UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

University of Modena and Reggio Emilia
XXXVIII cycle of the International Doctorate School in
"Information and Communication Technologies (ICT)"

Scaling AI for Oral and Dental Image Analysis

Luca Lumetti

Supervisor: Prof. Costantino Grana
Co-Supervisor: Prof. Federico Bolelli
PhD Course Coordinator: Prof. Luigi Rovati

Review committee:

Prof. Giuseppe Serra, Università degli Studi di Udine
Prof. Marco Bertini, Università degli Studi di Firenze

To my family, mentors, and colleagues
who made this journey possible.

Tesi di dottorato finanziata dall'Unione Europea - Next Generation EU, Missione 4, componente 2 "Dalla Ricerca all'Impresa" - Investimento 3.3 "Introduzione di dottorati innovativi che rispondono ai fabbisogni di innovazione delle imprese e promuovono l'assunzione dei ricercatori dalle imprese".

Abstract

Cone-beam computed tomography (CBCT) is central to contemporary dental and maxillofacial care, yet progress in automated analysis has been constrained by the paucity of large, publicly available voxel-level datasets. This thesis addressed that bottleneck by creating an open, extensible ecosystem, combining datasets, annotation tooling, algorithmic advances, and by demonstrating how those elements interacted cyclically to accelerate research and clinical translation.

The Maxillo dataset was the first of its kind, providing 91 densely annotated volumes plus 256 sparsely annotated scans for the annotation of the Inferior Alveolar Canal. The ToothFairy series built on this foundation: the first ToothFairy release provided 443 CBCT volumes (153 with dense 3D annotations); ToothFairy2 expanded to 480 volumes with 42 semantic classes for comprehensive maxillofacial segmentation; with ToothFairy3 it further grew the corpus to 532 volumes and 77 classes. Complementing volumetric CBCTs, the Bits2Bites dataset delivered 200 registered intra-oral scan pairs with multi-label occlusion annotations. All resources were openly released to enable reproducible benchmarking and downstream development.

To scale annotations without sacrificing clinical fidelity, I developed semi-automated annotation tools and a rigorous quality-control pipeline that combined model priors with expert refinement. Crucially, dataset creation, tooling, and model development proceeded cyclically: additional data enabled stronger models; stronger models powered faster, higher-quality annotation tools; and improved tools in turn produced larger, better datasets, forming the core intellectual contribution of this work.

On this data foundation, I advanced volumetric segmentation methods: modules inspired by transformer architectures that explicitly encoded spatial patch relationships to preserve voxel detail while aggregating long-range context, and adaptations of State-Space Models (Mamba) for efficient, high-accuracy 3D segmentation.

Finally, I introduced U-Net Transplant, a model-merging framework that proposed novel techniques to update and specialize clinical models without full retraining, reducing redeployment cost, storage, and privacy exposure.

Collectively, this ecosystem delivered the largest open CBCT benchmark for maxillofacial segmentation to date and a matched set of methods and tools that materially improved accuracy, efficiency, and lifecycle management of clinical AI, enabling faster, safer, and more reproducible dental AI research and deployment.

Contents

List of Abbreviations

ACDC	Automated Cardiac Diagnosis Challenge
Adam	Adaptive Moment Estimation (optimizer)
BraTS	Brain Tumor Segmentation Challenge
CBCT	Cone-Beam Computed Tomography
CNN	Convolutional Neural Network
CT	Computed Tomography
DDPM	Denoising Diffusion Probabilistic Model
DICOM	Digital Imaging and Communications in Medicine
DSC	Dice Similarity Coefficient
FDI	Fédération Dentaire Internationale (World Dental Federation)
FID	Fréchet Inception Distance
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
HD95	95th Percentile Hausdorff Distance
HIPAA	Health Insurance Portability and Accountability Act
HiPPO	High-order Polynomial Projection Operators
IAC	Inferior Alveolar Canal
IACAT	Inferior Alveolar Canal Annotation Tool
IOS	Intra-Oral Scanner / Intra-Oral Scan
ISO	International Organization for Standardization
ITK-SNAP	Insight Segmentation and Registration Toolkit - SNAP
KL	Kullback-Leibler (divergence)
MAE	Masked Autoencoder
MICCAI	Medical Image Computing and Computer-Assisted Intervention
MRI	Magnetic Resonance Imaging
MSD	Medical Segmentation Decathlon
NIfTI	Neuroimaging Informatics Technology Initiative
ODIN	Oral and Dental Image aNalysis (workshop)
OPG	Orthopantomography
RAS	Right-Anterior-Superior (coordinate system)
SGD	Stochastic Gradient Descent
STL	Stereolithography (file format)
TUS	Testicular Ultrasound
U-Net	Encoder-decoder architecture for segmentation
USCL	Ultrasound Contrastive Learning
ViT	Vision Transformer
WHO	World Health Organization

1. Introduction

1.1 Motivation and Clinical Context

Artificial intelligence is transforming healthcare delivery across virtually every medical specialty, yet dentistry and maxillofacial medicine have lagged behind despite representing one of the most common healthcare interactions worldwide. Over 3.5 billion people suffer from oral diseases globally, and dental conditions rank among the most prevalent non-communicable diseases affecting humanity [90]. AI-powered diagnostic and treatment planning systems hold tremendous potential to address this burden: they could democratize access to specialist-level expertise in underserved regions, reduce diagnostic errors and inter-observer variability, accelerate clinical workflows in overburdened practices, and enable personalized treatment strategies tailored to individual patient anatomy.

The scope of potential AI applications in dentistry and orthodontics is vast. Caries detection from bitewing and periapical radiographs represents the highest-volume diagnostic task, with billions of images acquired annually worldwide. Orthodontic treatment planning requires integration of cephalometric analysis, dental arch assessment, and facial aesthetics evaluation—a complex multi-factorial decision process ideally suited for AI augmentation. Periodontal disease staging benefits from automated bone loss quantification. Implant surgery demands precise preoperative planning to avoid critical anatomical structures. Oral pathology screening could leverage AI to flag suspicious lesions for specialist review. Each of these applications presents distinct computational challenges, yet all share a common prerequisite: annotated training data at scale.

Dental imaging encompasses a diverse array of modalities, each presenting unique opportunities and challenges for automated analysis. Two-dimensional radiographs—including bitewing, periapical, panoramic (OPG), and cephalometric images—remain the most ubiquitous imaging modality, with established diagnostic protocols and vast historical archives. Cone-Beam Computed Tomography (CBCT) provides three-dimensional volumetric imaging essential for implantology, orthognathic surgery, and endodontic treatment planning [28], enabling visualization of complex anatomical structures with submillimeter spatial resolution [82, 83]. Intra-Oral Scanners capture high-resolution 3D surface mod-

els of dental arches, revolutionizing orthodontic diagnosis, treatment monitoring, and digital workflows for prosthodontics. Clinical photographs document soft tissue conditions, facial aesthetics, and treatment outcomes.

Each modality requires specialized processing pipelines, domain-specific architectures, and dedicated training data. CBCT volumes demand memory-efficient 3D architectures. Intra-oral scans require surface-based geometric deep learning approaches that respect mesh topology. Two-dimensional radiographs and intra-oral photographs can exploit existing foundation models. Yet the potential for cross-modal integration represents perhaps the greatest untapped opportunity: combining CBCT's volumetric bone information with intra-oral scans' precise surface geometry could enable comprehensive treatment planning; integrating radiographic caries detection with clinical photographs could provide holistic diagnostic workflows; fusing cephalometric analysis with 3D surface models could revolutionize orthodontic treatment planning. Such multi-modal fusion architectures remain largely unexplored in dental AI.

Despite this promise, AI adoption in dentistry remains limited compared to fields like radiology and pathology. The fundamental barrier is infrastructural rather than algorithmic: while modern deep learning architectures have demonstrated superhuman performance on many visual recognition tasks, dental AI lacks the large-scale, publicly available, expertly annotated datasets that have catalyzed progress in other domains. ImageNet transformed computer vision; the BraTS challenge advanced brain tumor segmentation; the Medical Segmentation Decathlon established benchmarks across ten anatomies—yet dental imaging until recently lacked comparable resources.

The clinical stakes for accurate automated analysis are substantial across dental subspecialties. In implantology, precise localization of the Inferior Alveolar Canal (IAC)—the bony channel housing the neurovascular bundle that innervates the lower lip, chin, and teeth [45, 71]—is mandatory to avoid iatrogenic nerve injury during implant placement, a complication affecting patient quality of life. In orthodontics, accurate assessment of dental occlusion, skeletal relationships, and growth patterns guides treatment selection and outcome prediction. In oral and maxillofacial surgery, comprehensive segmentation of teeth, bone, sinuses, and adjacent structures enables patient-specific surgical guides and prosthetic design. Manual analysis of these structures across thousands of patients is time-consuming, subjective, and prone to inter-observer variability, creating bottlenecks that limit both clinical throughput and research progress.

This thesis addresses the data scarcity problem through a systematic program of dataset creation, tool development, and algorithmic innovation, demonstrating how these elements interact synergistically to accelerate the entire field of dental AI.

1.2 The Synergistic Cycle of Medical AI Development

The central thesis of this dissertation is that progress in medical AI follows a **synergistic cycle** wherein advances in annotation tools, datasets, and models mutually reinforce one another. This cyclical relationship, illustrated in Fig. 1.1, captures the fundamental dynamic that has driven the research presented in this thesis.

Tools Enable Datasets. The creation of large-scale medical imaging datasets is fundamentally bottlenecked by annotation. Unlike natural images where crowdsourcing can rapidly collect labels from non-experts, medical image annotation requires trained specialists whose time is expensive and limited. A single volumetric scan contains hundreds of slices, and voxel-level annotation of complex anatomical structures can require hours of expert effort. Specialized annotation software addresses this challenge by providing domain-optimized interfaces, semi-automated assistance, and quality control mechanisms that dramatically reduce the time and expertise required to create high-quality ground truth. The development of IACAT (Inferior Alveolar Canal Annotation Tool) exemplifies this principle: by providing spline-based annotation workflows tailored to tubular anatomical structures, IACAT reduced annotation time from hours to minutes per volume, making the creation of ToothFairy feasible.

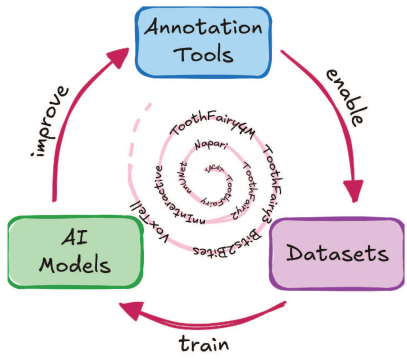


Figure 1.1: The cycle of dental AI development. Tools enable datasets, datasets train AI models, and models improve tools. This cycle is the foundation of the research presented in this thesis.

Datasets Enable Models. The availability of large, publicly accessible datasets with high-quality annotations is the foundation upon which modern deep learning systems are built. Performance improvements are inextricably linked to data quality and quantity, as neural networks learn statistical regularities from training examples. Equally important, public datasets enable reproducible research and fair algorithmic comparison—without standardized benchmarks, claims of superiority remain unverifiable and the field cannot make cumulative progress. The ToothFairy series of datasets enabled training and rigorous evaluation of segmentation architectures, while also empowering other research groups to develop complementary annotations of new structures.

Models Enable Tools. Trained neural networks complete the cycle by enhan-

cing annotation workflows. Rather than creating segmentations from scratch, human experts can review and refine model predictions, shifting their role from laborious manual delineation to efficient quality assurance. This human-in-the-loop paradigm simultaneously accelerates annotation and maintains clinical accuracy, enabling the iterative expansion of datasets and the development of more capable models that further accelerate annotation workflows. The open release of tools, datasets, and models fostered international collaboration, culminating in the ODIN Workshop series (Oral and Dental Image aNalysis, `odin-workshops.org`) and driving continued expansion of ToothFairy4M to support additional modalities and worldwide multi-center collaboration. This synergy between tools, datasets, and models has been the foundation of the research presented in this thesis.

1.3 Research Challenges

The development of effective AI systems for dental and medical image analysis faces several fundamental challenges that this thesis addresses.

1.3.1 Data Scarcity

Medical imaging datasets are notoriously difficult to obtain and annotate. Privacy regulations such as HIPAA and GDPR impose strict requirements on patient data handling, while ethical review processes create lengthy approval timelines. The need for expert annotators—whose time is expensive and limited—further constrains dataset creation. These barriers affect all medical imaging domains, but dental imaging faces particular challenges: the diversity of imaging modalities (CBCT, panoramic radiographs, intra-oral scans, periapical X-rays), the fine-grained nature of dental anatomy (32 individual teeth, each with distinct structures), and the relative youth of the field compared to domains like brain or cardiac imaging.

The consequences of data scarcity are profound. Prior to this work, research on structures such as the inferior alveolar canal relied exclusively on proprietary datasets, making claims of superiority unverifiable and preventing cumulative scientific progress. Public benchmarks that have transformed fields like natural image classification and autonomous driving remained absent in dental AI. The cost of expert annotation limited most studies to dozens rather than hundreds of samples, far below the scale required for robust deep learning. Perhaps most critically, the absence of multi-modal datasets—combining CBCT with intra-oral scans, or linking imaging to clinical outcomes—has prevented research into the integrative analyses that hold greatest clinical promise.

1.3.2 Architectural Limitations

The dominant paradigms in medical image segmentation each present significant limitations when applied to volumetric dental imaging. Convolutional neural networks (CNNs), exemplified by the U-Net architecture [77] and its self-configuring variant nnU-Net [44], excel at capturing local patterns through hierarchical feature extraction. However, their inherently local receptive fields struggle with long-range dependencies, which are critical for structures like the inferior alveolar canal that course through the entire mandible and require global anatomical context for accurate segmentation.

Vision transformers [23] address the locality limitation through self-attention mechanisms that enable each spatial position to attend to all others. Yet the quadratic computational complexity of self-attention, $\mathcal{O}(n^2)$ with respect to sequence length n, makes transformers impractical for high-resolution 3D volumes where n can exceed millions of voxels. Additionally, transformers require substantially more training data than CNNs to achieve competitive performance, exacerbating the data scarcity problem.

A critical gap in current architectures is support for multi-modal data. Clinical decision-making integrates information from multiple sources—CBCT for bone and nerve visualization, intra-oral scans for surface geometry, radiographs for caries detection—yet existing architectures process each modality independently, missing opportunities for synergistic learning.

1.3.3 Training and Deployment Efficiency

Beyond accuracy, clinical deployment of medical AI faces practical constraints that academic research often overlooks. Training models from scratch for each new task and dataset is computationally expensive. Once deployed, models cannot easily be updated or adapted to evolving clinical needs—adding a new structure class or accommodating a new scanner typically requires complete retraining. Maintaining separate models for each clinical task creates significant storage and deployment overhead in clinical IT systems, while the need for rapid adaptation to changing requirements conflicts with the lengthy retraining cycles of conventional approaches. Privacy regulations may prevent centralized data collection across institutions, necessitating distributed or modular approaches that current training paradigms do not support well. Model merging offers an elegant solution: combining multiple task-specific models into a single unified model through arithmetic operations alone, without additional training, enabling modular development where new capabilities can be added or removed incrementally. These efficiency challenges motivate the model merging and transfer learning strategies explored in this thesis.

1.4 Contributions

The contributions presented in this dissertation are described below, a compact list with urls are also available in appendix A.5.

1.4.1 Annotation Tools

Development of IACAT (Inferior Alveolar Canal Annotation Tool), a specialized software for efficient 3D annotation of tubular anatomical structures in CBCT volumes. IACAT employs Catmull-Rom splines for centerline tracing and an α-shape algorithm for volumetric mask generation, reducing annotation time by an order of magnitude compared to slice-by-slice approaches.

Development of ToothFairy4M, a web-based platform for collaborative annotation of multi-modal medical data with integrated AI models that enable model-assisted annotation workflows. ToothFairy4M has served as a cornerstone for international collaboration, facilitating the synergy between tools, datasets, and models and significantly accelerating the entire medical AI development cycle.

1.4.2 Datasets and Benchmarks

Creation and public release of multiple large-scale datasets:

- **ToothFairy**: 443 CBCT volumes with sparse (2D) and dense (3D) annotations of the inferior alveolar canal—the largest public dataset of its kind;
- **ToothFairy2**: Expansion to 480 volumes, each providing 3D annotations for 42 anatomical classes including mandible, maxilla, individual teeth, sinuses, and canals;
- **ToothFairy3**: Further expansion to 532 volumes with 77 anatomical classes (including pulp cavities, root canals, incisive nerves, and lingual foramen), introducing both computational efficiency evaluation and an interactive segmentation task for the IAC;
- **Bits2Bites**: 200 intra-oral scan pairs with multi-label occlusion annotations;
- **TesticulUS**: Testicular ultrasound classification synthetic dataset.

1.4.3 Architectural Innovation

Development of novel architectures addressing the fundamental limitations of patch-based 3D medical image segmentation:
Memory-Augmented Transformer for IAC Segmentation:

- Introduction of absolute positional encoding to overcome patch-based learning limitations;

- Memory-augmented transformer encoder for enhanced contextual understanding;
- State-of-the-art performance on the ToothFairy IAC segmentation benchmark.

Mamba-Based Architectures for General 3D Segmentation:

- Identification of the "initial hidden state problem" in applying sequential models to volumetric data;
- Proposal of bidirectional and multidirectional processing strategies to address spatial context limitations;
- Achievement of state-of-the-art performance on multiple benchmarks (BrainTumor, Synapse, ACDC) while maintaining linear computational complexity.

1.4.4 Training Efficiency

Introduction of model merging strategies for 3D medical segmentation:

- First application of task vector arithmetic to volumetric medical imaging;
- Discovery that "stable" pretraining (encouraging flat loss landscape minima) enables effective model merging with +18 Dice points improvement;
- Framework for modular development and deployment of clinical AI systems.

1.4.5 Open Science

All datasets, code, pretrained models, and challenge infrastructure have been publicly released to foster reproducibility and accelerate community research.

1.5 Thesis Organization

This thesis is organized as follows:

Chap. 2 provides essential background on medical image segmentation, covering traditional approaches, deep learning architectures (CNNs, transformers, state-space models), and relevant training strategies.

Chap. 3 introduces IACAT and ToothFairy4M, the annotation tools developed to enable efficient dataset creation. This chapter demonstrates the first link in the synergistic cycle: how better tools enable better datasets.

Chap. 4 presents the ToothFairy, ToothFairy2, and ToothFairy3 datasets, describing data collection, annotation protocols, challenge organization, and benchmark results. This chapter establishes the second link: how datasets enable model development.

Chap. 5 presents two complementary architectural innovations: a memory-augmented transformer with absolute positional encoding for IAC segmentation (PosPadUNet3D), and Mamba-based architectures for general 3D segmentation. This chapter demonstrates the third link: how datasets enable novel architectural innovations.

Chap. 6 explores model merging strategies, showing how stable pretraining enables effective task vector arithmetic for modular model development.

Chap. 7 extends the developed methods to diverse modalities, including testicular ultrasound classification and occlusal analysis from intra-oral scans, demonstrating broad applicability.

Chap. 8 synthesizes the contributions, analyzing how they collectively demonstrate the synergistic cycle and discussing broader implications.

Chap. 9 summarizes key findings and outlines directions for future research.

2. Background and Related Work

This chapter provides the theoretical foundations and reviews prior work relevant to the contributions of this thesis. We begin with an overview of medical image segmentation, then discuss the evolution of deep learning architectures from convolutional networks through transformers to state-space models. Finally, we review transfer learning and model merging techniques.

2.1 Medical Image Segmentation

Medical image segmentation is the task of partitioning a medical image into semantically meaningful regions, typically corresponding to anatomical structures or pathological entities. Unlike natural image segmentation, medical imaging presents unique challenges: low contrast between tissues, noise from acquisition physics, significant inter-patient anatomical variability, and the need for voxel-level precision in clinical applications. Moreover, oftentimes medical images can be large 3D volumes (CT, MRI) or high-resolution 2D images (WSI), making both the annotation and the segmentation task very challenging.

2.1.1 Traditional Methods

Prior to the deep learning era, medical image segmentation relied on classical computer vision techniques:

Thresholding and Region Growing exploit intensity differences between structures. While effective for high-contrast boundaries (*e.g.*, bone versus soft tissue in CT), these methods fail when intensity distributions overlap.

Statistical Shape Models (SSM) [49] encode prior knowledge about anatomical shape variation. An SSM represents a structure as a mean shape plus a linear combination of principal modes of variation learned from a training set. While powerful for constrained anatomies, SSMs struggle with pathological deformations or structures with high shape variability.

Graph-based Methods formulate segmentation as energy minimization over a graph where nodes represent pixels/voxels and edges encode spatial relationships. Graph cuts and random walks provide globally optimal solutions under certain energy formulations but require careful design of unary and pairwise potentials.

Atlas-based Segmentation propagates labels from annotated template images to new subjects via image registration. Multi-atlas approaches [56] improve robustness by combining propagated labels from multiple templates. However, registration quality degrades for structures with high anatomical variability.

To the extent of annotating 3D data, previous works have relied on the use of proprietary software such as Photoshop [73] and Invivo5 [89], which can be tedious, time-consuming, and not tailored for the specific task. Moreover, even when they propose a novel methodology to annotate such data, they do not release the source code of their implementation [53, 55, 46].

2.1.2 Evaluation Metrics

Quantitative evaluation of medical image segmentation is not trivial: different downstream goals (e.g., accurate volumetry, precise boundaries for intervention planning, or robust detection of small lesions) imply different notions of what constitutes a "good" segmentation. Recent work has therefore emphasized that metric choices should be *problem-aware* and aligned with the underlying domain interest, and that misleading conclusions can arise not only from suboptimal metric selection but also from how metrics are *applied and aggregated* across a dataset. The *Metrics Reloaded* framework explicitly addresses these issues by proposing a structured "problem fingerprint" (capturing, for example, whether boundary, volume, or center(line) accuracy matters most) and by cataloguing common pitfalls in metric selection and metric application.[1] [66]

In the context of voxel-wise 3D (semantic) segmentation, evaluation metrics are often grouped into *overlap-based* measures, which summarize volumetric agreement, and *boundary-based* (surface-distance) measures, which quantify spatial deviations between predicted and reference surfaces. Metrics Reloaded highlights that overlap-based metrics are shape/contour-unaware and can behave unintuitively for small structures or large size variability; it therefore recommends complementing overlap-based scores with boundary-based metrics whenever boundary quality is relevant. [66]

Dice Similarity Coefficient (DSC) is the most widely used overlap-based metric in medical image segmentation and measures volumetric overlap between prediction P and ground truth G:

$$\mathrm{DSC}(P, G) = \frac{2|P \cap G|}{|P| + |G|}. \tag{2.1}$$

[1] `https://metrics-reloaded.dkfz.de/`

DSC ranges from 0 (no overlap) to 1 (perfect agreement). In multi-class settings, DSC is typically computed per class and then averaged (e.g., macro-averaging) to avoid dominance of large structures.

95th Percentile Hausdorff Distance (HD95) is a boundary-based metric that measures surface discrepancy in physical units (e.g., millimeters when voxel spacing is taken into account). It summarizes the worst-case deviation while being more robust to single outliers than the maximum Hausdorff distance:

$$\mathrm{HD95}(P, G) = \max\left\{d_{95}(\partial P, \partial G),\ d_{95}(\partial G, \partial P)\right\}, \tag{2.2}$$

where $d_{95}(A, B)$ denotes the 95th percentile of distances from points on surface A to their nearest neighbors on surface B (and $\partial(\cdot)$ denotes the object boundary/surface). In practice, HD95 is particularly informative when small boundary shifts matter clinically, even if volumetric overlap remains high.

Beyond metric choice, *metric application* can substantially affect reported performance. Metrics Reloaded explicitly notes pitfalls such as inappropriate aggregation in the presence of hierarchical data (e.g., multiple volumes per patient or multi-center datasets), as well as edge cases like empty references or empty predictions that require consistent handling to avoid biased averages. [66]

In this thesis, we report DSC and HD95 throughout, combining an overlap-based measure of volumetric agreement with a boundary-based measure of spatial error. This pairing is common in medical image segmentation benchmarks and aligns with the general recommendation to use complementary overlap- and boundary-focused metrics when evaluating segmentation quality.

2.2 Medical Datasets

Large-scale medical image datasets have been pivotal in moving computer-aided diagnosis from hand-crafted feature engineering to representation learning, particularly through 2D benchmark collections that lowered barriers to entry and standardized evaluation. In radiography, chest X-ray corpora such as ChestX-ray14, CheXpert, and MIMIC-CXR enabled weakly supervised and multi-label classification from report-derived labels, highlighting the practical trade-off between cohort size and label noise, and motivating methods that explicitly model uncertainty, missingness, and domain shift across hospitals and acquisition devices [86, 43, 48]. In dermatology and ophthalmology, datasets like ISIC skin lesion archives and diabetic retinopathy fundus collections fostered strong baselines for 2D detection and grading while emphasizing the importance of careful curation, class imbalance handling, and patient-level splits to avoid overly optimistic performance [15, 33]. In digital pathology, gigapixel whole-slide imaging benchmarks (e.g., CAMELYON) pushed the community toward multiple-instance learning, tiling strategies, and scalable feature extraction pipelines that anticipate many of the computational concerns later encountered in volumetric learning [58]. Across these 2D settings, the maturation of data formats (DICOM at

acquisition and common research exports), annotation practices (image-level labels, bounding boxes, pixel masks), and evaluation protocols (AUC, sensitivity at fixed specificity, challenge leaderboards) created a methodological template for related work sections, while also exposing recurrent limitations such as inconsistent label definitions, limited external validity, and restricted access driven by privacy governance and licensing constraints.

In contrast, 3D medical image datasets center on volumetric modalities such as CT and MRI, where the clinically meaningful signal is distributed across slices and is confounded by heterogeneous voxel spacing, anisotropy, scanner-specific intensity characteristics, and varying fields of view; these properties strongly influence both preprocessing toolchains and architectural choices. Public resources and challenges—including TCIA-hosted collections, LIDC-IDRI for lung nodules, LUNA16 for detection, BraTS for brain tumor segmentation, and the Medical Segmentation Decathlon as a multi-task aggregation—have shaped common practice around converting DICOM series to research-friendly formats (often NIfTI), resampling to normalized spacing, intensity standardization (e.g., Hounsfield windowing for CT), and robust data augmentation to mitigate small cohort sizes and site effects [3, 79, 29, 2].

2.3 Deep Learning Revolution

Large-scale datasets and the introduction of fully convolutional networks (FCNs) enabled the deep learning revolution in medical image segmentation.

U-Net [77] established the dominant paradigm for medical image segmentation. Its encoder-decoder architecture with skip connections enables precise localization by combining high-level semantic features from the decoder with fine-grained spatial information from the encoder. The symmetric structure facilitates gradient flow during training and produces segmentation maps at the original input resolution.

3D Extensions. Volumetric medical images (*e.g.*, CT, MRI) naturally extend to 3D architectures. V-Net [68] introduced residual connections and the Dice loss for volumetric segmentation. 3D U-Net [13] directly extended the U-Net architecture to 3D, enabling end-to-end learning from volumetric data.

nnU-Net [44] represents the current gold standard for medical image segmentation. Rather than proposing a novel architecture, nnU-Net provides a self-configuring framework that automatically adapts preprocessing, augmentation, architecture, and postprocessing to each dataset. Its success demonstrates that careful engineering often outweighs architectural novelty.

2.3.1 Convolutional Neural Networks

Convolutional neural networks form the backbone of most medical image segmentation systems. Their success stems from three key properties:

Local Connectivity. Each neuron connects only to a local region of the input, controlled by the kernel size. This inductive bias exploits the spatial structure of images—nearby pixels tend to be more related than distant ones.

Weight Sharing. The same kernel weights are applied at all spatial locations, dramatically reducing parameter count and enabling translation equivariance.

Hierarchical Features. Stacking convolutional layers with pooling operations builds hierarchical representations—early layers capture edges and textures, while deeper layers encode increasingly abstract semantic concepts.

Receptive Field Limitations The effective receptive field of a CNN determines the spatial context available for each prediction. For a network with L layers using $k \times k$ kernels, the theoretical receptive field grows linearly as $L(k-1)+1$. Pooling and strided convolutions expand the receptive field more rapidly.

However, the *effective* receptive field—the region that actually influences predictions—is substantially smaller than the theoretical maximum [63]. Information from distant regions is progressively attenuated through successive layers. This limitation is problematic for tasks requiring global context, such as:

- Segmenting large structures spanning the entire image;
- Distinguishing similar-appearing structures based on anatomical context;
- Maintaining topological consistency over extended regions.

Various architectural modifications have been proposed to expand the effective receptive field:

Dilated Convolutions insert gaps between kernel elements, expanding the receptive field without increasing parameters. DeepLab [11] popularized atrous spatial pyramid pooling (ASPP), which applies dilated convolutions at multiple rates to capture multi-scale context.

Attention Mechanisms enable feature maps to adaptively weight contributions from different spatial locations. Squeeze-and-excitation networks [40] introduced channel attention; subsequent work extended this to spatial and joint attention.

Non-local Operations [88] compute responses at each position as a weighted sum over all positions, directly modeling long-range dependencies. However, the quadratic complexity limits applicability to low-resolution feature maps.

2.3.2 Transformer Architectures

Transformers [85] revolutionized natural language processing through the self-attention mechanism, which models pairwise relationships between all elements in a sequence. Their adaptation to vision tasks [23] has achieved remarkable success.

Self-Attention Mechanism

Given an input sequence $X \in \mathbb{R}^{n \times d}$ with n tokens of dimension d, self-attention computes:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{2.3}$$

where $Q = XW_Q$, $K = XW_K$, and $V = XW_V$ are linear projections. Each output token is a weighted combination of value vectors, with weights determined by query-key compatibility.

Multi-head attention runs h parallel attention operations with different projections, concatenating and projecting the results:

$$\text{MultiHead}(X) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_O \tag{2.4}$$

Vision Transformers

Vision Transformer (ViT) [23] adapts transformers to images by treating non-overlapping patches as tokens. An image $I \in \mathbb{R}^{H \times W \times C}$ is divided into $N = HW/P^2$ patches of size $P \times P$, which are linearly projected to token embeddings. Positional embeddings encode spatial location.

For 3D medical volumes, the patch-based approach extends naturally: a volume of size $H \times W \times D$ yields $N = HWD/P^3$ tokens. However, the token count grows cubically with resolution, making self-attention prohibitively expensive.

Transformers for Medical Segmentation

Several architectures adapt transformers for medical image segmentation:

TransUNet [10] combines a CNN encoder with a transformer at the bottleneck, using the CNN for local feature extraction and the transformer for global context modeling.

UNETR [36] uses a pure transformer encoder with skip connections to a CNN decoder, enabling direct application to 3D volumes through patch embedding.

Swin-UNETR [35] employs shifted window attention to reduce complexity from quadratic to linear in image size, enabling efficient processing of high-resolution 3D volumes.

nnFormer [98] interleaves local and global attention operations within the nnU-Net framework, achieving strong performance across multiple datasets.

Limitations

Despite their success, transformers face significant limitations in medical imaging:

Computational Complexity. Self-attention has $\mathcal{O}(n^2)$ complexity in sequence length. For a 512^3 volume with 8^3 patches, $n = 262144$ tokens, requiring substantial memory for attention matrices while having a degradation of the resolution due to the patching which is not ideal in medical scenarios.

Data Hunger. Transformers lack the inductive biases of CNNs (locality, translation equivariance) and require larger datasets to achieve competitive performance. Medical imaging datasets are typically orders of magnitude smaller than natural image benchmarks.

Position Encoding. Fixed or learned positional embeddings may not generalize well to volumes of different sizes or resolutions than those seen during training.

2.4 State-Space Models and Mamba

State-space models (SSMs) offer an alternative paradigm for sequence modeling that combines the long-range modeling capability of attention with the computational efficiency of recurrent networks.

Continuous State-Space Models

A continuous-time state-space model maps an input signal $x(t) \in \mathbb{R}$ to an output $y(t) \in \mathbb{R}$ through a latent state $h(t) \in \mathbb{R}^N$:

$$h'(t) = Ah(t) + Bx(t) \tag{2.5}$$

$$y(t) = Ch(t) \tag{2.6}$$

where $A \in \mathbb{R}^{N \times N}$ is the state transition matrix, $B \in \mathbb{R}^{N \times 1}$ is the input projection, and $C \in \mathbb{R}^{1 \times N}$ is the output projection.

Discretization

For application to discrete sequences, SSMs are discretized using a step size Δ and a discretization rule (typically zero-order hold):

$$\overline{A} = \exp(\Delta A) \tag{2.7}$$

$$\overline{B} = (\Delta A)^{-1}(\exp(\Delta A) - I)\Delta B \tag{2.8}$$

yielding the recurrence:

$$h_{t+1} = \overline{A}h_t + \overline{B}x_t \tag{2.9}$$

$$y_t = Ch_t \tag{2.10}$$

This formulation enables $\mathcal{O}(n)$ sequential processing while the convolutional view enables $\mathcal{O}(n \log n)$ parallel training.

HiPPO Initialization

Standard SSMs suffer from gradient instability over long sequences. The HiPPO (High-order Polynomial Projection Operators) theory [32] provides principled initialization for A that enables the hidden state to efficiently memorize input history. This initialization is crucial for modeling long-range dependencies.

Mamba: Selective State-Space Models

Mamba [31] introduces a *selection mechanism* that makes parameters B, C, and Δ input-dependent:

$$B = \text{Linear}_N(x) \tag{2.11}$$

$$C = \text{Linear}_N(x) \tag{2.12}$$

$$\Delta = \text{softplus}(\text{Linear}_1(x) + \text{Parameter}) \tag{2.13}$$

This selection mechanism enables content-based filtering—the model can learn to selectively propagate or forget information based on the input, unlike fixed linear SSMs. Combined with an efficient parallel scan implementation (S6), Mamba achieves linear complexity in sequence length, content-based reasoning through input-dependent dynamics, and long-range modeling via HiPPO-initialized state matrices.

A key innovation of Mamba is its hardware-aware algorithm design that achieves substantial speedups on modern GPUs. The selective scan operation cannot be parallelized as easily as convolutions or attention, but Mamba introduces a work-efficient parallel scan algorithm that exploits GPU memory hierarchy. By keeping the SSM state in fast SRAM rather than slower HBM (High Bandwidth Memory), Mamba avoids the memory bottleneck that limits many sequence models. This kernel fusion approach yields up to 3× faster training compared to optimized transformer implementations at sequence lengths of 2K–8K tokens, with the advantage growing for longer sequences due to the linear versus quadratic complexity difference.

Mamba-2 [19] further improves efficiency by establishing a theoretical connection between selective state-space models and attention mechanisms through the Structured State Space Duality (SSD) framework. This insight enables even more efficient implementations that leverage optimized matrix multiplication routines while maintaining the linear complexity of SSMs, achieving 2–8× speedups over the original Mamba implementation.

Mamba for Vision

Shortly after the release of Mamba, many researchers tried to replace vision transformers with mamba layers, achieving promising results [99, 64, 100]. Adapting Mamba to vision requires addressing the fundamental mismatch between

sequential models and spatial data, which was often ignored in the initially released papers. A 3D volume $V \in \mathbb{R}^{H \times W \times D \times C}$ must be flattened to a sequence before processing. The choice of flattening order—*e.g.*, (h, w, d), (h, d, w), or (d, w, h)—affects which spatial relationships are captured efficiently.

This "directionality problem" motivates the architectural innovations presented in Chap. 5.

2.4.1 Pretraining and Transfer Learning

Transfer learning leverages knowledge acquired from data-rich source domains to improve learning on data-scarce target tasks. Rather than training models from random initialization, transfer learning initializes networks with representations learned elsewhere, providing a strong inductive bias that accelerates convergence and improves generalization. This paradigm has become essential in medical imaging, where it often means the difference between a model that learns meaningful representations and one that merely overfits to limited training data.

Supervised Pretraining

The dominant transfer learning approach trains models on large-scale labeled datasets before fine-tuning on medical imaging tasks. ImageNet [20], containing over 14 million natural images across thousands of categories, has served as the primary source domain since the deep learning revolution began.

The success of ImageNet pretraining initially appeared surprising given the substantial domain gap between natural photographs and medical scans. Natural images contain color, texture, and object semantics fundamentally different from the grayscale intensity distributions and anatomical structures in medical imaging. CT and MRI modalities encode physical tissue properties through carefully calibrated acquisition protocols, producing images with characteristics entirely unlike everyday photography.

Despite this gap, empirical evidence consistently demonstrates that ImageNet-pretrained features transfer effectively to medical tasks. The explanation lies in the hierarchical nature of deep networks: early layers learn general-purpose features—edge detectors, blob filters, color contrasts—that prove useful across domains. These low-level representations remain relevant even in medical imaging, where edges delineate anatomical boundaries and intensity gradients indicate tissue transitions. Deeper layers capture increasingly domain-specific semantics that may transfer less directly, but the strong initialization of early layers alone provides substantial benefit.

For 3D medical imaging, the standard approach inflates 2D ImageNet-pretrained filters to 3D by replicating weights along the depth dimension and averaging [34]. While crude, this initialization outperforms training from scratch by

providing structure in early layers while allowing volumetric features to develop through fine-tuning.

Self-Supervised Pretraining

Self-supervised learning offers an attractive alternative: learning representations from unlabeled data through carefully designed pretext tasks. Medical imaging provides abundant unlabeled scans—hospitals generate thousands of studies daily—while annotations remain scarce. Self-supervised methods exploit this unlabeled data to learn general-purpose representations transferable to downstream tasks.

Contrastive Learning learns representations by maximizing agreement between differently augmented views of the same image while minimizing agreement between views from different images. SimCLR [12] pioneered this approach in computer vision, using a simple framework: augment each image twice, project representations to a lower-dimensional space, and optimize a contrastive loss that pulls together embeddings from the same image while pushing apart embeddings from different images.

MoCo (Momentum Contrast) [37] improved efficiency by maintaining a queue of negative examples and using a momentum-updated encoder to produce consistent representations. This design enables contrastive learning with smaller batch sizes while maintaining a large pool of negative samples crucial for learning discriminative features.

One of the latest advancements in contrastive learning is the DINO method [7, 72, 80], today at its third version, which eliminates the need for negative samples altogether. DINO employs a teacher-student framework where the student network learns to match the teacher's output on different augmented views of the same image. The teacher is updated as an exponential moving average of the student, providing stable targets that guide representation learning.

Masked Image Modeling takes inspiration from masked language modeling in NLP [21]. MAE (Masked Autoencoders) [38] masks random patches of an input image and trains an encoder-decoder to reconstruct the missing content. The asymmetric design—a large encoder processes only visible patches while a small decoder reconstructs the full image—enables efficient training by reducing computation on masked tokens.

Reconstruction forces the encoder to learn semantically meaningful representations: successfully predicting missing anatomical structures requires understanding spatial relationships and tissue appearance patterns. Unlike contrastive methods that require careful negative sampling, masked modeling provides a straightforward self-supervised objective applicable to any dataset.

Domain-Specific Pretraining exploits structure unique to medical imaging. Rotation prediction, where models classify the rotation angle applied to an image, teaches orientation-aware features. Context restoration predicts missing regions given their surroundings, similar to masked modeling but with structured

masking patterns. Multi-view learning in volumetric data treats axial, sagittal, and coronal slices as different views of the same anatomy, applying contrastive or consistency losses across planes.

Models Genes [30] demonstrated large-scale medical self-supervised pretraining by training on 86 million unlabeled CT images using contrastive learning. The resulting representations transferred effectively across multiple anatomies and pathologies, suggesting that sufficient scale and appropriate pretext tasks can yield universal medical imaging features.

Model Merging

Model merging combines multiple trained models into a single model without additional training [42, 67, 81, 95]. The majority of the approaches leverages task vectors [42, 76], which represent modifications to a pre-trained model introduced during fine-tuning for a specific task. These vectors can be added to tune the model's functionalities. Given a pretrained model with parameters θ_{pre} and task-specific models with parameters θ_{task_i}, task vectors are defined as:

$$\tau_i = \theta_{\text{task}_i} - \theta_{\text{pre}} \tag{2.14}$$

Merged models can be constructed through arithmetic operations on task vectors:

$$\theta_{\text{merged}} = \theta_{\text{pre}} + \sum_i w_i \tau_i \tag{2.15}$$

The literature explores various strategies for determining the best approach to merge different task vectors together [42, 94, 67, 27, 17], to reduce interference and maximize performance on all tasks, but few to none have explored the step prior to the merging: the pretraining phase that yields the initial model used as a base.

2.5 Summary

This chapter has established the technical foundations for the contributions of this thesis:

- Tools for efficient and accurate annotation of 3D medical images are essential to create high-quality datasets, and existing open-source solutions are lacking;
- Large-scale, diverse datasets are critical for training robust medical image segmentation models, yet publicly available datasets remain limited in size and scope;
- Medical image segmentation has evolved from classical methods through CNNs, transformer-based and mamba-based architectures, but there is not yet a clear best choice;

- CNNs excel at local feature extraction but struggle with long-range dependencies, Transformers capture global context but suffer from quadratic complexity and large data requirements, State-space models (Mamba) offer linear complexity with content-based reasoning;
- Transfer learning and model merging enable efficient training with limited data and in a continual-learning setting, a common scenario in medical imaging.

The following chapters build on these foundations to present novel contributions in annotation tools, datasets, architectures, and training strategies.

3. Annotation Tools for Medical Imaging

This chapter presents IACAT (Inferior Alveolar Canal Annotation Tool), a specialized software developed to enable efficient annotation of tubular anatomical structures in CBCT volumes, and ToothFairy4M, a web-based platform for collaborative annotation of multi-modal medical data with integrated AI models. IACAT, IACAT 2.0, and ToothFairy4M are a perfect demonstration of the first and last links in the synergistic cycle: specialized tools dramatically accelerate dataset creation, enabling the large-scale annotation efforts that produced the ToothFairy datasets described in Chap. 4, and the integration of AI models in the annotation process enables model-assisted annotation workflows that further improves the annotation process.

3.1 Introduction

The creation of large-scale medical imaging datasets is fundamentally bottlenecked by annotation. Unlike natural images, where crowdsourcing platforms can rapidly collect labels from non-experts, medical image annotation requires trained specialists whose time is expensive and limited. A single CBCT volume may contain 400+ slices, and voxel-level annotation of complex structures can require hours of expert effort.

The Inferior Alveolar Canal (IAC) presents particular annotation challenges:

- **Tubular geometry**: The IAC is a thin, elongated structure that curves through the mandible, requiring annotation in three dimensions;
- **Low contrast**: The boundary between the canal and surrounding cancellous bone is often poorly defined, particularly in regions with trabecular bone;
- **Discontinuities**: Imaging noise and patient-specific anatomy can create apparent gaps in the canal;

- **Bilateral structures**: Each patient has left and right canals that must be annotated separately.

General-purpose annotation tools such as ITK-SNAP [96] or 3D Slicer [25], while powerful for many applications, are not optimized for these challenges. They typically require slice-by-slice annotation, which is both time-consuming and prone to inter-slice inconsistencies.

3.2 IACAT Design Philosophy

IACAT was designed around three core principles:

Efficiency over complexity: The tool should minimize the number of user interactions required while maintaining clinical accuracy. Rather than annotating every voxel, experts define key control points that are interpolated into dense annotations.

Anatomically-informed visualization: The user interface should present views that align with clinical reasoning. For the IAC, this means panoramic views of the dental arch and cross-sectional views perpendicular to the canal course.

Real-time feedback: Annotations should be visualized immediately in 3D, allowing experts to identify and correct errors during the annotation process rather than in post-hoc review.

3.3 Technical Implementation

3.3.1 Spline-Based Centerline Definition

The annotation workflow begins with defining the IAC centerline. Rather than requiring experts to manually identify the canal in each slice, IACAT uses a two-stage approach:

Automatic initialization: The tool generates a panoramic view from the CBCT volume (analogous to an orthopantomogram) and applies image processing techniques to propose an initial centerline estimate. This estimate serves as a starting point that experts can refine.

Control point refinement: The centerline is represented as a Catmull-Rom spline defined by a sequence of control points. Catmull-Rom splines pass through all control points while maintaining C^1 continuity, producing smooth curves that match the natural curvature of anatomical structures. Experts can add, remove, or relocate control points to correct the proposed centerline.

The Catmull-Rom spline through control points $P_0, P_1, \ldots, P_n$ is defined

piecewise between consecutive points. For the segment between P_i and P_{i+1}:

$$C(t) = \frac{1}{2}\begin{bmatrix}1 & t & t^2 & t^3\end{bmatrix}\begin{bmatrix}0 & 2 & 0 & 0\\ -1 & 0 & 1 & 0\\ 2 & -5 & 4 & -1\\ -1 & 3 & -3 & 1\end{bmatrix}\begin{bmatrix}P_{i-1}\\ P_i\\ P_{i+1}\\ P_{i+2}\end{bmatrix} \tag{3.1}$$

where $t \in [0, 1]$ parameterizes position along the segment.

3.3.2 Cross-Sectional View Annotation

Once the centerline is defined, IACAT generates Cross-Sectional Views (CSVs) perpendicular to the canal at regular intervals along the spline. These views present the IAC as a roughly circular cross-section, simplifying the annotation task. Within each CSV, experts annotate the canal boundary using a closed Catmull-Rom spline. The tool provides:

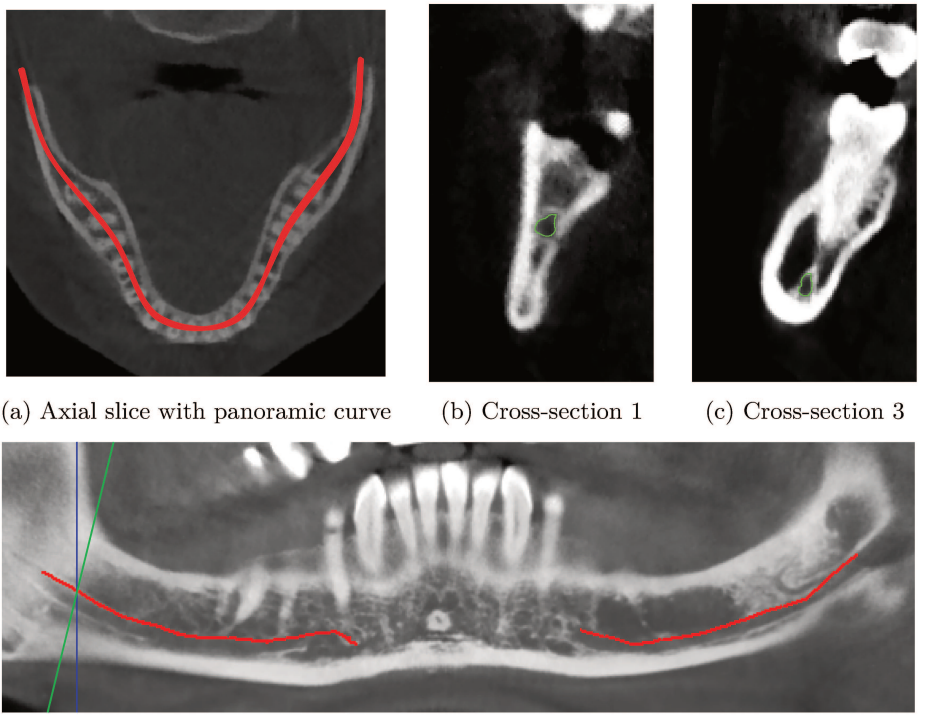
- Zoomed views centered on the expected canal location;
- Adjustable window/level for optimal contrast;
- Localized contrast enhancement to reveal boundaries in low-contrast regions.

3.3.3 Volumetric Mask Generation

The collection of 2D cross-sectional annotations must be converted to a dense 3D segmentation mask. IACAT employs the α-shape algorithm [24] for this purpose, providing a principled approach to surface reconstruction from point samples.

The α-shape is a generalization of the convex hull that captures shape at a specified scale. Formally, given a point set $S \subset \mathbb{R}^3$ and a parameter $\alpha > 0$, the α-shape is defined through the Delaunay triangulation. A simplex (edge, triangle, or tetrahedron) from the Delaunay triangulation belongs to the α-shape if its circumradius is at most $1/\alpha$. Equivalently, for each simplex, there exists an empty ball of radius $1/\alpha$ passing through all vertices of the simplex. The parameter α controls the level of detail: as $\alpha \to 0$, the α-shape approaches the convex hull; as $\alpha \to \infty$, it captures increasingly fine surface detail until eventually degenerating to the point set itself. These steps are illustrated in Fig. 3.2.

The volumetric mask generation proceeds through four steps. First, points are sampled uniformly along each cross-sectional contour at spacing determined by the desired output resolution. Second, these 2D contour points are transformed to 3D coordinates using the known position and orientation of each cross-sectional plane along the centerline spline. Third, the α-shape of the resulting 3D point cloud is computed, with α automatically selected based on the



(a) Axial slice with panoramic curve (b) Cross-section 1 (c) Cross-section 3

(d) Panoramic planes with orthogonal cross-sections

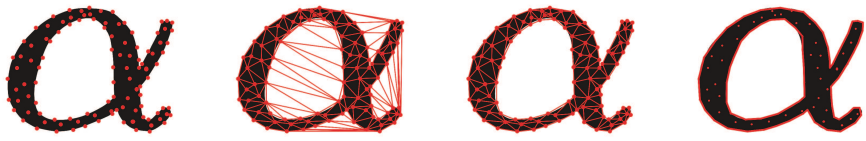
Figure 3.1: IACAT annotation workflow for efficient mandibular canal segmentation. **(a)** An axial slice from the CBCT volume showing the panoramic base curve (red line) crossing the mandible. **(d)** Panoramic plane visualization showing the complete IAC trajectory and orthogonal cross-section planes. **(b)**–**(c)** Cross-sectional views perpendicular to the IAC at different positions along the canal, with boundaries highlighted in green for annotation.

spacing between cross-sections to ensure watertight surfaces without excessive smoothing. Finally, the α-shape surface is voxelized to produce a binary segmentation mask aligned with the original CBCT volume.

3.4 IACAT 2.0: AI-Assisted Annotation

Based on experience from the initial ToothFairy annotation campaign, IACAT 2.0 incorporated several enhancements:

Model-assisted initialization: A trained PosPadUNet3D model [14] provides initial segmentation proposals for each CSV. Experts can accept, modify, or reject these proposals, substantially reducing annotation time for straightforward



(a) Point Set (b) Triangulation (c) α-complex (d) α-shape

Figure 3.2: The α-shape construction process. Starting from a point set (3.2a), the Delaunay triangulation (3.2b) is computed. Triangles with circumradius $\leq 1/\alpha$ form the α-complex (3.2c), whose boundary defines the α-shape (3.2d). This algorithm converts sparse cross-sectional annotations into watertight 3D surfaces.

Table 3.1: Annotation time comparison for IAC segmentation.

Method	**Time per Volume**	**Expert Interactions**
Slice-by-slice	2–4 hours	∼400 contours
IACAT 1.0	20–40 minutes	∼30 control points + 20 CSVs
IACAT 2.0	10–20 minutes	Refinement only

cases while maintaining full control.

Improved visualization: Enhanced contrast stretching algorithms reveal canal boundaries in regions where they are nearly invisible in raw intensity values. The tool applies localized histogram equalization while preserving global intensity relationships.

Quality control features: Automatic consistency checks flag potential annotation errors—*e.g.*, discontinuities between adjacent cross-sections, implausible variations in canal diameter, or deviations from expected anatomical course.

Multi-structure support: The annotation pipeline also integrated Napari [70] in the loop, an open-source multi-dimensional image viewer that provided flexible 3D editing capabilities for adding and correcting model predictions. This allowed annotators to efficiently handle additional structures beyond the IAC, such as teeth and mandible, directly within the same interface, enabling ToothFairy2 and ToothFairy3 datasets.

3.5 Annotation Efficiency

IACAT substantially improved annotation efficiency compared to slice-by-slice approaches:

These efficiency gains enabled the annotation of 153 densely-labeled volumes for ToothFairy up to the expansion to 77 classes in ToothFairy3, annotation efforts that would have been impractical with conventional tools.

3.6 Quality Validation

To assess annotation quality, two independent experts annotated a subset of 15 volumes. Inter-annotator agreement, measured by Dice coefficient between the two annotations, averaged 81%. This establishes an upper bound on algorithm performance—methods approaching this level achieve near-human accuracy.

The relatively modest inter-annotator agreement reflects the genuine difficulty of IAC annotation in challenging cases. Regions with low contrast, trabecular bone, or imaging artifacts may be interpreted differently by equally qualified experts. IACAT's annotation review workflow, where a second expert validates each annotation, helps mitigate but cannot eliminate this variability.

3.7 Impact: Enabling ToothFairy

IACAT's development was essential to the ToothFairy dataset. Without specialized tooling, the annotation of 443 volumes (153 with dense 3D labels) would have required thousands of hours of expert time, rendering the project infeasible.

More broadly, IACAT demonstrates a general principle: **investment in annotation infrastructure pays dividends throughout the research pipeline**. The time spent developing IACAT was recovered many times over through accelerated annotation, and the resulting datasets have enabled research contributions described in subsequent chapters.

This represents the first link in the synergistic cycle (Fig. 1.1): better tools enable better datasets, which in turn enable better models.

3.8 ToothFairy4M: A Multi-Modal Web Platform

Building on the success of IACAT, we developed ToothFairy4M, a comprehensive web-based platform that extends beyond CBCT annotation to integrate multiple imaging modalities and clinical data sources (Fig. 3.3). The platform addresses a fundamental limitation of single-modality approaches: clinical decision-making in orthodontics and maxillofacial surgery relies on synthesizing information from diverse sources, not isolated imaging studies.

ToothFairy4M supports six complementary data types per patient. Cone-beam computed tomography (CBCT) provides 3D volumetric information about bone structures, teeth, and canals. Orthopantomography (OPG) offers panoramic 2D visualization of the entire dental arch and surrounding structures. Cephalometric scans enable standardized lateral skull analysis for orthodontic treatment planning. intra-oral photographs capture clinical appearance including soft tissue conditions and aesthetics. intra-oral scans (IOS) provide high-resolution 3D surface models of dental arches for occlusion analysis. Finally,

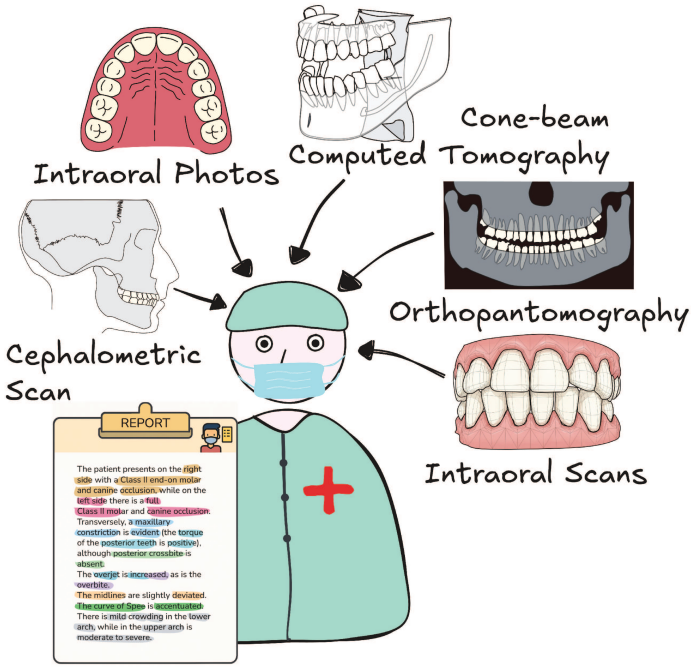


Figure 3.3: ToothFairy4M integrates multiple imaging modalities and clinical data per patient. The platform supports intra-oral photographs, cone-beam computed tomography (CBCT), orthopantomography (OPG), cephalometric scans, intra-oral scans (IOS), and structured textual reports with annotated clinical findings. This multi-modal approach enables comprehensive orthodontic and maxillofacial analysis.

structured textual reports contain clinical findings with annotated observations about occlusion, crowding, and treatment recommendations.

The platform integrates trained models from the ToothFairy series to provide automatic segmentation proposals, which clinicians can review and refine. This model-assisted workflow accelerates annotation while maintaining clinical accuracy, extending the IACAT philosophy to multiple modalities. Importantly, the multi-modal structure enables cross-modal learning: models trained on one modality can inform annotation in another, and multi-modal datasets enable research on correspondence between different imaging representations of the same patient.

ToothFairy4M has enabled international collaboration through the ODIN Workshop (Oral and Dental Image aNalysis), bringing together researchers and clinicians to advance automated dental image analysis. The platform continues to expand, incorporating additional data from collaborating institutions and

supporting new annotation tasks such as the Bits2Bites occlusion classification described in Chap. 7.

3.9 Summary

This chapter presented the annotation tools developed to enable large-scale medical imaging dataset creation. IACAT introduced spline-based annotation for the inferior alveolar canal, reducing expert interactions by an order of magnitude through cross-sectional view workflows aligned with anatomical structure and α-shape algorithms for robust volumetric mask generation. AI-assisted enhancements in IACAT 2.0 further accelerated annotation through model-based initialization.

ToothFairy4M, extended this foundation to a comprehensive multi-modal platform integrating CBCT, OPG, cephalometric scans, intra-oral photographs, intra-oral scans, and textual reports. This platform supports international collaboration and enables cross-modal research in dental image analysis. The platform is open-source and publicly available at `https://github.com/AImageLab-zip/ToothFairy4M`.

Together, these tools demonstrate how specialized annotation infrastructure breaks bottlenecks and accelerates the synergistic cycle of medical AI development. The datasets enabled by these tools are the subject of the next chapter.

IACAT is publicly available at: `https://github.com/AImageLab-zip/IAN_annotation_tool`

Dataset statistics and download information can be found at: `https://ditto.ing.unimore.it/dataset_view/`

4. Dataset Creation and Benchmarking

This chapter presents the three ToothFairy datasets, the largest public benchmarks for maxillofacial segmentation in CBCT volumes. These datasets, enabled by the tools described in Chap. 3, demonstrate the second link in the synergistic cycle: comprehensive datasets enable fair model comparison and drive architectural innovation.

4.1 Introduction

The development of deep learning algorithms for medical image segmentation is fundamentally constrained by data availability. While natural image datasets contain millions of annotated examples, medical imaging datasets typically number in the dozens to hundreds of samples. This scarcity is compounded by several factors:

- **Privacy regulations**: Patient data is protected by regulations such as HIPAA (US) and GDPR (EU), requiring ethical approval and anonymization for research use;
- **Annotation cost**: Expert annotation is expensive and time-consuming;
- **Institutional barriers**: Most datasets remain privately held within acquiring institutions, preventing reproducible research and fair algorithmic comparison.

Prior to this work, research on inferior alveolar canal segmentation relied exclusively on private datasets [46, 53, 54], making claims of superiority unverifiable and preventing cumulative scientific progress.

4.2 ToothFairy: IAC Segmentation Benchmark

4.2.1 Dataset Description

The ToothFairy dataset comprises 493 CBCT volumes:

- **Training set**: 443 volumes from the Affidea Center (Modena, Italy), acquired with a NewTom/NTVGiMK4 scanner (110 kV, 3 mA, 0.3 mm isotropic voxels);
- **Test set**: 50 volumes from Radboud University Medical Centre (Nijmegen, Netherlands), acquired with an i-CAT 3D system (0.4 mm voxels).

The training set includes two annotation types:

- **Sparse (2D) annotations**: All 443 volumes have the IAC marked on panoramic reconstructions, providing coarse localization;
- **Dense (3D) annotations**: 153 volumes have voxel-level ground truth generated using IACAT.

4.2.2 Demographics

The dataset includes patients aged 10–100 years, with 59% female. Age distributions differ between training (peaks at 20–30 and 60–70 years) and test sets (peak at 50–70 years), introducing realistic demographic shift that algorithms must handle.

4.2.3 Distribution Shift

The use of different scanners and acquisition sites for training and test sets introduces distribution shift. Analysis of intensity distributions reveals that training data exhibits different intensity ranges than test data. Successful algorithms must either:

- Apply robust preprocessing (*e.g.*, intensity clipping, z-score normalization);
- Use aggressive data augmentation during training;
- Learn features invariant to intensity distribution.

This intentional distribution shift ensures that benchmark results reflect generalization capability rather than overfitting to a single scanner or institution.

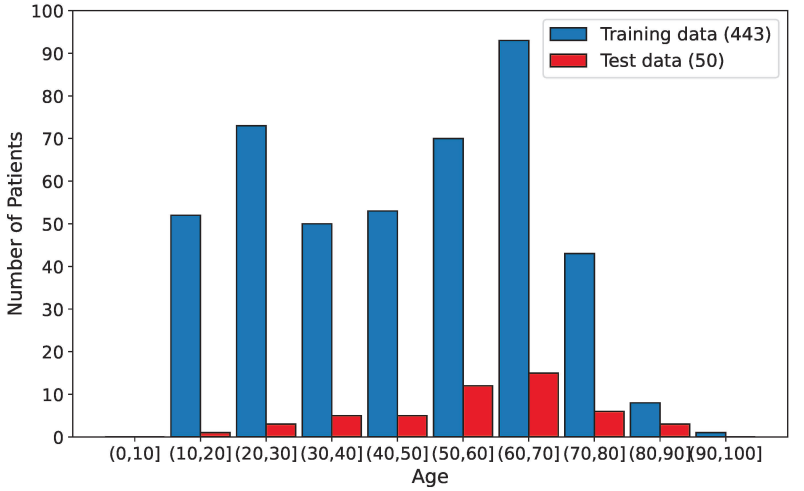


Figure 4.1: Demographic distribution of the ToothFairy dataset. The training set (Modena, Italy) and test set (Nijmegen, Netherlands) exhibit different age distributions, introducing realistic demographic shift that algorithms must handle for robust generalization.

4.2.4 The ToothFairy Challenge

ToothFairy was organized as a challenge at MICCAI 2023, attracting over 20 unique team submissions during the Final Test Phase.

Evaluation Protocol: Algorithms were evaluated using Dice Similarity Coefficient (DSC) and 95th percentile Hausdorff Distance (HD95), following recommendations from Maier-Hein et al. [65]. Final rankings were computed by averaging per-metric rankings across the test set.

Top Results: The winning method [59] achieved DSC 79.6% and HD95 4.49mm using a self-training approach with nnU-Net as the backbone. Key strategies among top performers included:

- nnU-Net-based architectures (used by 4/4 top teams);
- Semi-supervised learning to leverage sparse annotations;
- Strong data augmentation for domain generalization;
- Connected component analysis for postprocessing.

Table 4.1: Top results from the ToothFairy Challenge. Arrows indicate optimization direction: ↑ higher is better, ↓ lower is better.

Rank	Team	DSC (%) ↑	HD95 (mm) ↓
1	Liu *et al.*	79.6	4.49
2	Wang *et al.*	78.9	4.64
3	Wodzinski *et al.*	78.6	6.28
4	Kirchhoff *et al.*	78.4	5.59
Inter-annotator		81.0	3.21

4.2.5 Impact

ToothFairy established the first reproducible benchmark for IAC segmentation. The dataset has been downloaded over 1,000 times from 21 different countries and has been used in numerous subsequent publications. The challenge remains open for post-challenge submissions on the private test set, enabling ongoing fair benchmarking.

4.3 ToothFairy2: Multi-Structure Segmentation

4.3.1 Motivation

Clinical applications require segmentation of multiple anatomical structures, not just the IAC. Surgical planning for dental implants, for example, requires understanding the spatial relationships between teeth, alveolar bone, maxillary sinuses, and canals.

ToothFairy2 extends the original dataset to comprehensive multi-structure segmentation.

4.3.2 Dataset Description

ToothFairy2 comprises 480 CBCT volumes with annotations for 42 semantic classes:

- **Bones**: Mandible (lower jawbone), maxilla (upper jawbone);
- **Teeth**: 32 individual teeth following FDI [1] notation (8 per quadrant);
- **Canals**: Left and right inferior alveolar canals;
- **Sinuses**: Left and right maxillary sinuses;
- **Airways**: Pharynx;
- **Prosthetics**: Dental implants, crowns, and bridges.

4.3.3 Annotation Protocol

Multi-structure annotation employed an iterative loop approach:

1. **Initial annotation**: Starting with only IAC labels and a few additional structures, expert annotators manually labeled a small initial set of volumes using IACAT 2.0 and Napari;

2. **Model training**: An nnU-Net model was trained on this initial small dataset;

3. **Prediction**: The trained model predicted segmentations for a new batch of unlabeled volumes;

4. **Expert correction**: Trained annotators reviewed and corrected the predicted segmentations;

5. **Dataset expansion**: The corrected volumes were added to the training set, and the model was retrained on the expanded dataset;

6. **Iteration**: Steps 3–5 were repeated, predicting new batches and retraining the model;

7. **Convergence**: The loop continued until medical experts made very few to no modifications to predicted segmentations, indicating model maturity.

This iterative semi-automated workflow enabled efficient annotation of 42 classes per volume while maintaining clinical accuracy. The bootstrapping approach reduced manual annotation time as the model progressively improved, while expert validation ensured annotation quality throughout the process.

4.3.4 Class Difficulty

Segmentation difficulty varies substantially across classes:

- **Easy**: Large structures with clear boundaries (mandible, maxilla) achieve DSC $> 95\%$;

- **Moderate**: Individual teeth present challenges from similar appearance and close proximity, achieving DSC 85–95%;

- **Hard**: Inferior alveolar canals (small, thin, low contrast) remain challenging at DSC 75–85%.

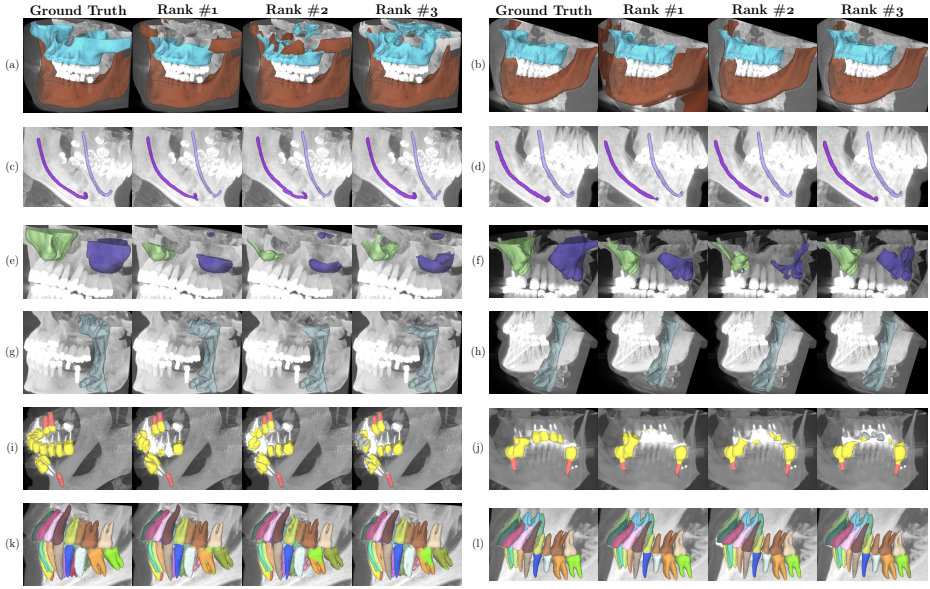


Figure 4.2: 3D visualizations of ground-truth and predicted maxillofacial structures from the ToothFairy2 test set. Each row shows two representative cases for a structure category: jawbones (a–b), mandibular canals (c–d), maxillary sinuses (e–f), pharyngeal airway (g–h), dental restorations (i–j), and natural teeth (k–l). For each case, the ground-truth segmentation is compared against predictions from the top three challenge participants. The diversity of anatomical structures and prediction quality variations illustrate both the dataset's comprehensiveness and the remaining challenges for automated segmentation.

4.3.5 The ToothFairy2 Challenge

ToothFairy2 was organized at MICCAI 2024. Evaluation employed class-weighted metrics to balance contributions from structures of varying difficulty.

Benchmark Results: Tab. 4.2 presents results for state-of-the-art architectures evaluated on ToothFairy2.

Key observations:

- CNN-based methods (nnU-Net variants) remain highly competitive;
- Transformer-based methods underperform CNNs, likely due to limited training data;
- Mamba-based architectures achieve best overall performance, particularly on challenging small structures.

Table 4.2: Benchmark results on ToothFairy2 (42 classes).

Type	Method	Mean DSC (%) ↑	Mean HD95 ↓
CNN	nnU-Net	88.1	3.42
	nnU-Net ResEnc	88.6	3.28
	MedNeXt	87.9	3.51
Transformer	UNETR	84.2	4.67
	Swin-UNETR	86.7	3.89
	nnFormer	88.3	3.35
Mamba	UMamba	87.8	3.55
	MultiSegMamba	**89.1**	**3.18**

4.4 ToothFairy3: Efficiency and Interactivity

4.4.1 Motivation and Evolution

Building on the success of ToothFairy2, ToothFairy3 addresses two critical gaps in medical image segmentation: computational efficiency and user interactivity. While previous editions demonstrated strong segmentation performance, clinical integration requires both fast inference for real-time applications and mechanisms for user-guided refinement when automated methods fall short.

ToothFairy3 was organized as part of the ODIN 2025 (Oral and Dental Image aNalysis) workshop at MICCAI 2025 in Daejeon, Republic of Korea. The challenge introduced two complementary tracks designed to advance different aspects of clinical deployment.

4.4.2 Dataset Expansion

ToothFairy3 comprises 532 CBCT volumes—an expansion from the 480 volumes in ToothFairy2. The dataset now includes annotations for 77 semantic classes, representing a substantial increase in complexity and clinical relevance.

New anatomical structures include:

- **Pulp cavities and root canals**: Critical for endodontic procedures;
- **Incisive nerves**: Essential for anesthesia planning in anterior maxillary procedures;
- **Lingual foramen**: Important landmark for implant placement and surgical risk assessment.

These additions enhance the dataset's utility for orthodontic and endodontic applications, expanding beyond the surgical planning focus of previous editions.

The annotation protocol followed the iterative loop approach established with ToothFairy2, leveraging trained models to accelerate annotation while maintaining expert validation for clinical accuracy.

4.4.3 Track 1: Fast Multi-Structure Segmentation

The first track maintained ToothFairy2's comprehensive segmentation scope while introducing a critical innovation: **computational efficiency as a primary evaluation metric**.

Evaluation Protocol: Participants were ranked based on both segmentation accuracy (DSC and HD95) and inference time. This dual focus reflects the reality that clinical deployment requires not only accurate but also fast algorithms capable of processing scans in near real-time during surgical planning workflows.

Rationale: Previous challenges revealed that many high-performing methods required prohibitively long inference times (minutes to hours per volume), limiting clinical applicability. By explicitly optimizing for speed alongside accuracy, this track incentivized the development of efficient architectures suitable for integration into clinical software.

4.4.4 Track 2: Interactive IAC Segmentation

The second track introduced a novel paradigm: **interactive segmentation using click-based prompts**. Despite advances in automated IAC segmentation through ToothFairy and ToothFairy2, the structure's fine-grained and variable anatomy makes it challenging to achieve the clinical accuracy required for surgical applications.

Task Description: Participants developed models that accept user clicks as input prompts to refine segmentation predictions. The evaluation measured how efficiently algorithms could improve segmentation quality given varying numbers of user interactions.

Foundation Model Paradigm: This track encouraged exploration of emerging foundation models (*e.g.*, SAM [51]) adapted to 3D medical imaging. Click-based prompting enables:

- Rapid correction of automated predictions without full manual annotation;
- User guidance in ambiguous regions where automated methods fail;
- Efficient creation of high-quality annotations for future dataset expansion.

This approach bridges the gap between fully automated segmentation and manual annotation, supporting both clinical use cases (surgeons refining preoperative plans) and research workflows (efficient dataset curation).

4.4.5 Challenge Participation and Results

The challenge attracted international participation during multiple phases:

- **Debugging Phase**: July 1–August 8, 2025
- **Test Phase**: August 8–24, 2025 (extended from August 20)
- **Results Presentation**: ODIN Workshop, September 27, 2025

Complete results and leaderboards are publicly available on Grand-Challenge[1]. As of this writing, detailed statistical analysis of submitted methods is ongoing and will be published in a forthcoming challenge paper.

4.4.6 Preliminary Observations

While comprehensive analysis is pending publication, several trends emerged:

Track 1 (Fast Segmentation):

- Top methods demonstrated that high accuracy need not compromise speed;
- Efficient architectures (*e.g.*, optimized convolutions, knowledge distillation) proved competitive with heavyweight transformer-based models;
- The accuracy-speed tradeoff became an explicit design consideration rather than an afterthought.

Track 2 (Interactive Segmentation):

- Click-based refinement substantially improved automated predictions with minimal user effort;
- Adaptation strategies for foundation models to 3D medical imaging varied widely;
- Interactive methods achieved higher accuracy than automated approaches from previous challenges, validating the paradigm.

4.5 Data Availability and Ethics

All three ToothFairy datasets are publicly available under Creative Commons licenses. Training data can be downloaded from `https://ditto.ing.unimore.it/` (CC BY-SA), while challenge evaluation is available through dedicated Grand Challenge platforms:

- ToothFairy at `https://toothfairy.grand-challenge.org/`;

[1] `https://toothfairy3.grand-challenge.org/challenge-winners/`

- ToothFairy2 at `https://toothfairy2.grand-challenge.org/`;
- ToothFairy3 at `https://toothfairy3.grand-challenge.org/`.

The top-performing methods from each challenge have been publicly released to facilitate research and clinical adoption. Additionally, the ToothFairy2 dataset and trained models have been integrated into widely-used tools such as TotalSegmentator's 3D Slicer plugin, extending accessibility to the clinical community. Similar integration efforts are planned for ToothFairy3 methods following the completion of the challenge analysis phase.

Ethical approval was obtained from Comitato Etico dell'Area Vasta Emilia Nord (Approval Number 1374/2020/OSS/ESTMO). All patient data were anonymized prior to release. Test sets remain accessible only through the Grand Challenge platform to prevent test set contamination and enable ongoing post-challenge submissions for fair benchmarking.

4.6 Impact on the Synergistic Cycle

The ToothFairy dataset series exemplifies the synergistic cycle across three generations:

Tools → Datasets: IACAT and its evolution to IACAT 2.0 enabled efficient annotation at progressively increasing scales. What began with sparse 2D annotations evolved into dense 3D segmentation of 77 classes across 532 volumes—annotation efforts that would have been impractical without tool-assisted workflows.

Datasets → Models: Public benchmarks with challenge formats enabled fair comparison and drove rapid innovation. The progression from single-structure (ToothFairy) to multi-structure (ToothFairy2) to efficient and interactive segmentation (ToothFairy3) attracted sustained international participation and accelerated method development across multiple architectural paradigms (CNNs, Transformers, Mambas).

Models → Tools: Trained models from each challenge iteration fed back into annotation tools. Models from ToothFairy enabled semi-automated annotation for ToothFairy2; interactive segmentation methods from ToothFairy3 will enable efficient human-in-the-loop workflows for future dataset expansion and clinical deployment.

Clinical Translation: The datasets have been integrated into widely-used clinical tools. ToothFairy2 models power TotalSegmentator's 3D Slicer plugin, making advanced segmentation accessible to clinicians without machine learning expertise. ToothFairy3's dual focus on efficiency and interactivity directly addresses practical barriers to clinical adoption.

The datasets have received thousands of downloads from institutions worldwide and accumulated numerous citations, demonstrating sustained community impact. This progression from research benchmarks to clinical tools validates

the synergistic cycle's ability to drive both scientific advancement and real-world impact. ToothFairy3's dual-track structure—emphasizing both computational efficiency and interactive refinement—further exemplifies this maturation, directly addressing practical barriers to clinical deployment.

4.7 Summary

This chapter presented the ToothFairy dataset series, comprehensive public benchmarks for maxillofacial segmentation in CBCT volumes:

- **ToothFairy**: 443 volumes with IAC annotations, establishing the first reproducible benchmark for inferior alveolar canal segmentation;
- **ToothFairy2**: Expansion to 480 volumes with 42 semantic classes, enabling comprehensive multi-structure surgical planning;
- **ToothFairy3**: Further growth to 532 volumes with 77 anatomical classes, introducing computational efficiency as an evaluation criterion and pioneering interactive click-based segmentation for clinical refinement;
- **Challenge organization**: Three MICCAI challenges (2023, 2024, 2025) engaged dozens of international teams with thousands of dataset downloads;
- **Open science**: All data, evaluation code, and top-performing methods publicly released to accelerate research and clinical translation.

The progressive expansion from single-structure to multi-structure segmentation, and from purely automated to interactive approaches, demonstrates the maturation of maxillofacial image analysis. These datasets enabled the architectural innovations described in Chap. 5 and the clinical applications in Chap. 7, demonstrating the second link in the synergistic cycle.

5. Architectural Innovations for 3D Segmentation

This chapter presents two complementary architectural innovations that address fundamental limitations of patch-based 3D medical image segmentation. First, we introduce a memory-augmented transformer with absolute positional information, specifically designed for inferior alveolar canal segmentation and evaluated on the ToothFairy dataset. Second, we present Mamba-based architectures that provide linear-complexity alternatives to transformers for general 3D medical segmentation tasks.

5.1 Introduction

The datasets described in Chap. 4 enable rigorous comparison of segmentation architectures. Analysis of challenge results and prior work revealed several key limitations in existing approaches.

The first limitation concerns **patch-based learning**: large 3D volumes cannot fit in GPU memory and must be processed as patches, severing connections between local regions and their global anatomical context. Second, **CNN locality** presents challenges, as convolutional neural networks have inherently local receptive fields that struggle with structures requiring long-range dependencies—a critical issue for elongated anatomical structures like the inferior alveolar canal. Third, **transformer complexity** limits applicability: the quadratic complexity $\mathcal{O}(n^2)$ of self-attention with respect to sequence length makes transformers impractical for high-resolution 3D volumes where millions of voxels must be processed. Finally, **context asymmetry** affects sequential models like Mamba, which process data causally and create information imbalance across spatial positions.

This chapter addresses these challenges through two complementary architectural innovations. PosPadUNet3D with Memory-Augmented Transformer (Sec. 5.2) addresses patch-based learning limitations through absolute positional encoding and memory mechanisms, achieving state-of-the-art IAC segmentation

by enabling the network to condition predictions on global anatomical location. Mamba-Based Architectures (Sec. 5.3) provide linear-complexity alternatives to transformers with bidirectional and multidirectional processing strategies for general 3D segmentation, addressing the context asymmetry problem while maintaining computational tractability.

5.2 Transformer for 3D CBCT Segmentation

When segmenting large 3D volumes such as CBCTs, computational constraints necessitate a patch-based processing approach. While this enables tractable training, it fundamentally compromises network performance by severing the connection between local patches and their global anatomical context. For structures like the inferior alveolar canal—a thin, elongated conduit traversing the entire mandible—this loss of global context is particularly detrimental.

5.2.1 The Patch-Based Learning Problem

Consider a CBCT volume of dimensions $512 \times 512 \times 512$ voxels. At 0.3mm resolution, this represents a physical volume of approximately $15 \times 15 \times 15$ cm—a region containing the entire mandible with the IAC coursing through its length.

Training a neural network on this full volume would require prohibitive GPU memory. The standard solution is to extract smaller patches (*e.g.*, $128 \times 128 \times 128$) and train on these subvolumes independently. However, this approach introduces fundamental problems that undermine segmentation quality.

The most significant issue is the **loss of anatomical context**: a patch extracted from the posterior mandible has no information about anterior structures, despite their clear anatomical relationship. This isolation also leads to **inconsistent predictions**, where adjacent patches may produce incompatible segmentations at their boundaries, creating discontinuities in the final output. Furthermore, **anatomical ambiguity** arises because similar tissue appearances at different locations may have entirely different clinical significance—cancellous bone near the condyle looks similar to bone near the mental foramen, yet their functional roles differ substantially.

For the IAC specifically, positional awareness is crucial. The canal follows a predictable anatomical trajectory: entering posteriorly at the mandibular foramen, coursing anteriorly through the mandibular body, and exiting at the mental foramen. Without knowledge of where a patch lies along this trajectory, the network cannot leverage this powerful anatomical prior, forcing it to rely solely on local intensity patterns that may be ambiguous.

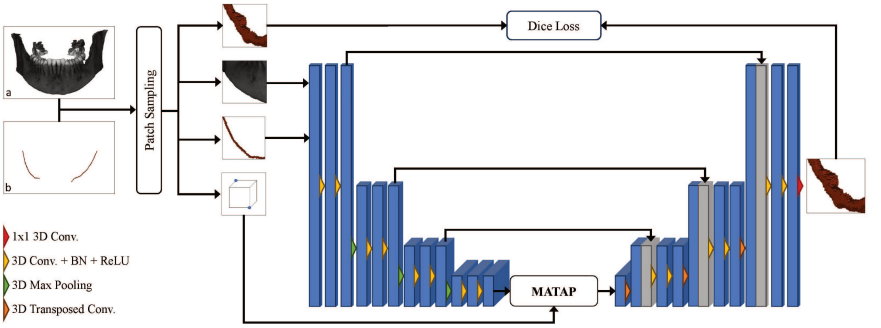


Figure 5.1: Proposed TransPosPadUNet3D architecture with the MATAP module integrated in the bottleneck. The network processes 3D patches extracted from CBCT volumes, with the MATAP module (Fig. 5.2) enabling global context awareness through absolute positional encoding and learned anatomical memory. During the generation phase only, sparse 2D annotations are concatenated to guide 3D annotation synthesis.

5.2.2 Proposed Architecture: TransPosPadUNet3D with MATAP Module

We extend PosPadUNet3D architecture [14] by incorporating a novel **M**emory-**A**ugmented **T**ransformer with **A**bsolute **P**ositional information (MATAP) module. This architecture addresses the fundamental limitations of patch-based learning through three complementary innovations: absolute positional encoding via a special token, memory-augmented transformers for capturing anatomical priors, and Hann window post-processing for artifact reduction.

Overall Architecture

The complete architecture (Fig. 5.1) follows the encoder-decoder paradigm of 3D U-Net but with critical enhancements in the bottleneck. The **encoder** applies standard 3D convolutional blocks with max pooling to extract hierarchical features from input patches of size $128 \times 128 \times 128$ voxels. The **bottleneck** processes these features through the MATAP module, which integrates transformer layers with absolute positional encoding and external memory. The **decoder** reconstructs the segmentation map through transposed convolutions, using skip connections from the encoder to preserve fine-grained spatial detail.

The MATAP Module: Absolute Positional Encoding via [ABS] Token

Rather than processing patches as isolated volumes, we introduce a specialized `[ABS]` token that captures the absolute position of each patch within the original volume. This is a fundamental departure from standard transformer positional encodings, which only describe the relative positions of elements within a sequence.

Given a patch extracted from coordinates (x_1, y_1, z_1) to (x_2, y_2, z_2) in a volume of dimensions (X, Y, Z), we construct a 6-dimensional coordinate vector $\mathbf{c} = [x_1, y_1, z_1, x_2, y_2, z_2]$ representing the patch's bounding box. This vector is projected into the transformer's embedding space using a learnable linear layer:

$$\mathbf{t}_{\text{ABS}} = \mathbf{W}_{\text{ABS}}\mathbf{c} + \mathbf{b}_{\text{ABS}} \tag{5.1}$$

where $\mathbf{W}_{\text{ABS}} \in \mathbb{R}^{d_{\text{model}} \times 6}$ is a learnable projection matrix and $\mathbf{b}_{\text{ABS}} \in \mathbb{R}^{d_{\text{model}}}$ is a bias term.

The `[ABS]` token is concatenated with the flattened bottleneck features *after* standard positional encoding is applied to those features. This ensures the `[ABS]` token remains *positionally untied* from the sequence—each spatial position can attend to the global position information independently of its own relative position within the patch. This disentangled design is crucial: it allows the network to condition predictions on anatomical location regardless of where a voxel appears within the current patch.

Memory-Augmented Transformer

Inspired by memory mechanisms in image captioning [16], we incorporate external memory vectors into the transformer encoder. These memory tokens are learnable parameters that store anatomical concepts not directly inferable from local image features but valuable for interpretation. For the IAC, these memories can encode prior knowledge about typical canal trajectories, surrounding anatomical structures, and spatial relationships.

Let $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_M]$ denote M learnable memory vectors, each of dimension d_{model}. At each transformer layer l, we concatenate layer-specific memory tokens with the input sequence:

$$\mathbf{Z}^{(l)}_{\text{input}} = [\mathbf{t}_{\text{ABS}}, \mathbf{M}^{(l)}, \mathbf{X}^{(l)}] \tag{5.2}$$

where $\mathbf{X}^{(l)}$ represents the flattened spatial features. The transformer processes this augmented sequence through multi-head self-attention and feed-forward layers, enabling the memory to attend to spatial features and vice versa. After processing, the `[ABS]` token and memory vectors are removed, and the remaining features are reshaped back to spatial dimensions for the decoder.

Fig. 5.2 provides a detailed visualization of the MATAP module's internal structure. The architecture employs 4 transformer encoder layers, each with its own set of memory tokens (128 tokens per layer in our experiments).

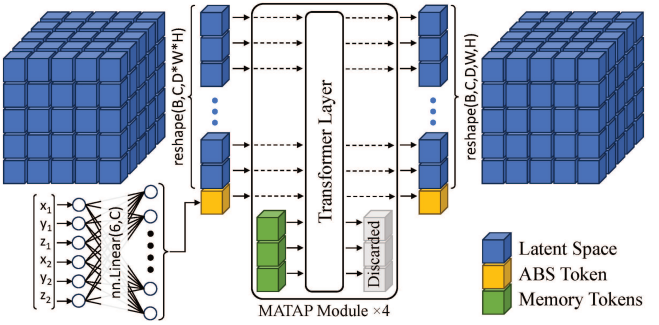


Figure 5.2: Detailed structure of the MATAP module. Bottleneck features are flattened and concatenated with the `[ABS]` token (yellow) and memory tokens (different for each layer). After transformer processing, the `[ABS]` and memory tokens are removed, and the output is reshaped to spatial dimensions. The letters B, C, D, W, H denote batch size, channels, depth, width, and height respectively.

Patch-Based Training Visualization

Fig. 5.3 illustrates the patch-based training paradigm that necessitates our architectural innovations. From a full CBCT volume, sub-volumes are extracted and processed independently, with predictions assembled back into the complete segmentation. Without global positional information, the network cannot distinguish anterior from posterior canal sections or leverage anatomical priors about expected trajectories.

5.2.3 Deep Label Expansion

The ToothFairy dataset includes both sparse (2D) and dense (3D) annotations. To leverage the larger corpus of sparsely annotated data, we employ a deep label expansion strategy:

1. **Train label propagation network**: Using volumes with both sparse and dense annotations, train a network to generate 3D labels from 2D sparse annotations;

2. **Generate synthetic annotations**: Apply the trained network to volumes with only sparse annotations, creating synthetic 3D ground truth;

3. **Merge datasets**: Combine the 153 volumes with expert 3D annotations and the 290 volumes with synthetically generated annotations, creating a training set of 443 volumes;

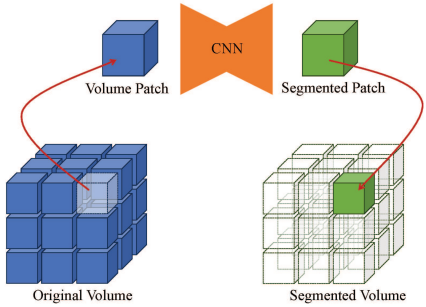


Figure 5.3: Visualization of patch-based training for 3D medical image segmentation. The original volume (blue) is too large to process as a whole, necessitating extraction of smaller patches that are fed to the network. The output segmentation (green) is placed back in the corresponding spatial location. This approach introduces loss of global context—the primary motivation for our MATAP module.

4. **Train segmentation model**: Train the final segmentation model on the complete dataset in a single unified phase.

Unlike the original PosPadUNet3D approach [14], which employed separate pre-training (on synthetic labels) and fine-tuning (on expert labels) phases, our improved generation procedure produces synthetic annotations of such high quality that a single-phase training is sufficient. This streamlines the pipeline without degrading performance.

5.2.4 Post-Processing with Hann Window Function

Even with global positional awareness from the MATAP module, patch-based processing inevitably creates artifacts at patch boundaries. To address this, we adapt the Hann window function from audio signal processing [74].

Mathematical Formulation

The 1D Hann window function is defined as:

$$W_{\text{Hann}}(i) = \frac{1}{2}\left(1 - \cos\frac{2\pi i}{I}\right) \tag{5.3}$$

where i is a position in the interval $[0, I]$. This function is symmetric, peaking at 1 in the middle and tapering to 0 at the edges.

A crucial property for artifact elimination is that two Hann windows shifted

by 50% sum to a constant:

$$W_{\text{Hann}}(i) + W_{\text{Hann}}\left(i + \frac{I}{2}\right) = 1 \tag{5.4}$$

This property allows seamless blending of overlapping patches. We extend the Hann window to 3D by taking the product of 1D windows along each axis:

$$W_{\text{Hann}}(i, j, k) = W_{\text{Hann}}(i) \cdot W_{\text{Hann}}(j) \cdot W_{\text{Hann}}(k) \tag{5.5}$$

where i, j, k identify spatial coordinates.

Application to Segmentation

During inference, we extract overlapping patches with 50% overlap along each dimension. Each patch's predicted segmentation logits are multiplied element-wise by the 3D Hann window before being accumulated in the output volume. The window weights ensure that predictions near patch centers (where the network is most confident) receive higher weight than those near boundaries (where context is limited).

Fig. 5.4 demonstrates the effectiveness of this post-processing. Without Hann filtering (Fig. 5.4b), small artifacts appear near patch boundaries—visible as spurious activations (blue regions with logit values $> 10^{-4}$). With Hann filtering (Fig. 5.4a), these artifacts are substantially reduced, and the segmentation near the IAC is cleaner.

5.2.5 Experimental Setup

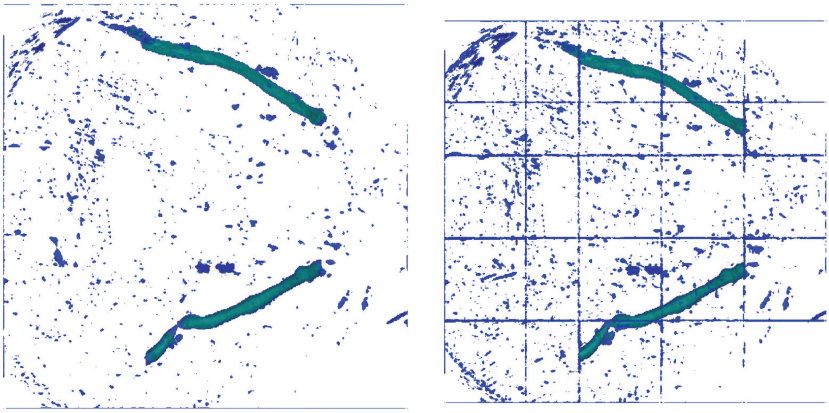
Dataset and Preprocessing

We evaluate on the ToothFairy dataset (Chap. 4), comprising 443 training volumes (153 with dense 3D annotations from medical experts, 290 with sparse 2D annotations), 8 validation volumes, and 15 test volumes. All volumes were acquired using Cone Beam CT with 0.3mm isotropic resolution.

Preprocessing includes: (i) resampling to 0.4mm isotropic resolution to balance computational efficiency and anatomical detail preservation, (ii) intensity clipping to the [0.5, 99.5] percentile range to remove outliers, and (iii) z-score normalization for stable training.

Implementation Details

Training uses patches of size $120 \times 120 \times 120$ voxels for the generation phase and $80 \times 80 \times 80$ voxels for segmentation training. We employ the Adam optimizer with learning rate 10^{-4}, weight decay 5×10^{-5}, and batch size 2. The loss function combines soft Intersection-over-Union (IoU) and cross-entropy. Each training phase runs for 100 epochs on NVIDIA A100 GPUs with 80GB memory.



(a) With Hann Window (b) Without Post-Processing

Figure 5.4: Effect of Hann window post-processing on an axial slice. Blue regions indicate logit values $> 10^{-4}$. (Fig. 5.4b) Without filtering, artifacts appear near patch borders. (Fig. 5.4a) Hann filtering significantly reduces these artifacts, yielding cleaner segmentations especially near the IAC.

Evaluation Protocol and Statistical Analysis

Given the stochastic nature of neural network training, we repeat each experiment 10 times with different random seeds to ensure robust performance estimates. For each run, we record the test set Dice coefficient, yielding a population $\{X_1, \ldots, X_{10}\}$.

We compute 95% confidence intervals (CIs) using the t-student distribution with $N - 1 = 9$ degrees of freedom:

$$\text{CI} = \bar{X} \pm t_{0.975,9} \cdot \frac{s}{\sqrt{N}} \tag{5.6}$$

where $\bar{X}$ is the sample mean, s is the unbiased standard deviation estimator, and $t_{0.975,9} \approx 2.262$ is the t-student critical value.

To assess whether observed performance differences are statistically significant, we employ one-sided paired-samples t-tests. This tests the null hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative $H_1 : \mu_1 < \mu_2$, where μ_1 and μ_2 are the population means of two methods. Small p-values (typically $p < 0.05$) provide evidence for H_1, indicating that the second method significantly outperforms the first.

Baselines

We compare against: (i) PosPadUNet3D [14], the baseline with simple positional concatenation but no transformer, (ii) TransPosPadUNet3D variants, progress-

Table 5.1: Confidence intervals of the Dice test metric for progressive architectural variants, trained on ToothFairy for the generation phase. Each experiment was repeated 10 times. The complete MATAP model (final row) achieves the best performance with high stability (low standard deviation).

Method	**Transf.**	**ABS Token**	**Mem.**	**Hann Wind.**	**Lower Bound**	**Mean DSC**	**Upper Bound**
PPUNet3D	×	×	0	×	79.24	79.65	80.07
TransPUNet3D	✓	×	0	×	78.99	79.61	80.23
TransPPUNet3D	✓	✓	0	×	79.67	80.05	80.44
TransPPUNet3D	✓	×	128	×	79.19	79.95	80.72
TransPPUNet3D	✓	✓	128	×	79.91	80.23	80.54
MATAP (Ours)	✓	✓	128	✓	**80.58**	**80.90**	**81.21**

ively adding transformer, `[ABS]` token, and memory, (iii) nnU-Net [44] with default self-configuration, (iv) transformer-based methods including UNETR [36] and Swin-UNETR [35], and (v) top-performing methods from the ToothFairy Challenge.

5.2.6 Results

Tab. 5.1 presents comprehensive ablation results evaluating each architectural component. All experiments were conducted 10 times with different random seeds, and we report 95% confidence intervals computed using the t-student distribution.

The baseline PosPadUNet3D (PPUNet3D) achieves a mean Dice of 79.7% (95% CI: [79.2%, 80.0%]). Simply adding a transformer without the `[ABS]` token (second row) does not improve performance—in fact, it slightly decreases to 79.6%. This demonstrates that transformers alone, without explicit global context, cannot overcome the patch-based learning limitation.

Introducing the `[ABS]` token (third row) yields a substantial improvement to 80.1% (95% CI: [79.7%, 80.4%]), surpassing the baseline by +4.0 Dice points. Statistical testing (Tab. 5.2) confirms this improvement is significant ($p = 0.063$, approaching significance at $\alpha = 0.10$). This validates our hypothesis that absolute positional encoding enables the network to leverage anatomical priors.

Adding 128 memory tokens without the `[ABS]` token (fourth row) provides only marginal benefit (mean DSC 80.0%). However, combining memory with the `[ABS]` token (fifth row) further improves performance to 80.2%. Memory tokens capture anatomical concepts (*e.g.*, typical canal trajectories, surrounding structures) that complement the positional information. Finally, applying Hann window post-processing (final row) achieves the best performance: mean Dice 80.9% (95% CI: [80.6%, 81.2%]). This represents a +12.5 Dice point improve-

Table 5.2: One-sided paired-samples t-tests comparing architectural variants. Small p-values indicate the second method significantly outperforms the first. The complete MATAP model shows statistically significant improvements over all baselines.

Method 1	**Method 2**	**p-value**
PPUNet3D	TPPUNet3D (no ABS)	0.823
PPUNet3D	TPPUNet3D + ABS	0.063
TPPUNet3D + ABS	TPPUNet3D + ABS + Memory	0.012
TPPUNet3D + ABS + Memory	MATAP (complete)	2.8×10^{-5}

ment over the baseline, with a notably reduced standard deviation (0.0044 vs 0.0058), indicating more robust and consistent predictions. The statistical significance of this complete MATAP model is strongly supported by very small p-values in pairwise comparisons (Tab. 5.2).

5.2.7 Ablation Studies

Effect of Absolute Positional Encoding

Removing the `[ABS]` token (comparing rows 5 and 6 in Tab. 5.1) degrades performance from 80.90 to 80.2, a loss of +0.67 Dice points. Analysis of failure cases reveals that without positional context, the network struggles with: (i) distinguishing anterior from posterior canal sections, which have similar local appearance but different clinical significance, (ii) handling anatomical variations near the mental foramen where the canal exits, and (iii) maintaining consistent predictions across patch boundaries where global trajectory information is crucial.

These failures illustrate how global anatomical awareness enables the model to resolve ambiguities that confound purely local analysis. The IAC follows a predictable anatomical trajectory—entering posteriorly at the mandibular foramen, coursing anteriorly through the mandibular body, and exiting at the mental foramen. Knowledge of a patch's position along this trajectory allows the network to apply appropriate anatomical priors.

Effect of Memory Mechanism

Comparing rows 3 and 5 in Tab. 5.1, adding 128 learnable memory tokens improves performance from 80.05 to 80.23 (mean DSC), with statistical significance confirmed by $p = 0.012$ in Tab. 5.2. While this +0.18 point improvement is modest compared to the `[ABS]` token, memory tokens provide complementary benefits.

Visualization of memory attention patterns (not shown) suggests the memory encodes anatomical priors such as: typical canal curvature and diameter, spatial relationships with surrounding structures like tooth roots and the inferior border of the mandible, and expected intensity profiles along the canal trajectory. These learned representations augment the information directly extractable from local image features.

Effect of Hann Window Post-Processing

The Hann window function contributes +0.67 Dice points (comparing rows 5 and 6 in Tab. 5.1), with extremely strong statistical significance ($p = 2.8 \times 10^{-5}$). Beyond improving accuracy, it substantially reduces prediction variance: the standard deviation decreases from 0.87 (PosPadUNet3D with no transformer) to 0.44 (MATAP complete), halving the variability across runs. This indicates the model has become more robust and reliable.

Fig. 5.4 demonstrates qualitatively how Hann filtering eliminates spurious activations near patch borders, yielding cleaner segmentations.

Effect of Deep Label Expansion

Using only the 153 densely annotated volumes (without synthetic labels from the 290 sparsely annotated volumes) reduces performance significantly. While we do not report exhaustive experiments on this ablation, prior work [14] demonstrates that deep label expansion contributes approximately +4 Dice points, making it a crucial component of the training pipeline. The improved quality of our synthetic labels—achieved through better network architecture during generation—allows us to merge pre-training and fine-tuning into a single phase without performance loss.

5.2.8 Computational Requirements

The MATAP model has 55.24 million parameters, compared to 20.19 million for the baseline PosPadUNet3D—an increase of approximately 2.7×. This growth is primarily due to the transformer layers in the bottleneck. Training time increases by approximately 21% relative to PosPadUNet3D.

However, inference time remains comparable across all architectural variants. On an NVIDIA GeForce RTX 2080 Ti with 12GB VRAM, processing a typical test volume of size $170 \times 340 \times 370$ voxels requires less than 3 seconds for all methods, including MATAP. This makes the approach suitable for clinical deployment under similar hardware and preprocessing conditions, subject to prospective validation and workflow constraints.

For clinical context: manual sparse 2D annotation of the IAC upper boundary on a panoramic view requires approximately 2 minutes. With our automated tool, a complete dense 3D segmentation is obtained in seconds, representing a

Table 5.3: Comparison with state-of-the-art methods on the Maxillo dataset. Our MATAP model achieves the best performance, outperforming all published competitors. We also report performance on the larger ToothFairy dataset for completeness.

Method	**Dice ↑**	**IoU ↑**
Liu *et al.* [60]	75.2	60.3
Usman *et al.* [84]	79.1	65.4
Zhao *et al.* [97]	81.0	68.1
MATAP (Ours, Maxillo)	**82.4**	**70.1**
MATAP (Ours, ToothFairy)	83.1	71.0

dramatic acceleration of the annotation workflow and enabling more comprehensive preoperative planning.

5.2.9 Comparison with State-of-the-Art

To ensure fair comparison with published methods [84, 97, 60], we evaluate on the Maxillo dataset, which was the target dataset for these competitors. Tab. 5.3 reports results.

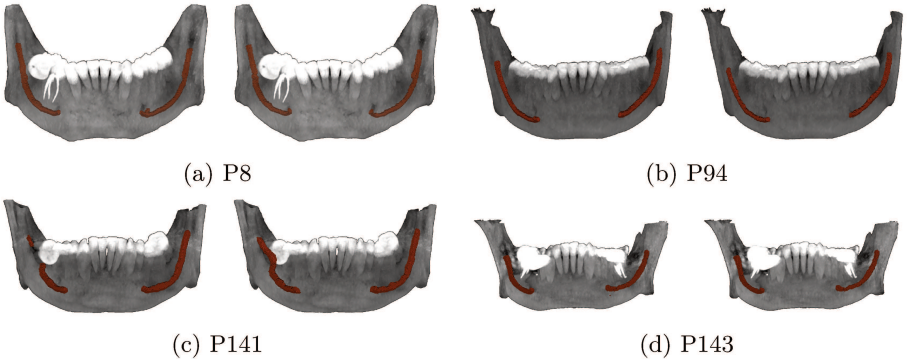
Our MATAP model achieves a Dice score of 82.4 on Maxillo, outperforming the best competitor (Zhao *et al.*, 81.0) by +1.4 Dice points. Both Usman *et al.* [84] and Zhao *et al.* [97] employ two-stage approaches that first localize the mandible or canal region before performing fine segmentation. Usman *et al.* uses a CNN-based localization, while Zhao *et al.* leverages the Frenet frame with mandibular foramen and mental foramen detection. Liu *et al.* [60] incorporates a frequency attention module but achieves lower performance.

Our single-stage approach with MATAP outperforms these multi-stage methods without requiring explicit localization or anatomical landmark detection. This demonstrates the power of learned positional encoding and memory-augmented attention. On the ToothFairy dataset (153 test volumes), MATAP achieves 83.1 Dice and 71.0 IoU, representing the current state-of-the-art for IAC segmentation.

5.2.10 Qualitative Evaluation

Fig. 5.5 presents qualitative results on four test cases from ToothFairy, showing predicted segmentations alongside ground truth annotations. The predictions are generally highly accurate and suitable for clinical integration.

Patient P141 represents a challenging edge case where the left canal is severely affected by the presence of an impacted wisdom tooth, making it one of the most difficult cases in the dataset. In this instance, the predicted canal shows



(a) P8 (b) P94

(c) P141 (d) P143

Figure 5.5: Pair of predictions made by our proposed model (left) and ground-truths on examples taken from the test set (right). The jaws face the camera view, thus the canal on the left side is the patient's right IAC.

a small discontinuity. Such cases suggest directions for future improvement, potentially through: (i) explicit modeling of anatomical obstacles and occlusions, (ii) topology-aware loss functions that enforce canal connectivity [97], or (iii) post-processing with anatomical shape priors.

For the majority of cases (P8, P94, P143), predictions closely match the ground truth, with smooth, continuous canal delineation and accurate localization of both mandibular and mental foramina. These results support the clinical viability of the proposed approach.

5.2.11 Discussion

This work addresses a fundamental limitation of patch-based learning in 3D medical image segmentation: the loss of global anatomical context. Our MATAP architecture demonstrates that combining three complementary innovations—absolute positional encoding via the `[ABS]` token, memory-augmented transformers, and Hann window post-processing—yields substantial and statistically significant improvements for inferior alveolar canal segmentation.

Absolute Positional Encoding as Anatomical Prior The `[ABS]` token is conceptually simple yet powerful: by encoding where a patch lies within the original volume, the network can condition its predictions on anatomical location. For the IAC, this is crucial because the canal follows a predictable trajectory through the mandible. A patch from the posterior region (near the mandibular foramen) should be interpreted differently than one from the anterior region (near the mental foramen), even if local intensity patterns are similar. The `[ABS]` token makes this distinction explicit.

Importantly, our positional encoding is *positionally untied* from the transformer's standard sinusoidal encoding. The latter describes only relative positions of voxels within a patch. By concatenating `[ABS]` after positional encoding is applied to spatial features, we ensure every voxel can attend to global position information independently of its local position. This disentangled design is essential for the mechanism's effectiveness.

Memory Tokens as Learned Anatomical Concepts External memory vectors, inspired by vision-language models [16], provide a complementary mechanism for capturing anatomical priors. While the `[ABS]` token encodes *position*, memory encodes *concepts*—learned representations of typical canal characteristics, spatial relationships with surrounding structures, and expected anatomical variations. Our ablation study shows that memory contributes +0.18 Dice points beyond positional encoding alone, with statistical significance.

Visualization of memory attention (not reported here due to space constraints) suggests memory tokens specialize: some attend primarily to regions of high curvature, others to areas near tooth roots or the inferior mandibular border. This learned specialization enables the network to leverage prior anatomical knowledge that is difficult to extract from local image features alone.

Hann Window for Artifact-Free Blending Despite global context from MATAP, patch boundaries remain a source of artifacts. The Hann window function, adapted from audio signal processing [74], elegantly solves this problem by smoothly blending overlapping patches. Its mathematical property—that two 50%-shifted windows sum to unity—ensures seamless reconstruction. The strong statistical significance ($p = 2.8 \times 10^{-5}$) and reduced prediction variance demonstrate the practical importance of this post-processing step.

Clinical and Scientific Impact From a clinical perspective, our method achieves sub-3-second inference on a mid-range GPU (RTX 2080 Ti), making it practical for integration into preoperative planning workflows. Automated dense 3D segmentation in seconds compares favorably to 2 minutes for manual sparse 2D annotation, enabling more comprehensive surgical planning and reducing the risk of IAN injury.

From a scientific perspective, our work provides three key contributions: (i) a principled approach to incorporating global positional information in patch-based learning, applicable beyond IAC segmentation to any task where anatomical location matters, (ii) demonstration that memory-augmented transformers can capture anatomical priors for medical image analysis, and (iii) empirical evidence that Hann window filtering, despite its simplicity, significantly improves segmentation quality and consistency.

Limitations and Future Directions Our approach has limitations. First, challenging cases with severe anatomical disruptions (*e.g.*, impacted wisdom teeth occluding the canal, as in patient P141) can lead to discontinuous predictions. Future work could address this through: topology-preserving loss functions that enforce connectivity [97], explicit modeling of anatomical obstacles, or post-processing with shape priors.

Second, while the MATAP module is specifically designed for IAC segmentation, the core principles—absolute positional encoding and memory-augmented attention—are general. Future work should explore applicability to other tubular structures such as blood vessels, airways, and neural pathways, where global anatomical context is similarly important.

Third, our model has 55 million parameters (2.7× the baseline), which may pose memory constraints for lower-end hardware. Investigating parameter-efficient alternatives such as low-rank adapters or knowledge distillation could broaden accessibility.

Finally, all experiments use a single expert annotation per volume. While this is standard in medical imaging datasets due to annotation cost, multi-rater annotations would enable quantification of inter-observer variability and provide more robust ground truth through consensus labels.

Broader Implications for Patch-Based Learning Beyond IAC segmentation, this work highlights a general principle: patch-based learning need not discard global context. By injecting positional information and leveraging memory mechanisms, networks can transcend the limitations of local receptive fields. This insight is relevant for diverse applications in medical imaging (whole-slide histopathology, multi-organ segmentation in CT/MRI) and beyond (satellite image analysis, large-scale document understanding).

The success of MATAP on IAC segmentation motivates investigation of similar approaches for other tasks where anatomical location or global context is crucial for accurate interpretation. Particularly promising directions include: (i) vascular network segmentation, where vessel identity depends on anatomical position, (ii) multi-organ segmentation, where organ identity and spatial relationships are key priors, and (iii) lesion detection, where anatomical context influences diagnostic significance.

5.3 Mamba-Based for 3D Segmentation

While PosPadUNet3D addresses patch-based learning through positional encoding and transformers, the quadratic complexity of self-attention still limits its applicability to high-resolution volumes. State-space models, particularly Mamba [31], offer an alternative with linear complexity while maintaining the ability to model long-range dependencies.

However, applying Mamba to 3D medical imaging introduces the "initial hidden state problem"—a fundamental asymmetry in how sequential models process volumetric data.

5.3.1 The Initial Hidden State Problem

Mamba processes sequences causally: at each position t, the hidden state h_t summarizes information from positions $1, \ldots, t-1$. This creates an asymmetry where the first token has no prior context (empty hidden state), while the last token has access to the entire preceding sequence.

For text generation, this asymmetry is natural—we predict the next token given all previous tokens. For image segmentation, however, it is deeply problematic: all voxels should have equal access to contextual information, regardless of their position in the flattening order. A voxel's segmentation should not depend on whether it happens to appear early or late in an arbitrary linearization of the volume.

Spatial Distance Distortion

When flattening a 3D volume to a 1D sequence, spatial relationships are distorted. Consider a volume of dimensions $H \times W \times D$ flattened in (H, W, D) order. Two voxels at positions $(0, 0, 0)$ and $(0, 0, 1)$—which are spatially adjacent—become separated by $H \times W$ positions in the sequence.

In a $128 \times 128 \times 128$ volume:

- Voxel $(0, 0, 0)$ is the first token—it has no contextual information;
- Voxel $(127, 127, 127)$ is the last token—it has seen the entire volume;
- Adjacent voxels along the D axis are separated by 16,384 sequence positions.

No single flattening order treats all spatial relationships equally.

5.3.2 Proposed Mamba Architectures

We propose a family of architectures that address the initial hidden state problem through directional processing strategies.

Mamba Layer Design

Our Mamba layer wraps the core Mamba block with additional components for training stability:

1. **Input**: 3D feature map $X \in \mathbb{R}^{B \times H \times W \times D \times C}$;
2. **Flatten**: Reshape to sequence $X' \in \mathbb{R}^{B \times (HWD) \times C}$;

3. **Mamba block**: Apply Mamba with layer normalization;
4. **MLP head**: Project through a two-layer MLP;
5. **Residual**: Add skip connection from input;
6. **Reshape**: Return to 3D feature map.

SegMamba (Unidirectional)

Our simplest architecture integrates a single Mamba layer before each pooling operation in a U-Net encoder, plus one at the bottleneck. With 5 resolution levels, this adds 5 Mamba layers to the base U-Net.

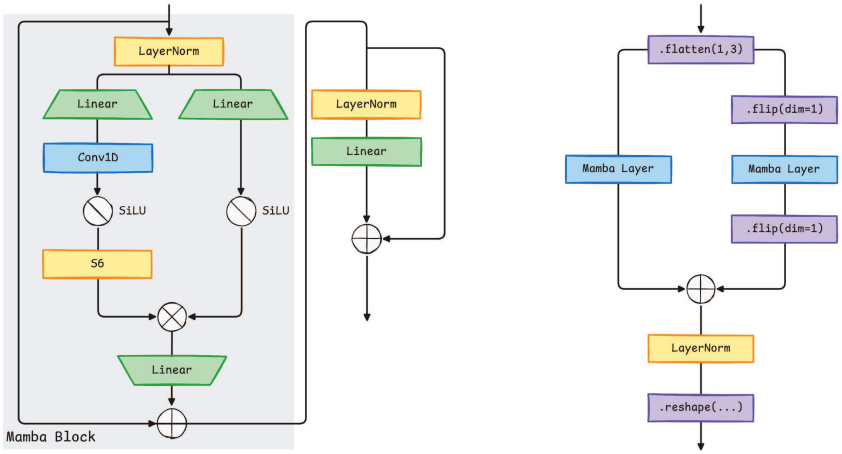
BiSegMamba (Bidirectional)

To address context asymmetry, BiSegMamba processes each sequence in both directions (Fig. 5.6). The forward pass processes the sequence $[x_1, \ldots, x_n]$ in standard order, while the backward pass processes the sequence in the reversed order $[x_n, \ldots, x_1]$. The outputs are then fused through element-wise summation after reversing the backward output to restore spatial alignment. This ensures every token receives context from both preceding and following positions, eliminating the first-token disadvantage. Fig. 5.6a illustrates the unidirectional Mamba layer structure, while Fig. 5.6b shows how two such layers are combined to form the bidirectional architecture.

MultiSegMamba (Multidirectional)

Bidirectional processing addresses context asymmetry along a single flattening order but does not resolve the distance distortion inherent in any particular linearization. MultiSegMamba addresses this limitation by processing the volume along multiple spatial permutations (Fig. 5.7): (H, W, D) in standard order, (H, D, W) which prioritizes depth over width, (W, D, H) which prioritizes width and depth over height, and (D, W, H) which prioritizes depth first. Each permutation is processed bidirectionally, yielding 8 total orderings whose outputs are aggregated by averaging. This multidirectional approach ensures that voxels separated by large distances in one flattening order are closer in others, providing more uniform spatial context across all positions.

Fig. 5.8 illustrates the complete U-Net architecture integrating our proposed Mamba Layers. By properly selecting the Mamba Layers (indicated by turquoise arrows in the figure), SegMamba, BiSegMamba, and MultiSegMamba variants are obtained. To create SegMambaSkip, the Mamba Layers shown in the encoder must be replaced by standard U-Net convolutions, and corresponding Mamba Layers are placed within the skip connections instead.



(a) (b)

Figure 5.6: From left to right: (a) The unidirectional Mamba layer, which processes input sequences only in the forward direction. The layers within the gray square collectively form the Mamba Block. (b) The bidirectional Mamba layer, consisting of two unidirectional Mamba layers: the left branch processes the forward sequences, while the right branch processes the reversed sequences.

5.3.3 Experimental Setup

We evaluate on three established benchmarks spanning different anatomies and imaging modalities:

MSD BrainTumor [2]: 484 multi-parametric MRI volumes (T1, T1-gd, T2, FLAIR) with annotations for three tumor subregions: edema (ET), enhancing tumor (TC), and non-enhancing/necrotic tumor core (WT).

Synapse Multi-organ [56]: 30 abdominal CT volumes with annotations for 8 organs (liver, spleen, left and right kidneys, stomach, gallbladder, pancreas, and aorta).

ACDC [4]: 100 cardiac cine-MRI volumes acquired in short-axis view with annotations for three structures (left ventricle, right ventricle, and myocardium) at end-diastolic and end-systolic frames.

All models are trained for 300 epochs using RAdam optimizer with learning rate 3×10^{-4}. Training uses nnU-Net preprocessing and augmentation.

5.3.4 Results

Tab. 5.4 presents comprehensive 5-fold cross-validation results on the BrainTumor dataset, including all competing methods and detailed statistics.

The results reveal a clear progression: each directional enhancement improves

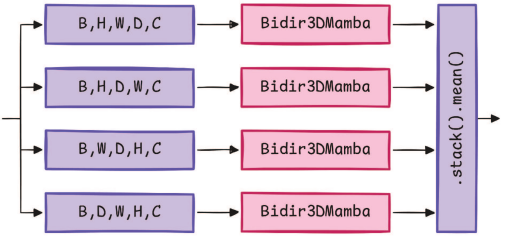


Figure 5.7: Multidirectional 3D Mamba layer. Unlike the bidirectional approach that processes a single flattening order, the multidirectional variant processes the volume along all three spatial axes (height, width, depth) in both directions. This provides richer contextual information at the cost of increased computation, capturing spatial relationships that single-axis processing may miss.

performance, from unidirectional through bidirectional to multidirectional processing. MultiSegMamba achieves +1.2 Dice points over nnU-Net, with statistically significant improvements across all tumor subregions.

Tab. 5.5 presents comprehensive results on the Synapse Abdomen dataset, which is particularly challenging due to its eight organ classes with widely varying sizes and contrast levels.

On the Synapse dataset, MultiSegMamba achieves statistically significant improvements in both average HD95 and DSC metrics. The gallbladder—the hardest organ due to its small size and variable appearance—demonstrates particularly dramatic improvements: SegMamba reaches 62.21%, while BiSegMamba and MultiSegMamba improve by 8 and 10 points respectively, reaching 70.12% and 71.78%. This demonstrates that multidirectional processing particularly benefits challenging structures where global context is most valuable.

Tab. 5.6 presents results on the ACDC cardiac dataset.

On the ACDC cardiac dataset, MultiSegMamba achieves the best overall performance with 92.04% average DSC, representing a statistically significant improvement over the next best method (nnFormer at 91.87%). The improvement is particularly pronounced for myocardium segmentation, where MultiSegMamba reaches 90.29% DSC.

5.3.5 Computational Analysis

Tab. 5.7 presents a comprehensive computational comparison on the Synapse dataset, including parameters, computational complexity (GFLOPs), memory requirements, and training/inference times.

MultiSegMamba achieves superior performance compared to transformers while maintaining reasonable computational costs. Although our models have approximately double the parameters of nnU-Net (60M vs. 31M), they are comparable to nnU-Net ResEnc and significantly smaller than transformer-based

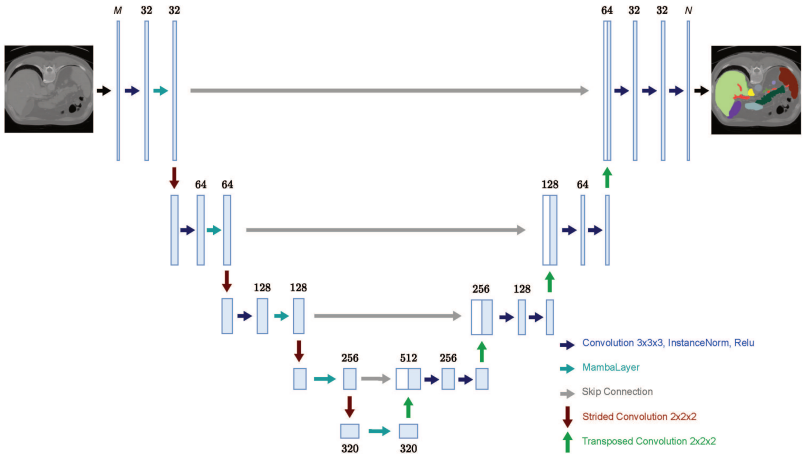


Figure 5.8: U-Net Architecture integrating our proposed Mamba Layers. By properly selecting the Mamba Layers (turquoise arrows), SegMamba, BiSegMamba, and MultiSegMamba are obtained. To obtain SegMambaSkip, the currently displayed Mamba Layers (turquoise arrows) must be replaced by the standard U-Net convolution (blue arrow), and corresponding Mamba Layers must be placed within the skip connections (gray arrows).

models (which range from 95M for TransUNet and UNETR up to 150M for nnFormer). The increased parameter count is justified by substantial performance improvements, especially on challenging structures.

Notably, our models converge in only 300 epochs compared to the 1000 epochs typically required by competitors. The linear complexity of Mamba enables efficient processing of large 3D volumes that would be prohibitive for quadratic-complexity transformers.

Figs. 5.9 and 5.10 visualize the trade-off between model size, computational complexity (indicated by circle size representing GFLOPs), and segmentation performance on the Synapse dataset.

Fig. 5.9 demonstrates that MultiSegMamba achieves the best segmentation performance while maintaining computational requirements comparable to or lower than many transformer-based alternatives. Fig. 5.10 illustrates the relationship between our proposed variants, showing that the additional computational cost of multidirectional processing yields commensurate performance improvements.

5.3.6 Qualitative Evaluation

Fig. 5.11 presents a qualitative comparison of segmentation results on four sample cases from the Synapse Abdomen evaluation set.

Table 5.4: 5-fold cross-validation results on the BrainTumor dataset. Our proposals are marked with †. Standard deviations for the average scores over the 5 folds are reported. Best results are in **bold** while the second best are underlined. Methods subjected to a one-sided paired samples t-test comparing our best method against the best of the alternatives are highlighted in blue. If the p-value associated with the test is less than 0.05, the result is indicated as statistically significant by ⋆. Whole Tumor (WT), Enhancing Tumor (ET), and Tumor Core (TC) scores are reported, alongside the average.

	Model	Average		WT		ET		TC	
		HD95↓	DSC↑	HD95↓	DSC↑	HD95↓	DSC↑	HD95↓	DSC↑
CNNs	nnU-Net [44]	4.53	85.74	4.21	91.15	3.89	80.76	5.47	85.29
	nnU-Net ResEnc [44]	4.12	85.60	3.71	89.93	3.72	80.86	4.93	86.01
	MedNeXt-M-K3 [78]	6.35	85.27	4.59	90.84	6.57	80.88	7.89	84.1
	MedNeXt-M-K5 [78]	6.67	84.79	4.93	88.93	6.75	79.97	8.33	85.47
Transformers	TransUNet [10]	13.18	64.14	14.42	70.16	10.80	54.31	14.31	67.94
	TransBTS [87]	9.83	69.72	10.32	78.22	10.20	57.26	8.97	73.68
	CoTr [91]	9.96	68.21	9.25	74.81	9.58	55.14	11.04	74.67
	UNETR [36]	9.04	70.92	8.03	78.85	9.83	58.06	9.25	75.85
	Swin-UNet [6]	9.98	67.95	8.85	76.43	10.31	57.21	10.77	70.2
	Swin-UNETR [35]	6.77	84.07	7.13	88.92	7.54	79.93	5.63	83.34
	LeViT-UNet-384s [93]	8.56	70.06	8.20	77.06	8.60	58.10	8.89	75.03
	MISSFormer [41]	9.21	83.08	8.40	88.21	9.57	79.71	9.64	81.33
	nnFormer [98]	4.05	86.34	3.47	91.28	4.24	**81.76**	4.46	85.97
Mamba	UMamba Bot [64]	**3.80**	86.35	3.49	92.10	3.80	80.04	4.10	86.9
	UMamba Enc [64]	4.17	86.16	3.63	92.30	4.44	79.72	4.43	86.46
	SegMambaSkip†	4.53	85.25	3.61	92.11	5.43	78.85	4.54	84.79
	SegMamba†	3.82	86.66	3.66	92.26	3.83	80.77	3.96	86.96
	BiSegMamba†	3.85	85.75	**3.38**	**92.43**	**3.46**	79.60	4.70	85.21
	MultiSegMamba†	3.84	**86.70**⋆	3.72	92.09	3.88	80.84	**3.93**	**87.18**

As evident from the figure, all Mamba-based variants perform qualitatively well, but the models leveraging multiple directions (BiSegMamba and MultiSegMamba) are less prone to errors when dealing with fine-grained details and ambiguous boundaries. This confirms the quantitative results: multidirectional processing provides more robust segmentation by ensuring each voxel receives contextual information from multiple spatial arrangements.

5.3.7 Clinical Metrics for Cardiac Segmentation

For cardiac segmentation datasets like ACDC, clinical metrics provide essential validation beyond standard segmentation metrics. Accurate segmentation of the left ventricle enables calculation of the ejection fraction, a critical indicator of cardiac function that quantifies the percentage of blood pumped out with each

Table 5.5: 5-fold cross-validation results on the Synapse Abdomen dataset. Our proposals are marked with †. For space constraints, single class results only report the Dice score. Best results are in **bold** while the second best are underlined. Methods subjected to statistical testing are highlighted in blue.

	Model	Average		Aorta	Gallb.	L.Kidn.	R.Kidn.	Liver	Pancr.	Spleen	Stom.
		HD95↓	DSC↑								
CNNs	nnU-Net [44]	10.91	86.21	91.65	70.01	86.67	85.75	96.11	83.22	90.69	85.55
CNNs	nnU-Net ResEnc [44]	7.70	86.61	89.94	64.20	90.79	91.18	**97.36**	79.48	92.03	**87.93**
CNNs	MedNeXt-M-K3 [78]	18.99	85.71	92.44	**72.75**	87.62	86.21	97.15	81.17	90.30	77.93
CNNs	MedNeXt-M-K5 [78]	17.30	86.00	92.15	71.66	87.89	87.43	96.91	80.26	90.95	80.78
Transformers	TransUNet [10]	32.27	77.24	86.88	62.59	81.35	76.98	94.45	55.57	84.97	75.12
Transformers	TransBTS [87]	11.98	83.27	91.95	62.24	86.91	87.15	96.67	71.91	91.62	77.7
Transformers	CoTr [91]	9.35	84.67	**92.77**	63.07	87.98	86.84	92.75	78.63	94.54	80.76
Transformers	UNETR [36]	19.15	78.1	89.75	55.81	85.71	84.71	94.00	60.23	84.47	70.14
Transformers	Swin-UNet [6]	22.02	79.06	85.65	66.46	83.03	79.37	94.02	56.57	90.67	76.72
Transformers	Swin-UNETR [35]	11.02	83.64	91.22	66.48	87.09	86.62	95.99	68.79	95.72	77.19
Transformers	LeViT-UNet [93]	16.80	78.38	87.52	61.77	84.04	79.87	92.80	59.20	88.84	73.03
Transformers	MISSFormer [41]	18.50	81.87	86.48	68.92	85.56	81.60	94.24	65.44	91.70	80.99
Transformers	nnFormer [98]	11.14	86.56	91.63	69.85	86.61	86.55	96.97	**83.68**	90.72	86.44
Mamba	UMamba Bot [64]	7.35	86.88	89.88	60.14	89.99	94.37	96.81	82.33	95.66	85.88
Mamba	UMamba Enc [64]	7.83	87.82	89.57	65.20	89.46	94.84	96.97	83.35	**96.80**	86.40
Mamba	SegMambaSkip†	6.29	88.26	89.64	69.04	93.40	**94.91**	96.80	79.61	96.45	86.19
Mamba	SegMamba†	7.91	87.48	89.59	62.21	93.65	94.81	96.82	80.72	95.22	86.85
Mamba	BiSegMamba†	5.99	88.29	91.02	70.12	92.98	94.32	96.94	79.08	96.26	85.58
Mamba	MultiSegMamba†	**5.98**	**88.93**	91.36	71.78	**94.00**	94.88	95.76	80.65	96.22	86.77

contraction.

For each patient, two MRI scans are provided: end-diastolic (Dd, when the ventricle is maximally filled) and end-systolic (Ds, after contraction). The ejection fraction is computed as:

$$EF = \frac{Dd - Ds}{Dd} \times 100\% \tag{5.7}$$

Tab. 5.8 compares predicted end-diastolic volumes (Dd) and ejection fractions (EF) against ground truth using Pearson correlation (ρ), mean absolute error (MAE), and bias with standard deviation.

All methods demonstrate exceptional precision in predicting both Dd and EF. For Dd measurements, the Pearson correlation exceeds 0.995 for all models, with MAE values below 7.5 ml (representing a relative error of approximately 3.8% given an average ground truth Dd of 196 ml). For ejection fraction, correlations exceed 0.976 with MAE below 3.2%.

These results confirm that our Mamba-based architectures not only achieve

Table 5.6: 5-fold cross-validation results on the Automatic Cardiac Diagnosis (ACDC) dataset. Our proposals are marked with †. The evaluation metric is the DSC (%). Best results are in bold while the second best are underlined. Methods subjected to statistical testing are highlighted in blue.

	Model	Average	RV	Myo	LV
CNNs	nnU-Net [44]	91.42	90.10	88.74	95.41
	nnU-Net ResEnc [44]	90.84	89.17	88.52	94.84
	MedNeXt-M-K3 [78]	91.64	89.43	89.77	95.72
	MedNeXt-M-K5 [78]	90.70	88.50	88.88	94.73
Transformers	TransUNet [10]	89.75	88.88	84.66	95.70
	TransBTS [87]	91.29	90.42	87.94	95.51
	CoTr [91]	90.90	89.17	88.34	95.18
	UNETR [36]	88.72	85.55	86.48	94.12
	Swin-UNet [6]	89.97	88.29	85.61	**96.01**
	Swin-UNETR [35]	91.36	90.48	87.84	95.75
	LeViT-UNet [93]	90.21	89.78	87.10	93.75
	MISSFormer [41]	87.73	86.55	85.24	91.42
	nnFormer [98]	91.87	**90.78**	89.37	95.46
Mamba	UMamba Bot [64]	90.44	87.67	88.76	94.89
	UMamba Enc [64]	90.07	87.34	88.23	94.65
	SegMambaSkip†	91.49	89.58	89.51	95.39
	SegMamba†	91.33	89.37	89.40	95.22
	BiSegMamba†	91.50	89.46	89.66	95.37
	MultiSegMamba†	**92.04**	90.39	**90.29**	95.44

superior segmentation metrics but also maintain clinical reliability for derived measurements that directly inform patient care decisions.

5.3.8 PosPadUNet3D vs. Mamba Architectures

The two architectural innovations presented in this chapter address complementary challenges with different design philosophies.

PosPadUNet3D is best suited for domain-specific applications where anatomical priors can be explicitly leveraged. The positional encoding enables the model to condition predictions on global anatomical location, while the memory mechanism learns domain-specific concepts that inform interpretation. Although the architecture requires patch-based processing, the combination of positional encoding and memory substantially mitigates the resulting limitations.

Mamba architectures take a more general-purpose approach. They provide linear-complexity alternatives to transformers that scale efficiently to large volumes. Multidirectional processing addresses spatial relationships without domain-specific assumptions. The architectures are applicable across anatomies without requiring modification, making them suitable for settings where domain expertise is limited or where a single model must handle diverse structures.

Table 5.7: Computational comparison on the Synapse dataset. Our proposals are marked with †. The number of parameters is expressed in millions [M] and VRAM in gigabyte [GB]. Training and inference times, expressed in hours [h] and seconds [s], respectively, are obtained on an Nvidia A100 with 80GB of memory. All competitor models were trained for 1000 epochs, as recommended by most of their original papers, while our method achieved convergence in only 300 epochs. Inference time is the average across all test volumes.

	Models	Params	GFLOPs	VRAM	Tr.	Inf.
CNNs	nnU-Net [44]	30.64	410.11	7.65	9.2	21.8
	nnU-Net ResEnc [44]	57.50	502.49	10.00	10.0	22.2
	MedNeXt-M-K3 [78]	32.65	248.03	15.32	67.6	153.6
	MedNeXt-M-K5 [78]	34.75	308.01	18.85	218.3	416.9
Transf.	TransUNet [10]	96.07	88.91	16.25	26.5	73.9
	CoTr [91]	50.12	369.22	8.10	18.6	41.4
	UNETR [36]	92.49	75.76	15.29	15.4	39.5
	Swin-UNETR [35]	62.83	384.20	13.91	22.0	38.7
	nnFormer [98]	150.50	213.41	9.73	8.2	20.6
Mamba	UMamba Bot [64]	41.95	156.32	13.55	22.0	54.2
	UMamba Enc [64]	42.85	231.18	26.42	37.9	89.3
	SegMambaSkip†	62.36	486.92	29.26	12.6	93.5
	SegMamba†	61.49	480.90	25.61	12.7	99.6
	BiSegMamba†	64.75	494.17	27.31	16.5	134.1
	MultiSegMamba†	68.46	527.56	36.92	18.2	149.0

Mamba Architectures:

- General-purpose solution for 3D medical segmentation;
- Linear complexity enables processing larger volumes;
- Multidirectional processing addresses spatial relationships;
- Applicable across anatomies without domain-specific modifications.

For IAC segmentation specifically, PosPadUNet3D achieves superior performance due to its exploitation of anatomical priors. For general 3D segmentation across diverse anatomies, Mamba architectures provide a more versatile solution.

5.3.9 Lessons for Architectural Design

Both approaches share common insights:

1. **Context matters**: Global anatomical context is essential for accurate segmentation;

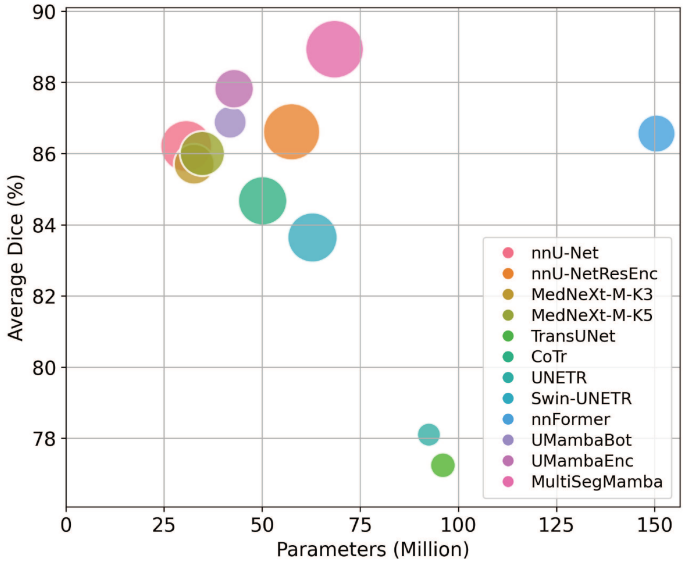


Figure 5.9: Deployment model size and average DSCs across our best model and competitors on Synapse. Circle size indicates GFLOPS.

2. **Position is informative**: Encoding spatial location improves predictions;
3. **Directionality is crucial**: Sequential models require careful handling of processing order.

5.4 Completing the Cycle

Trained segmentation models from both architectures can be integrated into annotation workflows, completing the synergistic cycle. PosPadUNet3D provides automatic IAC segmentation proposals within IACAT, reducing expert effort from manual delineation to verification and correction. Similarly, Mamba models accelerate multi-structure annotation for ToothFairy2 by generating initial predictions that annotators refine. This represents the third link in the cycle: better models enable better tools, which in turn create better datasets for training the next generation of models.

5.5 Summary

This chapter presented two complementary architectural innovations that address fundamental limitations of existing approaches to 3D medical image seg-

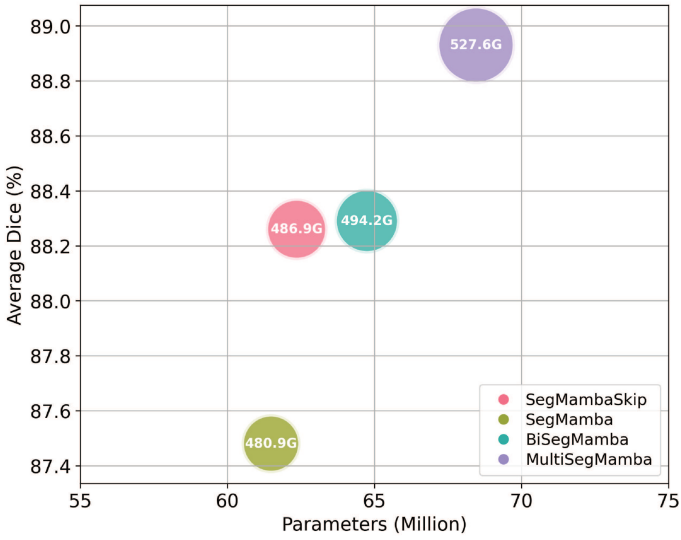


Figure 5.10: Deployment model size and average DSCs across our models on Synapse. Circle size indicates GFLOPS.

mentation.

PosPadUNet3D combines a memory-augmented transformer with absolute positional encoding to address the patch-based learning problem. By enabling the network to condition predictions on global anatomical location and store learned anatomical priors, the architecture achieves state-of-the-art IAC segmentation with +1.6 Dice improvement over the ToothFairy Challenge winner.

The Mamba-based architectures address the complementary challenge of computational efficiency through bidirectional and multidirectional processing strategies that resolve the initial hidden state problem inherent in sequential models. These architectures achieve state-of-the-art performance on BrainTumor, Synapse, and ACDC benchmarks while maintaining linear computational complexity that enables processing of large 3D volumes.

Source code for PosPadUNet3D is publicly available at `https://github.com/AImageLab-zip/ToothFairy`, and Mamba implementations at `https://github.com/LucaLumetti/TamingMambas`.

While these architectural innovations achieve state-of-the-art performance, training a new model from scratch for each clinical task remains computationally expensive and requires substantial annotated data. The next chapter addresses this limitation through model merging strategies, demonstrating how stable pretraining enables efficient task vector arithmetic for modular development of clinical AI systems without full retraining.

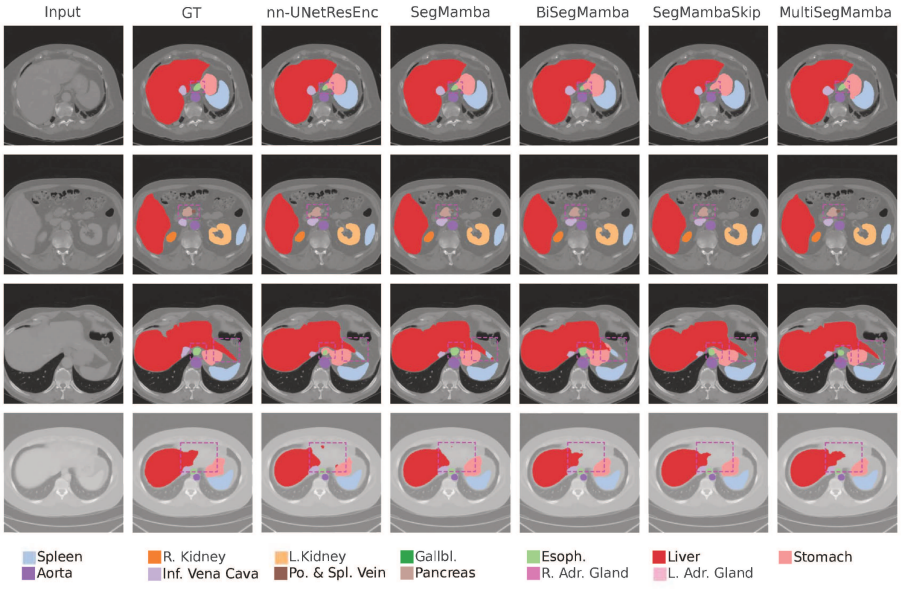


Figure 5.11: Visualization of segmentation results for four sample cases from the Synapse Abdomen evaluation set. Annotation errors are marked with magenta-dashed boxes. The figure is best viewed in color and zoomed in. From left to right: Input, Ground Truth (GT), nnU-Net ResEnc, SegMamba, BiSegMamba, SegMambaSkip, and MultiSegMamba.

Table 5.8: Clinical Metrics of the Left Ventricle for the test set of the ACDC Dataset. Our proposals are marked with †. This table presents a comparative evaluation of the end-diastolic volumes (Dd) and the derived ejection fractions (EF) between the ground truth measurements and the AI predictions. The metrics include the Pearson correlation (ρ), mean absolute error (MAE), and the mean difference with standard deviation (reported as bias $\pm$ σ).

Model	Left Ventricle Dd			Left Ventricle EF		
	ρ	$bias \pm \sigma$ ml	MAE ml	ρ	$bias \pm \sigma$ %	MAE %
nnU-Net ResEnc [44]	99.6%	-3.5±8.6	6.415	99.0%	-1.7±2.8	2.392
nnFormer [98]	99.6%	-2.6±8.0	5.576	99.0%	-1.6±2.7	2.461
SegMambaSkip†	99.6%	-1.8±7.6	5.133	99.0%	-1.8±2.8	2.602
SegMamba†	99.5%	-3.9±8.9	6.771	97.9%	-1.0±4.0	3.058
BiSegMamba†	99.5%	-4.3±8.9	7.451	97.6%	1.5±4.3	3.226
MultiSegMamba†	99.6%	-2.9±8.0	5.936	99.1%	-1.9±2.7	2.496

6. Model Merging and Training Efficiency

This chapter introduces model merging strategies for 3D medical image segmentation, enabling modular development and efficient adaptation of clinical AI systems. We demonstrate that "stable" pretraining—encouraging flat loss landscape minima through hyperparameter selection—dramatically improves merging effectiveness.

6.1 Introduction

Clinical deployment of medical AI faces practical constraints beyond pure performance. Training from scratch for each new task requires substantial computation and labeled data, making it impractical to develop separate models for every clinical application. Model management presents additional challenges, as maintaining separate models for each task creates significant storage and deployment overhead in clinical IT systems. Clinical needs evolve continuously, requiring models that can be updated and extended without full retraining from scratch. Furthermore, privacy regulations may prohibit centralized data collection, necessitating distributed approaches where models are trained locally and then combined.

Model merging offers an elegant solution to these challenges: combine multiple task-specific models into a single unified model without additional training. This chapter explores whether model merging, which has proven successful in 2D natural image classification, transfers to the more challenging domain of 3D medical segmentation.

6.2 Task Vectors and Model Merging

We deal with a neural net $f(\cdot;\boldsymbol{\theta})$ designed for 3D segmentation, like 3D U-Net. The model has weights $\boldsymbol{\theta} \in \mathbb{R}^m$ and takes 3D images as input $\boldsymbol{x} \in \mathbb{R}^{H\times W\times D}$. The output is a 3D map of class distributions $p_\theta(\boldsymbol{y}|\boldsymbol{x})$, one for each voxel $\boldsymbol{y}$ in $\mathcal{Y} \in$

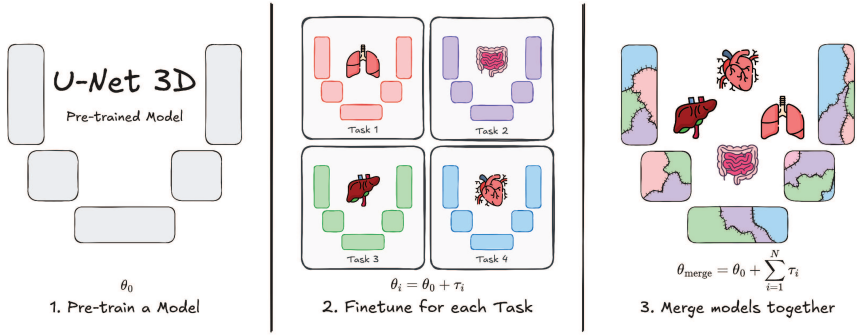


Figure 6.1: Overview of model merging for 3D medical image segmentation. Task-specific models fine-tuned from a shared pretrained checkpoint can be combined through arithmetic operations on their task vectors, creating a unified multi-task model without additional training.

$\mathbb{R}^{H\times W\times D\times C}$. We study a **multi-task learning framework** comprising T segmentation tasks, denoted as $\mathcal{T}$. Each task $t \in \mathcal{T}$ is associated with a dataset $\mathcal{D}_t$ of n_t training samples, sampled from a task-specific distribution $p_t(\boldsymbol{x}, \boldsymbol{y})$. Despite variations in these distributions (*e.g.*, different anatomical parts segmented in each task), all share a common loss function $\ell(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y})$ (*e.g.*, the cross-entropy loss), defined as the negative log-likelihood $\ell(\boldsymbol{\theta}|\boldsymbol{x}, \mathcal{Y}) = -\sum_{\boldsymbol{y}\in\mathcal{Y}} \log p_\theta(\boldsymbol{y}|\boldsymbol{x})$.

Model Merging. To learn multiple segmentation tasks, we consider training a distinct set of weights for each task independently. We organize these models within a pool $\mathcal{P} = \{f(\cdot;\boldsymbol{\theta}_t) \mid \boldsymbol{\theta}_t \triangleq \boldsymbol{\theta}_0 + \boldsymbol{\tau}_t\}_{t\in\mathcal{T}}$ that can be expanded to accommodate for new tasks. Importantly, each model $f(\cdot;\boldsymbol{\theta}_t)$ is initialized from a shared set of **pre-trained weights** $\boldsymbol{\theta}_0$ and fine-tuned for its respective task. The displacement in weight space $\boldsymbol{\tau}_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}_0$ is called *task vector* [42] and, intuitively, it represents a direction in which the loss decreases for the t-th task.

As we discuss further, the models in the pool $\mathcal{P}$ can be selected and combined in arbitrary ways to construct a (personalized) multi-task model. The simplest approach to achieve this is by simply averaging the weights within the pool:

$$f_{\mathcal{P}} \triangleq f(\cdot;\boldsymbol{\theta}_{\mathcal{P}}) \quad \text{s.t.} \quad \boldsymbol{\theta}_{\mathcal{P}} = \boldsymbol{\theta}_0 + \textstyle\sum_{t=1}^{T} w_t \boldsymbol{\tau}_t, \quad \sum_{t=1}^{T} w_t = 1. \tag{6.1}$$

By adjusting the coefficients w_t, we can specialize the merged model toward specific tasks, deprioritizing others. Conversely, for a model that maintains a balance across all tasks, a uniform weighting scheme, $w_t = 1/T$, can be used.

The **goal** is to design an approach that learns and combines multiple 3D segmentation models, ensuring that the resulting merged model performs well across a combined set of tasks. To assess multi-tasking, we define the **empirical**

risk, *i.e.*, the average loss $\hat{\ell}(\boldsymbol{\theta}|\mathcal{D})$ over the union of all training tasks:[1]

$$\hat{\ell}(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{\sum_{t=1}^{T} n_t} \sum_{\boldsymbol{x},\boldsymbol{y} \in \bigcup_{t=1}^{T} \mathcal{D}_t} \ell(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y}) \approx \mathbb{E}_{\substack{t\sim\mathcal{T} \\ \boldsymbol{x},\boldsymbol{y}\sim p_t(\boldsymbol{x},\boldsymbol{y})}} [\ell(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y})] \tag{6.2}$$

Research Question. While 2D image classification tasks can benefit from a variety of pre-trained models (*e.g.*, CLIP and DINO), 3D medical segmentation tasks face the absence of similar pre-trained models. In this respect, *how can we develop pre-trained models for 3D segmentation that facilitate model merging*?

6.2.1 Model Merging from a Pre-training Perspective

Following [75], we analyze model merging through the lens of the Taylor approximation of the loss function. Specifically, we indicate as $\ell_{\text{cur}}(\boldsymbol{\theta})$ the second-order approximation of the empirical risk, centered around the pre-trained weights $\boldsymbol{\theta}_0$:

$$\hat{\ell}_{\text{cur}}(\boldsymbol{\theta}) = \hat{\ell}(\boldsymbol{\theta}_0) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathrm{T}} \nabla \hat{\ell}(\boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathrm{T}} \mathbf{H}_{\hat{\ell}}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0), \tag{6.3}$$

with $\nabla \hat{\ell}(\boldsymbol{\theta}_0) \triangleq \nabla_{\boldsymbol{\theta}} \hat{\ell}(\boldsymbol{\theta}_0)$ and $\mathbf{H}_{\hat{\ell}}(\boldsymbol{\theta}_0) \triangleq \nabla^2_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_0)$ indicating the gradient and the Hessian around $\boldsymbol{\theta}_0$. Based on [75], assuming that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ is a local minimum for the empirical risk $\hat{\ell}(\boldsymbol{\theta})$ across all tasks, the Hessian is positive semi-definite. It follows that the second-order approximation $\hat{\ell}_{\text{cur}}(\boldsymbol{\theta})$ of the empirical risk is locally convex. Utilizing Jensen's inequality (valid for convex functions) we can establish the following relationship between the merged model and the individuals:

$$\hat{\ell}_{\text{cur}}(\boldsymbol{\theta}_{\mathcal{P}} = \boldsymbol{\theta}_0 + \textstyle\sum_{t=1}^{T} w_t \boldsymbol{\tau}_t) \leq \sum_{t=1}^{T} w_t \hat{\ell}_{\text{cur}}(\boldsymbol{\theta}_t = \boldsymbol{\theta}_0 + \boldsymbol{\tau}_t). \tag{6.4}$$

This inequality is informative because the term on the right offers a kind of worst-case upper limit for the performance of the merged model. In particular, the empirical risk $\hat{\ell}_{\text{cur}}(\boldsymbol{\theta}_{\mathcal{P}})$ of the merged model is constrained by the convex combination of the empirical risks associated with each individual model. This implies that if each individual model $\boldsymbol{\theta}_t$ performs accurately across all tasks, there are certain assurances regarding the risk level of the merged model $\boldsymbol{\theta}_{\mathcal{P}}$.

However, the issue with Eq. (6.4) is that, under a scenario with specialized models trained on separate tasks, we cannot ensure that each individual model $\boldsymbol{\theta}_t$ performs well across all tasks. Indeed, as $\boldsymbol{\theta}_t$ is trained exclusively on its specific distribution $p_t(\boldsymbol{x},\boldsymbol{y})$, its empirical risk is likely high for other data distributions $p_{t'\neq t}(\boldsymbol{x},\boldsymbol{y})$ ($\rightarrow$ **low out-of-distribution performance**). For this reason, the following augmented optimization problem was proposed [75] for the t-th learner:

$$\underset{\boldsymbol{\theta}_t}{\text{minimize}} \quad \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim p_t(\boldsymbol{x},\boldsymbol{y})} [\ell_{\text{cur}}(\boldsymbol{\theta}_t|\boldsymbol{x},\boldsymbol{y})] + \mathcal{D}_{\text{KL}}(p_{\boldsymbol{\theta}_0}(\boldsymbol{y}|\boldsymbol{x})||p_{\boldsymbol{\theta}_t}(\boldsymbol{y}|\boldsymbol{x})). \tag{6.5}$$

[1] To simplify the notation, we will no longer explicitly denote the dependence of the loss on the data and write the individual loss and the empirical risk as $\ell(\boldsymbol{\theta})$ and $\hat{\ell}(\boldsymbol{\theta})$.

Table 6.1: Stable *vs.* plastic training regimes, metrics, and corresponding hyperparameters: Batch size (BS), Dropout (DO), and Learning rate (LR). λ_i correspond to the eigenvalues of $\mathbf{H}_{\hat{\ell}}(\boldsymbol{\theta}_0)$.

Regime	Dataset	BS	DO	LR	Dice↑	$\sum \boldsymbol{\lambda_i}\downarrow$	$\boldsymbol{\lambda_1}\downarrow$
Stable	Cui	4	0.5	10^{-3}	34.93	0.57	0.02
Plastic		8	0.0	10^{-4}	42.68	40.71	6.00
Stable	AMOS	4	0.5	10^{-3}	43.76	2.30	0.03
Plastic		8	0.0	10^{-4}	46.87	58.46	0.05

In essence, the out-of-distribution performance of each model is preserved through additional regularization provided by the term $\mathcal{D}_{\mathrm{KL}}(\cdot)$, which acts explicitly on out-of-distribution examples $\boldsymbol{x}, \boldsymbol{y} \sim p_{t' \neq t}(\boldsymbol{x}, \boldsymbol{y})$. The $\mathcal{D}_{\mathrm{KL}}(\cdot)$ term **aligns** the predictions $p_{\boldsymbol{\theta}_t}(\boldsymbol{y}|\boldsymbol{x})$ of the individual model $f(\cdot; \boldsymbol{\theta}_t)$ to those generated by the pre-trained model $\boldsymbol{\theta}_0$. By doing so, the individual model can achieve at least the performance level of the pre-trained model on external distributions $p_{t' \neq t}(\boldsymbol{x}, \boldsymbol{y})$, effectively reducing the upper bound on the right side of Eq. (6.4).

6.2.2 The Role of the Training Regime of the Pre-trained Model

While the authors of [75] drew inspiration from Eq. (6.5) to design a data-free regularization term, we take a different approach that avoids introducing explicit regularization. Instead, we focus on analyzing the roles of the training regime of the pre-trained model.

Thesis. We hypothesize that the tendency of the fine-tuned model $\boldsymbol{\theta}_t$ to retain pre-training knowledge is linked to the curvature of the pre-trained point $\boldsymbol{\theta}_0$ within the landscape of the empirical risk $\hat{\ell}(\cdot)$. To show that, we approximate the $\mathcal{D}_{\mathrm{KL}}(\cdot)$ as in [9]: if $\boldsymbol{\theta}_t - \boldsymbol{\theta}_0 \to 0$, the $\mathcal{D}_{\mathrm{KL}}(\cdot)$ term is close to the *distance* between $\boldsymbol{\theta}_t$ and the pre-training weights $\boldsymbol{\theta}_0$:

$$\mathcal{D}_{\mathrm{KL}}(p_{\boldsymbol{\theta}_0}(\boldsymbol{y}|\boldsymbol{x}) \;||\; p_{\boldsymbol{\theta}_t}(\boldsymbol{y}|\boldsymbol{x})) \approx \tfrac{1}{2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_0)^{\mathrm{T}} \mathbf{H}_{\hat{\ell}}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_t - \boldsymbol{\theta}_0). \tag{6.6}$$

The weight distance is not isotropic but instead influenced by the Hessian of the empirical risk evaluated at $\boldsymbol{\theta}_0$. Thanks to Eq. (6.6) and the positive semi-definiteness of the Hessian around $\boldsymbol{\theta}_0$, we can establish a **bound** on $\mathcal{D}_{\mathrm{KL}}(\cdot)$:

$$\mathcal{D}_{\mathrm{KL}}(\dots) \approx \frac{1}{2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_0)^{\mathrm{T}} \mathbf{H}_{\hat{\ell}}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_t - \boldsymbol{\theta}_0) \leq \frac{1}{2}\lambda_1 \left\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\right\|^2 = \frac{1}{2}\lambda_1 \left\|\boldsymbol{\tau}_t\right\|^2, \tag{6.7}$$

where λ_1 is the **maximum eigenvalue** of the Hessian $\mathbf{H}_{\hat{\ell}}(\boldsymbol{\theta}_0)$ around the pre-training weights. The result is that the degradation in out-of-distribution performance relative to the pre-trained model is controlled by: *i)* the norm of the task vector, and *ii)* the maximum eigenvalue λ_1 of the Hessian. Notably, the entire spectrum of eigenvalues has been crucial in analyzing the geometry of the

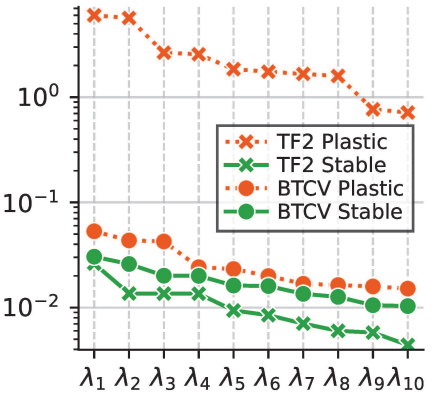


Figure 6.2: Top 10 eigenvalues (↓) for each experiment. Each eigenvalue represents a direction in the loss landscape, with larger values indicating steeper changes in loss. Sorting them reveals the directions with the steepest loss gradients. Stable training (green) produces a substantially flatter loss landscape around the pretrained weights compared to plastic training (orange).

loss landscape and its impact on generalization capabilities [22, 50]. Moreover, the maximum eigenvalue has been extensively used to characterize the width of a local minima [39, 50, 69]. In particular, a larger maximum eigenvalue suggests that the loss landscape is steeper along at least one dimension, which corresponds to a *sharper minimum*. Conversely, smaller eigenvalues suggest *wider minima* because the surface of the loss function changes less drastically in those directions. Hence, to sum up, for a fixed task vector $\boldsymbol{\tau}_t$, the wider the curvature of the pre-trained model, the lower the loss in out-of-distribution performance during fine-tuning, and the better fine-tuned individual models will merge.

6.2.3 Biasing the Base Pre-Trained Model Towards Wide Minima

Building on this analytical finding, we propose modifying the training regime of the base pre-trained model to bias optimization toward wider minima. To do so, the approach is simple: inspired by [69], we act on some key hyperparameters—like batch size, dropout, and learning rate—that have been shown to affect generalization and the geometry of the minimum [26, 57, 92]. Following the terminology in [69], we define two distinct pre-training regimes, namely *stable* (wide minima) *vs. plastic* (sharp minima). The *stable* pre-training regime employs a small batch size, a higher learning rate, and increased dropout. In contrast, the *plastic* pre-training follows conventional self-supervised learning best practices, including as large as possible batch size, no dropout, and lower learning rates.

Table 6.2: Details of the datasets used in our experiments. Data is not resampled, but it is preprocessed with z-score normalization and patch-based training.

Dataset	Modality	Volumes	Structs	Shape
AMOS [47] (**pre-training**)	CT	240	15	$148 \times 533 \times 560$
BTCV Abdomen [56]		30	13	$125 \times 512 \times 512$
Cui [18] (**pre-training**)	CBCT	151	42	$322 \times 402 \times 402$
ToothFairy2 [5]		480	42	$169 \times 356 \times 375$

To analyze the effects of these hyperparameters, a preliminary result is reported in Tab. 6.1. We pre-train two base models (the one within the stable regime and the other in the plastic one) on two datasets for 3D medical image segmentation, namely AMOS [47] and Cui [18]. We then evaluate the average Dice on the corresponding test sets and compare the Hessian's eigenvalues as a proxy for the width of the pre-training optimum. Following [8], the Hessian's eigenspectrum is calculated with the trace of the empirical Fisher Information Matrix (FIM) [52], as a (diagonal) approximation of the intractable Hessian. As observed, the performance of the two base models (stable *vs.* plastic) is comparable across both datasets; however, the stable model achieves a remarkably lower trace (Fig. 6.2). This indicates that manipulating hyperparameters is a simple yet effective way to influence the geometry of the solution attained by the pre-trained model.

6.3 Experiments and Results

Datasets and Task Splits. Considering four public datasets, we categorize experiments into two settings based on the target anatomical regions: *i) abdominal datasets* (AMOS [47] and BTCV Abdomen [56]) and *ii) maxillofacial datasets* (Cui [18] and ToothFairy2 [5, 62]). The summary characteristics are provided in Tab. 6.2. In the abdominal scenario, we use AMOS for pre-training and four BTCV classes (Liver, Spleen, Kidney, and Stomach) to create four tasks. In the maxillofacial scenario, we use Cui for pre-training and ToothFairy2 for fine-tuning, with four tasks based on Mandible, Pharynx, Teeth, and Canals.

Training. We perform stable and plastic pre-training for both AMOS and Cui according to the setup in Tab. 6.1. To perform fine-tuning, we replace the final $1 \times 1 \times 1$ convolution with a new one; for the rest of the layers, we fine-tune the corresponding parameters $\boldsymbol{\theta}_0$ through a task vector $\boldsymbol{\tau}_t$ (initialized at zero). We optimize with *AdamW* [61] and a weight decay penalty of 0.1 to discourage large task vector norms. Training runs for 10 epochs.

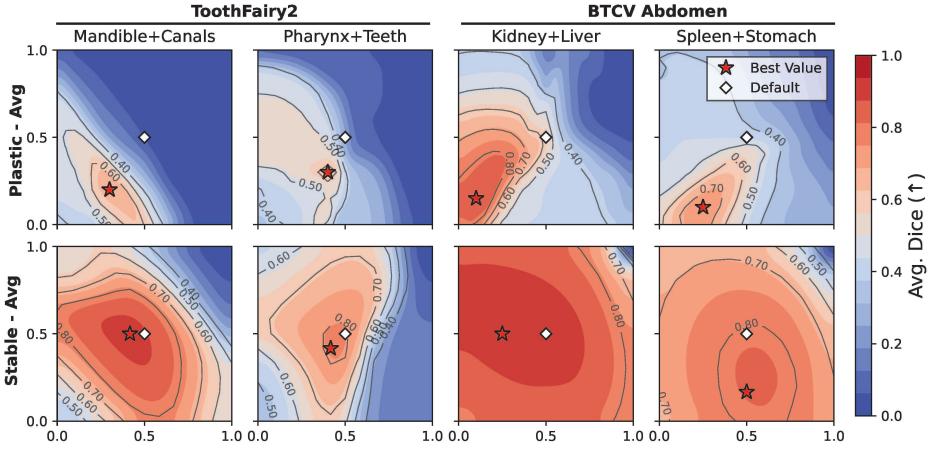


Figure 6.3: Effect of task vector weights on merged model performance. Each heatmap shows average Dice score when combining task vectors τ_1 and τ_2 with weights w_1 (x-axis) and w_2 (y-axis). The star marks the optimal weight combination; the diamond indicates default equal weights. Top row: plastic (standard) pretraining produces sharp minima where merged performance degrades rapidly. Bottom row: stable (SAM-based) pretraining produces flat minima where performance remains high across a wide range of weight combinations.

6.3.1 Impact of Pre-Training Regime on Model Merging

In each plot of Fig. 6.3, we consider a pair of tasks (*e.g.*, Mandible + Canals) and evaluate the Dice score of the merged model while varying merging coefficients w_1 and w_2. By comparing plastic (first row) *vs.* stable (second row) pre-training, we can say that stable pre-training allows for remarkably robust performance, exhibiting lower sensitivity to the merging coefficients—a feature that, in real-world applications, reduces the overhead associated with hyperparameter tuning. As further proof, for the stable regime the uniform weighing scheme ◇ ($w_{1,2} = 0.5$) is always closer to the best configuration ★ (found by hyperparameter tuning on a held-out set).

After examining a scenario where pairs of tasks are merged, we extend our analysis to a setting with four task vectors. We report in Fig. 6.4 the results (Dice score) on each task separately and also the average (**Overall**). Beyond comparing stable *vs.* plastic pre-training, we also examine their impact on TIES Merging [94], a well-established alternative to uniform averaging. The results in Fig. 6.4 show that, in both settings, the performance of the merged model is primarily influenced by the type of pre-training rather than the merging method. This is evidenced by the significant performance gains achieved with stable pre-training (*e.g.*, with uniform averaging, yielding an improvement of +18.60 on ToothFairy2 and +16.28 on BTCV).

Table 6.3: Performance scores obtained from pairwise task vector merging.

Dataset	w	Merging Strategy	Spl. Kid.	Spl. Liv.	Spl. Sto.	Kid. Liv.	Kid. Sto.	Liv. Sto.	Avg.
BTCV Abdomen	Default ◇	Average [42]	91.41	92.18	80.61	90.85	77.18	80.91	85.52
		TIES [94]	82.80	90.76	76.83	88.69	58.56	76.69	79.05
	Best ★	Average [42]	92.64	92.22	82.09	91.01	78.97	81.80	86.45
		TIES [94]	92.42	91.88	81.55	91.07	77.72	81.04	85.95
	-	Joint	91.40	93.34	78.38	92.31	91.79	88.86	89.35

Dataset	w	Merging Strategy	Mand. IACs	Mand. Teeth	Mand. Phar.	IACs Teeth	IACs Phar.	Teeth Phar.	Avg.
ToothFairy2	Default ◇	Average [42]	89.54	82.55	87.70	56.08	57.08	81.27	75.70
		TIES [94]	88.70	79.89	88.96	58.51	63.44	73.72	75.54
	Best ★	Average [42]	89.67	82.78	91.89	68.16	75.19	81.27	81.49
		TIES [94]	89.24	82.33	91.97	67.94	68.91	81.07	80.24
	-	Joint	98.75	97.33	98.26	83.04	97.10	93.61	94.68

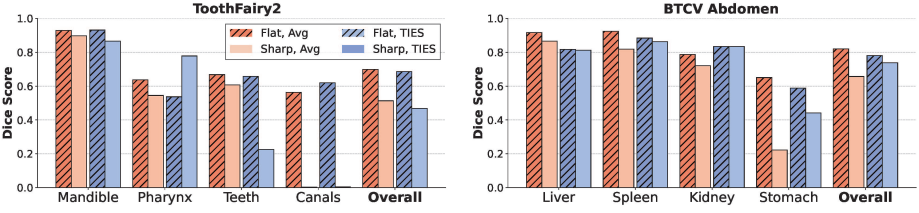


Figure 6.4: Task-wise performance after merging four distinct task vectors with weight averaging and TIES. The Overall bars aggregate results across tasks.

Further Comparative Analysis. To assess the effectiveness of model merging for 3D segmentation, we include a reference approach representing re-training from scratch, where the pre-trained model is fine-tuned on both classes jointly. As shown in Tab. 6.3, in BTCV Abdomen, Kidney+Stomach shows the largest drop w.r.t. the joint training ($\sim$ 18.60 Dice score), while other pairs achieve similar performance, indicating effective merging. In contrast, the gap is significantly larger in ToothFairy2, likely due to greater variation in the shape, size, and intensity values of maxillofacial structures. We conjecture that such an increased variability leads to higher interference when merging the relative task vectors.

7. Extending to Diverse Modalities and Tasks

This chapter demonstrates that the methods developed for CBCT analysis generalize to diverse imaging modalities and clinical tasks. We present applications to testicular ultrasound classification and occlusal analysis from intra-oral scans, validating the broad applicability of the synergistic cycle framework.

7.1 Introduction

The preceding chapters developed tools, datasets, and models in the context of maxillofacial CBCT imaging. A natural question arises: do these contributions generalize beyond the primary domain?

This chapter addresses generalization along multiple axes. First, we explore a different imaging **modality**, moving from CT/CBCT to ultrasound, which involves fundamentally different physics and artifact characteristics. Second, we shift from segmentation to classification as the target **task**, requiring different output structures and evaluation paradigms. Third, we transition from volumetric data to 3D meshes and point clouds, demonstrating flexibility across data **representations**.

7.2 Testicular Ultrasound Classification

7.2.1 Clinical Context

Testicular ultrasound (TUS) imaging is a key non-invasive tool for evaluating male reproductive health by assessing tissue characteristics such as parenchymal inhomogeneity, an emerging biomarker for male infertility. However, subjective image interpretation and complex tissue patterns hinder reliable, standardized assessment, highlighting the need for automated classification tools.

Progress in this area is hampered by the lack of large, publicly available datasets. Ethical and privacy concerns limit data sharing, resulting in small,

institution-specific datasets that constrain deep learning model development and hinder generalization. Label noise compounds this difficulty, as the subjective nature of echotexture interpretation introduces inconsistency in annotations across different radiologists. Furthermore, a substantial domain gap exists between general image datasets and ultrasound imaging, making transfer learning from pretrained models suboptimal without careful adaptation.

7.2.2 Dataset and Label Filtering

All experiments are conducted on an in-house dataset collected at the Antonio Nalin Center of Baggiovara Hospital in Modena, Italy, using Esaote MyLab25 Gold and MyLab XPro80 ultrasound systems. The unlabeled dataset comprises 25,792 images including 1,664 testicular scans and 24,126 thyroid scans used for pretraining. Additionally, 880 testicular images from 220 patients have inhomogeneity/homogeneity labels available, with approximately 20-80% class distribution.

A significant challenge is the inherent noisiness in image-label pairing. Ultrasound examinations are dynamic, with clinicians relying on real-time video evaluation, but only static screenshots are saved during clinical practice. These images may not accurately reflect actual homogeneity characteristics, introducing label noise.

To address this, we developed an automatic filtering procedure. Using a three-fold cross-validation schema, we train a ResNet-18 classifier and record per-sample loss values across epochs. Samples flagged as "suspicious" when their training loss exceeds 1.0 on at least three occasions are either discarded or have their labels flipped. This evaluation is repeated with different pretrained initializations across multiple random seeds, and samples consistently flagged across all evaluations (72 images total) were corrected, yielding improved classification performance.

7.2.3 Pretraining Strategies

For the classification task, we employ a ResNet-18 backbone, as more complex architectures tend to overfit given the limited labeled dataset size. We explore effective pretraining approaches combining supervised and self-supervised learning.

The pretraining objective combines contrastive learning with supervised classification. The contrastive component uses the SimCLR framework, maximizing agreement between differently augmented views of the same image while minimizing agreement between different images. Simultaneously, a supervised classification head predicts the anatomical region (thyroid or testicle), providing additional learning signal. The final pretraining loss combines these components with a weighting factor of 0.2 for the supervised term.

Table 7.1: TUS classification results by pretraining strategy (three-fold cross-validation).

Pretraining	**Accuracy (%)**	**F1-Score (%)**
None (train from scratch)	73.93 ± 6.80	56.05 ± 7.11
ImageNet	83.89 ± 2.15	67.89 ± 1.85
USCL (ultrasound-specific)	75.46 ± 2.64	57.43 ± 1.63
UD (testicles only)	81.09 ± 2.95	65.44 ± 2.32
UD (thyroids only)	80.13 ± 2.83	63.92 ± 2.46
UD (combined)	$\mathbf{86.78 \pm 2.21}$	$\mathbf{73.17 \pm 1.55}$

We compare multiple pretraining sources: ImageNet provides general visual features from natural images; USCL provides ultrasound-specific features from lung and liver imaging; and our unlabeled dataset provides domain-matched features from the same acquisition systems.

7.2.4 Synthetic Data Generation

To address data scarcity while respecting privacy, we explore synthetic data generation using Denoising Diffusion Probabilistic Models (DDPMs). The diffusion model operates through forward and reverse phases: the forward phase incrementally adds Gaussian noise to images over multiple timesteps, while the reverse phase trains a U-Net to reconstruct original images by progressively removing noise.

Quality assessment of generated images employs three metrics: improved precision measuring the fraction of synthetic samples falling within the real data manifold; recall measuring coverage of the real data distribution; and Fréchet Inception Distance (FID) capturing both fidelity and diversity.

We developed a filtering method based on precision to ensure synthetic image quality. Computing the real data manifold from feature embeddings of real images, we retain only synthetic images whose embeddings lie within this manifold. This filtering increased precision from 79.68% to 90.1% while reducing the dataset from approximately 20,000 to 9,000 samples.

7.2.5 Results

The results in Tab. 7.1 reveal several important findings. Pretraining consistently improves performance compared to training from scratch, with accuracy gains of approximately 10 percentage points. Interestingly, ultrasound-specific pretraining from different acquisition systems (USCL) does not outperform ImageNet, suggesting that domain mismatch between acquisition systems can be more detrimental than the modality gap between natural and ultrasound images.

When images originate from the same acquisition system, anatomical region matters less—pretraining on testicles versus thyroids yields similar downstream performance on testicular classification. The best results emerge from combining both anatomical regions during pretraining, enabling the integration of contrastive and supervised losses and yielding +2.89 accuracy and +5.28 F1-score improvements over ImageNet.

Additional experiments confirm that synthetic data can effectively substitute real images during pretraining. Models pretrained on filtered synthetic data achieve competitive performance, demonstrating the practical utility of synthetic data when access to real data is restricted.

7.2.6 Implications

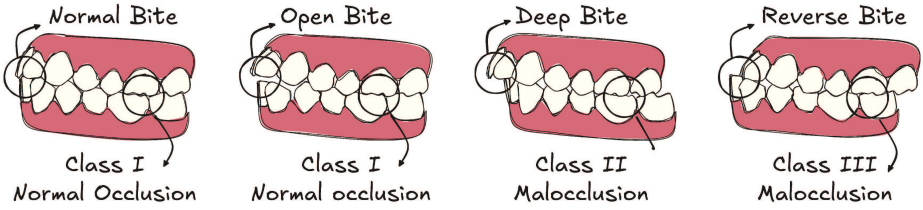
Synthetic data generation offers a promising path to privacy-preserving medical AI development. Rather than sharing sensitive patient data, institutions can share trained generative models that capture statistical properties without exposing individual patient information. This approach generalizes to other sensitive imaging domains where data sharing is restricted, extending the synergistic cycle by establishing synthetic data generation as a "tool" that enables dataset creation without privacy concerns.

7.3 Occlusal Classification from Intra-Oral Scans

7.3.1 Clinical Context

Orthodontic treatment planning requires comprehensive assessment of dental occlusion—how upper and lower teeth meet when the jaw closes. Traditional assessment follows the Angle classification system and evaluates multiple dimensions of tooth alignment. The sagittal dimension characterizes the anteroposterior relationship between dental arches, classified as Class I (normal), Class II (mandibular retrusion), or Class III (mandibular protrusion). The vertical dimension assesses overbite severity, ranging from deep bite (excessive overlap) through normal to open bite (insufficient overlap). The transverse dimension evaluates crossbite patterns, which may be bilateral, unilateral, or absent. Finally, midline assessment determines whether the dental midlines are centered or deviated.

Manual assessment of these characteristics is inherently subjective and time-consuming, contributing to inter-clinician variability in treatment planning. The advent of intra-oral scanners now provides high-resolution 3D surface models of dental arches, creating opportunities for automated, objective analysis.



(a) (b) (c) (d)

Figure 7.1: Occlusion types in the Bits2Bites dataset. (a) Normal Class I occlusion with proper tooth alignment. (b) Open bite with anterior vertical gap despite normal sagittal relationship. (c) Deep bite with excessive vertical overlap (Class II malocclusion). (d) Reverse bite with lower teeth protruding anteriorly (Class III malocclusion).

7.3.2 The Bits2Bites Dataset

We introduce Bits2Bites, the first publicly available dataset for occlusal classification from intra-oral scans. The dataset comprises 200 pairs of registered upper and lower dental arches in STL format, with separate high-resolution meshes spatially aligned to preserve the true occlusal relationship. All scans are transformed to a shared RAS (Right-Anterior-Superior) reference frame. Meshes average 92,201 vertices and 182,444 faces, with bounding-box dimensions of approximately 66mm width, 54mm depth, and 18mm height.

Scans were acquired using two different intra-oral scanners, Carestream and 3Shape TRIOS, to capture variability in acquisition technologies. The scans were selected randomly without filtering criteria to reflect natural clinical diversity.

Annotations were performed by an orthodontic specialist with five years of clinical experience. Each scan pair includes detailed occlusion labels across multiple dimensions: sagittal classifications (Class I, Class II edge-to-edge, Class II full, Class III) provided separately for left and right sides; vertical anterior-posterior relationships (normal, deep bite, reverse bite, open bite); transverse relationships (normal, crossbite, scissor bite); and midline alignment (centered, deviated). This multi-label annotation scheme enables clinically meaningful classification across sagittal, vertical, and transverse planes.

For Bits2Bites, both the source-code of the model[1] and the dataset[2] are publicly available.

[1] https://github.com/AImageLab-zip/Bits2Bites
[2] https://ditto.ing.unimore.it/bits2bites

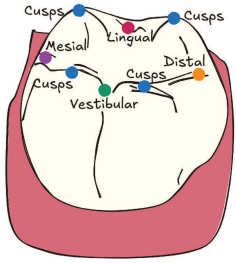


Figure 7.2: Anatomical landmarks annotated for each tooth in the Bits2Bites dataset. Each tooth is labeled with cusps (occlusal surface peaks), mesial and distal contact points (proximal surfaces), lingual (tongue-facing) and vestibular (cheek-facing) surfaces. These landmarks enable precise quantification of occlusal relationships.

Table 7.2: Bits2Bites macro-averaged F1-score by learning strategy (5-fold cross-validation).

Model	Strategy	F1-Score	Time (s)
PointTransformerV3	Single-Task (STL)	0.64 ± 0.13	1.10
PointTransformerV3	Multi-Task (MTL)	0.63 ± 0.03	0.11
SPUNet	Single-Task (STL)	0.60 ± 0.13	0.50
SPUNet	Multi-Task (MTL)	0.58 ± 0.05	0.05

7.3.3 Multi-Task Learning Framework

We formulate occlusal classification as a multi-task learning problem to leverage inherent relationships between different occlusal traits. The framework evaluates two state-of-the-art point cloud processing backbones: PointTransformerV3 and SPUNet. Both backbones process raw 3D point clouds directly, with task-specific classification heads branching from the shared encoder for each of the five occlusal tasks.

An ablation study on input features compares three configurations: raw 3D mesh only, automatically-predicted landmark coordinates only, and combined mesh plus landmark features. The combined approach yields best performance, though the landmark-only models are remarkably efficient—training takes approximately 15 minutes compared to over 2 hours for full mesh processing.

7.3.4 Results

The experimental results in Tab. 7.2 demonstrate a clear trade-off between performance and efficiency. Single-task learning, with dedicated models for each

Table 7.3: Comparison of imaging modalities in this thesis.

	CBCT	**Ultrasound**	**Intra-oral**
Data format	3D volume	2D image	3D mesh
Resolution	0.3–0.4 mm	Variable	<0.1 mm
Contrast	Good (bone)	Low	Surface only
Artifacts	Metal, motion	Speckle, shadows	Occlusion, reflections
Task	Segmentation	Classification	Classification

task, achieves marginally higher F1-scores but requires training and maintaining five distinct models per backbone. Multi-task learning provides an efficient and scalable solution with 10× faster inference time, well-suited for clinical applications where speed is critical.

Per-task analysis reveals that anterior bite (vertical) classification achieves highest F1-scores around 0.74–0.77, while midline assessment proves most challenging at 0.46–0.49. PointTransformerV3 consistently outperforms SPUNet across configurations, establishing it as the more robust architecture for this domain.

7.3.5 Clinical Applicability

Automated occlusal classification has significant potential for clinical practice. Objective, reproducible assessment would reduce inter-clinician variability that currently affects treatment planning consistency. Rapid automated screening could streamline orthodontic referral workflows, allowing general dentists to efficiently identify patients requiring specialist consultation. Quantitative tracking of occlusal traits throughout treatment would enable precise monitoring of progress, supporting evidence-based adjustments to treatment protocols.

7.4 Cross-Domain Analysis

7.4.1 Comparing Modalities

Despite the substantial differences between imaging modalities summarized in Tab. 7.3, common principles emerge that transcend specific domains. Public datasets consistently accelerate research progress regardless of modality, enabling reproducible benchmarking and lowering barriers to entry for new research groups. Pretraining—whether supervised, self-supervised, or synthetic—consistently improves performance across all applications, confirming that learned representations transfer even across significant domain gaps. Multi-task learning proves effective whenever related prediction targets share underlying structure, as demonstrated in both CBCT segmentation and occlusal classification.

7.4.2 The Synergistic Cycle in New Domains

Both applications presented in this chapter demonstrate the synergistic cycle operating in new domains. For testicular ultrasound, diffusion models serve as tools that generate synthetic datasets, which in turn train classification models. These trained classifiers could eventually provide pseudo-labels for unlabeled data, further expanding the training corpus. Future work will investigate conditioning the generative process to produce synthetic images with explicit homogeneity labels, enabling direct fine-tuning on generated data.

For Bits2Bites, the public dataset enables model development by the broader research community. As trained models mature, they could assist annotation for future data collection, reducing the annotation burden and enabling dataset expansion. The multi-label annotation scheme and standardized coordinate frame facilitate direct comparison of methods and foster reproducible research in automated orthodontic diagnosis.

7.5 Summary

This chapter demonstrated the broad applicability of thesis contributions across diverse imaging modalities and clinical tasks. The testicular ultrasound application showed that synthetic data generation enables privacy-preserving AI development, with combined pretraining on real ultrasound data achieving 86.78% accuracy and 0.73 F1-score. The occlusal classification work introduced the first public dataset for this task and demonstrated that point cloud architectures achieve F1-scores of 0.63–0.64 across five occlusal traits. Cross-domain analysis revealed consistent patterns: public datasets accelerate research, pretraining consistently helps, and multi-task learning leverages shared structure regardless of the specific imaging modality.

These results validate the synergistic cycle framework as a general paradigm for medical AI development, applicable beyond the primary CBCT domain to any imaging context where tools, datasets, and models can mutually reinforce progress.

8. Discussion

Working in an under-researched domain naturally led to exploiting a synergistic cycle where improvements in tools, datasets, and models mutually reinforce each other. Rather than proposing this as a theoretical framework, this thesis demonstrates how such a cycle emerged organically from practical necessity and proved instrumental in advancing maxillofacial AI. This chapter reflects on the broader implications of this work, examining recurring themes, discussing impact, acknowledging limitations, and outlining directions for future research.

8.1 The Synergistic Cycle in Action

The contributions described in preceding chapters illustrate a self-reinforcing cycle that accelerates progress through three key mechanisms:

Infrastructure investment compounds returns. IACAT's spline-based annotation workflow reduced annotation time by an order of magnitude, transforming infeasible projects into tractable ones. This single tool enabled the ToothFairy datasets, which in turn enabled architectural innovations and attracted international collaboration. The thousands of hours saved through efficient annotation represent value that propagates throughout the research pipeline.

Public resources amplify community impact. The ToothFairy datasets have been downloaded over 6,000 times and cited in dozens of papers. Challenge participation from over 50 international teams brought diverse approaches that no single research group could generate. Standardized benchmarks replaced unverifiable claims on private data with reproducible comparisons, allowing the community to identify which innovations genuinely advance the field. Open science transforms individual contributions into community resources whose scientific value extends far beyond the original creators.

AI assistance creates positive feedback loops. Trained models integrated into annotation tools (IACAT 2.0, ToothFairy4M) fundamentally change the annotator's role from creator to reviewer. Expert effort shifts from laborious delineation to efficient verification, dramatically increasing throughput while maintaining quality. The resulting higher-quality annotations improve future model generations, completing the cycle: better models make better tools, which

make better datasets, which train even better models.

8.2 Recurring Themes

Several themes recur across chapters:

8.2.1 Open Science Accelerates Progress

Every major contribution in this thesis has been publicly released. The IA-CAT source code enables other research groups to annotate their own datasets using the same efficient workflows. The ToothFairy and ToothFairy2 datasets provide standardized benchmarks for fair comparison. The Mamba architecture implementations allow direct reproduction and extension of our results. The Bits2Bites and TesticulUS benchmarks extend the open science philosophy to new modalities. Even the challenge evaluation infrastructure has been made available to support ongoing benchmarking.

This openness has enabled reproduction, extension, and improvement by the community. The scientific value of an unreleased dataset or closed-source model is fundamentally limited—claims cannot be verified, methods cannot be compared, and progress cannot accumulate. Open science transforms individual contributions into community resources that benefit the entire field.

8.2.2 Efficiency Matters

Clinical deployment requires not just accuracy but efficiency. Mamba's linear complexity enables processing large volumes that would exhaust GPU memory with the quadratic complexity of transformer attention mechanisms. Model merging fundamentally transforms how clinical AI systems are developed and maintained: instead of training separate models for each anatomy or retraining from scratch when adapting to new structures, practitioners can combine task-specific vectors to create multi-task models without prohibitive computational costs. This modular approach reduces both the initial development burden and the ongoing maintenance overhead as clinical needs evolve. Crucially, stable pre-training produces better pretrained models that serve as robust starting points for subsequent specialization, enabling rapid iteration and adaptation. Synthetic data generation enables privacy-preserving development by allowing models to be trained on artificially generated examples that capture statistical properties without exposing real patient information.

These efficiency considerations are often secondary in academic research, where access to computational resources is assumed and privacy concerns are deferred to future work or solved by not publishing the data. In clinical translation, however, efficiency becomes primary: models must run on available hardware,

updates must be deployable without prohibitive retraining costs, and development must proceed without compromising patient privacy.

8.2.3 Pretraining Bridges Data Gaps

Pretraining consistently improves performance across the domains explored in this thesis. Stable pretraining—using hyperparameters that encourage convergence to flat minima—enables effective model merging, allowing task vectors to be combined without destructive interference. Synthetic pretraining on diffusion-generated images matches real data for testicular ultrasound classification, demonstrating that learned representations transfer even when training examples are artificially created. Domain-specific pretraining on unlabeled ultrasound data outperforms ImageNet transfer, highlighting the value of in-domain exposure even without task-specific labels.

These findings suggest that **how we pretrain may matter as much as what architecture we use**. The choice of pretraining strategy—stable versus plastic optimization, real versus synthetic data, generic versus domain-specific sources—can determine whether downstream models succeed or fail, independent of architectural sophistication.

8.3 Broader Impact

8.3.1 Clinical Applications

The methods developed in this thesis have direct clinical applicability across multiple dental and maxillofacial subspecialties. In surgical planning, automated IAC segmentation identifies the neurovascular structures that surgeons must avoid during mandibular procedures, reducing the risk of iatrogenic nerve injury. For implant placement, multi-structure segmentation maps the spatial relationships between teeth, alveolar bone, maxillary sinuses, and canals, enabling patient-specific planning that accounts for individual anatomy. In orthodontics, automated occlusal classification provides objective diagnosis of malocclusion, reducing inter-clinician variability and supporting standardized treatment planning.

Clinical validation and regulatory approval remain as future work, but the technical foundations are now established. The accuracy achieved by our methods approaches inter-annotator agreement, suggesting that automated systems could provide reliable support for clinical decision-making.

8.3.2 Research Community

The datasets and challenges have established new research directions that continue beyond this thesis. Maxillofacial AI has transitioned from a fragmented

field relying on private datasets to a coherent research community with shared benchmarks, enabling cumulative progress through fair comparison. Mamba for medical imaging, pioneered in part by this work, has become an active research area with numerous follow-up papers exploring variations, extensions, and applications beyond those presented here. Model merging for 3D segmentation, previously unexplored, is now established as feasible, opening new possibilities for modular and efficient clinical AI development.

8.3.3 Methodological Advances

Beyond specific applications, this thesis demonstrates general principles that transfer to other domains facing similar challenges. The synergistic cycle framework provides a conceptual lens for understanding how tools, datasets, and models interact, applicable to any medical imaging domain where annotation bottlenecks, data scarcity, and architectural limitations constrain progress. The directionality solutions developed for Mamba—bidirectional and multidirectional processing strategies—address fundamental issues that arise whenever sequential models are applied to spatial data, relevant to any 3D imaging modality. The connection between loss landscape geometry and model merging effectiveness provides principled guidance for pretraining strategies, with implications extending beyond medical imaging to any domain where modular model development is desirable.

8.4 Future Directions

8.4.1 Short-Term (1–2 years)

Several research directions are ready for immediate investigation. The ToothFairy4M annotation platform can be extended beyond maxillofacial imaging to other clinical domains, such as brain MRI tumor segmentation, where AI-assisted annotation would accelerate the creation of training datasets in modalities that similarly benefit from automated processing support. Interactive segmentation models represent a natural evolution of the annotation workflow, enabling clinicians to iteratively correct model predictions through intuitive interactions rather than manual delineation from scratch.

Multi-modal integration offers transformative potential by combining medical images with paired clinical text reports. This integration would enable bidirectional generation tasks: synthesizing CBCT or intra-oral scans from textual descriptions, and automatically generating draft reports from input images. Text-to-image generation addresses fundamental dataset limitations by creating synthetic examples of rare diseases, underrepresented conditions, and diverse patient demographics—including variations in age and ethnicity—thereby mitigating bias and improving model generalization across heterogeneous popula-

tions. Draft reports would provide pre-compiled structured documentation that clinicians can verify and refine with minimal effort, reducing administrative burden while maintaining clinical accuracy. This multi-modal capability forms the foundation for medical-domain multi-modal large language models—specialized conversational AI systems that function as expert assistants for dental or broader medical contexts, analogous to ChatGPT but trained for clinical workflows.

8.4.2 Medium-Term (3–5 years)

Medium-term research directions require more substantial investment but offer transformative potential. Foundation models trained on diverse medical imaging data could provide powerful pretrained representations that transfer across anatomies, modalities, and tasks, analogous to the impact of large language models on natural language processing. Federated learning would enable training on distributed data without centralization, addressing privacy constraints that currently prevent multi-institutional collaboration while allowing models to benefit from broader patient populations. Equally important, clinical adoption requires clear accountability frameworks that define who is responsible when AI outputs influence decisions—developers, institutions, and clinicians must have well-scoped roles, escalation paths, and human oversight aligned with the system's intended use. Practical auditability (e.g., logging model versions and inputs/outputs) is essential to support incident review and to generate the evidence required for high-risk clinical AI. Clinical validation through prospective studies would evaluate real-world performance and safety, establishing the evidence base required for regulatory approval and widespread clinical adoption.

8.4.3 Long-Term Vision

The long-term vision extends toward fundamentally new capabilities. Clinical deployment and regulatory approval represent critical milestones that would transition research prototypes into clinical practice. While sophisticated models exist in research settings, they remain unavailable to practicing clinicians due to privacy constraints and institutional approval requirements. Obtaining regulatory clearance and establishing privacy-preserving deployment frameworks would unlock access to real-world data at scale, enabling continuous model evaluation, refinement, and validation in authentic clinical environments. This deployment would reinforce the core objective of medical AI: not to replace clinical expertise, but to support practitioners by reducing error rates, streamlining workflows, and extending specialized care to underserved regions where medical resources are limited. By decreasing the time required for diagnostic interpretation and treatment planning, these systems have the potential to reduce healthcare costs while improving access and quality of care.

Self-improving systems would leverage deployment feedback to continuously enhance performance, learning from clinician corrections and clinical outcomes

observed in practice. Personalized medicine would adapt pretrained models to individual patients, accounting for anatomical variation, disease progression, and treatment response in ways that population-level models cannot capture. Seamless clinical integration would achieve unobtrusive incorporation into diagnostic workflows, where AI assistance becomes as natural and essential as any other clinical tool. Ultimately, autonomous annotation would close the synergistic cycle completely, with models capable of generating high-quality training data for their own improvement without human intervention—a capability that, combined with quality assurance mechanisms, could accelerate dataset expansion beyond current limitations.

8.5 Ethical Considerations

8.5.1 Privacy

Medical imaging data is inherently sensitive, and this thesis has taken privacy seriously throughout. All datasets were anonymized following ethical approval, with identifying information removed and facial features obscured where applicable. Synthetic data generation, explored in the testicular ultrasound application, offers privacy-preserving alternatives that enable model development without exposing real patient information. However, emerging research suggests that model weights may encode patient-specific information recoverable through adversarial attacks, implying that even releasing trained models requires care to avoid unintended privacy violations.

8.5.2 Bias and Fairness

AI systems can perpetuate or amplify biases present in training data. The ToothFairy datasets reflect the demographics of their source institutions in Italy and the Netherlands, which may not match deployment populations in other geographic or socioeconomic contexts. Performance variations across demographic groups—including age, sex, and ethnicity—should be systematically evaluated before clinical deployment, as models may underperform for underrepresented subgroups. Clinical validation must include diverse patient populations to ensure that automated systems benefit all patients equitably rather than exacerbating existing healthcare disparities.

8.5.3 Clinical Responsibility

Automated systems support but do not replace clinical judgment. Final decisions must remain with qualified clinicians who can integrate AI predictions with patient history, clinical examination, and professional experience. Failure

modes and limitations must be transparently communicated to users, as overconfidence in automated systems can lead to errors that manual analysis would have caught. Regulatory approval is required before clinical deployment, ensuring that systems meet safety and efficacy standards before affecting patient care. The methods developed in this thesis provide technical foundations, but responsible clinical translation requires additional validation, documentation, and oversight.

8.6 Summary

This chapter synthesized the thesis contributions and their broader implications. The synergistic cycle framework captures the fundamental dynamics of medical AI development, explaining how tools, datasets, and models interact to accelerate progress. Open science, computational efficiency, and pretraining strategies emerge as recurring themes that transcend specific applications. The clinical and research impact extends beyond the particular methods and datasets presented, establishing new research directions and community resources that will continue to generate value. Acknowledged limitations point toward concrete future research directions, from fully automated annotation to foundation models for medical imaging. Throughout, ethical considerations—privacy, fairness, and clinical responsibility—remain central to responsible development and deployment.

The following chapter concludes the thesis with key takeaways and final remarks.

9. Conclusion

This thesis has advanced artificial intelligence for maxillofacial image analysis through a comprehensive body of work spanning tools, datasets, architectures, and training strategies. The contributions form a synergistic cycle where each component enables and amplifies the others.

9.1 What Was Delivered

This thesis produced concrete, publicly available resources that continue to serve the research community:

Software: IACAT and ToothFairy4M annotation platforms, reducing expert annotation time from hours to minutes and enabling AI-assisted workflows that have annotated thousands of volumes.

Data: ToothFairy (153 CBCT volumes with 3D IAC segmentation, 290 CBCT volumes with 2D IAC segmentation), ToothFairy2 (480 volumes, 3D annotation of 42 anatomical classes), and ToothFairy3 (532 volumes, 3D annotation of 77 anatomical classes)—the largest public maxillofacial imaging benchmarks, with 6,000+ downloads and ongoing citations. TesticulUS (10k synthetic ultrasound images) and Bits2Bites (200 IOS scan pairs with occlusal classification) datasets.

Architectures: PosPadUNet3D (memory-augmented transformers with positional encoding, +1.6 Dice improvement) and Mamba variants (bi- and multi-directional processing, linear complexity, state-of-the-art across multiple benchmarks).

Training Methods: Model merging via stable pretraining (flat loss landscape minima through sharpness-aware minimization, +18 Dice improvement, modular clinical system development). First model merging application in medical imaging and segmentation tasks in general.

Community: Three MICCAI challenges with 50+ international teams, and Oral and Dental Image aNalysis (ODIN) MICCAI workshop, with 25+ attending the event and the establishment of a dedicated community.

Open Science: Open-source code, reproducible benchmarks, and challenge platforms that remain active beyond this thesis.

9.2 Key Takeaways

1. **Tools are foundational**: Specialized annotation software dramatically accelerates dataset creation. Investment in tools pays dividends throughout the research pipeline.

2. **Data enables progress**: Public benchmarks accelerate the entire research community. The value of open data extends far beyond the original creators.

3. **Architecture matters**: Domain-specific designs (PosPadUNet3D for tubular structures) and general-purpose innovations (Mamba with linear complexity) both advance the state of the art. Exploring new architectures is essential and allows us to build custom solutions for specific problems.

4. **Maintenance burden is reducible**: Model merging enables modular development and maintenance of clinical AI systems, eliminating the need to retrain from scratch when adapting to new anatomies. Stable pretraining produces robust pretrained models that serve as superior starting points for specialization, accelerating iteration and deployment.

5. **Generalization is possible**: Methods transfer across imaging modalities (CBCT, ultrasound, intra-oral scans) and clinical tasks (segmentation, classification).

6. **The cycle amplifies progress**: Improvements in any component benefit all others, creating positive feedback that accelerates research.

9.3 Final Remarks

Medical AI stands at an inflection point. The combination of large-scale public datasets, novel architectures, and efficient training strategies is enabling capabilities that were recently impossible. Yet substantial work remains: clinical validation, regulatory approval, workflow integration, and addressing edge cases that arise in real-world deployment.

This thesis has contributed to the foundations on which clinical applications will be built. By releasing tools, datasets, code, and models, we have enabled others to extend this work. The true measure of scientific contribution is not papers published but progress enabled—and the thousands of downloads of ToothFairy data, the citations of our methods, and the challenge participants who pushed beyond our baselines suggest that this work has enabled meaningful progress.

The synergistic cycle continues. Better tools will create better datasets. Better datasets will train better models. Better models will improve tools. And through this cycle, medical AI will advance toward its ultimate goal: improving patient care.

A. Appendix

A.1 Additional Activities During My PhD

A.1.1 Teaching Activities

During my PhD period, I have been a teaching assistant for the course "Strutture Dati e Algoritmi" (Data Structures and Algorithms) for the academic year 2022/2023, 2023/2024, 2024/2025, and 2025/2026 (total of 500+ hours) at the University of Modena and Reggio Emilia under the supervision of Prof. Federico Bolelli. I have also served as honorary fellow for the course "Fondamenti di Informatica" (Foundations of Computer Science) and "Multimedia Data Processing" courses held by Prof. Costantino Grana at the University of Modena and Reggio Emilia where I mainly contributed to the preparation and the benchmarking of the exams.

A.1.2 Organized Challenges and Workshops

I have led the organization of the following events:

- **ToothFairy Challenge @ MICCAI 2023**: The first edition focused on the segmentation of the Inferior Alveolar Canal (IAC) in CBCT volumes. The challenge addressed the critical need for accurate IAC identification in dental implant planning and maxillofacial surgery, providing the first public benchmark with both sparse 2D and dense 3D annotations. The challenge dataset consisted of 443 CBCT volumes (only 153 with dense 3D labels) and engaged international research teams in developing deep learning approaches for this clinically important structure. Results were published in IEEE Transactions on Medical Imaging.

- **ToothFairy2 Challenge @ MICCAI 2024**: The second edition significantly expanded the scope to multi-structure segmentation, including 42 anatomical classes such as the mandible, maxilla, individual teeth, pharynx, and canals. This comprehensive segmentation task provided the first publicly available fully annotated CBCT dataset for maxillofacial structures, enabling

cross-disciplinary applications in head and neck surgery, clinical practice, and anesthesiology. Results are currently under review at Medical Image Analysis.

- **ToothFairy3 Challenge @ MICCAI 2025**: The third edition introduced two novel tracks within the ODIN 2025 challenge cluster. Task 1 expanded the dataset to 77 anatomical classes, including pulp cavities, incisive nerves, and the lingual foramen relevant for orthodontic procedures, while incorporating computational efficiency as a primary evaluation metric alongside accuracy. Task 2 introduced an innovative interactive segmentation track for the IAC using click-based prompting, bridging the gap between fully automated methods and clinical applicability through user-guided refinement with minimal input.

- **ODIN 2025 Workshop @ MICCAI 2025**: The first edition of the ODIN (Oral and Dental Image aNalysis) workshop, held at MICCAI 2025 in Daejeon, Republic of Korea. centered on advancing oral and dental image analysis through large-scale, clinically grounded challenges covering CBCT and intra-oral scan data for segmentation, labeling, and registration tasks. The workshop brings together multiple established challenge tracks—such as ToothFairy and 3DTeethSeg—to address increasingly complex anatomical structures, multi-modal data, and real-world clinical constraints including accuracy–latency trade-offs and human-in-the-loop interaction. By combining rigorous benchmarking, diverse multi-center datasets, and focused discussion around clinical applicability, ODIN 2025 aims to foster reproducible, efficient, and deployable AI solutions for dental diagnosis, treatment planning, and surgical support

- **ODIN Challenge @ MICCAI 2026** ***(under review)***: Fourth edition of the our challenge series, which marks a change in the name as it serves as an umbrella name for multiple challenges. It introduces a benchmark for end-to-end multi-modal clinical report generation in oral and dental imaging, targeting the transformation of CBCT volumes, intra-oral scans, and photographs into structured, clinically meaningful text reports. The challenge comprises two tasks: ToothFairy4 for maxillofacial and surgical report generation from CBCT, and Bite2Text for orthodontic report generation from intra-oral scans and photographs. Both challenges are designed with multi-center training data and fully hidden test sets to rigorously assess robustness and generalization under domain shift. Evaluation prioritizes clinical correctness and completeness through expert-informed metrics and blinded clinician review, complemented by standard language metrics, with the goal of advancing deployable, reproducible multi-modal AI systems for dental and maxillofacial care.

- **ODIN 2026 Workshop @ MICCAI 2026** ***(under review)***: second edition of the ODIN (Oral and Dental Image aNalysis) workshop, focused on

advancing clinically grounded, multi-structure and multi-modal AI for oral and dental image analysis, addressing challenges such as small siloed datasets, limited generalization, and weak translation to real clinical workflows. The workshop brings together researchers and clinicians through invited keynotes, peer-reviewed oral and poster sessions, and challenge-aligned discussions covering CBCT, intra-oral scans, 2D radiographs, photographs, and clinical reports, with an emphasis on robustness, multi-center evaluation, and trustworthiness. A central contribution is the launch of the ODIN Data Collection and Annotation Initiative, a privacy-preserving community platform for creating and sharing high-quality multi-modal dental datasets to enable reproducible, generalizable, and clinically relevant research.

- **R2MI 2026 – Reproducible Research in Medical Imaging @ MICCAI 2026** ***(under review)*** is dedicated to improving reproducibility, transparency, and long-term reusability in medical imaging AI research by addressing gaps in reporting, software portability, evaluation practices, and data constraints. The workshop combines invited keynotes, oral presentations of accompanying reproducibility papers, and interactive discussions to showcase practical tools, infrastructures, and policies for building robust and reproducible medical imaging pipelines. A key innovation is the introduction of an LLM-assisted reproducibility assessment tool, alongside concrete outcomes such as reproducibility labels, community checklists, and a shared roadmap for the MICCAI community.

A.1.3 Industrial Collaborations

During my PhD period, I have been involved in the following industrial projects:

- **PBL S.p.A - Pharmaceutical Quality Control (€25,000.00)**: A collaboration with PBL S.p.A (Parma, Italy) focused on developing AI-powered computer vision algorithms to detect impurities in pharmaceutical vials' liquid. The project aimed to enhance quality control processes in pharmaceutical manufacturing by automating the inspection of vials for contaminants, reducing manual inspection time while maintaining high detection accuracy.

- **W&H S.r.l - Clinical CBCT Segmentation (€25,000.00)**: A collaboration with W&H S.r.l (Bergamo, Italy). aimed at developing and integrating AI-powered tools for real-time CBCT image segmentation directly into their manufactured CBCT machines. This partnership enabled the translation of research innovations into clinical practice, providing dentists and maxillofacial surgeons with immediate automated anatomical structure identification during image acquisition.

A.1.4 Grants

The following are the grants that I wrote and was awarded within the AImageLab group:

- **FAR 2024 - FOMO (€70,400.00)**: The project entitled "Synthetic Data: A Solution to Medical Imaging Limitations" will be mainly devoted to designing, implementing, and testing artificial intelligence tools for the massive generation of realistic synthetic data as an ethical alternative to using sensitive patient data. The project will involve collaborating with different research groups led by Prof. Alexandre Anesi, Prof. Laura Bertoni, and Prof. Giulia Besutti.
- **FARD 2024 - CURIOSITY DRIVEN projects (€25,350.00)**: The grant will be mainly devoted to the study and development of artificial intelligence algorithms for the segmentation of Maxillofacial structures. Moreover, within the AImageLab laboratory, I was also funded for the call "POTENZIAMENTO INFRASTRUTTURE" under the same grant.
- **FARD 2023 - CURIOSITY DRIVEN projects (€10,000.00)**: The grant will be mainly devoted to the study and development of artificial intelligence algorithms for the advanced analysis of confocal and whole-slide images to support the clinical practice.
- **FARD 2022 - STARTER KIT (€10,000.00)**: The grant will be mainly devoted to the development of ML/AI algorithms for classifying skin lesion images from mobile devices.

A.1.5 Conference and Journal Reviewer

I have been a reviewer for the following journals and conferences:

- *IEEE International Symposium on Biomedical Imaging (ISBI)*
- *IEEE International Conference on Computer Vision (ICCV)*
- *International Conference on Image Analysis and Processing (ICIAP)*
- *International Conference on Pattern Recognition (ICPR)*
- *British Machine Vision Conference (BMVC)*
- *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*
- *International Conference on Machine Learning (ICML)*
- *Neural Information Processing Systems (NeurIPS)*

- *International Conference on Learning Representations (ICLR)*
- *IEEE Transactions on Medical Imaging (TMI)*
- *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*
- *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*
- *IEEE Winter Conference on Applications of Computer Vision (WACV)*

A.1.6 Conferences and Summer Schools attended

I have attended the following conferences and summer schools:

- *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, Nashville, USA, 2025
- *28th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, Daejeon, South Korea, 2025
- *27th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, Marrakesh, Morocco, 2024
- *26th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, Vancouver, Canada, 2023
- *23rd International Conference on Image Analysis and Processing (ICIAP)*, Rome, Italy, 2025
- *21st International Conference on Image Analysis and Processing (ICIAP)*, Udine, Italy, 2023
- *20th International Conference on Image Analysis and Processing (ICIAP)*, Lecce, Italy, 2022
- *International Computer Vision Summer School (ICVSS)*, Scicli, Italy, 2023
- *ELLIS Summer School on Large-Scale AI for Research and Industry*, Modena, Italy, 2023

A.2 List of Publications

This thesis is based on the following publications (* means equal contribution):

1. Luca Lumetti, Vittorio Pipoli, Federico Bolelli, Costantino Grana, "Annotating the Inferior Alveolar Canal: The Ultimate Tool," *International Conference on Image Analysis and Processing (ICIAP)*, 2023.

2. Luca Lumetti*, Vittorio Pipoli*, Federico Bolelli, Elisa Ficarra, Costantino Grana, "Enhancing Patch-Based Learning for the Segmentation of the Mandibular Canal," *IEEE Access*, vol. 12, pp. 79014–79024, 2024.

3. Federico Bolelli, Luca Lumetti, Shankeeth Vinayahalingam, Mattia Di Bartolomeo, Arrigo Pellacani, Kevin Marchesini, Niels Van Nistelrooij, Pieter Van Lierop, Tong Xi, Yusheng Liu, Rui Xin, Tao Yang, Lisheng Wang, Haoshen Wang, Chenfan Xu, Zhiming Cui, Marek Wodzinski, Henning Müller, Yannick Kirchhoff, Maximilian R Rokuss, Klaus Maier-Hein, Jaehwan Han, Wan Kim, Hong-Gi Ahn, Tomasz Szczepański, Michal K Grzeszczyk, Przemyslaw Korzeniowski, Vicent Caselles-Ballester, Xavier Paolo Burgos-Artizzu, Ferran Prados Carrasco, Bram van Ginneken, Alexandre Anesi, Costantino Grana, "Segmenting the Inferior Alveolar Canal in CBCTs Volumes: the ToothFairy Challenge," *IEEE Transactions on Medical Imaging*, 2024.

4. Luca Lumetti*, Vittorio Pipoli*, Kevin Marchesini*, Elisa Ficarra, Costantino Grana, Federico Bolelli, *et al.*, "Accurate Voxel-Level 3D Medical Image Segmentation with Mambas," *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2025.

5. Luca Lumetti*, Vittorio Pipoli*, Kevin Marchesini*, Elisa Ficarra, Costantino Grana, Federico Bolelli, *et al.*, "Taming Mambas for 3D Medical Image Segmentation," *IEEE Access*, 2025.

6. Federico Bolelli*, Kevin Marchesini*, Niels van Nistelrooij, Luca Lumetti, Vittorio Pipoli, Elisa Ficarra, Shankeeth Vinayahalingam, Costantino Grana, "Segmenting Maxillofacial Structures in CBCT Volumes," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5238–5248, 2025.

7. Luca Lumetti*, Giacomo Capitani*, Elisa Ficarra, Simone Calderara, Costantino Grana, Angelo Porrello, Federico Bolelli, "U-Net Transplant: The Role of Pre-training for Model Merging in 3D Medical Segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2025.

8. Nicola Morelli, Kevin Marchesini, Luca Lumetti, Daniele Santi, Costantino Grana, Federico Bolelli, "Enhancing Testicular Ultrasound Classification through Synthetic Data," *International Conference on Image Analysis and Processing (ICIAP)*, 2025.

9. Luca Lumetti*, Lorenzo Borghi*, Francesca Cremonini, Federico Rizzo, Costantino Grana, Luca Lombardo, Federico Bolelli, "Bits2Bites: Intra-oral Scans Occlusal Classification," *MICCAI Workshop on Oral and Dental Image aNalysis (ODIN)*, 2025.

10. Luca Lumetti*, Vittorio Pipoli*, Federico Bolelli, Elisa Ficarra, Costantino Grana, "Location Matters: Harnessing Spatial Information to Enhance the Segmentation of the Inferior Alveolar Canal in CBCTs," *International Conference on Pattern Recognition (ICPR)*, 2024.

11. Mattia Di Bartolomeo, Arrigo Pellacani, Federico Bolelli, Marco Cipriano, Luca Lumetti, Sara Negrello, Stefano Allegretti, Paolo Minafra, Federico Pollastri, Riccardo Nocini, Giacomo Colletti, Luigi Chiarini, Costantino Grana, Alexandre Anesi, "Inferior alveolar canal automatic detection with deep learning CNNs on CBCTs: development of a novel model and release of open-source dataset and algorithm," *Applied Sciences*, vol. 13, no. 5, p. 3271, 2023.

More publications have been made during my PhD period but are not included in this thesis as they are not directly related to the research presented in this thesis.

1. Jianning Li, Zongwei Zhou, Jiancheng Yang, Antonio Pepe, Christina Gsaxner, Gijs Luijten, Chongyu Qu, Tiezheng Zhang, Xiaoxi Chen, Wenxuan Li, Marek Wodzinski, Paul Friedrich, Kangxian Xie, Yuan Jin, Narmada Ambigapathy, Enrico Nasca, Naida Solak, Gian Marco Melito, Viet Duc Vu, Afaque R Memon, Christopher Schlachta, Sandrine De Ribaupierre, Rajnikant Patel, Roy Eagleson, Xiaojun Chen, Heinrich Mächler, Jan Stefan Kirschke, Ezequiel De La Rosa, Patrick Ferdinand Christ, Hongwei Bran Li, David G Ellis, Michele R Aizenberg, Sergios Gatidis, Thomas Küstner, Nadya Shusharina, Nicholas Heller, Vincent Andrearczyk, Adrien Depeursinge, Mathieu Hatt, Anjany Sekuboyina, Maximilian T Löffler, Hans Liebl, Reuben Dorent, Tom Vercauteren, Jonathan Shapey, Aaron Kujawa, Stefan Cornelissen, Patrick Langenhuizen, Achraf Ben-Hamadou, Ahmed Rekik, Sergi Pujades, Edmond Boyer, Federico Bolelli, Costantino Grana, Luca Lumetti, Hamidreza Salehi, Jun Ma, Yao Zhang, Ramtin Gharleghi, Susann Beier, Arcot Sowmya, Eduardo A Garza-Villarreal, Thania Balducci, Diego Angeles-Valdez, Roberto Souza, Leticia Rittner, Richard Frayne, Yuanfeng Ji, Vincenzo Ferrari, Soumick Chatterjee, Florian Dubost, Stefanie Schreiber, Hendrik Mattern, Oliver Speck, Daniel Haehn, Christoph John, Andreas Nürnberger, João Pedrosa, Carlos Ferreira, Guilherme Aresta, António Cunha, Aurélio Campilho, Yannick Suter, Jose Garcia, Alain Lalande, Vicky Vandenbossche, Aline Van Oevelen, Kate Duquesne, Hamza Mekhzoum, Jef Vandemeulebroucke, Emmanuel Audenaert, Claudia Krebs, Timo Van Leeuwen, Evie Vereecke, Hauke Heidemeyer, Rainer Röhrig, Frank Hölzle, Vahid Badeli, Kathrin Krieger, Matthias Gunzer, Jianxu Chen, Timo Van Meegdenburg, Amin Dada, Miriam Balzer, Jana Fragemann, Frederic Jonske, Moritz Rempe, Stanislav Malorodov, Fin H Bahnsen, Constantin Seibold, Alexander Jaus, Zdravko Marinov, Paul F Jaeger, Rainer Stiefelhagen, Ana Sofia Santos, Mariana Lindo, André Ferreira, Victor Alves, Michael Kamp, Amr Abourayya, Felix Nensa,

Fabian Hörst, Alexander Brehmer, Lukas Heine, Yannik Hanusrichter, Martin Weßling, Marcel Dudda, Lars E Podleska, Matthias A Fink, Julius Keyl, Konstantinos Tserpes, Moon-Sung Kim, Shireen Elhabian, Hans Lamecker, Dženan Zukić, Beatriz Paniagua, Christian Wachinger, Martin Urschler, Luc Duong, Jakob Wasserthal, Peter F Hoyer, Oliver Basu, Thomas Maal, Max JH Witjes, Gregor Schiele, Ti-chiun Chang, Seyed-Ahmad Ahmadi, Ping Luo, Bjoern Menze, Mauricio Reyes, *et al.*, "MedShapeNet – a large-scale dataset of 3D medical shapes for computer vision," *Biomedical Engineering/Biomedizinische Technik*, 2025.

2. Jun Ma, Feifei Li, Sumin Kim, Reza Asakereh, Bao-Hiep Le, Dang-Khoa Nguyen-Vu, Alexander Pfefferle, Muxin Wei, Ruochen Gao, Donghang Lyu, Songxiao Yang, Lennart Purucker, Zdravko Marinov, Marius Staring, Haisheng Lu, Thuy Thanh Dao, Xincheng Ye, Zhi Li, Gianluca Brugnara, Philipp Vollmuth, Martha Foltyn-Dumitru, Jaeyoung Cho, Mustafa Ahmed Mahmutoglu, Martin Bendszus, Irada Pflüger, Aditya Rastogi, Dong Ni, Xin Yang, Guang-Quan Zhou, Kaini Wang, Nicholas Heller, Nikolaos Papanikolopoulos, Christopher Weight, Yubing Tong, Jayaram K Udupa, Cahill J Patrick, Yaqi Wang, Yifan Zhang, Francisco Contijoch, Elliot McVeigh, Xin Ye, Shucheng He, Robert Haase, Thomas Pinetz, Alexander Radbruch, Inga Krause, Erich Kobler, Jian He, Yucheng Tang, Haichun Yang, Yuankai Huo, Gongning Luo, Kaisar Kushibar, Jandos Amankulov, Dias Toleshbayev, Amangeldi Mukhamejan, Jan Egger, Antonio Pepe, Christina Gsaxner, Gijs Luijten, Shohei Fujita, Tomohiro Kikuchi, Benedikt Wiestler, Jan S Kirschke, Ezequiel de la Rosa, Federico Bolelli, Luca Lumetti, Costantino Grana, Kunpeng Xie, Guomin Wu, Behrus Puladi, Carlos Martín-Isla, Karim Lekadir, Victor M Campello, Wei Shao, Wayne Brisbane, Hongxu Jiang, Hao Wei, Wu Yuan, Shuangle Li, Yuyin Zhou, Bo Wang, *et al.*, "Efficient medsams: Segment anything in medical images on laptop," *arXiv preprint arXiv:2412.16085*, 2024.

3. Gianpaolo Bontempo, Luca Lumetti, Angelo Porrello, Federico Bolelli, Simone Calderara, Elisa Ficarra, "Buffer-MIL: Robust multi-instance learning with a buffer-based approach," *International Conference on Image Analysis and Processing*, 2023.

4. Federico Bolelli, Stefano Allegretti, Luca Lumetti, Costantino Grana, "A State-of-the-Art Review with Code about Connected Components Labeling on GPUs," *IEEE Transactions on Parallel and Distributed Systems*, 2024.

5. Gabriele Rosati, Kevin Marchesini, Luca Lumetti, Federica Sartori, Beatrice Balboni, Filippo Begarani, Luca Vescovi, Federico Bolelli, Costantino Grana, "Identifying Impurities in Pharma Vials' Liquid," *International Conference on Pattern Recognition (ICPR)*, 2024.

6. Federico Bolelli, Luca Lumetti, Kevin Marchesini, Ettore Candeloro, Cost-

antino Grana, "Investigating the ABCDE Rule in Convolutional Neural Networks," *International Conference on Pattern Recognition (ICPR)*, 2024.

A.3 Internship Experience

My doctoral research was complemented by two mandatory internship experiences that provided invaluable opportunities to apply and extend my research on medical AI in real-world industrial and clinical settings. The first internship fulfilled the mandatory requirement for an international research experience, while the second was required as part of my industrial PhD sponsorship.

A.3.1 Relu - Leuven, Belgium

The international internship was conducted at Relu, a Belgian AI technology company headquartered in Leuven with an office in Boston, USA. Founded in 2019, Relu builds cloud-based artificial intelligence tools to automate and scale dental design, imaging, and treatment planning workflows for dental labs, clinics, and software partners. The company leverages advanced computer vision and machine learning to perform tasks such as CBCT segmentation, intra-oral scan alignment, 3D model generation, and automated design of dental appliances including surgical guides, retainers, and night guards. Relu's products—including Relu® Cloud, Relu® Creator, and Relu® Engine—have achieved key regulatory milestones such as FDA 510(k) clearance and European CE certification, and are trusted by hundreds of dental labs and thousands of clinicians across Europe and the United States.

The internship represented a natural extension of my PhD research, as Relu's core mission of automating dental imaging analysis through AI directly aligned with the technical expertise I had developed through the ToothFairy projects and architectural innovations presented in this thesis. My work at Relu focused on improving existing deep learning pipelines and developing novel multi-modal approaches that bridged the gap between academic research and production deployment.

A significant portion of my efforts concentrated on enhancing the robustness and efficiency of Relu's existing deep learning models. This involved comprehensive retraining campaigns to improve performance metrics, implementing more rigorous evaluation frameworks with expanded metric suites, and establishing stable ranking systems for model selection. These ranking systems provided clear decision boundaries for determining when a newly trained model should replace an existing production model, addressing a critical challenge in maintaining quality assurance for deployed AI systems in clinical workflows.

Beyond optimization of existing pipelines, I contributed to the development of novel capabilities. One project addressed the challenge of bracket removal from intra-oral scans captured while patients wore orthodontic appliances. This

required combining deep learning segmentation techniques with classical computer vision methods to accurately identify and digitally remove brackets, enabling the underlying tooth geometry to be recovered for treatment planning purposes.

The most substantial contribution involved developing a multi-modal pipeline for generating dental emergence profiles—the critical transition contour where teeth emerge from the gingival tissue. Accurate modeling of emergence profiles is essential for designing restorations and prosthetics that integrate naturally with surrounding soft tissue. The pipeline integrated multiple processing stages: segmentation of teeth from CBCT volumes, segmentation of teeth from intra-oral scans, registration between these complementary modalities, and application of computer vision techniques to synthesize the emergence profile from the fused information. This multi-modal approach exemplifies the synergistic potential of combining volumetric bone information from CBCT with precise surface geometry from intra-oral scans—precisely the type of cross-modal integration I identified as an untapped opportunity in Sec. 1.1.

Across these projects, I developed and deployed over ten deep learning models for segmentation, registration, and classification tasks across 3D CBCT volumes, intra-oral scans, meshes, and point clouds. The emergence profile pipeline alone eliminated approximately 95% of manual annotation effort, saving 8–10 clinician hours per week and increasing patient data throughput by a factor of twenty. These efficiency gains demonstrate the transformative potential of AI in clinical dental workflows when research innovations are successfully translated to production systems.

A.3.2 Miliaris - Modena, Italy

The second internship was conducted at Miliaris, the industrial sponsor of my doctoral research, as part of the mandatory requirements for an industry-funded PhD position. Miliaris is an Italian software development company based in Modena that specializes in creating custom web and mobile applications for businesses and public institutions. Founded in the mid-2000s, Miliaris describes itself as "digital artisans," emphasizing user-centered design and tailored technical solutions. Their portfolio spans healthcare appointment systems, payment integration platforms, sports medical visit booking services, and professional streaming solutions, demonstrating breadth across diverse application domains.

During my time at Miliaris, I initiated and led a collaborative research project with the Hospital of Sassuolo, a city near Modena, focused on developing an intelligent patient monitoring system for hospital wards. The project addressed a critical clinical challenge: reducing response times to adverse events occurring in patient rooms. Traditional monitoring relies on nurse call buttons and periodic rounds, creating delays that can compromise patient safety during critical incidents.

I engineered a real-time, privacy-compliant monitoring system that leveraged multi-modal Vision-Language foundation models to detect critical events from continuous video streams. The system achieved real-time performance with 250 millisecond latency while maintaining over 80% precision in event detection—balancing the need for rapid alerts against the clinical cost of false positives that could lead to alert fatigue among healthcare staff. Privacy compliance was paramount in the system design, as video monitoring in healthcare settings raises significant ethical and regulatory concerns. The architecture processed visual information on-device without storing patient-identifiable imagery, extracting only high-level semantic event descriptors for alert generation.

To translate detections into clinical action, I designed and deployed a responsive web application that integrated with the hospital's existing pager infrastructure through a custom API. This integration ensured seamless adoption within established clinical workflows rather than requiring staff to monitor separate alert systems. The complete pipeline—from event detection to clinical notification—reduced alert-to-response time from approximately 20 minutes (typical interval between nursing rounds) to under 10 seconds, representing a 120-fold improvement that could prove lifesaving in time-critical scenarios such as patient falls or sudden deterioration.

While this project diverged from the dental imaging focus of my primary research, it demonstrated the broad applicability of the computer vision and deep learning expertise developed throughout my PhD. The challenges of deploying AI in clinical environments—including real-time performance requirements, privacy constraints, integration with legacy systems, and the critical importance of reliability—parallel those encountered in dental AI deployment. Both domains require not only algorithmic innovation but also careful attention to clinical workflows, regulatory requirements, and the human factors that determine whether technical capabilities translate into improved patient outcomes.

These internship experiences collectively reinforced a central theme of this thesis: the gap between academic research and clinical deployment requires sustained attention to practical constraints that extend far beyond algorithmic performance metrics. Efficiency, regulatory compliance, integration with existing workflows, explainability to clinical users, and continuous quality assurance in production environments are prerequisites for translating research innovations into tangible clinical value.

A.4 Advisor for Master and Bachelor Students

During my PhD period, I supervised the following Master and Bachelor students:

- **Master Thesis - Matteo Lugli (2025/2026)**: "Automated Dental Bracket Landmark Prediction Using Deep Learning."

- **Bachelor Thesis - Lorenzo Borghi (2024/2025)**: "Bits2Bites: Intra-oral Scans Occlusal Classification."
- **Bachelor Thesis - Mattia Gualtieri (2024/2025)**: "Dal Voxel al Web: Modulo Three.js per il Rendering Web-based di Volumi CBCT."
- **Bachelor Thesis - Matteo Ferrari (2023/2024)**: "Progettazione e Integrazione del Modulo Speech-to-Text in un Sistema Distribuito: il Caso ToothFairy4M."
- **Master Thesis - Ettore Candeloro (2022/2023)**: "Skin Lesion Classification Explained with Generative Adversarial Networks."
- **Master Thesis - Gabriele Rosati (2022/2023)**: "Prediction of Kidney Failure with Deep Neural Networks Fusing WSI and Immunofluorescence Images."

A.5 Open Source Contributions

All code, datasets, and models from this thesis are publicly available:

A.5.1 Datasets

- **ToothFairy**: `https://ditto.ing.unimore.it/toothfairy`
- **ToothFairy2**: `https://ditto.ing.unimore.it/toothfairy2`
- **ToothFairy3**: `https://ditto.ing.unimore.it/toothfairy3`
- **TesticulUS**: `https://ditto.ing.unimore.it/testiculus/`
- **Bits2Bites**: `https://github.com/AImageLab-zip/Bits2Bites`

A.5.2 Challenge Platforms

- **ToothFairy Challenge**: `https://toothfairy.grand-challenge.org/`
- **ToothFairy2 Challenge**: `https://toothfairy2.grand-challenge.org/`
- **ToothFairy3 Challenge**: `https://toothfairy3.grand-challenge.org/`

A.5.3 Organized Workshops

- **ODIN at MICCAI 2025**: `https://odin-workshops.org/2025`

A.5.4 Code Repositories - Models

- **TamingMambas**: https://github.com/LucaLumetti/TamingMambas
- **IAC Segmentation**: https://github.com/AImageLab-zip/alveolar_canal
- **U-Net Transplant**: https://github.com/AImageLab-zip/UNetTransplant
- **ToothFairy2 Benchmarks**: https://github.com/AImageLab-zip/ToothFairy2-Benchmark

A.5.5 Code Repositories - Tools

- **IACAT**: https://github.com/AImageLab-zip/IAN_annotation_tool
- **ToothFairy Challenges Code**: https://github.com/AImageLab-zip/ToothFairy
- **Bits2Bites**: https://github.com/AImageLab-zip/Bits2Bites
- **ToothFairy4M**: https://github.com/AImageLab-zip/ToothFairy4M
- **TesticulUS**: https://github.com/AImageLab-zip/TesticulUS
- **ToothFairy Napari Plugin**: https://github.com/LucaLumetti/napari-toothfairy-annotator

Bibliography

[1] Dentistry — designation system for teeth and areas of the oral cavity, 2016. URL `https://www.iso.org/standard/68292.html`. International standard defining the FDI two-digit dental notation. 35

[2] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022. 13, 61

[3] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011. 13

[4] Olivier Bernard, Alain Lalande, Clement Zotti, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018. 61

[5] Federico Bolelli, Kevin Marchesini, Niels van Nistelrooij, Luca Lumetti, Vittorio Pipoli, Elisa Ficarra, Shankeeth Vinayahalingam, and Costantino Grana. Segmenting Maxillofacial Structures in CBCT Volumes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 77

[6] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. In *European Conference on Computer Vision*, pages 205–218. Springer, 2022. 64, 65, 66

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-

supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 19

[8] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-SGD: Biasing Gradient Descent into Wide Valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12), 2019. 77

[9] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *European Conference on Computer Vision*, 2018. 75

[10] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. TransUNet: Transformers make strong encoders for medical image segmentation. In *arXiv preprint arXiv:2102.04306*, 2021. 15, 64, 65, 66, 67

[11] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018. 14

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020. 19

[13] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 424–432, 2016. 13

[14] Marco Cipriano, Stefano Allegretti, Federico Bolelli, Federico Pollastri, and Costantino Grana. Improving segmentation of the inferior alveolar nerve through deep label propagation. *IEEE Access*, 10:54934–54945, 2022. 26, 46, 49, 51, 54

[15] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 12

[16] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-Memory Transformer for Image Captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10578–10587, 2020. 47, 57

[17] Donato Crisostomi, Marco Fumero, Daniele Baieri, Florian Bernard, and Emanuele Rodola. c^2m^3: Cycle-consistent multi-model merging. *Advances in Neural Information Processing Systems*, 37:28674–28705, 2024. 20

[18] Zhiming Cui, Yu Fang, Lanzhuju Mei, Bojun Zhang, Bo Yu, Jiameng Liu, Caiwen Jiang, Yuhang Sun, Lei Ma, Jiawei Huang, et al. A fully automatic AI system for tooth and alveolar bone segmentation from cone-beam CT images. *Nature Communications*, 13(1), 2022. 77

[19] Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024. 17

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE conference on Computer Vision and Pattern Recognition*, 2009. 18

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 19

[22] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp Minima Can Generalize For Deep Nets. In *International Conference on Machine Learning*, 2017. 76

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 5, 14, 15

[24] Herbert Edelsbrunner, David Kirkpatrick, and Raimund Seidel. On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, 29(4):551–559, 1983. 25

[25] Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean-Christophe Fillion-Robin, Sonia Pujol, Christian Bauer, Dominique Jennings, Fiona Fennessy, Milan Sonka, et al. 3d slicer as an image computing platform for the quantitative imaging network. *Magnetic resonance imaging*, 30(9):1323–1341, 2012. 24

[26] Jonathan Frankle, David J. Schwab, and Ari S. Morcos. The Early Phase of Neural Network Training. In *International Conference on Learning Representations Workshop*, 2020. 76

[27] Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodola. Task singular vectors: Reducing task interference in model merging. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18695–18705, 2025. 20

[28] Kiran Garlapati, D. B. G. Babu, N. C. S. K. Chaitanya, Kalyan Guduru, Suresh Babu Remidi, and D. Srilakshmi. Role of cone beam computed tomography in dentistry: A review. *Journal of Dr. NTR University of Health Sciences*, 6(3):139–145, 2017. 1

[29] Mina Ghaffari, Arcot Sowmya, and Ruth Oliver. Automated brain tumor segmentation using multi-modal brain scans: a survey based on models submitted to the brats 2012–2018 challenges. *IEEE reviews in biomedical engineering*, 13:156–168, 2019. 13

[30] Florin C Ghesu, Bogdan Georgescu, Awais Mansoor, Youngjin Yoo, Dominik Neumann, Pragneshkumar Patel, Reddappagari Suryanarayana Vishwanath, James M Balter, Yue Cao, Sasa Grbic, et al. Contrastive self-supervised learning from 100 million medical images with optional supervision. *Journal of Medical Imaging*, 9(6):064503–064503, 2022. 20

[31] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 17, 58

[32] Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. HiPPO: Recurrent memory with optimal polynomial projections. In *Advances in Neural Information Processing Systems*, volume 33, pages 1474–1487, 2020. 17

[33] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMAf*, 316(22):2402–2410, 2016. 12

[34] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 18

[35] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, and Daguang Xu. Swin UNETR: Swin transformers for semantic

segmentation of brain tumors in MRI images. In *International MICCAI Brainlesion Workshop*, pages 272–284, 2022. 15, 52, 64, 65, 66, 67

[36] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu. UNETR: Transformers for 3D medical image segmentation. In *IEEE Winter Conference on Applications of Computer Vision*, pages 574–584, 2022. 15, 52, 64, 65, 66, 67

[37] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 19

[38] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 19

[39] Sepp Hochreiter and Jürgen Schmidhuber. Flat Minima. *Neural Computation*, 9(1), 1997. 76

[40] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 14

[41] Xiaohong Huang, Zhifang Deng, Dandan Li, and Xueguang Yuan. MISSFormer: An Effective Medical Image Segmentation Transformer. *arXiv preprint arXiv:2109.07162*, 2021. 64, 65, 66

[42] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations Workshop*, 2023. 20, 73, 79

[43] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019. 12

[44] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2): 203–211, 2021. 5, 13, 52, 64, 65, 66, 67, 71

[45] Joe Iwanaga and R. Shane Tubbs. Anatomy of the inferior alveolar nerve: A comprehensive review. *Clinical Anatomy*, 34(4):578–589, 2021. 2

[46] Joel Jaskari, Jaakko Sahlsten, Jorma Järnstedt, Helena Mehtonen, Kalle Karhu, Osku Sundqvist, Ari Hietanen, Vesa Varjonen, Vesa Mattila, and Kimmo Kaski. Deep learning method for mandibular canal segmentation in dental cone beam computed tomography volumes. *Scientific reports*, 10 (1):5842, 2020. 11, 32

[47] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, et al. AMOS: ALarge-Scale Abdominal Multi-Organ Benchmark for Versatile Medical Image Segmentation. *Neural Information Processing Systems*, 35, 2022. 77

[48] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019. 12

[49] Dagmar Kainmüller, Thomas Lange, and Hans Lamecker. Shape constrained automatic segmentation of the liver based on a heuristic intensity model. *MICCAI Workshop on 3D Segmentation in the Clinic*, pages 109–116, 2009. 10

[50] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In *International Conference on Learning Representations Workshop*, 2017. 76

[51] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 39

[52] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *National Academy of Sciences*, 2017. 77

[53] Gloria Hyunjung Kwak, Eun-Jung Kwak, Jae Min Song, Hae Ryoun Park, Yun-Hoa Jung, Bong-Hae Cho, Pan Hui, and Jae Joon Hwang. Automatic mandibular canal detection using a deep convolutional neural network. *Scientific Reports*, 10(1):5711, 2020. 11, 32

[54] Pierre Lahoud, Mostafa EzEldeen, Thomas Beznik, Mattias Zeven, Constantinus Tieghem, Manon Zijderveld, et al. Artificial intelligence for fast and accurate 3-dimensional tooth segmentation on cone-beam computed tomography. *Journal of Endodontics*, 47(5):827–835, 2021. 32

[55] Pierre Lahoud, Siebe Diels, Liselot Niclaes, Stijn Van Aelst, Holger Willems, Adriaan Van Gerven, Marc Quirynen, and Reinhilde Jacobs. Development and validation of a novel artificial intelligence driven tool for accurate mandibular canal segmentation on cbct. *Journal of dentistry*, 116:103891, 2022. 11

[56] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, 2015. 11, 61, 77

[57] Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020. 76

[58] Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob Van de Loo, Rob Vogels, et al. 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. *GigaScience*, 7 (6):giy065, 2018. 12

[59] Yuxin Liu et al. Self-training for inferior alveolar canal segmentation. In *MICCAI ToothFairy Challenge*, 2023. 34

[60] Zhiyang Liu, Dong Yang, Minghao Zhang, Guohua Liu, Qian Zhang, and Xiaonan Li. Inferior Alveolar Nerve Canal Segmentation on CBCT Using U-Net with Frequency Attentions. *Bioengineering*, 11(4):354, 2024. 55

[61] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations Workshop*, 2019. 77

[62] Luca Lumetti, Vittorio Pipoli, Federico Bolelli, Elisa Ficarra, and Costantino Grana. Enhancing Patch-Based Learning for the Segmentation of the Mandibular Canal. *IEEE Access*, 2024. 77

[63] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016. 14

[64] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024. 17, 64, 65, 66, 67

[65] Lena Maier-Hein, Matthias Eisenmann, Annika Reinke, Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, Hrvoje Bogunovic, Andrew P. Bradley, Aaron Carass, et al. Why rankings of biomedical image analysis

competitions should be interpreted with care. *Nature Communications*, 9 (1):5217, 2018. 34

[66] Lena Maier-Hein, Annika Reinke, Patrick Godau, et al. Metrics reloaded: recommendations for image analysis validation. *Nature Methods*, 21(2): 195–212, 2024. doi: 10.1038/s41592-023-02151-z. 11, 12

[67] Michael S Matena and Colin A Raffel. Merging Models with Fisher-Weighted Averaging. *Advances in Neural Information Processing Systems*, 2022. 20

[68] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision*, pages 565–571, 2016. 13

[69] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Razvan Pascanu, and Hassan Ghasemzadeh. Understanding the Role of Training Regimes in Continual Learning. In *Advances in Neural Information Processing Systems*, 2020. 76

[70] napari contributors. napari: a multi-dimensional image viewer for Python. `https://github.com/napari/napari`, 2019. 27

[71] Hoa Nguyen and John Smith. Clinical considerations for inferior alveolar nerve preservation. *Journal of Oral and Maxillofacial Surgery*, 81:245–256, 2023. 2

[72] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 19

[73] Jin Seo Park, Min Suk Chung, Sung Bae Hwang, Yong Sook Lee, and Dong-Hwan Har. Technical report on semiautomatic segmentation using the adobe photoshop. *Journal of digital imaging*, 18(4):333–343, 2005. 11

[74] Nicolas Pielawski and Carolina Wählby. Introducing Hann windows for reducing edge-effects in patch-based image segmentation. *PloS one*, 15(3): e0229839, 2020. 49, 57

[75] Angelo Porrello, Lorenzo Bonicelli, Pietro Buzzega, Monica Millunzi, Simone Calderara, and Rita Cucchiara. A Second-Order Perspective on Model Compositionality and Incremental Learning. In *International Conference on Learning Representations*, 2025. 74, 75

[76] Filippo Rinaldi, Giacomo Capitani, Lorenzo Bonicelli, Donato Crisostomi, Federico Bolelli, Elisa Ficarra, Emanuele Rodolà, Simone Calderara, and

Angelo Porrello. Update Your Transformer to the Latest Release: Re-Basin of Task Vectors. In *International Conference on Machine Learning*, 2025. 20

[77] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015. 5, 13

[78] Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F Jaeger, and Klaus H Maier-Hein. MedNeXt: Transformer-driven Scaling of ConvNets for Medical Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2023*, pages 405–415. Springer, 2023. 64, 65, 66, 67

[79] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas Van Den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical image analysis*, 42:1–13, 2017. 13

[80] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 19

[81] Derek Tam, Mohit Bansal, and Colin Raffel. Merging by Matching Models in Task Parameter Subspaces. *Transactions on Machine Learning Research*, 2024. 20

[82] Juan Carlos Tapia and Jordi Peraire. Cone beam computed tomography in oral and maxillofacial surgery: Current applications. *Oral and Maxillofacial Surgery Clinics of North America*, 35(1):1–12, 2023. 1

[83] Anna Tereshchuk, Aldo Bruno Giannì, and Marina Codari. Advances in cbct imaging for dental and maxillofacial applications. *Dentomaxillofacial Radiology*, 51(4), 2022. 1

[84] Muhammad Usman, Azka Rehman, Amal Muhammad Saleem, Rabeea Jawaid, Shi-Sub Byon, Sung-Hyun Kim, Byoung-Dai Lee, Min-Suk Heo, and Yeong-Gil Shin. Dual-Stage Deeply Supervised Attention-Based Convolutional Neural Networks for Mandibular Canal Segmentation in CBCT Scans. *Sensors*, 22(24):9877, 2022. 55

[85] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 14

[86] Hongyu Wang and Yong Xia. Chestnet: A deep neural network for classification of thoracic diseases on chest radiography. *arXiv preprint arXiv:1807.03058*, 2018. 12

[87] Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. TransBTS: multi-modal Brain Tumor Segmentation Using Transformer. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021*, pages 109–119. Springer, 2021. 64, 65, 66

[88] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 14

[89] Andre Weissheimer, Luciane MacEdo De Menezes, Glenn T Sameshima, Reyes Enciso, John Pham, and Dan Grauer. Imaging software accuracy for 3-dimensional analysis of the upper airway. *American Journal of Orthodontics and Dentofacial Orthopedics*, 142(6):801–813, 2012. 11

[90] World Health Organization. Global oral health status report: towards universal health coverage for oral health by 2030, 2022. 1

[91] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention*, pages 171–180. Springer, 2021. 64, 65, 66, 67

[92] Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent escapes from sharp minima exponentially fast. *arXiv preprint arXiv:2002.03495*, 2020. 76

[93] Guoping Xu, Xuan Zhang, Xinwei He, and Xinglong Wu. LeViT-UNet: Make Faster Encoders with Transformer for Medical Image Segmentation. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 42–53. Springer, 2023. 64, 65, 66

[94] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. TIES-Merging: Resolving Interference When Merging Models. In *Neural Information Processing Systems*, 2024. 20, 78, 79

[95] Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. AdaMerging: Adaptive Model Merging for Multi-Task Learning. In *International Conference on Learning Representations Workshop*, 2024. 20

[96] Paul A Yushkevich, Yang Gao, and Guido Gerig. Itk-snap: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. In *2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 3342–3345. IEEE, 2016. 24

[97] Huanmiao Zhao, Junhua Chen, Zhaoqiang Yun, Qianjin Feng, Liming Zhong, and Wei Yang. Whole mandibular canal segmentation using transformed dental CBCT volume in Frenet frame. *Heliyon*, 9(7), 2023. 55, 56, 58

[98] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnFormer: Interleaved transformer for volumetric segmentation. In *arXiv preprint arXiv:2109.03201*, 2021. 15, 64, 65, 66, 67, 71

[99] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024. 17

[100] Qinfeng Zhu, Yuanzhi Cai, Yuan Fang, Yihan Yang, Cheng Chen, Lei Fan, and Anh Nguyen. Samba: Semantic segmentation of remotely sensed images with state space model. *Heliyon*, 10(19), 2024. 17