Towards Generalisable Inverse Modelling in Near-Infrared Diffuse Optical Tomography: A Two-Stage CNN-Transformer with Latent Alignment

MSc Dissertation

Max Hart

School of Computer Science

College of Engineering and Physical Sciences

University of Birmingham

2024-25

Abstract

Acknowledgements

Abbreviations

ACB	Apple Banana Carrot

Contents

Abstract ii

Acknowledgements iii

Abbreviations iv

List of Figures viii

List of Tables ix

1 Introduction 1

1.1 Background and Motivation 1

1.2 Near-Infrared Diffuse Optical Tomography (NIR-DOT) 1

1.3 Challenges in NIR-DOT Reconstruction 1

1.4 Research Objectives and Contributions 1

1.5 Dissertation Structure 2

2 Literature Review 3

2.1 Physics of NIR Light Propagation 3

2.2 Classical NIR-DOT Reconstruction Methods 3

2.3 Deep Learning for Medical Image Reconstruction 3

2.4 CNN Approaches in NIR-DOT 3

2.5 Transformers in Medical Imaging 4

2.6 Hybrid Architectures and Multi-Stage Training 4

2.7 Generalisation Challenges in DOT and Inverse Problems 4

2.8 Research Gap and Opportunity 4

3 Synthetic Phantom Data Generation 5

3.1 Physics-Based Forward Modelling 5

3.2 Geometric Phantom Construction 5

3.3 Optical Property Assignment 5

3.4 Surface Extraction and Probe Placement 5
3.5 Noise Model 6
3.6 Dataset Composition and Preprocessing 6

4 Hybrid CNN-Transformer Architecture 7
4.1 Architectural Overview 7
4.2 Stage 1: CNN Autoencoder Design 7
4.2.1 Encoder Architecture 7
4.2.2 Decoder Architecture 7
4.2.3 Residual Connections and Latent Space 7
4.3 Stage 2: Measurement Embedding and Transformer 8
4.3.1 Spatially-Aware Embedding 8
4.3.2 Transformer Encoder Design 8
4.3.3 Latent Alignment Strategy 8
4.4 Integration Strategy 8

5 Two-Stage Training Methodology 9
5.1 Training Strategy Overview 9
5.2 Stage 1: CNN Autoencoder Pre-Training 9
5.2.1 Loss Function and Optimisation 9
5.2.2 Hyperparameter and Setup 9
5.3 Stage 2: Transformer Enhancement Training 9
5.3.1 Frozen Decoder Approach 10
5.3.2 Latent Alignment Objective 10
5.3.3 Scheduler and Optimisation Strategy 10
5.4 Data Augmentation and Undersampling Strategy 10
5.5 Implementation Details 10

6 Experimental Results and Analysis 11
6.1 Experimental Setup 11
6.1.1 Dataset Preparation 11
6.1.2 Evaluation Metrics 11
6.1.3 Baseline Methods 11
6.2 Stage 1 Results: CNN Autoencoder Performance 11
6.3 Stage 2 Results: Transformer Enhancement 11
6.4 Comparative Analysis 12
6.5 Visualisation and Interpretation 12

7 Discussion 13
7.1 Key Findings and Insights 13
7.2 Clinical Implications of Generalisable DOT Models 13
7.3 Limitations of Current Approach 13
7.4 Computational Efficiency Considerations 13

8 Conclusion and Future Work 14
8.1 Summary of Contributions . . . 14
8.2 Future Research Directions . . . 14
8.3 Potential Clinical Applications . . . 14

A Implementation Details 16

B Additional Experimental Results 17

C Mathematical Derivations 18

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Background and Motivation

DOT uses NIR light to infer tissue absorption (μ_a) and reduced scattering (μ_s'), but mapping sparse boundary measurements to 3D volumes is severely ill-posed. This work motivates a modern, data-driven approach that retains physical plausibility while achieving fast, reliable reconstructions suitable for clinical workflows. We position the problem within quantitative NIR-DOT and argue for architectures that learn robust, geometry-aware representations.

1.2 Near-Infrared Diffuse Optical Tomography (NIR-DOT)

We briefly review photon migration in highly scattering media and the diffusion approximation used in practice for NIR-DOT. The forward model maps tissue optical properties to detector measurements (amplitude and phase), whereas the inverse task estimates volumetric μ_a and μ_s' from those measurements. The modality offers non-ionising, cost-effective, functional contrast compared with MRI/CT.

1.3 Challenges in NIR-DOT Reconstruction

The inverse problem is underdetermined, unstable to noise, and sensitive to modelling mismatches and probe geometry. Limited source–detector pairs and exponential attenuation exacerbate ambiguity when reconstructing 64^3 volumes. Robust solutions must integrate data-driven learning with constraints consistent with photon transport physics.

1.4 Research Objectives and Contributions

We propose a two-stage CNN–Transformer pipeline with latent alignment that reconstructs 3D optical property maps from frequency-domain measurements. Key contributions include: a physics-aware synthetic data pipeline, a 3D CNN autoencoder for spatial features, a transformer encoder for measurement sequences with spatial embeddings, and a teacher–student latent alignment objective enabling decoder reuse. We evaluate in raw units with clinically meaningful metrics.

1.5 Dissertation Structure

The thesis proceeds from background and related work to data generation, model architecture, training methodology, and results, before discussing limitations and future directions. Appendices compile implementation details, additional results, and mathematical derivations. This mirrors the experimental pipeline from simulation to evaluation.

CHAPTER 2

Literature Review

2.1 Physics of NIR Light Propagation

Summarises the diffusion approximation to the radiative transfer equation and boundary conditions used in tissue. Highlights how modulation frequency enables separation of absorption and scattering through amplitude and phase. Notes typical parameter ranges and assumptions.

2.2 Classical NIR-DOT Reconstruction Methods

Covers iterative FEM/Jacobian-based approaches with Tikhonov/TV regularisation and their computational burden. Discusses sensitivity to priors and geometry and the trade-off between speed and quantitative accuracy. Motivates learning-based accelerations.

2.3 Deep Learning for Medical Image Reconstruction

Reviews supervised and self-supervised paradigms, unrolled networks, and autoencoders for inverse problems. Emphasises data fidelity terms versus learned priors and the risk of hallucination. Positions physics-aware pipelines as a middle ground.

2.4 CNN Approaches in NIR-DOT

Summarises voxel-space CNN decoders and encoder–decoders trained on simulated data. Strengths: spatial locality and parameter efficiency; limitations: handling variable probe geometry and long-range dependencies. Notes typical losses and evaluation metrics.

2.5 Transformers in Medical Imaging

Outlines attention mechanisms for long-range dependency modelling in sequences and volumes. Discusses tokenisation strategies, positional encodings, and computational considerations. Motivates transformers for processing measurement sequences.

2.6 Hybrid Architectures and Multi-Stage Training

Surveys CNN–Transformer hybrids and teacher–student schemes used to stabilise training and reuse pretrained modules. Two-stage protocols can disentangle spatial representation learning from measurement encoding. Highlights applicability to NIR-DOT.

2.7 Generalisation Challenges in DOT and Inverse Problems

Describes geometry shift, noise shift, and simulator-to-real gaps that degrade performance. Argues for explicit spatial embeddings and randomised simulators to reduce shortcut learning. Frames ”generalisation across probe layouts” as a core objective.

2.8 Research Gap and Opportunity

Existing DL-DOT often assumes fixed geometries or end-to-end training that conflates roles. We identify a gap for a path-agnostic model combining a CNN spatial prior with a transformer measurement encoder, trained via latent alignment. This sets up our method.

CHAPTER 3

Synthetic Phantom Data Generation

3.1 Physics-Based Forward Modelling

Describes frequency-domain diffusion forward solves (e.g., 140 MHz) producing amplitude and phase per source–detector pair. Notes tetrahedral meshing and unit conventions. Establishes the mapping from μ_a, μ_s' volumes to measurements.

3.2 Geometric Phantom Construction

Details $64 \times 64 \times 64$ voxel volumes with ellipsoidal tissues and randomly placed tumour inclusions. Random rotations and shape parameters reduce bias and encourage invariance. Controls ensure inclusions are well within tissue boundaries.

3.3 Optical Property Assignment

Specifies physiologically plausible ranges for μ_a and μ_s' and scaling for tumours relative to background. Clarifies refractive index assumptions if fixed. Ensures values remain within clinical bounds during synthesis.

3.4 Surface Extraction and Probe Placement

Explains surface-aware placement of 50 sources and 20 detectors per source with SDS in 10–40 mm. Encodes each measurement with [log-amplitude, phase, src_xyz, det_xyz]. Clarifies pairing logic and avoidance of overlaps.

3.5 Noise Model

States additive noise levels for amplitude/phase (e.g., relative Gaussian on log-amp; small absolute Gaussian on phase). Justifies where noise is injected (pre/post transforms) and why these levels are realistic. Notes any fixed seeds for repeatability.

3.6 Dataset Composition and Preprocessing

Summarises dataset size, splits, and standardisation (per-channel z-score for volumes and features). Undersampling: always select exactly 256 from 1000 measurements per phantom. Mentions file formats and loader behaviour.

CHAPTER 4

Hybrid CNN-Transformer Architecture

4.1 Architectural Overview

Presents the two-stage pipeline: Stage 1 3D CNN autoencoder learns a 256-D latent; Stage 2 transformer encodes 256 measurement tokens with spatial awareness and maps to the same latent, feeding the frozen decoder. High-level flow from measurements to reconstruction.

4.2 Stage 1: CNN Autoencoder Design

The encoder uses progressive downsampling with residual blocks to compress 2-channel volumes to a 256-D latent; the decoder upsamples back to 64^3 outputs. The design balances capacity (7 M params) and efficiency for stable pretraining.

4.2.1 Encoder Architecture

Lists channel progression (e.g., 16→32→64→128→256), kernels/strides, and normalisation/activation choices. Notes global pooling or bottleneck projection to reach 256-D.

4.2.2 Decoder Architecture

Describes transposed-conv (or upsample+conv) stages reversing the encoder to recover μ_a and μ_s'. Mentions final activation/constraints if any.

4.2.3 Residual Connections and Latent Space

Explains residual paths for stable gradients and the rationale for a compact 256-D latent that later serves as the target for Stage 2. Notes parameter count and memory footprint.

4.3 Stage 2: Measurement Embedding and Transformer

The measurement branch embeds [log-amp, phase] while a spatial branch embeds [src_xyz, det_xyz]; fused tokens form a 256-D sequence. An L-layer, H-head transformer with dropout processes the sequence before pooling.

4.3.1 Spatially-Aware Embedding

Clarifies separate MLPs for signals and positions and how they are fused (concat + MLP). Emphasises geometry awareness as key to path-agnostic generalisation.

4.3.2 Transformer Encoder Design

States d_{model}, number of layers/heads, MLP ratio, positional encoding, and pooling (e.g., global average or learned queries). Notes typical head dimension.

4.3.3 Latent Alignment Strategy

The student transformer latent (256-D) is trained to match the frozen teacher latent via RMSE; the CNN decoder is reused, typically frozen. Optionally validate end-to-end recon during training.

4.4 Integration Strategy

Summarises tensor shapes from tokens → transformer → latent → decoder. Explains module freezing/unfreezing policy and how mismatches are handled (adapters if used/not used).

CHAPTER 5

Two-Stage Training Methodology

5.1 Training Strategy Overview

Outlines Stage 1 pretraining on ground-truth volumes followed by Stage 2 transformer training with the decoder frozen. Highlights benefits: stability, modularity, and reuse of a high-quality spatial prior.

5.2 Stage 1: CNN Autoencoder Pre-Training

Uses RMSE on standardised volumes with AdamW and OneCycleLR for rapid convergence. Mixed precision and sensible batch sizes accelerate training without sacrificing stability.

5.2.1 Loss Function and Optimisation

Defines the reconstruction loss, weight decay, gradient clipping, and AMP/bf16 choices. Notes early stopping criteria if used.

5.2.2 Hyperparameter and Setup

Records epochs, batch size, peak LR and schedule shape, seed, and checkpointing. Mentions channels-last/compile flags if applicable.

5.3 Stage 2: Transformer Enhancement Training

Loads the frozen decoder and trains only the transformer (and pooling head) to match the teacher latent; evaluation uses decoder outputs. Scheduler uses linear warm-up with cosine decay; EMA can smooth updates.

5.3.1 Frozen Decoder Approach

Justifies freezing to preserve Stage 1 decoder quality and reduce overfitting. Notes any optional fine-tuning at lower LR.

5.3.2 Latent Alignment Objective

RMSE between student and teacher latents drives learning; periodic full recon checks track end-to-end quality. Cosine similarity can be monitored but is not required.

5.3.3 Scheduler and Optimisation Strategy

Describes warm-up, cosine parameters, AdamW groups, and regularisation choices. Mentions EMA decay schedule if used.

5.4 Data Augmentation and Undersampling Strategy

Always subsample 256 of the 1000 measurements per phantom to set a fixed sequence length and provide augmentation. Rotation/randomisation come from phantom generation rather than image-space augmentations.

5.5 Implementation Details

Notes hardware (e.g., A100), mixed precision, dataloader workers, logging (e.g., W&B/TensorBoard), and run structure. Emphasises determinism settings for validation.

CHAPTER 6

Experimental Results and Analysis

6.1 Experimental Setup

Describes dataset splits, number of phantoms, and pre/post-processing used during evaluation. Lists key hyperparameters for both stages and any fixed seeds.

6.1.1 Dataset Preparation

Specifies training/validation/test proportions and how measurement subsampling is handled at eval time. Notes standardisation fitted on train only.

6.1.2 Evaluation Metrics

RMSE (per-channel and total), Dice for inclusion overlap, and contrast ratio in raw units. Optionally report latent RMSE for Stage 2 training curves.

6.1.3 Baseline Methods

Compares Stage 2 against Stage 1 reconstructions to quantify gains. Mentions any classic iterative baseline if available; otherwise focuses on within-pipeline comparison.

6.2 Stage 1 Results: CNN Autoencoder Performance

Reports convergence behaviour, best checkpoints, and qualitative slices. Typically sharper μ_a than μ_s' owing to modality physics; note typical artefacts.

6.3 Stage 2 Results: Transformer Enhancement

Presents improvement over Stage 1 with the latent alignment strategy. Shows validation curves and sample reconstructions demonstrating enhanced global coherence.

6.4 Comparative Analysis

Summarises gains in metrics and discusses where improvements concentrate (e.g., boundaries, global structure). Notes trade-offs in runtime/memory.

6.5 Visualisation and Interpretation

Includes representative axial/coronal/sagittal slices and, if available, attention summaries or latent distributions. Focuses on interpretability and failure modes.

CHAPTER 7

Discussion

7.1 Key Findings and Insights

The hybrid pipeline achieves coherent 3D reconstructions with a compact latent space, and Stage 2 consistently outperforms Stage 1. Spatially-aware tokenisation is central to geometry generalisation.

7.2 Clinical Implications of Generalisable DOT Models

Path-agnostic inference reduces per-device retraining costs and better matches handheld/bedside scenarios. Real-time potential arises from a single forward pass once trained.

7.3 Limitations of Current Approach

Training uses only synthetic data; μ_s' remains harder to recover sharply; memory footprint is non-trivial for 64^3 volumes. Generalisation to real hardware requires calibration and domain adaptation.

7.4 Computational Efficiency Considerations

Notes throughput, mixed-precision benefits, and the cost of forward solves during data generation. Outlines profiling hotspots and possible compression strategies.

CHAPTER 8

Conclusion and Future Work

8.1 Summary of Contributions

Reiterates the two-stage design, latent alignment, and physics-aware simulator as an integrated pipeline. Emphasises quantitative evaluation in raw units and improved reconstruction fidelity.

8.2 Future Research Directions

Small, concrete items: (i) tissue-patch context integration for local anatomy (not implemented here), (ii) dynamic sequence undersampling beyond fixed 256/1000, (iii) multi-wavelength DOT and real-data validation. Briefly note uncertainty estimation.

8.3 Potential Clinical Applications

Envisions portable screening and intra-operative monitoring with rapid reconstructions. Highlights requirements for regulatory, calibration, and robustness testing.

Bibliography

APPENDIX A

Implementation Details

Centralises exact hyperparameters, environment, and command-line invocations for reproducibility. Include config snippets and file layout.

APPENDIX B

Additional Experimental Results

Provides extra slices, metric tables, and training curves that support claims but are too detailed for the main text.

APPENDIX C

Mathematical Derivations

Includes diffusion equation forms, unit conversions, and any loss/metric derivations used for completeness.