-filled PVC targets (chance level: 25%). The model achieves 84% on the 2-choice task (chance level: 50%).

A direct numerical comparison between these figures is not meaningful: the

4-choice task is substantially more demanding than the 2-choice task, and the dolphin’s alternative set includes target types beyond SQ and FF, whereas the model’s 2-choice evaluation is restricted to distinguishing between these two shapes only. The two evaluation settings also differ in other important respects: the dolphin matches against physical 3D objects presented visually in air, whereas the model matches against 2D occupancy maps using the Dice similarity metric; the dolphin’s decision process integrates echoic information with visual perception and memory, whereas the model performs a purely computational comparison; and the dolphin must contend with the full complexity of the live experimental environment.

Despite these differences, both results are well above their respective chance levels, confirming that the echo waveforms in this experimental setup carry sufficient target-discriminative information for reliable shape-based matching. The model’s performance validates the computational feasibility of extracting this information through learned processing, complementing the behavioral evidence that dolphin neural processing accomplishes a similar feat.

On the synthetic data, the model achieves 93% (2-choice) and 81% (4-choice) accuracy on a substantially different task involving diverse, previously unseen target shapes. While a direct comparison with dolphin performance is not applicable in this setting, these results establish a baseline for the shape discrimination ability of a learned acoustic reconstruction system and demonstrate that the approach generalizes beyond the two-target regime of the real data experiment.

5.4 Limitations

Several limitations of the current work should be acknowledged:

Resolution. The spatial resolution of the reconstruction is limited to a 10×10 pixel grid, which captures only the coarse spatial structure of the target. Fine-grained features such as thin protrusions, small holes, or detailed boundary contours cannot be resolved at this resolution. This coarseness is a deliberate design choice that balances shape distinguishability against the solvability of the inverse problem: increasing the grid resolution would increase the number of unknowns relative to the information content of the echo signals, requiring either more hydrophone channels,

higher signal-to-noise ratio, wider angular diversity between source positions, shorter source-to-target distances, or stronger learned priors to maintain reconstruction quality.

Dimensionality. The reconstruction is limited to two dimensions. The real targets are three-dimensional objects constructed from PVC pipes, and the 2D occupancy map represents only a planar projection of the target’s cross-sectional structure. This simplification ignores depth-dependent scattering effects that may carry additional shape information. As noted in the problem formulation (Section 3.1), the extension to 3D is conceptually straightforward, since depth information can be inferred from travel time measurements, but recovering lateral structure from echoes remains the more fundamental challenge that the 2D formulation addresses.

Forward Model Simplicity. The acoustic simulator employs a point scatterer model with single-bounce reflections and geometric spreading loss. This simplified forward model neglects physical phenomena including diffraction, multiple scattering, shadow effects, and frequency-dependent scattering cross-sections that are present in real acoustic environments. The gap between the simulated and real acoustic physics means that a model trained on synthetic data cannot be directly deployed on real recordings without retraining, limiting the practical utility of the synthetic training pipeline in its current form.

Limited real target diversity. The real data experiment uses only two target shapes, which is insufficient to evaluate the model’s ability to reconstruct novel targets from real echolocation signals. While the pixel masking strategy elicits spatially varying per-pixel predictions, the two-target setting cannot distinguish between genuine per-pixel occupancy estimation and a learned spatial discriminability map between the two training targets. Evaluating true reconstruction capability on real data would require experiments with a substantially larger and more varied set of target geometries.

5.5 Future Directions

The results and limitations of this work suggest several directions for future research.

Extending the reconstruction from 2D to 3D is a natural next step. The current formulation fixes the target at a known depth plane; a volumetric extension would predict occupancy over a 3D voxel grid, using travel time information to resolve the depth dimension in addition to the lateral structure recovered in the present work. The multi-view aggregation mechanism already present in the architecture is well-suited to this extension, as observations from different source positions provide complementary depth perspectives.

A particularly promising direction is the incorporation of physics-informed constraints into the learning framework. The current approach is purely data-driven: the network learns the inverse mapping entirely from training examples without explicit knowledge of the forward acoustic model. As discussed by Kamyab et al. [18], embedding a differentiable forward model into the training loop can provide a powerful form of regularization for learned inverse problems. In the present context, this would involve computing the predicted echo waveforms from the model’s reconstructed occupancy map using the acoustic propagation model (Equation 3.4), and penalizing discrepancies between these predicted echoes and the observed input. Such a physics-informed loss would encourage the model to produce reconstructions that are not only visually plausible but also physically consistent with the observed acoustic data, potentially improving both accuracy and generalization.

Improving the realism of the synthetic data pipeline offers a complementary path toward better real-data performance. The current simulator uses a simplified point scatterer model that does not account for multipath reflections, frequency-dependent scattering, or the acoustic properties of the experimental enclosure. Incorporating these effects into the simulator would reduce the domain gap between synthetic and real data, potentially enabling synthetic-to-real transfer learning in which a model pre-trained on diverse synthetic targets is fine-tuned on the limited real data available. This transfer learning approach would allow the model to leverage the shape diversity of the synthetic training set while adapting to the acoustic characteristics of the real experimental environment.

CHAPTER 5. DISCUSSION

Finally, expanding the real data evaluation to include a larger and more diverse set of target shapes would provide a more rigorous test of the model’s reconstruction capability on physical echolocation signals. The current 2-target setup is sufficient to demonstrate that target-discriminative information is present in the real echo data, but evaluating generalization to novel shapes requires targets that the model has not encountered during training. Future experiments could be designed with this computational evaluation in mind, using a larger target library that supports both the dolphin’s behavioral task and the model’s reconstruction assessment.

Chapter 6

Conclusion

This dissertation investigated whether a learned model can reconstruct the shape of an underwater target directly from multi-channel echo waveforms, drawing inspiration from the shape perception abilities demonstrated by echolocating dolphins. We formulated the problem as an acoustic inverse problem, mapping raw waveforms recorded by a 16-channel hydrophone array to a pixel-wise binary occupancy map of the target on a 2D grid. To solve this inverse problem, we developed an end-to-end deep learning pipeline comprising a physics-based synthetic data generator, a CNN+MLP network architecture with per-channel encoding, cross-sensor fusion, and multi-view aggregation, and a multiple-choice evaluation protocol modeled after the dolphin EV-MTS behavioral paradigm. We evaluated the system on both synthetic data with diverse unseen target shapes and real data from dolphin echolocation recordings collected at Ocean Park Hong Kong.

The synthetic and real experiments address different aspects of the problem and together provide a more complete picture than either would alone. On synthetic data, where diverse target shapes can be generated with known ground truth, the model achieves 93% accuracy on 2-choice and 81% on 4-choice matching trials, both well above chance. These results are obtained on targets drawn from the test set that the model has never encountered during training, demonstrating that the network learns a generalizable inverse mapping from echoes to spatial structure rather than memorizing target-specific associations. The reconstructed occupancy maps recover the dominant spatial features of each target, including overall silhouette and spatial extent, at a level of fidelity sufficient to discriminate the correct target from distractors in the majority of trials. This establishes that multi-channel echo waveforms encode sufficient geometric information to support

pixel-wise shape reconstruction. On real data, the limited number of available targets (two) precludes evaluation of true reconstruction on novel shapes. When trained with a pixel masking strategy that forces independent per-pixel predictions, the model achieves 84% accuracy on the 2-choice task while producing spatially structured probability maps rather than fixed templates. The spatial pattern of the predictions, with higher confidence in regions shared between the two targets and lower confidence in target-specific regions, provides evidence that the echo waveforms carry spatially localized information beyond global target identity. However, whether this per-pixel behavior reflects true occupancy estimation or a spatial discriminability map between the two training targets cannot be resolved without a larger target library. The synthetic experiment establishes that learned shape reconstruction from echo waveforms is achievable in principle; the real data experiment confirms that the spatially localized acoustic information required for such reconstruction survives physical propagation conditions; the missing piece is evaluating generalization to diverse real targets.

As a first attempt at learned acoustic shape reconstruction from echolocation signals, this work is necessarily limited in spatial resolution, dimensionality, forward model fidelity, and the diversity of real targets available for evaluation. Nevertheless, the results demonstrate that the end-to-end pipeline from raw echo waveforms to explicit shape prediction is feasible, and that it produces reconstructions of sufficient quality to perform shape-based matching on targets the model has never encountered. No prior work has produced direct, pixel-wise shape predictions from echo waveforms in this setting; the closest comparable approach, sparsity-aware array processing, yields spatial reflectivity maps that require human interpretation and do not generalize to novel target geometries. By establishing this baseline, the present work provides a foundation on which subsequent improvements in simulator realism, reconstruction resolution, and real-data diversity can build incrementally toward a practical learned reconstruction system for underwater acoustic sensing.

Bibliography

- [1] W. W. L. Au, “The Sonar of Dolphins”, New York, NY: Springer New York, 1993.
- [2] W. W. L. Au and D. A. Pawloski, “Cylinder wall thickness difference discrimination by an echolocating atlantic bottlenose dolphin”, *Journal of Comparative Physiology A*, vol. 170, no. 1, pp. 41–47, Jan. 1992.
- [3] W. W. L. Au and C. W. Turl, “Material composition discrimination of cylinders at different aspect angles by an echolocating dolphin”, *The Journal of the Acoustical Society of America*, vol. 89, no. 5, pp. 2448–2451, May 1991.
- [4] A. Baggeroer, W. Kuperman, and P. Mikhalevsky, “An overview of matched field methods in ocean acoustics”, *IEEE Journal of Oceanic Engineering*, vol. 18, no. 4, pp. 401–424, Oct. 1993.
- [5] A. Brunetto, S. Hornauer, S. X. Yu, and F. Moutarde, “The audio-visual batvision dataset for research on sight and sound”, in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2023, pp. 1–8. arXiv: **2303.07257** [cs].
- [6] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, “3d-r2n2: A unified approach for single and multi-view 3d object reconstruction”, in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham: Springer International Publishing, 2016, pp. 628–644.
- [7] J. H. Christensen, S. Hornauer, and S. X. Yu, “Batvision with gcc-phat features for better sound to vision predictions”, *ArXiv*, Jun. 2020.
- [8] J. H. Christensen, S. Hornauer, and S. Yu, “Batvision: Learning to see 3d spatial layout with two ears”, Mar. 2020. arXiv: **1912.07011** [cs].

BIBLIOGRAPHY

- [9] D. Colton and R. Kress, “Ill-posed problems”, in *Inverse Acoustic and Electromagnetic Scattering Theory*, D. Colton and R. Kress, Eds., Cham: Springer International Publishing, 2019, pp. 111–136.
- [10] C. M. DeLong, W. W. L. Au, D. W. Lemonds, H. E. Harley, and H. L. Roitblat, “Acoustic features of objects matched by an echolocating bottlenose dolphin”, *The Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1867–1879, Mar. 2006.
- [11] C. M. DeLong, W. W. L. Au, and S. A. Stamper, “Echo features used by human listeners to discriminate among objects that vary in material or wall thickness: Implications for echolocating dolphins”, *The Journal of the Acoustical Society of America*, vol. 121, no. 1, pp. 605–617, Jan. 2007.
- [12] M. Fink, “Time reversal of ultrasonic fields. i. basic principles”, *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 39, no. 5, pp. 555–566, Sep. 1992.
- [13] R. Gao, C. Chen, Z. Al-Halah, C. Schissler, and K. Grauman, “Visualechoes: Spatial image representation learning through echolocation”, in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., Cham: Springer International Publishing, 2020, pp. 658–676.
- [14] R. E. Hansen, “Introduction to synthetic aperture sonar”, in *Sonar Systems*, IntechOpen, Sep. 2011.
- [15] L. M. Herman, A. A. Pack, and M. Hoffmann-Kuhnt, “Seeing through sound: Dolphins (*tursiops truncatus*) perceive the spatial structure of objects through echolocation”, *Journal of Comparative Psychology*, vol. 112, no. 3, pp. 292–305, 1998.
- [16] M. Hoffmann-Kuhnt, A. Ho, H. Vishnu, M. A. Chitre, E. Matrai, and A. Tse, “Buzz that rod! acoustic analysis and patterns of the echolocation signals of a blindfolded bottlenose dolphin performing a horizontal and vertical angular resolution task”, in *Society for Marine Mammalogy Biennial Conference*, San Francisco, CA, USA, Dec. 2015.

BIBLIOGRAPHY

- [17] P. Hou, X. Li, M. Hoffmann-Kuhnt, M. A. Chitre, and M. Henz, “Decoding dolphin echolocation: Individual and material-based variability in click patterns”, in *OCEANS 2025 - Great Lakes*, Sep. 2025, pp. 1–5.
- [18] S. Kamyab, Z. Azimifar, R. Sabzi, and P. Fieguth, “Deep learning methods for inverse problems”, *PeerJ Computer Science*, vol. 8, e951, May 2022.
- [19] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition”, *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
- [20] Y. Lin, S. Feng, J. Theiler, Y. Chen, U. Villa, J. Rao, J. Greenhall, C. Pantea, M. A. Anastasio, and B. Wohlberg, “Physics and deep learning in computational wave imaging”, Oct. 2024. arXiv: **2410.08329** [cs].
- [21] D. B. Lindell, G. Wetzstein, and V. Koltun, “Acoustic non-line-of-sight imaging”, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 6773–6782.
- [22] D. Ma, Y. Li, T. Ma, and A. M. Pascoal, “The state of the art in key technologies for autonomous underwater vehicles: A review”, *Engineering*, Aug. 2025.
- [23] S. M. Mousavi and G. C. Beroza, “Deep-learning seismology”, *Science*, vol. 377, no. 6607, eabm4470, Aug. 2022.
- [24] J. Ning, F. Han, and J. Zou, “A direct sampling method and its integration with deep learning for inverse scattering problems with phaseless data”, *SIAM Journal on Scientific Computing*, vol. 47, no. 2, pp. C343–C368, Apr. 2025.
- [25] G. Ongie, A. Jalal, C. A. Metzler, R. G. Baraniuk, A. G. Dimakis, and R. Willett, “Deep learning techniques for inverse problems in imaging”, *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 39–56, May 2020.
- [26] A. A. Pack, L. M. Herman, M. Hoffmann-Kuhnt, and B. K. Branstetter, “The object behind the echo: Dolphins (*tursiops truncatus*) perceive object shape globally through echolocation”, *Behavioural Processes*, vol. 58, no. 1, pp. 1–26, May 2002.

BIBLIOGRAPHY

- [27] A. A. Pack and L. M. Herman, “Sensory integration in the bottlenosed dolphin: Immediate recognition of complex shapes across the senses of echolocation and vision”, *The Journal of the Acoustical Society of America*, vol. 98, no. 2, pp. 722–733, Aug. 1995.
- [28] H. L. Roitblat, W. W. L. Au, P. E. Nachtigall, R. Shizumura, and G. Moons, “Sonar recognition of targets embedded in sediment”, *Neural Networks, Automatic Target Recognition*, vol. 8, no. 7, pp. 1263–1273, Jan. 1995.
- [29] N. Sakurai and M. Tomonaga, “Hearing "number"?: Relative quantity judgments through the echolocation by bottlenose dolphins (*tursiops truncatus*)”, *JOURNAL OF COMPARATIVE PSYCHOLOGY*, vol. 138, no. 4, pp. 214–231, Nov. 2024.
- [30] K. Sun, W. Cui, and C. Chen, “Review of underwater sensing technologies and applications”, *Sensors (Basel, Switzerland)*, vol. 21, no. 23, p. 7849, Nov. 2021.
- [31] J. Sutton, “Underwater acoustic imaging”, *Proceedings of the IEEE*, vol. 67, no. 4, pp. 554–566, Apr. 1979.
- [32] B. Van Veen and K. Buckley, “Beamforming: A versatile approach to spatial filtering”, *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.
- [33] H. Vishnu, M. Hoffmann-Kuhnt, M. Chitre, A. Ho, and E. Matrai, “A dolphin-inspired compact sonar for underwater acoustic imaging”, *Communications Engineering*, vol. 1, no. 1, pp. 1–11, Jun. 2022.
- [34] C. Wei, M. Hoffmann-Kuhnt, W. W. L. Au, A. Z. H. Ho, E. Matrai, W. Feng, D. R. Ketten, and Y. Zhang, “Possible limitations of dolphin echolocation: A simulation study based on a cross-modal matching experiment”, *Scientific Reports*, vol. 11, no. 1, p. 6689, Mar. 2021.
- [35] R. Zaheer, Q. V. Phung, I. Ahmad, A. Aziz, D. Habibi, Y. Rong, and W. K. Hasan, “A review on underwater beamforming: Techniques, challenges, and future directions”, *Journal of Ocean Engineering and Science*, vol. 11, no. 1, pp. 58–77, Feb. 2026.