Deep Learning-Based Deformable Registration of Transrectal Ultrasound to MRI for Prostate Imaging

Presented by:
Shaw Campbell

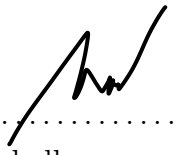
Prepared for:
Associate Professor Fred Nicolls
Dept. of Electrical and Electronics Engineering
University of Cape Town

Submitted to the Department of Electrical Engineering at the University of Cape Town in partial fulfilment of the academic requirements for a Bachelor of Science degree in Electrical and Computer Engineering

October 23, 2025

Declaration

1. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.
2. I have used the IEEE convention for citation and referencing. Each contribution to, and quotation in, this report from the work(s) of other people has been attributed, and has been cited and referenced.
3. This report is my own work.
4. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as their own work or part thereof.

Signature:  .....

Shaw Campbell

Date:23/10/2025.....

Terms of Reference

This project, supervised by Prof. Fred Nicolls, falls within the area of deep learning for deformable image registration. The work aims to explore how modern learning-based methods can be applied to the problem of non-rigid medical image alignment, particularly in multimodal contexts.

The scope of the project is intentionally broad to allow independent investigation and critical engagement with current techniques. The student is expected to identify a suitable registration setting, develop specific research questions, and implement and evaluate learning-based registration models.

The project objectives are to:

- Develop an understanding of traditional and learning-based deformable registration frameworks.
- Review and analyze existing unsupervised and weakly supervised registration methods.
- Implement a deep learning-based registration pipeline using a relevant medical dataset.
- Evaluate model performance using standard metrics and qualitative analysis.

The direction of the research and specific experimental focus are determined by the student in consultation with the supervisor. This document outlines the general research area, objectives, and expectations for independent academic work.

Acknowledgments

Make relevant acknowledgements to people who have helped you complete or conduct this work, including sponsors or research funders.

Abstract

This study investigates weakly supervised deformable registration of transrectal ultrasound (TRUS) and magnetic resonance imaging (MRI) for prostate imaging. The work focuses on evaluating the effectiveness of intensity-based similarity metrics versus label-driven loss formulations, the impact of affine preprocessing, and the role of data augmentation in improving model robustness. Using the MuRegPro dataset and a modified VoxelMorph framework, the research systematically explores how architectural and preprocessing choices affect multimodal registration accuracy. The findings demonstrate that intensity-based unsupervised registration using metrics such as mean squared error and normalized cross-correlation fails to capture anatomical correspondence across modalities, while weakly supervised models driven by label similarity produce anatomically consistent results. Affine preprocessing was shown to be essential in achieving stable deformation fields, and non-rigid augmentation improved robustness without degrading generalization. These results establish a practical foundation for anatomically faithful multimodal registration and highlight key methodological considerations for similar applications in medical image analysis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background to the Study

Medical image registration is the process of spatially aligning two or more images so that corresponding anatomical structures coincide. It forms the backbone of numerous clinical applications, including image-guided surgery, radiotherapy planning, and diagnostic fusion. Traditional registration methods, such as those implemented in ANTs and Elastix, formulate the problem as an optimization of a similarity metric under smoothness constraints. While effective in mono-modal cases such as MRI-MRI or CT-CT, these methods struggle when applied to multimodal data, where voxel intensities differ fundamentally in distribution and physical meaning [2].

The growing adoption of deep learning has shown promise in medical image registration by allowing networks to learn mappings between image pairs directly. Frameworks such as VoxelMorph [6] demonstrated that convolutional neural networks (CNNs) could achieve registration accuracy comparable to classical optimization-based methods while reducing runtime from minutes to seconds for MRI-CT registration. However, these models rely heavily on intensity-based similarity metrics such as mean squared error (MSE) or normalized cross-correlation (NCC), which assume that corresponding regions across images have similar voxel intensities - an assumption that does not necessarily hold in multimodal settings such as TRUS-MRI registration.

Among multimodal applications, MRI-TRUS registration of the prostate is of particular clinical importance. MRI provides high-resolution soft-tissue contrast and enables precise localization of tumors, while TRUS offers real-time imaging during biopsy and ablation

procedures. Accurately registering TRUS to MRI can therefore improve diagnostic yield and therapeutic targeting. Yet, the two modalities exhibit drastically different signal characteristics, making conventional similarity metrics unreliable. Recent studies [7, 8] have proposed weakly supervised, label-driven approaches that replace intensity-based losses with anatomical overlap constraints, but the relative importance of affine preprocessing and augmentation in this context remains unclear with respect to all datasets. This study investigates these aspects through controlled experiments using the µ-RegPro dataset [9].

1.1.1 Objectives of this Study

Problems to be Investigated

This research addresses three key questions in the domain of deep learning-based multimodal image registration:

1. **Is affine preprocessing essential for MRI–TRUS registration for data sets that appear roughly registered already?**
   While classical pipelines use affine alignment to remove global transformations before deformable refinement, deep networks theoretically have the capacity to learn both if the affine misalignment is small enough. This study tests whether excluding explicit affine preprocessing degrades performance for the µ-RegPro dataset, where the data looks roughly aligned already.

2. **Are traditional similarity metrics suitable for TRUS-MRI registration?**
   Common losses such as MSE and NCC measure intensity similarity but ignore structural correspondence. Their utility in TRUS-MRI registration is tested by comparing intensity-driven and label-driven loss formulations.

3. **Does affine data augmentation improve model robustness?**
   Augmentation is vital in medical imaging due to limited data availability. However, it is unclear whether small affine perturbations improve learning when affine alignment is already performed separately. The study compares non-rigid-only augmentation against combined affine-non-rigid perturbations using µ-RegPro metrics.

Purpose of the Study

The overarching goal of this study is to determine the necessary components of a reliable deep learning pipeline for multimodal MRI-TRUS registration of prostate scans within the context of the µ-RegPro dataset. By isolating and evaluating affine preprocessing, similarity loss formulation, and augmentation strategy, this research clarifies where deep learning should replace-and where it must complement-classical methods. The results are significant not only for prostate imaging but for multimodal registration more broadly, as they establish practical guidelines for developing robust, data-efficient registration models. This contributes to the wider goal of achieving clinically deployable, anatomically consistent multimodal fusion systems.

1.1.2 Scope and Limitations

Scope

The scope of this project is limited to multimodal deformable registration of transrectal ultrasound (TRUS) to magnetic resonance imaging (MRI) prostate volumes using deep learning. The study focuses on the µ-RegPro dataset, which provides paired MRI-TRUS scans, segmentation masks, and anatomical landmarks. The VoxelMorph architecture serves as the baseline model. The research investigates:

- The necessity of affine preprocessing prior to deformable registration for a dataset that looks roughly aligned with no affine alignment.
- The suitability of intensity-based versus label-driven similarity metrics.
- The effect of non-rigid and affine data augmentation on registration robustness.
- Evaluation using µ-RegPro metrics, including Dice, TRE, Hausdorff distance, and Jacobian smoothness.

Limitations

This study is limited by several practical constraints. First, only the training and validation sets of µ-RegPro are publicly available; the hidden test set was not accessible,

restricting external validation. Second, the dataset size (73 paired cases) limits the statistical power of model comparisons and the ability to train large, high-capacity networks. Third, due to computational resource limits and time constraints, experiments were limited to affine and deformable CNN architectures without exploring more advanced probabilistic or transformer-based models. Finally, while results provide insight into the design of multimodal registration systems, generalization to other organs or modalities (e.g., MRI-CT, PET-CT) or other datasets requires further study.

1.2 Plan of development

The remainder of this thesis is organized as follows. The Introduction provides a broad overview of the challenges in multimodal medical image registration and motivates the need for weakly supervised learning approaches, setting the context for TRUS-MRI alignment in prostate imaging. The Literature Review surveys relevant work in deformable registration, emphasizing intensity-based, label-driven, and hybrid deep learning approaches, as well as prior efforts in affine preprocessing and the use of weak supervision.

The Experimental Setup outlines the dataset, preprocessing procedures, model architecture, computational setup, and overall methodology used to test the research hypotheses. The Hypothesis Testing Experiments section presents the sequence of experiments conducted to evaluate unsupervised, weakly supervised, and augmented models, as well as the investigation into affine preprocessing. Each subsection describes the experimental design, results, and discussion of findings in a structured manner. The code used for this project can be found in the GitHub repository listed in Appendix F.

Finally, the Conclusion summarizes the main outcomes and contributions of this study, while the Future Recommendations section proposes potential avenues for extending this work, including improvements in model architecture, dataset diversity, and integrated registration frameworks.

Chapter 2

Literature Review

2.1 Introduction

Image registration is the process of establishing a spatial correspondence between two or more images of the same or related anatomical structures [1]. The transformation that maps one image to another can be categorized as rigid, affine, or deformable [2]. Rigid and affine transformations address global spatial differences using translation, rotation, scaling, and shearing and are typically represented by a 3 by 4 matrix for 3D volumes. These transformations are often sufficient to correct patient repositioning or gross motion between acquisitions. In contrast, deformable registration seeks to model local, non-linear shape variations caused by physiological motion, tissue deformation, or differing imaging conditions, using dense displacement fields that describe voxel-level correspondence throughout the volume. While rigid and affine registration have achieved maturity across many clinical contexts, deformable registration remains challenging, particularly when the images originate from different modalities [3].

In multimodal registration, the task extends beyond aligning spatial coordinates to reconciling inherently different image contrasts and acquisition characteristics. For instance, magnetic resonance imaging (MRI) offers superior soft-tissue contrast, while computed tomography (CT) excels in visualizing bone and high-density structures [4]. Fusing such complementary modalities enables integrated visualization and analysis, which is essential in treatment planning, surgical navigation, and radiotherapy. Similarly, in prostate interventions, the fusion of high-resolution MRI with real-time transrectal ultrasound (TRUS) is critical to guide biopsy and ablation procedures. However, the intensity relationships between modalities such as MRI and TRUS are complex and non-linear, making traditional

similarity-based registration unreliable [3].

In recent years, deep learning has emerged as a transformative approach to medical image registration. Instead of iteratively optimizing for each image pair, neural networks can learn a parametric function that predicts the spatial transformation directly, substantially reducing computational time while retaining accuracy. Zou et al. [3] provide a comprehensive overview of deep learning-based deformable registration methods, providing insight into the deep learning methods, regions of interest (ROIs) and evaluation metrics popular among researchers.

Their analysis reveals that, despite remarkable progress, most studies to date have focused on unimodal registration tasks - approximately 50% involving MRI-MRI and 30% CT-CT, while multimodal scenarios remain significantly underexplored. The majority of research has also been concentrated on brain imaging, which accounts for nearly half of all studies, whereas anatomically and physically complex regions such as the prostate comprise a smaller fraction. This distribution underscores the comparative difficulty of multimodal deformable registration and the scarcity of well-curated multimodal datasets.

The same review highlights that Dice similarity coefficient (DSC)-based evaluation is used in nearly half of all published studies. Since Dice metrics require anatomical label annotations, this trend suggests a shift toward methods that leverage anatomical supervision either explicitly or implicitly to overcome the ambiguity of intensity-based alignment. Zou et al. [3] also report that unsupervised and weakly supervised deep learning methods collectively dominate current research efforts, accounting for over half of the surveyed approaches. This reflects a growing recognition that obtaining voxel-level ground-truth deformation fields is impractical in medical imaging, where annotated data are costly and scarce. MRI acquisition alone can require hours of scanner time and expert supervision, while labeling or segmentation of each volume must be performed by trained clinicians, further limiting dataset availability.

2.2 The CuRIOUS2018 Challenge

A key reference point for multimodal registration research in the ultrasound to MRI space is the CuRIOUS2018 Challenge, which benchmarked MRI-intraoperative ultrasound (iUS) registration algorithms for brain-shift correction during tumor resection [5]. The challenge addressed a central obstacle in multimodal registration - the lack of common datasets and standardized evaluation metrics - by releasing paired MRI-iUS scans with manually

annotated landmarks and ranking submissions by the mean Euclidean distance between corresponding points.

Unlike segmentation challenges, CuRIOUS2018 deliberately avoided Dice or intensity-based similarity measures, reflecting the field's recognition that no single intensity metric (e.g., MSE, NCC, MI) consistently captures correspondence across imaging modalities. The organizers noted that there was no accepted standard metric for multimodal registration and therefore relied exclusively on geometric error between landmarks to compare algorithms. This decision implicitly underscores a key limitation of similarity-driven registration: even when two images appear well aligned in intensity space, the anatomical alignment can remain inaccurate. This speaks to the great challenge of assessing success in this space, necessitating qualitative assessment of results and inclusion of basic evaluation metrics such as mean Euclidean distance. In the context of this thesis, it seems appropriate to evaluate on as many metrics as possible.

Most of the top-ranked teams used intensity-driven or feature-based optimization approaches such as LC^2 or self-similarity context (SSC) features. These methods achieved mean post-registration errors around 1.5-2 mm, demonstrating that, with careful feature design and handcrafted optimization, robust alignment was possible. However, the challenge also revealed the fragility of these metrics in multimodal contexts: performance varied markedly across test cases, and the only deep-learning submission that initially excelled on training data (mean $\pm$ SD = 1.21 $\pm$ 0.55 mm) collapsed to 5.70 $\pm$ 2.93 mm on unseen test data, exposing severe overfitting and poor generalization. The authors attributed this to the model's reliance on intensity correspondence learned from limited data rather than modality-invariant anatomical cues. This problem is addressed in this thesis by investigating the effectiveness of intensity metrics and data augmentation to enhance a models ability to generalize.

An additional insight from the challenge concerned affine preprocessing. Several teams observed that affine transformations alone sometimes outperformed nonlinear deformable methods in pre-resection scans. This was explained by the fact that much of the MRI-iUS misalignment originated from inaccurate patient-to-MRI registration rather than local tissue deformation. Accordingly, teams that first performed affine alignment before deformable refinement achieved more stable results. This surprised teams, as the data seems roughly aligned with no affine preprocessing and one would think that a dense deformable registration model could fix any small affine misalignment.

2.3 VoxelMorph: Learning-based Deformable Registration

A major shift in deformable image registration emerged with the introduction of learning-based methods such as VoxelMorph, proposed by Balakrishnan *et al.* [6]. Unlike classical registration, which optimizes a similarity metric separately for each image pair, VoxelMorph formulates registration as a parametric learning problem in which a convolutional neural network learns to predict the spatial transformation directly from examples. Once trained, the network can register unseen image pairs in a single forward pass, eliminating the need for iterative optimization and dramatically improving computational efficiency. VoxelMorph addresses only the dense deformable vector field component of a registration pipeline.

Formally, given a moving image I_m and a fixed image I_f, VoxelMorph seeks a dense displacement field ϕ that warps I_m to align with I_f. The deformation field is defined as:

$$\phi = I_d + u$$

where I_d is the identity transform and u is the voxel-wise displacement vector predicted by a convolutional network $g_\theta(I_m, I_f)$ with trainable parameters θ. The warped image is obtained using a differentiable spatial transformer operator $\mathcal{W}$:

$$\tilde{I}_m = \mathcal{W}(I_m, \phi)$$

which performs trilinear interpolation to resample the moving image at the transformed coordinates.

Training proceeds by minimizing a composite loss function that balances image similarity and deformation regularity:

$$\mathcal{L} = \mathcal{L}_{\text{sim}}(I_f, \tilde{I}_m) + \lambda \, \mathcal{L}_{\text{smooth}}(u)$$

where λ controls the smoothness of the deformation. For unimodal images, the similarity term $\mathcal{L}_{\text{sim}}$ is typically the mean squared error (MSE) or local normalized cross-correlation (NCC), while the smoothness term penalizes large spatial gradients:

$$\mathcal{L}_{\text{smooth}}(u) = \sum_x \|\nabla u(x)\|^2$$

This ensures the learned deformation fields are smooth and anatomically plausible.

The mean squared error (MSE) and normalized cross-correlation (NCC) are two common similarity metrics used in image registration. MSE measures the voxel-wise intensity difference between the fixed image f and the warped moving image $m \circ \phi$, penalizing large discrepancies in intensity. It assumes that corresponding structures have similar intensity values, which makes it suitable for unimodal registration but unreliable for multimodal cases, where intensities may not correspond directly.

$$MSE(f, m \circ \phi) = \frac{1}{|\Omega|} \sum_{p \in \Omega} [f(p) - (m \circ \phi)(p)]^2$$

In contrast, NCC evaluates the linear relationship between intensity variations of the two images by normalizing for mean and variance. This makes NCC less sensitive to global intensity scaling or offset differences, and therefore more appropriate for multimodal registration tasks such as MRI-TRUS, where intensity distributions differ substantially between modalities. In the Voxelmorph paper, they experiment with NCC and MSE, because of their popularity, motivating this thesis to test on these two intensity metrics as well.

$$CC(f, m \circ \phi) = \sum_{p \in \Omega} \frac{\left(\sum_{p_i} (f(p_i) - \hat{f}(p)) \left[(m \circ \phi)(p_i) - \hat{m \circ \phi}(p) \right] \right)^2}{\left(\sum_{p_i} (f(p_i) - \hat{f}(p))^2 \right) \left(\sum_{p_i} [(m \circ \phi)(p_i) - \hat{m \circ \phi}(p)]^2 \right)}$$

While both metrics encourage intensity correspondence, NCC captures relative structural similarity rather than absolute intensity agreement, making it more robust for aligning images across modalities with differing contrast mechanisms.

The network adopts a U-Net-like encoder-decoder topology, where the encoder extracts hierarchical spatial features from the concatenated pair (I_m, I_f), and the decoder fuses coarse and fine features through skip connections to produce a dense displacement field. Like many papers in this field, I learning rate of 0.0005 and a lambda value of 0.5 was found to work well enough.

Although the base model is unsupervised, Balakrishnan *et al.* also explored the auxiliary use of anatomical labels during training by using the dice score of the labels for some training images, which improved the model performance.

VoxelMorph achieved registration accuracy comparable to traditional methods such as

ANTs and Elastix while reducing inference time from minutes to under a second per 3D volume pair for uni-modal and MRI-CT data. However, because its similarity metric relies on intensity correlation, its performance degrades in multimodal contexts where corresponding structures cannot be established based on intensity metrics. These limitations motivated the exploration of label-driven learning paradigms, where anatomical segmentations replace intensity similarity as the supervisory signal.

Voxelmorph is available as a library with tools like prebuilt registration CNNs, loss function metrics that can be plugged into Keras models such as NCC and MSE, spatial transformers and more. This thesis relied on their spatial transformer and loss metrics in particular.

2.4 Weakly Supervised Multimodal Registration: Hu *et al.* (2018)

Hu *et al.* [7] explicitly addressed this limitation in their seminal paper on weakly supervised convolutional neural networks (CNNs) for multimodal image registration. Their work represented a conceptual shift: rather than attempting to learn a universal intensity-based similarity metric, they proposed to replace intensity similarity with anatomical correspondence supervision derived from structural labels exclusively. This approach allows the network to learn transformations that align anatomically homologous regions, irrespective of local intensity patterns.

They formulated multimodal registration as a function g_θ that predicts a dense displacement field $\phi = g_\theta(I_m, I_f)$, warping the moving image I_m toward the fixed image I_f. Instead of computing loss directly on intensities, the network uses corresponding segmentation label maps L_m and L_f, minimizing the Dice similarity coefficient between the warped moving labels $\mathcal{W}(L_m, \phi)$ and the fixed labels L_f:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2|\mathcal{W}(L_m, \phi) \cap L_f|}{|\mathcal{W}(L_m, \phi)| + |L_f|}$$

This label-driven loss explicitly measures anatomical overlap rather than voxel intensity agreement, allowing the network to learn multimodal spatial relationships that reflect clinical relevance.

Their most important conclusion is that intensity similarity is an unreliable supervisory

signal for multimodal registration, as it may encourage visually plausible but anatomically incorrect alignments. By contrast, label-driven supervision enforces semantically meaningful correspondences, ensuring deformation fields respect organ boundaries and clinically relevant structures.

2.5 Label-Driven MRI-TRUS Registration: Zeng *et al.* (2020)

Building on the weakly supervised framework of Hu *et al.*, Zeng *et al.* [8] proposed a specialized deep learning pipeline for MRI-TRUS registration tailored to prostate radiotherapy. Their pipeline consists of three stages: segmentation networks for prostate masks, a 2D affine registration CNN predicting 12 affine parameters, and a 3D nonrigid registration network trained with weak anatomical supervision.

To overcome limited data, they used NiftyNet's augmentation layer combining elastic, affine, rotational, and flipping perturbations. This inspired the experimental design of this thesis, which explicitly isolated and evaluated the effects of affine and non-rigid augmentation.

They claim to have successfully achieved data segmentation, producing labels for their data without medical professionals. This means that if a good registration CNN is achieved, acquiring labels should not be too difficult in practice.

Their pipeline focuses on scalability in terms of labels. This is realized by making the model agnostic to the number and types of labels used by feeding in a random label during for each training pair. Even though the model learns from one random label per pair, they demonstrate that it learns to perform registration based on all of the labels at once during inference.

2.6 µ-RegPro Challenge

The µ-RegPro Challenge [9] was established to benchmark MRI-TRUS registration methods. Its dataset, derived from the SmartTarget Biopsy clinical trial, includes 108 MRI-TRUS pairs from 141 patients, with isotropic preprocessing, segmentation masks, and landmarks.

The public split includes 65 training and 8 validation cases.

The challenge defines a VoxelMorph-based baseline using MSE loss and L2 gradient regularization. Eight metrics comprehensively assess overlap, feature alignment, robustness, deformation smoothness, and computational efficiency: The final evaluation framework defines eight complementary metrics that together capture volumetric overlap, feature alignment accuracy, robustness to outliers, deformation smoothness, and computational efficiency. These are summarized as follows:

Dice Similarity Coefficient (DSC) [10, 11, 15]: Measures volumetric overlap between the warped TRUS prostate mask and the fixed MRI mask over the gland boundary, averaged across all cases. Values range from 0 to 1, with higher scores indicating better spatial correspondence. DSC reflects both localization and size agreement, which is critical in intraoperative augmented reality or lesion localization.

Robust Dice Similarity Coefficient (RDSC): Computed as the DSC averaged over the best-performing 68% of cases, mitigating the influence of outliers. The 68% threshold corresponds to one standard deviation in a Gaussian distribution, assuming normality of metric values.

Target Registration Error (TRE) [12]: Quantifies registration accuracy using anatomical landmarks identifiable in both MRI and TRUS images, such as the apex, base, urethra, visible lesions, zonal boundaries, and cysts. For each case, TRE is the root-mean-square (RMS) Euclidean distance between corresponding landmarks. The mean TRE across cases is normalized to the range [0, 1], assuming unregistered images define the maximum TRE. This directly measures clinically relevant alignment accuracy even when volumetric overlap appears high.

Robust Target Registration Error (RTRE): Averages the TRE over the best-performing 68% of cases to reduce the effect of extreme outliers or noisy annotations.

Robust Targets (RTs): Computes landmark error over the three lowest-error landmarks (out of five) per case, averaged across all test cases. This robustness criterion accounts for inconsistent or erroneous expert labels by excluding worst outliers.

95th Percentile Hausdorff Distance (95%HD) [15]: Captures boundary alignment accuracy by measuring the 95th percentile of distances between organ surface points from registered MRI and TRUS segmentations. Values are normalized to [0, 1] relative to unregistered cases. This reflects fidelity at organ boundaries, an important factor in

surgical guidance.

Standard Deviation of log-Jacobian Determinant (StDJD) [13, 14]: Evaluates deformation smoothness and physical plausibility. The Jacobian determinant describes local volumetric expansion or contraction; its logarithmic standard deviation measures irregularity in deformation. Smaller StDJD values indicate more stable and anatomically consistent warps.

Runtime: Represents the mean time per case to compute the deformation and warped image. Since µ-RegPro targets intraoperative applications such as MRI–TRUS fusion for biopsy and radiotherapy, real-time inference is a desirable property [15].

The final composite score is defined as:

$$\text{Score} = 0.2(\text{DSC}) + 0.1(\text{RDSC}) + 0.3(1 - \text{TRE}) + 0.1(1 - \text{RTRE}) + 0.1(1 - \text{RTs}) + 0.2(1 - 95\%\text{HD})$$

Chapter 3

Experimental Setup

3.1 Dataset Details

Very few MRI-TRUS datasets exist, and even fewer include anatomical labels suitable for deformable registration. Among those available, the µ-RegPro (micro-registration of the prostate) dataset was chosen for this work because of its simplicity, clear structure, and the availability of a working VoxelMorph baseline. While larger and more commonly used datasets such as NiftyReg exist, the µ-RegPro dataset offers a practical advantage by providing paired MRI-TRUS volumes, segmentation masks, and reference evaluation metrics that make it ideal for hypothesis-driven experimentation. The competition page reports that the best-performing registration method achieved an overall score of 0.862, showing that it is possible to perform accurate deformable registration on this dataset.

The µ-RegPro dataset is based on the SmartTarget Biopsy clinical trial and includes paired MRI and TRUS scans from 141 patients. Each case contains three-dimensional MRI and TRUS volumes, six anatomical segmentation masks, and landmark annotations identifying key prostate structures. The data are distributed as compressed npz files that can be easily loaded as NumPy arrays. After loading, all volumes were resized to 64×64×64 voxels to balance spatial resolution with memory and computational requirements. Each label volume had a shape of [1, 64, 64, 64, 6], where the first dimension is the batch channel and the last dimension corresponds to the six anatomical structures.

Figure 3.1 shows axial mid-slices of paired MRI and TRUS volumes with their corresponding label channels. Figure 3.2 shows sagittal mid-slices of paired MRI and TRUS volumes with their corresponding channel labels. From inspection, it can be seen that the data are

roughly aligned, suggesting that some form of coarse registration or standardization was applied before release. However, there are still visible local misalignments, particularly near the prostate boundaries, that justify the need for deformable registration. It was also observed that labels in channels 2–6 were very sparse across the dataset, often containing only a few voxels or being empty per class. This sparsity affects Dice-based evaluation since smaller regions are more sensitive to small spatial displacements. The first label channel, representing the prostate gland, was the most consistent and served as the primary structure used for quantitative evaluation and for guiding affine preprocessing. To show more instances of the aforementioned in the data, Appendix B shows similar axial and sagittal mid-slices for a few more pairs from the validation set.

To verify the degree of initial alignment, a simple affine check was performed. One MRI image was rotated by 90 degrees and registered to its corresponding TRUS image using ANTs affine alignment. The image was rotated back to its original orientation, confirming that the dataset already had a relatively consistent orientation and coarse alignment. This supported the decision to test models on mid-axial slices during early development to simplify debugging and visualization.

All images were intensity-normalized independently to ensure consistent value ranges across the dataset. No histogram equalization or domain adaptation was applied, as this could interfere with the natural intensity differences between modalities, which form part of the motivation for this study.

Only the training and validation splits of the dataset were available for use, consisting of 65 training and 8 validation cases. The test set remains private to the µ-RegPro challenge organizers. This limitation prevents direct comparison with leaderboard results but does not affect the ability to study the impact of affine preprocessing, weak supervision, or augmentation strategies within the available data.

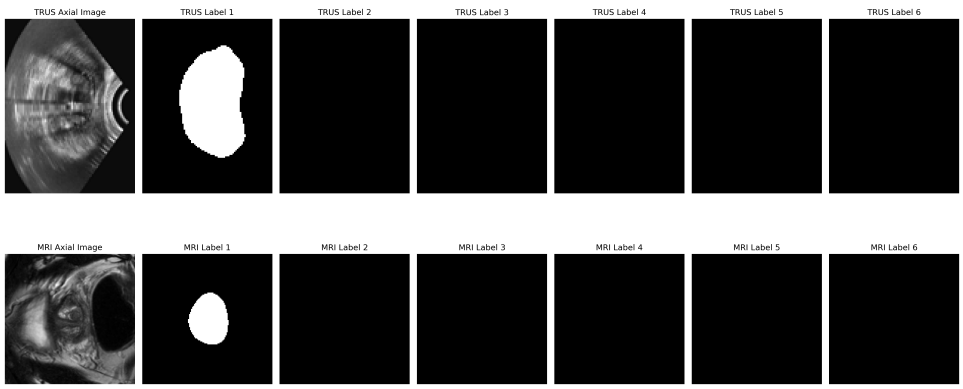


Figure 3.1: Axial midplane slices of paired MRI (bottom) and TRUS (top) volumes with their corresponding label channels

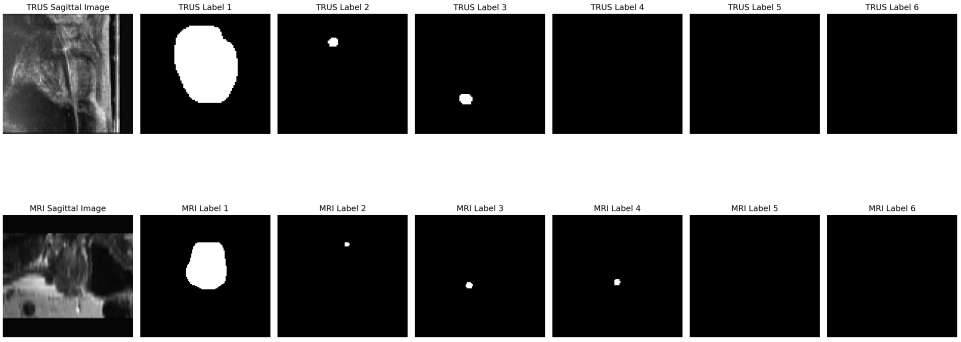


Figure 3.2: Sagittal midplane slices of paired MRI (bottom) and TRUS (top) volumes with their corresponding label channels

3.2 Computational Setup

All experiments were carried out on the university's high-performance computing (HPC) cluster. This platform was chosen because it provides up to 48 CPU cores with 8 GB of RAM per core and access to NVIDIA A100 GPUs. The high memory availability was important for handling the 3D medical imaging data. Although the cluster provides GPU acceleration, the dataset used in this project was relatively small, and GPU-based training did not significantly improve performance. In practice, training was faster on the CPU nodes, especially when the dataset was fully cached in the vast RAM provided by the HPC. Caching reduced the time per training epoch from about ten minutes to roughly one minute, which made repeated experiments and parameter tuning feasible. The HPC system also made it possible to queue and run multiple jobs at once, which allowed for broad hyperparameter exploration without manual monitoring. Each run was configured to automatically save its model checkpoints, logs, and evaluation metrics, enabling structured comparison between experiments. This setup provided a flexible and reliable environment for development and testing. The large available memory and multicore architecture ensured that data handling never became a bottleneck, and the ability to run several experiments in parallel helped accelerate the overall workflow.

The baseline model used is written in Python, leveraging Keras and Tensorflow to implement the major machine learning functionality. Many other python libraries are used for data processing and visualization. These tools were used further to modify the baseline as needed.

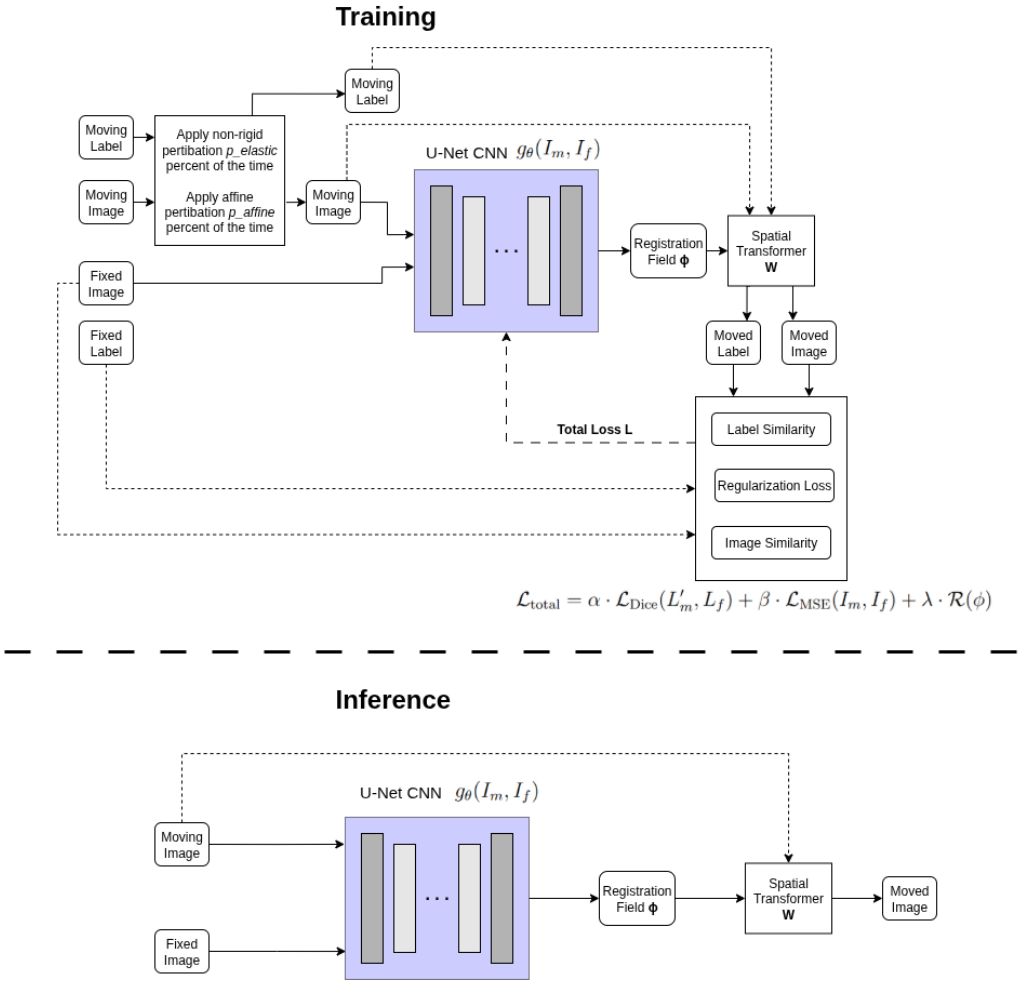


Figure 3.3: Overview of the training and inference procedure of the modified baseline model. The network predicts a dense displacement field ϕ to warp the moving image I_m towards the fixed image I_f. During training, both images and labels are used, while during inference only the image pair is required.

3.3 Baseline Model Verification

Figure 3.3 shows a schematic overview of the training and inference processes of the modified baseline model. Two versions of the model were implemented: a two-dimensional (2D) version for rapid prototyping and debugging, and a three-dimensional (3D) version used for the main hypothesis testing. Both versions share the same overall architecture and differ only in dimensionality and output channel count. For a more detailed visual of the model, Appendix C shows a more detailed diagram of the modified baseline model generated from Keras.

The model was based on the μ-RegPro VoxelMorph baseline and extended to allow control over the type of supervision and regularization. The loss function used for training is defined as:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{Dice}}(L'_m, L_f) + \beta \cdot \mathcal{L}_{\text{MSE}}(I_m, I_f) + \lambda \cdot \mathcal{R}(\phi) \tag{3.1}$$

The μ-RegPro VoxelMorph baseline did not come with functioning weak supervision, and had to be added by modifying the inputs and outputs of the model and data loaders. The outputs were then used to create the required loss function using the Voxelmorph provided dice and NCC or MSE loss. Keras does not allow inputs to be outputted from the model if they are not actually used in the model. The images are used to generate the dense deformation field, but since the labels do not actually form part of that process directly until the calculation of the dice loss, they had to be passed through a dummy function to prevent Keras from throwing an error.

where L'_m is the warped moving label, L_f the fixed label, I_m and I_f are the moving and fixed images, and ϕ is the dense deformation field predicted by the network. The coefficients α, β, and λ control the relative contributions of the Dice loss, intensity-based MSE or NCC loss, and deformation regularization term respectively. By setting α and β to zero and one respectively, unsupervised learning is implemented. By setting α and β to one and zero respectively, weakly supervised label driven learning is implemented. The regularization weight λ controls the smoothness penalty on the displacement field, with $\lambda = 0.05$ and the learning rate set to 0.0005 being popular in the literature.

The data loader provided with the baseline implementation was extended to support random data augmentation. Each moving image and label could be perturbed non-rigidly with probability p_{elastic} or affinely with probability p_{affine}. This enabled systematic

evaluation of augmentation strategies. Additionally, the data loader was modified to read from a pre-loaded python dictionary of the data rather than from secondary storage. The data loader provides a TRUS-MRI image and random label label pair when called during training. Since some labels are empty, the dice loss is pessimistic during training, hence curves showing the loss over training in this thesis do not reflect the actual training dice score achieved, but are still included to visualize the performance of the model over training. After training, the empty labels are ignored, and the true average dice score for the validation set is reported throughout this thesis.

During training, both the image and label volumes were required as inputs, while during inference only the image pair was used. After each epoch, the model computed the average Dice score over the validation set, which was used as a criterion for early stopping. Early stopping was implemented by monitoring the validation Dice history and halting training once no improvement beyond a small threshold was observed within a specified patience window of 5 epochs. This prevented overfitting and unnecessary computation. The model weights from the epoch achieving the highest validation Dice were automatically saved for later use.

To confirm that this mechanism functioned correctly, both the total loss and validation Dice were visualized after each epoch along with example visualizations of the fixed image, warped moving image, and corresponding labels. Early stopping consistently occurred shortly after overfitting became apparent. Variability between training runs using identical hyperparameters was low, typically less than 0.001 in the final Dice score, and therefore repeated runs were not deemed necessary.

Figure 3.4 shows the convolutional U-Net architecture used to implement $g_\theta(I_f, I_m)$. Each encoder block applies two convolutional layers followed by max pooling with a stride of two, halving the spatial resolution. The decoder mirrors this process using upsampling by a factor of two after each block. The number of feature channels increases toward the bottleneck and decreases thereafter. Internal residual connections were included within convolutional blocks to stabilize training by allowing partial feature reuse.

A wrapper function was developed to automate experimental configuration. It allowed adjustment of α, β, λ, p_{elastic}, and p_{affine}, and specified the output CSV file for logging evaluation metrics. Each experiment (e.g., "experiment_A") wrote its results to a unique file containing the μ-RegPro evaluation metrics computed for both the original and affinely preprocessed data. The first two rows of each file contained baseline metrics and the metrics computed for the original data for comparison. The metrics for the original data was used to normalize the metrics such that the final score is between zero and one.

Since the prostate gland label was the most consistent and least sparse among the six anatomical labels, Dice and Robust Dice (RDSC) scores were also recorded specifically for this label. These scores were consistently higher than those computed across all labels, likely due to the small size and sparsity of non-gland structures, which makes them more sensitive to downsampling and voxel misalignments.

To validate that the model performed deformable registration correctly, a controlled experiment was conducted using synthetic ultrasound-like images derived from MRI data. A function, denoted mri_to_us_like($\cdot$), was used to transform MRI volumes by applying anisotropic Gaussian blurring, multiplicative Rayleigh speckle noise, and logarithmic compression. The resulting synthetic ultrasound images preserved anatomical structure but exhibited realistic speckle and lower contrast. These images were then warped by a sinusoidal deformation field and used as the fixed inputs, while the original MRI images served as the moving inputs. The deformation field is shown in Figure 3.5.

Figure 3.6 shows the validation results after one training epoch for the synthetic MRI–ultrasound experiment. The warped MRI image visibly follows the sinusoidal deformation, and the predicted displacement field reproduces the sinusoidal pattern. This confirmed that the model could learn complex non-linear transformations and perform true deformable registration. Verifying this behavior was critical before proceeding to the main experiments, as it established confidence that the baseline implementation and training pipeline were functioning as intended.

3.4 Affine Preprocessing Experiments

For the same reasons that multimodal deformable registration between TRUS and MRI images is challenging, one cannot simply affinely register TRUS-MRI image pairs directly using tools such as ANTs. In which case, the images would be affinely registered, and their corresponding labels would be warped with the same affine transform. This is usually done in unimodal contexts, or simpler co-modal contexts like CT-MRI. ANTs relies on mutual information (MI) to align intensity distributions, which can fail in multimodal contexts due to the absence of consistent intensity correspondences between modalities. When ANTs was applied directly to TRUS-MRI image pairs, the resulting Dice scores were low, confirming the inadequacy of intensity-based affine registration for this task.

To overcome this, a label-driven affine registration approach was implemented. Instead of registering the TRUS and MRI images themselves, the gland labels were registered using

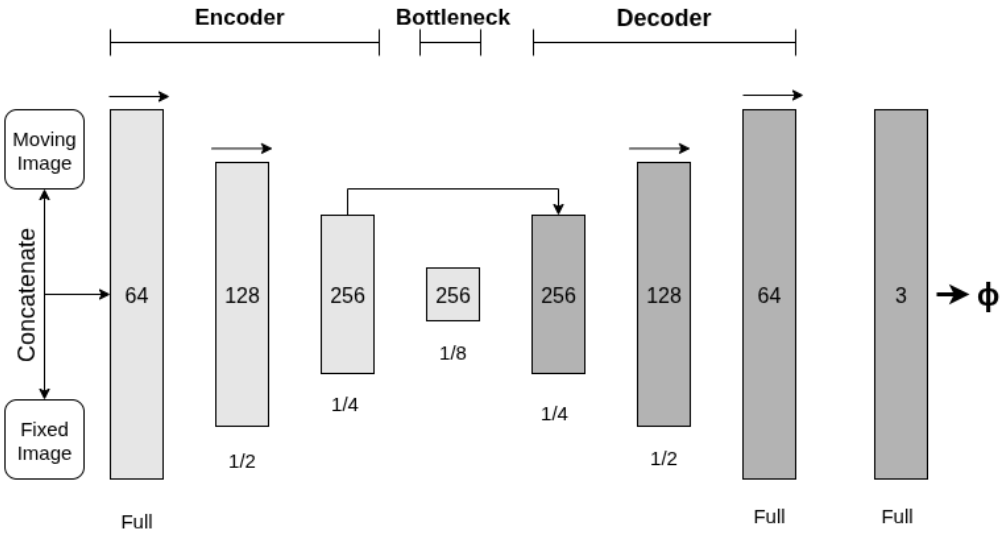


Figure 3.4: Convolutional U-Net architecture implementing $g_\theta(I_f, I_m)$. Skip connections preserve spatial detail between encoder and decoder layers, while internal residual feedback improves gradient flow. The network outputs a dense displacement field ϕ of two or three channels, depending on the registration dimensionality.

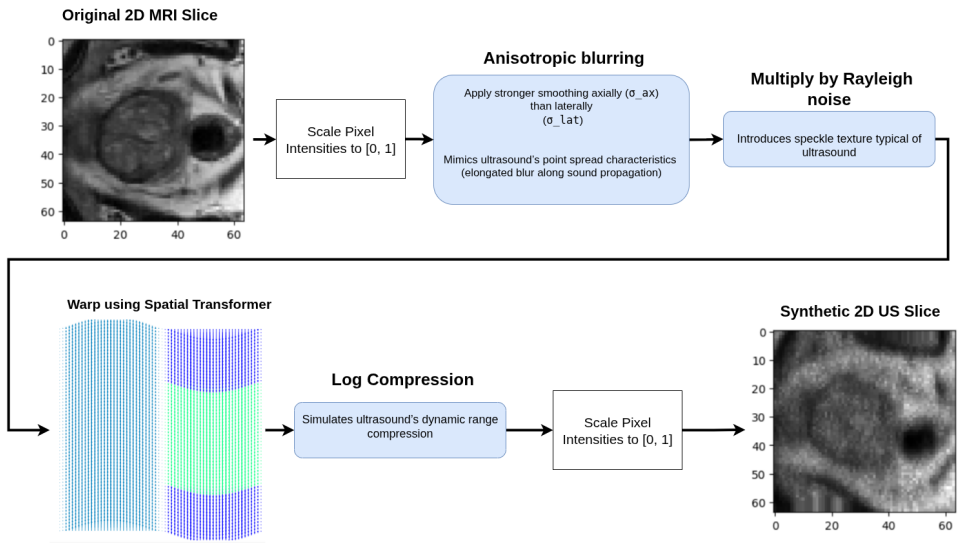


Figure 3.5: Schematic illustration of synthetic ultrasound generation from MRI. The MRI volume undergoes anisotropic Gaussian blurring, multiplicative Rayleigh speckle addition, and logarithmic compression to approximate ultrasound-like characteristics. The resulting synthetic ultrasound is then warped by a sinusoidal field to serve as a controlled multimodal registration test case.

ANTs' affine transform, and the resulting affine matrix was applied to the corresponding images and other labels. In this way, the affine alignment was entirely driven by label geometry rather than image intensity, effectively removing modality-related bias. The process and an example case are shown in Figure 3.7, which depicts the moving and fixed gland and cyst labels alongside the TRUS image from a validation pair. Qualitatively, the gland label is scaled along the x and y axes and translated, with little to no rotation or diagonal stretching. This observation is consistent with the hypothesis that the dataset is already roughly pre-aligned.

Even though the affine transformation was computed solely from the gland, the alignment of the cyst and other small structures also improved, as seen qualitatively in Figure 3.7. Quantitatively, this is reflected in the CSV snippet shown in Figure 3.8, where the Dice score for the gland label increased from 0.42 (without affine preprocessing) to 0.87 after ANTs affine registration. The mean Dice score across all labels improved from 0.29 to 0.35, and similar improvements were seen in the other µ-RegPro metrics. Appendix D shows additional slices from the validation set for labels 0 and 1.

The large discrepancy between the gland Dice score and the mean Dice score was investigated. Figure 3.9 shows two pairs of synthetic binary circles: one pair with a radius of 10 pixels and an offset of 10 pixels between centers, and another pair with a radius of 5 pixels and the same offset. The Dice score dropped from 0.46 to 0.012 as the circle size decreased, despite identical misalignment. This demonstrates that the Dice coefficient is highly sensitive to object size and that small labels contribute disproportionately low Dice scores even when alignment quality is visually good. This helps explain the limited improvement in the overall mean Dice score despite the strong improvement in gland alignment.

To further improve the average Dice score, all label masks were fused into a single composite volume and registered as one entity. However, this approach consistently yielded poorer alignment results, likely because fusing the labels obscured the geometric clarity of the gland, which dominates the label set and serves as the most reliable alignment target.

To remain faithful to the approach of Zeng *et al.* [8], a convolutional neural network (CNN) was also implemented for affine registration. The model was directly based on the design described in their paper, which states:

> "Magnetic resonance (MR) and TRUS labels were designated as moving and fixed labels, respectively, since we aim to use the warped preoperative MRI, which matches the intra-operative TRUS, as image guidance during

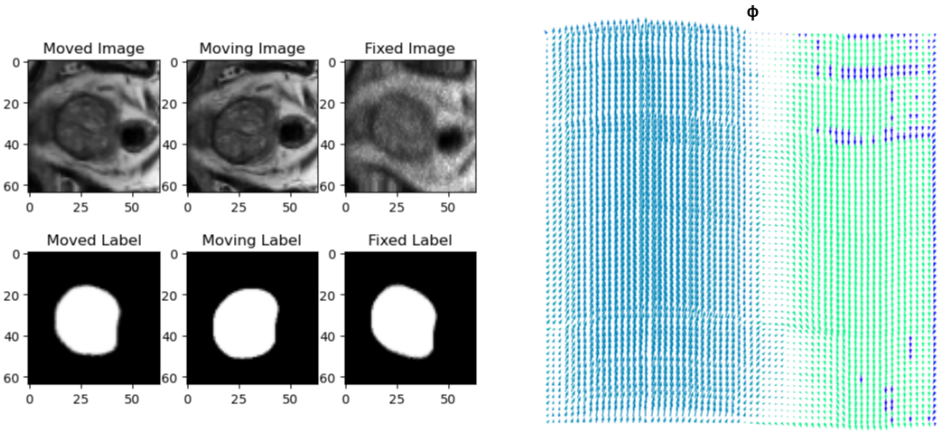


Figure 3.6: Validation results after one training epoch using synthetic MRI–ultrasound pairs. The model successfully learned to produce the sinusoidal deformation, producing a corresponding sinusoidal displacement field, confirming correct deformable registration behavior.

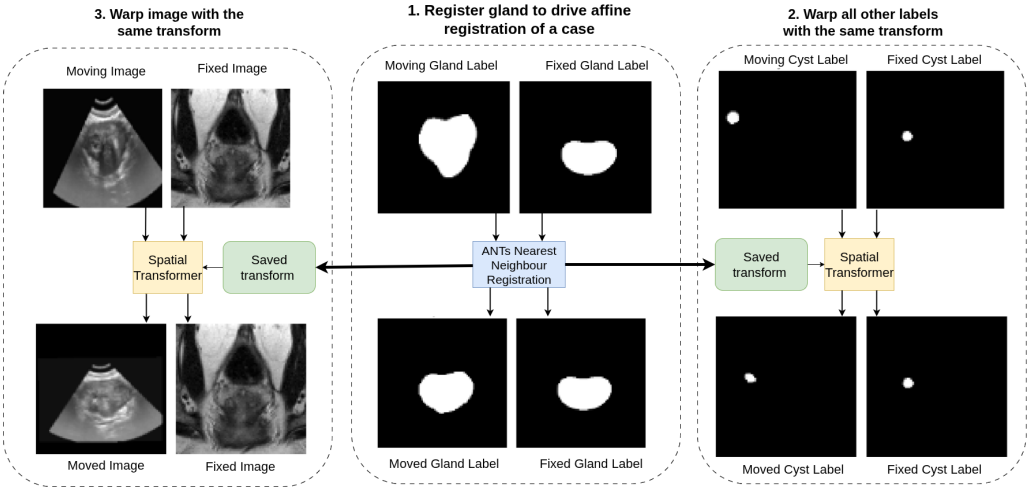


Figure 3.7: ANTs affine registration process and example results showing gland-driven affine alignment.

Model	Trial	DSC	RDSC	HD95	TRE	RTRE	RTs	StDJD	DSC Gland	RDSC Gland	Final Score
Nothing	nan	0.29	0.33	8.77	6.75	6.6	4.35	nan	0.42	0.49	0.09
ANTS	nan	0.35	0.38	2.35	2.46	2.22	1.11	nan	0.87	0.88	0.58

Figure 3.8: ANTs affine registration process and example results showing gland-driven affine alignment.

> the procedure. According to research (van de Ven *et al.*, 2015; Wang *et al.*, 2016), it is very helpful to set a good rigid or affine registration as an accurate initialization prior to subsequent nonrigid registration. Therefore, we designed a 2D CNN which takes 3D image as inputs to predict 12 affine transformation parameters to warp the original MR images and labels, as an initialization step of nonrigid registration. The automatic affine registration aimed to optimize the Dice similarity coefficient. . . "

The implemented model follows the same architecture shown in Figure 3.10, consisting of five convolutional layers with 128, 256, 512, 1024, and 12 filters, respectively, each followed by batch normalization and ReLU activation. These are followed by a flatten layer and a fully connected layer producing 12 outputs corresponding to the affine transformation parameters.

Initially, the model was trained using random label types from the dataset in the hopes that it would learn a general affine alignment strategy rather than one dominated by the gland geometry. However, training failed to converge, suggesting that the smaller labels do not contain sufficient or consistent geometric information to guide affine registration. Training and validation were therefore restricted to the gland label only.

Early training results were unstable. As shown in Figure 3.11, the model frequently produced extreme rotations and diagonal scalings that bore little resemblance to valid affine transformations. The affine matrices corresponding to these cases had large off-diagonal elements (highlighted in red), representing cross-axis scaling and shearing, which should ideally be close to zero for this dataset. Initially, these off-diagonal terms were penalized as part of the loss function to constrain the solution space, but this had little effect on training stability or convergence.

The only modification that meaningfully stabilized training was reducing the learning rate from 5×10^{-4} to 5×10^{-5}. With this smaller learning rate, the model converged smoothly, the loss decreased steadily (as shown in Figure 3.12), and the validation Dice score stabilized around 0.75 - an impressive result for affine preprocessing. The corresponding affine matrices had near-zero off-diagonal elements, and visual inspection of the moved labels confirmed that the CNN had learned to apply the same types of affine transformations identified earlier with ANTs: x and y scaling and translation without significant rotation.

Figure 3.13 shows an example validation case for the best performing epoch of the CNN model. The moved gland label has been scaled and translated in the x and y directions,

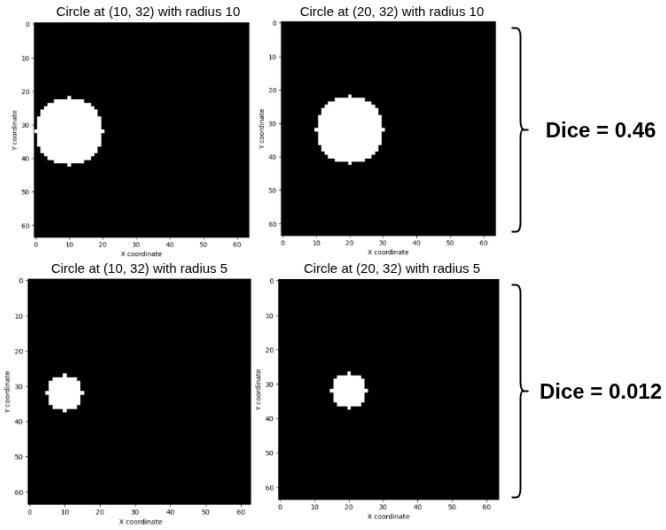


Figure 3.9: Synthetic Dice experiment showing sensitivity of Dice to small label size.

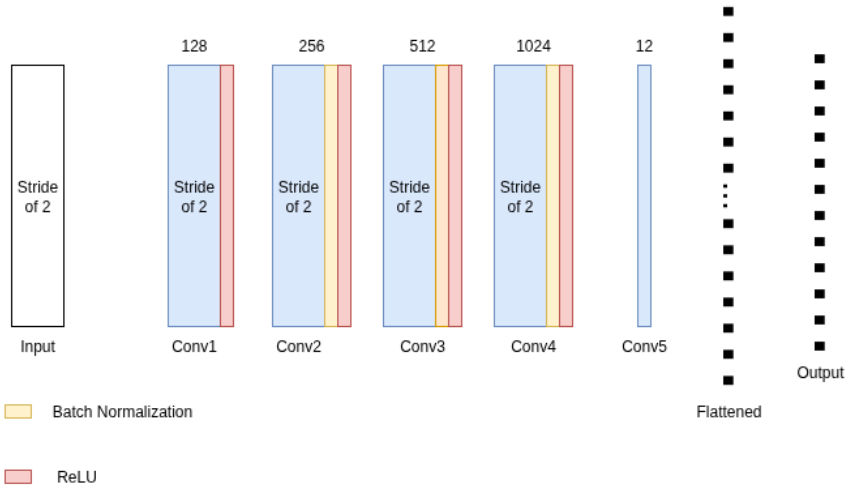


Figure 3.10: Affine CNN architecture used for label-driven affine registration based on [8].

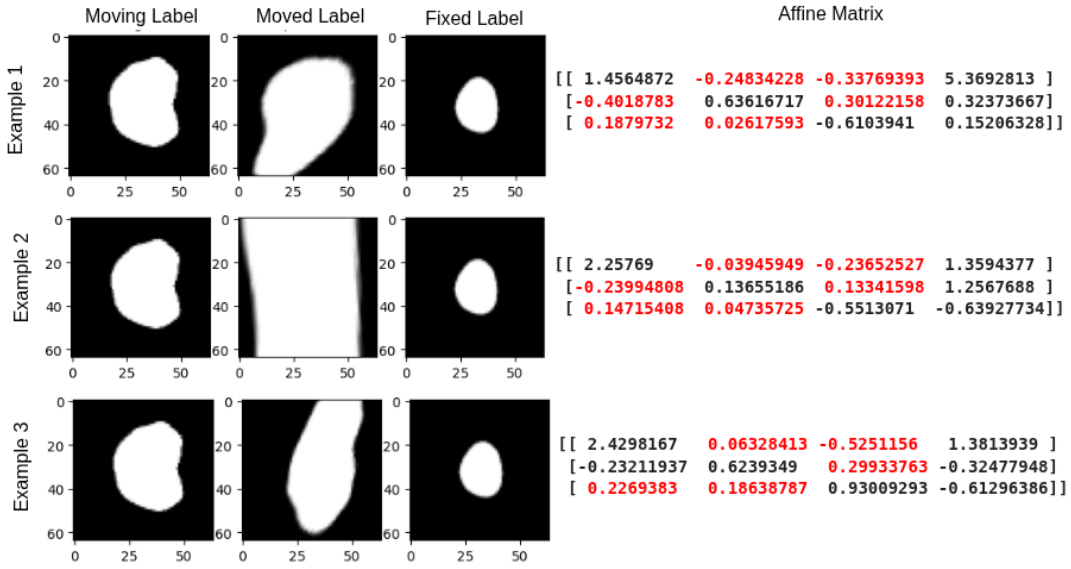


Figure 3.11: Example training failures of the affine CNN showing excessive rotation and scaling due to unstable learning.

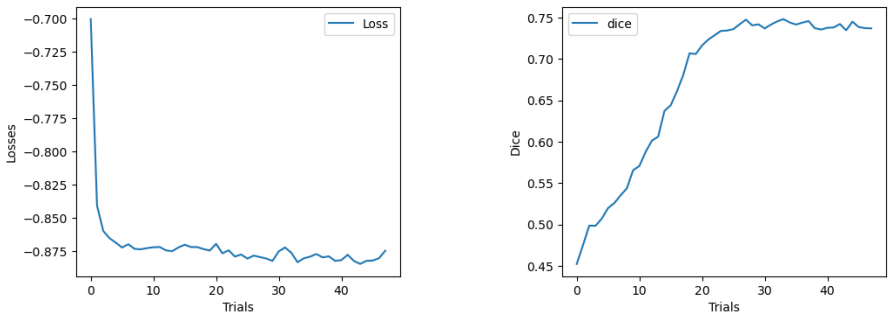


Figure 3.12: Loss and validation Dice curves for the affine CNN after reducing the learning rate to 5×10^{-5}.

consistent with the qualitative behavior observed for ANTs. The off-diagonal matrix elements are small, confirming that with the correct learning rate the model learned to avoid unrealistic transformations.

Although the CNN achieved a respectable Dice score and qualitatively similar transformations, it did not outperform ANTs, which remained more consistent and robust across the dataset. Therefore, ANTs was adopted as the standard affine preprocessing tool for all subsequent experiments in this study.

3.5 Affine and Non-Rigid Data Augmentation

To improve model robustness and investigate the effects of different augmentation strategies, affine and elastic (non-rigid) transformations were implemented directly in the data loading pipeline. These augmentations were applied to both the moving and fixed image-label pairs during training with adjustable probabilities, controlled by the variables `p_affine` and `p_elastic`.

The *affine augmentation* applied small, random in-plane rotations to simulate minor misalignments between MRI and TRUS scans. Specifically, each pair was rotated by a random angle uniformly sampled between $-5°$ and $+5°$ in the axial (xy) plane. When label maps were present, the same transformation was applied using nearest-neighbour interpolation to preserve discrete label values. This ensured spatial consistency between images and their corresponding segmentation labels while introducing realistic geometric variability in the training data. The operation can be viewed as a lightweight, label-consistent simulation of patient repositioning or probe angle variation during TRUS acquisition.

The *elastic deformation* served as a simple model of local tissue distortion. A smooth random displacement field was generated by filtering Gaussian noise with a kernel of standard deviation $\sigma = 20$ voxels and scaling the resulting displacements by $\alpha = 2$. This field was then added to the voxel grid, and the volume was resampled at the displaced coordinates using trilinear interpolation for images and nearest-neighbour interpolation for labels. This procedure produced smooth, spatially varying deformations that mimic anatomical variability and probe pressure effects, without introducing discontinuities or folding.

Both augmentation types were implemented in 3D to match the volumetric nature of the

data. The transformations were applied probabilistically before each training iteration, allowing each batch to present a slightly perturbed version of the data. This stochastic augmentation strategy aimed to expose the model to a wide range of plausible affine and non-rigid variations, improving generalisation and providing a mechanism for systematically studying the effects of affine versus non-rigid perturbations in later experiments.

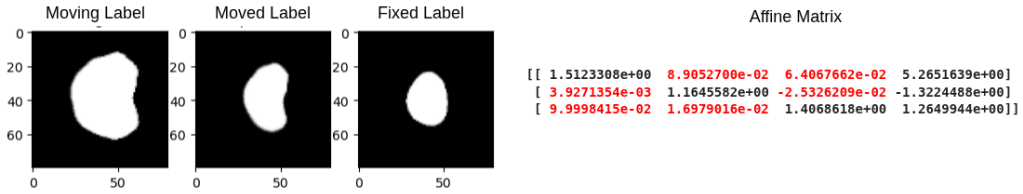


Figure 3.13: Example of a correctly converged affine CNN registration showing alignment similar to ANTs-based results.

Chapter 4

Hypothesis Testing Experiments

In theory, one could run all permutations of the model discussed, that is, all combinations of weakly supervised versus unsupervised training, affine preprocessing versus no affine preprocessing, no augmentation versus non-rigid augmentation versus combined non-rigid and affine augmentation, and various learning rates and regularization weights, to obtain a fully comprehensive understanding of what works best. However, such an exhaustive search is not feasible within the time constraints of this study. Instead, this section describes how the hypotheses were investigated hierarchically, validating more fundamental model characteristics earlier while continuously testing affine versus non-affine preprocessing. The goal was to identify which architectural and training choices most strongly influence registration behaviour before testing higher-level improvements such as augmentation. All experiments in this section were performed using the 3D version of the model.

4.1 Unsupervised Investigation

The first experiment aimed to assess the capacity of purely intensity-driven registration to learn meaningful spatial correspondences between MRI and TRUS images without any anatomical supervision. Based on the literature, it was anticipated that unsupervised models relying solely on similarity metrics such as mean squared error (MSE) or normalized cross-correlation (NCC) would struggle to capture multimodal correspondence due to the fundamentally different image formation physics of MRI and TRUS. Nevertheless, this experiment was conducted to qualitatively confirm these shortcomings and establish a reference point for subsequent weakly supervised experiments.

4.1.1 Experimental Procedure

Two versions of the unsupervised model were trained, one using MSE and the other NCC as the similarity metric. Both models employed the standard regularization weight $\lambda = 0.05$ and learning rate 5×10^{-4}, consistent with literature conventions. The total loss function was defined as

$$L_{total} = \beta \, \mathcal{L}_{sim}(I_m, I_f) + \lambda \, \mathcal{R}(\phi) \tag{4.1}$$

where $\mathcal{L}_{sim}$ represents either MSE or NCC, $\mathcal{R}(\phi)$ is the regularization term encouraging smoothness of the deformation field ϕ, and β is set to 1. Each model was trained both with and without affine preprocessing of the dataset to determine whether affine alignment improved convergence or semantic accuracy.

Instead of using early stopping, the models were allowed to train well past convergence (200 epochs), since premature termination could conceal how intensity-based metrics deform the image as training progresses. During training, both the warped images and corresponding labels were saved at intervals to visually evaluate the anatomical plausibility of the learned deformations.

4.1.2 Results and Discussion

Figure 4.1 shows the qualitative results for models trained with NCC and MSE, both with and without affine preprocessing. After 200 epochs, all models had converged in terms of both total and regularization losses, but their qualitative performance diverged. The warped TRUS images exhibited visually convincing alignment with the fixed MRI images; however, inspection of the warped gland labels revealed clear semantic errors. The glands were frequently distorted or displaced into regions of non-gland tissue, indicating that the models were matching image intensities rather than anatomical structures.

MSE-based registration produced coarse, blurred deformations with poor boundary alignment, consistent with its known sensitivity to differences in image contrast. NCC-based registration yielded smoother, higher-fidelity warps, but these too were anatomically meaningless. The model effectively "painted" bright areas of the TRUS image onto bright regions of the MRI image, regardless of true gland boundaries.

Affine preprocessing was hypothesized to improve results by removing global misalignment; however, no measurable benefit was observed. Both preprocessed and non-preprocessed datasets led to similarly incorrect mappings, confirming that intensity-based similarity alone is inadequate for this multimodal task.

Figure 4.2 shows the loss and validation Dice curves for the respective cases. The Dice scores plateaued early and remained extremely low, even after filtering out empty labels. The total and regularization losses stabilized well within 200 epochs, confirming that poor anatomical alignment was not caused by unstable optimization, but rather by the inherent inability of the loss function to encode cross-modality correspondence.

In summary, this experiment confirmed that purely intensity-driven similarity metrics are inadequate for multimodal deformable registration between MRI and TRUS images. Even though NCC performed marginally better than MSE, both methods failed to establish anatomically meaningful correspondences, regardless of affine preprocessing. These results strongly justify the use of anatomical labels to guide registration and motivate the transition toward weakly supervised approaches in the subsequent experiments.

4.2 Weakly Supervised Investigation

The previous experiment demonstrated that intensity-based similarity metrics cannot effectively drive deformable registration between MRI and TRUS images. This section investigates whether weak supervision, using anatomical labels to guide training, can achieve meaningful registration. Assuming that weak supervision performs well, the best-performing hyperparameters will be used in the following augmentation experiments. All experiments were conducted on both affinely preprocessed and non-affinely preprocessed datasets to further evaluate the necessity of affine alignment for this particular dataset. Although the data appears roughly aligned, the degree to which residual affine misalignment impacts performance remains uncertain.

4.2.1 Experimental Procedure

The weakly supervised model used the Dice loss with a regularization term, expressed as

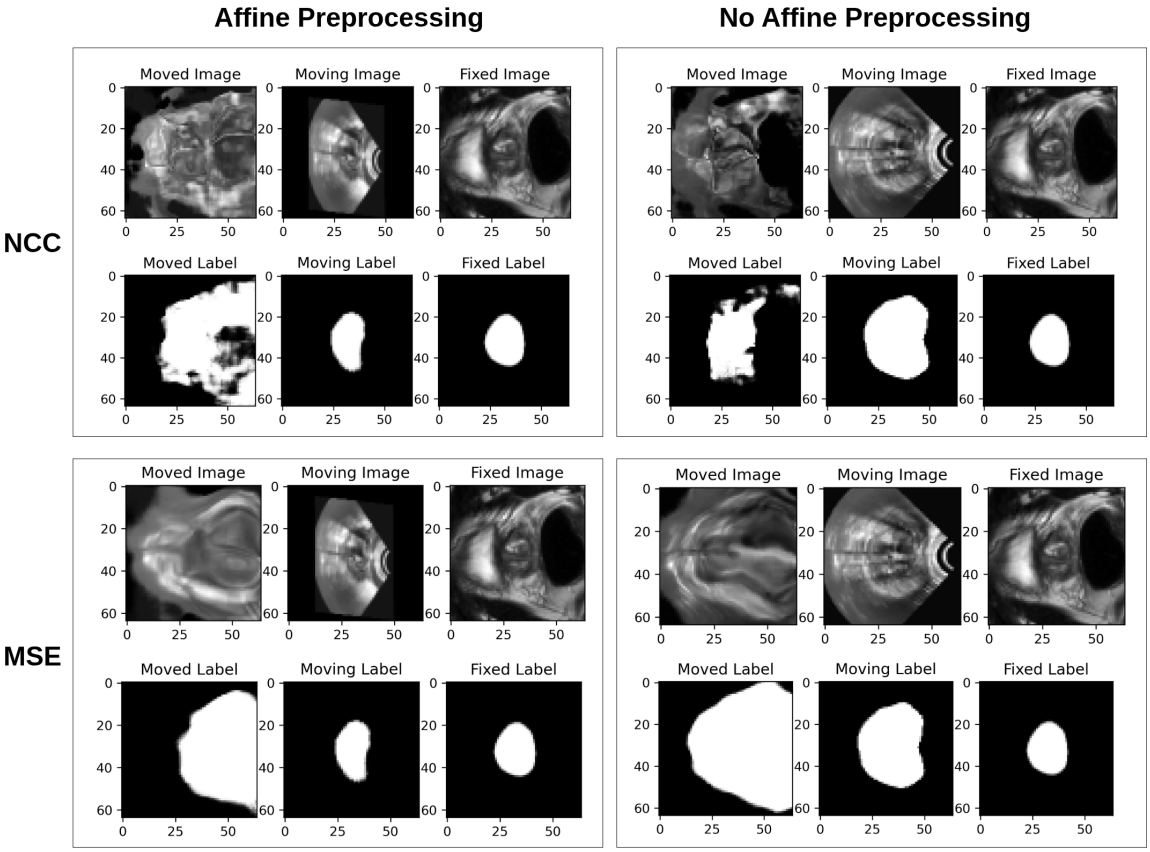


Figure 4.1: Qualitative comparison of unsupervised registration using MSE and NCC, with and without affine preprocessing. While the warped TRUS images visually resemble the MRI images, the corresponding gland labels reveal severe anatomical misalignment.

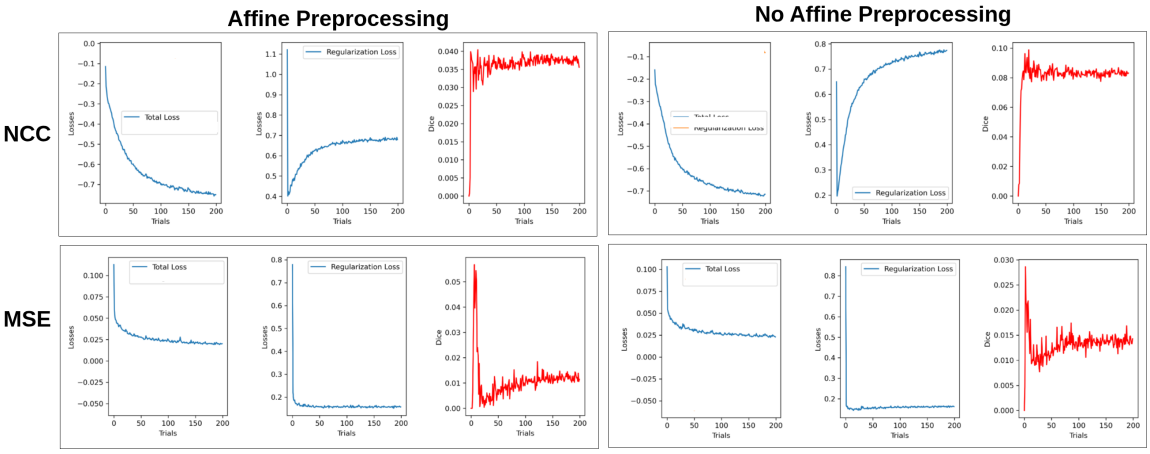


Figure 4.2: Training curves showing total loss, regularization loss, and validation Dice score for the unsupervised models. Both NCC and MSE converge rapidly, but the validation Dice remains significantly below that of the unregistered data.

$$L_{total} = \alpha \, \mathcal{L}_{Dice}(L'_m, L_f) + \lambda \, \mathcal{R}(\phi) \tag{4.2}$$

where α and β are weighting factors controlling the contribution of label-based and intensity-based terms respectively, and λ is the regularization coefficient controlling deformation smoothness. In all cases, α was set to 1.

To identify the best-performing configuration, models were trained using early stopping with a range of learning rates $\{0.01, 0.001, 0.0005, 0.00005\}$ and regularization weights $\{0.01, 0.05, 0.1, 0.2, 0.5\}$. After each run, quantitative metrics were computed and stored in a CSV file for analysis. The metrics included Dice similarity coefficient (DSC), relative DSC (RDSC), Hausdorff distance (HD95), target registration error (TRE), relative TRE (RTRE), registration time (RT), and standard deviation of the Jacobian determinant (StDJD), consistent with the µ-RegPro evaluation protocol.

4.2.2 Results and Discussion

Figure 4.3 shows an excerpt of the CSV file summarizing the evaluation metrics for the weakly supervised model across the different hyperparameter combinations, with and without affine preprocessing. The final scores for the affinely preprocessed models are markedly higher than for those trained without affine preprocessing. Looking at all of the scores for the models with no affine preprocessing, it can be seen that they did far worse than ANTs on all cases, and for all metrics. This suggests that even though the dataset appears visually aligned, residual affine misalignment still hinders the performance of the dense deformable model. This may suggest that the dense deformable model cannot be expected to do even small affine alignment, which would make affine augmentation not work.

Notably, the best-performing trials occurred relatively early, with validation Dice scores peaking around epoch 22, indicating that the model may not have had enough data to fully generalize. This further motivates the introduction of augmentation in later experiments. The gland DSC and RDSC score slightly decreased after deformable registration for the preprocessed cases, but all other metrics improved in some instances, including the best performing instance that had: preprocessing, a learning rate of 0.0005, and a lambda value of 0.1. The higher StDJD values observed in models without affine preprocessing suggest that these models required larger displacement vectors to compensate for residual global misalignment, reinforcing the need for affine preprocessing in this dataset.

Figure 4.4 shows qualitative results for the best-performing weakly supervised model, trained with a learning rate of 0.0005 and regularization weight of 0.1. The model successfully preserves the gland label shape and location, producing only subtle deformations consistent with fine local corrections rather than large affine transformations. This is consistent with the expectation that the dense deformable field primarily refines alignment rather than performing global registration.

Figure 4.5 presents the training curves for the same model, showing total loss, regularization loss, and validation Dice score. The validation Dice converges more gradually than in the unsupervised case and achieves approximately double the maximum Dice value, even though the displayed curves are pessimistic due to the inclusion of empty masks during training. Figure 4.6 shows the same curves for the model trained without affine preprocessing. The loss curves are considerably noisier and exhibit weaker convergence, further supporting the conclusion that the model struggled to learn appropriate transformations without affine alignment.

Overall, the results demonstrate that weak supervision significantly improves registration performance compared to purely unsupervised approaches. Moreover, affine preprocessing remains beneficial even when supervision is introduced, suggesting that this dataset still contains residual global misalignments that the deformable model cannot correct within its regularization constraints. These findings validate the use of weak supervision for multimodal TRUS-MRI registration and justify retaining affine preprocessing in subsequent experiments.

Figure 4.7 shows axial mid-slices of the dense displacement fields for the label-driven model's best-performing case and for the unsupervised models after their regularization loss converged (at least epoch 50, as seen in Figure 4.2). While vector fields themselves can be difficult to interpret quantitatively, the qualitative differences between models support earlier findings. The weakly supervised model produced a displacement field with much smaller vectors, fine enough that the checkerboard pattern from the convolutional kernels is visible. These small, localized displacements are consistent with the expected minor adjustments of deformable registration following affine alignment. In contrast, the unsupervised models produced smooth but large-magnitude displacements that are not anatomically meaningful, confirming that intensity-based similarity metrics alone tend to produce global misregistrations.

Model	Trial	DSC	RDSC	HD95	TRE	RTRE	RTs	StDJD	DSC Gland	RDSC Gland	Final Score
Nothing	nan	0.29	0.33	8.77	6.75	6.6	4.35	nan	0.42	0.49	0.09
ANTS	nan	0.35	0.38	2.35	2.46	2.22	1.11	nan	0.87	0.88	0.58
Preprocessed, lr=0.01, lambda=0.01	11	0.31	0.35	2.31	2.0	1.89	1.05	0.73	0.84	0.85	0.6
Preprocessed, lr=0.01, lambda=0.05	19	0.36	0.39	2.03	1.86	1.62	0.71	0.16	0.86	0.86	0.64
Preprocessed, lr=0.01, lambda=0.1	15	0.34	0.37	2.19	1.93	1.76	0.8	0.23	0.85	0.85	0.62
Preprocessed, lr=0.01, lambda=0.2	22	0.32	0.35	2.1	1.77	1.56	0.88	0.13	0.85	0.86	0.63
Preprocessed, lr=0.01, lambda=0.5	7	0.3	0.34	2.54	1.98	1.92	1.05	0.41	0.77	0.8	0.59
Preprocessed, lr=0.001, lambda=0.01	5	0.36	0.39	2.15	1.86	1.69	0.77	0.2	0.84	0.85	0.64
Preprocessed, lr=0.001, lambda=0.05	7	0.36	0.39	2.01	1.79	1.6	0.72	0.11	0.85	0.86	0.64
Preprocessed, lr=0.001, lambda=0.1	11	0.33	0.37	2.06	1.88	1.66	0.92	0.24	0.84	0.86	0.63
Preprocessed, lr=0.001, lambda=0.2	1	0.34	0.38	2.13	1.95	1.77	0.87	0.1	0.83	0.84	0.62
Preprocessed, lr=0.001, lambda=0.5	7	0.34	0.37	2.09	1.9	1.64	0.82	0.09	0.86	0.86	0.63
Preprocessed, lr=0.0005, lambda=0.01	11	0.34	0.38	2.2	1.93	1.7	0.8	0.51	0.85	0.85	0.62
Preprocessed. lr=0.0005. lambda=0.05	9	0.35	0.38	2.04	1.85	1.65	0.7	0.15	0.86	0.86	0.64
Preprocessed, lr=0.0005, lambda=0.1	12	0.36	0.4	2.03	1.8	1.57	0.62	0.16	0.85	0.86	0.65
Preprocessed, lr=0.0005, lambda=0.2	3	0.33	0.36	2.15	1.86	1.63	0.93	0.1	0.86	0.86	0.62
Preprocessed, lr=0.0005, lambda=0.5	8	0.35	0.39	2.05	1.77	1.58	0.74	0.08	0.85	0.86	0.64
Preprocessed, lr=5e-05, lambda=0.01	0	0.31	0.34	2.24	1.99	1.79	0.9	0.15	0.85	0.85	0.61
Preprocessed, lr=5e-05, lambda=0.05	16	0.36	0.39	2.04	1.8	1.57	0.78	0.33	0.85	0.85	0.64
Preprocessed, lr=5e-05, lambda=0.1	0	0.28	0.31	2.35	2.11	1.83	1.09	0.13	0.83	0.84	0.59
Preprocessed, lr=5e-05, lambda=0.2	11	0.33	0.35	2.15	1.87	1.6	0.78	0.2	0.85	0.85	0.63
Preprocessed, lr=5e-05, lambda=0.5	5	0.33	0.36	2.1	1.86	1.64	0.83	0.15	0.86	0.87	0.63
Not Preprocessed, lr=0.01, lambda=0.01	5	0.15	0.18	7.89	6.24	6.0	4.26	2.16	0.37	0.45	0.1
Not Preprocessed, lr=0.01, lambda=0.05	8	0.13	0.16	5.97	5.15	4.76	3.24	1.45	0.49	0.56	0.23
Not Preprocessed, lr=0.01, lambda=0.1	16	0.17	0.19	5.93	4.58	4.43	2.81	0.4	0.51	0.57	0.28
Not Preprocessed, lr=0.01, lambda=0.2	8	0.14	0.16	5.71	4.45	3.8	2.98	0.34	0.53	0.61	0.29
Not Preprocessed, lr=0.01, lambda=0.5	10	0.13	0.14	6.27	4.46	4.34	3.35	0.28	0.5	0.54	0.25
Not Preprocessed, lr=0.001, lambda=0.01	3	0.17	0.2	4.32	3.49	3.33	2.25	1.29	0.59	0.69	0.4
Not Preprocessed, lr=0.001, lambda=0.05	2	0.17	0.19	5.85	4.33	3.76	2.85	0.49	0.56	0.58	0.3
Not Preprocessed, lr=0.001, lambda=0.1	5	0.15	0.16	5.8	4.59	4.48	3.24	0.3	0.51	0.56	0.26
Not Preprocessed, lr=0.001, lambda=0.2	15	0.23	0.28	3.97	3.47	2.82	1.95	0.35	0.59	0.7	0.44
Not Preprocessed, lr=0.001, lambda=0.5	11	0.17	0.19	5.02	4.01	3.5	2.94	0.35	0.53	0.58	0.34
Not Preprocessed, lr=0.0005, lambda=0.01	1	0.14	0.16	6.5	5.08	4.93	3.63	0.66	0.46	0.51	0.21
Not Preprocessed, lr=0.0005, lambda=0.05	15	0.19	0.23	3.87	3.31	3.1	1.89	0.78	0.64	0.72	0.43
Not Preprocessed, lr=0.0005, lambda=0.1	1	0.12	0.14	6.69	5.17	4.89	3.46	0.32	0.47	0.55	0.2
Not Preprocessed, lr=0.0005, lambda=0.2	11	0.18	0.22	5.41	4.57	4.3	2.62	0.31	0.51	0.61	0.3
Not Preprocessed, lr=0.0005, lambda=0.5	4	0.15	0.18	6.13	4.45	3.94	2.99	0.14	0.51	0.57	0.28
Not Preprocessed, lr=5e-05, lambda=0.01	0	0.21	0.26	7.45	5.55	5.19	3.76	0.16	0.42	0.49	0.18
Not Preprocessed, lr=5e-05, lambda=0.05	13	0.18	0.2	4.42	3.46	2.92	2.54	0.46	0.61	0.66	0.4
Not Preprocessed, lr=5e-05, lambda=0.1	0	0.21	0.25	7.75	5.75	5.42	3.89	0.13	0.4	0.48	0.16

Figure 4.3: Excerpt of the µ-RegPro evaluation CSV for the weakly supervised model across hyperparameter combinations, with and without affine preprocessing. The final scores are consistently higher for affinely preprocessed data.

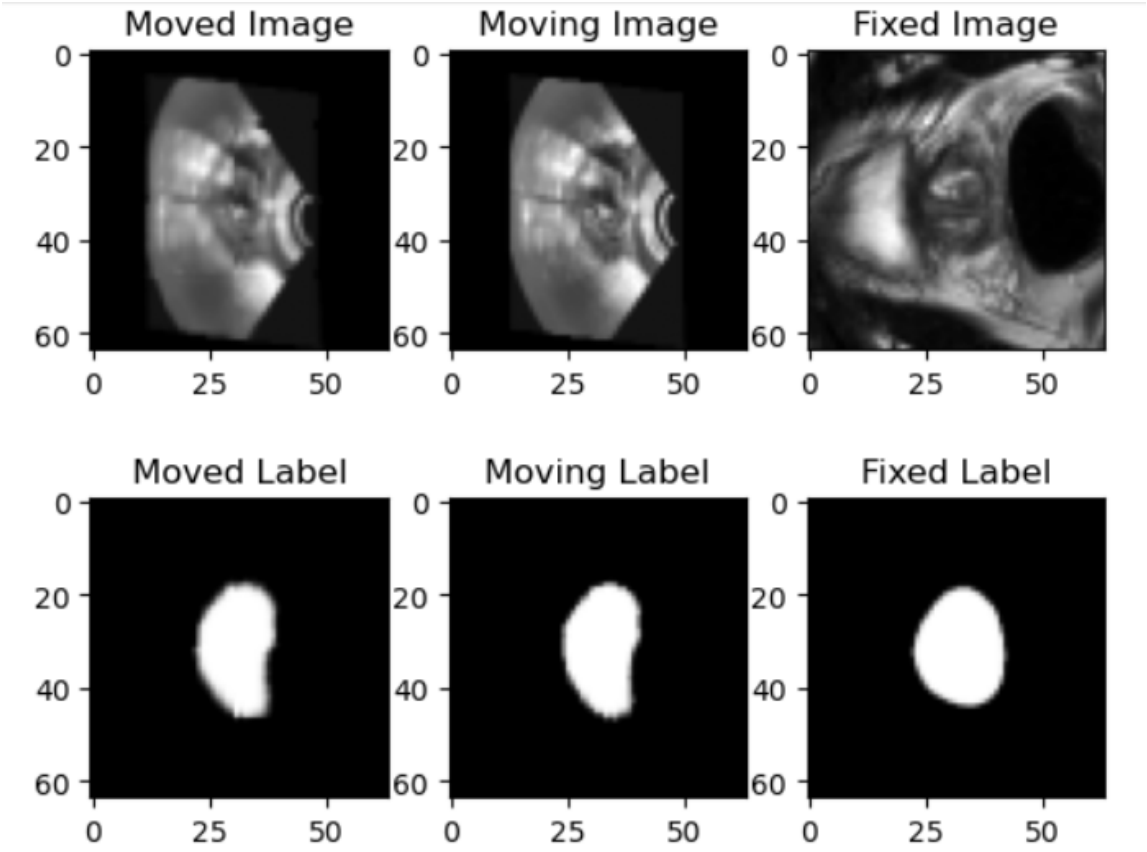


Figure 4.4: Qualitative comparison showing moving, fixed, and warped MRI-TRUS pairs and their corresponding labels for the best-performing weakly supervised model. The gland label remains stable, indicating that the model performs small local corrections rather than large deformations.

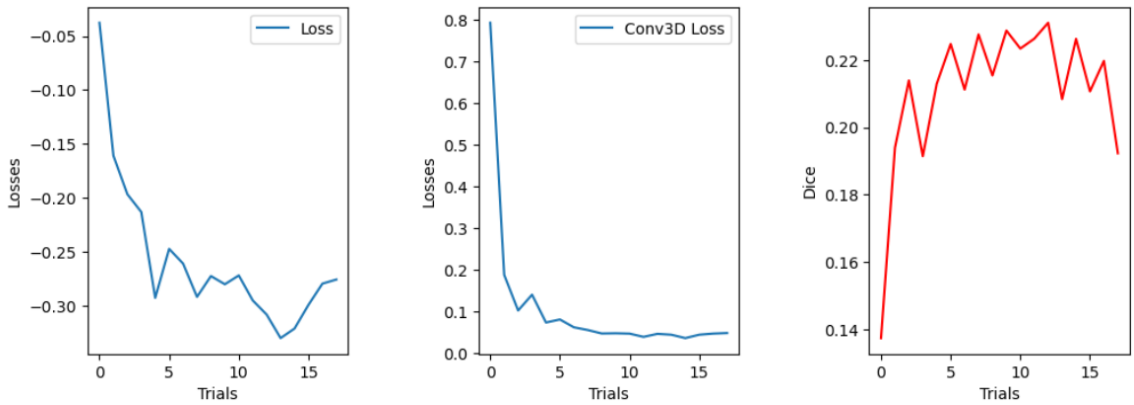


Figure 4.5: Training curves showing total loss, regularization loss, and validation Dice for the best-performing weakly supervised model. The validation Dice converges gradually and stabilizes at approximately twice the maximum of the unsupervised case.

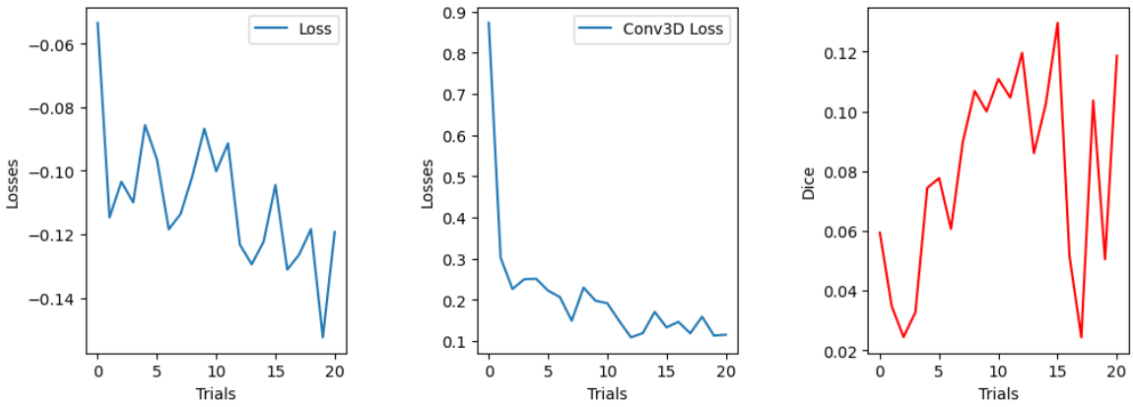


Figure 4.6: Training curves for the weakly supervised model trained without affine preprocessing. The curves are irregular and show reduced convergence, indicating difficulty in handling affine misalignment.

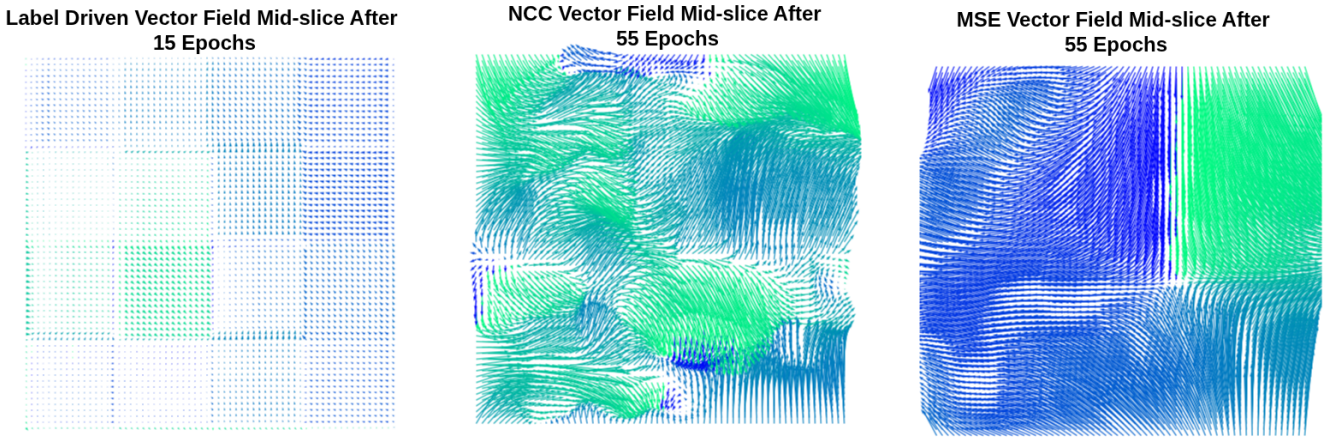


Figure 4.7: Axial mid-slices of the dense displacement fields (DDFs) for the best-performing weakly supervised model and the unsupervised models after regularization loss convergence. The weakly supervised model produces small, smooth displacement vectors consistent with local non-rigid alignment, while the unsupervised models exhibit large, irregular displacements indicative of unstable and anatomically implausible deformations.

4.3 Data Augmentation Investigation

In medical image registration, data scarcity often limits the generalization ability of models. Augmentation plays a crucial role in mitigating overfitting by effectively expanding the diversity of the available training data. Ideally, successful augmentation increases the variability of the data seen by the model, reduces the variation of performance across validation folds, and maintains low bias. This represents the classical bias-variance trade-off in machine learning.

4.3.1 Experimental Procedure

To quantify the effect of data augmentation on model robustness, the model from the previous section (the best-performing weakly supervised configuration) was modified to perform nine independent training runs with different validation partitions. The dataset, consisting of 72 total cases, was divided such that each run used eight consecutive cases for validation and the remaining for training. After each run, the µ-RegPro evaluation metrics were stored in arrays for analysis, allowing the generation of box-and-whisker plots for each metric. A more robust model should exhibit a smaller interquartile range in these plots, indicating reduced variation in performance across validation folds.

Three augmentation configurations were evaluated:

1. No augmentation ($p_{affine} = 0$, $p_{elastic} = 0$)
2. Elastic-only augmentation ($p_{affine} = 0$, $p_{elastic} = 0.7$)
3. Affine-only augmentation ($p_{affine} = 0.7$, $p_{elastic} = 0$)

Each experiment was performed using the same training parameters to isolate the effects of augmentation strategy.

4.3.2 Results and Discussion

Figure 4.8 shows an excerpt from the CSV file containing the µ-RegPro metrics for the three augmentation configurations. The results indicate that the average metric values

differ only slightly between the three experiments, suggesting that the bias of the model was not adversely affected by the introduction of augmentation. The model maintained comparable average performance while being exposed to more diverse training data.

Figure 4.9 presents box-and-whisker plots for each metric across the three augmentation strategies. The spread of the boxes represents the variability of model performance across the nine validation folds. The results show that the non-rigid (elastic-only) augmentation achieved the lowest spread for most metrics, demonstrating improved robustness and generalization. In contrast, the affine-only augmentation produced the largest variation across folds, indicating that small random affine perturbations negatively impacted training stability. This aligns with earlier findings that the model should not be expected to perform or compensate for affine misalignments in this dataset.

Overall, these results confirm that non-rigid augmentation successfully improves the robustness of the model without introducing additional bias. Conversely, affine augmentation degraded performance consistency, verifying that random affine perturbations are not appropriate for this dataset.

4.4 Discussion on Affine Preprocessing

The question of whether affine preprocessing was necessary for this dataset was central to several experiments, including the affine CNN, the intensity metric experiments, the weakly supervised registration experiments, and the data augmentation study. Although the TRUS and MRI volumes in the µ-RegPro dataset appeared roughly aligned, the results show that this apparent alignment is not sufficient for stable or meaningful deformable registration.

ANTs-based affine preprocessing provided clear improvements in Dice scores and overall registration quality, confirming that small but systematic affine misalignments existed between the MRI and TRUS volumes. The affine CNN experiments further highlighted this point - not because the network performed better than ANTs, but because it demonstrated how sensitive affine alignment is to the optimisation landscape. At a typical learning rate of 5×10^{-4}, the CNN failed catastrophically, producing excessive diagonal scaling and rotation. Only when the learning rate was reduced to 5×10^{-5} did it converge to transformations similar to those produced by ANTs. This finding reinforced that affine registration in this context is an inherently delicate problem, one that benefits from deterministic tools like ANTs rather than from a learned affine stage.

Model	Trial	DSC	HD95	TRE	RTs	StDJD	DSC Gland	RDSC Gland	Final Score
Nothing	nan	0.29	8.77	6.75	4.35	nan	0.42	0.49	0.09
ANTS	nan	0.35	2.35	2.46	1.11	nan	0.87	0.88	0.58
Affine and Elastic Augmentation	7.778	0.41	2.847	1.784	0.988	0.142	0.869	0.869	0.628
Elastic Augmentation	7.111	0.409	2.798	1.742	0.99	0.175	0.87	0.87	0.631
No Augmentation	7.111	0.408	2.822	1.768	0.95	0.164	0.87	0.87	0.63

Figure 4.8: Excerpt of the µ-RegPro evaluation CSV for the three augmentation configurations: no augmentation, elastic-only, and affine-only. Metric means remain similar across configurations, suggesting minimal change in bias.

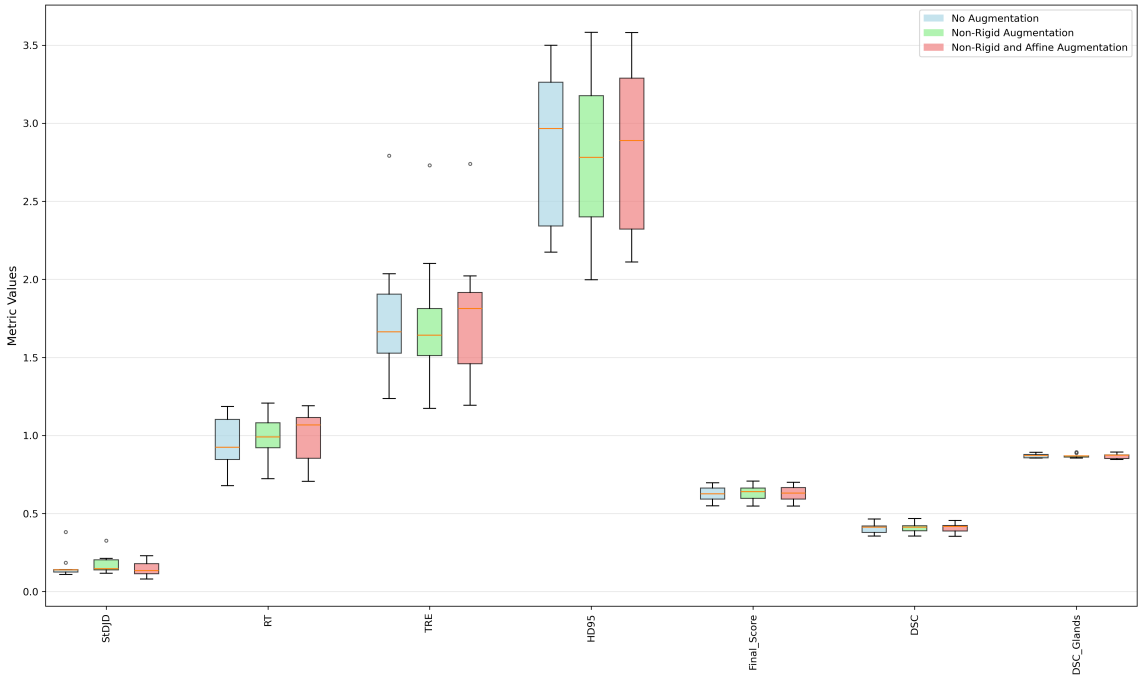


Figure 4.9: Box-and-whisker plots of µ-RegPro metrics for each augmentation strategy. Elastic-only augmentation achieves a narrower spread across folds, indicating improved robustness, while affine-only augmentation increases variability.

The unsupervised intensity metric experiments provided a complementary insight: affine preprocessing alone cannot compensate for the fundamental weaknesses of intensity-driven similarity metrics such as MSE and NCC. Even when preprocessing was applied, these models produced semantically incorrect deformations, confirming that intensity alignment does not equate to anatomical alignment.

The weakly supervised experiments gave stronger evidence for the necessity of preprocessing. Models trained without affine preprocessing displayed unstable loss curves, higher Jacobian variance, and visibly distorted deformation fields, whereas affinely preprocessed models converged faster and achieved consistently higher Dice scores. This showed that residual affine offsets, even if small, increase the burden on the deformable network and lead to instability.

Finally, the data augmentation experiments further supported this conclusion. When affine perturbations were included as part of the augmentation, the model's performance became less stable and more variable across folds, whereas purely non-rigid augmentation improved robustness without degrading accuracy. This behaviour indicates that the deformable model struggles with affine perturbations - precisely the type of misalignment that should already be resolved through preprocessing.

Taken together, these results demonstrate that affine preprocessing is not just beneficial but necessary for this dataset. Although the µ-RegPro volumes are approximately aligned, small residual affine discrepancies significantly reduce registration stability and generalisation. Preprocessing with ANTs provides a clean, deterministic correction step that allows the deformable model to focus on learning subtle, anatomically meaningful deformations rather than compensating for global misalignments.

Chapter 5

Conclusions

This study set out to investigate weakly supervised deformable registration of transrectal ultrasound (TRUS) and magnetic resonance imaging (MRI) using convolutional neural networks. The motivation arose from the persistent challenge of aligning multimodal prostate images, where conventional intensity-based similarity measures such as mean squared error and normalized cross-correlation are unreliable due to strong modality-dependent contrast and texture differences.

The experiments demonstrated that label-driven weak supervision provides a stable and effective means of guiding registration in this multimodal context. In particular, weak supervision based on the Dice similarity coefficient yielded anatomically plausible deformations and improved performance across most evaluation metrics, especially when combined with affine preprocessing. By contrast, unsupervised models using only intensity-based losses failed to produce meaningful anatomical alignment, even after prolonged training, confirming the inadequacy of such measures for TRUS-MRI registration.

Affine preprocessing was found to be a critical component in achieving stable and high-quality registration. Although the dataset appeared roughly aligned, residual scale and translation differences between modalities degraded deformable model performance when left uncorrected. The results indicated that performing a preliminary affine alignment - whether through a learned affine network or an established tool such as ANTs - simplifies the learning process and yields smoother deformation fields.

Data augmentation experiments further revealed that non-rigid (elastic) augmentations improve robustness and reduce variability across validation folds, while affine perturbations tended to degrade performance. These findings suggest that augmentations which mimic

realistic anatomical deformations are more beneficial than those which disrupt the global alignment between modalities.

Overall, this work provides a clear validation of weak supervision and affine preprocessing as key design choices for TRUS-MRI registration using deep learning. The results serve as a framework for practical registration pipelines in clinical or research applications where data is limited and annotations are available only for select structures.

Chapter 6

Recommendations

While the present work successfully established the efficacy of weakly supervised label-driven registration for TRUS-MRI alignment, several directions remain open for future research. The most immediate opportunity lies in refining the registration architecture to better capture local deformations while maintaining global consistency. Incorporating multi-resolution training or coarse-to-fine deformation fields may yield improved accuracy without sacrificing regularization stability.

The limited size of the µ-RegPro dataset constrained the generalization capacity of the models. Future studies would benefit from larger datasets or from the use of synthetic data generation strategies that preserve the physical realism of prostate motion. Moreover, since the test set of the µ-RegPro challenge is not publicly available, external validation on unseen patient data would be essential to establish true generalization performance.

A further avenue for exploration lies in hybrid loss formulations that combine weak supervision with modality-invariant feature learning. Feature extractors trained with adversarial or self-supervised objectives could help bridge the modality gap that limits intensity-based approaches. Similarly, attention-based mechanisms could be explored to focus deformation learning on anatomically relevant regions while ignoring irrelevant background structures.

Finally, the affine preprocessing step could be integrated directly into the registration framework as a joint optimization stage. This would allow affine and deformable transformations to be learned in a single network while maintaining modular interpretability. Combined with uncertainty estimation methods or probabilistic deformation models, such an approach could improve both the transparency and clinical reliability of multimodal registration

systems.

In conclusion, this work lays the foundation for a practical, explainable, and anatomically consistent registration framework. The findings reaffirm that even in deep learning-based deformable registration, domain understanding and careful preprocessing remain as crucial as the architecture itself.

Bibliography

[1] B. Zitová and J. Flusser, "Image registration methods: a survey," *Image and Vision Computing*, **vol. 21, no. 11**, pp. 977–1000, 2003.

[2] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: a survey," *IEEE Transactions on Medical Imaging*, **vol. 22, no. 8**, pp. 986–1004, 2003.

[3] J. Zou, B. Gao, Y. Song, and J. Qin, "A review of deep learning-based deformable medical image registration," *Frontiers in Oncology*, **vol. 12**, p. 1047215, 2022.

[4] B. Huang, F. Yang, M. Yin, X. Mo, and C. Zhong, "A review of multimodal medical image fusion techniques," *Computational and Mathematical Methods in Medicine*, **vol. 2020**, p. 8279342, 2020.

[5] Y. Xiao *et al.*, "Evaluation of MRI to Ultrasound Registration Methods for Brain Shift Correction: The CuRIOUS2018 Challenge," *IEEE Transactions on Medical Imaging*, **vol. 39, no. 3**, pp. 777–786, 2020.

[6] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "VoxelMorph: A learning framework for deformable medical image registration," *IEEE Transactions on Medical Imaging*, **vol. 38, no. 8**, pp. 1788–1800, 2019.

[7] Y. Hu, M. Modat, E. Gibson, W. Li, N. Ghavami, S. Ourselin, J. A. Noble, and T. Vercauteren, "Weakly-supervised convolutional neural networks for multimodal image registration," *Medical Image Analysis*, **vol. 49**, pp. 1–13, 2018.

[8] Q. Zeng *et al.*, "Label-driven magnetic resonance imaging (MRI)–transrectal ultrasound (TRUS) registration using weakly supervised learning for MRI-guided prostate radiotherapy," *Physics in Medicine & Biology*, **vol. 65, no. 13**, p. 135002, 2020.

[9] *μ-RegPro Challenge: Assessment Criteria*, Available: `https://muregpro.github.io/assessment.html`, Accessed: 2025.

[10] A. P. Zijdenbos, B. M. Dawant, R. A. Margolin, and A. C. Palmer, "Morphometric analysis of white matter lesions in MR images: Method and validation," *IEEE Transactions on Medical Imaging*, **vol. 13, no. 4**, pp. 716–724, 1994.

[11] J. Bertels, T. Eelbode, H. Berman, D. Vandermeulen, F. Maes, and P. Suetens, "Optimizing the Dice score and Jaccard index for medical image segmentation: Theory and practice," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2019.

[12] R. D. Datteri, B. M. Dawant, and P. C. Cao, "Evaluation of non-rigid registration accuracy using a realistic brain phantom," *Medical Image Analysis*, **vol. 16, no. 3**, pp. 633–647, 2012.

[13] A. D. Leow *et al.*, "Statistical properties of Jacobian maps and the diffeomorphic metric mapping of brain image registration," *Medical Image Analysis*, **vol. 11, no. 6**, pp. 643–656, 2007.

[14] S. Kabus *et al.*, "Evaluation of 4D-CT lung registration accuracy using vessel bifurcations," *Medical Image Analysis*, **vol. 13, no. 6**, pp. 715–728, 2009.

[15] F. Bianchi *et al.*, "Augmented reality visualization for image-guided interventions," *IEEE Transactions on Biomedical Engineering*, **vol. 68, no. 3**, pp. 902–912, 2021.

Appendix A

GA Tracking

GA	Requirement	Justification and section in the report
1	Problem-solving	The project required developing and evaluating several registration strategies to address the multimodal misalignment between TRUS and MRI. The problem was approached by formulating hypotheses, testing affine preprocessing, unsupervised, and weakly supervised models to find effective solutions. (Sections 4.1–4.4)
4	Investigations, experiments and data analysis	Extensive experiments were designed and executed to test different learning paradigms, hyperparameters, and preprocessing approaches. Results were analyzed both qualitatively and quantitatively using evaluation metrics such as DSC, TRE, and Jacobian determinant statistics. (Sections 4.1–4.3)
5	Use of engineering tools	The work made use of advanced computational tools, including Python, TensorFlow/Keras for model development, ANTs for classical registration, and the university's HPC cluster for large-scale training. (Sections 3.2–3.4)
6	Professional and technical communication (Long report)	The thesis clearly documents the motivation, methodology, results, and implications of the experiments in a structured and technically sound manner, following academic and engineering communication standards. (Entire report)
8	Individual work	All experimental implementation, code design, data processing, and analysis were performed independently by the author, including the design of experiments and interpretation of results. (Sections 3–5)
9	Independent learning ability	The study required self-directed learning in multimodal registration theory, deep learning frameworks, and performance evaluation metrics. Implementation and troubleshooting of model architectures and registration algorithms were conducted without prescriptive instruction. (Sections 2–4)

Table A.1: Graduate Attribute (GA) tracking table showing how each attribute was met throughout the project.

Appendix B

Additional Axial and Sagittal Slices From Validation Set

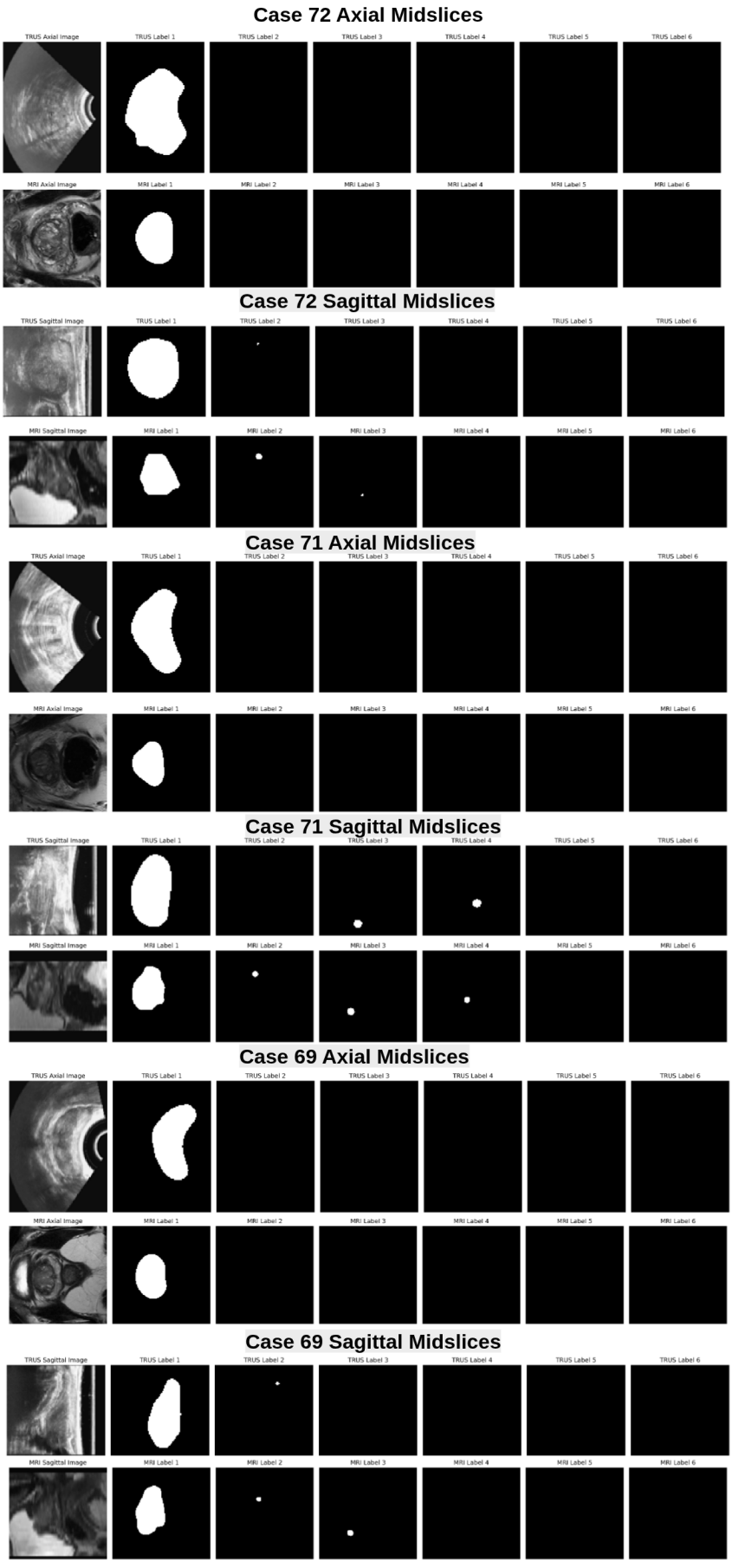


Figure B.1: Additional axial and sagittal mid-slices from the validation set, demonstrating that the gland label is the most robust label and that the rest are sparse or empty

Appendix C

Model Details

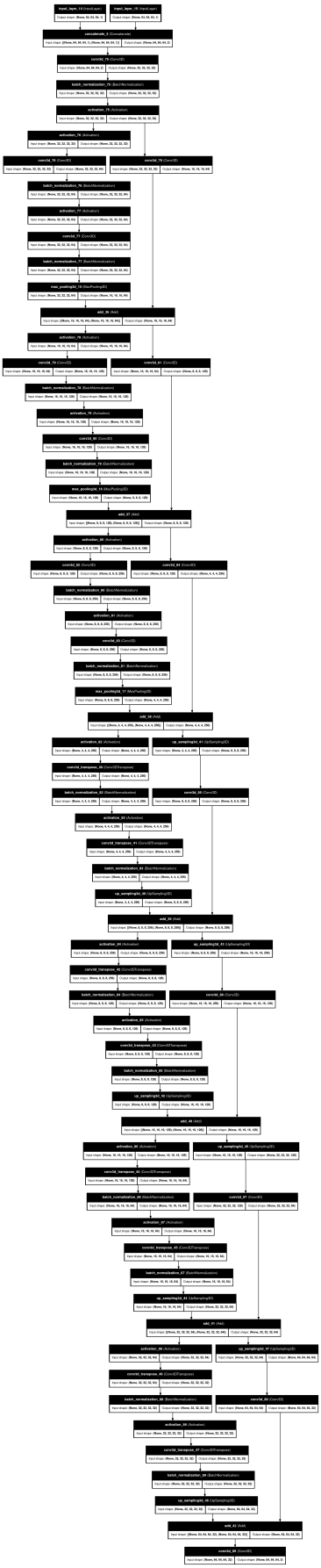


Figure C.1: Additional axial and sagittal mid-slices from the validation set, demonstrating that the gland label is the most robust label and that the rest are sparse or empty

Appendix D

More Examples of ANTs Registration from the Validation Set

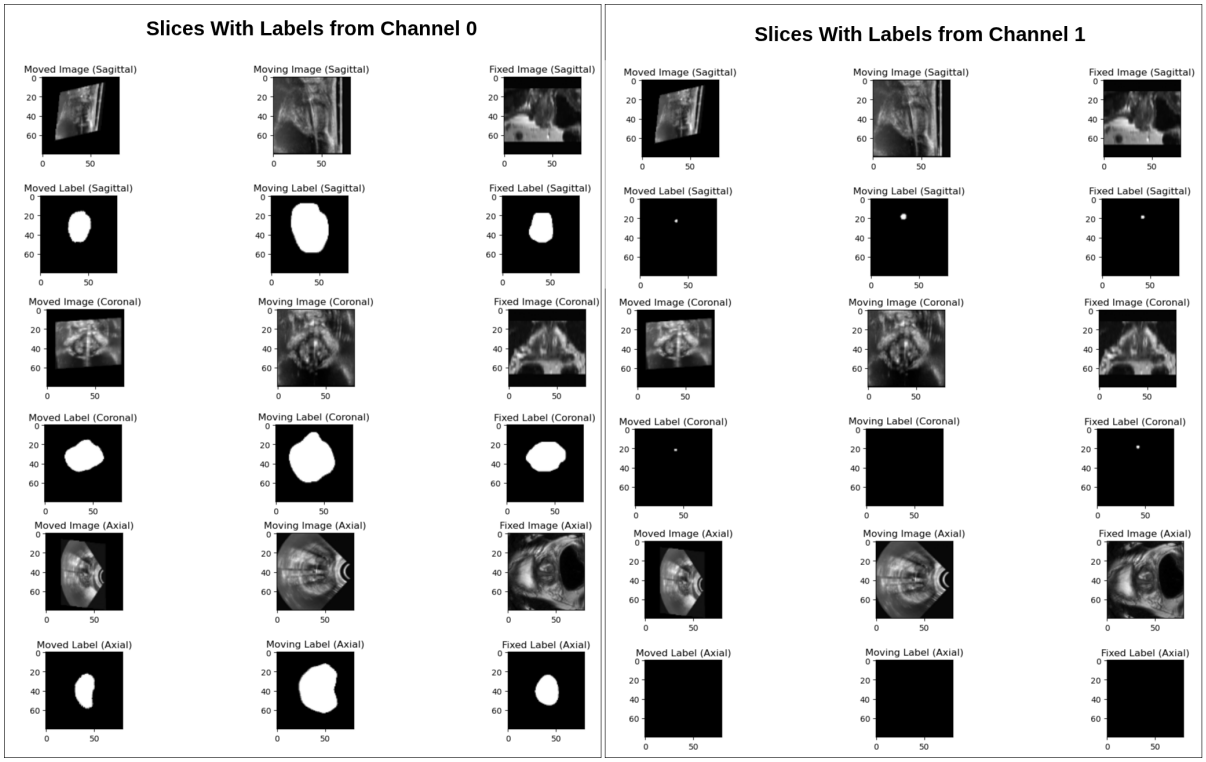


Figure D.1: Additional axial, sagittal and coronal slices of ANTs registered validation pairs, showing labels 1 and 0 to observe registration qualitatively.

Appendix E

AI Use

Large Language Models such as ChatGPT and DeepSeek were used to help create the general outline of the report, experiments and code in general. They were used extensively to find the right syntax for various applications in Python. In summary, they were used as a drafting or brain storming tool and a syntax search tool.

Appendix F

GitHub Link

The code used for this project can be found at GitHub Link