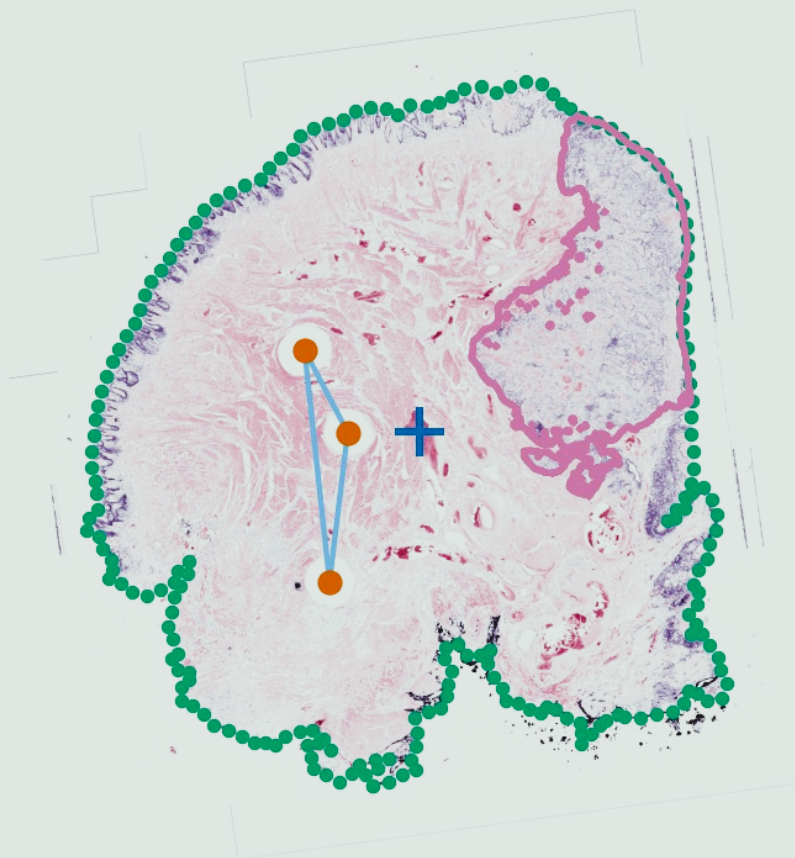


Bridging the Gap: Semi-Automated 3D Histopathology-to- Ultrasound Registration

Towards Intraoperative Margin Assessment in Oral Cancer Surgery

J.M.J. van Beurden



Bridging the Gap: Semi-Automated 3D Histopathology-to- Ultrasound Registration

Towards Intraoperative Margin Assessment in Oral
Cancer Surgery

by

J.M.J. van Beurden

Supervised by: Maarten van Alphen, Theo van Walsum, Luc Karssemakers, Andrea Borghesi and George Yiasemis
27-08-2026

Faculty of Technical Medicine: Delft, Leiden and Rotterdam

Preface

This thesis marks the end of my Master's in Technical Medicine, and with it the end of my student years, which took me from Delft to Rotterdam and finally to Amsterdam.

The project was quite specific, but I was given a lot of freedom in it, and I ended up exploring a broad set of topics, partly because of the diverse backgrounds in my supervisory team. I enjoyed how much of what I learned during my studies came back in this project.

I would like to thank my supervisors. Maarten, thank you for introducing me to this project and for your daily guidance, both technical and clinical. Your ideas, broad knowledge, and steady positive mindset kept me going. Theo, thank you for your critical eye and outsider's perspective; your questions pushed me into other directions as well. Luc, thank you for the clinical side and for reminding me who this work is ultimately for, whether in the OR, the clinic, or elsewhere. Andrea, we started on the exact same day, you as a PhD student, me as a Master's student, from completely different backgrounds, yet on the same projects. I learned a great deal from you, from your small lectures on AI and coding to testing the setup on everything from cherry tomatoes to lung specimens. George, although you were not one of my direct supervisors, thank you for showing up every time and for your consistently positive and valuable input.

I had heard good stories about the Antoni van Leeuwenhoek before I started, and the experience turned out to be even better than the stories. Everyone was eager to join a discussion and quick to help, and there was always a welcoming atmosphere. The other students at the Head and Neck Surgery department made my time there even more enjoyable.

Finally, I would like to thank my friends and family for their support and welcome distractions.

*J.M.J. van Beurden
Amsterdam, August 2026*

Abstract

Introduction: Surgical resection is the primary treatment for tongue squamous cell carcinoma (TSCC), but inadequate margins occur in up to 85% of cases in Dutch cohorts. Histopathology is the gold standard for margin assessment, but its result only becomes available days after surgery, leaving the surgeon without intra-operative feedback. A deep learning model segmenting the tumour in ex-vivo 3D ultrasound (US) could give intra-operative feedback, but needs training labels with a known tumour boundary in US space. Histopathology provides that boundary in a separate coordinate system, so transferring it requires image registration.

Methods: A semi-automated 3D-to-3D registration pipeline was developed for oral-cavity specimens. A 3D histopathology volume is reconstructed from serial 2D H&E slices, refined per slice using internal fiducial needles, and registered as a whole to the 3D US volume, held as the fixed reference. Registration estimates similarity transforms by aligning the specimen boundaries through signed distance fields. The pipeline was validated on a development and a held-out cohort without parameter re-tuning, using fiducial TRE, landmark TRE, NSD@1mm, and Dice; individual components were assessed through ablation and sensitivity experiments.

Results: On the development cohort ($n = 5$), the pipeline reached a mean fiducial TRE of 1.56 mm, a mean landmark TRE of 1.69 mm, an NSD@1mm of 0.89, and a whole-specimen Dice of 0.80. On the held-out cohort ($n = 7$), accuracy remained in similar range (2.13 mm, 2.23 mm, 0.87, 0.74).

Conclusion: To our knowledge, this is the first pipeline to register a reconstructed 3D histopathology volume to 3D US for oral-cavity specimens, with slice correspondence recovered semi-automatically rather than fixed by hand for each specimen. The current accuracy of roughly 2 mm limits direct margin assessment at the 1 mm threshold, but provides a scalable framework and an accuracy baseline for generating registered training data.

Contents

Preface	i
Abstract	ii
Nomenclature	v
General Introduction	1
0.1 Clinical context	1
0.2 Post-operative margin assessment	1
0.3 Problem statement	2
0.4 Technical background	3
0.4.1 US imaging	3
0.4.2 Image registration	4
0.4.3 Prior registration approaches	6
0.5 Aim and contribution	7
1 Introduction	9
2 Materials and Methods	11
2.1 Overview	11
2.2 Histopathology	11
2.2.1 Preprocessing and landmark extraction	11
2.2.2 Initial 2D registration	13
2.2.3 Fiducial-aware joint refinement	13
2.2.4 Volume reconstruction	14
2.3 Histopathology-to-US registration	15
2.3.1 Initial 3D alignment	15
2.3.2 3D similarity refinement	15
2.3.3 Per-slice 2D refinement	15
2.4 Composition into US space	15
3 Experiments and Results	17
3.1 Data	17
3.1.1 Cohort and demographics	17
3.1.2 Ultrasound acquisition	17
3.1.3 Histopathology acquisition	17
3.1.4 Annotations	17
3.2 Implementation	18
3.3 Evaluation metrics	19
3.3.1 Fiducial-based accuracy	19
3.3.2 Anatomical landmark accuracy	19
3.3.3 Specimen mask overlap and shrinkage	19
3.4 Experiments	19
3.4.1 Experiment 1: Ablation of pipeline components	19
3.4.2 Experiment 2: Regulariser sensitivity analysis	19
3.4.3 Experiment 3: 3D refinement transform model	20
3.4.4 Experiment 4: Comparison against manual initialisation	21
3.4.5 Experiment 5: Validation on the development cohort	22
3.4.6 Experiment 6: Validation on the evaluation cohort	23
3.5 Shrinkage	23
3.6 Qualitative results	23

4	Discussion	26
4.1	Principal findings	26
4.2	Interpretation of results	26
4.3	Comparison to prior work	27
4.4	Strengths and limitations	28
4.5	Clinical value	29
4.6	Future work	29
5	Conclusion	31
	References	32
A	Appendix: Full Results Experiments	36
B	Appendix: Flips and Rotations Results	38
C	Appendix: Per-Slice Information	39
D	Appendix: Error Direction Analysis	42
E	Appendix: Full Per-Slice 2D Refinement Results	44
F	Appendix: Full 2D Qualitative Results	48
G	Appendix: Standard Operating Procedure	53
H	Appendix: Adding and Running a New Patient in the Pipeline	57

Nomenclature

Abbreviations

Abbreviation	Definition
ASSD	Average Symmetric Surface Distance
DOF	Degrees of Freedom
DRMIME	Differentiable Mutual Information and Matrix Exponential
GUI	Graphical User Interface
H&E	Haematoxylin and Eosin
HD95	95th-percentile Hausdorff Distance
ICP	Iterative Closest Point
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
mTRE	mean fiducial-line Target Registration Error
NSD	Normalised Surface Dice
OR	Operating Room
OSCC	Oral Squamous Cell Carcinoma
PCA	Principal Component Analysis
PET	Positron Emission Tomography
PZT	Lead Zirconate Titanate
QC	Quality Control
RGB	Red, Green, Blue
SDF	Signed Distance Field
SOP	Standard Operating Procedure
SVD	Singular Value Decomposition
TRE	Target Registration Error
TRUS	Transrectal Ultrasound
TSCC	Tongue Squamous Cell Carcinoma
US	Ultrasound
WSI	Whole-Slide Image

Symbols

Symbol	Definition	Unit
$\mathbf{c}_k$	Centroid of the k -th fiducial line	
d	Depth of a reflector	[mm]
$\mathbf{d}_k$	Direction of the k -th fiducial line	
F	Fixed image	
$\mathcal{F}$	Set of slices optimised in the joint refinement	
$\mathbf{h}_{i,k}$	Fiducial-hole centroid in slice i , trajectory k	
i	Slice index	
k	Fiducial-trajectory index	
M	Moving image	
$\mathcal{M}$	Specimen mask	
n	Number of specimens (cohort size)	
p	Statistical significance (p -value)	
r_s	Spearman rank correlation coefficient	
R	Half of the larger side of the padded image	[mm]

Symbol	Definition	Unit
S_k	Set of slices containing fiducial trajectory k	
T	Similarity transform	
$\mathbf{t}_i$	Translation of slice i	[mm]
v	Speed of sound in tissue	[m/s]
$\mathbf{x}$	Voxel (point) coordinate	
Z	Acoustic impedance	[kg m ⁻² s ⁻¹]
Δt	Round-trip echo travel time	[s]
θ_i	Rotation angle of slice i	[deg]
μ	Joint-refinement regulariser weight	
ρ	Tissue density	[kg m ⁻³]
τ	Surface-distance tolerance (NSD)	[mm]
Φ	Signed distance field	
ω	Rotation-to-length weighting factor	
Ω	Shared voxel grid (image domain)	

General introduction

0.1. Clinical context

Within head and neck oncology, the oral cavity is the most commonly affected site, with tongue squamous cell carcinoma (TSCC) being the most frequent tumour type [2, 3, 4]. In the Netherlands, according to the registry figures from the Netherlands Comprehensive Cancer Organisation (IKNL), 976 patients were diagnosed with cancer of the oral cavity in 2025, with 421 of these cases involving a tumour of the tongue [5]. The five-year relative survival rate for these patients is around 60%, and both incidence and mortality are projected to rise in the coming decades [5, 6, 7]. Because the tongue is central to essential daily activities like speech, chewing, swallowing, and taste, treatment involves a trade-off: surgeons must achieve complete tumour removal while preserving as much functional tongue tissue as possible to protect the patient's quality of life [7].

Anatomically, the tongue is a complex muscular organ covered by a thin mucosal layer (Figure 1). It is composed of both extrinsic muscles, which control its position and overall movement, and intrinsic muscles, which contract to change its shape and give it its wide range of motion [8]. Dense vascular networks and hundreds of minor salivary glands run through this muscular structure [9, 10]. This mixed tissue composition affects how the tongue appears on imaging: on ultrasound (US), tongue cancer typically appears as a dark (hypoechoic) and irregular region [3, 11, 12] (Figure 2). Because surrounding healthy salivary glands and blood vessels can look similar, even experienced specialists show significant variation when trying to manually trace where the tumour ends and healthy tissue begins [3, 13].

Surgery is the primary curative treatment for resectable TSCC [14, 15]. During the procedure, the surgeon aims to remove the entire tumour along with a safety rim of healthy tissue, while striving to preserve as much functional tongue tissue as possible [3]. Once removed, the tissue is analysed under the microscope by a pathologist. According to the guidelines of the Royal College of Pathologists, surgical margins are classified based on the distance between the tumour cells and the outer cut surface: a margin of more than 5 mm is considered “clear” (adequate), 1 to 5 mm is “close,” and less than 1 mm is “positive” [16]. Notably, in daily clinical practice, up to 85% of oral cavity cancer resections result in inadequate (close or positive) margins as reported at two Dutch academic medical centres [17]. Inadequate margins are most strongly associated with local recurrence at the primary site, and through

poorer loco-regional control, with reduced overall survival [2, 18]. Correcting them typically requires a second surgery or adjuvant (chemo)radiotherapy, both of which impose a substantial physical and functional burden on top of the initial surgery [3, 17].

0.2. Post-operative margin assessment

After surgery, the resected specimen is sent to the pathologist, where the surgical margin is assessed histologically. This is the reference against which all intra-operative techniques are compared, and it is the gold standard for margin assessment [4, 19]. However, the process takes several days, and therefore, the surgeon has no feedback while the operation is still in progress [3, 20].

Before microscopic examination, the specimen goes through a fixed sequence of processing steps (Figure 3). First, the specimen is submerged in neutral buffered formalin for roughly 24 to 72 hours, depending on its thickness [10]. This preserves the tissue and prevents further decomposition. Next, in a step called grossing, the pathologist cuts the fixed specimen into parallel sections of about 4 mm thick, perpendicular to the specimen's long axis. The long axis is the anterior–posterior direction, oriented using surgical sutures placed during resection. Each section is placed in a small cassette, dehydrated, and embedded in paraffin wax, forming a hard block that can be cut cleanly. From each block, a microtome is used to shave off a single thin slice of only 4 μm , one-thousandth the thickness of the original section. This slice is mounted onto a glass slide, stained with haematoxylin and eosin (H&E) to colour the cell nuclei and cytoplasm, and finally digitised at high magnification into a whole-slide image (WSI) with a resolution of around 0.24 μm per pixel [3]. The pathologist views the WSI on a computer to measure the distance from the tumour to the cut surface.

Because slicing is done perpendicular to the long axis of the specimen, the position of each section along that axis is roughly known: sections are numbered in order, and their spacing corresponds to the 4 mm grossing interval. The exact position, however, still depends on the angle at which the specimen was cut, small deviations from truly perpendicular slicing cause the plane of each section to tilt slightly relative to the specimen [3, 21]. The original 3D coordinate system of the specimen can be disrupted during processing [3]. When a section is transferred into a paraffin block, the technician positions the tissue by hand, at an arbitrary

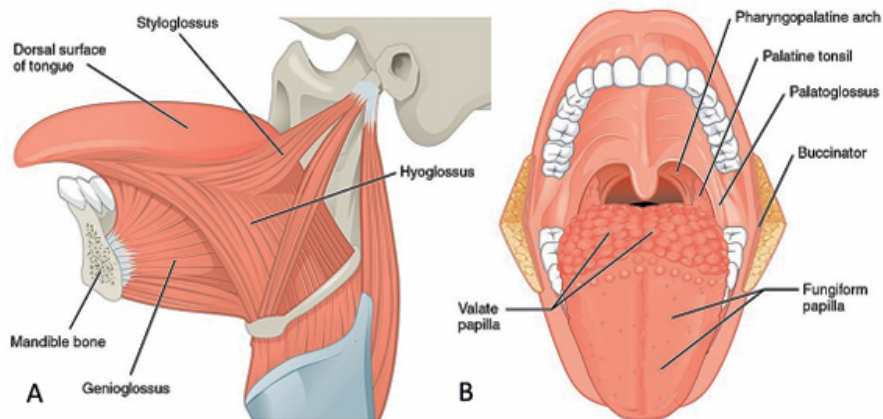


Figure 1: Anatomy of the tongue and oral cavity. (A) Side view showing the extrinsic muscles of the tongue (styloglossus, hyoglossus, genioglossus). Intrinsic muscles sit within the tongue itself. (B) Frontal view of the oral cavity, showing the tongue surface with its papillae and the surrounding structures. Adapted from OpenStax Anatomy and Physiology [1], licensed under CC BY 3.0.

rotation and offset within the mould. The 4 μm slice cut from that block therefore starts in a new, unrelated coordinate frame. On top of this, the tissue can shift within its cassette during dehydration, the microtome blade can shave the block at a slightly skewed angle, and the thin slice can be stretched, rotated or flipped upside down when it is picked up from the water bath [21]. By the time a whole-slide image is available, the exact in-plane position and orientation of the tissue relative to the original specimen are lost, along with which face of the slice lies against the glass.

The specimen also shrinks between resection and the final stained slide. In a study of TSCC specimens at the Netherlands Cancer Institute, Dijkhuis reported a cumulative shrinkage of approximately 29% from fresh tissue to the final stained slide, with most of this occurring immediately after resection due to tissue recoil rather than during formalin fixation [11]. Similar magnitudes have been reported by others [22, 23]. Because the 5 mm margin threshold is applied to measurements made on the shrunk tissue, the true fresh-tissue margin at the time of surgery is larger than the margin recorded on the histopathology form [24, 25]. This does not reduce to a single correction factor; shrinkage is non-uniform and anisotropic, varying between patients, between anatomical axes, and between tissue types, with dense tumour shrinking less than the surrounding muscle [11, 22]. To compensate for post-operative tissue shrinkage, surgeons routinely resect a wide macroscopic margin of 10 mm or more of healthy tissue around the tumour [3, 10].

0.3. Problem statement

Histopathology's multi-day delay means the gold-standard margin result is unavailable while the surgeon

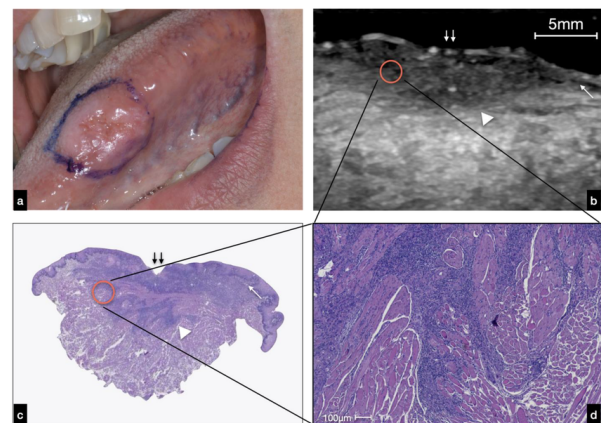


Figure 2: The same tongue tumour seen with different techniques. (a) Clinical view of a TSCC at the right tongue margin. (b) US evaluation showing an interruption of the covering mucosa (double arrow), with the red circle marking the infiltrative tumour region. (c) Histopathology of the same slice, the red circle marking the same region. (d) Zoomed-in histopathology view of the red-circled region. Reproduced from Di Stasio et al. [12], licensed under CC BY 4.0.

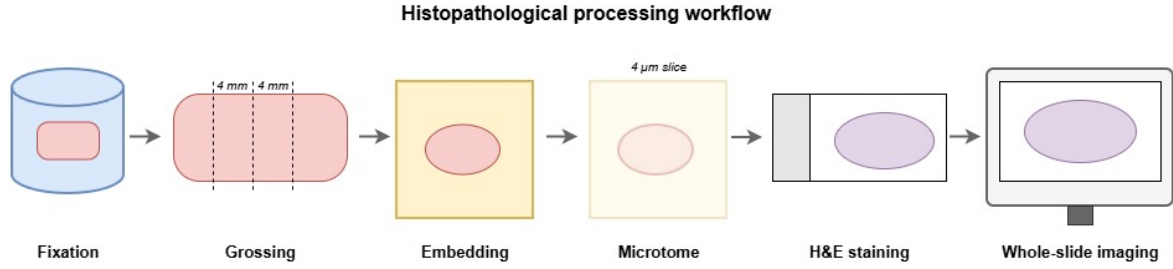


Figure 3: Histopathological processing workflow for a resected tongue specimen: formalin fixation, grossing into parallel ~ 4 mm sections, paraffin embedding, microtome slicing, H&E staining, and digitisation into a whole-slide image. Only one thin slice is retained per section.

can still act on it. Ex-vivo 3D US is fast enough to image the resected specimen and give feedback during surgery, but converting a raw 3D volume into a margin measurement still requires the tumour to be delineated slice by slice in the 2D images that make up the volume; a step that is time-consuming and highly dependent on the observer. Deep learning models such as the 3D nnU-Net can in principle perform this segmentation automatically within seconds, and early work has applied them to ex-vivo US margin assessment [26, 27]. Model performance, however, is bounded by the quality of the training labels, and for tongue cancer this is where the process currently breaks down.

The tumour boundary itself is often ambiguous on US (Figure 2 b): as noted, TSCC blends into structures with a similar appearance, so annotators vary in where they place the edge. Beyond this variability, the dark region on US also includes peritumoural inflammation, so the boundary an annotator draws tends to be wider than the true tumour extent on histopathology. For these reasons manual annotations of TSCC on US vary substantially between observers. When the authors in Bekedam et al. [26] trained a 3D nnU-Net for tumour segmentation using such manual annotations, the resulting margin measurements correlated only weakly with histopathology ($r = 0.12$) [26], too low to guide surgical decisions.

Histopathology provides an objective, microscopic-level boundary of the tumour, but sits in a different coordinate system than the US volume, as described in Section 0.2. To use histopathology annotations as ground truth for training, they must first be moved into the coordinate system of the US volume. Image registration is a general technique for spatially aligning data, and can be applied to this problem. For tongue specimens, a first step in this direction was made by Bekedam [3], who manually paired each histopathology slide with its corresponding US slice and performed a 2D-to-2D registration [28]. Manual pairing scales poorly, however, which limits its use for building the larger labelled datasets that deep learning models require.

This thesis addresses that gap by developing a semi-automated 3D-to-3D registration pipeline for oral-cavity specimens. Rather than pairing each slide to a US slice by hand, it aligns a reconstructed 3D histopathology volume as a whole to the corresponding 3D US volume, transferring the tumour annotation into US space to serve as a training label.

0.4. Technical background

0.4.1. US imaging

US imaging is based on the pulse-echo principle. Short pulses of high-frequency sound are emitted into the tissue, and the returning echoes are used to reconstruct an image. Both emission and reception rely on the piezoelectric effect: piezoelectric crystals inside the transducer, most commonly lead zirconate titanate (PZT), physically deform when a voltage is applied across them, generating a pressure wave, and conversely produce a voltage when they are struck by a returning wave [29]. The scanner alternates rapidly between these two modes for each element in the transducer.

Once a pulse is emitted, it propagates through soft tissue as a longitudinal wave at an approximately constant speed $v \approx 1540$ m/s [29]. At every interface between two tissues with different acoustic properties, part of the wave is reflected and part is transmitted. The relevant physical quantity is the acoustic impedance Z , defined as the product of tissue density ρ and the speed of sound in that tissue:

$$Z = \rho \cdot v. \quad (1)$$

For a wave meeting an interface between two tissues with impedances Z_1 and Z_2 at normal incidence, the fraction of intensity that is reflected is given by

$$R = \left(\frac{Z_2 - Z_1}{Z_2 + Z_1} \right)^2, \quad (2)$$

so a large impedance mismatch produces a strong echo, whereas nearly matched tissues reflect little of the incoming signal [29].

The scanner reconstructs a two-dimensional greyscale image, referred to as B-mode (brightness mode), by measuring the round-trip travel time Δt of each echo and computing the depth of the corresponding reflector:

$$d = \frac{v \cdot \Delta t}{2}. \quad (3)$$

The amplitude of each returning echo is then mapped to a pixel intensity: high-amplitude echoes appear as bright, hyperechoic pixels; low-amplitude echoes as darker, hypoechoic pixels; and regions producing no returning signal, such as fluid-filled cavities, as black (anechoic) [29]. A linear-array transducer sweeps this measurement across many parallel lines of sight by sequentially firing subsets of its elements, and the resulting 1D depth profiles are assembled side by side to form the final 2D image.

Because acoustic impedance depends on tissue microstructure, muscle, salivary glands, blood vessels, and connective tissue all produce their own characteristic patterns of internal echoes. TSCC appears as a hypoechoic and heterogeneous region, as introduced in Section 0.1. The hypoechoic appearance reflects that malignant tissue has a higher water content and a lower, more homogeneous acoustic impedance than the surrounding dense muscle, so relatively little sound is reflected back to the transducer [3]. The heterogeneous, speckled pattern within the tumour reflects the disorganised microstructure of a growing carcinoma, a chaotic mixture of tumour nests, inflammation, stromal reactions, and micro-vessels acting as sub-wavelength scatterers [3, 12].

Central to this thesis is that the intensity pattern inside a specimen on US is set by acoustic scattering and impedance contrast, whereas on H&E-stained histopathology it is set by how strongly different tissue components take up the stains. Although the same structures may be visible in both, their intensities do not correspond: a given tissue's brightness on US bears no consistent relation to its appearance on H&E [30]. What they do reliably share is the outer boundary.

Assessing the margin over the entire specimen requires a 3D US volume rather than a single 2D frame. In the ex-vivo setup developed at the Netherlands Cancer Institute [3], the freshly resected specimen is submerged in a 0.9% saline bath, and a high-frequency linear-array transducer is translated over it by a motorised rail (Figure 4), so that 2D B-mode frames are acquired at a fixed through-plane spacing of 0.5 mm and stacked directly into a regular 3D array without interpolation. Because the probe never touches the specimen and the saline both transmits the sound and supports the specimen, the tissue is not compressed, and the contour geometry in the acquired 3D volume reflects the specimen's true shape. This is what makes the 3D US volume a stable

reference frame into which histopathology annotations can be transferred.

0.4.2. Image registration

Image registration is the process of finding a spatial transformation that lines up one image with another [31, 32]. The two images play different roles: the fixed image F stays in place and defines the coordinate system, and the moving image M is transformed so that it lines up with the fixed image. Registration searches for the transform T that makes the two images look as similar as possible, measured by a similarity metric $\mathcal{S}$ over a shared domain Ω . This transform is found by an iterative optimiser that adjusts T to minimise $\mathcal{S}$.

$$T^* = \arg \min_T \mathcal{S}(F(x), M(T(x))), \quad x \in \Omega. \quad (4)$$

In this thesis, the 3D US volume represents the fixed image and the reconstructed 3D histopathology volume the moving image, so that the pathology annotations end up in the US coordinate frame.

Different transforms exist depending on how much freedom they give the image to move. A rigid transform only allows rotation and translation (six degrees of freedom (DOF) in 3D). Distances and angles stay the same, as if you were moving a solid object around. A similarity transform adds one isotropic scaling factor, so the object can also grow or shrink uniformly. An affine transform goes a step further and lets each axis stretch independently, and also allows shearing, but it is still a global transform applied everywhere in the same way. A deformable (non-rigid) transform lets each part of the image move on its own and can absorb local tissue deformations such as shrinkage or compression, but it is much harder to constrain and to check [31]. The registration in this thesis uses similarity transforms. The specimen envelope in the US is undeformed, and the dominant geometric difference between the two modalities is approximately global shrinkage of the histopathology tissue, which a similarity transform captures to first order. Because the registration is driven by the specimen contour, additional DOF are only weakly constrained in the interior, so a higher-DOF transform could distort internal tissue without changing the boundary the objective sees. We therefore compare the similarity baseline against a global affine model in the sensitivity analysis (Section 3.4.3).

The similarity metric decides which transform is considered best. The simplest option compares the voxel intensities of the two images directly. The mean squared error (MSE),

$$\text{MSE}(F, M \circ T) = \frac{1}{|\Omega|} \sum_{x \in \Omega} (F(x) - M(T(x)))^2, \quad (5)$$

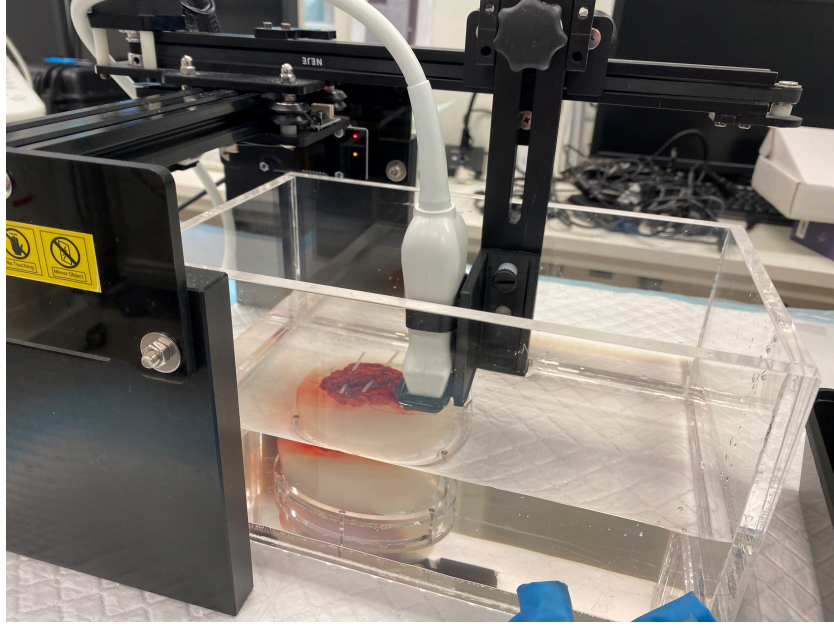


Figure 4: Motorised 3D US acquisition setup used at the Netherlands Cancer Institute. The resected specimen is submerged in a saline bath and imaged by a linear-array transducer translated over it by a motorised rail. Adapted from [3].

assumes that the two images should have the same intensity at the same location, which only works when both images come from the same modality. Normalised cross-correlation relaxes this to a linear relationship between intensities, and mutual information generalises it further to any statistical dependence, which is what makes mutual information the standard choice for multi-modal registration [31]. None of these metrics work well for aligning US to H&E histopathology though. As explained in Section 0.4.1, the two modalities do not share a reliable intensity relationship inside the specimen: they reflect different physical properties of the tissue and produce completely different internal patterns. What they do share is the outer boundary of the specimen. The registration in this thesis is therefore based on binary masks of the specimen envelope rather than on raw image intensities.

However, comparing binary masks directly does not work well either. The derivative of a step function is zero almost everywhere, so the optimiser has no gradient to follow whenever the two masks do not already overlap. A common way around this is to convert each mask into a Signed Distance Field (SDF) [33, 34].

The SDF $\Phi(x)$ gives, at every voxel, the distance to the closest point on the boundary $\partial\mathcal{M}$ of the specimen mask $\mathcal{M}$. The sign says whether the voxel sits inside

or outside the mask (Figure 5):

$$\Phi(x) = \begin{cases} -d(x, \partial\mathcal{M}) & \text{if } x \text{ is inside the mask,} \\ 0 & \text{if } x \text{ lies on the boundary } \partial\mathcal{M}, \\ +d(x, \partial\mathcal{M}) & \text{if } x \text{ is outside the mask,} \end{cases} \quad (6)$$

with $d(x, \partial\mathcal{M}) = \inf_{y \in \partial\mathcal{M}} \|x - y\|_2$.

A true signed distance field also satisfies $\|\nabla\Phi\|_2 = 1$ [34], which means its gradient has the same size everywhere and points straight towards the nearest boundary. This gives the optimiser a stable direction to move in, even when the two envelopes do not yet overlap. Because both SDFs encode the same geometric quantity, the distance to the specimen boundary, they are directly comparable value-for-value, and MSE becomes an appropriate metric again, now on the distance fields rather than raw intensities:

$$T^* = \arg \min_T \frac{1}{|\Omega|} \sum_{\mathbf{x} \in \Omega} (\Phi_{\text{US}}(\mathbf{x}) - \Phi_{\text{H}}(T(\mathbf{x})))^2, \quad (7)$$

where Φ_{US} and Φ_{H} are the SDFs of the US and histopathology specimen masks.

The quality of a registration is judged by how well corresponding structures align after the transform is applied. No single metric captures this fully, so registration accuracy is commonly reported with several complementary measures [35]. Three types of metric are used in this thesis. The Target Registration Error (TRE) measures the Euclidean distance between corresponding landmarks in the two images after registration [36]. For a landmark located at $x_{F,i}$ in the fixed image and

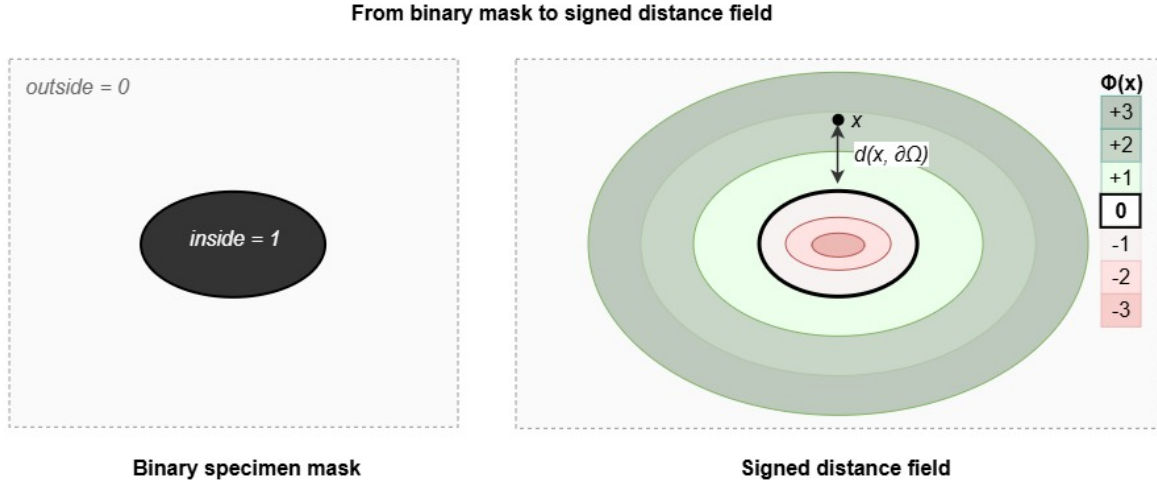


Figure 5: Illustration of a signed distance field. Left: a binary mask of a specimen cross-section. Right: the corresponding signed distance field. Unlike the binary mask, the signed distance field varies smoothly across the whole volume.

its corresponding landmark at $x_{M,i}$ in the moving image, the TRE is

$$\text{TRE}_i = \|\mathbf{x}_{F,i} - T(\mathbf{x}_{M,i})\|_2 \quad (8)$$

and is reported as the mean over all validation landmarks, in millimetres. TRE is computed on anatomical landmarks that are not used to drive the multimodal registration: they are annotated independently and withheld from the optimisation of T , providing an accuracy measure that is independent of the fit. The Dice similarity coefficient measures the volumetric overlap between two segmentations X and Y [37, 38]:

$$\text{Dice}(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|}, \quad (9)$$

ranging from 0 (no overlap) to 1 (perfect overlap). It is sensitive to overall placement but forgiving of small local errors. The normalised surface Dice (NSD) measures the fraction of the boundary surface of one segmentation that lies within a given tolerance τ of the other [39]:

$$\text{NSD}(X, Y; \tau) = \frac{|\{x \in \partial X : d(x, \partial Y) \leq \tau\}| + |\{y \in \partial Y : d(y, \partial X) \leq \tau\}|}{|\partial X| + |\partial Y|},$$

where ∂X and ∂Y are the boundary surfaces of the two segmentations. Unlike Dice, NSD is directly interpretable in millimetres through its tolerance parameter. In this thesis τ is set to 1 mm, corresponding to the clinical margin below which a resection is considered positive.

0.4.3. Prior registration approaches

For tongue specimens, Bekedam et al. [28] manually paired each histopathology slide with an US slice and fitted a 2D–2D transform per pair. Most other image-to-histopathology registration work comes from the prostate. Two studies register to: Schalk et al. [30] registered 3D prostate models to Transrectal Ultrasound (TRUS) with an affine iterative closest point (ICP) initialisation and elastic surface refinement, and Jager et al. [40] extended this to a rigid ICP initialisation with a finite-element deformation model that also compensates for probe pressure. The others register to MRI, either as the target or as an intermediate: Rusu et al. [41] reconstructed prostatectomy specimens in 3D and applied consecutive 2D transforms to pre-surgical MRI, Pensa et al. [42] fused micro-US and MRI by landmark registration and mapped pathology with a capsule-based thin-plate spline, and Ghezzi et al. [43] bridged histology to in-vivo PET/MRI through an ex-vivo MRI with fiducial strands and B-spline mapping. Table 3 summarises the reported accuracies. All of these either pair slices manually or fix the pairing at acquisition; none registers a reconstructed 3D histopathology volume to a 3D US volume with the pairing solved automatically, which is the gap this thesis addresses.

Table 3: Comparison of multimodal image-to-histopathology registration methods. The proposed pipeline (van Beurden) is contrasted with existing approaches by imaging setting, registration dimensionality, methodology, and target registration error (TRE), reported as the mean unless marked median (med.). † mean fiducial mTRE; ‡ mean anatomical landmark TRE; med., median.

Study	Setting	Dimensions	Approach	TRE (mm)	Cohort
Jager et al. (2025) [40]	Prostate; 3D TRUS (in-vivo)	3D–3D	Rigid (ICP) + non-rigid (FEM), probe comp.	3.5	$n = 20$
Schalk et al. (2016) [30]	Prostate; TRUS (in-vivo)	3D surface	Affine ICP + elastic surface	2.1	$n = 7$
Pensa et al. (2024) [42]	Prostate; microUS & MRI (in-vivo)	3D + 2D	3D landmark + 2D capsule (TPS)	3.9	$n = 15$
Bekedam et al. (2024) [28]	Tongue; 3D US (ex-vivo)	2D–2D	Landmark + fiducial matrix	0.42 (med.)	$n = 8$
Rusu et al. (2020) [41]	Prostate; MRI (in-vivo)	3D recon $\rightarrow$ 2D	2D affine + deformable	2.8	$n = 157$
Ghezzi et al. (2023) [43]	Prostate; PET/MRI (in-vivo)	2D $\rightarrow$ 3D	Semi-auto landmark + B-spline	1.59 (med.)	$n = 3$
van Beurden (2026), dev.	Tongue; 3D US (ex-vivo)	3D–3D	Fiducial 2D pre-align + 3D/2D similarity (SDF)	1.56 fid.† 1.69 lm.‡	$n = 5$
van Beurden (2026), eval.	Tongue; 3D US (ex-vivo)	3D–3D	Fiducial 2D pre-align + 3D/2D similarity (SDF)	2.13 fid.† 2.23 lm.‡	$n = 7$

0.5. Aim and contribution

The purpose of this work is to transfer gold-standard histopathology annotations of TSCC into the corresponding 3D US volume, so that they can serve as training labels for a downstream deep learning segmentation model. Building such a training set requires many specimens, so the transfer must be reliable and scalable. This leads to the central research question:

How can 2D histopathology annotations of tongue squamous cell carcinoma be reliably transferred into the 3D US volume of the same specimen?

This question is addressed through three objectives:

- Develop a semi-automated 3D-to-3D registration pipeline that reconstructs a 3D histopathology volume from serial 2D H&E slices and registers it as a whole to the corresponding 3D US volume of the same specimen (Chapter 2).
- Validate the pipeline on a development cohort and evaluate on a held-out cohort without re-tuning parameters (Chapter 3).
- Produce an initial, proof-of-concept dataset of registered histopathology-to-US labels for training downstream tumour segmentation models (Section 3.6 and Appendix F).

Clinical implications and directions for future work are discussed in Chapter 4.

Introduction

Oral squamous cell carcinoma (OSCC) is the most prevalent malignancy of the head and neck, of which tongue squamous cell carcinoma (TSCC) is the most common cancer of the oral cavity [2, 3, 44]. Surgical resection remains the primary curative treatment, with the oncological goal of achieving clear resection margins, defined as greater than 5 mm of tumour-free tissue [16]. However, obtaining clear margins is notoriously difficult; while international studies report inadequate margin rates between 30% and 65%, this rate reaches up to 85% in a retrospective cohort across two Dutch academic medical centres, where inadequate margins are strongly associated with increased local recurrence and reduced survival [17].

In current clinical practice, postoperative histopathological examination remains the gold standard for margin assessment [3, 4]. However, this assessment cannot be completed during surgery. As a result, surgeons receive no intraoperative feedback on margin status, and patients often wait several days before definitive margin assessment is available. Correcting an inadequate margin therefore requires adjuvant (chemo)radiotherapy or a second surgery [45]. Adjuvant (chemo)radiotherapy of the head and neck carries substantial morbidity, including mucositis, xerostomia, and an increased risk of osteoradionecrosis [13, 46]. These toxicities impair tongue mobility, speech, and swallowing, imposing a considerable functional and quality-of-life burden on the patient [4].

Re-resection, when feasible, is complicated because the wound bed has already been closed and changed by the time definitive histopathology is available. Beyond the uncertainty in relocating the involved margin [20, 47], re-excision is then performed around the scar with a wide macroscopic margin to be sure of clearing any residual tumour, removing considerably more tissue than a correct primary resection would have [10]. Intraoperative margin assessment would instead allow immediate extension of the resection within a single procedure, preserving anatomical orientation, maximising functional tongue tissue, and reducing the need for adjuvant treatment [3, 13].

Several intraoperative assessment tools have been developed. Frozen section analysis is widely used but

only a few margins can be analysed; more importantly, this procedure increases anaesthesia time by 30–60 minutes [4, 19, 48]. Fluorescence imaging visualises the resection surface but lacks depth penetration; the specimen must be cross-sectioned to determine the actual margin in millimetres, which disrupts anatomical orientation [4, 45]. Raman spectroscopy offers high sensitivity and specificity but relies on point measurements that may miss irregular tumour spread [49]. Magnetic Resonance Imaging (MRI) preserves anatomical orientation and visualises the entire specimen, but it is costly and presents significant logistical hurdles for real-time operating room integration [13, 50].

Ultrasound (US) is a highly accessible and cost-effective imaging modality capable of visualising deep tumour margins. However, conventional 2D US is highly operator-dependent, and freehand acquisition applies probe pressure to the specimen, deforming the tissue and distorting margin measurements. Two-dimensional images also fail to preserve the geometric relationships between adjacent slices, making it difficult to relocate specific margins. To address these limitations, a motorised 3D US setup was developed at the Netherlands Cancer Institute, in which the specimen is submerged in saline and scanned without probe contact, producing a stable, reproducible 3D volume of the entire specimen [51].

Prior studies have shown that ex-vivo US can visualise tongue tumours and that US-based margin measurements correlate with histopathology, supporting its use for intraoperative margin assessment [13, 28]. However, TSCC appears in US as a hypoechoic region with heterogeneous characteristics that can be difficult to distinguish from adjacent glandular or vascular tissue, and manual delineation is known to introduce substantial interobserver variability [28]. Automated tumour segmentation, however, requires labelled reference datasets in which the true tumour extent is known within the US volume. Histopathology provides the gold standard for tumour delineation, but is acquired postoperatively and in a separate coordinate system; constructing such a ground-truth reference dataset therefore requires transferring histopathologi-

cal ground truth into the US volume.

Histopathology and US cannot be compared directly. The two modalities image the same tissue but respond to different physical properties, acoustic scattering versus stain affinity, so their internal intensity patterns do not correspond [30]. The geometry is disrupted as well: during pathological processing, the specimen is cut into sections and slices are digitised in a fixed order, but the position and orientation of each slice within the original specimen is lost. In addition, the specimen deforms during handling, fixation, and sectioning, so its geometry at the time of histological digitisation no longer matches its geometry at the time of US acquisition. Mapping pathological annotations onto US therefore requires more than a rigid transformation: each slice must be localised within the volume, and tissue deformation between the two modalities must be compensated.

This thesis presents, to our knowledge, the first semi-automated 3D-to-3D registration pipeline for oral-cavity specimens that reconstructs a 3D histopathology volume from serial 2D H&E slices and registers it to the intraoperative 3D US volume, held as the fixed reference. This transfers the histopathological ground truth into US space, so that pathological tumour and margin annotations can serve as training labels for data-driven tools for intraoperative margin assessment in tongue cancer surgery. This framework is one component of an envisioned intraoperative workflow; the full Standard Operating Procedure (SOP) is outlined in Appendix G, assuming all components are trained and integrated. The main contributions are: (i) a registration method that recovers the slice correspondence between histopathology and 3D US semi-automatically, enabling application at the cohort sizes needed for downstream deep learning; (ii) a multi-metric validation framework applied to a development cohort and a held-out cohort without parameter re-tuning; and (iii) an initial dataset of registered histopathology-to-US labels for training downstream tumour segmentation models for intraoperative margin assessment in tongue cancer surgery.

The remainder of this manuscript is organised as follows. Chapter 2 describes the registration pipeline. Chapter 3 presents the experiments, including sensitivity and ablation analyses and a comparison against manual initialisation, with validation on the development cohort and evaluation on the held-out cohort. Chapter 4 discusses the principal findings, compares them with prior work, and outlines directions for future work.

Materials and Methods

2.1. Overview

The registration pipeline proposed in this work is designed to map 2D H&E histopathology images into the coordinate system of the 3D US imaging of the corresponding ex-vivo specimen. The pipeline consists of eight sequential steps divided into two main stages (Figure 2.1). In the first stage (steps 1–5), the 2D histopathology slices are preprocessed, landmark features are extracted, and the slices are stacked forming a reconstructed 3D volume. In the second stage (steps 6–8), this reconstructed histopathology volume is aligned to the 3D US volume, which acts as the fixed reference system.

Two primary design choices guide this pipeline. First, the US volume is kept as the fixed reference while the histopathology volume acts as the moving image. The US geometry is spatially intact and free from physical distortion, whereas the histopathology tissue undergoes shrinkage, tearing, and potential tissue loss during fixation and slicing. Isotropic scale factors are introduced during both the 3D and per-slice 2D registrations to account for the global and in-plane shrinkage. The labels generated by the pipeline are intended to train a deep learning model for tumour and margin segmentation directly in the US volume. Second, the registration is driven by the specimen masks rather than the raw modality intensities. Multi-modal similarity metrics such as mutual information can align modalities with different contrast mechanisms when consistent structural correlates exist across them. B-mode US and H&E histopathology do not offer such correlates within the specimen. What is reliably shared between the two modalities is the tissue boundary. The registration therefore operates on SDFs computed from the binary specimen masks, giving a smooth optimisation surface that is robust to small errors in the mask boundary.

2.2. Histopathology

Prior to scanning the specimen with the 3D US set-up and bringing it to histopathology, up to three intravenous-tube fiducial needles are inserted through the specimen. The resulting circular holes appear

in each 2D H&E slice and are used as landmarks throughout the following steps. Figure 2.2 illustrates the histopathological fixation and slicing process.

2.2.1. Preprocessing and landmark extraction

The preprocessing and landmark extraction steps are illustrated in Figure 2.3. All numerical parameters for these steps are reported in Section 3.2. Preprocessing produces, for each slice, a binary tissue mask and a padded RGB image. The mask drives landmark extraction and the geometry of the subsequent stacking and registration; the RGB image is carried through unchanged and warped alongside the mask, contributing to quality-control overlays and the final rendered output.

Each slice was first downsampled to a fixed working resolution, coarse enough to keep processing tractable while preserving the fiducial holes and tissue contour. Near-black scanner-void pixels were converted to white so that they were not classified as tissue by the subsequent threshold. A binary tissue mask was then generated by global RGB thresholding. Small connected components caused by debris were removed and small internal holes were filled by a fixed area threshold. This threshold is set well below the cross-sectional area of the fiducial-tube tracts, so that the fiducial holes are preserved for later detection. All slices in the cohort were padded to a common image size required for stacking (Section 2.2.2). Inter-slice staining variation was reduced with the Macenko stain normalisation method [53], taking the slice with the largest stained tissue area in the cohort as reference.

Landmark extraction produces two sets of per-slice features used by the registration, the tissue centre and the fiducial-hole centroids, together with a sampled outer tissue contour retained for visual inspection only. The tissue centre was defined as the centre of mass of the binary mask and used to centre each slice within the common padded image (Section 2.2.2). The outer tissue contour was extracted with a marching-squares algorithm [54], applied to the largest connected component of the mask, and sub-sampled at uniform stride along the contour-point sequence for visualisation. Fiducial-hole centroids were detected as small enclosed regions

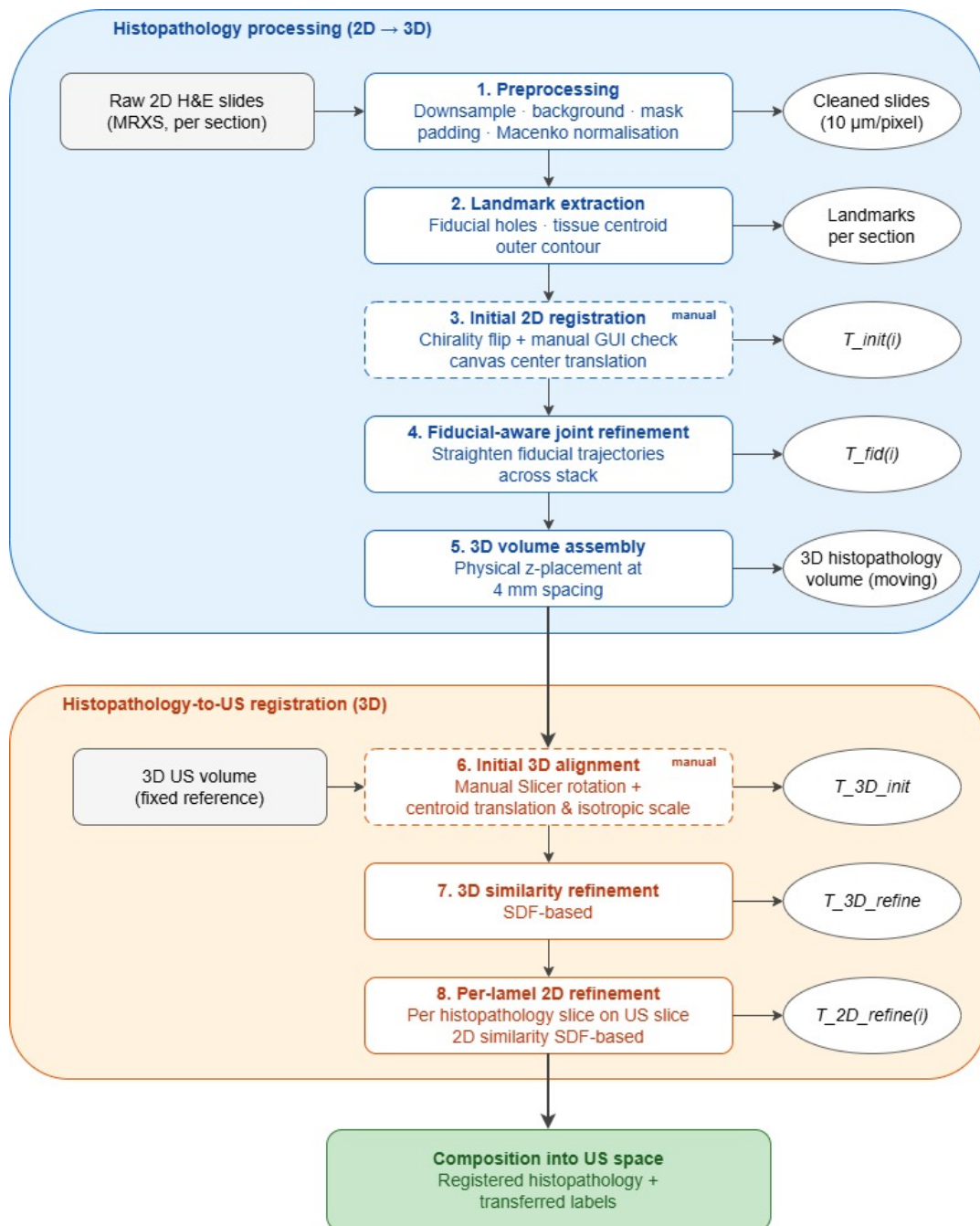


Figure 2.1: Overview of the registration pipeline. Left: the inputs to each step. Middle: the eight numbered steps, with arrows showing the flow between them; steps 1–5 reconstruct the 3D histopathology volume and steps 6–8 register it to the US volume. Right: the output of each step. Steps 3 and 6 include a manual review or initialisation; all other steps are automated.

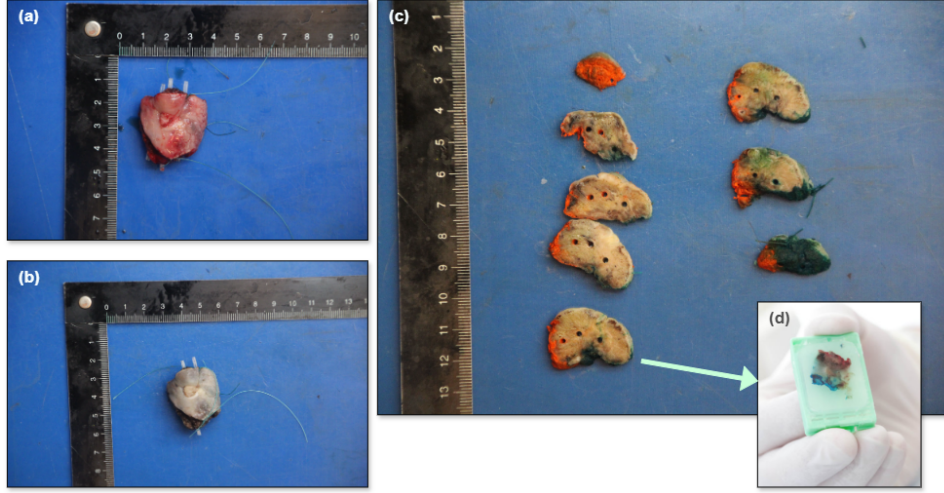


Figure 2.2: Histopathological processing of a resected tongue specimen (patient 2). (a) The fresh specimen prior to fixation, tumour facing up. Three orientation sutures are visible (long-long: anterior, top of the image; short-short: medial; long-short: posterior, bottom of the image) and the three inserted fiducials protrude from the top. (b) The same specimen after 72 hours of formalin fixation. (c) The fixed specimen cut into parallel sections, with inked surfaces marking orientation. (d) A paraffin-embedded tissue block from a different specimen. Panel (d) reproduced from [52].

within the tissue mask. The mask was inverted to find enclosed regions; the border-connected background region was excluded so that only true internal holes remained. Candidate holes were then retained if they satisfied area and circularity criteria chosen to preserve circular fiducial-tube cross-sections while rejecting elongated tears; if more than three candidates remained, only the three largest by area were kept. The detected fiducial-hole centroids were subsequently used to constrain the joint refinement of Section 2.2.3. Detection performance across the cohort is reported in Section 3.1.1.

2.2.2. Initial 2D registration

The initial 2D registration gives each slice into a canonical orientation and centring, so that fiducial holes can be matched across the stack during the joint refinement (Section 2.2.3). The chirality, meaning whether the slice is mirrored, arising from face-up or face-down mounting on the glass slide, of each slice was determined using the signed area of the triangle formed by its three fiducial holes. This was possible if at least three holes were detected and their configuration was non-degenerate. The canonical orientation was defined as the majority chirality across the stack, and slices with opposite chirality were vertically flipped. These flip decisions, together with any required discrete 90° rotation, could manually be altered through a graphical user interface (GUI); the frequency of manual overrides across the cohort is reported in Section 3.2. Each slice was then translated so that its tissue centre coincided with the image centre, yielding the initial 2D transform $T_{\text{init}}(i)$ per slice i .

2.2.3. Fiducial-aware joint refinement

The fiducial-aware joint refinement corrected the per-slice 2D poses so that the stacked fiducial holes traced straight lines through the 3D volume, matching the physically straight needles from which they originate. This straightness provided a geometric constraint on the alignment beyond what the initial alignment could enforce.

The slice with the most detected fiducial holes was chosen as the anchor, held fixed as the reference; ties were resolved toward the stack centre. For each non-anchor slice i containing at least one fiducial hole (single-hole slices are retained; see below), an additional rigid transform $T_{\text{fid}}(i) = (\theta_i, \mathbf{t}_i)$ was applied on top of the initial alignment of Section 2.2.2 and optimised against:

$$L = \sum_{k=1}^K \sum_{i \in S_k} \|(I - \mathbf{d}_k \mathbf{d}_k^\top)(T_i(\mathbf{h}_{i,k}) - \mathbf{c}_k)\|^2 + \mu \sum_{i \in \mathcal{F}} [\omega \theta_i^2 + \|\mathbf{t}_i\|^2], \quad (2.1)$$

where the first term minimises the perpendicular distance from each transformed fiducial centroid $T_i(\mathbf{h}_{i,k})$ to the best-fit 3D line $(\mathbf{c}_k, \mathbf{d}_k)$ computed for the k -th fiducial trajectory. This line is re-estimated at each iteration by total-least-squares fitting through the fiducial points [55]: the singular value decomposition (SVD) of the centred points gives the direction $\mathbf{d}_k$ (first right-singular vector) and $\mathbf{c}_k$ (the centroid). The second term is a small regularisation penalty, biasing each pose to

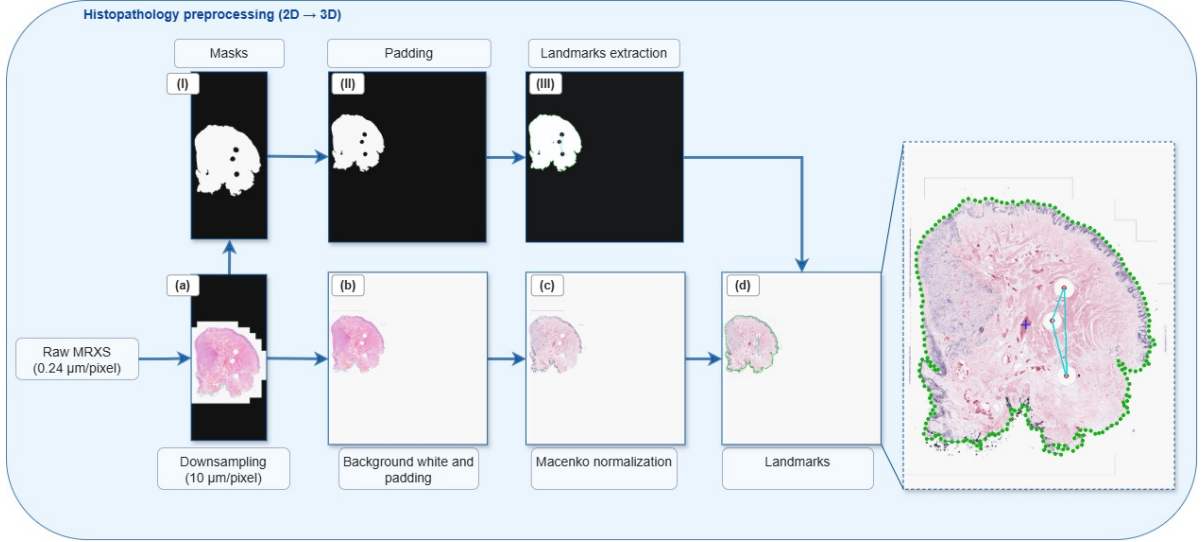


Figure 2.3: 2D preprocessing applied to each histopathology slice before 3D reconstruction. Top row (I–III): the binary tissue mask, shown as the raw mask (I), padded to the shared image size (II), and at the landmark-detection stage (III). Bottom row (a–d): the corresponding image data: (a) raw whole-slide scan downsampled to 10 $\mu\text{m}/\text{pixel}$ from the native 0.24 $\mu\text{m}/\text{pixel}$; (b) scanner background set to white and padded to the shared image size, tissue anchored top-left; (c) Macenko stain normalisation to a fixed H&E reference; (d) the preprocessed slice with detected landmarks: fiducial-hole centroids (red circles, connected by light blue lines forming a triangle), tissue centre (dark blue cross), and outer tissue contour (green points).

ward its initial alignment, so smaller pose changes are preferred. It is applied over the set of optimised slices $\mathcal{F} = \{i \neq i_{\text{anchor}} : \text{slice } i \text{ has } \geq 1 \text{ detected hole}\}$; slices without detected fiducial holes were excluded from the optimisation and held at the pose assigned by the initial 2D registration. This penalty keeps single-hole slices to their initial pose: a single hole contributes only two independent components of perpendicular distance to the fitted line, two equations for the three pose parameters (θ_i, t_x, t_y) , so the fiducial term alone leaves one degree of freedom unconstrained, which the regularisation constrains. The weighting factor $\omega = (R\pi/180)^2$, where R is half of the larger side of the shared padded image in mm, converts the rotation angle from degrees to a length scale: a rotation by θ degrees moves a point at radius R by $R\theta\pi/180$ mm, so a rotation by θ degrees and a translation of $R\theta\pi/180$ mm contribute equally to the regularisation.

Correspondence between fiducial holes across slices was established by starting from the anchor slice and tracking outwards, using the Hungarian algorithm [56] to match each subsequent slice’s holes.

2.2.4. Volume reconstruction

The refined 2D slices were assembled into a 3D volume by placing each slice at its anatomical z-position, so that this reconstructed 3D volume can be registered to the US volume (Section 2.3). Each slice was positioned using a per-specimen section-mapping protocol, and its 2D mask was swept through the 4 mm slab thickness from which the slice was cut. The result was a 3D

volume in which the specimen mask forms a continuous stack of slabs at their physical z-locations (Figure 2.4). The z-placement for slice i is denoted $T_z(i)$.

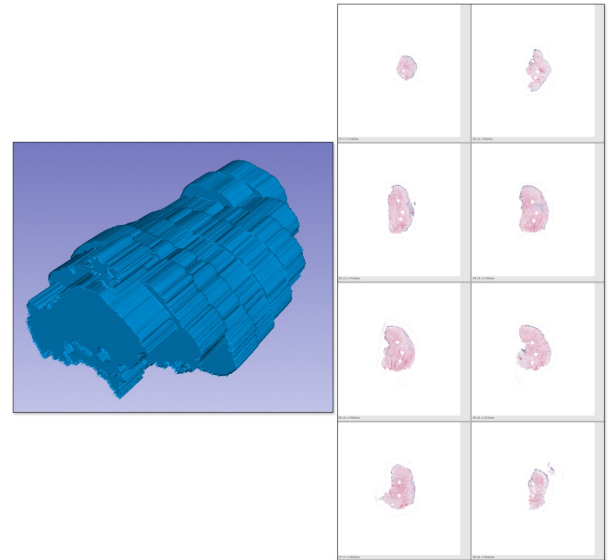


Figure 2.4: Assembled histopathology volume for an example specimen (patient 2), the same specimen as Figure 2.2, shown next to its registered slices disassembled. The volume is rendered in 3D Slicer. Each 4 mm slab corresponds to one physical histopathology slice, placed at its physical z-location, so the specimen mask forms a continuous stack of slabs at 4 mm z-spacing. The front slab of the 3D volume corresponds to the last image of the disassembled slices.

2.3. Histopathology-to-US registration

The reconstructed 3D histopathology volume was registered to the US coordinate system in three steps: an initial 3D alignment, a global 3D similarity refinement, and a per-slice 2D refinement.

2.3.1. Initial 3D alignment

An initial alignment was obtained by combining a manual rotation with automatically computed translation and scale. In 3D Slicer, the histopathology volume was interactively aligned onto the US volume until their specimen masks coincided, and only the rotational component of the resulting transform was retained. Translation was subsequently computed as the vector between the centres of mass of the two specimen masks, and an isotropic scale factor was included, computed from the ratio of the specimen mask volumes. The resulting transform is denoted $T_{3D,init}$.

2.3.2. 3D similarity refinement

The 3D similarity refinement recovers the residual position, orientation, and isotropic scale of the specimen envelope between the two modalities, starting from the initial alignment $T_{3D,init}$. Both binary specimen masks were resampled onto a shared voxel grid Ω and converted into SDFs. A similarity transform T was then optimised by minimising the MSE between the two SDFs:

$$T_{3D,refine} = \arg \min_{T \in \text{Sim}(3)} \frac{1}{|\Omega|} \sum_{\mathbf{x} \in \Omega} (\Phi_{US}(\mathbf{x}) - \Phi_H(T(\mathbf{x})))^2, \quad (2.2)$$

where Φ_{US} and Φ_H are the SDFs of the US and histopathology specimen masks, Ω is the shared voxel grid in the US coordinate system, and $\text{Sim}(3)$ denotes the 3D similarity transform group (rotation, translation, isotropic scale). The MSE was estimated by stochastic subsampling at each iteration, and minimised by regular-step gradient descent within a multi-resolution image pyramid; implementation parameters are given in Section 3.2.

2.3.3. Per-slice 2D refinement

Each histopathology slice was refined in 2D against a corresponding US slice to correct for per-slice residual misalignment and in-plane shrinkage variation. Because the histopathology slice's pose in US space usually does not coincide with a US acquisition plane, both the US and histopathology slices were resampled as thin single-voxel 2D planes along the plane defined by the histopathology slice's current 3D pose. The two

2D slices were then registered using their SDFs and a 2D similarity transform, following the same objective as the 3D refinement (Equation 2.2) but in 2D. The optimised transform for slice i is denoted $T_{2D,refine}(i)$. An example of this refinement step can be seen in Figure 2.6.

2.4. Composition into US space

The transforms produced by the reconstruction and registration stages are composed to map each histopathology slice into the US coordinate system. For a point $\mathbf{x}_i$ in the native 2D system of slice i , embedded in 3D as $(x, y, 0)^\top$, its position in US space is given by:

$$\mathbf{x}_{US} = T_{3D}(i) T_{recon}(i) \mathbf{x}_i, \quad (2.3)$$

where

$$T_{recon}(i) = T_z(i) \tilde{T}_{fid}(i) \tilde{T}_{init}(i)$$

composes the initial 2D registration, fiducial-aware joint refinement, and z-placement into the 3D histopathology volume (Section 2.2.4), and

$$T_{3D}(i) = \tilde{T}_{2D,refine}(i) T_{3D,refine}$$

maps the reconstructed histopathology volume into US space (Section 2.3). Here $T_{3D,refine}$ denotes the inverse of the optimised transform from Equation 2.2, which maps points from the US grid Ω into histopathology space and already incorporates the initial 3D alignment (Section 2.3.1). The per-slice refinement $\tilde{T}_{2D,refine}(i)$ is applied in the same inverse sense.

A tilde marks a 2D transform applied inside 3D: it moves points within one plane and leaves the perpendicular axis unchanged. For $\tilde{T}_{init}(i)$ and $\tilde{T}_{fid}(i)$, this plane is the z-axis/xy-plane of the histopathology reconstruction volume (Section 2.2.4). For $\tilde{T}_{2D,refine}(i)$, it is slice i 's own plane in US space, as established by $T_{3D,init}$ and $T_{3D,refine}$.

The final output volume was constructed by applying Equation 2.3 to each slice. Each slice was taken in its native 2D coordinates, transformed through the full pipeline, and mapped into the US voxel grid as a tilted slab, extruded as in Section 2.2.4 but now at its registered 3D pose, with no interpolation between neighbouring physical slices (Figure 2.5).

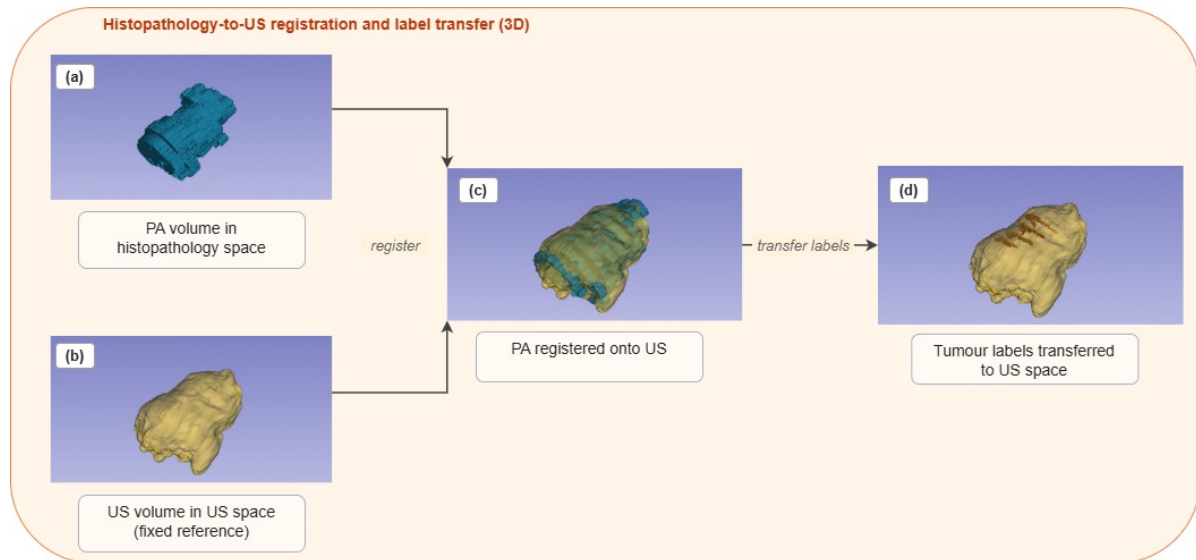


Figure 2.5: Final output volume for patient 3, showing the composed histopathology volume mapped into US space via Equation 2.3. Each slab is one histopathology slice placed at its registered 3D pose, 4 mm thick to match the physical section thickness.

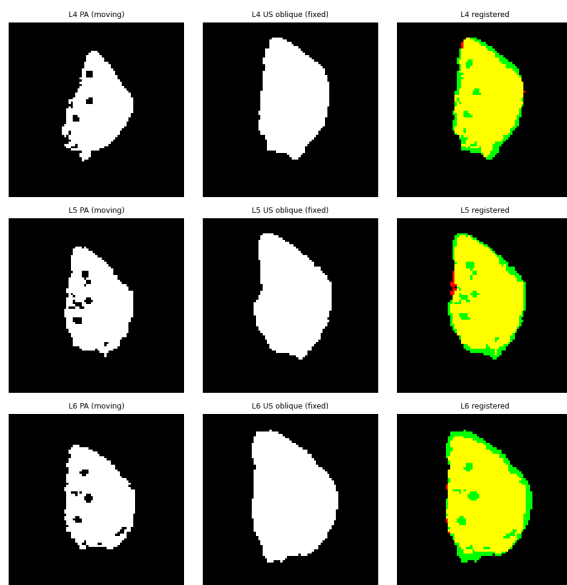


Figure 2.6: Per-slice 2D refinement for slices 4, 5 and 6 of patient 3. Left: histopathology specimen mask before 2D refinement. Middle: corresponding US specimen mask, resampled along the histopathology slice's plane, used as the fixed target. Right: registered histopathology mask (red) overlaid on the US mask (green).

Experiments and Results

3.1. Data

3.1.1. Cohort and demographics

Twelve patients with TSCC were included retrospectively from the Netherlands Cancer Institute (Antoni van Leeuwenhoek, Amsterdam, The Netherlands), with one resection specimen per patient. The dataset was split into two groups: five patients were used for the development and initial validation of the registration pipeline, while the remaining seven patients, from an earlier phase of the same project [11], were held out for final evaluation. Tissue acquisition and the reuse of clinical imaging data were approved by the institutional ethics committee under protocol number NL-78804.031.21 [51].

The development cohort (three male, two female) had a mean age of 60.6 years (27–83) with three T1 and two T2 tumours. The evaluation cohort (five male, two female) had a mean age of 64.0 years (36–86) with five T1, one T2 and one T3 tumour. Per-patient demographics are given in Table 3.1.

Three intravenous-tube fiducials were inserted through the specimen prior to US acquisition to serve as fiducial markers in all patients except two development patients; patient 4, which had fiducials inserted after the US scan, and 5, which never had fiducials inserted, both reasons remain unknown. Between 6 and 12 histopathology slices were obtained per specimen with a median of 9; per-slice area and fiducial hole count for every specimen are listed in Appendix C.1, which also reports the per-slice registration quality. Automated fiducial-hole detection recovered 196 of 205 true holes across the fiducial-bearing patients (miss rate 4.4%), with 1 false-positive (0.5%). Three misses occurred in the development cohort and 6 in the evaluation cohort.

3.1.2. Ultrasound acquisition

For each specimen, one ex-vivo 3D US volume was acquired directly after resection using a motorised acquisition workflow developed at the Netherlands Cancer Institute [51]. Each specimen was placed in a saline-filled container. A BK5000 US system (I14C51, BK Medical, Herlev, Denmark) was used with a 18 MHz transducer. The transducer was mounted on a mo-

torised linear rail (adapted from Master 2s, NEJE, China) and translated step-wise across the specimen, producing a 3D volume with a nominal through-plane slice spacing of 0.5 mm. Specimen masks of the reconstructed US volumes were obtained automatically using a previously trained 3D nnU-Net segmentation model [3], which reached a Dice of 0.96 and a 95th-percentile boundary error of 1.2 mm on held-out data. The acquisition set-up is shown in Figure 4.

3.1.3. Histopathology acquisition

Following US acquisition, each specimen was cut into 4 mm sections along the anterior–posterior axis. From each section, a 4 μ m slice was cut, H&E-stained, and digitised as a whole-slide image in MRXS format using a 3DHistech Panoramic 1000 slide scanner (3DHistech Ltd., Budapest, Hungary) at a native resolution of 0.24 μ m/pixel.

3.1.4. Annotations

Two sets of manual annotations were used in this study, both performed by the first author. In each specimen containing fiducial markers, the centre-line of each fiducial hole was traced in 3D Slicer as an open curve in the US volume, with control points placed at approximately 0.5 mm spacing; the corresponding histopathology centre-line was obtained by manually placing one control point per slice at the visible fiducial hole in the reconstructed histopathology volume, yielding control points spaced at the 4 mm inter-slice interval. Additionally, 10 anatomically distinct point landmarks were placed per specimen in both modalities along the outer border of the specimen. Landmarks were placed at features whose identity could be independently re-confirmed in both modalities. In three evaluation-cohort specimens, individual slices contained two anatomically separate tissue fragments on a single glass slide. Anatomical landmarks were not placed on these slices, as no reliable point-to-point correspondence with the US volume could be established. These slices were retained in the assembly and registration stages, so they still contributed to the 3D alignment of neighbouring slices; the reported landmark TRE therefore excludes points on these slices

Table 3.1: Per-patient demographics and data characteristics. Fiducials histo/US indicates whether traced needle markers were available in the histopathology and US volumes, respectively.

Patient	Age (y)	Sex	T-stage	Sections (n)	Fiducials histo/US
<i>Development cohort</i>					
1	27	M	T1	10	yes/yes
2	73	M	T1	8	yes/yes
3	58	M	T2	9	yes/yes
4	62	F	T2	7	yes/no
5	83	F	T1	12	no/no
Mean	60.6 (27–83)			9.2 (7–12)	
<i>Evaluation cohort</i>					
6	74	F	T1	11	yes/yes
7	44	M	T2	9	yes/yes
8	70	F	T1	11	yes/yes
9	74	M	T1	10	yes/yes
10	86	M	T1	6	yes/yes
11	36	M	T1	9	yes/yes
12	64	M	T3	6	yes/yes
Mean	64.0 (36–86)			8.9 (6–11)	
Overall mean	62.6			9.0	

but reflects registration behaviour that includes them.

In addition to these validation annotations, the tumour was delineated on each histopathology whole-slide image by the first author and reviewed by a pathologist, who indicated corrections where needed. These tumour annotations were not used for validation; they were transferred into US space by the pipeline to produce the first set of registered tumour labels, presented as qualitative and preliminary margin-transfer results (Section 3.6).

3.2. Implementation

The pipeline was implemented in Python 3.14.4. Whole-slide images were read with OpenSlide and tifffile; image processing used NumPy, SciPy, scikit-image, and Pillow. Optimisation of the fiducial-aware joint refinement (Section 2.2.3) was performed with L-BFGS-B in SciPy [57]. The 3D and per-slice 2D registrations (Sections 2.3.2 and 2.3.3) were performed with SimpleITK [58]. Reconstructed volumes and registrations were inspected in 3D Slicer (version 5.10.0) [59].

For the configuration parameters, the histopathology slices were downsampled to 10 $\mu\text{m}/\text{pixel}$. Background void pixels were turned to white using a threshold of $\text{RGB} \leq 50$, and the binary tissue masks were generated with a threshold of $\text{RGB} \geq 230$. Both values were chosen empirically by checking the mask quality across the cohort. Internal regions were kept as fiducial holes if their area was between 10,000 and 500,000 pixel^2

and their circularity ($4\pi \cdot \text{Area}/\text{Perimeter}^2$) was above 0.6. The outer tissue contour was sampled at 200 evenly spaced points, sufficient to represent the slice boundary at a level of detail adequate for visual inspection.

The regulariser weight for the joint refinement was set to $\mu = 10^{-2}$, chosen to be small enough that the fiducial term dominates whenever two or three holes are present, while still constraining slices with only one detected hole. The pipeline’s stability across a range of μ values is validated empirically in Experiment 2 (Section 3.4.2).

The 3D and 2D similarity refinements resampled their SDFs onto a 0.5 mm isotropic grid. The multi-resolution scheme used three levels (shrink factors $\{4, 2, 1\}$, Gaussian smoothing sigmas $\{2, 1, 0\}$, coarse to fine); the MSE metric was estimated on a random 30% voxel subsample per iteration, with a fixed seed for reproducibility. The regular-step gradient descent optimiser started with a step length of 1.0, halved whenever the gradient direction reversed (indicating overshoot), and stopped once the step fell below 10^{-4} or 500 iterations were reached. For the final composition into US space (Section 2.4), each slice was rendered as a 4 mm-thick tilted slab.

Across the twelve patients, the automatic chirality-based flip decision was overridden manually in 38 of 108 slices (35.2%), and a 90° rotation was applied in 55 of 108 (50.9%). The majority orientation was inverted relative to the automatic canonical choice in

three patients (4, 5 and 11); in the remaining nine, the automatic canonical orientation matched the majority of the manual decisions. Per-patient counts are given in Appendix B.

The full pipeline runs in approximately 5 minutes of automated compute per specimen on an Intel Core i5-6500 workstation with 16 GB RAM. The manual review and initialisation steps (Section 2.3.1) add roughly 10–20 minutes per specimen. The full procedure for adding and running a new patient through the pipeline is documented in Appendix H.

3.3. Evaluation metrics

Registration accuracy was quantified at three levels.

3.3.1. Fiducial-based accuracy

The line-based target registration error (mTRE) was computed on the traced fiducial centre-lines. For each histopathology control point, the perpendicular distance to the warped US centre-line was reported; the axial component was excluded. The mean over all control points per specimen was reported as the mTRE, with the (sample) standard deviation (SD) indicating the spread across control points.

3.3.2. Anatomical landmark accuracy

For each landmark pair, the Euclidean distance in US space between the warped histopathology landmark and its US counterpart was computed. The mean over all landmarks per specimen was reported as the landmark TRE, with the SD indicating the spread across landmarks.

3.3.3. Specimen mask overlap and shrinkage

Whole-specimen mask agreement after the 3D refinement was quantified using the Dice and the NSD at 1 mm tolerance [39]. Both metrics were computed at the native resolution of the US specimen mask, comparing it directly against the composed histopathology-in-US volume described in Section 2.4. Both metrics are reported throughout.

Shrinkage was estimated from the scale factors solved for during registration: the whole-specimen scale from the 3D registration (isotropic) and the median per-slice scale from the 2D refinements (in-plane). Isotropic shrinkage was taken from the 3D registration before the per-slice 2D refinement, so that it reflects uniform tissue shrinkage rather than the local corrections applied in that later step.

3.4. Experiments

The following experiments examine the pipeline’s design choices and then validate it. Experiments 1–3 test the design prior to validation: whether both refinement stages contribute to accuracy, how sensitive the pipeline is to the regulariser weight, and whether the similarity transform is preferable to an affine model in the 3D refinement. Experiments 4–6 validate the full pipeline against manual initialisation and on the development and held-out evaluation cohorts. Across these experiments, whole-specimen overlap (Dice), surface- and point-based accuracy (NSD@1mm, landmark TRE), and fiducial-line error (mTRE) did not always move together, as discussed in Chapter 4.

3.4.1. Experiment 1: Ablation of pipeline components

An ablation study was conducted to evaluate the individual contributions of two refinement stages: the fiducial-aware joint refinement (step 4, Section 2.2.3) and the per-slice 2D refinement (step 8, Section 2.3.3).

Three variations were compared on the five development patients: the full pipeline, a variant without step 4, and a variant without step 8. All remaining settings matched Section 3.2. Accuracy was quantified with fiducial mTRE (patients 1–3), landmark TRE (all patients), and whole-specimen Dice and NSD@1mm.

The landmark TRE and fiducial mTRE results are shown in Figure 3.1; the full results are provided in Appendix A (Table A.1). Removing step 4 altered global Dice scores and NSD@1mm values by no more than 0.4 percentage points in four of five patients; the exception was patient 4, where Dice and NSD@1mm shifted by 3.9 and 3.6 percentage points respectively. It increased mTRE in all three fiducial-bearing patients (mean +0.61 mm, range 0.54–0.69 mm). It increased landmark TRE in four of the five patients and left it unchanged in patient 5 (mean +0.46 mm, range 0.00–1.01 mm across all five).

Removing step 8 increased landmark TRE in all five patients without exception (mean +1.91 mm, range 0.48–3.06 mm) and increased mTRE in all three fiducial-bearing patients (mean +0.95 mm, range 0.29–1.41 mm). Its effect on Dice was mixed: an increase in patients 1 and 4, and a decrease in patients 2, 3, and 5 (0.78–0.84 vs. 0.72–0.85 at baseline). NSD@1mm, in contrast, was lower in every patient (0.66–0.78 vs. 0.85–0.94 at baseline).

3.4.2. Experiment 2: Regulariser sensitivity analysis

The sensitivity of the pipeline to the joint-refinement regulariser weight μ (Section 2.2.3) was assessed, and the stable operating range identified.

The full pipeline was run on the three fiducial-bearing

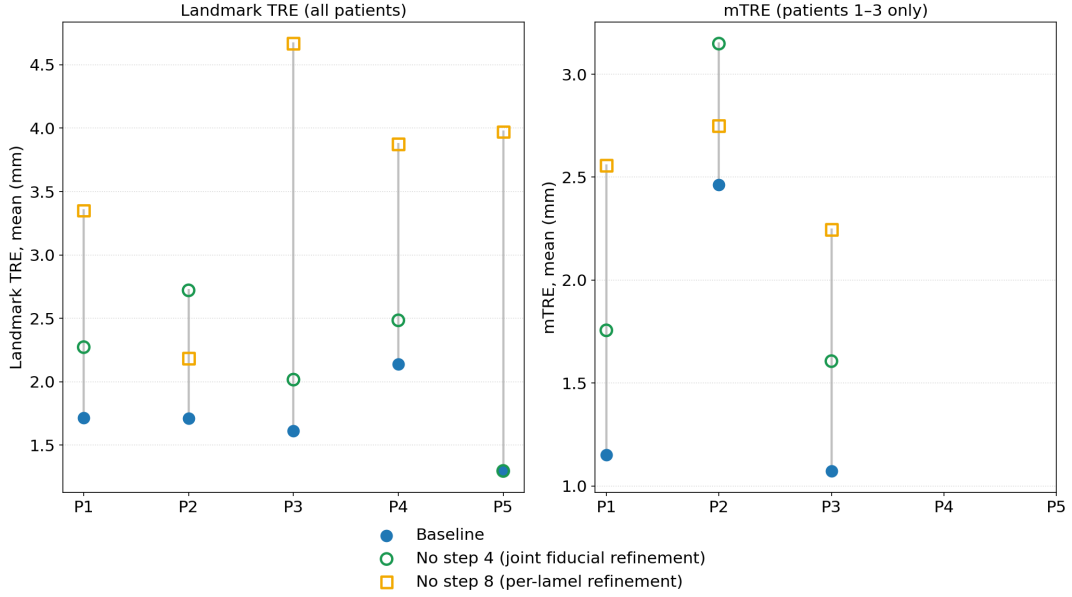


Figure 3.1: Landmark TRE (top, mean [mm]) and fiducial mTRE (bottom, [mm]) across the three ablation configurations.

development patients (patients 1–3) for five values of μ : $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^0\}$, with all other settings as in Section 3.2. For each value, the composed-transform mTRE along the traced fiducial centre-lines (Section 3.3.1) was computed.

The resulting fiducial mTREs are summarised in Figure 3.2; the full results are in Appendix A (Table A.2). The pipeline remains stable across $\mu \in [10^{-2}, 1]$, with individual mTRE values shifting by no more than 0.7 mm. At $\mu = 10^{-3}$, two of the three patients still fall within this stable range, while the third already shows clear degradation; registration performance degrades sharply for all patients only at $\mu \leq 10^{-4}$.

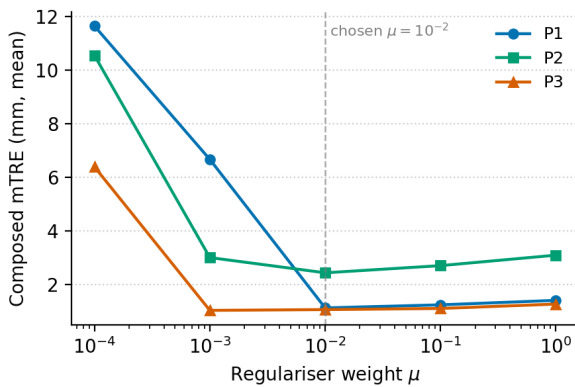


Figure 3.2: Fiducial mTRE (mean [mm]) as a function of the joint-refinement regulariser weight μ (log scale), for the three development-cohort patients with traced fiducials. The dashed vertical line marks the chosen operating value, $\mu = 10^{-2}$.

Figure 3.3 illustrates this operating point directly for one patient (Patient 3): joint refinement at $\mu = 10^{-2}$

visibly tightens each fiducial track around its fitted 3D line relative to the pre-refinement pose.

3.4.3. Experiment 3: 3D refinement transform model

The 3D refinement step (Section 2.3.2) uses a similarity transform by default. This experiment tested whether relaxing that constraint to a full affine transform (independent per-axis scale and shear) improves registration accuracy.

The 3D refinement was run in similarity and affine mode for all five development patients, with all other settings as in Section 3.2. Accuracy was assessed with fiducial mTRE (patients 1–3), landmark TRE (all patients), and whole-specimen Dice and NSD@1mm.

The landmark TRE results are summarised in Figure 3.4; the full results can be found in Appendix A (Table A.3).

For patients 1–4, NSD@1mm changed by at most 0.05, while landmark TRE was consistently higher under the affine model (mean increases of 0.43–1.98 mm). For patient 5, NSD@1mm was lower for affine (0.84 affine vs. 0.94 similarity), and landmark TRE rose from 1.30 mm to 18.51 mm (max 31.01 mm). The similarity transform was retained as the default.

Figure 3.5 illustrates this failure mode for patient 5: corresponding pairs are visually closer together with similarity, while with affine the error vectors are larger and point in multiple directions.

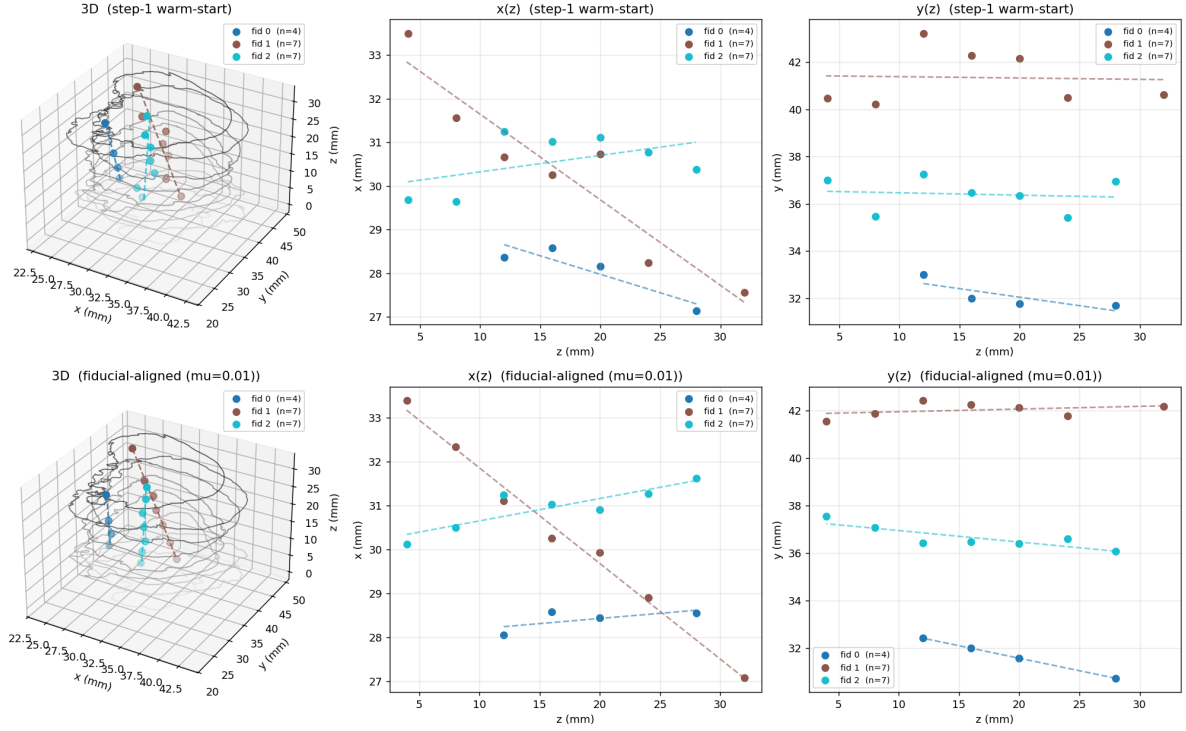


Figure 3.3: Effect of the joint fiducial refinement (step 4, $\mu = 10^{-2}$) on the 3D linearity of the fiducial tracks for patient 3. Three fiducials (fid 0/1/2), detected in 4, 7, and 7 of the imaged slices respectively. Top row: initial per-slice pose before fiducial refinement. Bottom row: after joint fiducial refinement. Left: 3D fiducial hole centroids per slice, coloured per fiducial, with the fitted 3D line per fiducial (dashed). Middle/right: the same fit projected onto the x - z and y - z planes.

3.4.4. Experiment 4: Comparison against manual initialisation

The manual 3D–3D initialisation alone (step 6) was compared against the full pipeline.

The manual initialisation transform (manual rotation with automatic translation and scale) was applied to each development patient without further refinement and compared against the full pipeline output. Accuracy was assessed with fiducial mTRE (patients 1–3), landmark TRE (all patients), Dice, and NSD@1mm.

Landmark TRE is shown in Figure 3.6; the full results are provided in Appendix A (Table A.4). Manual-only Dice was higher than the full pipeline for two of five patients (patients 1 and 4) and lower for the remaining three, no consistent direction across the cohort (0.75–0.81 manual vs. 0.72–0.85 full pipeline). NSD@1mm was lower for manual-only alignment in every patient (0.63–0.71 vs. 0.85–0.94), and landmark TRE was higher for all five patients, with the largest increase in patient 5, from 1.30 mm to 9.97 mm. Fiducial mTRE was likewise higher for manual-only alignment in the three fiducial-bearing patients (mean 1.29 mm, range 0.95–1.83 mm).

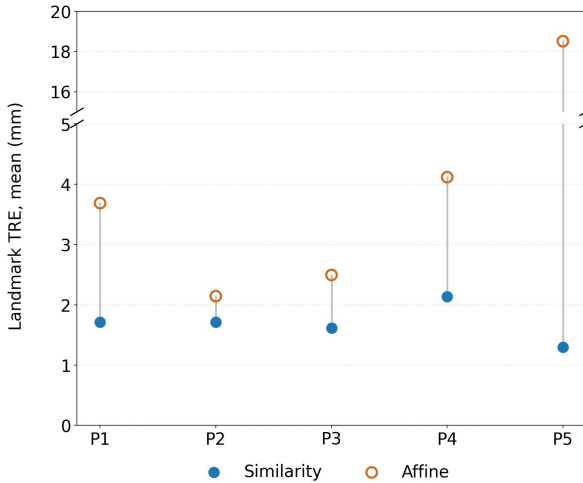


Figure 3.4: Landmark TRE (mean [mm]) for the similarity and affine transform models, all five development-cohort patients.

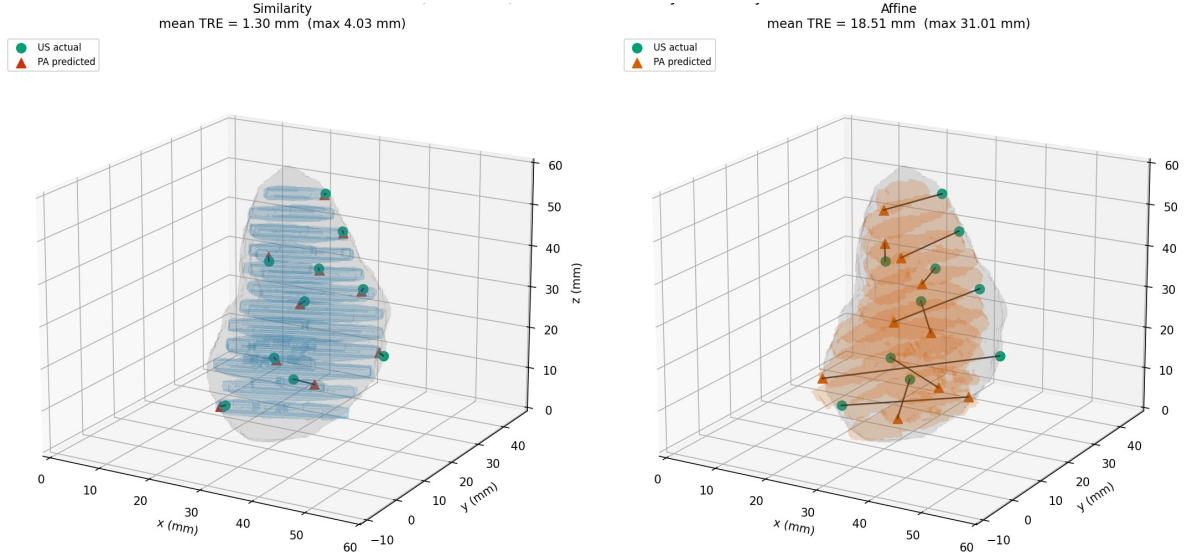


Figure 3.5: Patient 5: ground-truth (US, green) and predicted (histopathology-warped, triangle) landmark positions against the fixed US specimen shape (grey), for the similarity (left) and affine (right) transform models. Black lines connect corresponding landmark pairs.

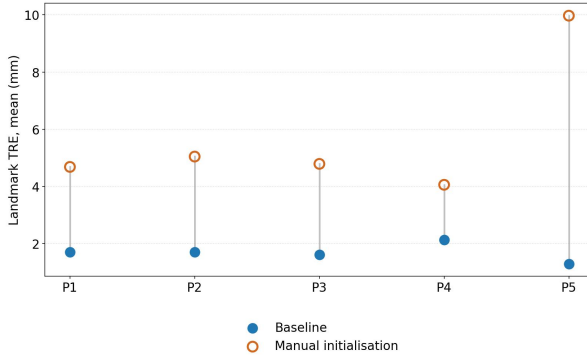


Figure 3.6: Landmark TRE (mean, mm) for manual initialisation alone and the baseline (full pipeline), all five development-cohort patients.

3.4.5. Experiment 5: Validation on the development cohort

The full pipeline was validated on all five development-cohort patients using the performance metrics detailed in Section 3.3.

The pipeline was executed using the baseline parameters from Section 3.2. For the three specimens containing fiducial markers (patients 1–3), fiducial mTRE was calculated for the final composed transform. Landmark TRE and global mask overlap were calculated across all five patients.

Point-based accuracy and whole-specimen mask overlap are given in Table 3.3. The cohort mean landmark TRE was 1.69 mm and the highest patient mean was 2.14 mm (patient 4). The cohort-mean whole-

specimen Dice was 0.80 and the cohort-mean NSD@1mm was 0.89.

Landmark TRE was not uniform along the specimen: it was lowest in the mid-region and elevated toward both tips (Figure 3.7). Mean landmark TRE in the tip regions (normalised position < 0.2 or > 0.8) was 2.27 mm ($n = 19$), compared with 1.34 mm in the mid-region ($0.2\text{--}0.8$, $n = 31$); this difference was statistically significant (Mann–Whitney $U = 452$, $p = 0.0017$).

To examine why accuracy drops at the tips, three factors were tested across all available specimens. Tip slices carried fewer detected fiducial holes than mid-region slices (across the eleven fiducial-bearing specimens: mean 1.23 vs. 2.35; Mann–Whitney $U = 416.5$, $p < 0.0001$, $n_{\text{tip}} = 31$, $n_{\text{mid}} = 62$). Hole-to-fiducial labelling ambiguity, measured by the nearest-neighbour ratio from the Hungarian assignment step (Section 2.2.3), increased with walking distance from the anchor slice ($r_s = 0.51$, $p < 0.0001$, $n = 197$). Tip sections were also smaller than mid-region sections (all twelve specimens: mean 170.2 vs. 276.1 mm²; Mann–Whitney $U = 385.5$, $p < 0.0001$).

A post-hoc analysis of the signed displacement vectors (Appendix D) shows an inward radial bias of ~ 0.44 mm across the pooled 50 landmarks (Wilcoxon $p < 0.0001$; sign test $p = 0.0003$). A smaller bias was also found along the posterior axis (mean -0.30 mm, $p = 0.046$); no bias was found along the other two Cartesian axes.

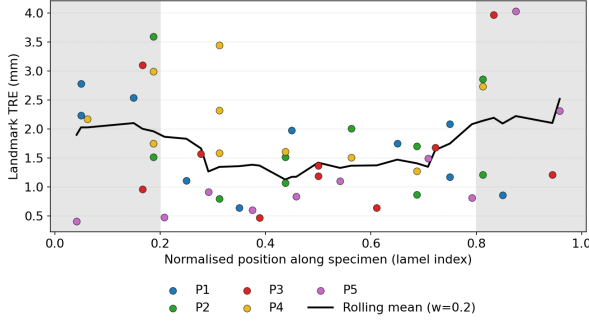


Figure 3.7: Landmark TRE against normalised position along the specimen (slice index, rescaled per patient by total section count), across the five development-cohort specimens. Each point is one landmark, coloured by patient; the black line is a rolling mean (window 0.2). Shaded regions mark the specimen tips (position < 0.2 and > 0.8).

3.4.6. Experiment 6: Validation on the evaluation cohort

To assess generalisation beyond the development cohort, the full pipeline was further validated on seven evaluation-cohort patients, using the same performance metrics detailed in Section 3.3.

The complete pipeline was executed using the baseline parameters from Section 3.2, unchanged from the development cohort. Fiducial mTRE was calculated for the final composed transform in all seven specimens; four contained three traced fiducial markers and three (patients 9, 10, and 12) contained two. Landmark TRE and global mask overlap were also calculated across all seven patients.

Table 3.4 summarises the point-based accuracy results and the whole-specimen mask overlap results. The cohort mean landmark TRE was 2.23 mm and the cohort mean fiducial mTRE was 2.13 mm, both comparable to, though somewhat higher than, the development cohort (1.69 mm and 1.56 mm respectively). Mean whole-specimen Dice (0.74) and mean NSD@1mm (0.87) were slightly lower than on the development cohort (0.80 and 0.89 respectively). Per-slice 2D refinement QC overlays for all twelve patients are shown in Appendix E.

3.5. Shrinkage

Table 3.2 shows the isotropic and in-plane shrinkage estimates for both cohorts (see Section 3.3.3 for metric definitions). Average isotropic shrinkage was comparable between the two cohorts, at 10.1% and 10.9% respectively. The average in-plane tissue shrinkage was 6.3% for the development cohort (ranging from 3.7% to 9.2%) and 8.0% for the evaluation cohort (ranging from 3.7% to 12.0%), indicating a broadly similar but somewhat more variable degree of in-plane shrinkage

in the unseen specimens.

3.6. Qualitative results

To illustrate the pipeline’s intended downstream use, the composed transform (Equation 2.3) was used to transfer the histopathology tumour annotation into US space for each development-cohort specimen. Figures 3.8 and 3.9 show transferred tumour in US space next to the original histopathology annotation for four example slices; Figure 3.10 shows the transferred tumour rendered within the full 3D US volume for all five specimens. Full 2D results of all five development cohort specimens are shown in Appendix F.

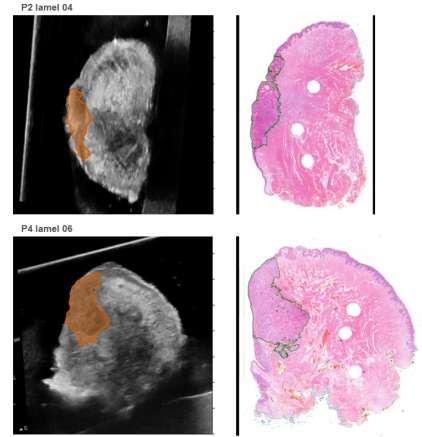


Figure 3.8: Two successful examples, one per row: transferred tumour (orange) on the US volume (left) and the original tumour annotation (green outline) on the histopathology slide (right). Top: patient 2 (slice 4); bottom: patient 4 (slice 6).

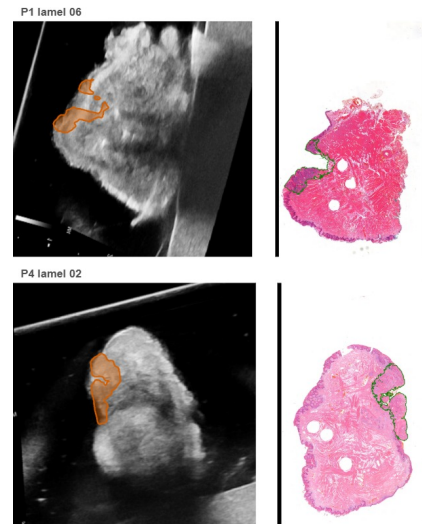


Figure 3.9: Two less successful examples, laid out as in Figure 3.8. Top: patient 1 (slice 6); bottom: patient 4 (slice 3).

Table 3.2: In-plane and isotropic shrinkage estimates for the development and evaluation cohorts (%).

Patient	Isotropic (%)	In-plane (%)
<i>Development cohort</i>		
1	21.4	9.2
2	4.0	3.7
3	4.3	7.2
4	11.7	6.3
5	9.0	5.2
Mean	10.1	6.3
<i>Evaluation cohort</i>		
6	4.5	3.7
7	13.8	12.0
8	7.2	5.2
9	12.5	6.9
10	9.9	9.4
11	12.9	8.0
12	15.8	10.4
Mean	10.9	8.0

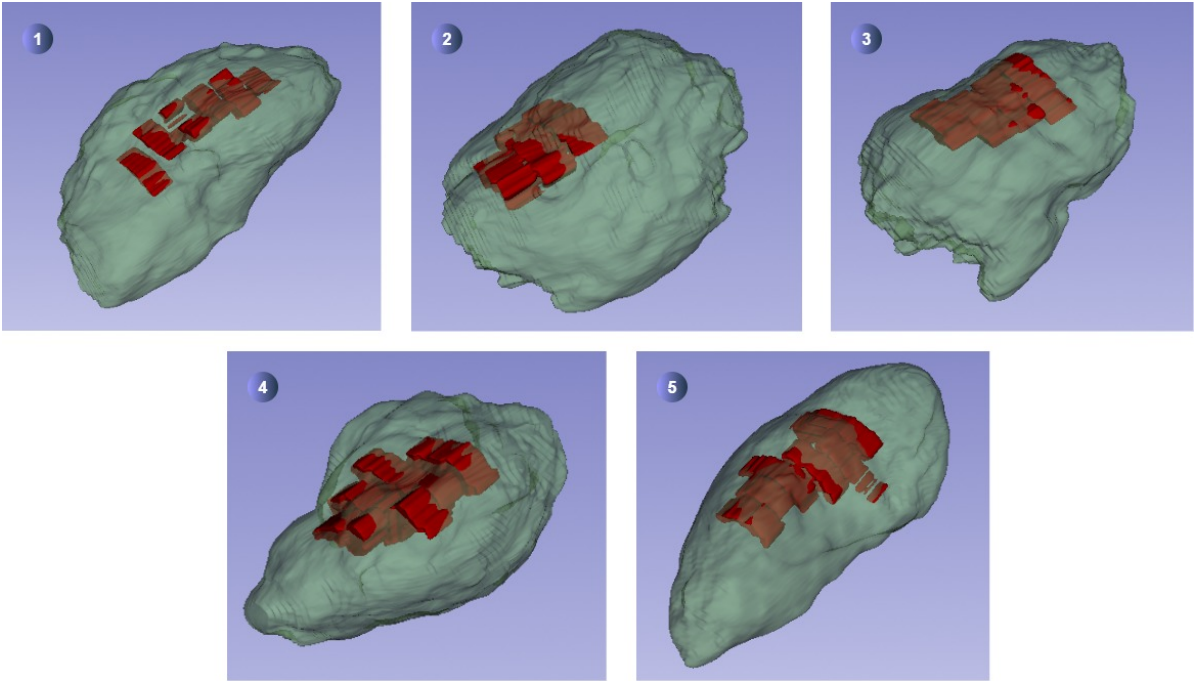


Figure 3.10: Transferred tumour region (red) rendered within the full registered US specimen volume (green), for all five development-cohort patients (1–5).

Table 3.3: Whole-specimen mask overlap and point-based accuracy after the full pipeline on the development cohort. NSD is reported at a 1 mm tolerance; mTRE and landmark TRE are mean (SD) in mm. mTRE is available for patients 1–3 only.

Patient	Dice	NSD@1mm	mTRE (mm)	Landmark TRE (mm)
1	0.72	0.89	1.15 (0.54)	1.71 (0.73)
2	0.84	0.88	2.46 (1.49)	1.71 (0.90)
3	0.85	0.90	1.07 (0.70)	1.61 (1.10)
4	0.75	0.85	–	2.14 (0.72)
5	0.84	0.94	–	1.30 (1.11)
Mean	0.80	0.89	1.56	1.69

Table 3.4: Whole-specimen mask overlap and point-based accuracy after the full pipeline on the evaluation cohort. NSD is reported at a 1 mm tolerance; mTRE and landmark TRE are mean (SD) in mm. mTRE is available for all patients.

Patient	Dice	NSD@1mm	mTRE (mm)	Landmark TRE (mm)
6	0.70	0.82	4.09 (2.65)	2.94 (1.08)
7	0.73	0.88	2.00 (0.99)	3.46 (1.21)
8	0.72	0.85	2.69 (1.86)	1.83 (0.60)
9	0.77	0.91	0.88 (0.57)	1.97 (0.90)
10	0.74	0.84	2.60 (1.53)	1.72 (0.80)
11	0.77	0.90	1.06 (0.51)	2.00 (0.70)
12	0.76	0.88	1.57 (1.12)	1.72 (0.94)
Mean	0.74	0.87	2.13	2.23

Discussion

4.1. Principal findings

This work presented a pipeline for registering 2D H&E histopathology slices to the 3D US coordinate frame of ex-vivo TSCC resection specimens, transferring the histopathology tumour annotation into the US volume. A 3D volume is reconstructed from the 2D histopathology slices and registered to the 3D US volume, which serves as the fixed reference. To our knowledge, this is the first 3D-to-3D registration pipeline for oral-cavity specimens that recovers slice correspondence semi-automatically rather than by manual pairing.

On the five-patient development cohort, the composed transform reached a mean fiducial TRE of 1.56 mm, a mean landmark TRE of 1.69 mm, and a mean whole-specimen NSD@1mm of 0.89. On the seven-patient held-out cohort, both point accuracy and surface overlap were slightly lower but in a similar range (mTRE 2.13 mm, mean landmark TRE 2.23 mm, NSD@1mm 0.87) with no parameter re-tuning. Thus, the pipeline generalises beyond the development set. Isotropic shrinkage was consistent between cohorts, 10.1% for the development cohort and 10.9% for the evaluation cohort, and mean in-plane shrinkage was 6.3% and 8.0% respectively.

Four findings emerged; first, each refinement stage improved point accuracy consistently across the cohort: removing the fiducial refinement raised fiducial mTRE in all three fiducial-bearing patients (+0.54–0.69 mm), and removing the per-slice 2D refinement raised landmark TRE in all five patients (+0.48–3.06 mm) and mTRE in the three fiducial-bearing patients (+0.29–1.41 mm). Both shifts are large relative to the 1 mm surgical margin tolerance the downstream model targets. Second, relaxing the 3D refinement from similarity to affine raised landmark TRE in every development patient (+0.43 to +17.2 mm) without a compensating gain in NSD@1mm; the similarity model was retained. Third, whole-specimen Dice did not move consistently across patients in any of the four ablation or comparison experiments, whereas NSD@1mm dropped in every patient when a whole refinement stage was removed (mean -0.17 for the per-slice 2D refinement, -0.23 for manual initialisation only) and was insensitive to removing the fiducial refinement and the affine

transform-class change. Fourth, registration accuracy was not uniform along the specimen: pooled landmark TRE was higher in the outer 20% at each end than in the middle 60% (2.27 vs. 1.34 mm, Mann–Whitney $p = 0.0017$; Figure 3.7).

4.2. Interpretation of results

The registration has only three sources of information: the specimen contour, the fiducial-needle geometry, and the operator’s discrete decisions. The experiments in Chapter 3 test these sources, though not fully separately.

Of the two ablated refinement stages, removing the per-slice 2D refinement hurt landmark TRE more than removing the fiducial refinement (mean +1.9 vs. +0.5 mm), so the per-slice step carries more of the point-accuracy improvement. It recovers a local scale for each section rather than one global value, and these local scales vary substantially within a single specimen, which a single global fit could not capture (Appendix C.1). This is consistent with sections shrinking by different amounts during fixation. The per-slice rotations in this step can also absorb residual errors from the histopathology reconstruction stage.

Relaxing the 3D refinement from similarity to affine degraded landmark TRE without a compensating gain in surface agreement. For the specimen with the largest error, the affine fit converged to a rotation roughly 138° different from the similarity fit, about the specimen’s own long axis. A twist around that axis barely changes the outer silhouette but displaces every internal point, so the mask objective does not penalise it. Across the cohort, this extra transform freedom on a mask input therefore does not improve accuracy: the objective cannot constrain the interior, however much freedom the transform has. Progress needs interior constraint, not more transform freedom, realistically a learned deformable registration that recovers structure from the image content inside the specimen rather than from the boundary alone. That requires interior signal the current mask objective discards, and a larger dataset, since such a model has far more degrees of freedom than the twelve specimens here can constrain

without overfitting.

Registration accuracy was not uniform along the specimen. Tip sections consistently performed worse (Section 3.4.5). Three factors, all concentrated at the tips, plausibly contribute: fewer detected fiducial holes, smaller section area, and correspondence ambiguity that grows with distance from the anchor slice (Section 2.2.3, Appendix C.1). Clinically the mid-region is the most critical, as tumour usually sits there rather than at the tips. Tip accuracy still matters, though: a whole-volume registration is worth more than a partial one, and the tips supply the tumour-free slices a downstream model needs as negative examples, so poor tip registration risks noisy negative labels. Downstream training could weight voxels by expected registration reliability (for example by section size and distance from the anchor) rather than trusting all labels equally.

The fiducial term contributes only where holes exist, and its strength depends on how many. Within the fiducial term, two holes are the minimum needed to identify the in-plane pose (rotation and 2D translation); a third adds no new constraint but improves robustness. A third hole could in principle also resolve the slice's chirality, since two holes leave a flip ambiguous; but chirality is fixed earlier in the orientation review, so the fiducial refinement need only solve a rigid rotation and translation, not a reflection. One hole fixes only the translation, leaving rotation for the regulariser; with zero holes the slice is held entirely at its prior pose (Equation 2.1). This is why performance is stable from $\mu = 10^{-2}$ to $\mu = 1$ but degrades sharply below it: the regulariser is what protects a weakly identified section from being pulled toward a wrongly assigned fiducial. Requiring at least two detected holes, or weighting each section by hole-to-fiducial assignment confidence, would handle this per-section rather than through a single global regulariser weight.

The manual-initialisation comparison (Section 3.4.4) shows that a coarse manual initial alignment alone does not reach the same accuracy as the full pipeline. Whole-specimen Dice was comparable, but point accuracy was worse in every patient and NSD@1mm collapsed. This does not make the manual step redundant. The automated refinement is a local optimiser that only corrects small residual misalignment from wherever it starts, and a mask-based objective can favour a wrong but similarly-shaped pose, for example a rotated or mirrored specimen, over the correct one. The manual step exists to start the optimiser close to the correct pose, so that refinement corrects the residual rather than converging to a wrong but similarly-shaped one; a grossly wrong initial pose, such as a mirror or flip, would not be recovered. The other manual decisions have no refinement stage acting on them at all, so they must be correct exactly. Chirality flips and

90° rotations were overridden in a substantial share of slices (Section 3.2), and a mirrored slice or one placed at the wrong z cannot be corrected by any later step. These decisions have to be correct at input; the manual burden is small for now. Keeping them manual also acts as a quality-control checkpoint, but could be automated if the pipeline is scaled to larger cohorts.

Whole-specimen Dice and NSD@1mm capture different aspects of the alignment. NSD@1mm scores the outer boundary against a fixed 1 mm tolerance, independent of specimen size. On the other hand, Dice measures interior overlap and is volume-dependent: the same absolute error costs more Dice on a smaller structure, contributing to lower Dice at the tips regardless of alignment. Dice is also known to be less sensitive to boundary misalignment than surface-based metrics generally: once two masks broadly overlap, Dice changes little even as the boundary shifts by several millimetres [39], which is why NSD@1mm was introduced. Its 1 mm tolerance maps directly onto the surgical margin threshold, making boundary agreement at that distance the clinically meaningful quantity.

The ablations reproduced this difference in sensitivity. NSD@1mm dropped substantially and consistently wherever registration clearly failed (omitting the per-slice 2D refinement and manual initialisation alone) moving with landmark TRE in every patient. Dice changed only marginally across the same configurations. Dice is therefore a coarse check that the specimen sits roughly in the right region; NSD@1mm is the more sensitive check on whether the boundary sits within the clinical tolerance. Had conclusions been drawn from Dice alone, none of the four experiments would have shown a meaningful effect, and three of the four would have pointed the wrong way. This discussion therefore rests primarily on fiducial mTRE, landmark TRE, and NSD@1mm; Dice is reported for continuity with prior work.

4.3. Comparison to prior work

This work's result cannot be compared directly against an in-domain state of the art: no published method registers a reconstructed 3D histopathology volume to a 3D US volume for oral-cavity specimens with the slice pairing solved automatically (Table 3, Section 0.4.3). The reported accuracies fall into two groups.

The first sets the slice correspondence by hand or by constraining acquisition so the imaging plane matches the cutting plane, then registers only the resulting in-plane alignment in 2D (Bekedam et al. [28], Rusu et al. [41], Pensa et al. [42], and Ghezzi et al. [43]). Because the correspondence is fixed beforehand rather than solved, their TRE reflects only that in-plane alignment,

not the pairing itself, and it spans a wide range (0.42–3.9 mm) that does not track problem difficulty. The two values that fall below the TREs reported here come from manual pairing on a small cohort (Bekedam et al. [28], 0.42 mm) and a five-step PET/MRI bridging protocol on three patients (Ghezzi et al. [43], 1.59 mm). The pipeline instead recovers the pairing as part of the 3D transform and absorbs any pairing error into the reported TRE, so these numbers are not comparable.

The second group registers to in-vivo US and corrects probe-induced deformation with a deformable step (Schalk et al. [30] 2.1 mm, Jager et al. [40] 3.5 mm), evaluated against the same type of landmark or marker reference used here. The landmark TREs (1.69 mm dev, 2.23 mm eval) fall in the same range using similarity transforms alone, because the ex-vivo saline-bath acquisition removes the probe–tissue contact those deformable steps exist to correct. This is the most meaningful comparison, since these methods report a landmark TRE against a comparable type of ground truth.

Fixed pairing was avoided because manual or acquisition-time pairing does not scale to the cohort sizes segmentation training needs.

4.4. Strengths and limitations

The main strength of the method concerns the correspondence. Slice-to-volume correspondence between the histopathology sections and the US volume is recovered semi-automatically, rather than being fixed manually per specimen. Manual pairing scales linearly with cohort size, whereas the pipeline can be reused as the institutional dataset grows. Bekedam [3] identified this as a bottleneck in the same clinical setting and proposed automated registration as the way to close that gap. Automating the correspondence is a step towards the data needs of a downstream segmentation model, which typically plateaus in the range of several hundred training cases for nnU-Net on 3D medical volumes [60]. On the registration itself, the pipeline is driven by specimen masks converted to SDFs instead of greyscale intensities, which avoids having to relate two modalities whose intensities encode fundamentally different physical quantities rather than a shared signal that merely differs in contrast.

The evaluation used three metrics with different dependencies. Fiducial mTRE is anchored on the inserted needles, is independent of the specimen mask, and gives control points inside the specimen rather than only on the outer border. Landmark TRE is also independent of the mask and, unlike mTRE, applies to every patient. NSD@1mm is scored against the mask but is calibrated to the 1 mm surgical margin threshold. The pipeline generalised to a seven-patient held-out co-

hort without parameter re-tuning, and its components were tested through ablation, sensitivity analysis, and a manual-initialisation comparison (Sections 3.4.1–3.4.4).

The main limitation of the method itself was the transform class. The similarity transform cannot model the non-linear deformation from shrinkage variation, tearing, or partial section loss, and the affine relaxation showed that adding linear freedom does not help. A deformable transform is not a drop-in fix either, since the SDF input constrains only the outer boundary; addressing this requires a different input to the registration, discussed in Section 4.6.

The principal study limitation was the scale of the data, in both the number of specimens and the number of sampled slices per specimen. The cohorts were small, with five development and seven evaluation specimens, and only one histopathology slice was retained per 4 mm section. Denser sampling is constrained by what the pathology workflow can process, though the institutional protocol has since moved to 2 mm sections. This small number of specimens limits the accuracy of the error estimates and prevents subgroup analysis by, for example, tumour stage or section count. It also means the dataset is too small to train a learned registration model, the direction a deformable approach would require. The held-out cohort indicates that the parameters were not overfitted to the development data, but does not remove the limitation of the sample size. Both cohorts were also acquired at the same institution with the same US set-up, so generalisation was tested across patients, not across institutions.

The mask and the histopathology reconstruction were further study limitations. Both the registration refinements and the overlap metrics used the same automatically generated nnU-Net US mask. The refinements align to the mask, not to the true tissue boundary, and Dice and NSD@1mm are also scored against that mask. Registration error and mask error are therefore mixed together, and a perfect registration cannot score higher than the mask allows. The mask’s own held-out accuracy is reported in Section 3.1.2 (Dice of 0.96 and a 95th-percentile boundary error of 1.2 mm). This affects the metrics unequally. mTRE and landmark TRE are annotated directly on the US image, independently of the registration mask; those placements carry their own uncertainty, but it is uncorrelated with the fit. Dice and NSD@1mm, in contrast, are scored against the mask that also drives the registration, so mask error contaminates them most, because its error is shared with what is being evaluated. Retraining the nnU-Net on more data would tighten both the registration and its evaluation.

A second source of error the registration cannot recover concerns the histopathology reconstruction itself. Only one 4 μm slice was kept per 4 mm section, so each slice represented tissue it did not sample. These errors enter the moving volume and cannot be recovered by any later step. Halving the sectioning interval to 2 mm, as the institutional protocol now does, reduces this. A separate preprocessing assumption is that each slide carries one connected tissue mass; in three evaluation-cohort specimens (patients 6, 7 and 8), individual slices contained anatomically separate fragments, which were placed at a single z -position. This still contributed to the 3D alignment of neighbouring sections.

Finally, both the fiducial centre-lines and the anatomical landmarks were annotated by a single observer (Section 3.1.4), so the observer's localisation error is embedded in the reported mTRE and landmark TRE and cannot be separated from the registration error. With a single observer this error could not be measured directly, since there is no second set of annotations to compare against. Bekedam et al. [28] estimated the same source of error at approximately 0.5 mm in the same modality, and cross-modality landmark studies in a comparable setting report per-point localisation errors below 0.5 mm [61]. The annotation contribution is therefore likely in that order, small relative to the reported TRE values of 1.5–2.5 mm but not negligible. An inter-rater comparison on a subset of specimens would quantify it directly. The anatomical landmarks were also placed only on the outer border, since no internal anatomical structures could be matched with certainty across both modalities so accuracy in the deeper tissue where the tumour margins lie is not independently verified, except along the fiducial tracks.

4.5. Clinical value

The pipeline does not measure margins directly. It produces the labelled training data a downstream US segmentation model would need, and margins can then be measured from the model's output. Its clinical value is therefore one step removed: what matters is not the 1 mm margin tolerance itself, but how much label error a segmentation model can absorb and still detect the margin.

At a mean landmark TRE of 1.69–2.23 mm across the two cohorts, the labels carry roughly 2 mm of error against a 1 mm clinical requirement. A model trained on these labels would have to be more accurate than its own training data to detect a 1 mm margin. How much error the task can absorb depends not only on its magnitude but on the error distribution and the amount of training data. Random error that scatters around the true boundary averages out across many

examples, so a model can still recover the correct margin from enough noisy labels; a systematic shift does not average out and propagates directly into the predictions. The first step is therefore to check whether the residual error is systematic or random. An analysis on the development cohort (Appendix D) shows an inward radial bias of ~ 0.44 mm, significant under both the Wilcoxon signed-rank and sign tests. Part of this bias is consistent with the isotropic scale slightly under-recovering histopathology shrinkage. However, a smaller directional component was also found along one Cartesian axis (posterior, $p = 0.046$), which a purely isotropic scale error would not produce; the cause of this axis-specific component is not yet understood and is left as an open question. If confirmed on a larger cohort, the bias should be reduced at the registration stage. The random component of the error, meanwhile, can still be absorbed by the downstream model through confidence-weighted label propagation, without requiring registration itself to improve.

The 5 mm clear-margin threshold is far less demanding: a ~ 2 mm error is small relative to it, so even at current accuracy the labels are informative about whether that boundary has been breached. The pipeline should be read as a proof of concept rather than a finished tool. It produces usable histopathology labels in US space now; reaching the label quality a 1 mm task needs will likely require registration improvements. The pipeline already establishes what any margin-prediction model in this setting depends on: (semi-)automatic 3D registration of histopathology labels into US space.

4.6. Future work

Two directions follow from this work, differing in whether they improve the registration itself or accept its current accuracy and compensate for it when building the training set.

The first direction keeps the current registration and instead compensates for its residual error when transferring labels into US space. This requires characterising that error more fully: the systematic bias should be characterised further, and since accuracy is highest near fiducial-bearing sections and drops toward the tips, treating every registered voxel as equally certain is not justified. The error can then be fed into the downstream model through a soft label-propagation scheme, in which confidence decays with distance from a sampled plane, letting the model weight training voxels by registration reliability rather than trusting all labels equally. Both characterising the error and training the downstream model on it require more data than the current development set, which the volume-reconstruction improvements below could help supply.

The second direction refines the registration. Replac-

ing the similarity model with a local deformable one would let it absorb non-linear tissue behaviour. In particular shrinkage that varies across tissue types, which a single global transform cannot capture. A deformable transform can be substituted into the current pipeline, but would not help: because the inputs are SDFs of the specimen mask, the objective constrains only the outer boundary, leaving the interior of the displacement field free to distort without penalty. Different tissue types are visible in US but hard for a human observer to distinguish reliably. A learned, unsupervised approach could instead find these interior structures directly from the imaging data. On this basis a regularised architecture such as DRMIME [62] could fit local displacement fields without manual guidance. Like the first direction, a learned registration needs substantially more specimens than the current cohort, so it becomes feasible only as the dataset grows.

Reducing reliance on the fiducials is a related question. The trained US model will already be fiducial-free by construction, since the fiducials exist only to build the training set. Within data collection for this training set, needle insertion is an invasive extra step in the specimen-handling workflow, so reducing reliance on it is desirable. The fiducials currently supply the only interior point correspondence the registration has, so removing them is not straightforward on the current pipeline. However, if a learned deformable method can recover interior structure directly from the imaging data as described above, the reliance on fiducials for interior anchoring would reduce. Finding a registration technique that reaches sufficient accuracy without the fiducials would therefore be a worthwhile direction, and would simplify the specimen-handling workflow.

Two reconstruction gaps remain, and closing them could additionally yield usable training data for the first direction as well as better registration in the second direction. First, the 4 mm sectioning leaves empty planes between sampled sections. The institutional protocol has since changed to 2 mm, halving the gap. Learned inter-section interpolation (for example a diffusion-based model synthesising intermediate appearance) could fill the remainder to give a continuous labelled volume rather than the current sparse stack. Second, where a section is torn, lost, or splits into anatomically separate fragments, no transform can recover correspondence. These cases need upstream detection and handling, such as targeted re-processing for sections with tears, or multi-fragment sections could be separated into distinct z -positions.

Finally, in terms of generalisability and scaling, extending the pipeline to other acquisition sites or clinical specialisms would first require testing on data that does not yet exist for this set-up. This is left as a longer-term direction rather than a near-term priority.

5

Conclusion

In this thesis, a semi-automated 3D-to-3D registration pipeline was developed and validated that aligns a reconstructed ex-vivo histopathology volume to the 3D specimen ultrasound for oral-cavity specimens. The pipeline reached a mean landmark TRE of 1.69 mm on the development cohort ($n = 5$) and 2.23 mm on the held-out cohort ($n = 7$) without parameter re-tuning, recovering the slice correspondence semi-automatically rather than by manual pairing. This roughly 2 mm error currently limits direct margin assessment at the 1 mm threshold. By replacing manual slice-by-slice pairing, the pipeline provides a scalable framework and an accuracy baseline for building the registered datasets a downstream margin-prediction model will need.

References

- [1] OpenStax. *Anatomy and Physiology*. Licensed under CC BY 3.0. OpenStax, Rice University, 2013. URL: <https://openstax.org/details/books/anatomy-and-physiology>.
- [2] Elisa Bellini et al. “Impact of surgical margins status on survival outcomes in oral cavity squamous cell carcinoma: a systematic review and meta-analysis”. In: *Acta Otorhinolaryngologica Italica* 45.2 Suppl 1 (2025), S2–S14. DOI: 10.14639/0392-100X-suppl.1-45-2025-N1044.
- [3] Nicolaas Maarten Bekedam. “Waves of precision: Intra-operative 3D ultrasound for margin assessment of resected tongue cancer”. ISBN: 978-94-6506-882-4. PhD thesis. Universiteit van Amsterdam, 2025.
- [4] Fatemeh Makouei et al. “3D Ultrasound and MRI in Assessing Resection Margins during Tongue Cancer Surgery: A Research Protocol for a Clinical Diagnostic Accuracy Study”. In: *Journal of Imaging* 9.9 (2023), p. 174. DOI: 10.3390/jimaging9090174.
- [5] Integraal Kankercentrum Nederland (IKNL). *Kanker van het hoofd-halsgebied in Nederland: Mondholtekanker*. Oct. 2025. URL: <https://iknl.nl/kanker-van-het-hoofd-halsgebied/mondholte>.
- [6] B. A. C. van Dijk et al. “Trends in oral cavity cancer incidence, mortality, survival and treatment in the Netherlands”. In: *International Journal of Cancer* 139.3 (2016), pp. 574–583. DOI: 10.1002/ijc.30107.
- [7] Yunhan Tan et al. “Oral squamous cell carcinomas: state of the field and emerging directions”. In: *International Journal of Oral Science* 15 (2023), p. 44. DOI: 10.1038/s41368-023-00249-w.
- [8] Ira Sanders and Liancai Mu. “A Three-Dimensional Atlas of Human Tongue Muscles”. In: *The Anatomical Record* 296.7 (2013), pp. 1102–1114. DOI: 10.1002/ar.22711.
- [9] Rob Noorlag. “Ultrasound-guided oral cancer surgery”. Utrecht, The Netherlands. PhD thesis. Universiteit Utrecht, 2022.
- [10] Saman Warnakulasuriya and John S. Greenspan. *Textbook of Oral Cancer: Prevention, Diagnosis and Management*. Cham, Switzerland: Springer Nature Switzerland AG, 2020. ISBN: 978-3-030-32315-8. DOI: 10.1007/978-3-030-32316-5.
- [11] Nynke J. Dijkhuis. “Towards improved intraoperative resection margin assessment in tongue carcinoma: A study on tissue shrinkage, tumor segmentation and ultrasound calibration”. Master thesis. Enschede, The Netherlands: University of Twente, 2025.
- [12] Dario Di Stasio et al. “High-definition ultrasound characterization of squamous carcinoma of the tongue: A descriptive observational study”. In: *Cancers* 14.3 (2022), p. 564. DOI: 10.3390/cancers14030564.
- [13] Klijs J. de Koning et al. “Application and accuracy of ultrasound-guided resections of tongue cancer”. In: *Oral Oncology* 133 (2022), p. 106023. DOI: 10.1016/j.oraloncology.2022.106023.
- [14] David G. Pfister et al. *Head and Neck Cancers, Version 4.2024, NCCN Clinical Practice Guidelines in Oncology*. 2024. URL: https://www.nccn.org/professionals/physician_gls/pdf/head-and-neck.pdf.
- [15] Nederlandse Vereniging voor Keel-Neus-Oorheelkunde (NVKNO). *Landelijke Richtlijn Hoofd-halstumoren (Adjuvante behandeling van het hoofd-halsgebied)*. <https://richtlijndatabase.nl/>. Accessed: 2026-07-29. 2023.
- [16] Keith Hunter et al. *Dataset for the histopathological reporting of carcinomas of the oral cavity*. Tech. rep. G110. London, UK: The Royal College of Pathologists, 2023. URL: <https://www.rcpath.org/resourceLibrary/g110-dataset-for-histopathology-reporting-of-carcinomas-of-the-oral-cavity.html>.
- [17] R. W. H. Smits et al. “Resection margins in oral cancer surgery: Room for improvement”. In: *Head & Neck* 38.S1 (2016), E2197–E2203. DOI: 10.1002/hed.24075.

- [18] A. Rubert et al. "Impact of surgical margins on recurrence and survival rate in patients with oral squamous cell carcinoma: A systematic review and meta-analysis". In: *Medicina Oral, Patología Oral y Cirugía Bucal* 31.2 (2026), e307–e316. DOI: 10.4317/medoral.27784.
- [19] Arpan K. Shah. "Postoperative pathologic assessment of surgical margins in oral cancer: A contemporary review". In: *Journal of Oral and Maxillofacial Pathology* 22.1 (2018), pp. 78–85. DOI: 10.4103/jomfp.JOMFP_185_16.
- [20] Elisa M. Barroso et al. "Performance of Intraoperative Assessment of Resection Margins in Oral Cancer Surgery: A Review of Literature". In: *Frontiers in Oncology* 11 (2021), p. 628297. DOI: 10.3389/fonc.2021.628297.
- [21] Harald C. Groen et al. "Three-dimensional registration of histology of human atherosclerotic carotid plaques to in-vivo imaging". In: *Journal of Biomechanics* 43.10 (2010), pp. 2087–2092. DOI: 10.1016/j.jbiomech.2010.04.005.
- [22] L. A. Umstattd et al. "Shrinkage in oral squamous cell carcinoma: An analysis of tumor and margin measurements in vivo, post-resection, and post-formalin fixation". In: *American Journal of Otolaryngology* 38.6 (2017), pp. 660–662. DOI: 10.1016/j.amjoto.2017.08.011.
- [23] C. Burns and M. Gorina Faz. "An Analysis of Tumor Margin Shrinkage in the Surgical Resection of Squamous Cell Carcinoma of the Oral Cavity". In: *Cureus* 13.5 (2021), e15329. DOI: 10.7759/cureus.15329.
- [24] H. A. El-Fol et al. "Significance of post-resection tissue shrinkage on surgical margins of oral squamous cell carcinoma". In: *Journal of Cranio-Maxillofacial Surgery* 43.4 (2015), pp. 475–482. DOI: 10.1016/j.jcms.2015.01.009.
- [25] Patrizia Morbini et al. "The impact of histology and molecular biology in refining the prognostic value of surgical margins in head and neck oncology – A narrative review". In: *Acta Otorhinolaryngologica Italica* 45.2 Suppl 1 (2025), S137–S157. DOI: 10.14639/0392-100X-suppl.1-45-2025-N1141.
- [26] Nicolaas M. Bekedam et al. "Implementing a deep learning model for automatic tongue tumour segmentation in ex-vivo 3-dimensional ultrasound volumes". In: *British Journal of Oral and Maxillofacial Surgery* 62.3 (2024), pp. 284–289. DOI: 10.1016/j.bjoms.2023.12.017.
- [27] Fabian Isensee et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature Methods* 18.2 (2021), pp. 203–211. DOI: 10.1038/s41592-020-01008-z.
- [28] N. M. Bekedam et al. "Improved ground truth annotation by multimodal image registration from 3D ultrasound to histopathology for resected tongue carcinoma". In: *European Archives of Oto-Rhino-Laryngology* (2024). DOI: 10.1007/s00405-024-08979-1.
- [29] Thomas L. Szabo. *Diagnostic Ultrasound Imaging: Inside Out*. 2nd ed. Academic Press (Elsevier), 2014. ISBN: 978-0-12-396487-8.
- [30] Stefan G. Schalk et al. "3D surface-based registration of ultrasound and histology in prostate cancer imaging". In: *Computerized Medical Imaging and Graphics* 47 (2016), pp. 29–39. DOI: 10.1016/j.compmedimag.2015.11.001.
- [31] Aristeidis Sotiras, Christos Davatzikos, and Nikos Paragios. "Deformable medical image registration: A survey". In: *IEEE Transactions on Medical Imaging* 32.7 (2013), pp. 1153–1190. DOI: 10.1109/TMI.2013.2265603.
- [32] J. B. Antoine Maintz and Max A. Viergever. "A survey of medical image registration". In: *Medical Image Analysis* 2.1 (1998), pp. 1–36. DOI: 10.1016/S1361-8415(98)80001-7.
- [33] Stanley Osher and James A. Sethian. "Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations". In: *Journal of Computational Physics* 79.1 (1988), pp. 12–49. DOI: 10.1016/0021-9991(88)90002-2.
- [34] Mark Sussman, Peter Smereka, and Stanley Osher. "A level set approach for computing solutions to incompressible two-phase flow". In: *Journal of Computational Physics* 114.1 (1994), pp. 146–159. DOI: 10.1006/jcph.1994.1155.
- [35] Rogier R. Wildeboer et al. "Accurate validation of ultrasound imaging of prostate cancer: a review of challenges in registration of imaging and histopathology". In: *Journal of Ultrasound* 21.3 (2018), pp. 197–207. DOI: 10.1007/s40477-018-0311-8.

- [36] J. Michael Fitzpatrick, Jay B. West, and Calvin R. Maurer. "Predicting error in rigid-body point-based registration". In: *IEEE Transactions on Medical Imaging* 17.5 (1998), pp. 694–702. DOI: 10.1109/42.736021.
- [37] Lee R. Dice. "Measures of the amount of ecologic association between species". In: *Ecology* 26.3 (1945), pp. 297–302. DOI: 10.2307/1932409.
- [38] Abdel Aziz Taha and Allan Hanbury. "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool". In: *BMC Medical Imaging* 15.1 (2015), p. 29. DOI: 10.1186/s12880-015-0068-x.
- [39] Stanislav Nikolov et al. "Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study". In: *Journal of Medical Internet Research* 23.7 (2021), e26151. DOI: 10.2196/26151.
- [40] Auke Jager et al. "Development and validation of a framework for registration of whole-mount radical prostatectomy histopathology with three-dimensional transrectal ultrasound". In: *BMC Urology* 25.73 (2025). DOI: 10.1186/s12894-025-01736-4.
- [41] Mirabela Rusu et al. "Registration of presurgical MRI and histopathology images from radical prostatectomy via RAPSODI". In: *Medical Physics* 47.9 (2020), pp. 4177–4188. DOI: 10.1002/mp.14337.
- [42] Jake Pensa et al. "Evaluation of prostate cancer detection using micro-ultrasound versus MRI through co-registration to whole-mount pathology". In: *Scientific reports* 14.1 (2024), p. 18910.
- [43] Samuele Ghezzi et al. "[68Ga]Ga-PSMA and [68Ga]Ga-RM2 PET/MRI vs. Histopathological Images in Prostate Cancer: A New Workflow for Spatial Co-Registration". In: *Bioengineering* 10.8 (2023), p. 953. DOI: 10.3390/bioengineering10080953.
- [44] American Cancer Society. *Risk Factors for Oral Cavity and Oropharyngeal Cancers*. Last Revised: March 23, 2021. 2021.
- [45] Jaron G. de Wit et al. "EGFR-targeted fluorescence molecular imaging for intraoperative margin assessment in oral cancer patients: a phase II trial". In: *Nature Communications* 14.1 (Aug. 2023), p. 4952. DOI: 10.1038/s41467-023-40324-8. URL: <https://doi.org/10.1038/s41467-023-40324-8>.
- [46] Herve Y. Sroussi et al. "Common oral complications of head and neck cancer radiation therapy: mucositis, infections, saliva change, fibrosis, sensory dysfunctions, dental caries, periodontal disease, and osteoradionecrosis". In: *Cancer Medicine* 6.12 (2017), pp. 2918–2931. DOI: 10.1002/cam4.1221.
- [47] Cornelia G. F. van Lanschot et al. "Relocation of inadequate resection margins in the wound bed during oral cavity oncological surgery: A feasibility study". In: *Head & Neck* 41.7 (2019), pp. 2159–2166. DOI: 10.1002/hed.25690.
- [48] Sallie M. Long et al. "Use of Intraoperative Frozen Section to Assess Final Tumor Margin Status in Patients Undergoing Surgery for Oral Cavity Squamous Cell Carcinoma". In: *JAMA Otolaryngology–Head & Neck Surgery* 148.10 (2022), pp. 911–917. DOI: 10.1001/jamaoto.2022.2131.
- [49] Y. Aaboubout, M. R. Nunes Soares, T. C. Bakker Schut, et al. "Intraoperative assessment of resection margins by Raman spectroscopy to guide oral cancer surgery". In: *Analyst* 148.17 (2023), pp. 4116–4126. DOI: 10.1039/d3an00650f.
- [50] Stefan C. A. Steens et al. "Evaluation of tongue squamous cell carcinoma resection margins using ex-vivo MR". In: *International Journal of Computer Assisted Radiology and Surgery* 12.5 (2017), pp. 821–828. DOI: 10.1007/s11548-017-1524-6.
- [51] N. M. Bekedam et al. "Intra-operative resection margin model of tongue carcinoma using 3D reconstructed ultrasound". In: *Advances in Oral and Maxillofacial Surgery* 4 (2021), p. 100154. DOI: 10.1016/j.adoms.2021.100154.
- [52] Geneticist Biotech. *Paraffin-Embedded Tissue: What Is It?* Accessed: August 2026. 2023. URL: <https://geneticistusa.com/blog/paraffin-embedded-tissue-what-is-it>.
- [53] Marc Macenko et al. "A Method for Normalizing Histology Slides for Quantitative Analysis." In: vol. 9. June 2009, pp. 1107–1110. DOI: 10.1109/ISBI.2009.5193250.
- [54] Stéfan van der Walt et al. "scikit-image: image processing in Python". In: *PeerJ* 2 (2014), e453. ISSN: 2167-8359. DOI: 10.7717/peerj.453. URL: <http://dx.doi.org/10.7717/peerj.453>.

- [55] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. 2nd ed. Cambridge University Press, 2004.
- [56] H. W. Kuhn. “The Hungarian method for the assignment problem”. In: *Naval Research Logistics Quarterly* 2.1–2 (1955), pp. 83–97. DOI: 10.1002/nav.3800020109.
- [57] Pauli Virtanen et al. “SciPy 1.0: fundamental algorithms for scientific computing in Python”. In: *Nature Methods* 17.3 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2. URL: <https://doi.org/10.1038/s41592-019-0686-2>.
- [58] Bradley Lowekamp et al. “The design of simpleITK”. In: *Frontiers in neuroinformatics* 7 (Dec. 2013), p. 45. DOI: 10.3389/fninf.2013.00045.
- [59] Andriy Fedorov et al. “3D Slicer as an Image Computing Platform for the Quantitative Imaging Network”. In: *Magnetic Resonance Imaging* 30.9 (2012), pp. 1323–1341. DOI: 10.1016/j.mri.2012.05.001.
- [60] Matteo Cavicchioli et al. “Main challenges on the curation of large scale datasets for pancreas segmentation using deep learning in multi-phase CT scans: Focus on cardinality, manual refinement, and annotation quality”. In: *Computerized Medical Imaging and Graphics* 117 (2024), p. 102434. DOI: 10.1016/j.compmedimag.2024.102434.
- [61] Yiming Xiao et al. “REtroSpective Evaluation of Cerebral Tumors (RESECT): A clinical database of pre-operative MRI and intra-operative ultrasound in low-grade glioma surgeries”. In: *Medical Physics* 44.7 (2017), pp. 3875–3882. DOI: 10.1002/mp.12268.
- [62] Abhishek Nan et al. “DRMIME: Differentiable Mutual Information and Matrix Exponential for Multi-modal Image Registration”. In: *Proceedings of the Third Conference on Medical Imaging with Deep Learning*. Ed. by Tal Arbel et al. Vol. 121. Proceedings of Machine Learning Research. PMLR, 2020, pp. 527–543. URL: <https://proceedings.mlr.press/v121/nan20a.html>.

A

Appendix: Full Results Experiments

Table A.1: Ablation of the fiducial-aware joint refinement (step 4) and per-slice 2D refinement (step 8) against the full baseline pipeline. mTRE and landmark TRE are mean error (mm); mTRE is available only for patients 1–3 (traced fiducial centre-lines). Dice is whole-specimen volume overlap; NSD@1mm is surface agreement at 1 mm tolerance.

Patient	Variant	Dice	NSD@1mm	mTRE (mm)	Landmark TRE (mm)
1	Baseline	0.72	0.89	1.15	1.71
	No step 4	0.72	0.89	1.76	2.27
	No step 8	0.78	0.66	2.56	3.35
2	Baseline	0.84	0.88	2.46	1.71
	No step 4	0.85	0.88	3.15	2.72
	No step 8	0.84	0.78	2.75	2.19
3	Baseline	0.85	0.90	1.07	1.61
	No step 4	0.85	0.91	1.61	2.02
	No step 8	0.80	0.73	2.25	4.67
4	Baseline	0.75	0.85	–	2.14
	No step 4	0.79	0.89	–	2.48
	No step 8	0.81	0.70	–	3.87
5	Baseline	0.84	0.94	–	1.30
	No step 4	0.84	0.94	–	1.30
	No step 8	0.81	0.72	–	3.97

Table A.2: Composed-transform mTRE (mean, mm) for the three fiducial-bearing development-cohort specimens across values of the joint-refinement regulariser weight μ . The value used elsewhere in this study is $\mu = 10^{-2}$.

Patient	$\mu = 10^{-4}$	$\mu = 10^{-3}$	$\mu = 10^{-2}$	$\mu = 10^{-1}$	$\mu = 1$
1	11.64	6.67	1.15	1.25	1.42
2	10.55	3.01	2.46	2.71	3.10
3	6.39	1.05	1.07	1.12	1.28

Table A.3: Similarity versus affine transform model for the 3D refinement step (Section 2.3.2). mTRE and landmark TRE are mean values (mm); mTRE is available for patients 1–3 (traced fiducial centre-lines). Dice is whole-specimen volume overlap; NSD@1mm is surface agreement at 1 mm tolerance, both after the full pipeline.

Patient	Mode	Dice	NSD@1mm	mTRE (mm)	Landmark TRE (mm)
1	Similarity	0.72	0.89	1.15	1.71
	Affine	0.70	0.89	1.54	3.69
2	Similarity	0.84	0.88	2.46	1.71
	Affine	0.87	0.85	2.38	2.14
3	Similarity	0.85	0.90	1.07	1.61
	Affine	0.88	0.87	1.53	2.49
4	Similarity	0.75	0.85	–	2.14
	Affine	0.78	0.90	–	4.12
5	Similarity	0.84	0.94	–	1.30
	Affine	0.84	0.84	–	18.51

Table A.4: Manual initialisation alone (step 6) versus the full pipeline. mTRE and landmark TRE are mean error (mm); mTRE is available for patients 1–3 only. Dice is whole-specimen volume overlap; NSD@1mm is surface agreement at 1 mm tolerance.

Patient	Variant	Dice	NSD@1mm	mTRE (mm)	Landmark TRE (mm)
1	Full pipeline	0.72	0.89	1.15	1.71
	Manual only	0.75	0.65	2.98	4.69
2	Full pipeline	0.84	0.88	2.46	1.71
	Manual only	0.77	0.66	3.41	5.05
3	Full pipeline	0.85	0.90	1.07	1.61
	Manual only	0.81	0.71	2.17	4.79
4	Full pipeline	0.75	0.85	–	2.14
	Manual only	0.78	0.66	–	4.06
5	Full pipeline	0.84	0.94	–	1.30
	Manual only	0.76	0.63	–	9.97

B

Appendix: Flips and Rotations Results

Table B.1: Per-patient orientation review during step 3 (initial 2D registration). Flip and 90°-rotation counts summarise the manual GUI corrections applied on top of the automatic chirality-based flip detection. The second-to-last column report agreement between the manual decision and a re-computed automatic label: *same* counts slices where they matched, *diff* where they did not. The last column gives the specimen’s majority orientation relative to the automatic canonical choice: *flipped* where most slices had to be flipped, *not flipped* otherwise.

Cohort	Patient	slices <i>n</i>	Flipped		Rotated 90°		Auto agree same / diff	Majority orientation
			<i>n</i>	%	<i>n</i>	%		
Development	1	10	4	40.0	7	70.0	8 / 2	Not flipped
	2	8	3	37.5	4	50.0	3 / 5	Not flipped
	3	9	2	22.2	4	44.4	7 / 2	Not flipped
	4	7	4	57.1	3	42.9	1 / 6	Flipped
	5	12	10	83.3	11	91.7	2 / 10	Flipped
Evaluation	6	11	1	9.1	7	63.6	9 / 2	Not flipped
	7	9	2	22.2	4	44.4	6 / 3	Not flipped
	8	11	0	0.0	2	18.2	11 / 0	Not flipped
	9	10	0	0.0	7	70.0	10 / 0	Not flipped
	10	6	1	16.7	4	66.7	5 / 1	Not flipped
	11	9	8	88.9	2	22.2	1 / 8	Flipped
	12	6	3	50.0	0	0.0	3 / 3	Not flipped

Appendix: Per-Slice Information

Table C.1: Per-slice z -position, normalised position, area, fiducial hole count, and per-slice 2D registration quality. Dice and HD95 compare the histopathology slice mask against the US slice’s tissue mask in the same working plane (Section 2.3.3); HD95 is the 95th-percentile boundary-to-boundary distance, computed in each direction (histopathology-to-US and US-to-histopathology) and reported as the larger of the two. Scale and Rotation are the per-section 2D similarity transform’s recovered scale factor (US→histopathology) and in-plane rotation. Position is z normalised by the specimen’s total reconstructed length (z/z_{extent}), so it is comparable across patients of different lengths.

Patient	Slice	z (mm)	Position	Area (mm ²)	n holes	Dice	HD95 (mm)	Scale	Rotation (°)
P1	1	0.0	0.00	73.5	1	0.725	2.92	0.833	-66.6
P1	2	4.0	0.10	115.5	2	0.761	5.00	0.806	-25.5
P1	3	8.0	0.20	288.0	3	0.923	7.22	0.932	-13.8
P1	4	12.0	0.30	324.5	3	0.888	6.53	0.936	-10.1
P1 [†]	5	16.0	0.40	354.2	3	0.911	6.99	0.902	-33.7
P1	6	20.0	0.50	335.2	3	0.846	6.70	0.927	-13.9
P1	7	24.0	0.60	276.8	3	0.859	5.18	0.914	-23.9
P1	8	28.0	0.70	233.2	3	0.871	5.00	0.919	-28.0
P1	9	32.0	0.80	121.5	0	0.775	3.14	0.801	1.8
P1	10	36.0	0.90	89.8	0	0.759	2.50	0.835	3.2
P2	1	0.0	0.00	102.5	2	0.701	2.98	0.990	-0.3
P2	2	4.0	0.13	178.8	3	0.851	4.52	0.974	4.8
P2	3	8.0	0.25	275.0	3	0.928	6.02	0.963	-1.2
P2 [†]	4	12.0	0.38	312.8	3	0.911	6.50	0.945	-5.6
P2	5	16.0	0.50	322.0	2	0.886	6.77	0.920	-6.2
P2	6	20.0	0.63	315.2	3	0.887	6.10	0.925	-8.1
P2	7	24.0	0.75	281.2	2	0.911	5.25	0.963	9.3
P2	8	28.0	0.88	177.2	2	0.796	2.92	1.039	16.5
P3	1	0.0	0.00	86.2	0	0.645	2.50	1.141	-23.0
P3	2	4.0	0.11	156.8	2	0.838	3.50	1.015	-49.7
P3	3	8.0	0.22	194.8	2	0.888	4.00	0.938	-25.6
P3	4	12.0	0.33	226.5	3	0.903	3.78	0.922	-14.7
P3 [†]	5	16.0	0.44	256.5	3	0.902	4.47	0.928	-2.5
P3	6	20.0	0.56	272.5	3	0.894	5.00	0.883	-0.8
P3	7	24.0	0.67	314.0	2	0.908	5.50	0.873	19.5
P3	8	28.0	0.78	327.2	2	0.907	4.50	0.920	0.0
P3	9	32.0	0.89	265.5	1	0.850	4.47	1.101	-1.3
P4	1	4.0	0.13	222.8	3	0.878	3.54	1.012	35.1
P4	2	8.0	0.25	268.2	3	0.897	4.55	0.957	10.0
P4 [†]	3	12.0	0.38	323.8	3	0.883	6.06	0.937	10.5
P4	4	16.0	0.50	374.5	3	0.901	7.39	0.912	-4.3
P4	5	20.0	0.63	398.5	3	0.907	7.43	0.922	-10.5
P4	6	24.0	0.75	340.0	2	0.870	6.07	0.875	-14.8

Patient	Slice	z (mm)	Position	Area (mm ²)	n holes	Dice	HD95 (mm)	Scale	Rotation (°)
P4	7	28.0	0.88	334.5	0	0.870	3.58	1.005	-6.7
P5 [†]	1	0.0	0.00	68.5	0	0.896	1.00	0.988	-19.3
P5	2	4.0	0.08	140.8	0	0.861	2.37	0.998	0.0
P5	3	8.0	0.17	174.5	0	0.932	1.12	0.955	-55.1
P5	4	12.0	0.25	196.2	0	0.914	2.50	0.941	-11.3
P5	5	16.0	0.33	203.8	0	0.902	1.52	0.903	-45.6
P5	6	20.0	0.42	260.8	0	0.936	1.50	0.936	-45.0
P5	7	24.0	0.50	304.0	0	0.938	5.75	0.932	-9.3
P5	8	28.0	0.58	307.2	0	0.931	2.94	0.907	-0.5
P5	9	32.0	0.67	368.8	0	0.917	6.30	0.959	6.9
P5	10	36.0	0.75	337.5	0	0.883	3.12	0.962	1.5
P5	11	40.0	0.83	191.8	0	0.832	3.61	0.904	-17.5
P5	12	44.0	0.92	175.0	0	0.845	3.78	0.978	8.1
P6	4	16.0	0.31	135.0	3	0.335	4.51	1.241	24.9
P6 [†]	5	20.0	0.38	168.0	3	0.749	2.83	1.087	34.4
P6	6	24.0	0.46	200.2	3	0.847	4.33	1.016	49.3
P6	7	28.0	0.54	214.8	3	0.851	4.53	0.948	16.0
P6	8	32.0	0.62	237.5	2	0.855	5.50	0.933	-4.3
P6	9	36.0	0.69	215.2	3	0.754	7.16	0.922	-26.5
P6	10	40.0	0.77	326.5	1	0.839	6.56	0.978	-52.8
P6	11	48.0	0.92	108.0	0	0.748	4.96	0.948	12.4
P7	1	0.0	0.00	65.8	0	0.604	2.26	1.087	32.0
P7	2	4.0	0.11	93.8	0	0.579	5.59	0.896	-4.6
P7	3	8.0	0.22	252.8	3	0.837	6.75	0.876	44.9
P7	4	12.0	0.33	268.0	3	0.803	7.64	0.848	26.5
P7 [†]	5	16.0	0.44	320.2	3	0.877	8.77	0.847	0.8
P7	6	20.0	0.56	279.0	3	0.841	7.38	0.880	-8.7
P7	7	24.0	0.67	192.2	3	0.826	5.39	0.851	13.1
P7	8	28.0	0.78	138.8	2	0.811	3.56	0.953	7.7
P7	9	32.0	0.89	84.8	1	0.778	1.50	1.072	29.5
P8	1	0.0	0.00	193.5	2	0.533	5.01	1.146	8.0
P8	2	8.0	0.17	227.0	3	0.852	6.61	0.928	52.6
P8	3	12.0	0.25	266.2	3	0.846	6.09	0.925	16.1
P8 [†]	4	16.0	0.33	274.2	3	0.795	7.16	0.891	14.5
P8	5	20.0	0.42	332.5	1	0.756	7.15	0.960	36.9
P8	6	24.0	0.50	325.2	2	0.686	7.09	1.022	-7.1
P8	7	28.0	0.58	314.5	2	0.809	7.16	0.928	3.5
P8	8	32.0	0.67	342.2	2	0.810	8.06	0.948	-7.8
P8	9	36.0	0.75	333.5	1	0.734	8.00	0.989	8.7
P8	10	40.0	0.83	249.5	0	0.807	5.50	0.889	-13.1
P8	11	44.0	0.92	182.8	0	0.863	2.34	0.975	-5.6
P9	1	0.0	0.00	60.2	0	0.778	2.12	0.961	-34.8
P9	2	4.0	0.10	142.8	2	0.837	2.55	0.976	-2.8
P9	3	8.0	0.20	209.2	1	0.855	4.57	0.926	1.3
P9	4	12.0	0.30	256.8	2	0.861	5.15	0.919	0.1
P9 [†]	5	16.0	0.40	279.2	2	0.835	5.90	0.894	-3.0
P9	6	20.0	0.50	339.8	2	0.867	6.02	0.918	-5.5
P9	7	24.0	0.60	343.2	2	0.872	6.11	0.934	-8.3
P9	8	28.0	0.70	316.5	2	0.835	6.07	0.967	-3.7
P9	9	32.0	0.80	248.2	2	0.792	6.58	0.928	-22.7
P9	10	36.0	0.90	219.5	1	0.770	4.00	0.956	21.8
P10	1	0.0	0.00	84.2	1	0.672	3.01	0.976	8.4

Patient	Slice	z (mm)	Position	Area (mm ²)	n holes	Dice	HD95 (mm)	Scale	Rotation (°)
P10	2	4.0	0.17	219.5	2	0.783	3.51	0.933	8.9
P10 [†]	3	8.0	0.33	257.2	3	0.822	4.98	0.870	-0.4
P10	4	12.0	0.50	290.8	2	0.844	7.81	0.848	-3.9
P10	5	16.0	0.67	311.0	0	0.833	7.64	0.880	4.3
P10	6	20.0	0.83	308.0	2	0.775	4.92	1.065	36.0
P11	1	0.0	0.00	178.0	0	0.742	2.53	0.968	72.5
P11	2	4.0	0.11	306.8	3	0.883	7.50	0.920	19.9
P11	3	8.0	0.22	290.5	3	0.835	7.94	0.921	18.3
P11	4	12.0	0.33	280.5	3	0.876	6.94	0.891	-5.1
P11 [†]	5	16.0	0.44	271.8	3	0.891	6.95	0.895	-27.6
P11	6	20.0	0.56	248.8	3	0.878	7.73	0.882	-45.1
P11	7	24.0	0.67	194.8	2	0.833	6.85	0.865	-45.5
P11	8	28.0	0.78	163.8	2	0.833	5.39	0.921	-64.3
P11	9	32.0	0.89	66.0	1	0.550	2.95	0.973	-10.5
P12	1	0.0	0.00	278.2	2	0.848	4.03	0.909	7.2
P12	2	4.0	0.17	325.2	2	0.870	6.29	0.878	-0.6
P12 [†]	3	8.0	0.33	342.5	2	0.848	5.75	0.910	-13.4
P12	4	12.0	0.50	305.2	0	0.900	2.70	0.877	0.8
P12	5	16.0	0.67	252.0	1	0.810	6.74	0.882	39.1
P12	6	20.0	0.83	180.2	0	0.817	3.55	0.935	4.3

[†] anchor section (held fixed, $\delta = 0$ by construction). Patient 6 has 11 slices for 13 physical sections: sections 3 and 12 were each placed on the same slide as a neighbouring section rather than on their own, so they do not appear as separate rows. The first three slices fell entirely outside the US volume's extent (zero overlapping US pixels), so step 8's 2D fit (Section 2.3.3) could not run; the table therefore starts at slice 4.

D

Appendix: Error Direction Analysis

The landmark TRE metric reported in the main text (Section 3.3.2) is magnitude-only. This appendix checks whether the residual landmark errors also have a systematic direction, since a systematic bias would propagate directly into any segmentation model trained on registered labels, whereas random error partly averages out (Section 4.5).

For each landmark pair, the signed displacement vector was defined as $\mathbf{d} = \mathbf{p}_{\text{histo,warped}} - \mathbf{p}_{\text{US}}$, so that $\|\mathbf{d}\|$ equals the per-landmark TRE. Each $\mathbf{d}$ was decomposed into a radial component relative to the US specimen centroid, $d_{\text{radial}} = \mathbf{d} \cdot \mathbf{r} / \|\mathbf{r}\|$ with $\mathbf{r} = \mathbf{p}_{\text{US}} - \mathbf{c}_{\text{US}}$, and into components along the US coordinate axes. Non-zero bias was tested with the Wilcoxon signed-rank test and the sign test, both per-patient and pooled across the cohort.

The analysis shows a small, consistent inward radial bias: the warped histopathology landmarks sit slightly closer to the specimen centroid than their US counterparts (mean ~ 0.44 mm). This radial bias does not vary systematically along the specimen’s long axis, and reached individual significance in only two of the five development patients. A separate, smaller bias (~ 0.30 mm) was found along the scanner’s posterior axis, a fixed world-frame direction independent of specimen orientation. Both effects are small against the overall landmark TRE (~ 2 mm) and should be re-evaluated on a larger cohort before any automated correction is applied at the registration stage.

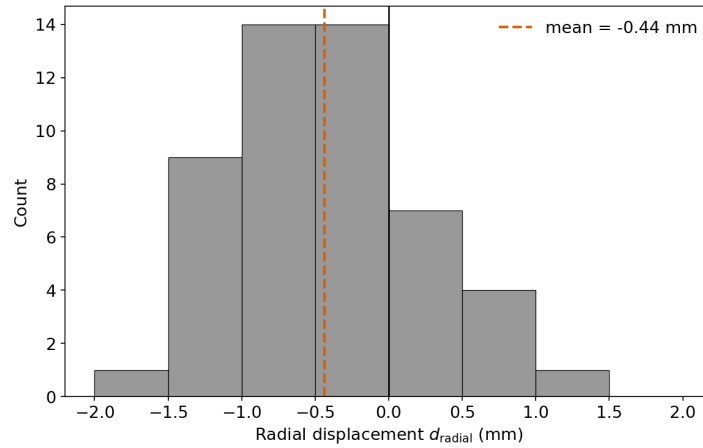


Figure D.1: Cohort-pooled distribution of the radial displacement d_{radial} across all 50 development landmarks (median: -0.47 mm; 38 of 50 landmarks displaced inward). Negative values indicate inward displacement, toward the specimen centroid.

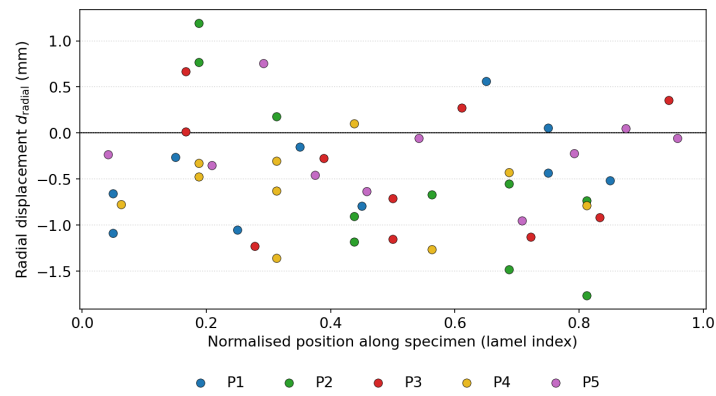
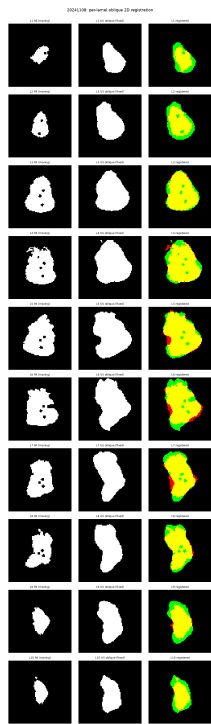


Figure D.2: Radial displacement per landmark plotted against normalised long-axis position (section index, rescaled per patient by total section count) and coloured by patient. No systematic trend along the long axis is visible (Spearman's $r_s = -0.12$, $p = 0.41$). Along the Cartesian axes, only the posterior axis showed a minor bias (mean: -0.30 mm, $p = 0.046$; other axes $p \geq 0.66$).

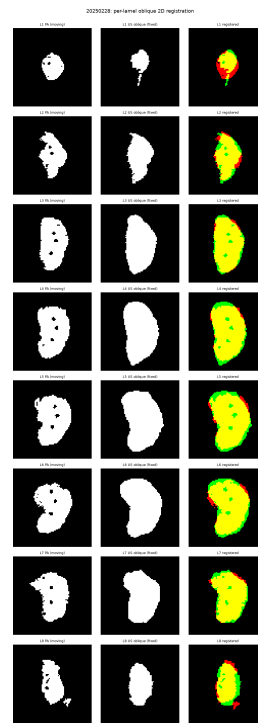
E

Appendix: Full Per-Slice 2D Refinement Results

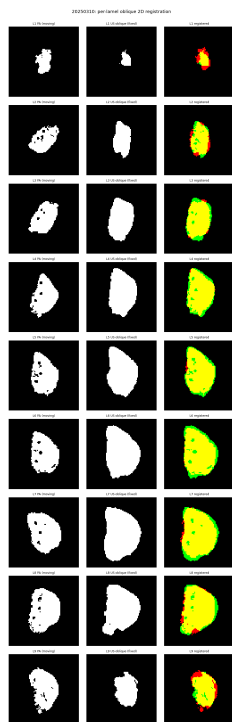
Figures E.1–E.3 show the per-slice 2D refinement QC overlay (Section 2.3.3) for all twelve patients.



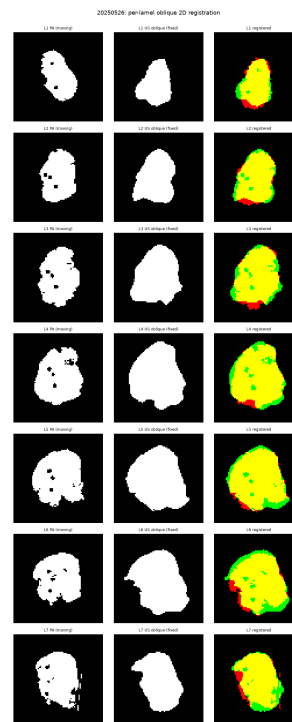
(a) Patient 1



(b) Patient 2

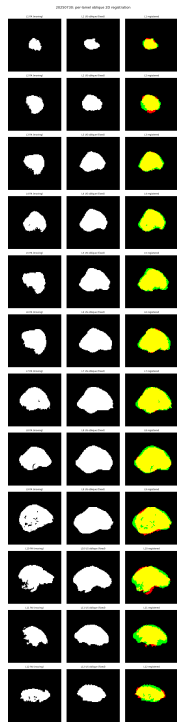


(c) Patient 3

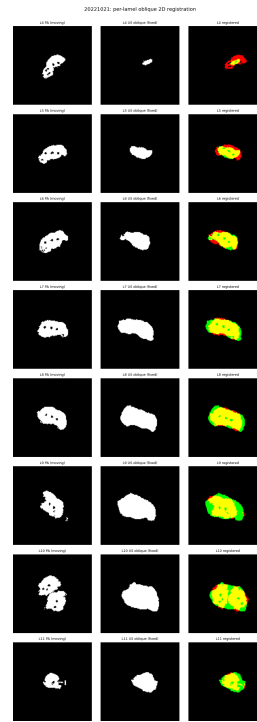


(d) Patient 4

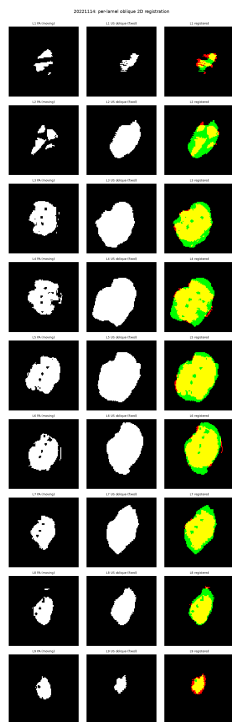
Figure E.1: Per-slice 2D refinement QC overlays, patients 1–4 (development cohort).



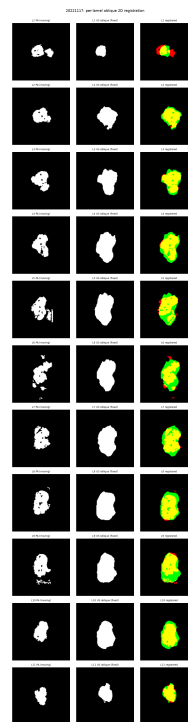
(a) Patient 5



(b) Patient 6

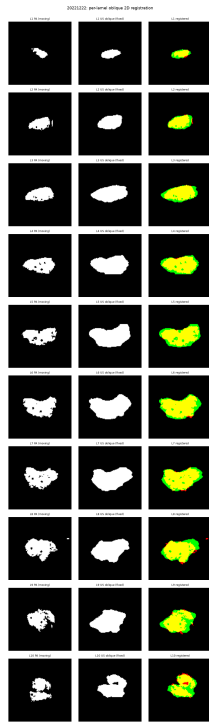


(c) Patient 7

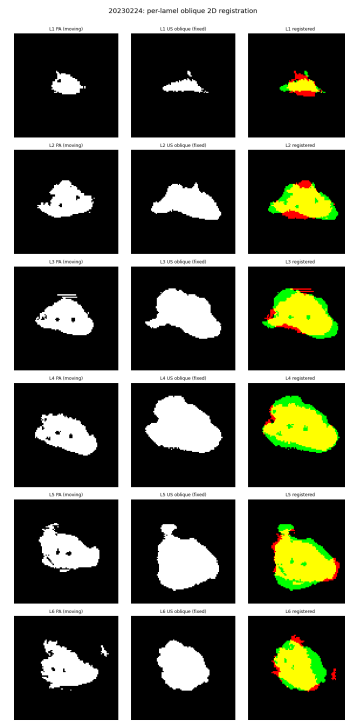


(d) Patient 8

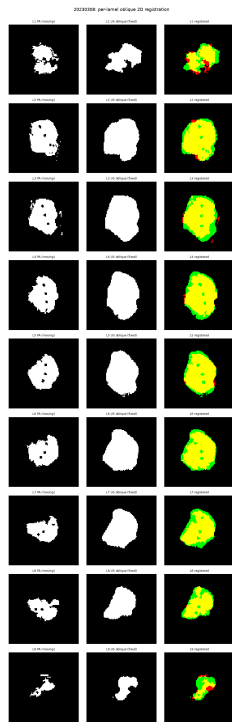
Figure E.2: Per-slice 2D refinement QC overlays, patients 5 (development cohort) and 6–8 (evaluation cohort).



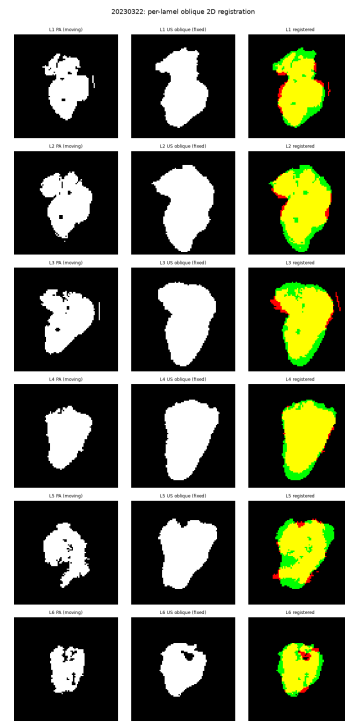
(a) Patient 9



(b) Patient 10



(c) Patient 11



(d) Patient 12

Figure E.3: Per-slice 2D refinement QC overlays, patients 9–12 (evaluation cohort).

F

Appendix: Full 2D Qualitative Results

Figures F.1–F.5 show the transferred histopathology tumour region overlaid on the corresponding 2D US slice for every histopathology slice with a tumour annotation, one page per development-cohort patient.

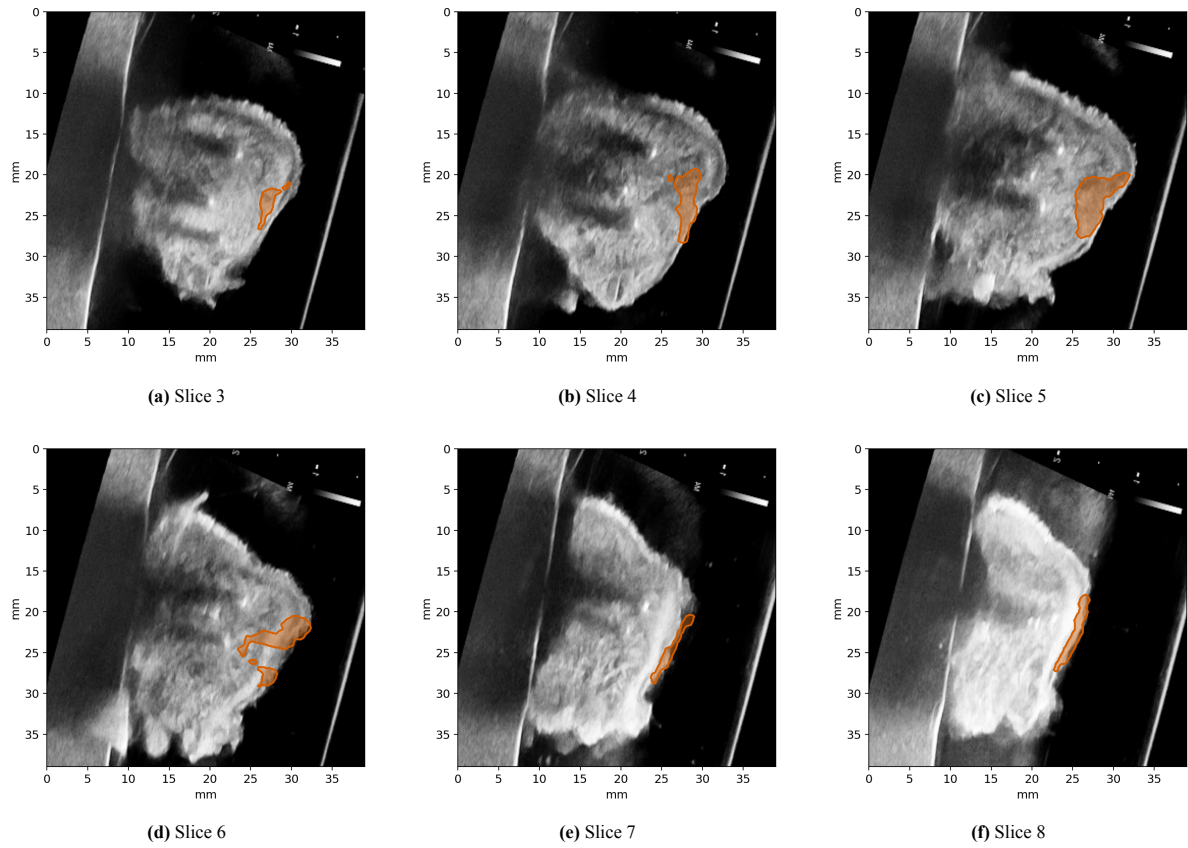


Figure F.1: Transferred tumour region in US space, patient 1.

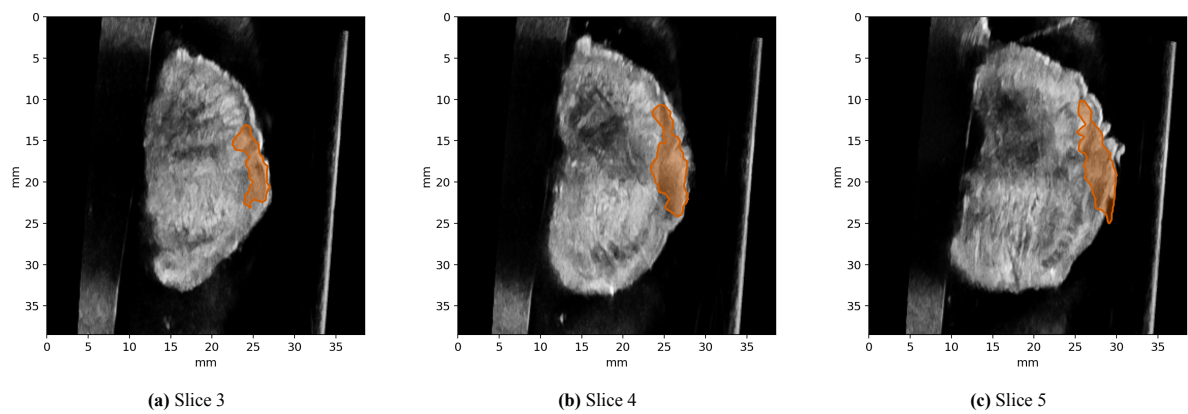


Figure F.2: Transferred tumour region in US space, patient 2.

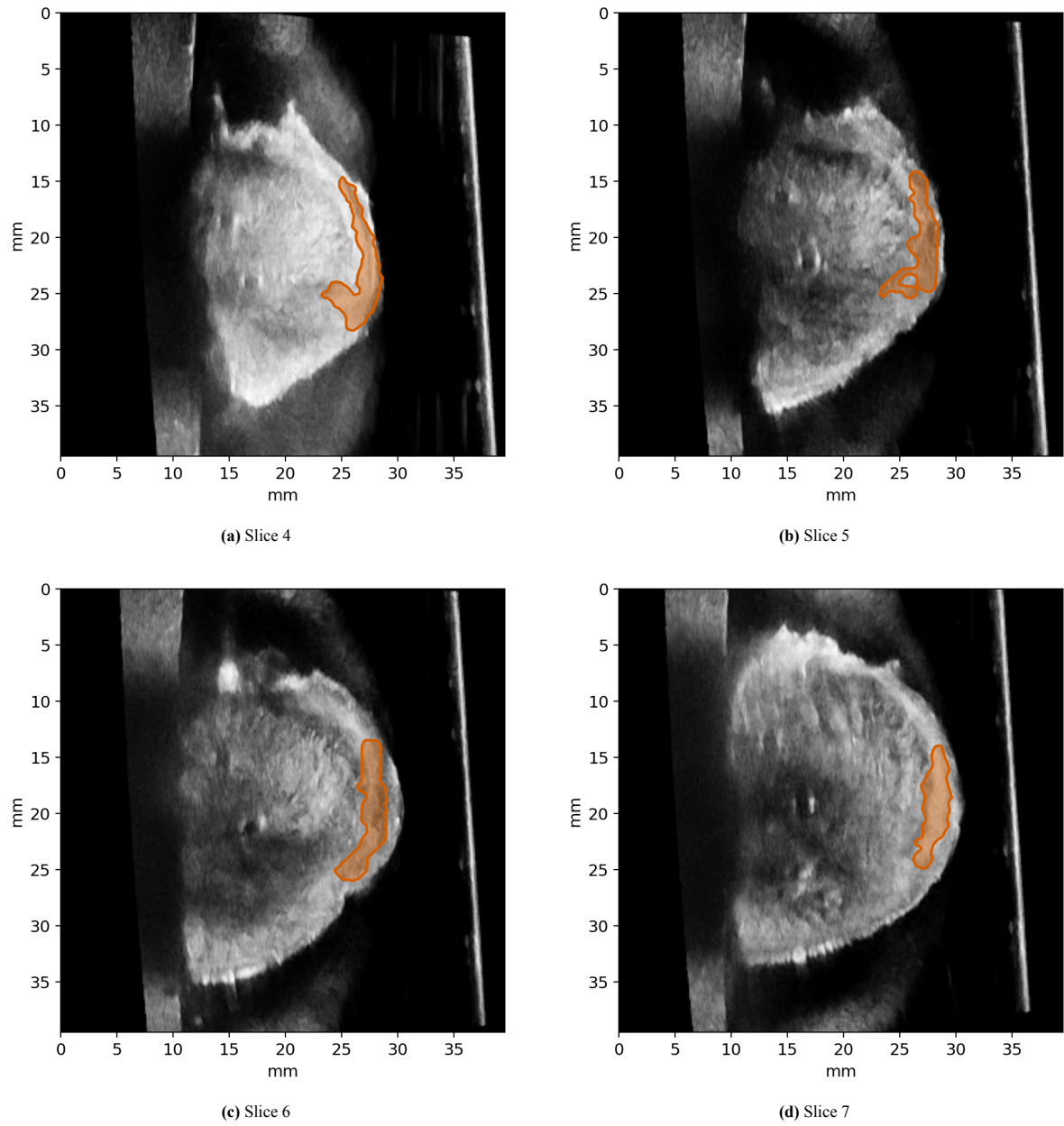


Figure F.3: Transferred tumour region in US space, patient 3.

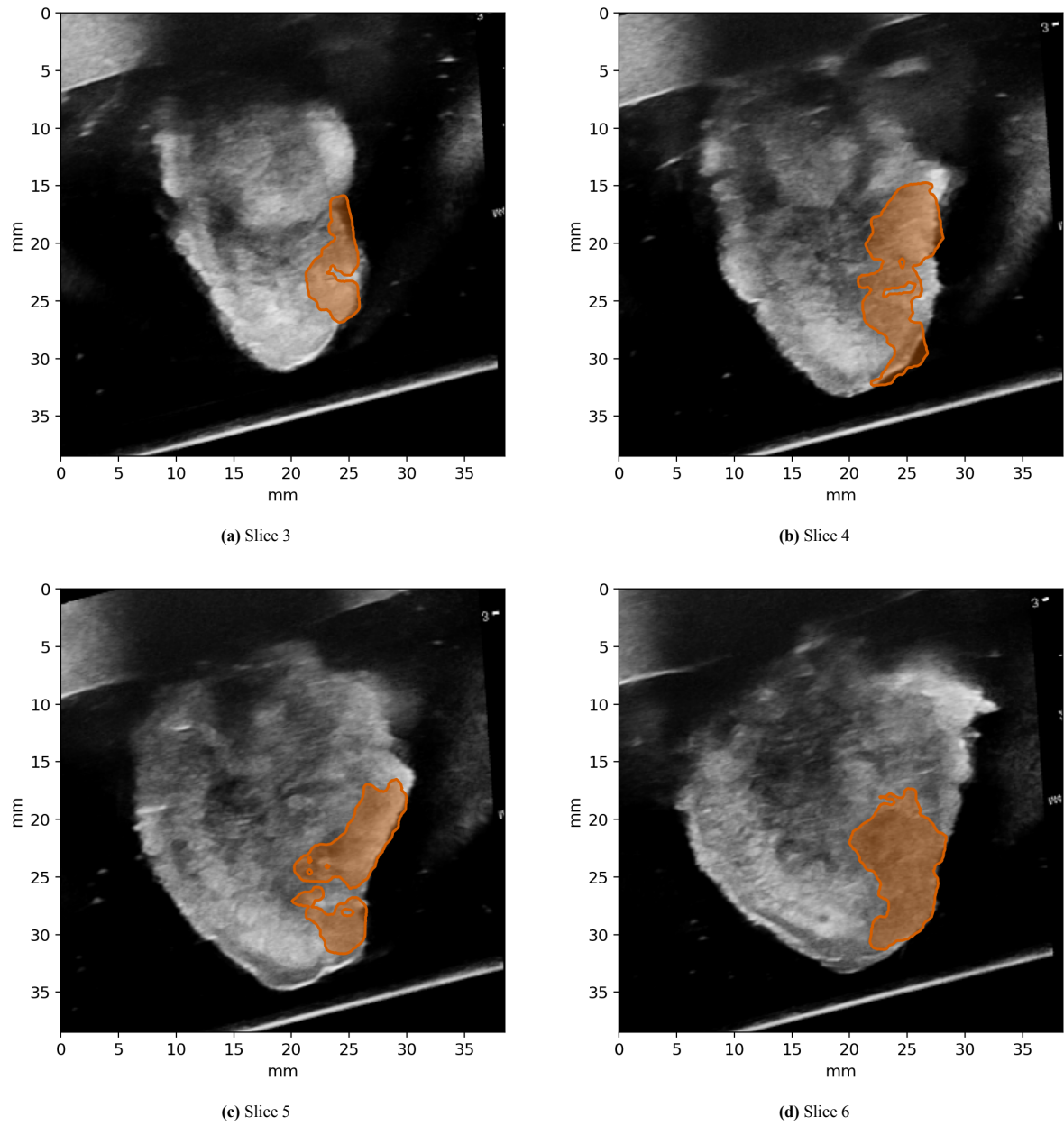


Figure F.4: Transferred tumour region in US space, patient 4.

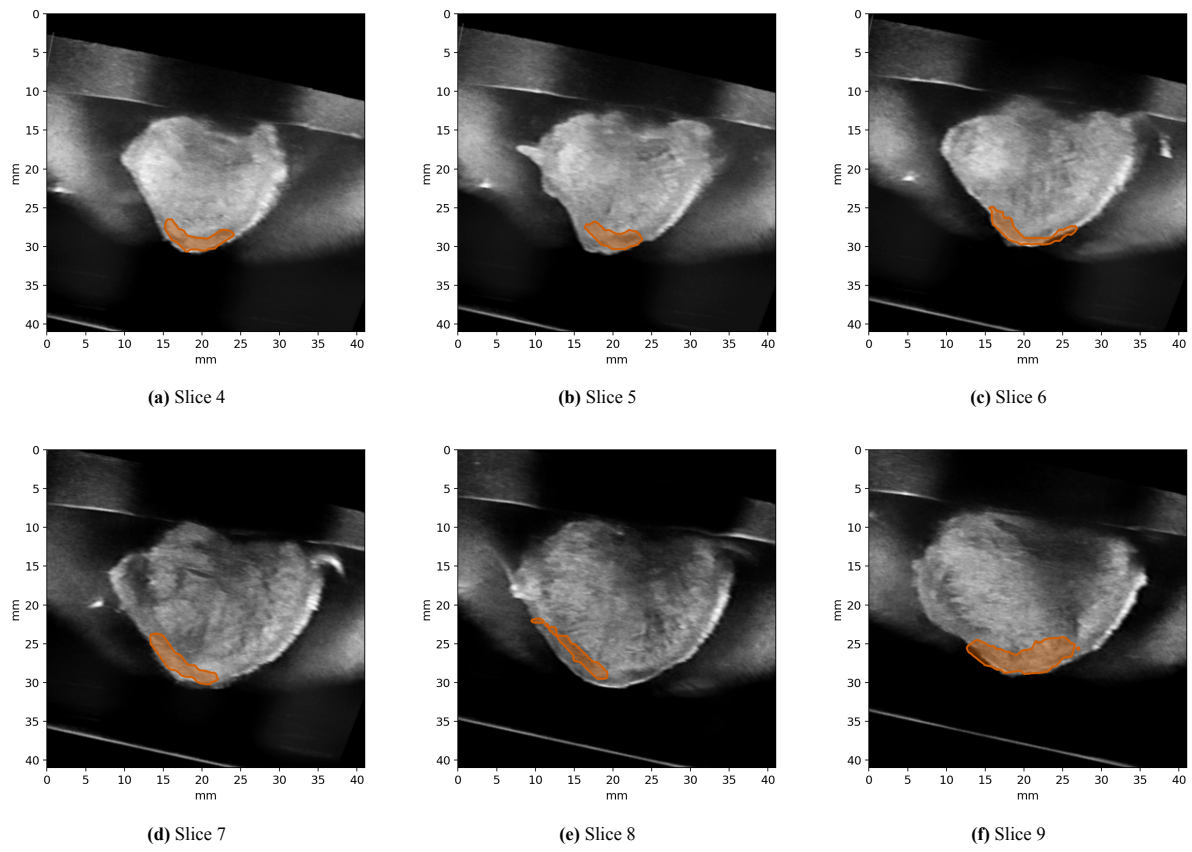


Figure F.5: Transferred tumour region in US space, patient 5.



Appendix: Standard Operating Procedure

Standard Operating Procedure

Ex-Vivo 3D Ultrasound Tongue Specimen Acquisition and Post-Processing (BK5000 System)

Document ID: SOP-3DUS
Date: 27-08-2026
Version: 1.0

Target audience

- 3D lab and AI for oncology researchers
- Head and neck oncology surgeons
- OR nurses

Introduction / background

The primary objective of this procedure is to intraoperatively assist the surgeon in achieving a clear pathological resection margin (>5.0 mm) during tongue squamous cell carcinoma (TSCC) surgery.

Aiming for a microscopic margin of >5.0 mm requires a larger margin of about 7.3 mm on the fresh ex-vivo specimen, which combines the effects of tissue shrinkage, ultrasound resolution, and uncertainty in where the margin is measured (roughly 1.0 cm in vivo). Ex-vivo 3D ultrasound (3D US) provides rapid, volumetric feedback on possible close or positive margins to help the surgeon decide if an immediate additional resection is required.

Abbreviations

OR	Operating room
TSCC	Tongue Squamous Cell Carcinoma
NaCl	Sodium Chloride
2D	Two-Dimensional
3D US	Three-Dimensional Ultrasound (BK5000)
AI	Artificial Intelligence
GUI	Graphical User Interface
DVI	Digital Visual Interface
HDMI	High-Definition Multimedia Interface
USB	Universal Serial Bus

Roles and responsibilities

Role	Responsibilities
Head and neck surgeon	Performs (re-)resection and marks specimen's anatomical orientation with 3 surgical sutures (long/long, long/short, short/short). Remains sterile in the operating field. Communicates with the researcher to guide visualisation in feedback interface. Assesses prediction and interprets need for intraoperative re-resection (sole decision-making authority).
Researcher (3D lab and AI for oncology)	Orders (if applicable) and collects items. Coordinates and prepares the location for the set-up. Operates the motorised rail and the BK5000 scanner. Runs the AI prediction and analysis. Manipulates the 3D interface to present multi-planar cross-sections to the sterile surgeon. Cleans up set-up.
OR nurse	Is responsible for specimen delivery to pathology.

Materials

Hardware

- Laptop + laptop charger
- USB-hub, Magewell, DVI-HDMI converter
- Motor-arm system, motor-arm system power cable, motor-arm system USB cable
- Ultrasound: BK5000 system, US-probe 18L5
- Mobile table, 3x power connectors

Consumables

- 1x Transparent scanning container/tray with a gel pad inlay
- 1x Cellulose mat (OR storage)
- 3x 16G Grey infusion needles (OR)
- 1x Parker Ultrasound gel pad (Order via secretary in packs of 6)
- Three to five bottles of 500 mL NaCl 0.9% (OR storage)
- Surgical gloves (OR storage)

Preparation

Set-up

1. Position mobile table in OR preparation room or other designated scanning area. Lay the cellulose mat on the right side of the table.
2. Place the scanning container on top of the cellulose mat. Insert the US gel pad into the designated ring at the bottom of the container.
3. Position the motorised rail system over the container, ensuring the track is aligned to allow the probe to translate along the full length of the container without obstruction.
4. Place the 18L5 probe in the custom probe holder.
5. Fill the container with three to five bottles of sterile 0.9% NaCl saline, depending on the expected size of the specimen. The specimen will need to be fully submerged once it's placed.

6. Place the laptop on the left side of the table. Connect the laptop charger, the motorised system power cable, and the BK5000 US system's power cable to power connectors.
7. Connect the motorised system to the laptop using its USB cable. Connect the BK5000 system's video output to the Magewell capture card using the DVI-to-HDMI converter and HDMI cable. Plug the Magewell's USB cable into the laptop's USB-hub, and connect the hub to the laptop.
8. Plug the 18L5 transducer connector into the active probe port of the BK5000. Switch on the BK5000 system. Make sure the right settings are in place.

Software (see README.md for installation)

1. On the laptop launch PyCharm.
2. Activate the environment, if not yet activated automatically.
3. Run the main acquisition script: `GUI_pyqt.py`
4. Open Windows File Explorer and verify that the laptop has active connection privileges to the network database folder: `\\clin-storage\3DLab\BK Medical 3D US\Database3D`
5. Output for each acquisition:
 - (a) `Stacked3Dvolume_US<n>_<date>_0000.nii.gz` – the 3D volume with (x, y, z) spacing
 - Deep Learning Pipeline: This reconstructed 3D US volume serves as the raw input for the trained deep learning model (nnU-Netv2). The model automatically segments the specimen boundary and the hypoechoic tumour region. From these predicted segmentations, the surgical margins are retrieved and visualised.
 - (b) `Stacked3Dvolume_US_info<n>_<date>_0000.nii.gz` – a single 2D snapshot with x/y spacing
 - (c) `Image_header_US<n>_<date>.json` – acquisition metadata
 - (d) `logbook_US_Acq_<n>_<timestamp>.log` – detailed log

The files are saved in `Database3D/Motorized_Acq_<date>/US_Acq_<n>/`.

Procedure description

1. Specimen resection
2. Specimen placement
3. Motorised 3D US scan
4. AI segmentation
5. Present multi-planar map
6. Surgeon decision

Phase 0: Preparation

- The set-up is prepared.
- The specimen is resected with the goal of >5 mm microscopic margin.
- The specimen's orientation is marked with 3 different kinds of sutures.

Phase A: Setup and scan (duration: < 5 min)

- The specimen is placed on a US gel pad inside the container filled with 0.9% saline until the specimen is completely submerged.
- The specimen is scanned.

Phase B: Automated image processing (duration: < 2 min)

- The 3D US volume is automatically segmented using a trained deep learning model.
- The system calculates every distance from the tumour edge to the specimen surface.

Phase C: Sterile feedback and anatomical relocation (duration: < 5 min)

- The specimen, tumour and margins are registered to the patient's preoperative MRI and visualised in a multi-planar interface in the OR.
- If all margins are found to be clear → no re-resection.
- If surgeon decides margins are not all clear → intraoperative re-resection.

Visualisation and Intra-operative Feedback

1. The surgeon remains fully sterile in the operating field and verbally directs the non-sterile researcher to manipulate the 3D interface.
2. The researcher aligns the reconstructed 3D specimen volume using the physical surgical sutures to ensure correct coordinate mapping back to the patient's wound bed.
3. The interface displays the reconstructed ultrasound volume in three aligned orthogonal 2D planes (axial, sagittal, coronal).
4. Resection margins are calculated automatically using the Signed Distance Field (SDF) and displayed as a 3D surface model with a standardised colormap:
 - Red (< 5.0 mm): Inadequate margin (close or positive). High recurrence risk.
 - Orange/Yellow (5.0 – 7.3 mm): Borderline margin (may fall below 5 mm once shrinkage and measurement uncertainty are included).
 - Green (> 7.3 mm): Clear margin.

Script handling and image processing

1. Run the GUI script. A Graphical User Interface will open.
2. In this GUI you select a new patient with the date of today, you select the right camera index and right serial port for the motorised rail, you select the right region of interest and make sure the settings are correct in the BK5000 (depth, width, frequency) and in the GUI (step size, length, depth).
3. Connect the motor by clicking on 'connect motor', test the motor by clicking on 'test motor' and finally, after the probe is positioned at the start of the specimen, run the motor by clicking on 'scan'.
4. Scan a few times and decide in 3D Slicer which scan is best.
5. Create a 3D volume with segmentations by running the trained nnU-Net. Display the segmented specimen, tumour and margins in the multi-planar interface with three 2D planes.

Training

- The researcher has been trained by a predecessor and must be fully competent in operating the set-up.
- The participating surgeons must undergo a brief introductory briefing on interpreting the margin colormaps.

H

Appendix: Adding and Running a New Patient in the Pipeline

Checklist: introducing one new patient, start to finish

The numbers below are script numbers (the filename prefixes, e.g. 9_PA_reconstruction.py), not the method steps of Chapter 2. Script 1–5 is preprocessing, script 6 is landmark extraction, and so on.

Do these **in this order**: several later scripts depend on a file you create by hand from an earlier script's output, so a new patient can't be run in one loop the way a fully staged one can.

Place the raw data

Under 1_US&PA\1_ULTRON\<new_patient>:

- PA\<slidename>.mrxs + its companion data folder (PA\<slidename>\Data0000.dat, Slidedat.ini, etc.) , one pair per physical label that was sectioned.
- US\Stacked3Dvolume_<new_patient>_0000.nii.gz , the US volume.
- US\mask_s-label.nii.gz , the US specimen mask (exported labelmap).

Register the patient in code: two separate places

A patient is processed only if their ID is in both of these hardcoded lists:

- **patient_registry.py** → add <new_patient> to ULTRON_PATIENT_IDS:

```
1 ULTRON_PATIENT_IDS = [  
2     "20221021", "20221114", "20221117", "20221222",  
3     "20230224", "20230308", "20230322",  
4     "<new_patient>",      # add here  
5 ]
```

Nothing else needs to change in this file for a new patient. UltronPatientEntry's paths are all derived directly from the patient ID (uniform naming scheme, no per-patient hand-typed fields needed).

- **1-5_PA_preprocessing_u.py** → add <new_patient> to PATIENT_ALLOWLIST:

```
1 PATIENT_ALLOWLIST = {  
2     "20221021", "20221114", "20221117", "20221222",  
3     "20230224", "20230308", "20230322",  
4     "<new_patient>",      # add here  
5 }
```

Run preprocessing (scripts 1–5, whole cohort)

```
1 python 1-5_PA_preprocessing_u.py
```

This processes every patient in the allowlist in one pass. Already-processed patients are skipped quickly at script 1 (existing downsampled output); the new patient gets fully processed through Macenko normalisation. No `--patient` flag needed (and using it would skip re-estimating things you actually want recomputed here).

Scripts 6–8 for the new patient only

```
1 python 6_PA_landmarks_u.py --root . <new_patient>
```

Check the QC overlay PNGs (fiducial holes, centroid, contour) look sane.

```
1 python 7_PA_initialisation_u.py --root . <new_patient>
```

In the GUI: check every slice’s orientation, not just the flip flag. Sections can come out mirrored, most will already be correct, but the automatic fiducial-chirality flip detection isn’t reliable on every slice. Left-click toggles flip, right-click rotates 90°. Compare each slice’s orientation against its neighbours before applying.

```
1 python 8_PA_fiducial_alignment_u.py --patient <new_patient> --root .
```

Optional QC, exports fiducials into Slicer:

```
1 python 8_fiducial_visualize_slicer_u.py --patient <new_patient> --root .
```

Create the lamel-mapping CSV, needs the pathology report

Before script 9 can run, create `lamel_mappings\<new_patient>.csv`:

```
1 v_number, lamel_number, notes
2 1,1,anterior end
3 2,2,
4 3,4,lesion starts here
5 ...
```

This is real per-patient data from the pathology/gross-sectioning report (which V-numbered slide is which true physical lamel, including gaps where a lamel wasn’t sectioned), not something that can be inferred or defaulted.

Run script 9 (3D reconstruction)

```
1 python 9_PA_reconstruction.py --patient <new_patient>
```

(No `--root` flag on this script, it uses its own hardcoded root already.) This produces `step9_reconstruction\<new_patient>\volume_sparse_gray.nii.gz` / `volume_sparse_mask.nii.gz`, the histo volume you need for the next step.

Create the manual Slicer 3D–3D init: needs script 9’s output

This can only happen now, since it needs the reconstructed histo volume that script 9 just produced, alongside the raw US volume. In Slicer: load both, manually align the histopathology specimen onto the US specimen (rough rigid alignment), and export/save the transform as an `.h5` file into:

`step10_initialisation_3D3D\<new_patient>\<any-name>.h5`

Any filename not starting with `init_` works: script 10’s auto-discovery falls back to any single non-`init_`-prefixed `.h5` in that folder if the canonical name doesn’t match.

Scripts 10–13 for the new patient

```
1 python 10_3D3D_initialisation.py --patient <new_patient> --root . --init-mode oriented
```

Always `--init-modeoriented`: reuses your manual Slicer rotation, re-anchors translation to the current reconstruction, rather than the ambiguous automatic PCA fallback.

```
1 python 11_3D3D_refinement.py --patient <new_patient> --root .
```

Default `--modesimilarity`, leave it unless you have a specific reason to suspect non-isotropic shrinkage.

```
1 python 12_2D2D_refinement.py --patient <new_patient> --root .
2 python 13_reconstruction_pa_in_us.py --patient <new_patient> --root .
```

Script 13 prints the final Dice/ASSD/NSD numbers for this patient.

Validation annotation (needles + landmarks): manual Slicer work, then run 14a2/14c2

In Slicer, trace the fiducial needle tracks and place landmark points on both the PA and US volumes, exporting each as its own markup file:

```
step14_validation\<new_patient>\needles\fid1_PA.json
step14_validation\<new_patient>\needles\fid1_US.json
step14_validation\<new_patient>\needles\fid2_PA.json
step14_validation\<new_patient>\needles\fid2_US.json
...
step14_validation\<new_patient>\landmarks\lm01_PA.json
step14_validation\<new_patient>\landmarks\lm01_US.json
...
```

Then:

```
1 python 14a2_validate2.py --patient <new_patient> --root .
2 python 14c2_landmarks_visualize.py --patient <new_patient> --root .
```

14a2 prints an mTRE-based pass/borderline/fail decision against the project's own protocol thresholds (< 2 mm sound, 2–3 mm borderline, > 3 mm troubleshoot).

What can and can't be batched for a new patient

Preprocessing (scripts 1–5) and scripts 6–8 can run unattended, except for script 7's GUI pause. Scripts 9–13 cannot be pre-looped for a brand-new patient, because script 9 needs the lamel CSV (which needs the pathology report) and script 10 needs a manual Slicer init (which needs script 9's own output), there is a hard dependency chain through human-created files in the middle of the automated part. Once both prerequisite files exist, scripts 10–13 (and 6–9, if rerunning after a correction) can be looped freely.

```
1 # Scripts 6-9 only, once lamel CSV exists:
2 python 6_PA_landmarks_u.py --root . <new_patient>
3 python 7_PA_initialisation_u.py --root . <new_patient> # GUI pause
4 python 8_PA_fiducial_alignment_u.py --patient <new_patient> --root .
5 python 9_PA_reconstruction.py --patient <new_patient>
6
7 # --- create the manual Slicer .h5 init here, using script 9's output ---
8
9 # Scripts 10-13, once the manual .h5 exists:
10 python 10_3D3D_initialisation.py --patient <new_patient> --root . --init-mode oriented
11 python 11_3D3D_refinement.py --patient <new_patient> --root .
12 python 12_2D2D_refinement.py --patient <new_patient> --root .
13 python 13_reconstruction_pa_in_us.py --patient <new_patient> --root .
```

Full loop, all currently-staged patients, scripts 6 through 13

For patients that **already have** their lamel CSV and manual .h5 init in place (i.e., not the “just introduced” case above), the whole 6–13 chain can run back to back:

```
1 $patients = @("20221021", "20221114", "20221117", "20221222", "20230224", "20230308", "20230322")
2 foreach ($p in $patients) {
3     Write-Output "=== $p ==="
4     python 6_PA_landmarks_u.py --root . $p
5     python 7_PA_initialisation_u.py --root . $p # GUI opens - review orientation/
        mirroring before continuing, slices need same orientation as before.
6     python 8_PA_fiducial_alignment_u.py --patient $p --root .
7     python 9_PA_reconstruction.py --patient $p # needs lamel_mappings\ $p.csv to already
        exist
8     python 10_3D3D_initialisation.py --patient $p --root . --init-mode oriented # needs a
        manual .h5 in step10_initialisation_3D3D\ $p\
```



```

9  python 11_3D3D_refinement.py --patient $p --root .
10 python 12_2D2D_refinement.py --patient $p --root .
11 python 13_reconstruction_pa_in_us.py --patient $p --root .
12 }

```

Add the new patient's ID to this list only after you've completed the checklist above at least once for them. Since script 7 is interactive, the loop still pauses at every patient's GUI review regardless.

Required input files, technical file reference

Every file and registry target below must exist in the layout shown for the workflow to run.

Raw data (staged before script 1)

Path root: 1_US&PA\1_ULTRON\<patient>\

File	Notes	Used by
PA\<slidename>.mrxs	One digital file per physical lamel (e.g. 279339_T22_11105_III1_HE.mrxs).	Script 1
PA\<slidename>\ (companion dir)	Contains Data0000.dat and Slidedat.ini; required for the OpenSlide parser.	Script 1
US\Stacked3Dvolume_<patient>_0000.nii.gz	3D ultrasound volume (raw reconstruction).	Scripts 10–13
US\mask_s-label.nii.gz	US specimen mask (exported labelmap).	Scripts 10–13

User-generated files

Path root: 2_Preprocessed_ULTRON\

File	Notes	Used by
lamel_mappings\<patient>.csv	CSV with headers v_number, lamel_number, notes. Maps which V-number is which physical lamel, including gaps where a lamel wasn't sectioned. Needs the pathology report.	Script 9
step10_initialisation_3D3D\<patient>\<any-name>.h5	SimpleITK transform holding the manual alignment. Needs script 9's 3D grey reconstruction.	Script 10 (--init-mode oriented)

Manual quality-control annotations

Path root: 2_Preprocessed_ULTRON\step14_validation\<patient>\

File	Notes	Used by
needles\fid<k>_PA.json / needles\fid<k>_US.json	Slicer markups curve export (.mrk.json). One pair per needle (k = 1, 2, 3, ...).	14a2_validate2.py
landmarks\lm01_PA.json / landmarks\lm01_US.json	Slicer markups fiducial point lists.	14c2_landmarks_visualize.py

These annotations are traced by hand in 3D Slicer; there is no automatic generation.

Patient registry control gates

Target	Effect if missing
patient_registry.py → ULTRON_PATIENT_IDS	Master allowlist. If the patient is excluded, scripts 10–14 crash with a KeyError.
1-5_PA_preprocessing_u.py → PATIENT_ALLOWLIST	Step-gating list. If missing, the patient is silently skipped (no error raised).

Note: apex orientation

Every patient's `histo_apex_at` in `patient_registry.py` defaults to "unknown". This only matters for script 10's automatic pca init mode, which uses it to disambiguate anterior/posterior when set, since this project always uses `--init-modeoriented` instead (which gets its orientation from your manual Slicer init, not from `histo_apex_at`), leaving it as "unknown" for a new patient is fine and doesn't need to be set.