# Label-efficient Learning for Medical Images




**Yiwen Li**

St.Anne's College

University of Oxford

A thesis submitted for the degree of

*Doctor of Philosophy*

Michealmas Term 2024

**Abstract**

Despite the improvements deep-learning methods brought to a great variety of computer vision tasks, the requirement of large-scale annotated datasets hinders many of their real-world applications. In the management of prostate cancer, where segmentation and registration of lower-pelvic imaging play indispensable roles in various stages, automation of these tasks is limited by the availability of samples and annotations. Legal frameworks impose stringent requirements on data collection, necessitating patient consent, which may be withheld due to fears of personal information disclosure and its potential impact on patients' social and professional lives. Rare pathological conditions, such as prostatic adenoid cystic carcinoma, further complicate dataset collection. Additionally, annotations are scarce due to the significant effort required by experienced clinicians with expertise in anatomy and pathology. As a result, previous studies have focused narrowly on specific prostate regions, leaving many structures of planning interest, such as the seminal vesicle and neurovascular bundle, with insufficient labels. Furthermore, generalisation failure has been widely acknowledged due to the limited training samples and their high variability, making domain shifts between training and application data common.

This thesis proposes approaches to problems when large-scale annotated datasets matching the distribution of target evaluation data are unavailable. Three distinct yet highly relevant applications are investigated, including 2D segmentation on computer vision data, 3D segmentation on medical images and 3D segmentation-label-driven image registration. A novel self-supervised task that generates pseudo-classes for general few-shot semantic segmentation is proposed to encourage the learning of a more discriminative feature space while adapting to novel target classes. A fully 3D prototypi-

cal few-shot segmentation algorithm with a spatial registration mechanism and a support mask conditioning module is proposed for few-shot cross-institutional segmentation. A semi-weakly-supervised registration pipeline with two novel commutative image perturbation methods and corresponding consistency rules designed specifically for semi-weakly-supervised image registration is proposed. This thesis describes the details of methodology development and validation of the individual applications, using open data sets in computer vision and, in the case of the latter two medical applications, real-world clinical T2-weighted magnetic resonance (MR) images from prostate cancer patients.

This thesis also describes a collection of manual annotations of eight anatomical structures on 589 T2-weighted MR images, which were acquired, preprocessed and published to aid the reproducibility of the presented studies. This also contributes to future urological and radiological applications for male pelvic organs and other structures. A set of enabling features for medical image registration, used in the third application, have been documented, tested and contributed to the open-source codebase MONAI (monai.io) [1], as part of its official 0.5.0 release.

This thesis is submitted to the Department of Engineering Science, University of Oxford, in fulfilment of the requirements for the degree of Doctor of Philosophy. This thesis is entirely my own work, and except where otherwise stated, describes my own research.

*Yiwen Li, March 2024*

**Acknowledgements**

I am profoundly grateful to my respected supervisor, Professor Victor Prisacariu, for his invaluable guidance and support throughout the entirety of my doctoral journey. Additionally, I would like to express my deepest gratitude to Professor Yipeng Hu, who not only guided me into medical imaging but also contributed as significantly as a supervisor. His expertise, patience and encouragement have profoundly enriched my doctoral experience.

I extend my sincere appreciation to my collaborators for their invaluable contributions, particularly Yuanguan, Wesley and Henry. Additionally, I would like to express gratitude to Wenqi for providing me with the valuable opportunity to contribute to MONAI.

Warm thanks are extended to my dear friends in the Active Vision Lab, including Zirui, Shuda, Theo, Michael, Marcelo, Lixiong, Xinghui, and Wenjing. Together, we have shared countless memorable moments, from shared meals to frisbee and picnics in the sunshine. Wholehearted thanks to everyone accompanied me during this arduous path, especially Krystal, for always listening to my complaints and responding with warmth.

Lastly, I express my heartfelt gratitude to my family. Despite their occasional urge to push me forward, their unwavering support and backing have provided me solace during the most challenging times. Thank you to my grandmother who always stand by my side unconditionally. Though you did not live to see me graduate, I believe you would have been proud of me.

While this list is by no means exhaustive, it serves as a humble acknowledgement of the countless individuals whose support, guidance, and encouragement have played indispensable roles in shaping my academic journey.

# Table of Contents

**1 Introduction** **1**

1.1 Thesis Outline . . . . . 4

1.2 Publication . . . . . 4

**2 Prostate Cancer and Imaging** **6**

2.1 Prostate Cancer . . . . . 6

2.2 Prostate Imaging . . . . . 7

2.2.1 Ultrasound . . . . . 7

2.2.2 Magnetic resonance imaging . . . . . 8

2.2.3 Other imaging . . . . . 10

2.3 Prostate Cancer Management . . . . . 11

2.3.1 Diagnosis . . . . . 11

2.3.2 Cancer Staging . . . . . 12

2.3.3 Active Surveillance . . . . . 13

2.3.4 Treatment . . . . . 14

2.4 Machine Learning Applications . . . . . 16

2.4.1 Cancer Detection and Aggressiveness Characterisation 16

2.4.2 Lower-pelvic Structure Segmentation . . . . . 18

2.4.3 Registration . . . . . 19

2.4.4 Challenges . . . . . 20

2.4.1 Limited Number of Samples . . . . . 20
2.4.2 Limited Number of Annotations . . . . . 21
2.4.3 Domain Shift across Institutions . . . . . 22
2.5 Conclusion . . . . . 22

**3 Literature Review** **24**
3.1 Image Semantic Segmentation . . . . . 24
3.1.1 Overview . . . . . 24
3.1.2 Task Definition . . . . . 25
3.1.3 Deep Learning Engine for Image Semantic Segmentation 26
3.1.1 Model Output . . . . . 26
3.1.2 Dataset . . . . . 26
3.1.3 Training . . . . . 26
3.1.4 Evaluation . . . . . 28
3.1.4 Literature Review . . . . . 29
3.1.1 Classical Methods . . . . . 30
3.1.2 Deep-learning image segmentation . . . . . 32
3.1.3 Summary . . . . . 38
3.2 Image Registration . . . . . 40
3.2.1 Overview . . . . . 40
3.2.2 Task Definition . . . . . 41
3.2.3 Literature Review . . . . . 41
3.2.1 Iterative image registration . . . . . 41
3.2.2 Deep-learning image registration . . . . . 44
3.2.3 Affine Transformation Matrix . . . . . 50
3.2.4 Dense Displacement Field . . . . . 50
3.2.5 Summary . . . . . 51
3.3 Label-efficient learning . . . . . 52

3.3.1 Overview . . . 52
3.3.2 Few-shot learning . . . 52
3.3.3 Semi-supervised learning . . . 59
3.3.1 Deep generative methods . . . 59
3.3.2 Consistency regularisation methods . . . 60
3.3.3 Pseudo-labelling methods . . . 63
3.3.4 Knowledge prior methods . . . 64

**4 Few-shot Semantic Segmentation with Self-supervision from Pseudo-classes 65**

4.1 Task Definition . . . 66
4.2 Background . . . 67
4.2.1 Dataset . . . 67
4.2.2 Episodic Training and Evaluation . . . 69
4.3 Introduction . . . 69
4.4 Related Work in Superpixel Segmentation . . . 73
4.5 Method . . . 73
4.5.1 Prototype Extraction . . . 73
4.5.2 Self-supervision from Pseudo-classes . . . 75
4.6 Experiments . . . 77
4.6.1 Implementation Details . . . 77
4.6.2 Results . . . 78
4.6.3 Ablation Studies . . . 83
4.6.1 Model Architecture . . . 83
4.6.2 Pseudo-class Generation . . . 84
4.6.3 Effect of Novel Class in Training Background 85
4.6.4 Effect of $\alpha$ Parameter . . . 86
4.7 Conclusion and Discussion . . . 86

**5 Prototypical few-shot segmentation for cross-institution male pelvic structures with spatial registration 88**

5.1 Introduction . . . 90

5.2 Task Definition . . . 94

5.2.1 Dataset . . . 95

5.2.2 Evaluation . . . 96

5.3 Related Work in Cross-institution learning . . . 97

5.4 Method . . . 100

5.4.1 Local Prototype Extraction . . . 100

5.4.2 Spatial Registration Mechanism . . . 103

5.4.3 Support Mask Conditioning Module . . . 106

5.4.4 Loss . . . 106

5.4.5 Multiple-shot Evaluation . . . 107

5.5 Experiments . . . 107

5.5.1 Data Set . . . 107

5.5.2 Implementation Details . . . 110

5.5.3 Compared Baseline Networks . . . 111

5.5.4 Ablation Studies . . . 113

5.5.1 Ablation on different modules . . . 113

5.5.2 Ablation study on the number of shots . . . 115

5.5.3 Ablation on varying training data availability 116

5.6 Results . . . 117

5.7 Discussion . . . 125

5.8 Conclusion . . . 126

**6 Semi-weakly-supervised neural networks for medical image registration 127**

6.1 Introduction . . . 128

6.1.1 Learning-based medical image registration . . . . . 128
6.1.2 Semi-weakly-supervised registration . . . . . 129
6.1.3 Commutative property for transformation prediction . 131
6.1.4 Summary of contributions . . . . . 133
6.2 Literature Review . . . . . 133
6.2.1 Medical image registration . . . . . 133
6.2.2 Semi-supervised learning . . . . . 135
6.3 Task Definition . . . . . 136
6.4 Method . . . . . 138
6.4.1 Weak supervision loss on labelled data . . . . . 138
6.4.2 Consistency following weight perturbation . . . . . 139
6.4.3 Image perturbation for registration . . . . . 141
6.4.1 WarpDDF . . . . . 141
6.4.2 RegCut . . . . . 144
6.4.3 Combining WarpDDF and RegCut . . . . . 146
6.5 Experiments . . . . . 147
6.5.1 Dataset and Implementation Details . . . . . 147
6.5.2 Ablation Studies on Registration Performance . . . . . 148
6.5.3 Computational Atlas for Pelvic MR Images . . . . . 148
6.6 Results . . . . . 150
6.7 Conclusion and Discussion . . . . . 153

**7 Conclusion** **155**

7.1 Few-shot Semantic Segmentation . . . . . 155
7.2 Few-shot Cross-institution Segmentation . . . . . 157
7.3 Semi-weakly Supervised Registration . . . . . 158
7.4 Conclusion . . . . . 159

**Bibliography** **160**

# Chapter 1

# Introduction

The data-driven representation learning enabled by deep neural networks (DNNs) has led to improvements in a wide range of computer vision tasks. However, most methods rely on large datasets representative of the task and with full annotations, which are expensive and sometimes infeasible to collect. Especially in the field of medial imaging, the data-related problem is three-fold. Firstly, the number of collectable samples are limited due to patient privacy protection and the rare occurrence of some uncommon pathological appearances. Secondly, when a sufficient number of samples is collected, annotation relies on expertise in anatomical knowledge and significant effort in pixel-level labelling. Thirdly, generalisation failures occur due to the domain shifts between the training dataset and real-application data often caused by discrepancies in the image acquiring protocols.

This thesis investigates methodologies to address these challenges for segmentation and registration, two tasks with wide medical applications, through label-efficient approaches such as few-shot learning and semi-supervised learning.

Segmentation of clinically interesting structures can be used to aid diagnosis decisions, drive longitudinal analysis of physiologic changes, calculate dose usage and support more accurate registration. Automating this task can largely reduce the costs in expertise and intra- and inter-observer variations of manual segmentation. Recent supervised segmentation methods demonstrate promising segmentation results but mostly rely on large data sets with full annotations. Few-shot segmentation, which segments unseen classes using only a few labelled examples provides a direction for addressing the data challenges.

In this thesis we examine several limitations of few shot segmentation. First, we find this task to be more challenging when query images contain other semantic classes besides the target class, pointing out the necessity of learning a more discriminative feature space. Chapter 4 proposes a novel self-supervised task that generates random pseudo-classes in the background of the query images based on superpixel segmentation to provide extra training data that would otherwise be unavailable when predicting individual target classes. This work [2] has been published with 2021 British Machine Vision Conference (BMVC).

Along with the challenge of collecting large labelled datasets, in the field of medical imaging, domain shifts caused by different image acquiring protocols often lead to generalisation problems. A fully 3D prototypical few-shot segmentation algorithm is proposed in Chapter 5 for training networks that can be adapted effectively to clinically interesting structures that are absent in training, using only a few labelled images from a different institute. Specifically, a *spatial registration mechanism* and a *support mask conditioning module* are proposed for, respectively, compensating the spatial variability and further utilisation of annotation available from the support images. The

preliminary version of this work has been accepted as a International Symposium on Biomedical Imaging conference paper [3] and its full version has been accepted as a journal paper by Medical Image Analysis [4].

Image registration is the process of establishing spatial correspondence between images and, in medical imaging, can be applied to both intra-patient and inter-patient. Longitudinal intra-patient registration enables the tracking of radiological evolution. Multi-model intra-patient registration combines information across different imaging modalities. Inter-patient registration promotes intra-operative motion modelling for generating patient-specific motion data and enables group-wise registration for population studies. Like in the segmentation task, the cost in expertise and inter-/intra- observer variations would be significantly reduced by automating the labeling process. Weak-supervision from segmented corresponding regions-of-interest (ROIs) has been proven effective for supplementing registration training as we discuss in the literature review, but entails significant specialised effort to produce the data labels.

Chapter 6 describes a novel semi-weakly-supervised registration pipeline, uses labelled image pairs to improve model performance, when only a small corresponding-ROI-labelled dataset is available. Specifically, we propose two types of augmentation methods, i.e for network weights and image resampling, such that consistency-based unsupervised losses can be applied on unlabelled data. This work is currently submitted to IEEE Transactions on Medical Imaging and under review.

## 1.1 Thesis Outline

This thesis first provides the clinical background about prostate cancer and imaging, which has been adopted as the application field by the second and third contributions, in Chapter 2. Literature reviews of the two major tasks (semantic segmentation in section 3.1 and registration in section 3.2) and two label-efficient learning approaches (few-shot learning and semi-supervised learning in section 3.3) are described in Chapter 3. The three following chapters respectively detail the three technical contributions - the proposed pipelines for few-shot semantic natural image segmentation in Chapter 4, few-shot cross-institutional medical image segmentation in Chapter 5 and semi-weakly-supervised medical image registration in Chapter 6.

## 1.2 Publication

The following publications are discussed in this thesis:

- **Li, Y.**, Data, G.W.P., Fu, Y., Hu, Y. and Prisacariu, V.A.(2021). Few-shot Semantic Segmentation with Self-supervision from Pseudo-classes. In *The British Machine Vision Conference (BMVC).*
- **Li, Y.**, Fu, Y., Yang, Q., Min, Z., Yan, W., Huisman, H., Barratt, D., Prisacariu, V.A. and Hu, Y. (2022). Few-shot Image Segmentation for Cross-institution Male Pelvic Organs using Registration-Assisted Prototypical Learning. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI) (pp. 1-5). IEEE.*
- **Li, Y.**, Fu, Y., Gayo, I.J., Yang, Q., Min, Z., Saeed, S.U., Yan, W., Wang, Y., Noble, J.A., Emberton, M. and Clarkson, M.J. (2023).

Prototypical Few-shot Segmentation for Cross-institution Male Pelvic Structures with Spatial Registration. In *Medical Image Analysis, 90, p.102935.*

- **Li, Y.**, Fu, Y., Gayo, I.J., Yang, Q., Min, Z., Saeed, S.U., Yan, W., Wang, Y., Noble, J.A., Emberton, M. and Clarkson, M.J. (2024). Semi-weakly-supervised neural network training for medical image registration.

The following works I contributed to during my DPhil are not included since they are only loosely connected to this thesis:

- Howard-Jenkins, H., **Li, Y.** and Prisacariu, V.A. (2020). GroSS: Group-Size Series Decomposition for Grouped Architecture Search. In European Conference on Computer Vision (pp. 18-33). Cham: Springer International Publishing.

- Cardoso, M., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murray, B., Myronenko, A., Zhao, C. and Yang, D., MONAI: An open-source framework for deep learning in healthcare. (2022). arXiv preprint arXiv:2211.02701.

- Fu, Y., **Li, Y.**, Saeed, S.U., Clarkson, M.J. and Hu, Y. (2023). A Recycling Training Strategy for Medical Image Segmentation with Diffusion Denoising Models. arXiv preprint arXiv:2308.16355.

- Fu, Y., **Li, Y.**, Saeed, S.U., Clarkson, M.J. and Hu, Y., 2023. Importance of Aligning Training Strategy with Evaluation for Diffusion Models in 3D Multiclass Segmentation. arXiv preprint arXiv:2303.06040.

# Chapter 2

# Prostate Cancer and Imaging

The challenges in data collection, label annotation and generalisation encourage the investigation of label-efficient approaches to promote the real-life application of deep-learning-based methods, the medical imaging field. Prostate cancer (PCa), a prevalent and durable disease with a wide range of potential deep-learning applications, has been chosen as the focus of this thesis's second and third contributions. This chapter provides the background of prostate cancer, describing the imaging modalities used for prostate imaging in Section 2.2, management of prostate cancer in Section 2.3 and potential machine learning applications in Section 2.4

## 2.1 Prostate Cancer

Prostate cancer (PCa), accounting for 9% of all cancer deaths among men, is now recognised as one of the most prevalent diseases facing male population [5]. In several regions, including the USA, Australia, New Zealand, and Western and Northern Europe, it is now the most commonly diagnosed non-cutaneous malignancy in men [6]. However, prostate cancer is a slow-

progressing disease, with a higher than average five-year survival rate of above 90%. The age is also an important risk factor of prostate cancer. Therefore, many patient spend a long time living with prostate cancer and may pass away due to other reasons [7].

This above-discussed long duration and high variation in comorbidity of the prostate cancer substantially complicate the management of the disease. Different diagnosis, grading, staging and treatment options have been debated for their varying sensitivity, specificity, efficacy and complication. For example, over-diagnosis and under-diagnosis from the widely-adopted sextant biopsy have both been reported for different patient cohorts [8], while over-treatment for traditional radical prostatectomy has recently been recognised [9].

## 2.2 Prostate Imaging

Imaging plays an important role in the diagnosis, management and treatment of prostate cancer. Multiple modalities with different characteristics have been employed for prostate imaging in various clinical tasks. This section describes their advantages and disadvantages, as well as their targeted applications.

### 2.2.1 Ultrasound

Sonography is created by using a probe to send pulses of ultrasound into tissue and display echos differentiated due to different reflection properties. For prostate imaging, the ultrasound probe is inserted into the rectum to generate transrectal ultrasonography (TRUS). TRUS has been used to guide biopsy acquiring histopathological samples as a mainstay urological prac-

tice due to its cost-effective visualisation of the prostate gland. However, many lesions are now known to be isoechoic, which are not visible at all on ultrasound images. Even for hypoechoic lesions reflecting significantly less sound echoes than normal tissue which are detectable by ultrasound, the low signal-to-noise ratio of grayscale ultrasound images leads to less accurate delineation of soft tissues and difficulty in reliably differentiating cancerous from non-cancerous regions (with a 40% detection rate of prostate cancer) [10]. A variety of new ultrasound-based imaging techniques such as share-wave elastograph, color doppler ultrasound, contrast-enhanced ultrasound, micro-ultrasound and their combinations have been proposed to supplement grayscale ultrasound for higher resolution and better sensitivity.

### 2.2.2 Magnetic resonance imaging

Magnetic resonance imaging (MRI) provides a clear depiction of zonal anatomy and soft tissue resolution superior to any other imaging modality to date. Prostate MRI is typically performed on a 1.5- or 3-T magnet using a multichannel pelvic phased-array receiver coil. The use of an endorectal coil, which increases signal-to-noise ratio at the cost of patient discomfort, image artifacts, and longer examination times, is still debatable.

The Prostate Imaging Reporting and Data System (PI-RADS) provides a structured reporting scheme for radiologists to evaluate suspected prostate cancer based on prostate multiparametric MRI (mpMRI) which combines anatomic (T1-weighted and multiplanar T2-weighted images) and functional sequences (diffusion-weighted imaging with apparent diffusion coefficient maps and dynamic contrast-enhanced imaging). T1-weighted and T2-weighted MRI respectively allow the magnetisation to recover and decay before measuring the signal by changing the repetition and echo times. As a result,

T1-weighted MRI enhances the signal of the fatty tissue and suppresses the signal of the water while T2-weighted MRI enhances the signal of the water. For prostate imaging, T1-weighted MRI is mainly used to survey the entire pelvis for enlarged pelvic lymph nodes and possible bone metastases, and T2-weighted MRI shows detailed zonal anatomy and can be used to evaluate extracapsular extension and seminal vesicle invasion. Diffusion-weighted imaging (DWI) maps the diffusion process of molecules, mainly water, in biological tissues, to reflect interactions with obstacles such as macro-molecules, fibres and membranes. Apparent diffusion coefficient (ADC) maps reduce T2 weighting by taking multiple conventional DWI images with different DWI weighting for more specific diffusion representation. Prostate DWI is an indicator of cellularity and tumour aggressiveness. Dynamic contrast-enhanced (DCE) MR perfusion relies on T1-shortening effects of gadolinium-based contrast agent. Perfusion parameters could be calculated by evaluating T1 shortening on rapidly repeated T1 imaging after injection of intravenous contrast bolus. Persistent increased signal intensity in DCE MRI indicates permeable blood vessel characteristics of tumour tissue, which could be used to assist the interpretation of T2-weighted MRI and DWI in the detection of high-risk prostate cancer and surveillance status post-treatment. The multiparametric approach increases the predictive power of magnetic resonance imaging for diagnosis [8].

Compared to TRUS, MRI is a relatively new, expansive and less prevalent diagnostic tool for prostate cancer, but its ability to identify small yet clinically significant tumours has attracted increasing amounts of research interest.

### 2.2.3 Other imaging

CT, with inferior soft-tissue contrast and lack of molecular information, is mainly used for the evaluation of nodal and distant metastases. Though has a low yield compared with more advanced hybrid imaging methods like PET/CT and relies on morphologic criteria, CT is still currently recommended for patients with intermediate to high-risk PCa in the guidelines of the American Urological Association due to its high availability [11]. Moreover, the ability to measure electron density makes CT the conventional modality for radiotherapy planning.

Positron emission tomography (PET) uses radioactive tracers to visualise and measure changes in metabolic processes and other physiological activities. PET scanners can incorporate a CT scanner or MRI scanner to acquire sequential images from both devices in the same session, combining functional imaging provided by PET with structural information from CT or MRI into a superposed image. PET/CT has gained increasing importance in prostate cancer detection with several tracers demonstrating efficacy in various clinical settings. Sodium 18F-Fluoride (NAF) PET/CT reflects the increased blood flow, bone remodelling, and osteoblastic activity related to osseous metastases [12]. Choline PET/CT is mainly used for restaging in the setting of biochemical recurrence (BCR) [13]. Prostate-specific membrane antigen (PSMA) PET/CT shows a robust performance even at low prostate-specific antigen (PSA) levels and has become the front-runner of the currently available tracers [14].

## 2.3 Prostate Cancer Management

The management of the prostate cancer is complicated for its long duration and high variation in comorbidity. Taking into account the risks and potential impact on the patient's health-related quality of life, individuals diagnosed with prostate cancer may opt for active surveillance or undergo various treatments depending on the stage of the cancer.

### 2.3.1 Diagnosis

Men with clinical suspicion of prostate cancer evidenced by threshold prostate-specific antigen (PSA) or abnormal digital rectal examination (DRE) are recommended biopsy for further diagnosis. Early diagnosis of prostate cancer relies on DRE with transrectal biopsies targeting palpable lesions. The development of transrectal ultrasound (TRUS) enabled direct visualisation of the prostate with a rectal ultrasound probe. While not an accurate detection method alone, TRUS enabled better guidance for transrectal biopsies. Standard sextant biopsy protocols and extended-core biopsy protocols respectively involved taking three and ten to twelve biopsy cores from the prostate. Systematic biopsies increase sensitivity for prostate cancer detection with greater numbers of cores but fail to discriminate well between clinically significant and clinically insignificant cancers.

mpMRI, which relies on the detection of morphological changes in anatomical and functional radiomics to diagnose the presence and extent of cancer [15], was introduced to further improve diagnostic accuracy. mpMRI consists of a combination of MRI sequences - anatomical sequences providing details of the prostate, in particular T2-weighted (T2W) images, and functional se-

quences including diffusion-weighted imaging (DWI) and dynamic contrast enhancement (DCE), to assess for features suspicious for potential prostate cancer. It became the first line imaging for patients found to have a raised PSA level or abnormal DRE prior to biopsy. It could be used not only for locating suspicious targeting regions but also for screening purposes using either the Prostate Imaging-Reporting and Data System (PI-RADS) [16] or Likert scoring systems that aim to quantify the likelihood of clinically significant prostate cancer being present on a scale of 1 to 5. Studies demonstrated that a diagnostic pathway including risk assessment with MRI before biopsy and MRI-targeted biopsy in the presence of a lesion suggestive of cancer was superior to the diagnostic pathway of standard TRUS-guided biopsy [17]. According to the PROMIS trial, Using mpMRI to triage men might allow 27% of patients to avoid a primary biopsy and diagnosis of 5% fewer clinically insignificant cancers [8]. Moreover, for clinical T-staging, MRI demonstrates higher accuracy in the assessment of uni- or bilobar disease (T2), extracapsular extension and seminal vesicle invasion (T3), as well as the invasion of adjacent structures (T4) compared with DRE, TRUS and computed tomography (CT). T2-weighted MRI shows a 97% sensitivity in prostate cancer localisation [18].

### 2.3.2 Cancer Staging

Prostate cancer is staged using the Tumour, Node, Metastasis (TNM) system. Used alongside PSA levels and histological grade, TNM helps to risk stratify disease and determine appropriate treatments. Tumour (T) assessment scores are based on the size or area of the cancer. The T-stage assessment is primarily clinical based on DRE findings. mpMRI could also be used for the assessment and has been shown to be highly specific but moderately

sensitive for picking up adverse features including extraprostatic extension (T3a) and seminal vesicle invasion (T3b) [19]. Nodal (N) staging assesses the absence(N0)/presence(N1) of regional lymph node (LN) metastasis and upstages non-regional lymph nodes to M1a [20]. Metastases are classified by the absence (M0)/presence (M1) of distant metastases which could be further stratified into M1a for non-regional lymph nodes, M1b for bony metastases, and M1c for metastases for other sites based on location.

Patients at high risk of advanced disease are traditionally examined for nodal and bone metastasis respectively with contrast CT and technetium 99 (Tc99) bone scan. Recently, whole-body MRI and PET/CT, have been shown to be superior in the detection of bony metastasis [21, 22].

Patients with N1M0 have a high risk of recurrence and therefore should be considered for adjuvant therapies dependent on nodal volume [20], making accurate nodal staging important. Though may lead to higher morbidity, extended pelvic lymph node dissection (ePLND) during radical prostatectomy remains the gold standard of local nodal staging [23] and found to be significantly more sensitive than PSMA PET/CT [24].

### 2.3.3 Active Surveillance

For patients with low and intermediate-risk prostate cancer, standard therapies may be undesirable due to the significant risk and morbidity to the patient's health-related quality of life with potential impact on sexual, urinary, and bowel functions. Active surveillance can manage low/intermediate-risk prostate cancer to delay definitive treatment until progression.

Active surveillance follow-up protocols vary. In the UK, the National Institute for Clinical Excellence (NICE) recommend PSA to be checked every

3-4 months in the first year and six monthly thereafter, with annual digital rectal examinations and a repeat MRI at 12-18 months with further MRIs and/or repeated biopsies indicated by concerning examination features or PSA kinetics.

The Prostate Cancer Radiological Estimation of Change in Sequential Evaluation (PRECISE) criteria standardised the reporting of MRI examinations performed to assess for disease progression in patients on active surveillance for prostate cancer. It scores the likelihood of disease progression on a scale of 1-5, where 1 or 2 represent disease regression, 3 is a stable disease, and 4 or 5 are varying degrees of radiological progression, with evidence showing that it performs well with a high specificity and positive predictive value (PPV). However, two studies suggest omitting biopsies and relying on mpMRI alone has insufficient diagnostic accuracy to exclude disease progression even when studies utilized the PRECISE criteria. Another emerging method is using PSMA PET/CT to reduce repeated biopsies with one study has showing potential to reduce false positives and improve negative predictive value (NPV) compared with mpMRI.

### 2.3.4 Treatment

Once prostate cancer is found, localised, and imaged, a treatment plan is required. Standard therapies includes surgery organ deprivation and radiotherapy.

Radical prostatectomy - complete surgical removal of the prostate - remains the most frequently used treatment option for men with localised prostate cancer. Even after TRUS-guided biopsy, surgeons have limited information about the tumour. MRI taken before prostatectomy helps to make appropri-

ate surgery decisions (to resect or spare the neurovascular bundles [25], predict intra-operative blood loss and the time to stable unitary continence [26]).

Radiation therapy simultaneously decreases the intensity of the treatment to the whole prostate gland and increases the intensity to the identified tumour. MRI, providing superior soft tissue contrast, is helpful in treatment plans as it spare the normal tissue in the surrounding pelvic organs [27].

For patients with low and intermediate-risk prostate cancer, standard therapies can replaced with focal therapies. Focal therapy is a selective ablation targeted to specific sites in the prostate gland to reduce the chance of injury to adjacent organs and resultant side effects [28]. Most prevalent minimally invasive thermal ablation methods such as cryotherapy, laser ablation and high-intensity-focused ultrasound (HIFU) are targeted by multiparametric MRI which is currently considered the optimum approach to accurately localise and contour the target lesion [29]. Moreover, near real-time MRI thermometry and subsequent dose estimations assist tumour targeting and treatment parameter adjustment. For percutaneous cryoablation, MRI enables not only high soft-tissue resolution but also high accuracy in distinguishing normal and frozen tissue [30]. Laser-induced interstitial thermal therapy (LITT) is inherently MRI-compatible as simultaneous deposition of light energy and MRI signal acquisition could be performed without degrading the MRI image. Though TRUS guidance has been adopted for focused ultrasound ablation for many years [31], inadequate and over-ablation due to TRUS's limitation in visualising the cancer focus (especially when the focus is small) [32, 33] encouraged the adoption of MRI-integrated systems capable of thermal monitoring and lesion/zone localisation. Accurate delineating regions of interest is vital in focal therapy to avoid both inadequate treatments leading to the recurrence of cancer and over-treatment leading to morbidities

such as impotency, haematuria, dysuria, urine retention and spontaneous closure of rectal fistula. Automated multi-parametric MRI segmentation could potentially assist the treatment with high accuracy and reduced inter- and intra-observer variability.

## 2.4 Machine Learning Applications

The development of machine learning algorithms analysing medical to assist clinical decision-making for a variety of diseases (including prostate cancer) is an active area of research. This section describes the application scenarios as well as task definitions of the most prominent prostate cancer-targeted machine-learning tasks.

### 2.4.1 Cancer Detection and Aggressiveness Characterisation

The detection of cancer lesions and grading aggressiveness of the detected lesions from non-invasive images (MRI, ultrasound image) is imperative in various stages of prostate cancer management from diagnosis, active surveillance to treatment. At the diagnosis stage, patients with aggressive cancer detected and localised on non-invasive images could benefit from MRI-targeted TRUS-guided biopsy with higher precision and less biopsy needle sample to take, and patients with no cancer detected or characterised as indolent cancer according to non-invasive images could safely avoid biopsy and its side-effects. After the biopsy, accurate automated cancer aggressiveness grading on the acquired histopathology images could speed up pathologists' grading process as well as alleviate inter- and intra-pathologist variability in Gleason grading. The treatment of aggressive prostate cancer could be priori-

tised and better treatment decisions (whether to perform active surveillance, focal therapy or radical prostatectomy) could be made knowing the cancer location and extent of aggressiveness. These above applications require Machine learning algorithms accurately detecting, localising and characterising lesions as aggressive or indolent from non-invasive images have the potential to be included in routine clinical care for patient management and treatment planning.

Cancer detection machine learning algorithms analysing non-invasive images can be divided into two major tasks: lesion classification and lesion detection. Lesion classification models [34, 35, 36] classify lesions outlined by radiologists by clinical significance. Ground truth labels are acquired through pathology confirmation from prostate biopsy or surgery. Lesion detection [37, 38, 39, 40, 41, 42, 43, 44, 45, 46] - the task to predict a pixel-level probability (segmentation) of cancer distribution for the prostate, highlighting highly suspicious areas - is more challenging than lesion classification. Three main types of ground-truth labels adopted by lesion detection methods include 1) radiologist outlines of PI-RADS 3 or above lesions with [37, 38, 47, 40] or without [41, 42, 46] pathology confirmation, 2) pathologist outlines of cancer on whole-mount histopathology images mapped to pre-operative MRI [47, 39], and 3) Gleason pattern labels on whole-mount histopathology images from deep learning algorithms mapped on to MRI [45, 43, 44]. The evaluation could be patient-level, lesion-level or pixel-level.

Moreover, histopathology images could also be analysed with machine learning algorithms to segment cancer regions and automate Gleason grading. The input digital histopathology images could be derived from prostate needle core biopsies or radical prostatectomy specimens which ground-truth labels

are either from pathology reports [48, 49, 50] or pixel-level annotations by experienced pathologists [51, 52, 53]. The evaluation could be pixel-level, region-level, slice-level or patient-level based on label type. Studies show machine learning algorithms achieved a ranking comparable to that of international experts in prostate pathology [54] and could assist pathologists to significantly improve Gleason grading [49]. This accurate performance motivated the mapping of segmentation predicted by deep learning algorithms to MRI as ground-truth label for MRI lesion detection [45, 43, 44].

### 2.4.2 Lower-pelvic Structure Segmentation

Besides the cancer lesions, accurate delineation and contouring of lower-pelvic structures is also essential for various prostate cancer management applications. More accurate MRI-targeted TRUS-guided biopsy is enabled by fusing MRI and TRUS images with multi-modal registration driven by segmentation maps [55, 56, 57, 58]. PI-RADS scoring system requires localising the lesion to either transition zone or peripheral zone, which depends on segmenting both zones [59]. Segmentation at different time points could be used for longitudinal analysis of patients undergoing active surveillance [60, 61, 56, 62, 63]. Daily adaptive treatment plan relies on the segmentation of lower pelvic structures to correct compromised alignment caused by movement in the position of the clinical target volume (CTV) due to physiologic changes in bladder and rectum volume [64]. Moreover, gland volume could be estimated from the segmentation for drug-induced prostate volume change measuring [65], correlation with cancer volume [66] and significant cancer detection [67, 68].

To this end, deep learning on male lower-pelvic structure segmentation has been promoted by the organisation of challenges such as the Prostate MR

Image Segmentation 2012 challenge (PROMISE-12) [69], the NCI-ISBI 2013 Challenge [70], the medical segmentation decathlon (MSD-prostate) [71], all providing annotated datasets. Popular regions of interest include prostate capsule, central gland and peripheral zone and other surrounding regions.

### 2.4.3 Registration

Registration could be applied to intra-patient, longitudinal [72] and cross-modalities [73, 74] image pairs as well as inter-patient image pairs.

Aligning images of the same patient acquired at different time points helps to estimate the spatial transformation of various structures, especially prostate glands, for tracking radiological evolution. Multi-modal registration enables fusion of images of the same patient across different modalities for information combination. When biopsy is performed during diagnosis, real-time TRUS image urologists used to guide the biopsy needle suffer from low signal-to-noise ratio and artefacts which hinder clinicians from reliably distinguishing cancerous from normal regions. To address this issue, suspicious lesions could be located from MRI and then projected to TRUS images for biopsy targeting through fusion with MRI-TRUS registration. During treatment, CT is widely used for dose planning and delivery of prostate cancer radiotherapy. However, its low tissue contrast makes identifying prostate boundaries, especially the interface between prostate, bladder and rectum difficult. MRI, on the other hand, provides superior soft tissue contrast and is ideal for delineation but may suffer geometric distortion effects and lack necessary physical information such as electron density for dose calculation. Registering MRI to CT substantially improves contouring accuracy and therefore high treatment efficacy. For research interest, mapping whole-mount histopathology images acquired from radical prostatectomy to pre-surgical MRI promotes

existing MRI interpretation and provides training/evaluation data for MRI cancer detection algorithms [45, 43, 44]. Inter-patient registration promotes intra-operative motion modelling for generating patient-specific motion data and enables group-wise registration for population studies.

The alignment of prostate images for the above purposes requires a significant amount of time from experienced clinicians with expertise anatomy knowledge. Machine learning algorithms could be adopted to automate, accelerate and standardise the procedure.

### 2.4.4 Challenges

Deep-learning algorithms often depend on large-scale fully-annotated datasets, which can be challenging to collect due to limited sample availability and high annotation costs. Even when such datasets are accessible, there is often a loss in performance due to domain shifts caused by differing image-acquisition protocols between training and evaluation data, as recognised in previous studies [75]. This section highlights the challenges of data scarcity from various perspectives.

#### 2.4.1 Limited Number of Samples

In the UK, the collection and publication of a medical image dataset must adhere to legal frameworks such as the Data Protection Act (DPA) 2018 [76], which incorporates the principles of the EU General Data Protection Regulation (GDPR) [77], and the Common Law Duty of Confidentiality (CLDC) [78]. These regulations require that data collection and processing be fair, lawful, and transparent. Consent must be obtained under specific conditions outlined by the CLDC. Such consent must be freely given, specific, informed, unambiguous, and withdrawable at any time.

To ensure anonymity, metadata containing personally identifiable information (PII) should be removed. However, even with full anonymisation, patients may still have concerns about the potential inadvertent disclosure of their personal and sensitive health information. There is a societal stigma associated with prostate cancer, and patients may worry about how this disclosure could affect their social and professional relationships. For instance, revealing a prostate cancer diagnosis could impact a patient's current or future employment opportunities and insurance coverage, as employers or insurers might view them as a higher risk. The public accessibility of their medical information might add stress or anxiety to patients already dealing with the emotional burden of a cancer diagnosis.

Even with complete patient consent, certain pathological appearances may be rare and potentially absent in many institutions. For example, prostatic stromal sarcoma accounts for less than 0.1% of prostate malignancies [79], and prostatic adenoid cystic carcinoma has been reported in fewer than 150 cases [80].

### 2.4.2 Limited Number of Annotations

Limited annotations are available for lower-pelvic structures crucial for planning purposes. Previous studies, such as INDEX [81] and PROMISE12 [69], have solely segmented the prostate capsule. Similarly, NCI-ISBI [70] segmented solely the central gland (CG) and the peripheral zone (PZ). Despite the significance of these central gland and peripheral zone segments for PI-RAD scoring, biopsy targeting, and treatment planning, other structures of planning interest remain unaddressed. For instance, improper treatment leading to seminal vesicle damage or neurovascular bundle injury can result in infertility or erectile dysfunction, respectively. Additionally, delineating the

surrounding structures of the prostate can enhance registration accuracy.

This thesis addresses this gap by annotating and collecting labels for eight key anatomic structures (bladder, bone, central gland, neurovascular bundle, obturator internus, rectum, seminal vesicle, and peripheral zone) from 589 T2-weighted images acquired from 7 different institutions, each involving the same number of patients. Completing this task required an estimated one thousand observer-hours (more than one and a half hour for each 3D image) from biomedical imaging researchers with at least 2 years of experience in the annotation of medical image data, posing a significant challenge for most local hospitals if conventional supervised learning methods were adopted for tool validation or adaptation. This underscores the clinical relevance and efficiency of the proposed few-shot segmentation approach.

### 2.4.3 Domain Shift across Institutions

Moreover, the performance loss when generalising deep learning models across institutions is well-recognised [75]. The generalisation failure roots from the different imaging protocols adopted by different institutions with variations in terms of scanners (manufacturers and field strengths), varying field-of-view and anisotropic voxels dimensions.

## 2.5 Conclusion

Prostate cancer, one of the most prevalent diseases faced by the male population, is complicated to manage due to its long duration and high variability. Non-invasive imaging (MRI and ultrasound) plays important roles throughout prostate cancer management stages from diagnosis, staging, active surveillance to treatment. Prominent tasks that could be potentially as-

sisted or delivered by machine learning algorithms for improved efficiency and accuracy, as well as reduced inter-/intra-observer variability, include cancer detection and following aggressiveness characterisation, lower-pelvic structure segmentation and intra-/inter-patient registration. Yet, training deep-learning algorithms for the above-mentioned tasks still requires a sufficient amount of data and annotations which, like in many other medical applications, could be expensive if feasible to obtain. Moreover, domain shifts between training and evaluation data caused by various image-acquiring protocols could lead to failure in generalising algorithms for real-life clinical usage. This thesis investigates approaches to address the insufficient amount of data, cost involved in annotations and domain shifts in prostate cancer-related tasks (such as lower-pelvic structure segmentation and registration) that could potentially benefit other medical applications facing the same problems. Specifically, Chapter 5 proposes a fully 3D prototypical few-shot segmentation algorithm for few-shot cross-institutional segmentation and Chapter 6 proposes a semi-weakly-supervised registration pipeline. Both approaches proved their efficacy on a collection of manual annotations of eight lower-pelvic anatomical structures on 589 T2-weighted MR images.

# Chapter 3

# Literature Review

This thesis explores label-efficient learning approaches towards different challenges. Specifically, two tasks with wide medical applications (segmentation and registration) are targeted and two label-efficient learning approaches (few-shot learning and semi-supervised learning) are adopted. This chapter outlines literature reviews of the two tasks (semantic segmentation in Section 3.1 and registration in Section 3.2) and two label-efficient learning approaches (in Section 3.3).

## 3.1 Image Semantic Segmentation

### 3.1.1 Overview

Semantic segmentation is the task of partitioning an image into segments where each pixel is assigned a label corresponding to a specific object class. The goal is to classify each pixel in the image into a predefined category, such as "road," "building," "tree," "person," etc. This technique is used to understand the content of an image at a pixel level, providing detailed information

about the shapes and locations of objects within the image. Semantic segmentation is widely used in various applications such as defect detection [82], self-driving vehicles [83, 84] and medical image analysis [85, 86]. Specifically, in the management of prostate cancer, segmentation of lower-pelvic structures are crucial for various prostate cancer management applications, including enabling more precise MRI-targeted TRUS-guided biopsy, facilitating PI-RADS scoring system requirements and allowing segmentation-based longitudinal analysis for active surveillance patients.

### 3.1.2 Task Definition

As both 2D (in Chapter 4) and 3D (in Chapter 5) segmentation tasks have been investigated in this thesis, the subsequent definitions are tailored for 3D scenarios but can be modified for 2D applications. For a $C$-class segmentation task, given a $\Lambda$-channel image $I \in \mathbb{R}^{\Lambda \times W \times H \times D}$ (of width $W$, height $H$ and depth $D$), the algorithm is tasked to predict a set of segmentation masks $\{M_{(c)}\}_{c=1}^{C}$ of the same spatial size $\mathbb{R}^{W \times H \times D}$. Denote the components of $I$ and $M_{(c)}$ respectively as $i_{\lambda xyz}$ and $m_{(c)xyz}$ where the subscript $\lambda = 1, ..., \Lambda$ is used as channel index whilst $x = 1, ..., W$, $y = 1, ..., H$ and $z = 1, ..., D$ are used as location coordinate indices. Each component $m_{(c)xyz}$ of a segmentation mask $M_{(c)}$ should be assigned with a value of 1 if spatial location $(x, y, z)$ of the image $I$ belongs to the class $c$, 0 otherwise, i.e. $[m_{(1)xyz}, ..., m_{(C)xyz}]^{\top}$ be a voxel-wise one-hot vector.

### 3.1.3 Deep Learning Engine for Image Semantic Segmentation

#### 3.1.1 Model Output

Generally, given an image $I$, neural networks for semantic segmentation are formulated to predict a set of probability mask predictions $\{\bar{M}_{(c)}\}_{c=1}^{C} = g(I;\theta)$ all of the same spatial size $\mathbb{R}^{W\times H\times D}$ as the image $I$. Denote components of $\bar{M}_{(c)}$ as $\bar{m}_{(c)xyz}$ where the subscript $x = 1, ..., W$, $y = 1, ..., H$ and $z = 1, ..., D$ are used as location coordinate indices. Each component $\bar{m}_{(c)xyz}$ represents the predicted probability of spatial location $(x, y, z)$ belonging to the class $c$ (therefore $\sum_{c=1}^{C} \bar{M}_{(c)xyz} = 1$).

#### 3.1.2 Dataset

A standard fully-annotated dataset $\mathcal{D} = \{(I^i, \{M^i_{(c)}\}_{c=1}^{C})\}_{i=1}^{N}$ for fully-supervised segmentation training includes $N$ pairs of image-mask pairs, where $N$ should be sufficiently big to represent the distribution.

#### 3.1.3 Training

With a model $g(\cdot;\theta)$ parameterised by $\theta$, deep-learning methods approach the segmentation tasks by training the model parameter $\theta$ on the dataset $\mathcal{D}$. Specifically, for each iteration, an image-mask pair $(I, \{M_{(c)}\}_{c=1}^{C})$ is sampled from the dataset. The model parameter $\theta$ is trained to minimise loss representing the difference between the predicted set of probability masks $\{\bar{M}_{(c)}\}_{c=1}^{C} = g(I;\theta)$ and the ground-truth set of masks $\{M^i_{(c)}\}_{c=1}^{C}$. Commonly adopted losses for semantic segmentation include cross-entropy loss and Dice loss.

**Cross-Entropy Loss** Cross-entropy loss captures the dissimilarity between two probability distributions. For a single class $c$, the single-class cross-entropy loss is measured by

$$\mathcal{L}_{\text{ce}}(\bar{M}_{(c)}, M_{(c)}) = -\mathbb{E}_{x,y,z}(m_{(c)xyz} \log \bar{m}_{(c)xyz}) \tag{3.1}$$

Multi-class cross-entropy loss is then the average of single-class cross-entropy loss across all classes:

$$\mathcal{L}_{\text{CE}}(\{\bar{M}_{(c)}\}_{c=1}^{C}, \{M_{(c)}\}_{c=1}^{C}) = \mathbb{E}_{c=1}^{C}(\mathcal{L}_{\text{ce}}(\bar{M}_{(c)}, M_{(c)})) \tag{3.2}$$

While cross-entropy loss is particularly effective when the goal is to optimise for pixel-wise classification accuracy, it is vulnerable towards class imbalance.

**Dice** Dice loss is introduced to deep-learning-based segmentation by Milletari et. al. [87] based on the Dice coefficient, and aim to balance between foreground and background voxels (pixels). For a single class $c$, the single-class cross-entropy is measured by

$$\mathcal{L}_{\text{Dice}}(\bar{M}_{(c)}, M_{(c)}) = -\frac{1}{WHD} \sum_{c=1}^{C} \frac{2 \sum_{x=1}^{W} \sum_{y=1}^{H} \sum_{z=1}^{D} m_{(c)xyz} \bar{m}_{(c)xyz}}{\sum_{x=1}^{W} \sum_{y=1}^{H} \sum_{z=1}^{D} m_{(c)xyz} + \bar{m}_{(c)xyz}} \tag{3.3}$$

Multi-class Dice loss is then the average of single-class Dice loss across all classes:

$$\mathcal{L}_{\text{Dice}}(\{\bar{M}_{(c)}\}_{c=1}^{C}, \{M_{(c)}\}_{c=1}^{C}) = \mathbb{E}_{c=1}^{C} \mathcal{L}_{\text{Dice}}(\bar{M}_{(c)}, M_{(c)}) \tag{3.4}$$

As the overlap (true-positive) is weighted twice compare with the false-positive and false-negative predictions, Dice loss is more robust to class imbalance in cases where the segmentation masks are sparse or have small object sizes. It is also useful when dealing with imbalanced datasets or when the

segmentation task requires more precise delineation of object boundaries.

### 3.1.4 Evaluation

The set of probability mask predictions $\{\bar{M}_{(c)}\}_{c=1}^{C}$ could be transformed into a set of binary segmentation mask predictions $\{\hat{M}_{(c)}\}_{c=1}^{C}$ by assigning each voxel $(x, y, z)$ to the class with highest predicted probability. Denote components of $\hat{M}_{(c)}$ as $\hat{m}_{(c)xyz}$ where the subscript $x = 1, ..., W$, $y = 1, ..., H$ and $z = 1, ..., D$ are used as location coordinate indices, $\hat{m}_{(c)xyz}$ is assigned a value of 1 if $c = \arg\max_c \bar{m}_{(c)xyz}$, 0 otherwise.

The performance is then measured by comparing the predicted and ground-truth segmentation masks, $\{\hat{M}_{(c)}\}_{c=1}^{C}$ and $\{M_{(c)}\}_{c=1}^{C}$. Frequently adopted measurements include intersection of union (IoU), Dice score and Hausdorff distance.

**Intersection of Union (IoU)** IoU measures the ratio of the intersection of two masks to their combined area. For single-class (binary) masks

$$f_{\text{iou}}(\hat{M}_{(c)}, M_{(c)}) = \frac{\sum_{x=1}^{W}\sum_{y=1}^{H}\sum_{z=1}^{D}\hat{m}_{(c)xyz}m_{(c)xyz}}{\sum_{x=1}^{W}\sum_{y=1}^{H}\sum_{z=1}^{D}\hat{m}_{(c)xyz} + m_{(c)xyz} - \hat{m}_{(c)xyz}m_{(c)xyz}} \tag{3.5}$$

When multiple classes are involved, mean IoU is then the average of single-class IoU across all classes.

$$f_{\text{IoU}}(\{\hat{M}_{(c)}\}_{c=1}^{C}, \{M_{(c)}\}_{c=1}^{C}) = \mathbb{E}_{c=1}^{C}(f_{\text{iou}}(\hat{M}_{(c)}, M_{(c)})) \tag{3.6}$$

**Dice Score (Coefficient)** Instead of measuring the ratio of the intersection of two masks to their combined area, Dice score measures the ratio of

the intersection to the sum of areas. For single-class (binary) masks

$$f_{\text{Dice}}(\hat{M}_{(c)}, M_{(c)}) = \frac{1}{WHD} \sum_{c=1}^{C} \frac{2 \sum_{x=1}^{W} \sum_{y=1}^{H} \sum_{z=1}^{D} m_{(c)xyz} \hat{m}_{(c)xyz}}{\sum_{x=1}^{W} \sum_{y=1}^{H} \sum_{z=1}^{D} m_{(c)xyz} + \hat{m}_{(c)xyz}} \tag{3.7}$$

When multiple classes are involved, multi-class Dice score is then the average of single-class Dice score across all classes.

$$f_{\text{Dice}}(\{\hat{M}_{(c)}\}_{c=1}^{C}, \{M_{(c)}\}_{c=1}^{C}) = \mathbb{E}_{c=1}^{C}(\mathcal{L}_{\text{Dice}}(\hat{M}_{(c)}, M_{(c)})) \tag{3.8}$$

**95% Hausdorff Distance** Max Hausdorff distance measures the maximum distance of one set to the nearest point in another. On this basis, 95% Hausdorff distance restricts the calculation into the 95% percentile of the distances to eliminate the impact of a very small set of outliers.

### 3.1.4 Literature Review

In this section, we explore important works in the field of semantic segmentation. We begin by a concise overview of classical methods, which form the foundation of semantic segmentation, encompassing region-based, edge-based, and model-based techniques. Following this, we undertake a more detailed examination of deep learning methods, with a particular focus on the widely adopted Fully Convolutional Networks (FCN) architecture. Within this domain, we explore innovations to address two prominent challenges: the generation of high-resolution outputs constrained by memory limitations and the segmentation of multi-scale objects. Additionally, we look into the domain-specific intricacies of medical imaging segmentation for its unique challenges and specialised solutions.

### 3.1.1 Classical Methods

Early algorithms for semantic segmentation could be divided into three major groups: region-based methods, edge-based methods and model-based methods.

Region-based methods, including thresholding, region growth, region split & merge, watershed and clustering, rely on the similarity/homogeneity of pixels belonging to the same meaningful part of an image. Uniformity criteria such as intensity, colour, distance and texture are used to define regions. Thresholding generates binary output by comparing pixels/voxels with a threshold value. Popular strategies to determine threshold value include maximum entropy method [88], Otsu's threshold algorithm [89] and k-means clustering [90]. Thresholding algorithms perform well on images with sharp edges but are vulnerable to noise and obscure boundaries. Region growth algorithms start from a specific pixel/voxel and expand the region based on the connectivity with the neighbouring cells measured by similarity criteria [91, 92]. Region split and merge [93] approaches, on the other hand, first successively split images into quadrants based on homogeneous criteria and then merge divided regions. They are therefore less sensitive to the choice of seed point and suppress the high-frequency artifacts. Watershed methods [94, 95, 96, 97] view the image as a topographic surface with the intensity of each pixel representing its height and find watershed lines as segmentation. Standard watershed algorithms are vulnerable to over-segmentation when the signal-to-noise ratio is low, which could be mitigated by filtering unrelated neighbourhood lowest [94, 95]. Clustering gathers homogeneous data points into clusters. While k-means clustering hardly assigns each data point to a single cluster each time, the soft clustering approach adopts fuzzy

community function to assign a pixel to multiple groups [98]. The objective function could be altered according to task-dependent necessities for various applications [99, 100, 101, 102, 103].

Edge-based methods [104], on the other hand, rely on discontinuity to define boundaries between separate regions. Firstly, edge points are detected by finding abrupt changes or discontinuities. Edge points are mostly detected through convolution with filters for finding higher gradients (Prewitt [105], Sobel [106] and watershed [107]) or zero crossings in the second-order derivatives (Laplacian [108]). Secondly, closed-form boundaries are deducted from edges with algorithms such as active contour. Edge detection is vulnerable to noise, resulting in inadequate or irregular edges due to the comprising of discrete pixels. To this end, Gaussian smoothing is often applied before edge detection. However, the preprocessing itself may lead to fallacious edge detection and hinder the recognition of multi-resolution edges.

Model-based algorithms contain prior information such as shape and existence with models and are therefore sturdier against artifacts compared with conventional algorithms. Markov random field models [109, 110] model each pixel with a node with neighbouring property represented by edges and label the pixel with the class that maximises the joint probability. [110] proposed HMRF-EM framework which combines hidden Markov random field model and expectation maximisation. Other variations include combining Markov random field with techniques such as self-organising feature map [111]. Atlas-based approaches are usually adopted for medical image segmentation where the images to be segmented are topologically similar. Such approaches propagate labels from an atlas to the image to be partitioned through spatial correspondence. Here, an atlas is defined as the combination of an intensity image (template) and its segmented image. It could be a single-subject

atlas [112], which is often a volume image that has been selected from a data set to be representative of the objects to be segmented in other images, or a probabilistic atlas [113, 114] constructed on the basis of population to characterise the variability of anatomical structures and could easily incorporate new images for evolution. Graph-cut approaches [115] perform binary segmentation by solving a maximum flow (minimum energy) problem and defining a minimal cut of the graph. Multi-label segmentation could be enabled by applying the graph-cut iteratively to a sequence of binary problems as an approximation solution. Graph-cut approaches can exhibit unwanted "blockiness" artifacts and be biased toward producing a small contour.

For the above methods, there is no need of large datasets of example segmented images. Although machine learning methodology had been widely applied in the early classical segmentation methods, but they usually consider pixels/voxels within individual images as the data and model distributions between pixels/voxels, rather than between segmented images.

### 3.1.2 Deep-learning image segmentation

Since around the mid-2010s, deep neural networks have been adopted for computer vision tasks including semantic segmentation. A neural network mimics human brain operation to recognise underlying relationships from a set of data. Specifically, it consists of layers of interconnected nodes called neurons. Each node is a mathematical function which performs multiple linear regression followed by a nonlinear activation function. Given a defined loss function, network parameters are optimised to reduce the loss through backpropagation during training.

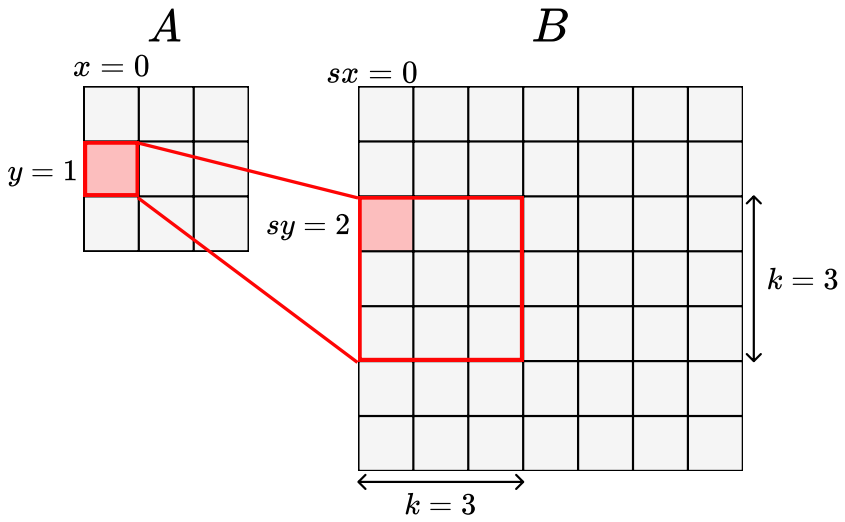



Figure 3.1: An example of stride 2 convolution with kernal size of 3 at location $x = 0, y = 1$

#### 3.1.2.1 Fully Convolutional Networks

[116] first proposed to adopt fully convolutional networks for image segmentation. Fully convolutional networks, with no fully connected layers, are made of convolution, pooling and activation functions, components that operate on local input regions and depend only on relative spatial coordinates. Specifically, denote $A_{x,y}$ and $B_{x,y}$ respectively as the data vector at location $(x, y)$ of the input and output matrix, these functions compute $B_{xy}$ by

$$A_{x,y} = f_{k,s}(\{B_{sx+\delta x, sy+\delta y}\}_{0 \leq \delta x, \delta y \leq k}) \tag{3.9}$$

where $k$ and $s$ refers to the kernal size and stride. 3.1 gives an example when $s = 2$ and $k = 3$.

This functional form is maintained under composition, with kernel size and

stride obeying the transformation rule

$$f_{k,s} \circ g_{k',s'} = (f \circ g)_{k'+(k-1)s',ss'} \tag{3.10}$$

Therefore, fully convolutional networks could operate on inputs of variant sizes and produce outputs of corresponding spatial dimensions.

ImageNet pretrained classifiers could be converted to fully convolutional networks by replacing fully connected layers with convolutions, and upsampling components such as unpooling and deconvolution could be adopted to recover the resolution for pixel-wise prediction. However, there are two major challenges in the application of fully convolutional networks to semantic segmentation - reduced feature resolution due to computational cost and existence of objects at multiple scales

#### 3.1.2.2 Reduced Resolution

In order to capture global and semantically meaningful high-level context, field-of-view and feature dimensions are increased as the network goes deeper and lead to rising computational cost which is often compensated by reducing the spatial size with subsampling through down-pooling and strided convolution. Though improved computational efficiency, subsampling reduces the prediction resolutions and coarsens the output.

To this end, the encoder-decoder structure is widely adopted for efficient use of computational power such that high-level features could be extracted without sacrificing output resolution - the encoder downsamples feature maps while extracting deep features with increasing number of channels and the decoder recovers the resolution through upsampling. [117] introduced DeConvNet, consisting two parts, an encoder adapted from VGG-16 backbone and

a decoder comprising deconvolution and unpooling layers which upsample the encoded feature vector to pixel-wise segmentation prediction. Moreover, local information could be passed from encoder blocks to decoder blocks at the same spatial scale to enable more precised prediction. For instance, [118] kept pooling indices computed in the max-pooling step for the corresponding encoder in SegNet to perform non-linear upsampling. [119] utilised skip connections between corresponding encoder and decoder blocks to utilise high resolution features during upsampling.

Instead of recovering high-resolution representations through upsampling, [120] proposed HRNet which maintains high-resolution representations through the encoding process. HRNet connects high-to-low-resolution convolutions in parallel and produces strong high-resolution representation by repeatedly conducting fusion across parallel convolutions.

Another approach for reducing computational cost is replacing large kernels with atrous convolutions which uses strides inside the kernel windows to achieve larger field-of-view with fewer parameters. DeepLab v1 [121] adopted atrous convolution to capture global context without pooling and therefore reduction in resolution. DeepLab v2 [122] further combined atrous convolution with spatial pyramid pooling to extract multi-scale features with filters at different sampling ratio.

#### 3.1.2.3 Multiscale Objects

While translational invariant, fully convolutional networks are not scale invariant. Challenges are caused by the existence of objects at multiple scales, motivating multiscale analysis to be deployed in various neural network architectures.

[123] proposed a multi-resolution reconstruction architecture using a low-resolution feature map to reconstruct a low-frequency segmentation map and refine it successively by adding higher-frequency details derived from higher-resolution feature maps.

Spatial pyramid pooling, which aggregates local features from divisions partitioned from fine-to-coarse levels, was one of the most successful classical methods. [124] introduced PSPNet with an encoder followed by a pyramid pooling module, which pools the encoded feature at different scales before up-sampling and concatenation to harvest different sub-region representations. DeepLab [122] further proposed to combine pyramid pooling with atrous convolution as atrous spatial pyramid pooling (ASPP), with multiple parallel atrous convolution layers with various dilation rates which capture objects as well as useful image context at multiple scales.

Featurised image pyramid, which computes features for various scales independently is another approach commonly adopted by classical methods for extracting hand-crafted features such as SIFT and HOG. [125] proposed to adapt and accelerate featurised image pyramid to deep-learning by introducing the Feature Pyramid Network naturally leveraged the pyramidal shape of convolutional neural networks. Feature pyramid network constructs feature pyramids including high-level semantic feature maps at all scales using a top-down architecture with lateral connections. Attention mechanism could be included to weight features at different scales. [126] softly weight the multi-scale features at each pixel location with attention. [127] introduced a feature pyramid attention module and a global attention upsample module to respectively generate attention for high-level and low-level features under the guidance of different scale context information and high-level features.

#### 3.1.2.4 Medical Image Segmentation

Segmentation is one of the fundamental computing tasks in medical imaging applications, found in diagnosis, treatment, and monitoring, and remains a research interest. Diagnosis of varieties of diseases can be assisted by quantifying the morphology, or its change, of multiple structures. For example, brain disorders, including Alzheimer's disease [128] and Parkinson [129], associates with abnormal volumes or shapes of neurological regions. Identifying brain structures is the key to many quantitative studies, such as functional activation mapping and brain development analysis [130]. Minimally invasive treatments often benefit from careful planning of both interventional instruments and guidance imaging, with respect to segmented patient-specific anatomical structures. In endoscopic pancreatobiliary procedures, image guidance that displays registered anatomical models outside the endoscopic field of view helps the surgeon during targeting and navigation [131].

Two most important architectures adopted by medical image segmentation are UNet [119] and VNet [87]. UNet [119] consists a contracting path to capture context and a symmetric expanding path to enable precise localisation, yielding a u-shaped architecture. Skip connections combine high resolution features from the contracting path with the upsampled output. VNet [87] also adopts symmetric encoder-decoder structure but replace pooling with strided convolutions. Many variant have been developed on this basis. 2D networks could realise 3D image segmentation by stacking continuous 2D slice predictions of a 3D image but fail to probe spatial information along the third dimension. 3D UNet [132] replaces 2D blocks with 3D blocks for volumetric segmentation. H-DenseUNet [133] introduced a hybrid approach with a 2D DenseUNet for intra-slice feature extraction and a 3D counterpart

for volumetric context aggregation. Both 2D and 3D networks are optimised jointly through a hybrid feature fusion layer. [134] redesigned skip pathways of UNet with nested dense convolutional blocks to reduce the semantic gap between the feature maps of the encoder and decoder sub-networks. [135] proposed Res-UNet by adding skip connection scheme to increase the depth of the network and weighted attention mechanism to concentrate at the target ROI area. [136] introduced attention UNet by using attention gates to highlight salient features and focus on target structures. [137] incorporated global aggregation blocks based on self-attention for downsampling and upsampling. nnU-Net [86] configures itself by modelling key design choices (including preprocessing, network architecture, training and post-processing) as a set of fixed parameters.

Different anatomical organs and structures have various features, so specific segmentation algorithms are designed by the target organ structure. For example, Res-UNet, designed for retinal vessel segmentation, adopted a weighted attention mechanism to focus on the target ROI and discard irrelevant noise to solve the great influence of noise on the vessel's shape.

### 3.1.3 Summary

Deep-learning image segmentation methods have achieved superior performance over classical methods in both natural and medical images. Some of these methods have reached performance levels comparable to human experts in clinical tasks such as MRI cardiac multi-structure segmentation [138], CT liver and lesion segmentation [139], and histopathology image segmentation [54]. Prostate segmentation is considered more challenging due to the high variability of prostate shape across patients and imaging modalities. Another widely recognized obstacle in prostate segmentation is the

lack of publicly available large annotated datasets. Additionally, existing open-source datasets have limitations, including selection bias, limited annotations, and a lack of demographic statistics [140, 141, 142].

## 3.2 Image Registration

### 3.2.1 Overview

Image registration is the process of spatially transforming two or more different images into one coordinate system with matched imaging contents. The correspondence establishing the problem of determining what a good match is can be application-dependent and, for some applications, still remain an open question.

In the medical imaging research area, the images in question can be either 2D, 3D, 4D or higher dimensional. Given a moving and a fixed image, the model is tasked to predict the correspondence from the moving image to the fixed image. Such correspondence can be useful in many clinical applications. Inter-subject image registration is the basis of the atlas-based segmentation algorithms [143] and population studies that are important for epidemiology research [144]. Intra-subject longitudinal image registration can track suspicious pathological regions for individuals by aligning images acquired from different time points [145]. Registering images from different modalities can combine complementary diagnostic information from different imaging sources to make informed clinical decisions or guide surgery and interventions [146].

Registration algorithms have been developed and adopted by many multidisciplinary researchers. Stemmed from computer vision algorithms such as optical flow and pose estimation, it proves its usefulness in medical image analysis.

### 3.2.2 Task Definition

Pairwise registration involves aligning one image (referred to as the moving image) with another (referred to as the fixed image). Given a pair of moving image $I^{\text{mov}} \in \mathbb{R}^{\Lambda \times W \times H \times D}$ and fixed image $I^{\text{fix}} \in \mathbb{R}^{\Lambda \times W \times H \times D}$, pairwise registration spatially aligns the moving image towards the fixed image. $\Lambda$ indicates the number of image channels. $W$, $H$ and $D$ indicates the width, height and depth of the 3D volumetric images respectively.

### 3.2.3 Literature Review

In this section, we undertake a review of significant contributions in medical registration, spanning from early iterative image registration methods to the latest advancements in deep learning-based approaches. We start from the initial developments of deep similarity-based registration and the emergence of reinforcement learning-based registration. Then, we detail the more prevalent single-step registration methods in state-of-the-art researches, including fully supervised, weakly supervised, and unsupervised techniques.

#### 3.2.1 Iterative image registration

In the past three decades, "classical" non-learning image registration algorithms have converged to a so-called three-component framework, where the pairwise registration is formulated as an *optimisation* problem of a parametric *transformation model*, where an image *similarity measure* is maximised between a fixed image and the transformed moving image. Early works optimise transformation parameters iteratively using numerical optimisers, such as those based on gradients or higher-order derivatives, with gradient-free methods such as simplex being also an option [147]. Group-wise registration

between three or more images is largely based on pairwise image registration algorithms. We outline the transformation models and similarity measures in the following sections.

#### 3.2.1.1 Transformation model

Transformations used for registration are often classified as rigid and non-rigid transformation.

**Rigid Transformation** The six-degree-of-freedom rigid transformation and its generalisation affine transformation represent very constrained spatial differences such as those caused by rotation, translation and scaling, and can be parameterised by a linear matrix multiplication. In addition, aligning 2D and 3D images may require a projective transformation. Rigid transformations are mainly applied for registration of rigid structures such as bones [148, 149, 150, 151] or pre-registration before more complex geometric transformation, as outlined below.

**Non-Rigid Transformation** Due to the deformability of anatomical parts, most medical image registration approaches adopt non-rigid transformation which can be divided into two types: free-form transformations and guided deformations.

Free-form deformation models enable any form of deformation [152]. Predefined control points are moved individually to optimise similarity measures. Transformations between control points can then be propagated from local deformations computed at control points through interpolators such as cubic B-spline.

Guided deformation models, on the other hand, constrain deformation by

physical models based on material properties. Elastic models assume registration targets as elastic solids and seek equilibrium between internal elastic forces that oppose deformation and external forces driven by similarity measures [153]. Thin-plate splines (TPS) based registration methods rely on deformable solid properties [154]. Like free-form deformation models, control points are defined to compute local deformations. Differently, transformations of neighbours are interpolated from local deformations through a thin-plate model which minimises the bending energy globally. In other words, the movement of a control point affects the entire deformation. Flow-based models, such as the demons algorithm, address registration as a motion problem - the content of an image moves continually towards another to minimise the energy of the adopted physical model [155].

#### 3.2.1.2 Similarity Measures

Similarity can be measured directly on intensities or features such as organ segmentation and landmarks. Simple sum-of-squared-differences (SSD) or, equivalently, mean-square-error (MSE) is frequently used to measure how well two images are aligned if both are acquired from the same type of scanner and using the same imaging protocol. The difference/error can be intensity difference [156] or spatial distance between corresponding structures [157]. More robust similarity measures dealing with multimodal image registration and noise-corrupted data have also been widely proposed, derived from statistics, information theory and imaging physics, such as those based on cross correlation [158] and mutual information [159].

A regularisation term computed on the deformation field is commonly added to the cost function as a trade-off between similarity and deformation constraint [160] for deformable image registration methods.

### 3.2.2 Deep-learning image registration

Recent development in deep-learning have inspired the investigation towards deep-learning-based registration methods.

#### 3.2.2.1 Deep Similarity-based Registration

Early methods inherit the iterative registration framework and replace hand-crafted a similarity metrics with neural network-learned features. [161] first proposed to learn basis filters extracting morphological signatures through unsupervised learning of a convolutional-stacked autoencoder to replace hand-crafted features for unimodal deformable 3D brain MR volumes registration. [162] and [163] adopted a CNN to learn dissimilarity metric from aligned and perturbed image pairs respectively for 3D T1/T2 brain MR volume registration and rigid transrectal US/MR registration. [164] proposed symmetrisation and dithering to correct the bias induced by misalignment and enforce single mode metric such that roughly aligned training data can also be used for metric learning.

#### 3.2.2.2 Reinforcement Learning-based Registration

Another stream of methods adopted reinforcement learning, replacing the pre-defined optimisation algorithm with a trained agent. [165] first proposed a greedy supervised approach for end-to-end training, coupled with an attention-driven hierarchical strategy for rigid registration of cardiac and abdominal 3D CT images and cone-beam CT (CBCT) images. [166] proposed an algorithm derived from Q-learning leveraging contextual information of medical images to resolve data ambiguities and improve robustness. [167] adopted a multi-agent system with an auto-attention mechanism for robust and efficient rigid spine x-ray/CT registration. The above reinforcement

learning-based registration typically involves a rigid transformation model to avoid the large number of possible actions involved in high-resolution deformation fields. [168] proposed to use a low-dimensional parametric statistical deformation model to minimise the possible actions for organ-focused non-rigid registration.

#### 3.2.2.3 Single-step Registration

Despite the performance improvement achieved by deep iterative registration and reinforcement learning approaches, the iterative transformation estimation leads to slow registration speed, especially for deformable transformation prediction with high-dimensional solution space. This motivated the development of neural networks to directly predict transformation in a single step.

Single-step registration methods can be divided into fully-supervised, unsupervised and weakly-supervised transformation estimation based on the availability of ground-truth transformation and label guidance.

**Fully-supervised approaches** are trained to minimise the difference between predicted and ground-truth transformation parameters. Such training requires image pairs with known transformations which in practice come either from the traditional/manual registration-generated transformation [169, 170, 171, 172] or from synthetic transformations [173, 174, 175, 176, 177, 178, 179, 180, 181].

[169] adopted a CNN to rigidly register 2D/3D X-ray attenuation maps and 2D X-ray images through hierarchical registration partitioning six transformation parameters into three groups. Ground truth transformation parameters are generated by a combination of manual registration followed by an intensity-based registration method using Powell's method combined with

gradient correlation. [176], [179] and [180] trained rigid brain MR volumes registration with image pairs consisting of an original image and its transformed version after a randomly sampled transformation.

A more complicated process is involved in generating ground truth for deformable registration. [170] adopted a UNet-like architecture to predict deformation field registering 2D/3D intersubject brain MR volumes. Unbiased atlases are created using Python Cellular Automata library (PyCA) as moving images and registered to all training images using large deformations diffeomorphic metric mapping (LDDMM) [182] shooting combined with a sum-of-squared intensity difference in PyCA as ground-truth transformation. [171] took traditional registration-generated transformation from previously segmented regions of interest instead of images as reference deformation to train a UNet-inspired network for 3D cardiac MR volume registration, encouraging the alignment of more global and structural information. [172] designed a CNN in a patch-based manner, predicting respective patch-pair displacement vectors which can be used to derive a whole-image deformation field though thin-plate spline (TPS) interpolation in an iterative manner. The ground-truth deformation field is generated by a two-stage registration with SyN [158] followed by diffeomorphic demons [183]. The quality of traditional registration-generated transformation is limited by the performance of the traditional registration algorithm. Ground-truth simulation approaches, though more challenging for deformable transformations, are motivated by a desire to enhance the size and diversity of datasets. Both [174] and [177] trained 3D CNNs for deformable intra-subject 3D lung CT registration with synthetically deformed clinical images. [174] generated synthetic DVFs with varying spatial frequency and amplitude. [177] combined multiple transformations at different scales to form realistic deformation without folding and

tearing. [175] learned a statistical appearance model from training images to train a FlowNet estimating the deformation field for the registration of 2D brain MRs and 2D cardiac MRs. [178] obtained small seed sets including deformation vector fields with a conventional registration method after manual validation to synthesise image pairs for 3d brain MR volume registration.

While supervised training of one-step registration models enabled efficient and robust predictions, their performances are strongly correlated to the quality of ground-truth training examples. The quality of traditional registration-generated transformation is constrained by the upper bound determined by the underlying traditional registration algorithm. Manual registration-generated transformations are labour expensive to acquire due to the demanding expertise required. Synthetic ground truths are accurate, but the simulated data may not faithful reflect real-world clinical data, particularly for deformable transformation models. These limitations inspired works investigating unsupervised and weakly-supervised approaches as described below.

**Unsupervised approaches** Provided similar or *corresponding* anatomical structures present in both moving and fixed images, the image registration task can be approached when no labels are provided by utilising image similarity metrics between warped and fixed images and, sometimes, regularisation terms for deformable models.

Commonly adopted image similarity measures include normalised cross-correlation and mean square error. [184] and [185] respectively trained a FCN and a CNN for deformable inter-subject registration of 3D brain MR volumes by minimising normalised cross correlation between the warped and fixed image. [186] also adopted normalised cross-correlation as its loss function, for

training a FCN to predict dense velocity field for inter-subject 4D cardiac cine-MR volume registration. [187] explored the adaptability of a CNN-based image registration model trained with normalised cross-correlation loss from a source domain to unseen target domains. [188] registered neuron images acquired with a scanning electron microscope through a coarse-to-fine approach with the mean square error between the warped and fixed volumes as the loss function. [189], minimising mean square error, both linearly and locally register inhale-exhale pairs of lung MR volumes. Similarly, [190] adopted a variational inference approach for registering T1-weighted brain MRI scans by minimising mean square error. There are also other similarity measures available. For instance, [191] proposed an unsupervised deformable 3D brain MR image registration framework adopting the upper bound of the sum of the squared difference between the warped and fixed images as the loss function. [192] took latent representations extracted by an unsupervised-trained autoencoder as the input of an FCN for deformable 2D cardiac cine-MR volume registration. The loss function is the sum of absolute errors between the warped and fixed images.

However, the difficulty in quantifying image similarity for images across different modalities limits the usage of image similarity-based works. For instance, anatomically distinct imaging structures, such as the inner-outer gland separation in the prostate observable on TRUS image and the central-peripheral zonal boundary observable on MR images, appear similar. This similarity results in a strong statistical correlation between them, often leading to false alignment when using image similarity measures [193].

**Weakly-supervised approaches** This challenge can be approached by introducing high-level information shared by images of different modalities.

When auxiliary information like segmentation maps and landmarks are available, they can be used to guide registration as weak supervision. The networks are encouraged to maximise the similarity between warped labels and fixed labels. Such weak supervision enables similarity quantification in multimodal cases.

For instance, [74, 73] trained lower-pelvic MR-TRUS multimodal registration under the guidance of full gland segmentation, structure-specific landmarks (including apex, base, urethra, visible lesions, junctions between the gland, vas deference and the seminal vesicles) and many other patient-specific point landmarks (such as calcifications and fluid-filled). [194] supervised cardiac motion tracking network with annotations for left and right ventricle cavity and myocardium to register MR images captured at end-diastole (ED) and end-systole (ES) time points. Similarly, VoxelMorph [195] guided the registration of T1-weighted brain MRI with structures segmented with FreeSurfer with at least 100 voxels in volume.

Despite addressing the challenges related to the reliability and costliness of ground truth labels, weakly supervised transformation estimation methods still demand a considerable volume of manually annotated data. The exploration of weakly supervised registration with a restricted number of annotated examples persists as an unresolved issue in current research.

**Transformations** Commonly adopted transformation models by deep-learning approaches include affine transformation matrix and dense displacement field (DDF).

Denote the components of the moving image $I^{\text{mov}}$ and fixed images $I^{\text{fix}}$ respectively as $i^{\text{mov}}_{xyz}$ and $i^{\text{fix}}_{xyz}$, where the subscript $\lambda = 1, ..., \Lambda$ is are used as channel index whilst $x = 1, ..., W$, $y = 1, ..., H$ and $z = 1, ..., D$ are used as

location coordinate indices. Therefore $i^{\text{mov}}_{xyz}$ and $i^{\text{fix}}_{xyz}$ respectively represents the channel $\lambda$ intensity at spatial location $(x, y, z)$ of $I^{\text{mov}}$ and $I^{\text{fix}}$ respectively. The most commonly adopted sets of transformation parameters can then be defined as outlined in the following sections.

### 3.2.3 Affine Transformation Matrix

The affine transformations which include translation, scaling, reflection, rotation, shear mapping, and their compositions - preserves lines and parallelism, but not necessarily Euclidean distances and angles. It can be represented with a transform matrix mapping the original spatial coordinates to new coordinates. For instance, multi-channel 3D moving image $I^{\text{mov}} \in \mathbb{R}^{\Lambda \times W \times H \times D}$ can be affine transformed into the warped image $I^{\text{warp}} \in \mathbb{R}^{\Lambda \times W' \times H' \times D'}$ with a transformation matrix $\tau \in \mathbb{R}^{3 \times 4}$

$$I^{\text{warp}} = I^{\text{mov}} \diamond \tau \tag{3.11}$$

where, the affine operation is denoted by "$\diamond$", transforming the left image (or mask) $I^{\text{mov}}$ using the right transformation matrix $\tau$, such that the spatial location $(x, y, z)$ in the original image $I^{\text{mov}}$ are projected to spatial location $(x', y', z')$ in the transformed image where

$$[x', y', z', 1]^\top = \tau [x, y, z, 1]^\top \tag{3.12}$$

### 3.2.4 Dense Displacement Field

A dense displacement field can be used to represent any kind of transformation without constraints and should be of the same spatial size $U \in \mathbb{R}^{3 \times W \times H \times D}$ as the fixed image $I^{fix} \in \mathbb{R}^{\Lambda \times W \times H \times D}$. The DDF components

$u_{dxyz}$, where $d \in \{1, 2, 3\}$, and $u_{1xyz}$, $u_{2xyz}$ and $u_{3xyz}$ indicates the $x$-, $y$- and $z$-displacement at voxel location $(x, y, z)$.

The DDF $U$ can be used to resample the moving image into a warped image

$$I_{(c)}^{\text{warp}} = I_{(c)}^{\text{mov}} \circ U \tag{3.13}$$

where, the image resampling operation is denoted by "$\circ$", resampling the left image (or mask) $I^{\text{mov}}$ using the right DDF $U$, such that the indexed values $i_{\lambda xyz}^{\text{warp}}$ in channel $\lambda$ at its $(x, y, z)$ resampled voxel locations are spatially-interpolated from the $(x+u_{1xyz}, y+u_{2xyz}, z+u_{3xyz})$ locations from the original image $I_{\lambda xyz}^{\text{mov}}$ coordinates. In practice, the inverse fixed-to-moving DDF is implemented, for resampling at regular fixed image grid locations avoiding numerical inverting.

### 3.2.5 Summary

While classical methods treat image registration as an optimisation without using prior knowledge from other image pairs, deep-learning-based transformation predictions improved performance by treating registration as a machine-learning problem relying on the training from a substantial number of example pairs. Methods uplifting data requirements are worth investigating for adapting deep-learning approaches to a wider range of applications. To this end, Chapter 6 investigates a novel semi-weakly-supervised registration approach, uses labelled image pairs to improve model performance, when only a small corresponding-ROI-labelled dataset is available.

## 3.3 Label-efficient learning

### 3.3.1 Overview

Recently label-efficient learning approaches aiming to achieve high performance with fewer labels have received attention for their ability to alleviate annotation and potentially data collection costs. Transfer learning repurposes pretrained model to related tasks. Active learning dynamically and incrementally acquires annotations for examples deemed to be informative by the model. Few-shot learning learns from a very limited number of annotated examples. Semi-supervised learning adopts prior information from large-sized unlabelled datasets for better learning performance on small-sized labelled datasets. This section focus on few-shot learning and semi-supervision, describing their major works and commonly adopted approaches.

### 3.3.2 Few-shot learning

Few-shot learning is a machine learning paradigm to learn from a limited number of examples with supervised information [196, 197, 198]. Specifically, a $K$-shot (few-shot) learning problem tasks a model to learn a *novel task* from $K$ examples. Note that though a limited number of examples are available for the novel task, sufficient samples of *base tasks* (different from the novel task) could still be available for training. For clarity, we denote the $K$ examples for the novel task as a *support set* and the sufficient samples of base tasks as the *base set*.

Few-shot learning enables learning for rare cases where supervised information is hard if not impossible to obtain (for instance rare pathological appearance only observed in rare instances as mentioned in chapter 2), as well

as when it is desirable to reduce data acquisition effort and computational costs.

While in standard supervised learning sufficient samples provide a good approximation of test data distribution, the small support set encountered by few-shot learning problems may lead to overfitting issues. To alleviate this problem, few-shot learning methods utilised prior knowledge to enhance learning from various perspectives. Such prior knowledge could be employed through multiple approaches, with the most prominent categories including data-based, model-based and algorithm-based approaches.

#### 3.3.0.1 Data-based few-shot learning

Data-based few-shot learning methods augment the support set with augmented support samples [199, 200, 201, 202], pseudo-labelled candidates from weakly-labelled or unlabelled datasets [203, 204] or aggregated and adapted samples from a similar dataset [205, 206]. Support samples could be augmented with prior knowledge of intra-task variability learnt from similar (base) tasks. Samples with the target label could be selected from weakly-labelled or unlabelled datasets through criteria such as label propagation [204] and classifiers learnt from the support set alone [203] or in addition to pseudo-labelled candidates selected through progressive sampling strategy. Samples from a dataset with similar (base) labels could be aggregated with weights based on similarity measures [205] and adapted through generative adversarial networks (GANs) trained to map samples from base classes to target classes [206].

One weakness of data-based approaches is their limited generalisability across domains as augmentation policies are often specifically tailored for the support set or training dataset in an ad-hoc manner. Moreover, both selecting as

well as aggregating and adapting samples occupy significant memory during inference for the storage of large (base) datasets.

#### 3.3.0.2 Model-based few-shot learning

Model-based few-shot learning methods constrain hypothesis (algorithm) complexity for a smaller hypothesis space and could be divided into multitask-learning, embedding learning, memory-based and generative modelling methods.

**Multitask learning** methods introduce auxiliary tasks to exploit task-generic information. Specifically, the parameters of each task are constrained by other tasks through parameter sharing or parameter tying. *Parameter sharing* shares part of the parameters from layers where similar features are expected among tasks. Early layers of the network [207] or embedding functions [208] are shared for tasks with inputs from the same domain (e.g. when charge prediction and legal attribute prediction are adopted as few-shot and auxiliary tasks, inputs of both tasks are criminal case descriptions.) On the other hand, autoencoders after initial feature encoding are shared for few-shot domain translation tasks [209, 210]. *Parameter tying* encourages parameters of different tasks to be similar, popularly through regularisation such as penalising pairwise task parameter differences [211] and other specifically designed regularisation terms [212].

**Embedding learning** approaches embed samples in a lower-dimensional space where similar samples are close together such that a sample can be classified based on its distance from support samples in the lower-dimensional space. Embedding learning methods could be described as task-specific, task-invariant and hybrid methods which respectively learn embedding spaces

from the support set, the base set and a combination of both. Task-invariant methods have speed advantages compared to most other methods as the trained network could be directly applied to the support set without further adaptation or training. Siamese networks, with shared weights between two pathways taking different inputs, is a commonly adopted architecture for embedding learning methods [213]. [214] first proposed *episodic training* paradigm to simulate the few-shot scenario while training - for a $K$-shot problem, each episode consists of a support set formed by $K$ samples of a sampled base task and a *query* sample of the same base task. Given the support set as examples, the model is optimised to maximise performance on the query sample. In $K$-shot scenarios when $K > 1$, instead of comparing the query with each of the support samples, Snell et al. [215] achieved more stable results while reducing computational cost by representing each class with a single prototype feature vector.

**Memory-based** methods extract knowledge from the support set and store it in an external memory. A new sample is then represented by a weighted average of contents extracted from the memory, limiting the hypothesis space. A key-value memory [216] is commonly adopted, i.e. each memory slot consists of a key-value pair. Both the writing of support samples into the memory and the reading of query samples from the memory weight memory slots by the similarity between the embedded sample and slot keys, make the quality of key-value pairs in the memory important.

**Generative modelling** methods estimate the probability distribution of novel classes from support samples with the help of prior knowledge. Specifically, generative modelling methods assume the existence of a latent variable whose distribution could be learned from other datasets. This latent variable

could represent manually chosen decomposable components [198, 217, 218], group-wise (with unsupervised grouping) shared prior [219, 220] or machine-learned representations [221, 222, 223, 224, 225, 226, 227].

#### 3.3.0.3 Algorithm-based few-shot learning

Algorithm-based methods use prior knowledge to influence the parameter tuning by either providing a good initialisation or learning an optimiser. The initialisation could either be learned from other tasks and then refined (through fine-tuning by regularisation [228, 229, 230, 231, 232, 233], learning aggregation weights [234, 228], or introducing new parameters [235, 236]) or through meta-learning [237, 238].

#### 3.3.0.4 Few-shot Semantic Segmentation

One-shot learning for semantic segmentation model (OSLSM) [239] first introduced the task of few-shot segmentation and demonstrated the effectiveness of episodic training, compared favourably with a fine-tuning baseline. PLNet [240] extracted class-wise prototypes from the support set and made predictions based on cosine similarity between pixels and prototypes. This two-step architecture was later inherited by many in this field.

Based on PLNet, extracting information from the support set was proposed to further improve the performance. For instance, PPNet [241] and RPMM [242] generated multiple prototypes for each class through clustering and Expectation-Maximisation. PGNet [243] and DAN [244] adopted the attention unit to adjust the foreground prototype according to query pixel features. OANet [245] incorporated an objectness segmentation module to calculate the objectness prior. FWB [246] introduced regularisation to suppress support image background activation and boost the prototype with an ensemble of experts.

Other improvements can be achieved by refining the feature comparison module. In particular, multi-scale comparisons were utilised to overcome spatial inconsistency [247, 248, 243, 242]. PANet [249] described a prototype alignment regularisation which encouraged the resulting segmentation model to perform few-shot learning in the reverse direction.

However, these methods share a common limitation with binary masks, as discussed during episodic training: pixels not belonging to the sampled target class in query images are all labelled as the same background class, regardless of their semantic differences or similarities. In the first contribution (Chapter 4) of this thesis, we tackle this issue by generating pseudo-classes through superpixels. This approach creates additional training classes and enables discrimination of these non-query classes from the background, thereby providing a more refined training signal.

#### 3.3.0.5 Few-shot Segmentation on Medical Images

Common challenges in data collection make medical images a popular domain for few-shot semantic segmentation.

[250] first adapted episodic training for medical image segmentation and inserted 'channel squeeze & spatial excitation' blocks in a conditional network to enable heavy interaction between conditioning and support branches. Most state-of-the-art methods adopted prototypical networks and focused on improving prototype extraction. [251] assigned weights to frame-level feature and pixel-level feature based on hierarchical attention metric. [252] introduced a context relation encoder (CRE) to capture local relation features between foreground and background regions with correlation. [253] adopted an anomaly detection approach which solely depends on relatively homogeneous foreground features and explicitly refrained from modelling relatively

heterogeneous background features. [254] proposed self-reference regularisation which enforces correct segmentation of support images based on support prototypes. [255] updated prototypes with a prototype refinement module which iteratively refines prototypes to reconstruct query features with prototype and predicted query foreground mask. [256] proposed a self-organised vector quantisation (SOVQ) which creates a new representation space where learnable local prototypes are updated by assigning feature points. [257] and [254] respectively adopted contrastive learning to impose intra-class cohesion and inter-class separation between support and query features.

In many medical applications, images share similar spatial layouts - for example, bladders typically appear at the top of the lower-pelvic images. This spatial layout prior knowledge has been utilised to enhance learning in many works. [250] inserted 'channel squeeze & spatial excitation' blocks in a conditional network which use spatial maps generated by the 'channel squeeze' conditional branch for 'spatial excitation' of the segmentation branch. [252] used support mask as the query mask initialisation in its recurrent mask refinement module. [258], [259] and [256] divided the image into grids and extracted local prototypes which could be spatially matched to query pixels/voxels. These methods rely on spatial layout priors that are vulnerable to misalignment between support and query images. To this end, [251] proposed dual-level feature adjustment to address spatial-temporal distribution bias across episodes. [260] introduced a global correlation module to bridge the connection of spatially distant objects.

Nevertheless, the misalignment of support and query data from different institutions, stemming from various field-of-view protocols, can lead to performance degradation. In the second contribution (Chapter 5) of this thesis, we propose to address this challenge with a novel spatial alignment mechanism.

### 3.3.3 Semi-supervised learning

Semi-supervised learning (SSL) is a learning paradigm associated with both labelled and unlabelled data. Specifically, SSL methods harness latent patterns from the additional large amounts of unlabelled data compared to supervised learning algorithms (trained solely on labelled data) to improve learning performance. This section introduces various SSL approaches including deep generative methods, consistency regularisation methods, graph-based methods and pseudo-labelling methods.

#### 3.3.1 Deep generative methods

Deep generative methods explore the distribution of training datasets to generate new examples with generative models such as generative adversarial networks and variational autoencoders. A generative adversarial network [261] consists of a generator synthesising fake samples by learning distribution from real data and a discriminator distinguishing real and fake samples. Combining the unsupervised GAN value function with supervised classification objective functions facilitates simple yet efficient semi-supervised learning. Further improvement could be achieved by including additional information [262, 263, 264], imposing constraints on generators [265, 266, 267] and introducing encoder modules [268, 269, 270]. A variational autoencoder (VAE) [271] learns to represent high-dimensional input data with a distribution of low-dimensional latent variables and could be trained on both labelled and unlabelled data as an unsupervised preprocessing step for a supervised classifier. Kingma and Welling [272] first proposed to include labels as additional latent variables and compute supervised loss during VAE training. Following works added additional auxiliary variables [273, 274], adopted mix-

ture models [275] and utilised different disentangled methods [276, 277].

### 3.3.2 Consistency regularisation methods

Consistency regularisation methods enforce consistency between predictions perturbed in various perspectives, including input perturbation, weight perturbation and layer perturbation.

**Input perturbations** encourages the model to be transformation-consistent for unlabelled data and can be applied through data additive noise [278], augmentation [279, 280] and adversarial training [281]. For semi-supervised segmentation, commonly used augmentation includes Gaussian noise, Gaussian blurring, random rotation, scaling and contrast variations. French et al. [282] first proposed to adapt CutOut and CutMix for consistency-based semi-supervised segmentation as image perturbation. CutOut augments an image by masking a rectangular region to zero. MixUp blended inputs and target labels of two randomly chosen examples are using a randomly chosen factor. CutMix, a combination of MixUp and CutOut, cuts out a rectangular region from one image and pastes it over another. Consistency regularisation encourages consistency between segmentation prediction of the CutOuted/CutMixed images and the CutOuted/CutMixed segmentation prediction of the original images. Both MixUp [283, 284] and CutOut [285] proved their efficacy in semi-supervised medical image segmentation. While CutMix achieved better results than CutOut, hard rectangular mixing changes adjacent contexts resulting in poor artificial labels. Olsson et al. [286] introduced ClassMix to eliminate this prediction uncertainty close to the mixing border by replacing rectangular masks with pseudo-binary masks of randomly sampled classes. Chen et al. [287] further incorporated aspects of CutMix and

ClassMix to control the complexity of the augmented data while attempting to be semantically correct and address the trade-off between complexity and correctness. The efficacy of input augmentation has also be proved from medical images. Huang [285] validated the efficacy of combining Gaussian blurring, Gaussian noise, CutOut, slice misalignment and contrast variation. Shu et al. [283] and Basak et. al [284] improved semi-supervised medical segmentation adopted MixUp as image perturbation.

For the task of semantic segmentation, these augmentations could be directly applied to images and their corresponding predicted segmentation masks (both from the image domain) to create matched pairs for segmentation tasks. However, for the registration task, where the input consists of a pair of moving and fixed images from the image domain and the output is a predicted transformation in the transformation domain, such consistency does not hold directly. In the third contribution (Chapter 6) of this thesis, we propose two commutative augmentations across the image and transformation domains that can be adopted for the registration task.

**Weight Perturbations** keep the same architecture but perturb weights (parameters) with temporal ensembling methods or adversarial perturbations [288]. Most consistency regularisation semi-supervised segmentation methods are based on Mean Teacher which forces consistency between the predictions of a student network and a teacher network. The weights of student and teacher networks are respectively updated by backpropagation and exponential moving average (EMA) of the student weights. More sets of weights could be included - An et al. [289] optimised three models circularly, training one network with the guidance of the other two networks at each optimisation step. Some following works investigated more complicated linkages

between differently weighted networks. Zeng et al. [290] proposed reciprocal learning with a teacher network generating pseudo labels and a student network learning from teacher-generated pseudo labels. The teacher is updated according to reciprocal feedback signals reflecting student performance from the labelled set. Chen et al. [291] proposed cross-pseudo supervision (CPS) using the one-hot pseudo segmentation map obtained from one network to supervise the other network.

**Layer Perturbations** could be achieved by varying layer connections with random and adversarial dropout [292, 288] and enforce consistency across different layer connections.

The performance of consistency regularisation methods is highly dependent on the quality of prediction on unlabelled data. Uncertainty or confidence estimation could be utilised to learn from more meaningful and reliable targets during training. [293] designed an uncertainty-aware scheme which estimates the uncertainty with the Monte Carlo Dropout. [294] introduced a confidence module trained on labelled data to predict the model confidence on unlabelled data. [295] proposed multi-scale uncertainty rectification which selects reliable voxels for loss calculation based on the KL-divergence between the scale-average prediction and the prediction at multiple scales. [296] attached an error estimation network to predict the CE loss map of the teacher's prediction.

Consistency could be encouraged not only directly on segmentation predictions but also on other representations including reconstruction from segmentation predictions by a generation network [297], segmentation contour represented by an elliptical descriptor [298], hierarchical features extracted by a discriminator network [299], attention maps [300, 301], class-vectors

mapped from segmentation predictions by a classification model [302], and prototype-based segmentation [303].

### 3.3.3 Pseudo-labelling methods

Pseudo-labelling methods usually consist of alternating steps of training and pseudo-labelling. During training, one or more models are trained on labelled data as well as pseudo-labelled data from the previous iterations. During pseudo-labelling, the models are used to infer previously unlabelled samples and generate pseudo-labels. While self-training methods [304] leverage the model's own prediction for pseudo-labelling, co-training methods [305, 306, 307, 308, 309, 310, 311] involve multiple learners and exploit disagreement among learners.

Pseudo-labelling for segmentation could be classified as direct or indirect generation. Direct generation chooses model predictions as pseudo-labels based on confidence levels. The confidence level could be measured globally by evaluating consistency between prediction before and after cross-domain augmentation with low-level amplitude spectrum from Fourier transformation [312], or locally for pixel-wise tasks through predicting the inconsistency between different misclassification costs [313]. The pseudo-labels could be then refined with strategies such as thresholding confidence values [314], self-ensembling strategy which gradually ensembles predictions of unlabelled data via moving average at each epoch [315], features and edges of superpixel maps [316] and dual error correction based on discriminative error predicted by a specifically trained network. Indirect generation is mostly based on label propagation such as prototypical learning that compares unlabelled images with prototypes extracted from labelled images [317], neighbour matching with a pseudo label estimator based on local manifold updated dynamically with labelled

data [318] and matching unlabelled image pixels to the semantics of the best fitting pixel in a reference set of sampled labelled images [319].

### 3.3.4 Knowledge prior methods

Knowledge priors (information about the problem available in addition to the training data) are helpful for learning new tasks especially when limited annotation is available. For instance, medical images inherently carry anatomical priors such as the shape and position of organs/structures, which have been adopted to improve representation ability by multiple semi-supervised approaches. Probability atlas maps, indicating the probability of specific organs/structures appearing at different locations, are widely adopted for medical image segmentation for usages such as weighting segmentation loss to focus on challenging samples [320], selecting partial high-confidence pixels [321] and initialising coarse masks to be refined by the deep contextual refinement module.

Moreover, with the large unlabelled dataset, priors could be learnt through an efficient proxy task pertaining such as reconstruction from perturbated images [285] or autoencoder latent representations [322] and contrastive learning using superpixels [285, 322]. These tasks could be used as not only pretraining but also regularisation. Auxiliary tasks such as self-supervised jigsaw puzzle [323], lesion region inpainting [324] and contrastive learning [325, 326, 327] could also be used for regularisation.

# Chapter 4

# Few-shot Semantic Segmentation with Self-supervision from Pseudo-classes

This chapter focuses few-shot semantic segmentation in general computer vision field. Despite the success of deep learning methods for semantic segmentation, few-shot semantic segmentation remains a challenging task due to the limited training data and the generalisation requirement for unseen classes. While recent progress has been particularly encouraging, we discover that existing methods tend to have poor performance in terms of meanIoU when query images contain other semantic classes besides the target class. To address this issue, we propose a novel self-supervised task that generates random pseudo-classes in the background of the query images, providing extra training data that would otherwise be unavailable when predicting individual target classes. To that end, we adopted superpixel segmentation for generating the pseudo-classes. With this extra supervision, we improved the

meanIoU performance of the state-of-the-art method by 2.5% and 5.1% on the one-shot tasks, as well as 6.7% and 4.4% on the five-shot tasks, on the PASCAL-5i and COCO benchmarks, respectively. This work [2] has been published with 2021 British Machine Vision Conference (BMVC).

We first detail the few-shot semantic segmentation's task definition in Section 4.1 and generally adopted deep learning engine in Section 4.2 before detailing the proposed innovation with its motivation in Section 4.3, related work in Section 4.4, method in Section 4.5 and experiments in Section 4.6.

## 4.1 Task Definition

As mentioned in Section 3.1, the goal of semantic segmentation is to classify each pixel in an image into semantically meaningful classes. The data-driven representation learning enabled by deep neural networks has led to accurate and compelling segmentation results [116, 328]. However, most segmentation methods rely on large datasets with full annotations, which are expensive and labour-intensive to collect. Furthermore, the predictable categories are often constrained to those that have been annotated in the dataset. It is, therefore, of great value to reduce the annotation burden by enabling unseen class prediction with only a few training examples. Specifically, given a small number of *support* images with ground truth binary masks of an unseen *target* class, the task is to segment the regions of the same class in unannotated *query* images. This is known as few-shot semantic segmentation (FSS) [239].

Specifically, for few-shot 2D natural image segmentation, denote $(I, M_{(c)})$ as an single-class image-mask pair, $I \in \mathbb{R}^{3\times W\times H}$ is a RGB image of width $W$, height $H$ and $M_{(c)}$ is the corresponding binary mask of class $c$. For $K$-shot semantic segmentation, the task is to predict the *novel* class $c \in \mathcal{C}_{\text{novel}}$ binary

mask of a query image $I^{\text{que}}$ given a support set $\mathcal{S} = \{(I^{\text{sup},i}, M_{(c)}^{\text{sup},i})\}_{k=1}^{K}$ including $K$ examples image-mask pairs of the class sampled from a set of novel classes isolated from training.

## 4.2 Background

### 4.2.1 Dataset

Though data and annotations related to novel (test) classes should be isolated from training, a *base dataset* could still be formed with image-mask pairs of base classes $\mathcal{C}_{\text{base}}$, given base classes and novel classes are mutually exclusive i.e. $\mathcal{C}_{\text{base}} \cap \mathcal{C}_{\text{novel}} = \emptyset$. Practically, a set of images $\mathcal{I}$ could be separated into disjoint training and test subsets $\mathcal{I}^{\text{train}}$ and $\mathcal{I}^{\text{test}}$. The base dataset $\mathcal{D}_{\text{base}}$ is formed with training images and the corresponding labels of the base classes.

$$\mathcal{D}_{\text{base}} = \bigcup_{I \in \mathcal{I}^{\text{train}}} \bigcup_{c \in \mathcal{C}_{\text{base}}} \{(I, M_{(c)})\}. \tag{4.1}$$

where $M_{(c)}$ is the class $c$ segmentation mask of image $I$.

Similarly, test images and the corresponding labels of novel classes form the testing novel dataset.

$$\mathcal{D}_{\text{novel}} = \bigcup_{I \in \mathcal{I}^{\text{test}}} \bigcup_{c \in \mathcal{C}_{\text{novel}}} \{\ (I,\ M(I, c))\ \} \tag{4.2}$$

In this way, data leakage is prevented as there is no overlapping between images inside the base and novel datasets, i.e.$\mathcal{I}^{\text{train}} \cap \mathcal{I}^{\text{test}} = \emptyset$.

**Algorithm 1** Episodic Training

**input :** Neural network $g(\cdot|\theta)$ with parameters $\theta$.
Learning rate $\alpha_{\text{lr}}$.
Number of shots $K$.
Base (training) dataset $\mathcal{D}_{base} = \bigcup_{I \in \mathcal{I}^{\text{train}}} \bigcup_{c \in \mathcal{C}_{\text{base}}} \{(I, M_{(c)})\}$
**output:** Trained model $g_\theta$.
**while** *Training* **do**
  Sample $c \sim \mathcal{C}_{base}$
  **while** *True* **do**
    Sample $I \in \mathcal{I}_{\text{train}}$
    **if** $M_{(c)}$ *satisfies* $\sum_{x=1}^{W} \sum_{y=1}^{H} \sum_{z=1}^{D} m^i_{(c)xyz} > 0$ **then**
      $(I^{\text{que}}, M^{\text{que}}_{(c)}) \leftarrow (I, M_{(c)})$
      Break
    **end**
  **end**
  $\mathcal{S} \leftarrow \{\}$
  $k \leftarrow 1$
  SAMPLED_SET $\leftarrow \{I^{\text{que}}\}$
  **while** $k \leq K$ **do**
    Sample $I \in \mathcal{I}_{\text{train}}$
    **if** $I \notin$ *SAMPLED_SET* **then**
      **if** $M_{(c)}$ *satisfies* $\sum_{x=1}^{W} \sum_{y=1}^{H} \sum_{z=1}^{D} m^i_{(c)xyz} > 0$ **then**
        $\mathcal{S} \leftarrow \mathcal{S} \cup (I, M_{(c)})$
        SAMPLED_SET $\leftarrow$ SAMPLED_SET $\cup \{I\}$
        $k \leftarrow k + 1$
      **end**
    **end**
  **end**
  Predict $\bar{M}^{\text{que}}_{(c)} = g(I^{\text{que}}, \mathcal{S}; \theta)$
  Compute loss: $L = \mathcal{L}(\bar{M}^{\text{que}}_{(c)}, M^{\text{que}}_{(c)})$
  Update model parameters: $\theta \leftarrow \theta - \alpha_{\text{lr}} \nabla_\theta L$
**end**

### 4.2.2 Episodic Training and Evaluation

Episodic training was first adapted to few-shot semantic segmentation by Shaban et. al. [239] and is now commonly adopted by state-of-the-art methods. Episodic training paradigm (detailed in Algorithm 1) simulates the few-shot task during training. Each *episode* consists of a query $(I^{\text{que}}, M^{\text{que}}_{(c)})$ and $K$ support $\mathcal{S} = \{(I^{\text{sup},i}, M^{\text{sup},i}_{(c)})\}_{k=1}^{K}$ image-mask pairs, for a base class $c \in \mathcal{C}_{\text{base}}$ sampled from base classes. Like in evaluation, given a query image $I^{\text{que}}$ and a support set $\mathcal{S}$ model $g(\cdot;\theta)$ parameterised by $\theta$ predicts the class $c$ probability mask $\bar{M}^{\text{que}}_{(c)} = g(I^{\text{que}}, \mathcal{S}; \theta)$ of the query image $I^{\text{que}}$. Model parameter $\theta$ is trained to minimise the difference between the probability mask prediction $\bar{M}^{\text{que}}_{(c)}$ and the ground-truth query mask $M^{\text{que}}_{(c)}$.

During evaluation, as detailed in Algorithm 2, all images from the novel dataset are considered query images once for each of the novel classes with a randomly sampled support set.

## 4.3 Introduction

While existing methods [249, 248] have achieved significant improvements over fine-tuning and foreground-background segmentation baselines [239], we observed a significant drop in performance when the *query* image contains multiple classes (Figure 4.1), *i.e.* the more classes (including both training and testing classes defined in the dataset) the image has, the more difficult is it to segment the *target* class precisely.

For example, on the 1-shot PASCAL-5i [329, 239] benchmark, the state-of-the-art (PFENet [248]) at the time of publication achieved a mean IoU score of 66.1% for *query* images with one class but only 32.7% for *query* images

**Algorithm 2** Few-shot Evaluation

**input :** Neural network $g(\cdot|\theta)$ with parameters $\theta$.
Number of shots $K$.
Novel dataset $\mathcal{D}_{\text{novel}} = \bigcup_{I \in \mathcal{I}^{\text{test}}} \bigcup_{c \in \mathcal{C}_{\text{novel}}} \{\ (I,\ M(I,c))\ \}$.
**output:** Evaluation Metric
**for** $I \in \mathcal{I}_{test}$ **do**
  **for** $c \in \mathcal{C}_{novel}$ **do**
    **if** $M_{(c)}$ *satisfies* $\sum_{x=1}^{W} \sum_{y=1}^{H} \sum_{z=1}^{D} m^{i}_{(c)xyz} > 0$ **then**
      $(I^{\text{que}}, M^{\text{que}}_{(c)}) \leftarrow (I, M_{(c)})$
      $\mathcal{S} \leftarrow \{\}$
      $k \leftarrow 1$
      SAMPLED_SET $\leftarrow \{I^{\text{que}}\}$
      **while** $k \leq K$ **do**
        Sample $I \in \mathcal{I}_{\text{test}}$
        **if** $I \notin$ *SAMPLED_SET* **then**
          **if** $M_{(c)}$ *satisfies* $\sum_{x=1}^{W} \sum_{y=1}^{H} \sum_{z=1}^{D} m^{i}_{(c)xyz} > 0$ **then**
            $\mathcal{S} \leftarrow \mathcal{S} \cup (I, M_{(c)})$
            SAMPLED_SET $\leftarrow$ SAMPLED_SET $\cup \{I\}$
            $k \leftarrow k + 1$
          **end**
        **end**
      **end**
      Predict $\bar{M}^{\text{que}}_{(c)} = g(I^{\text{que}}, \mathcal{S}; \theta)$
      Transform predicted probability mask $\bar{M}^{\text{que}}_{(c)}$ to segmentation mask $\grave{M}^{\text{que}}_{(c)}$
      Update metric by comparing $\grave{M}^{\text{que}}_{(c)}$ and $M^{\text{que}}_{(c)}$)
    **end**
  **end**
**end**

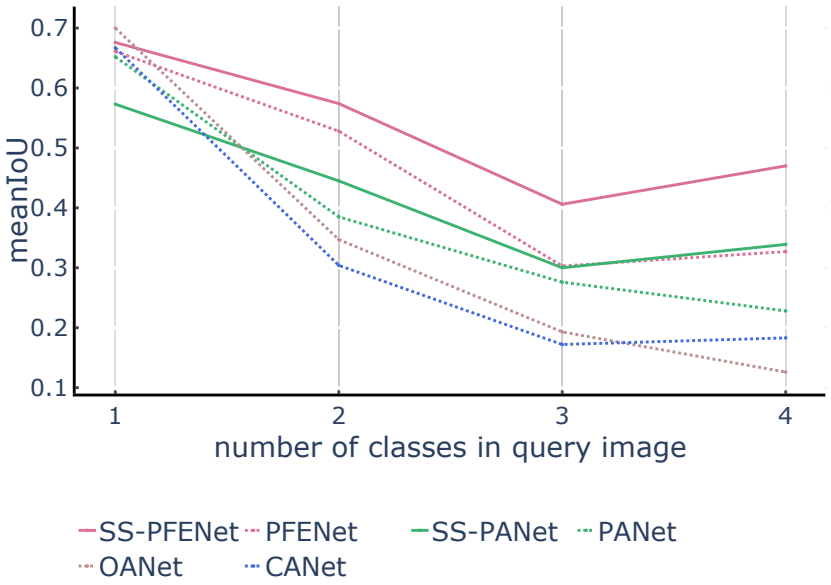



Figure 4.1: Performance of various few-shot semantic segmentation methods versus the number of semantic classes in the query image. Different colours represent different methods and the prefix "SS-" denotes models with the proposed self-supervision. Existing methods tend to perform similarly well when there is a single class in the query, but their performance tends to decrease when more classes are present.

with four classes. A similar trend has also been observed in the 5-shot results and with other benchmarks such as few-shot COCO [330].

We conjecture that the performance drop is caused by the presence of the other non-target classes in the *query* images during training; Here, the target class is the class of segmentation interest, whilst in binary segmentation tasks all the non-target classes are labelled as background regardless of their differences. Although the *target*-versus-*non-target* binary classification paradigm seems appropriate for images with a single target class, it may inhibit the model from learning meaningful representations from the *non-target* pixels. Given practical limitations such as finite data and imperfect labels, further discriminating these non-target classes can potentially aid classification of *target* pixels, similar to multi-task learning [331]. This is especially true in

few-shot learning, in which data-efficient model updating is sought-after.

Following the above intuition, we propose a novel self-supervised training strategy that generates pseudo-classes in the *query* images. For each episode, a pseudo-class is created by sampling superpixels with high activation from the *query* background. Together with the *query* image, an extra pseudo-image-mask pair can be generated to assist training. By guiding the model to discriminate the pseudo-class from the background, this training paradigm enforces the model to distinguish possible non-target classes (those unannotated in the training set) present in the background of the *query* images. As details presented in the remainder of this chapter, the proposed method leads to consistent improvement over different architectures and different datasets.

Our contributions are summarised as follows:

- We propose a novel self-supervised task that generates pseudo-classes for few-shot semantic segmentation, which encourages the models to learn a more discriminative feature space while adapting to *novel* target classes.
- We highlight the significant performance drop as the number of classes in *query* images increases for existing methods. To the best of our knowledge, it is the first time this issue is highlighted and investigated.
- We present an extensive set of 1-shot and 5-shot experiments that demonstrate the significant improvement from the proposed method, for both PASCAL-5i and COCO datasets. The performance on *query* images with multiple classes was improved and the imbalance between classes was mitigated.

## 4.4 Related Work in Superpixel Segmentation

Superpixel segmentation is an over-segmentation method that groups pixels coherently based on handcrafted features. It is widely used in image segmentation task to reduce computational costs [332, 333]. As superpixel provides additional local information, it has been used for few-shot semantic segmentation to compensate for the lack of annotated data. For example, SSL-ALPNet [334] used superpixel-based pseudo-labels to replace manual annotation. An important difference in our use of pseudo-label is to amend the performance drop when multiple classes present in the *query* image. PPNet [241] enriched support prototypes by extracting features from unlabelled support images based on superpixel segmentation. Rather than focusing on the support set, in this work, we sample superpixels to define pseudo-classes only on the background of the query images.

## 4.5 Method

As mentioned in Section 4.2.2, most state-of-the-art few-shot semantic segmentation methods follow the episodic training paradigm. For each iteration, an *episode*, consisting a *support* set $\mathcal{S} = \{(I^{\text{sup},i}, I^{\text{sup},i}_{(c)})\}_{k=1}^{K}$ and a *query* image-mask pair $(I^{\text{que}}, M^{\text{que}}_{(c)})$, where the *target* class $c \in \mathcal{C}_{\text{base}}$ is sampled from the *base* dataset $\mathcal{D}_{\text{base}}$.

### 4.5.1 Prototype Extraction

This method adopts a prototypical approach, which computes a prototype feature vector per class by averaging support features over locations belonging to that class. Segmentation prediction is made locally by comparing query

features with the class prototype.

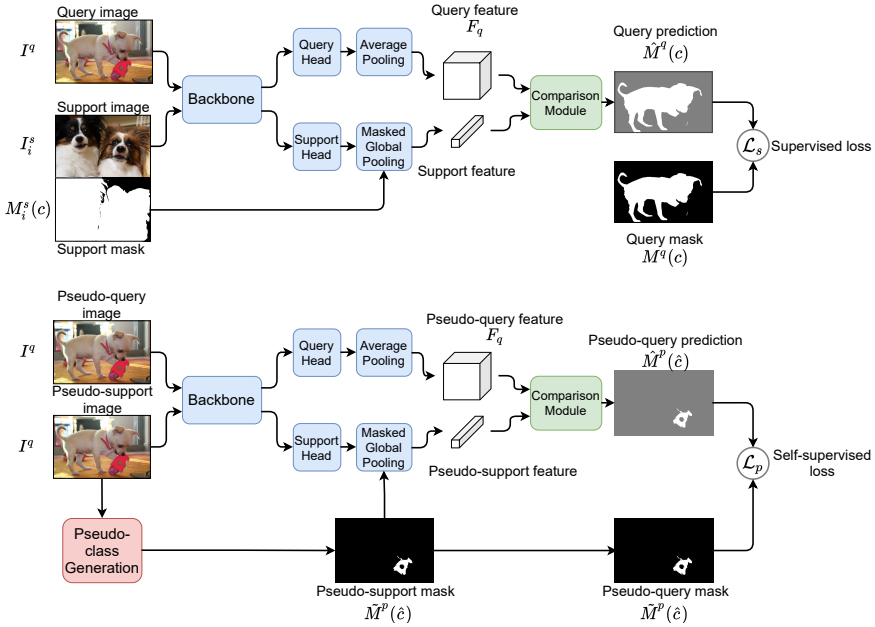



Figure 4.2: An illustration of training data flow. The upper diagram demonstrates the supervised pathway which predicts the query mask, conditioned on the given support example. The lower diagram demonstrates the self-supervised pathway which generates a pseudo-class and predicts the generated mask, conditioned on the pseudo-support example. Networks are shared between two pathways during training.

Specifically, for 2D RGB natural image few-shot segmentation, given an image $I \in \mathbb{R}^{3\times W\times H}$ with its corresponding mask $M_{(c)} \in \mathbb{R}^{W\times H}$, a feature map $F \in \mathbb{R}^{\Lambda\times W\times H}$ could be extracted by the feature extractor with $\Lambda$ channels. Denote the components of $M_{(c)}$ and $F$ respectively as $m_{(c)xy}$ and $f_{\lambda xy}$ where the subscript $\lambda = 1, ..., \Lambda$ is used as channel index whilst $x = 1, ..., W$ and $y = 1, ..., H$ are used as location coordinate indices.

A class prototype $h_{(c)}$ could be derived by averaging the support feature map $F^{\text{sup}}$ over locations belonging to class $c$.

$$h_{(c)} = \frac{\sum_{x=1}^{W}\sum_{y=1}^{H} m_{(c)xy}[f_{(c)xy}, ..., f_{\Lambda xy}]^\top}{\sum_{x=1}^{W}\sum_{y=1}^{H} m_{(c)xy}} \tag{4.3}$$

Predictions are made locally by comparing query features $F^{\text{que}}$ with the class prototype $h_{(c)}$ with the comparison module, resulting in a predicted query probability mask $\bar{M}^{\text{que}}_{(c)}$.

The model is optimised to reduce $\mathcal{L}_{\text{ce}}(\bar{M}^{\text{que}}_{(c)}, M^{\text{que}}_{(c)})$, the cross entropy loss between the predicted *query* probability mask $\bar{M}^{\text{que}}_{(c)}$ and the ground-truth *query* mask $M^{\text{que}}_{(c)}$.

Under this training paradigm as shown in the upper diagram in Figure 4.2, *non-target*-class objects with different semantic meanings in the *query* image are all regarded as the same class – background. The purposed self-supervision task aims to better utilise the information from the background pixels.

### 4.5.2 Self-supervision from Pseudo-classes

We propose a self-supervision task for better utilise the information from the background pixels.



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

Figure 4.3: Step-by-step breakdown of pseudo-mask generation for self-supervision. (a) Input query image. (b) Generate superpixels in the query image. (c) Remove superpixels that coincide with the original class ground truth. (d) Randomly sample superpixel with high corresponding feature activation, and convert it into a binary mask.

First, we apply a superpixel segmentation method (such as Felzenszwalb *et al.* [335]) to $I^{\text{que}}$. For each superpixel, denote its region as the pseudo-class $\dot{c}$ and the binary mask representing the region as the pseudo-mask $M_{(\dot{c})} \in \{0,1\}^{H \times W}$.

Secondly, we refine each pseudo-mask $M_{(\dot{c})}$ with the aid of the ground-truth mask $M^{\text{que}}_{(\dot{c})}$ by removing pixels of the *target* class $c$,

$$M_{(\dot{c})} \leftarrow M_{(\dot{c})} \odot (1 - M^{\text{que}}_{(c)}), \tag{4.4}$$

where $\odot$ denotes the element-wise multiplication.

Then, a class activation score $s(\dot{c})$ is calculated for each pseudo-class $\dot{c}$ by averaging the mask-size-interpolated extracted query feature $F^{\text{que}} \in \mathbb{R}^{\Lambda \times W \times H}$ over the pseudo-mask foreground and channels, specifically:

$$s(\dot{c}) = \frac{\sum_{\lambda=1}^{\Lambda} \sum_{x=1}^{W} \sum_{y=1}^{H} F^{\text{que}}_{(x,y,z)} M_{(\dot{c})xy}}{\sum_{\lambda=1}^{\Lambda} \sum_{x=1}^{W} \sum_{y=1}^{H} M_{(\dot{c})xy}} \tag{4.5}$$

Lastly, a pseudo-class $\grave{c}$ is randomly selected among the ones having the top-5 scores, so that the selected pseudo-class is more likely to be non-background while keeping diversity between pseudo-classes. The corresponding refined mask is denoted as $M_{(\grave{c})}$.

As shown in the lower diagram of Figure 4.2, the additional image-mask pair $(I^{\text{que}}, M_{(\grave{c})})$ serves as both *pseudo-support* and *pseudo-query*, forming a *pseudo support-query* pair for $I^{\text{que}}$ in the current training iteration. Specifically, the self-supervised task is defined as: given the *pseudo-support* example $(I^{\text{que}}, M_{(\grave{c})})$, the model is required to predict the same mask $M_{(\grave{c})}$ for the pseudo-class $\grave{c}$. The corresponding prediction is denoted as $\bar{M}_{(\grave{c})} \in [0, 1]^{H \times W}$. Similarly, we compute a cross-entropy loss $\mathcal{L}_{\text{ce}}(\bar{M}_{(\grave{c})}, M_{(\grave{c})})$, for this self-supervision.

The total loss is therefore a weighted sum of the supervised loss with the

target class $c$ and the self-supervised loss with the pseudo-class $\grave{c}$,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ce}}(\bar{M}^{\text{que}}_{(c)}, M^{\text{que}}_{(c)}) + \alpha \mathcal{L}_{\text{ce}}(\bar{M}_{(\grave{c})}, M_{(\grave{c})}), \tag{4.6}$$

where $\alpha_{\text{pseudo}}$ is a scaling coefficient to control the contribution of the self-supervised loss towards the overall objective.

## 4.6 Experiments

### 4.6.1 Implementation Details

We evaluated our methods on the PASCAL-5i [239] (which consists of PASCAL VOC 2012 [329] with extended annotations from SDS [336]) and COCO [330] datasets. Each dataset has 4 configurations (folds) of base-novel class splits, following OSLSM [239] and FWB [246] protocol. PFENet [248], the current state-of-the-art, was chosen as our baseline architecture. Felzenszwalb *et al.* [335] was used for superpixel segmentation with scale set to 100, sigma to 0.8 and min_size to 200. The derived superpixel segmetations were further refined by the training (base) class masks. The hyper-parameter $\alpha_{\text{pseudo}}$ was set to 0.5 in equation (4.6). We report the meanIoU metric for evaluation, which is the intersection-over-union (IoU) averaged over classes. For 5-shot evaluation, the individual 1-shot models were used without fine-tuning and we averaged the resulting prototype vectors of supports before feeding them to the comparison module. For brevity, the prefix "SS-" denotes models trained with the proposed self-supervision.

Table 4.1: meanIoU results on PASCAL-5i [239]. Best performance with the same model and overall are respectively bolded and starred. "SS-" refers to models with self-supervision.

| method | 1-shot | | | | | 5-shot | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | fold1 | fold2 | fold3 | fold4 | mean | fold1 | fold2 | fold3 | fold4 | mean |
| VGG-16 | | | | | | | | | | |
| OSLSM [239] | 33.6 | 55.3 | 40.9 | 33.5 | 40.8 | 37.5 | 50.0 | 44.1 | 33.9 | 41.4 |
| co_FCN [337] | 36.7 | 50.6 | 44.9 | 32.4 | 41.2 | 35.9 | 58.1 | 42.7 | 39.1 | 44.0 |
| AMP [338] | 41.9 | 50.2 | 46.7 | 34.7 | 43.4 | 41.8 | 55.5 | 50.3 | 39.9 | 46.9 |
| Sg-one [339] | 40.2 | 58.4 | 48.4 | 38.4 | 46.4 | 41.9 | 58.6 | 48.6 | 39.4 | 47.1 |
| FWBNet [246] | 47.0 | 59.6 | 52.6 | 48.3 | 51.9 | 50.9 | 62.9 | 56.5 | 50.1 | 55.1 |
| RPPM [242] | 47.1 | 65.8 | 50.6 | 48.5 | 53.0 | 50.0 | 66.5 | 51.9 | 47.6 | 54.0 |
| CRNet [340] | - | - | - | - | 55.2 | - | - | - | - | 58.5 |
| PANet [249] | 42.3 | 58.0 | 51.1 | 41.3 | 48.2 | 51.8 | **64.6** | 59.8 | 46.5 | 55.7 |
| SS-PANet | **49.3** | **60.8** | **53.9** | **45.2** | **52.3** | **54.5** | 64.5 | **62.3** | **52.0** | **58.3** |
| PFENet[248] | **56.9** | **68.2** | 54.4 | 52.4 | 58.0 | **59.0** | 69.1 | 54.8 | 52.9 | 59.0 |
| SS-PFENet | 54.5 | 67.4 | **63.4** | **54.0** | **59.8** | 56.9 | **70.0** | **68.3** | **62.1** | **64.3** |
| ResNet-50 | | | | | | | | | | |
| PPNet [241] | 48.6 | 60.6 | 55.7 | 46.5 | 52.9 | 58.9 | 68.3 | 66.8 | 58.0 | 63.0 |
| RPMM [242] | 50.2 | 66.9 | 52.6 | 50.7 | 55.1 | 56.3 | 68.3 | 54.5 | 51.0 | 57.5 |
| CANet [247] | 52.5 | 65.9 | 51.3 | 51.9 | 55.4 | 55.5 | 67.8 | 51.9 | 53.2 | 57.1 |
| CRNet [340] | - | - | - | - | 55.7 | - | | - | - | 58.8 |
| PGNet [243] | 56.0 | 66.9 | 50.6 | 50.4 | 56.0 | 57.7 | 68.7 | 52.9 | 54.6 | 58.5 |
| ORG [341] | 52.6 | 65.8 | 54.7 | 52.1 | 56.3 | 57.2 | 67.8 | 57.5 | 56.2 | 59.7 |
| OANet [245] | 56.9 | 66.4 | 57.1 | 50.7 | 57.8 | 59.6 | 68.9 | 57.8 | 54.4 | 60.2 |
| ASGNet [342] | 58.8 | 67.9 | 56.8 | 53.7 | 59.3 | 63.66 | 70.6 | 64.2 | 57.4 | 64.0 |
| RePRI [343] | 60.2 | 67.0 | 61.7 | 47.5 | 59.1 | **64.5**$^*$ | 70.8 | **71.7**$^*$ | 60.3 | 66.8 |
| SCL [324] | **63.0**$^*$ | **70.0**$^*$ | 56.5 | **57.7**$^*$ | 61.8 | **64.5**$^*$ | 70.9 | 57.3 | 58.7 | 62.9 |
| PFENet [248] | 61.7 | 69.5 | 55.4 | 56.3 | 60.7 | 63.1 | 70.7 | 55.8 | 57.9 | 61.9 |
| SS-PFENet | 58.9 | 69.9 | **66.4**$^*$ | **57.7**$^*$ | **63.2**$^*$ | 61.4 | **75.0**$^*$ | 70.5 | **67.7**$^*$ | **68.6**$^*$ |

### 4.6.2 Results

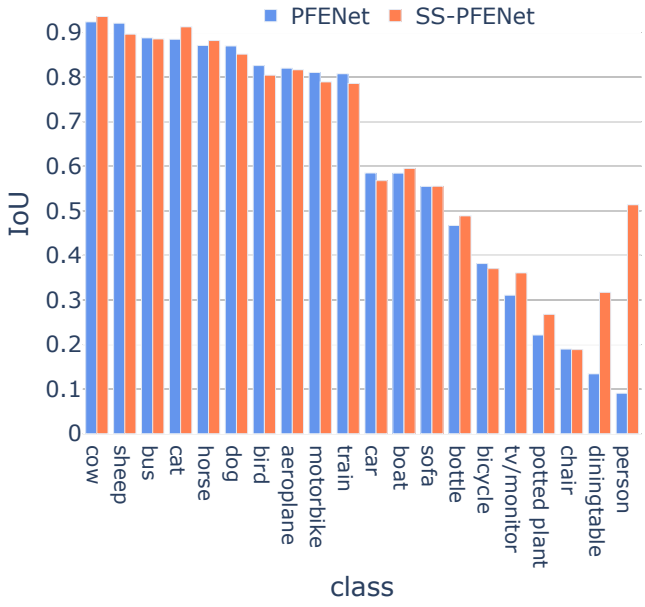
The meanIoU performances per fold for PASCAL-5i 1-shot and 5-shot tasks are presented in Table 4.1. By incorporating the self-supervision during the training of ResNet-50 PFENet, we achieved 63.2% meanIoU on the PASCAL-5i 1-shot task, improving its baseline performance by 2.5%. For the 5-shot task, we achieved 68.6% meanIoU with a 6.7% absolute improvement. This

Table 4.2: meanIoU results on COCO [329]. Best performance with the same model and overall are respectively bolded and starred. "SS-" refers to models with self-supervision.

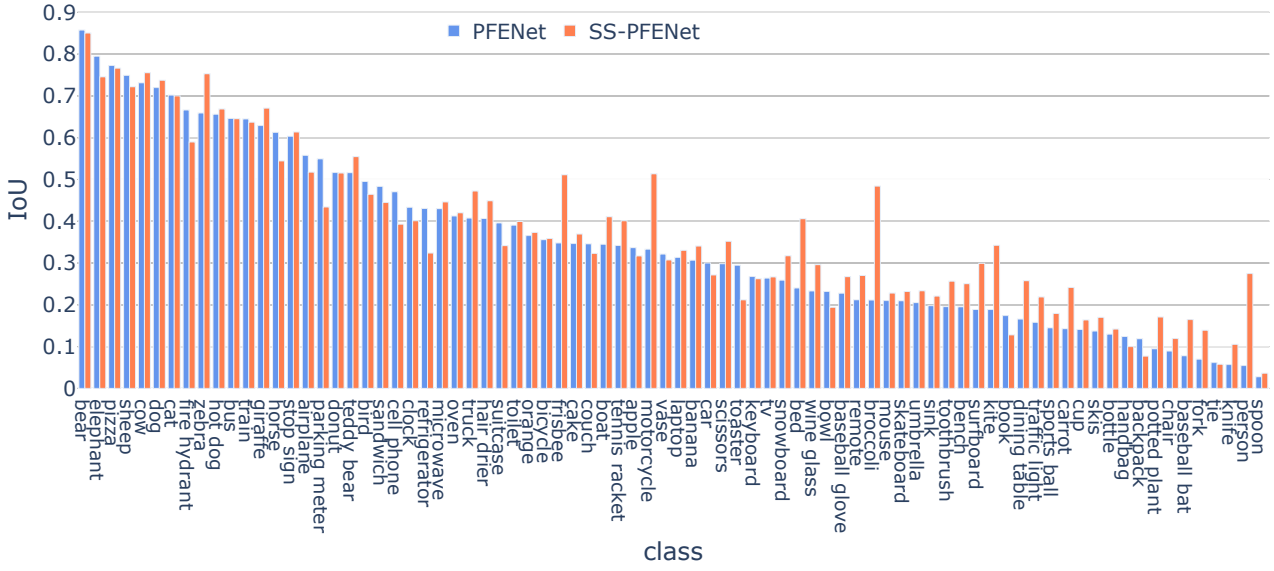
| | 1-shot | | | | | 5-shot | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| method | fold1 | fold2 | fold3 | fold4 | mean | fold1 | fold2 | fold3 | fold4 | mean |
| VGG-16 | | | | | | | | | | |
| FWBNet [246] | 18.4 | 16.7 | 19.6 | 25.4 | 20.0 | 20.9 | 19.2 | 21.9 | 28.4 | 22.6 |
| PANet [249] | 28.7 | 21.2 | 19.1 | 14.8 | 21.0 | 39.4 | 28.3 | 28.2 | 22.7 | 29.7 |
| SS-PANet | **29.8** | **21.2** | **26.5** | **28.5** | **26.2** | **36.7** | **41.0** | **37.6** | **35.6** | **37.7** |
| PFENet [248] | 34.3 | 33.0 | 32.3 | 30.1 | 32.4 | 38.5 | 38.6 | 38.2 | 34.3 | 37.4 |
| SS-PFENet | **35.6**$^*$ | **39.2**$^*$ | **37.6**$^*$ | **37.3**$^*$ | **37.5**$^*$ | **40.4**$^*$ | **45.8**$^*$ | **40.3**$^*$ | **40.7**$^*$ | **41.8**$^*$ |

sets a new state-of-the-art on the PASCAL-5i dataset for both 1-shot and 5-shot. On the COCO benchmark (Table 4.2), we achieved 37.5% and 41.8% meanIoU for 1-shot and 5-shot using VGG-16 PFENet and set the new state-of-the-art from our knowledge.

Additionally, we present the IoU values per class in Figure 4.4 for PFENet with and without self-supervision on both PASCAL-5i and COCO. Most of the improvement came from previously under-performing classes (i.e person, dining table, potted plant, etc.), and thus led to a more class-wise balanced performance. These enhancements may stem from the tendency of previously underperforming classes to co-occur with other classes in images. For instance, out of 548 testing images for the "person" class, 493 also contain non-person classes, yielding a proportion of 0.90. A quantitative analysis investigating the correlation between the performance improvement for each class $C$ and the presence of non-$C$ classes in the testing images yielded a positive Pearson correlation coefficient of 0.60 (as shown in Figure 4.8). Importantly, this improvement is not only anticipated in images with greater potential for enhancement but is also highly desirable in more intricate real-world scenarios where multiple classes are commonly encountered.

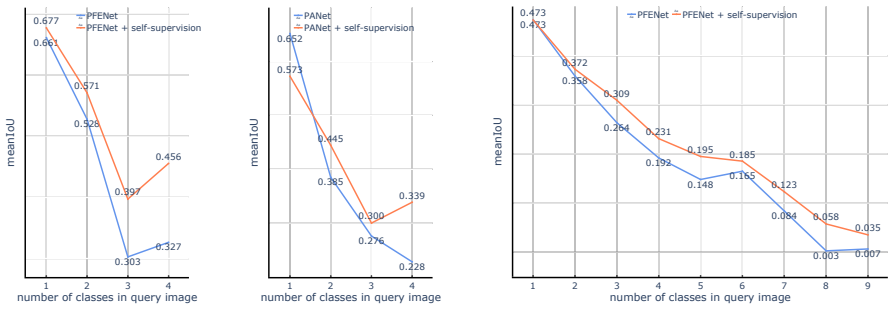



(a) PASCAL-5i




(b) COCO-20i

Figure 4.4: Class IoUs reached by PFENet with and without self-supervision on (a) PASCAL-5i and (b) COCO-20i dataset. X-axis corresponds to different classes in descending order of performance reached by PFENet without self-supervision.




(a) PASCAL-5i (b) PASCAL-5i (c) COCO

Figure 4.5: meanIoUs reached with and without superpixel self-supervision when different number of classes occur in the query image.

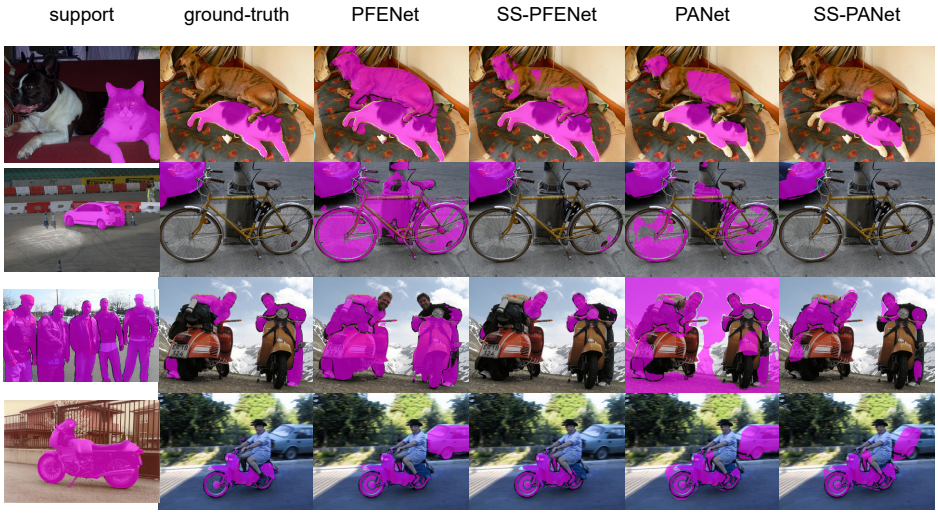



Figure 4.6: Qualitative results achieved by PFENet and PANet with and w/o the purposed self-supervision.

The meanIoU performances on images with the different number of classes are plotted in Figure 4.5. Here the number of classes is defined by the ground-truth labels. Therefore, for PASCAL and COCO, the maximum available numbers of classes in a query image are respectively 20 and 80. From Figure 4.5(a) and Figure 4.5(c), one can observe that the proposed method consistently improved the performance when multiple classes were presented in the query image over both PASCAL-5i and COCO dataset. This ex-



(a) PFENet (b) SS-PFENet (c) PFENet (d) SS-PFENet

Figure 4.7: Visual comparison between feature spaces from PFENet trained with and without the purposed self-supervision. (a) and (b) were trained on PASCAL-5i [239] excluding *fold2* classes, (c) and (d) were trained on COCO excluding *fold3* classes.

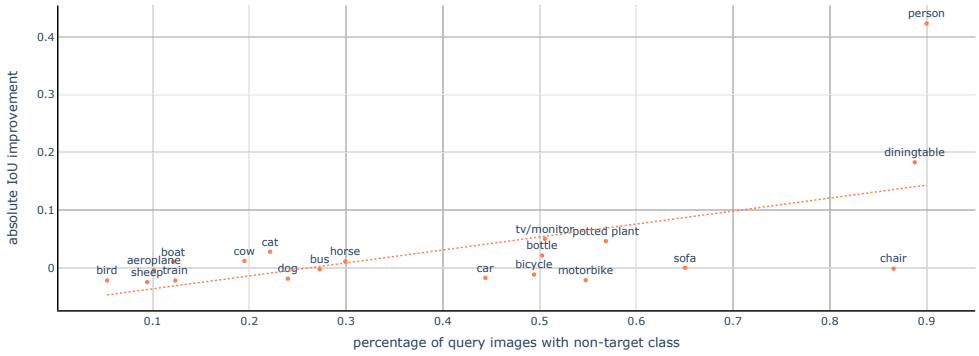



Figure 4.8: Absolute class IoU improvement reached by adding self-supervision to PFENet on PASCAL dataset. X-axis corresponds to the percentage of query images with non-target classes.

plains the higher relative improvement in performance achieved on COCO (11.8%) compare to PASCAL-5i (3.1%) when the same backbone (VGG-16) was adopted. The observed improvement in performance, both by the baseline and our methods in Figure 4.5(a), when the query includes four classes compared to three classes, is likely statistically insignificant. This is largely attributable to the limited number of samples containing either three or four classes in the PASCAL dataset.

Figure 4.6 shows some qualitative examples achieved by ResNet-50 PFENet

with and without the purposed self-supervision. When multiple objects exist in the query image, PFENet succeeds in detecting them from the background but fails to filter out non-target-class objects. This failure in discriminating class differences has been rectified by the proposed self-supervised task which is likely to increase class diversities due to introducing pseudo-classes.

To further understand the impact of the proposed self-supervised task, we used t-SNE [344] to visualise the foreground embedding from query images in Figure 4.7. Each dot corresponds to the global average pooled query foreground feature from a testing episode, and different colours represent different classes. The point clouds of different classes tend to be better disentangled, which qualitatively supports our theory and explains the observed results.

### 4.6.3 Ablation Studies

#### 4.6.1 Model Architecture

To test the applicability of the proposed self-supervision across different architectures, further experiments were performed on PANet [249], an intrinsically different architecture that adopts cosine similarity to compare both foreground and background prototypes with each pixel and make predictions. As shown in Table 4.1 and Table 4.2, the proposed method achieved 4.1% and 2.6% absolute improvement on PASCAL-5i 1-shot and 5-shot tasks as well as 5.2% and 8.0% absolute improvement on COCO 1-shot and 5-shot tasks. The consistent improvement on multi-class query images was also observed across architectures as shown in Figure 4.5(b).

### 4.6.2 Pseudo-class Generation

We compared the following three other different pseudo-class generation processes: 1) Gridding: divide the image into $10 \times 10$ grids evenly, 2) SLIC [345]: segment image through iterative clustering, setting compactness to 10 and n_segments to 100, 3) HED contour detector [346]: leverage deep learning supervision to detect edges. The meanIoU on the PASCAL-5i 1-shot task are reported in Table 4.3. Gridding leads to a significant drop of performance to 57.5%, lower than being without self-supervision. The possible reason may be the presence of multiple classes together with the background in the same grid, making the pseudo-classes misleading. We chose to use the second-best algorithm Felzenszwalb instead of HED because HED was pre-trained on the SDS dataset [336], part of our test dataset, which may cause potential information leakage.

Table 4.3: Performance of ResNet-50 PFENet on PASCAL-5i 1-shot task with self-supervision using different superpixel segmentation algorithms.

| method | PFENet | SS-PFENet | | | |
|---|---|---|---|---|---|
| | | gridding | SLIC | HED | Felzenszwalb |
| meanIoU | 60.7 | 57.5 | 62.9 | **63.3** | 63.2 |

We also compared three different pseudo-class sampling strategies: 1) select the superpixel with the highest activation score calculated in Eq. 4.5, 2) sample a superpixel with the top-5 activation scores, 3) sample a superpixel from the entire image randomly. The meanIoUs on the PASCAL-5i 1-shot task are reported in Table 4.4, with the "top-5" strategy performing the best. This is understandable as on the one hand, top-5 provides more pseudo-classes than top-1 for training; on the other hand, selecting a random superpixel without ranking may result in over-segmenting, for example, a background region, e.g. sky, to be different to other background regions, e.g. other superpixels

of the sky, thus misguiding the model.

Table 4.4: Performance of ResNet50-PFENet on PASCAL-5i task with self-supervision when choosing pseudo-class from superpixels with the top-n activation scores.

| method | PFENet | SS-PFENet | | |
|---|---|---|---|---|
| pesudo-class choice range | | all | top-5 | top-1 |
| meanIoU | 60.7 | 58.5 | **63.2** | 63.0 |

### 4.6.3 Effect of Novel Class in Training Background

Though there is no label of testing classes available during training, their existences in the training images could still lead to advantage when the sampled *pseudo-classes* overlap with those regions. This ablation study examines the performance of the purposed method without such advantage by masking all pixels belonging to the testing classes in training images so these testing class objects will not be part of any pseudo-classes during training. As shown in Table 4.5, the meanIoU decreased only by 0.1% and 0.4% respectively for PANet and PFENet architecture. This minor decrease indicates the ability of the proposed self-supervision to generalise to completely unseen novel classes.

Table 4.5: meanIoU results on PASCAL-5i [239]. All models adopts ResNet-50 as backbone and best performance achieved by same model are bolded. "SS-" refers to models with self-supervision and "-mask" refers to models trained with pixels belonging to novel classes masked.

| method | fold1 | fold2 | fold3 | fold4 | mean |
|---|---|---|---|---|---|
| PANet[248] | 42.3 | 58.0 | 51.1 | 41.3 | 48.2 |
| SS-PANet | **49.3** | **60.8** | 53.9 | **45.2** | **52.3** |
| SS-PANet-masked | 49.1 | 60.7 | **54.0** | 45.0 | 52.2 |
| PFENet[248] | **61.7** | 69.5 | 55.4 | 56.3 | 60.7 |
| SS-PFENet | 58.9 | **69.9** | **66.4** | **57.7** | **63.2** |
| SS-PFENet-masked | 58.8 | 69.7 | 65.4 | 57.3 | 62.8 |

### 4.6.4 Effect of $\alpha$ Parameter

Table 4.6: Performance of ResNet-50 PFENet on PASCAL-5i [239] with self-supervision using different $\alpha$ values.

| $\alpha$ | 0.0 | 0.25 | 0.50 | 0.75 | 1.0 |
|---|---|---|---|---|---|
| meanIoU | 60.7 | 62.5 | **63.2** | 61.2 | 59.5 |

We varied the self-supervised loss scaling factor $\alpha$ and reported the performance achieved on PASCAL-5i [239] in Table 4.6.$\alpha = 0.5$ appears to be the optimal value. This is reasonable because the pseudo labels generated by superpixel segmentation are inaccurate (over-segmentation) and thus less aligned with our objective than ground-truth annotations.

## 4.7 Conclusion and Discussion

In this work, we first raised the issue of the dependency between the few-shot segmentation performance and the inherent complexion of individual *query* images, specifically, the presence of additional classes. We have provided quantitative evidence showing accuracy decreases in segmentation for *novel* classes, as a result of more latent objects appearing in the background of the *target* class of interest.

We directly address the lack of background class supervision, by proposing a novel strategy that generates self-supervising pseudo-labels for the non-target classes. A set of extensive experimental results have shown that the proposed approach consistently improved the few-shot segmentation performance on images that contain multiple classes of objects, for different network architectures on multiple datasets. Finally, we showed that different classes were indeed better disentangled in the embedding space using our method.

The proposed methodology is not limited by the choice of pseudo-supervision generation method, here we proposed a superpixel-based approach that is simple and effective. Despite the existence of more sophisticated deep-learning techniques like the segment anything model (SAM) [347], we opt against their utilisation in this work to avoid biased assessments. The choice is made to prevent data leakage, given the overlap between training data employed in deep-learning methods and the evaluation dataset. Nevertheless, in domains such as medical imaging where such concerns are less prevalent, future works can investigate the feasibility of leveraging deep-learning frameworks for pseudo-class generation.

Moreover, aside from inter-class discrimination, future studies might investigate methods to enhance model predictability by accommodating shape, colour and other variations within classes.

# Chapter 5

# Prototypical few-shot segmentation for cross-institution male pelvic structures with spatial registration

This chapter continues with the few-shot segmentation problem while diverting into the medical imaging field and introducing the generalisation challenge caused by domain shifts across institutions.

The prowess that makes few-shot learning desirable in medical image analysis is the efficient use of the *support* image data, which are labelled to classify or segment new classes, a task that otherwise requires substantially more training images and expert annotations. This work describes a fully 3D prototypical few-shot segmentation algorithm, such that the trained networks can be effectively adapted to clinically interesting structures that are ab-

sent in training, using only a few labelled images from a different institute. First, to compensate for the widely recognised spatial variability between institutions in episodic adaptation of novel classes, a novel *spatial registration mechanism* is integrated into prototypical learning, consisting of a segmentation head and an spatial alignment module. Second, to assist the training with observed imperfect alignment, *support mask conditioning module* is proposed to further utilise the annotation available from the support images. Extensive experiments are presented in an application of segmenting eight anatomical structures important for interventional planning, using a data set of 589 pelvic T2-weighted MR images, acquired at seven institutes. The results demonstrate the efficacy in each of the 3D formulation, the spatial registration, and the support mask conditioning, all of which made positive contributions independently or collectively. Compared with the previously proposed 2D alternatives, the few-shot segmentation performance was improved with statistical significance, regardless whether the support data come from the same or different institutes. The preliminary version of this work has been accepted as a International Symposium on Biomedical Imaging conference paper [3] and its full version has been accepted as a journal paper by Medical Image Analysis [4].

We first explain the motivation for introducing the task of few-shot cross-institutional segmentation in Section 5.1 and the task definition in Section 5.2, followed by reviewing related works for cross-insitution learning in Section 5.3. The proposed method, experiments and results are detailed in in Section 5.4, 5.5 and 5.6.

## 5.1 Introduction

Multi-structure segmentation is one of the fundamental computing tasks in medical imaging applications, found in diagnosis, treatment, and monitoring, and remains a research interest. Diagnosis of varieties of diseases can be assisted by quantifying the morphology, or its change, of multiple structures. For example, brain disorders, including Alzheimer's disease [128] and Parkinson [129], are associated with abnormal volumes or shapes of neurological regions. Identifying brain structures is the key to many quantitative studies, such as functional activation mapping and brain development analysis [130]. Minimally invasive treatments often benefit from careful planning of both interventional instruments and guidance imaging, with respect to segmented patient-specific anatomical structures. In endoscopic pancreaticobiliary procedures, as previously reported, image guidance that displays registered anatomical models outside the endoscopic field-of-view helps the surgeon during targeting and navigation [131].

This work is primarily concerned with segmenting multiple organs and urologically interesting structures on T2-weighted MR images from prostate cancer patients, to plan targeted biopsy, focal therapy, and, increasingly, other therapeutic procedures such as radiotherapy. As described in Chapter 2, accurate segmentation of these structures helps with targeting suspected cancerous regions found in multiparametric MR imaging with respect to the prostate gland and avoiding vulnerable surrounding structures, such as rectum, bladder and neurovascular bundles, for minimising risks of infection, impotence and other potential injury and complications [28].

The data-driven representation learning enabled by deep neural networks has

led to promising segmentation results in multi-structure segmentation tasks, for example, in neuroimaging [348] and abdominal organs [349]. Automating this task reduces the current requirement of manual segmentation which is often associated with costs in expertise and intra- and inter-observer variations [350]. However, recent supervised segmentation methods mostly rely on large data sets with full annotations, subject to similar limitations in labelling, albeit only with training. On the other hand, few-shot image learning aims to classify unseen classes using only a few labelled examples [215, 351]. For medical image segmentation, such novel classes may represent new types of organ or anatomical regions, whose expert annotations are not available for large training data sets. Using our intended interventional planning application as an example, different biopsy or therapy procedures may require different anatomical structures and pathological regions to be annotated during the planning stage. Prostate zonal structures, if not routinely segmented for MR-planning of radiotherapy, may provide more precise target localisation in registration-assisted ultrasound-guided focal ablation or a new therapy.

The now well-recognised performance loss in cross-institute generalisation of deep learning models [75] has motivated a body of research such as domain adaptation [352, 353] and federated learning [354]. This work focuses on few-shot segmentation with the practically important cross-institution context (as shown in Figure 5.1), which aims to segment a *novel class* from a *query* image, given only a few *support* images and their binary masks of the novel class, from a *novel institution* where the limited labelled data are available. In other words, the model should be able to simultaneously adapt to both novel classes and novel institutions.

To improve the inter-class and inter-institute generalisation, this work first examines the manifestation of the performance-reducing inter-institute vari-

ability in prototypical few-shot image segmentation algorithms. In particular, we investigate the impact of spatial alignment, or the lack thereof, between support and query data from different institutes, a key component in such a prototypical learning paradigm, on the few-shot segmentation accuracy.

First, addressing one of the previously identified challenges of spatial inconsistency [248], also found in medical image applications [251, 260], we develop a *spatial registration mechanism* to align the support and query images prior to the comparison between the two. This spatial registration mechanism consists of a *segmentation head* and a *spatial alignment module*, trained end-to-end, and is motivated by medical-image-specific observations of the difference between intra- and inter-institution data characteristics, due to different scanners and local imaging protocols, discussed further in Section 5.4.

Second, we propose an additional *support mask conditioning module*, also trained end-to-end, to enforce the conditioning on the available novel class labels. The conditioning module is empirically designed to work together with the spatial registration mechanism, to maximise utilisation of the few and prized support masks.

In addition to the evaluation of efficacy due to each proposed component, this work also demonstrates the benefit of the proposed 3D formulation, replacing existing 2D neural network-based few-shot segmentation approaches [355, 334, 250, 356, 251, 252, 258, 260].

One widely identified problem in both developing and evaluating multi-structure segmentation algorithms is the lack of a sizable labelled data set. To the date of publication of this work, there were no multi-structure annotations publicly available for pelvic MR images. Through this work, all of our manual labels from open data sets have been made available at `https:`

//zenodo.org/record/7013610, for aiding the reproducibility of this work and, potentially, for other urologic or radiologic tasks concerning multiple pelvic anatomical structures.

Our contributions are summarised as follows:

- We introduced the cross-institution few-shot segmentation task to address the data scarcity problem specifically faced in medical applications.
- We first proposed to extend prototypical neural network to 3D for few-shot multi-class segmentation, which requires fewer parameters while achieving similar performance to its 2D counterpart.
- We developed an spatial registration mechanism and a support mask conditioning module, directly addressing the observed limitations in medical image few-shot segmentation, for improving generalisation across data from different institutions.
- We presented extensive ablation studies to investigate the impact of the proposed individual components, the increasing number of support data, the varying size of the training set, and the permutations in the available institutes.
- We published all expert annotations based on public image data sets at https://zenodo.org/record/7013610, which includes full segmentation of eight distinctive male lower pelvic structures on 589 3D MR images.

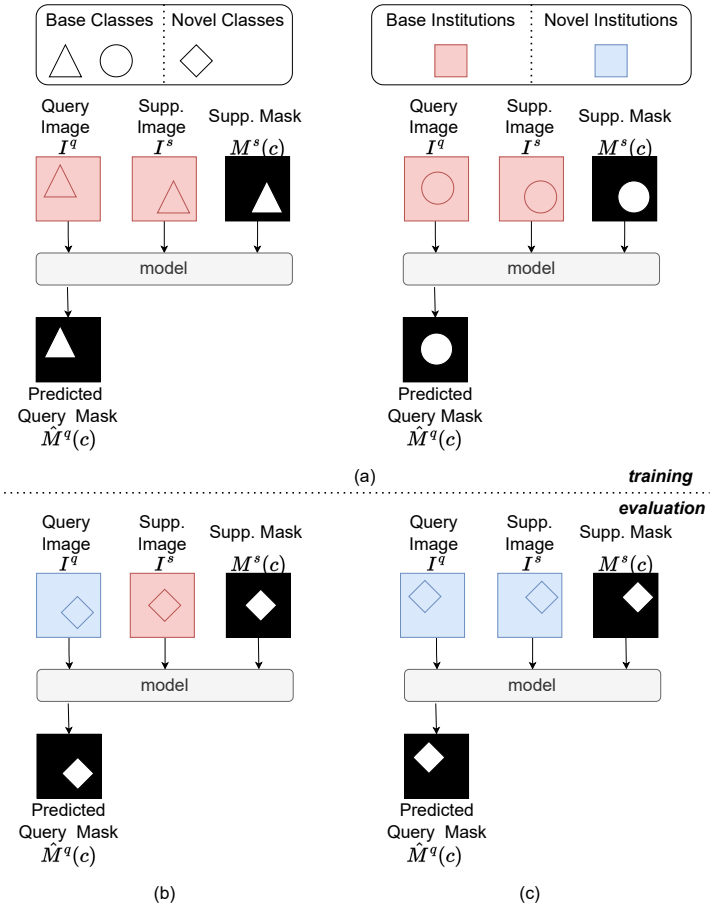



Figure 5.1: Visualisation of the proposed cross-institution few-shot segmentation task. As shown in (a), the task aims to train a model on the base data set including images from base institutions with corresponding masks on base classes and generalise to novel institutions as well as novel classes. The model is evaluated by segmenting novel classes from query images acquired by novel institutions, while support images come from either base or novel institutions, respectively shown in (b) and (c).

## 5.2 Task Definition

This section introduces the few-shot cross-institutional segmentation task focusing on few-shot segmentation with the practically important cross-institution context (as shown in Figure 5.1), which aims to segment a novel class from

a query image, given only a few support images and their binary masks of the novel class, from a *novel institution* where the limited labelled data are available. In other words, the model should be able to simultaneously adapt to both novel classes and novel institutions.

As mentioned in Section 4.1, $K$-shot few-shot segmentation tasks to predict the novel class $c \in \mathcal{C}_{\text{novel}}$ binary mask of a query image $I^{\text{que}}$ given a support set $\mathcal{S} = \{(I^{\text{sup},i}, M_{(c)}^{\text{sup},i})\}_{k=1}^{K}$ including $K$ examples image-mask pairs of a class $c \in \mathcal{C}_{\text{novel}}$ sampled from a set of novel classes isolated from training. Specifically, for few-shot 3D MRI segmentation, denote $(I, M_{(c)})$ as an single-class image-mask pair, $I \in \mathbb{R}^{W \times H \times D}$ is MRI of width $W$, height $H$ and depth $D$ with $M_{(c)} \in \mathbb{R}^{W \times H \times D}$ being the corresponding binary mask of class $c$.

Here one further variability, the institution from which the images are acquired, is introduced. Specifically, consider two disjoint sets of institutions $\mathcal{V}_{\text{base}}$ and $\mathcal{V}_{\text{novel}}$. We are interested in the performance of the model trained with data acquired from base institutions $\mathcal{V}_{\text{base}}$ when the support set $S$ is made of image-mask pairs acquired from a novel institution $u \in \mathcal{V}_{\text{novel}}$ isolated from training.

### 5.2.1 Dataset

For each institution $v \in \mathcal{V}$, $\mathcal{I}_v$ denotes the set of all images, which could be separated into disjoint training and test subsets $\mathcal{I}_v^{\text{train}}$ and $\mathcal{I}_v^{\text{test}}$. A base data set $\mathcal{D}_{\text{base}}$ is formed with training images and the corresponding labels of the base classes from all the base institutions.

$$\mathcal{D}_{\text{base}} = \bigcup_{v \in \mathcal{V}_{\text{base}}} \bigcup_{I \in \mathcal{I}_v^{\text{train}}} \bigcup_{c \in \mathcal{C}_{\text{base}}} \{(I, M_{(c)}\}. \tag{5.1}$$

where Similarly, test images and the corresponding labels of novel classes from all base institutions form another data set,

$$\mathcal{D}_{\text{base_ins_novel_cls}} = \bigcup_{v \in \mathcal{V}_{\text{base}}} \bigcup_{I \in \mathcal{I}_v^{\text{test}}} \bigcup_{c \in \mathcal{C}_{\text{novel}}} \{ \ (I, \ M_{(c)}) ) \ \} \tag{5.2}$$

together with the data set of images from the novel institutions and the novel classes' labels,

$$\mathcal{D}_{\text{novel_ins_novel_cls}} = \bigcup_{v \in \mathcal{V}_{\text{novel}}} \bigcup_{I \in \mathcal{I}_v} \bigcup_{c \in \mathcal{C}_{\text{novel}}} \{(I, \ M_{(c)})\} \tag{5.3}$$

a novel data set is built focusing on novel classes:

$$\mathcal{D}_{\text{novel}} = \mathcal{D}_{\text{base_ins_novel_cls}} \ \cup \ \mathcal{D}_{\text{novel_ins_novel_cls}}. \tag{5.4}$$

Therefore, the cross-institution segmentation task aims to train a model on the base data set $\mathcal{D}_{\text{base}}$ and generalise to the novel data set $\mathcal{D}_{\text{novel}}$ which contains novel classes on both base and novel institutions.

### 5.2.2 Evaluation

Following the few-shot setting described in [250] and [334], the model is tasked to segment a novel class $c \in \mathcal{C}_{\text{novel}}$ in a query image $I^{\text{que}}$ acquired from a novel institution $v^{\text{que}} \in \mathcal{V}_{\text{novel}}$ with only $K$ support examples $\{(I_k^{\text{sup}}, M_{(c),k}^{\text{sup}})\}_{k=1}^{K}$ from institution $v^{\text{sup}}$ which could be same as or different from the query institution $v^{\text{que}}$. The predicted query segmentation mask $\hat{M}_{(c)}^{\text{que}}$ is compared with the ground truth query label $M_{(c)}^{\text{que}}$ with a segmentation metric like Dice score [357].

During evaluation, as detailed in Algorithm 3, all images from the novel

institution were considered query images for evaluation for each of the novel classes. Results could be grouped by support institution $v^{\text{supp}}$ to analyse the effect of cross-institutinal effect.

## 5.3 Related Work in Cross-institution learning

The proposed cross-institution few-shot segmentation task aims to segment a novel class in images from a novel institution, with support images and labels from the same novel or other non-novel institutions. Although the objective is also to generalise on novel data, it differs from federated learning and domain adaptation due to their constraints on data privacy, accessibility, and availability. However, an optimal gain in efficient use of labelled data could be achieved by combining these methods with few-shot learning.

Federated learning is a learning paradigm that targets the problem of data governance and privacy by training algorithms collaboratively without the need for physically exchanging the data themselves, sometimes requiring the compliance of varying access policy [358]. It has been used in different medical imaging applications [354], but the focus is mainly on improving performance in trained classes without the need for generalisation in a novel class.

There exist three different domain adaptation methods: supervised, unsupervised, and semi-supervised. While sharing the same overarching objective, i.e. generalising to new domains such as novel classes, the supervised domain adaptation [359] does not explicitly focus on the data scarcity as few-shot learning. Unsupervised methods [360], on the other hand, often assume large-scale data set from the novel institution but without labels, thus paying different attention from that of cross-institution few-shot learning. Combin-

**Algorithm 3** Few-shot Cross-institutional Evaluation

**Input :** Neural network $g_\theta$ with parameters $\theta$
Images $\mathcal{I}_v^{\text{test}}$ for all base institutions $v \in \mathcal{V}_{base}$
Images $\mathcal{I}_v$ for all novel institutions $v \in \mathcal{V}_{novel}$
Masks for all novel classes in $\mathcal{C}_{\text{novel}}$

**Output:** Metrics for all, novel, and base institutions: ALL_METRIC, NOVEL_METRIC, and BASE_METRIC

**for** $I \in \bigcup_{v \in \mathcal{V}_{novel}} \mathcal{I}_v$ **do**
  **for** $v \in \mathcal{V}_{novel}$ **do**
    **for** $c \in \mathcal{C}_{novel}$ **do**
      **if** $M_{(c)}$ *satisfies* $\sum_{x=1}^{W}\sum_{y=1}^{H}\sum_{z=1}^{D} m^i_{(c)xyz} > 0$ **then**
        $(I^{\text{que}}, M^q_{(c)}) \leftarrow (I, M_{(c)})$
        $\mathcal{S} \leftarrow \{\}$
        $k \leftarrow 1$
        SAMPLED_SET $\leftarrow \{I^{\text{que}}\}$
        **while** $k \leq K$ **do**
          Sample $I^{\text{sup}} \in \mathcal{I}_v$ **if** $I \notin$ *SAMPLED_SET* **then**
            **if** $M_{(c)}$ *satisfies* $\sum_{x=1}^{W}\sum_{y=1}^{H}\sum_{z=1}^{D} m^i_{(c)xyz} > 0$ **then**
              $\mathcal{S} \leftarrow \mathcal{S} \cup (I, M_{(c)})$
              SAMPLED_SET $\leftarrow$ SAMPLED_SET $\cup \{I\}$
              $k \leftarrow k + 1$
            **end**
          **end**
        **end**
        Predict $\bar{M}^{\text{que}}_{(c)} = g(I^{\text{que}}, \mathcal{S}; \theta)$
        Transform predicted probability mask $\bar{M}^{\text{que}}_{(c)}$ to segmentation mask $\hat{M}^q_{(c)}$
        Comparing $\hat{M}^q_{(c)}$ and $M^q_{(c)}$)
        Update ALL_METRIC
        Update NOVEL_METRIC
      **end**
    **end**
  **end**

**for** $v \in \mathcal{V}_{base}$ **do**
  **for** $c \in \mathcal{C}_{novel}$ **do**
    **if** $M_{(c)}$ *satisfies* $\sum_{x=1}^{W}\sum_{y=1}^{H}\sum_{z=1}^{D} m^{i}_{(c)xyz} > 0$ **then**
      $(I^{\text{que}}, M^{q}_{(c)}) \leftarrow (I, M_{(c)})$
      $\mathcal{S} \leftarrow \{\}$
      $k \leftarrow 1$
      SAMPLED_SET $\leftarrow \{I^{\text{que}}\}$
      **while** $k \leq K$ **do**
        Sample $I^{\text{sup}} \in \mathcal{I}_v$ **if** $I \notin$ *SAMPLED_SET* **then**
          **if** $M_{(c)}$ *satisfies* $\sum_{x=1}^{W}\sum_{y=1}^{H}\sum_{z=1}^{D} m^{i}_{(c)xyz} > 0$ **then**
            $\mathcal{S} \leftarrow \mathcal{S} \cup (I, M_{(c)})$
            SAMPLED_SET $\leftarrow$ SAMPLED_SET $\cup \{I\}$
            $k \leftarrow k + 1$
          **end**
        **end**
      **end**
      Predict $\bar{M}^{\text{que}}_{(c)} = g(I^{\text{que}}, \mathcal{S}; \theta)$
      Transform predicted probability mask $\bar{M}^{\text{que}}_{(c)}$ to segmentation mask $\hat{M}^{q}_{(c)}$
      Comparing $\hat{M}^{q}_{(c)}$ and $M^{q}_{(c)}$)
      Update ALL_METRIC
      Update BASE_METRIC
    **end**
  **end**
**end**
**end**

ing both techniques, semi-supervised methods [361, 362] can leverage labelled and unlabelled data sets more efficiently.

Despite the distinct focuses, techniques and methodologies from federated learning and domain adaptation have indeed been considered in developing our cross-institute few-shot segmentation approach. For example, the spatial normalisation and divergence between features have commonly been adopted in federated learning [363] and feature-level domain adaptation [364], respectively.

## 5.4 Method

This work adopts the common episodic training paradigm (detailed in Section 4.2.2 and Algorithm 1) that simulates the few-shot task during training. Each *episode* consists of a query $(I^{\text{que}}, M^{\text{que}}_{(c)})$ and $K$ support $\mathcal{S} = \{(I^{\text{sup},i}, M^{\text{sup},i}_{(c)})\}_{k=1}^{K}$ image-mask pairs, for a base class $c \in \mathcal{C}_{\text{base}}$ sampled from base classes.

### 5.4.1 Local Prototype Extraction

To extract location-sensitive local prototypes [258], images of spatial size $W \times H \times D$ are partitioned into overlapping windows $g \in G$ of size $\alpha_w W \times \alpha_h H \times \alpha_d D$ with the equidistant spacing between window centres being half of the window size. As shown in Figure 5.2, for each window $g \in G$, two local prototype feature vectors $h^g_{(c)}$ and $h^g_{(0)}$ are calculated where class 0 refers to the background.

Specifically, for 3D MRI few-shot segmentation, given an image $I \in \mathbb{R}^{W \times H \times D}$ with its corresponding mask $M_{(c)} \in \mathbb{R}^{W \times H \times D}$, a feature map $F \in \mathbb{R}^{\Lambda \times W \times H \times D}$

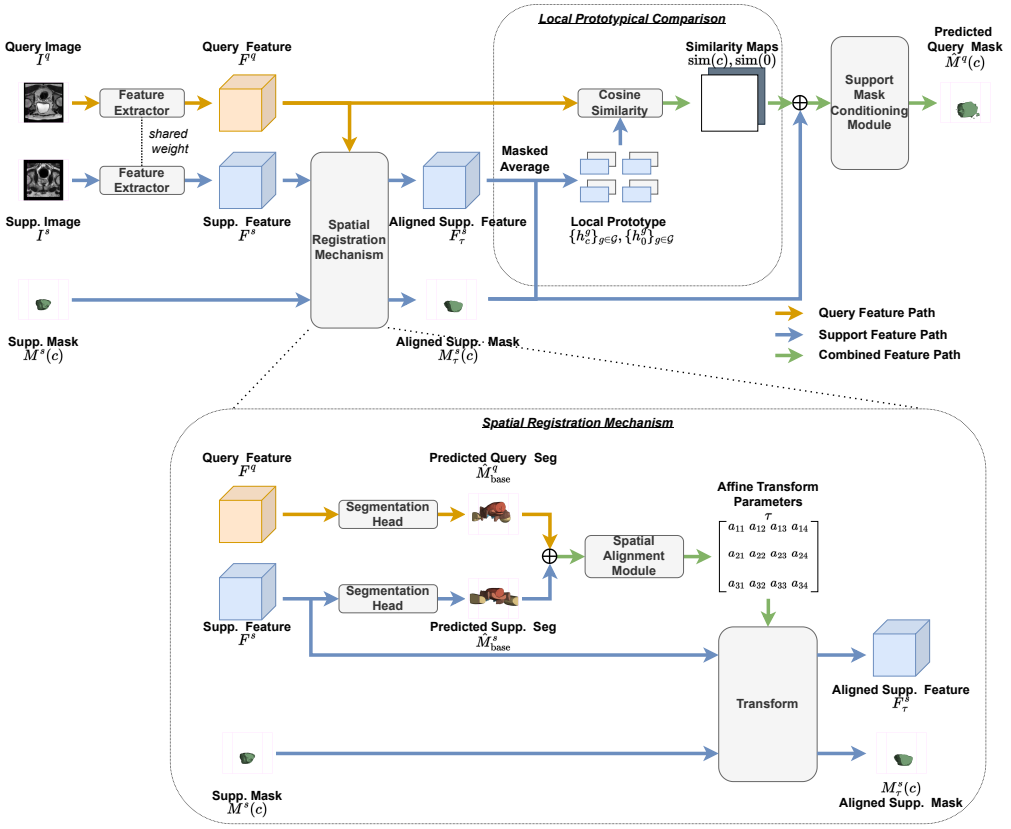



Figure 5.2: Overview of the proposed method. A shared feature extractor outputs query/support features from query/support images. The spatial registration mechanism spatially registers support feature/mask towards query. Foreground/background similarity maps are derived through local prototypical comparison, concatenated with the aligned support mask and processed by the support mask conditioning module to make the final prediction

could be extracted by the feature extractor with $\Lambda$ channels. Denote the components of $M_{(c)}$ and $F$ respectively as $m_{(c)xyz}$ and $f_{\lambda xyz}$ where the subscript $\lambda = 1, ..., \Lambda$ is used as channel index whilst $x = 1, ..., W$, $y = 1, ..., H$ and $z = 1, ..., D$ are used as location coordinate indices. A local class prototype $h^g_{(c)}$ for window $g$ could be derived by averaging the support feature map

$F^{\text{sup}}$ over locations belonging to class $c$ and within the window.

$$h^{g}_{(c)} = \frac{\sum_{(x,y,z)\in g} m^{\text{sup}}_{(c)xyz}[f^{\text{sup}}_{(c)xyz}, ..., f^{\text{sup}}_{\Lambda xyz}]^{\top}}{\sum_{(x,y,z)\in g} m^{\text{sup}}_{(c)xyz}} \tag{5.5}$$

For each query location $(x, y, z)$, $\mathcal{G}_{(x,y,z)}$ denote the set of all windows that contain the location: $\mathcal{G}_{(x,y,z)} = \{g \mid (x, y, z) \in g\}$. The local similarity between the local feature and the class is then calculated using the maximum cosine similarity over windows $g \in \mathcal{G}_{(x,y,z)}$ with the corresponding local prototype feature vectors $\{h^{g}_{(c)}\}_{g\in\mathcal{G}_{(x,y,z)}}$:

$$s_{(c)xyz} = \max_{g\in\mathcal{G}_{(x,y,z)}} \frac{[f^{\text{que}}_{(c)xy}, ..., f^{\text{que}}_{\Lambda xy}]^{\top} \cdot h^{g}_{(c)}}{\|[f^{\text{que}}_{(c)xy}, ..., f^{\text{que}}_{\Lambda xy}]^{\top}\|_2 \ \|h^{g}_{(c)}\|_2} \tag{5.6}$$

Predictions are made locally by comparing local query features $F^{\text{que}}$ with both the class and background prototype, resulting in a predicted query probability mask $\bar{M}^{\text{que}}_{(c)}$. Denote the background class as class 0, the class $c$ probability of location $(x, y, z)$ is derived by

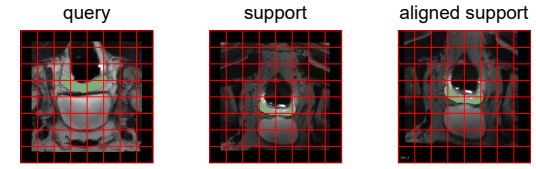
$$\bar{m}^{\text{que}}_{(c)xyz} = \frac{\exp(s_{(c)xyz})}{\exp(s_{(0)xyz}) + \exp(s_{(c)xyz})} \tag{5.7}$$

The model is trained to minimise the Dice loss between the predicted and ground-truth mask:

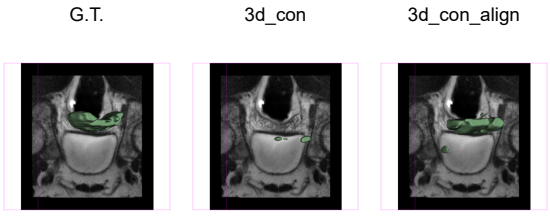
$$\mathcal{L}_{\text{fewshot}} = \mathcal{L}_{\text{Dice}}(\bar{M}^{\text{que}}, M^{\text{que}}) \tag{5.8}$$

### 5.4.2 Spatial Registration Mechanism

As discussed earlier, the differences in intra- and inter-institution variations pose challenges in the local prototypical network due to the varying image sizes, orientations, and voxel dimensions of the acquiring institution. The target structure in the query and support images can be distant (as in Figure 5.3). Therefore, they may not be included inside of the same or even adjacent windows, and this results in erroneous comparison between the query voxels of the structure and, thus irrelevant prototype vectors.




(a) ht




(b) ht

Figure 5.3: Qualitative results achieved by '3d_con' and '3d_con_align' on a misaligned query-support pair. (a) visualises the original query and support, as well as the support aligned towards query through the spatial registration mechanism. (b) visualises predictions made by '3d_con' and '3d_con_align' based on the same support-query pair.

However, while the absolute locations of target structures vary among images, the relative position between different structures remains consistent. This observation motivated spatial alignment of the query and the support images, $I^q$ and $I^{\text{sup}}$, illustrated in Figure 5.2, before extracting the local pro-

totype feature vectors. This spatial registration process is conjectured to alleviate the discrepancy between different institutions and therefore reduce the amount of cross-institution training data required for generalisation.

In this work, we consider an affine transformation to account for the above-discussed spatial difference with potential uncertainties due to variable imaging positioning, signal sampling and scanner calibration, although higher-degree transformation will also be of interest. Furthermore, to avoid repeated feature map extraction, we propose to apply the transformation directly on feature maps ($F^{\text{que}}$ and $F^{\text{sup}}$), rather than on images ($I^{\text{que}}$ and $I^{\text{sup}}$).

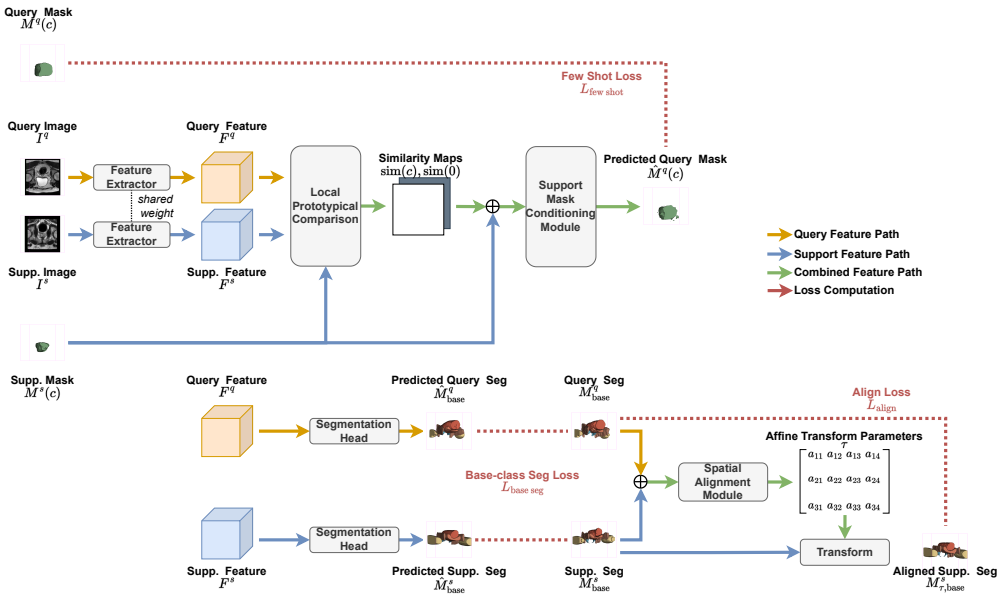



Figure 5.4: The training procedure of the proposed method. The model is trained to minimise the sum of few-shot loss (Equation 5.8), base class segmentation loss (Equation 5.10) and alignment loss (Equation 5.11).

The affine transformation prediction consists of two stages. Firstly, a shared base class segmentation head segments all base classes from the query and support feature maps, $F^{\text{que}}$ and $F^{\text{sup}}$, respectively. The multi-class predictions are denoted as $\{\hat{M}^{\text{que}}_{(c)}\}_{c\in\mathcal{C}_{\text{base}}}$ and $\{\hat{M}^{\text{sup}}_{(c)}\}_{c\in\mathcal{C}_{\text{base}}}$, for the query and sup-

port images, respectively. During evaluation, these predictions are concatenated and passed into the *spatial alignment module*, illustrated in Figure 5.2, which predicts an affine transformation matrix $\tau \in \mathbb{R}^{3\times4}$ of 12 degrees of freedom. During training, as illustrated in Figure 5.4, base classes segmentation masks $\{M^{\text{que}}_{(c)}\}_{c\in\mathcal{C}_{\text{base}}}$ and $\{M^{\text{sup}}_{(c)}\}_{c\in\mathcal{C}_{\text{base}}}$ are used for alignment prediction. Secondly, alignment $\tau$ is applied to the support feature map $F^{\text{sup}}$ and the label $\{\hat{M}^{\text{sup}}_{(c)}\}_{c\in\mathcal{C}_{\text{base}}}$, to obtain the aligned support feature map $F^{\text{sup}}_{\tau} = \tau \circ F^{\text{sup}}$ and the aligned support label for all base classes $\{M^{\text{sup}}_{\tau,(c)}\}_{c\in\mathcal{C}_{\text{base}}} = \{\tau \circ M^{\text{sup}}_{(c)}\}_{c\in\mathcal{C}_{\text{base}}}$, respectively. These aligned feature maps are then used to generate the local prototypes (as detailed in Section 5.4.1) for segmentation.

Two losses are defined for training the spatial registration mechanism. First, a Dice loss is defined for the multi-class segmentation:

$$\mathcal{L}_{\text{base seg}} = \mathcal{L}_{\text{Dice}}(\{\bar{M}^{\text{que}}_{(c)}\}_{c\in\mathcal{C}_{\text{base}}}, \{M^{\text{que}}_{(c)}\}_{c\in\mathcal{C}_{\text{base}}}) \tag{5.9}$$

$$+\mathcal{L}_{\text{Dice}}(\{\bar{M}^{\text{sup}}_{(c)}\}_{c\in\mathcal{C}_{\text{base}}}, \{M^{\text{sup}}_{(c)}\}_{c\in\mathcal{C}_{\text{base}}}) \tag{5.10}$$

The second loss aims to optimise alignment by minimising the Dice loss between the query label $\{M^{\text{que}}_{(c)}\}_{c\in\mathcal{C}_{\text{base}}}$ and the aligned support label $\{M^{\text{sup}}_{\tau,(c)}\}_{c\in\mathcal{C}_{\text{base}}}$ of all base classes:

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{Dice}}(\{M^{\text{que}}_{(c)}\}_{c\in\mathcal{C}_{\text{base}}}, \{M^{\text{sup}}_{\tau,(c)}\}_{c\in\mathcal{C}_{\text{base}}}) \tag{5.11}$$

In theory, the transformation could be applied the other way around, i.e. by applying the reverse alignment $\tau^{-1}$ to the query feature map $F^{\text{que}}$ resulting in the aligned query feature map $F^{\text{que}}_{\tau^{-1}} = \tau^{-1} \circ F^{\text{que}}$, to achieve spatial alignment between query and support features. A cycle-consistent two-way registration may also apply. However, once the aligned query mask $\hat{M}^{\text{que}}_{\tau^{-1},(c)}$ is predicted,

it needs to be inverted, in order to obtain the mask of the original query image $\hat{M}^{\text{que}}_{(c)} = \tau \circ \hat{M}^{\text{que}}_{\tau^{-1},(c)}$. The mechanism was adopted in this work for its computational efficiency without additional resampling in practice.

### 5.4.3 Support Mask Conditioning Module

During the prototype feature vector extraction using the novel class mask in the support image, the location-wise information may be made invariant to the transformation during the aggregation in Equation 5.5, which is designed for "normalising" cross-institution data but may also result in large spatial variability in the prediction. Therefore, a simple yet effective *support mask conditioning module* is proposed.

The support mask conditioning module takes as input a concatenation of the class similarity $s_{(c)}$, the background similarity $s_{(x)}$ and the aligned mask of the class in the support image $M^{\text{sup}}_{\tau,(c)}$, for the final prediction $\hat{M}^{\text{que}}_{(c)}$. Unlike the multiplication of support mask to support features for prototype calculation (in Equation 5.5), the direct use of support mask here provides a more direct route for the novel class information to the final segmentation task, similar to commonly designed shortcut layers for skipping networks.

Different from 3D medical segmentation algorithms using masks predicted from downsampled images to provide context information for high-resolution patches, the proposed method uses the segmentation of query images with location and shape information of the support mask.

### 5.4.4 Loss

Both the spatial registration mechanism and support mask conditioning module are trained with the original local prototypical network, with an overall

training loss function used in this study as follows:

$$\mathcal{L} = \mathcal{L}_{\text{fewshot}} + \mathcal{L}_{\text{base seg}} + \mathcal{L}_{\text{align}} \tag{5.12}$$

### 5.4.5 Multiple-shot Evaluation

Due to memory limitation, the training was carried out in one-shot paradigm, i.e. $K = 1$. During the evaluation, for the query image $I^q$ and each of the $K$ support images $I_k^{\text{sup}}$, the base class segmentation is predicted from the base class segmentation head, denoted by $\{\bar{M}_{(c)}^{\text{que}}\}_{c\in\mathcal{C}_{\text{base}}}$ and $\{\bar{M}_{(c)}^{\text{sup}}\}_{c\in\mathcal{C}_{\text{base}}}$, respectively. Among the $K$ support images, only the support image that is the most similar to the query image in terms of the cosine similarity on base class segmentation prediction, denoted by $I_{\grave{k}}^{\text{sup}}$, is chosen to calculate the local prototypes in (5.5). Precisely,

$$\grave{k} = \arg\max_k \frac{\bar{M}_{\text{base},k}^{\text{sup}} \cdot \bar{M}_{\text{base}}^{\text{que}}}{\|\bar{M}_{\text{base},k}^{\text{sup}}\|_2 \ \|\bar{M}_{\text{base}}^{\text{que}}\|_2}, \tag{5.13}$$

where the dot product and norm are calculated on the flattened predictions. The local prototypes are then calculated using the selected support example.

$$h_{(c)}^g = \frac{\sum_{(x,y,z)\in g} m_{(c)xyz}^{\text{sup},\grave{k}} [f_{(c)xyz}^{\text{sup},\grave{k}}, ..., f_{\Lambda xyz}^{\text{sup},\grave{k}}]^\top}{\sum_{(x,y,z)\in g} m_{(c)xyz}^{\text{sup},\grave{k}}} \tag{5.14}$$

## 5.5 Experiments

### 5.5.1 Data Set

State-of-the-art male lower-pelvic image datasets, such as INDEX [81], PROMISE12 [69], and NCI-ISBI [70], have focused exclusively on segmenting the prostate cap-

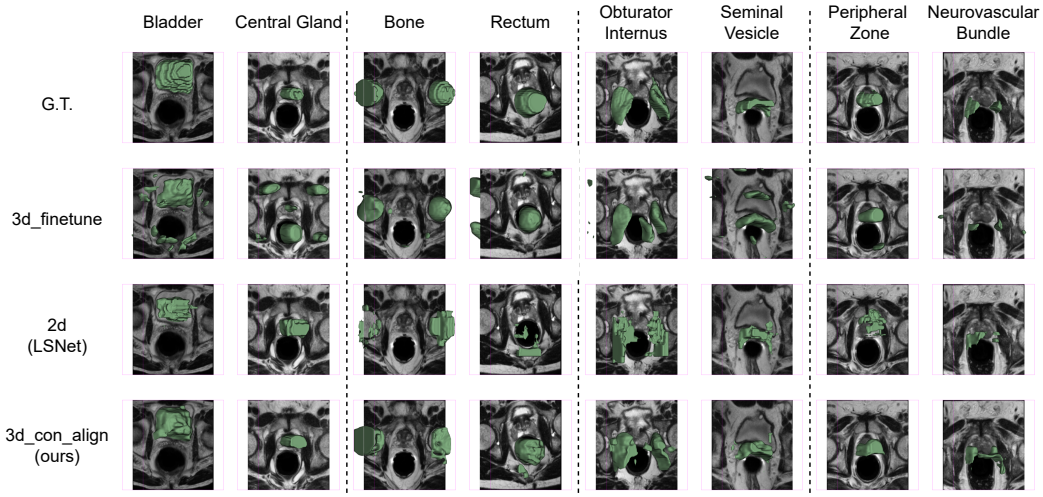



Figure 5.5: Qualitative result achieved by '3d_finetune' and '2d' baselines (detailed in Section 5.5.3) and our proposed method.

Table 5.1: The number of images acquired from each institution.

| | institution | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | total |
| # of images | 321 | 45 | 74 | 82 | 24 | 24 | 19 | 589 |

sule and its sub-regions, including the central gland (CG) and peripheral zone (PZ). Although these segments are crucial for PI-RAD scoring, biopsy targeting, and treatment planning, other anatomically relevant structures involved in planning have been largely overlooked.

This thesis addresses this gap by annotating and collecting labels for eight key anatomic structures (bladder, bone, central gland, neurovascular bundle, obturator internus, rectum, seminal vesicle, and peripheral zone) as shown in Figure 5.5. All eight structures are of planning interest to avoided during radio or focal therapy to avoid morbidity. Injury to the bladder during treatment can lead to cystitis, characterised by urinary frequency, pain, and hematuria. Damage to the femur bone may result in pathological fractures and osteonecrosis, leading to bone death. The neurovascular bundle, if in-

jured, can cause erectile dysfunction. Harm to the obturator internus muscle can cause pain, muscle weakness, and mobility issues. Rectal damage can result in proctitis, leading to rectal bleeding. Lastly, injury to the seminal vesicle can reduce semen production and cause infertility.

The data set includes 589 T2-weighted images acquired from the same number of patients collected by seven studies, INDEX [365], the SmartTarget Biopsy Trial [366], PICTURE [367], TCIA Prostate3T [368], Promise12 [69], TCIA ProstateDx (Diagnosis) [369] and the Prostate MR Image Database [370]. Further details are reported in the respective study references.

All segmentations were manually annotated by eight biomedical imaging researchers, with experience ranging from 2 to 10 years in the annotation of medical image data, each annotating a mixed-institution subset using an institution-stratified sampling. Each annotation has been reviewed at least once. The annotation process required around one thousand observer-hours, translating to over one and a half hours per 3D image. This extensive time requirement highlights the challenge for most local hospitals if they were to rely on conventional supervised learning methods for tool validation or adaptation, thereby underscoring the clinical relevance and efficiency of the proposed few-shot segmentation approach.

These images were divided into seven subsets based on the acquiring institution. The number of images acquired from each institution is anonymously summarised in Table 5.1. The cross-institution imaging protocols contain multiple scanners (two manufacturers with a mixed 1.5 and 3T field strengths), varying field-of-view and anisotropic voxels, in-plane voxel dimensions ranging between 0.3 and 1.0 mm and out-of-plane spacing between 1.8 and 5.4 mm.

The full segmentation masks and the derived intensity arrays from T2-weighted sequences, after pre-processing, used to produce the results in this study, are available at https://zenodo.org/record/7013610.

The eight lower-pelvic structures were randomly divided into four folds as shown in Table 5.2. In a cross-validation experiment, the classes contained in each fold were considered novel classes, the other three folds representing base classes. The images of each institution $v$ were then randomly sampled into training and testing subsets in a 3:1 ratio. A further validation set was formed with 12 images, 2 from each institution, randomly chosen from the novel data set $\mathcal{D}_{novel}$. Those images were excluded from the testing. Unless otherwise specified, the same data partitioning was used for all the results presented. All statistical conclusions are reported using paired Student's t-tests at the significance level of $\alpha_{t-test} = 0.05$.

Table 5.2: The eight structures are randomly divided into 4 folds.

| | structures |
|---|---|
| fold 1 | bladder, central gland |
| fold 2 | bone, rectum |
| fold 3 | obturator internus, seminal vesicle |
| fold 4 | peripheral zone, neurovascular bundle |

### 5.5.2 Implementation Details

All images were normalised, resampled and centre-cropped to an image size of $256 \times 256 \times 48$, with a voxel dimension of $0.75 \times 0.75 \times 2.5$ during pre-processing. Random rotation, translation and scaling were adopted for data augmentation during training.

The best training episode was chosen based on the performance of the validation sets. During the evaluation, all images from the novel institution were

considered query images for evaluation for each of the fold-specified novel classes. A binary Dice score for this novel class was calculated for each query image based on a sampled support image from each of the seven institutions, excluding the query. As described in Algorithm 3, the results were reported when support images were from 1) all institutions, 2) base institutions, and 3) novel institutions. The institution from which the support image comes is denoted as 'support ins'.

A 3D UNet was adopted as the feature extractor, whose architecture is detailed in Figure 5.7. $\alpha_w = \alpha_h = \frac{1}{8}$ and $\alpha_d = 1$ were selected for local prototypical comparison as detailed in Section 5.4.1. As shown in Figure 5.8, the support mask conditioning module comprised two convolutional layers. For the spatial registration mechanism, the base class segmentation head was a single convolutional layer, and the spatial alignment module was a GlobalNet [74] as specified in Figure 5.9. The models were trained using an Adam optimiser starting at a learning rate of $10^{-4}$ with a mini-batch size of 1. The implementation code has been released at https://github.com/kate-sann5100/CrossInstitutionFewShotSegmentation.

### 5.5.3 Compared Baseline Networks

For comparison, we report the results of the following baseline networks.

1. The '3d_finetune' network - The '3d_finetune' baseline implemented the same UNet as the feature extractor. It was pre-trained on the base data set to segment all base classes for 100 epochs using an Adam optimiser starting at a learning rate of $10^{-4}$. During evaluation, for each query, the pre-trained model is fine-tuned on the support images for 10 iterations before testing. Using a simple transfer learning strategy, this

baseline provides a reference as a "lower-bound" performance.

2. The '2d' network - LSNet [258] was adopted as the 2D episodic baseline. It adopted the same local-prototype comparison approach as detailed in Section 5.4.1 but using a 2D backbone based on ResNet-50 pre-trained on ImageNet, instead of 3D networks. To the best of our knowledge, this is the prototypical network closest to our work that has been proposed for medical image segmentation.

3. BiGRU [371] - Another recent few-shot medical segmentation method with a UNet-like network for 2D slice prediction and a bidirectional gated recurrent unit (GRU) for adjacent slices consistency.

4. The 'localnet (unsupervised)' network for multi-atlas segmentation - A non-rigid registration network, LocalNet [73], trained on the base data set images with no organ label supervision. Given a support-query pair, the model is trained to predict a dense displacement field that warps the support towards the query. For n-shot evaluation, $n$ dense displacement fields are predicted, each registering a support example towards the query. $n$ query mask predictions are derived by warping each support mask using the corresponding predicted dense displacement field. The $n$ query mask predictions make the final prediction through majority voting. The implementation was based on the open-source repository MONAI [1].

5. The 'localnet (supervised)' network for multi-atlas segmentation - A non-rigid registration network, LocalNet [73], trained on the base data set with masks from all classes including novel classes. Similar to the 'localnet(unsupervised)' network, the query mask is the warped support mask using the registration-predicted support-to-query dense dis-

placement field. The implementation was based on the open-source repository MONAI [1].

6. The '3d_supervised' network - A fully supervised 3D model was trained on the base data set images with masks from all classes. The results on the novel institution images are reported as an "upper-bound" performance.

### 5.5.4 Ablation Studies

#### 5.5.1 Ablation on different modules

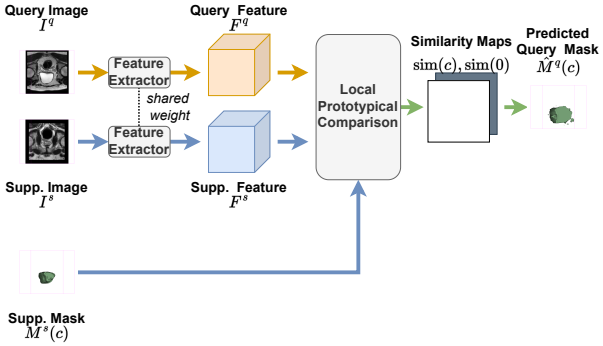
To assess the effectiveness of different modules in the proposed method, we report the results of the following variations.

1. The '3d' network variant - The proposed 3D local prototypical network, detailed in Section 5.4.1 and Figure 5.6(a), without the support mask conditioning or spatial registration.

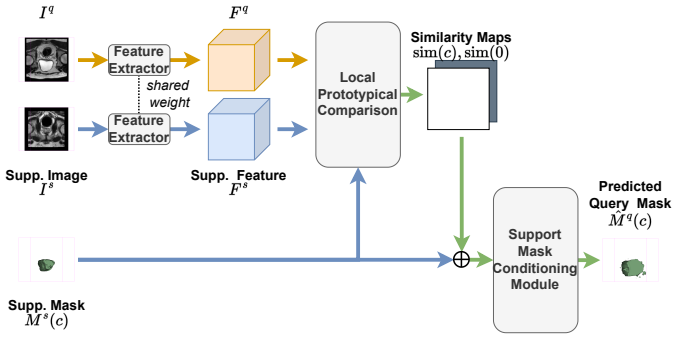
2. The '3d_con' network variant - The '3d' version with the support mask conditioning, but without the spatial registration, described in Section 5.4.3 and Figure 5.6(b)).

3. The '3d_align' network - The '3d' version of the proposed network (detailed in 5.6(c)) with the spatial registration mechanism, without the support mask conditioning.

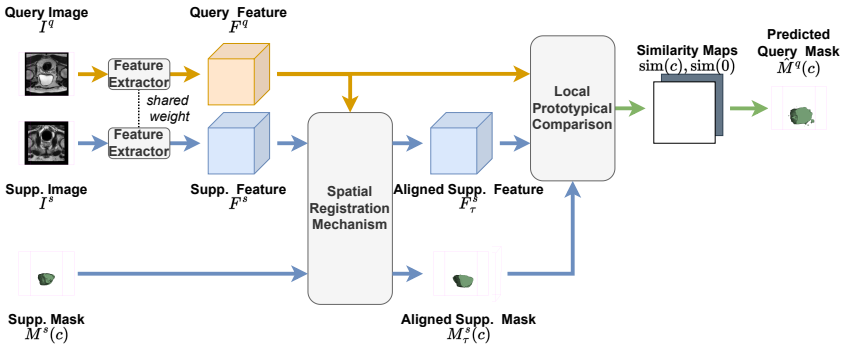
4. The '3d_con_align' network - The "complete" version of the proposed network with both the support mask conditioning module and the spatial registration mechanism, as shown in Figure 5.2.




(a) h




(b) h




(c) h

Figure 5.6: Illustration of the three model variations (a) '3d' (b) '3d_con' (c) '3d_align' experimented for ablation studies (Section 5.5.4)

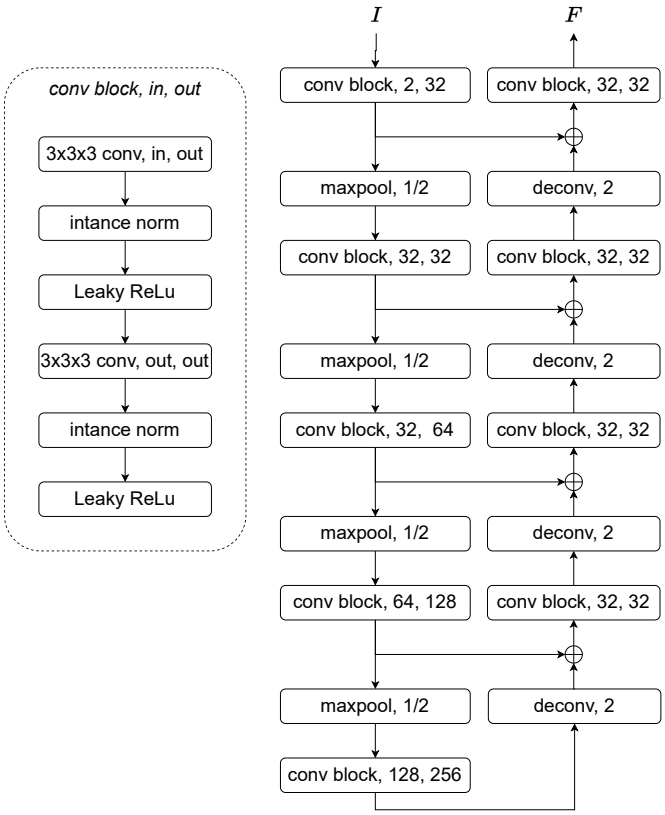



Figure 5.7: Feature extractor architecture.

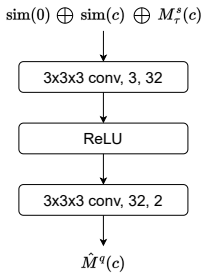



Figure 5.8: Support mask conditioning module architecture.

### 5.5.2 Ablation study on the number of shots

We report model performance when different numbers of support trios are available in each episode.

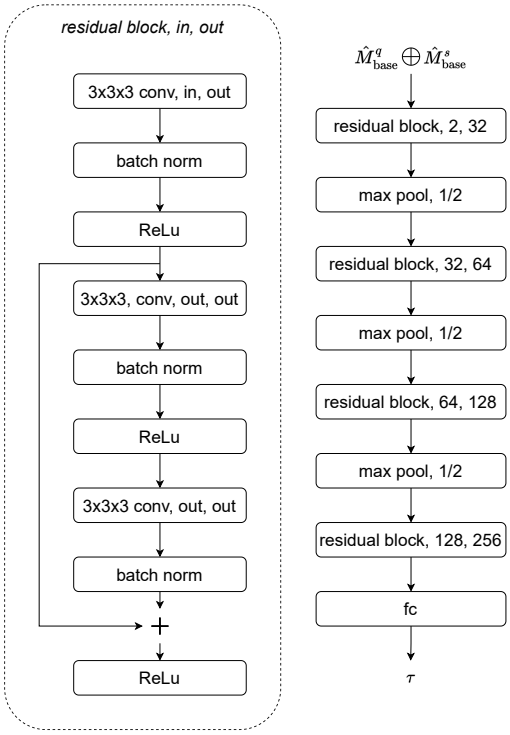



Figure 5.9: Spatial alignment module architecture.

### 5.5.3 Ablation on varying training data availability

To investigate the dependency of our proposed method on the training set, we report the performance of the proposed model (3d_con_align) trained on various availability in training sets. The 'half' and 'quarter' experiments respectively include $\frac{1}{2}$ and $\frac{1}{4}$ of the training subset $\mathcal{D}_u^{train}$ for each base institutions $u \in \mathcal{V}_{base}$, in order to test the impact of the training data set size on the few-shot segmentation performance. The same ratio between institutions was maintained in these experiments. Additionally, the 'half_single_ins' experiment tests the same number of images as the 'half' experiment, but all the images are sampled from the same institution (Institution 1).

To assess and compare the intra- and inter-institution generlisation, results are also reported when all the institutions, only the base institutions and

only the novel institutions were used as support institutions, denoted as 'all', 'base' and 'novel', respectively.

## 5.6 Results

Table 5.3: Dice score (%) achieved when institution 3 was adopted as the novel institution. 'support ins' refers to the institution from which the support images were sampled. Δ refers to the percentage difference between predictions made with support images from the base and novel institutions.

| model | support ins | Dice (%) | | | | | 95% Hausdorff distance (mm) | | | | | #param (million) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | fold1 | fold2 | fold3 | fold4 | mean | fold1 | fold2 | fold3 | fold4 | mean | |
| 3d_finetune | all | 48.07 | 44.78 | 43.92 | 29.25 | 41.50 | 61.47 | 69.78 | 60.33 | 57.02 | 62.15 | 5.75 |
| | base | 45.74 | 39.81 | 41.69 | 26.02 | 38.32 | 63.22 | 73.42 | 62.33 | 59.24 | 64.55 | |
| | novel | 62.03 | 74.56 | 57.31 | 48.63 | 60.63 | 50.98 | 47.96 | 48.32 | 43.72 | 47.74 | |
| | Δ | 26.26% | 46.60% | 27.25% | 46.48% | 36.80% | -24.02% | -53.09% | -28.99% | -35.51% | -35.21% | |
| BiGRU | all | 44.05 | 43.08 | 29.61 | 20.97 | 34.43 | 23.70 | 33.67 | 23.60 | 22.83 | 25.95 | 38.85 |
| | base | 44.05 | 42.74 | 29.21 | 20.62 | 34.15 | 23.73 | 34.06 | 23.85 | 23.20 | 26.21 | |
| | novel | 44.03 | 45.11 | 32.01 | 23.12 | 36.07 | 23.49 | 31.30 | 22.12 | 20.60 | 24.38 | |
| | Δ | -0.06% | 5.26% | 8.75% | 10.80% | 5.30% | -1.04% | -8.82% | -7.83% | -12.59% | -7.52% | |
| 2d | all | 49.23 | 44.51 | 36.34 | 27.44 | 39.38 | 24.03 | 30.86 | 29.51 | 30.77 | 28.79 | 23.63 |
| | base | 48.80 | 42.92 | 35.52 | 26.83 | 38.52 | 24.12 | 31.93 | 30.30 | 31.54 | 29.47 | |
| | novel | 51.80 | 54.07 | 41.28 | 31.08 | 44.56 | 23.52 | 24.39 | 24.83 | 26.13 | 24.72 | |
| | Δ | 5.80% | 20.62% | 13.94% | 13.68% | 13.55% | -2.54% | -30.92% | -22.03% | -20.72% | -19.24% | |
| 3d | all | 47.79 | 52.06 | 34.81 | 26.16 | 40.20 | 31.60 | 30.68 | 32.95 | 31.71 | 31.73 | 5.75 |
| | base | 47.89 | 51.07 | 34.12 | 25.44 | 39.63 | 31.69 | 31.03 | 33.35 | 32.19 | 32.07 | |
| | novel | 47.22 | 58.00 | 38.96 | 30.49 | 43.67 | 31.07 | 28.54 | 30.54 | 28.81 | 29.74 | |
| | Δ | -1.42% | 11.94% | 12.44% | 16.58% | 9.25% | -2.00% | -8.73% | -9.20% | -11.73% | -7.82% | |
| 3d_con | all | 58.34 | 46.11 | 40.43 | 30.94 | 43.96 | 27.31 | 27.15 | 27.33 | 26.38 | 27.04 | 5.75 |
| | base | 58.25 | 43.58 | 39.87 | 29.80 | 42.88 | 27.69 | 28.34 | 28.03 | 27.06 | 27.78 | |
| | novel | 58.84 | 61.30 | 43.81 | 37.78 | 50.43 | 25.04 | 20.02 | 23.14 | 22.30 | 22.63 | |
| | Δ | 1.00% | 28.91% | 8.99% | 21.12% | 14.98% | -10.56% | -41.54% | -21.12% | -21.31% | -22.76% | |
| 3d_align | all | 50.80 | 51.56 | 32.39 | 30.89 | 41.41 | 27.33 | 36.29 | 32.63 | 29.93 | 31.55 | 27.34 |
| | base | 50.60 | 51.17 | 31.25 | 30.82 | 40.96 | 27.35 | 36.82 | 33.01 | 30.42 | 31.90 | |
| | novel | 51.97 | 53.88 | 39.22 | 31.33 | 44.10 | 27.20 | 33.15 | 30.34 | 26.96 | 29.41 | |
| | Δ | 2.64% | 5.04% | 20.31% | 1.62% | 7.12% | -0.55% | -11.08% | -8.78% | -12.84% | -8.46% | |
| 3d_con_align | all | 59.36 | 60.38 | 45.73 | 37.60 | 50.77 | 22.66 | 29.63 | 27.92 | 26.41 | 26.66 | 27.35 |
| | base | 59.11 | 59.80 | 45.06 | 37.28 | 50.31 | 22.69 | 30.30 | 28.38 | 26.89 | 27.06 | |
| | novel | 60.86 | 63.87 | 49.77 | 39.54 | 53.51 | 22.49 | 25.63 | 25.19 | 23.54 | 24.21 | |
| | Δ | 2.89% | 6.37% | 9.47% | 5.70% | 5.98% | -0.91% | -18.21% | -12.65% | -14.26% | -11.79% | |
| localnet (unsupervised) | all | 25.64 | 38.99 | 21.83 | 11.20 | 24.42 | 28.74 | 27.21 | 24.78 | 26.97 | 26.93 | 5.94 |
| | base | 24.16 | 36.43 | 21.03 | 10.35 | 22.99 | 29.43 | 28.48 | 25.39 | 27.89 | 27.80 | |
| | novel | 34.52 | 54.37 | 26.61 | 16.34 | 32.96 | 24.60 | 19.55 | 21.10 | 21.47 | 21.68 | |
| | Δ | 30.01% | 33.00% | 20.98% | 36.67% | 30.25% | -19.63% | -45.70% | -20.32% | -29.93% | -28.22% | |
| localnet (supervised) | all | 72.70 | 72.14 | 52.84 | 42.65 | 60.08 | 19.36 | 22.81 | 21.78 | 24.00 | 21.99 | 5.94 |
| | base | 71.80 | 70.11 | 51.21 | 41.65 | 58.69 | 20.29 | 24.69 | 23.07 | 25.11 | 23.29 | |
| | novel | 78.11 | 84.33 | 62.61 | 48.61 | 68.42 | 13.77 | 11.57 | 14.03 | 17.29 | 14.17 | |
| | Δ | 8.09% | 16.86% | 18.20% | 14.31% | 14.21% | -47.33% | -113.31% | -64.37% | -45.26% | -64.39% | |
| 3d_supervised | N/A | 89.01 | 91.68 | 81.44 | 70.30 | 83.11 | 4.73 | 5.14 | 5.91 | 6.83 | 5.65 | 5.75 |

The Dice score and 95% Hausdorff distance achieved by variations of our

Table 5.4: Dice score (%) achieved when institution 4 was adopted as the novel institution. 'support ins' refers to the institution from which the support images were sampled. $\Delta$ refers to the percentage difference between predictions made with support images from the base and novel institutions.

| model | support ins | Dice (%) | | | | | 95% Hausdorff distance (mm) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | fold1 | fold2 | fold3 | fold4 | mean | fold1 | fold2 | fold3 | fold4 | mean |
| 3d_finetune | all | 37.61 | 29.23 | 33.06 | 20.25 | 30.04 | 59.83 | 71.87 | 50.66 | 51.62 | 58.49 |
| | base | 36.72 | 28.33 | 31.11 | 20.02 | 29.05 | 60.77 | 73.85 | 52.09 | 52.02 | 59.68 |
| | novel | 42.96 | 34.63 | 44.73 | 21.63 | 35.99 | 54.16 | 59.97 | 42.09 | 49.20 | 51.36 |
| | $\Delta$ | 14.53% | 18.17% | 30.45% | 7.43% | 19.29% | -12.21% | -23.13% | -23.75% | -5.74% | -16.22% |
| 2d | all | 42.09 | 29.00 | 32.49 | 24.46 | 32.01 | 29.52 | 42.65 | 28.04 | 31.97 | 33.05 |
| | base | 41.83 | 28.51 | 32.09 | 24.47 | 31.73 | 29.62 | 42.39 | 28.08 | 31.93 | 33.01 |
| | novel | 43.64 | 31.92 | 34.85 | 24.41 | 33.71 | 28.90 | 44.18 | 27.80 | 32.19 | 33.27 |
| | $\Delta$ | 4.14% | 10.68% | 7.92% | -0.26% | 5.87% | -2.50% | 4.05% | -1.00% | 0.80% | 0.79% |
| 3d_con_align | all | 57.58 | 37.42 | 37.42 | 30.18 | 40.65 | 23.02 | 37.56 | 30.22 | 28.52 | 29.83 |
| | base | 57.71 | 36.74 | 36.70 | 29.97 | 40.28 | 22.90 | 38.37 | 30.74 | 28.52 | 30.13 |
| | novel | 56.83 | 41.46 | 41.70 | 31.40 | 42.85 | 23.75 | 32.72 | 27.12 | 28.51 | 28.02 |
| | $\Delta$ | -1.54% | 11.39% | 11.99% | 4.56% | 6.00% | 3.58% | -17.26% | -13.36% | -0.04% | -7.52% |
| 3d_supervised | N/A | 88.31 | 88.44 | 80.49 | 70.78 | 82.00 | 5.14 | 7.72 | 6.49 | 6.71 | 6.51 |

proposed method as well as the baseline networks are presented for comparison. Table 5.3 and 5.4 summarise the network performances, with respect to different folds and the mean of all folds using institutions 3 and 4 as the novel institution, respectively. Other institutions were not used as novel institutions as they have either too many or too few samples such that the training or test set size will be too small. When Institution 3 is used as the novel institution (Table 5.3), the proposed method with both support mask conditioning module and spatial registration mechanism ('3d_con_align') outperformed 'BiGRU' by 16.34%, 16.16% and 17.44%, and outperformed the '2d' baseline by 11.39%, 11.79% and 8.95% in absolute Dice improvement when support images came from all, base and novel institutions, respectively. While the '3d_finetune' baseline achieved higher Dice than the '2d' baseline, its 95% Hausdorff distance is more than double of the '2d' baseline. This could be related to the commonly appearing false positive predictions as shown in Figure 5.5. Dice may still have small yet distant false positive predictions which caused higher Hausdorff distance. We refer the reader to our recent study for a discussion on this particular issue [372]. Similar improvements have

also been observed when Institution 4 was adopted as the novel institution. 8.64%, 8.55% and 9.15% absolute Dice improvements over the '2d' baseline were achieved by the proposed method, as support images came from all, base and novel institutions, respectively (Table 5.4).

The relative difference between performances when support images come from base and novel institutions, denoted as performance gap $\Delta$, are also summarised in Table 5.3 and Table 5.4. While achieving better performance overall, the proposed method reported smaller performance gap.

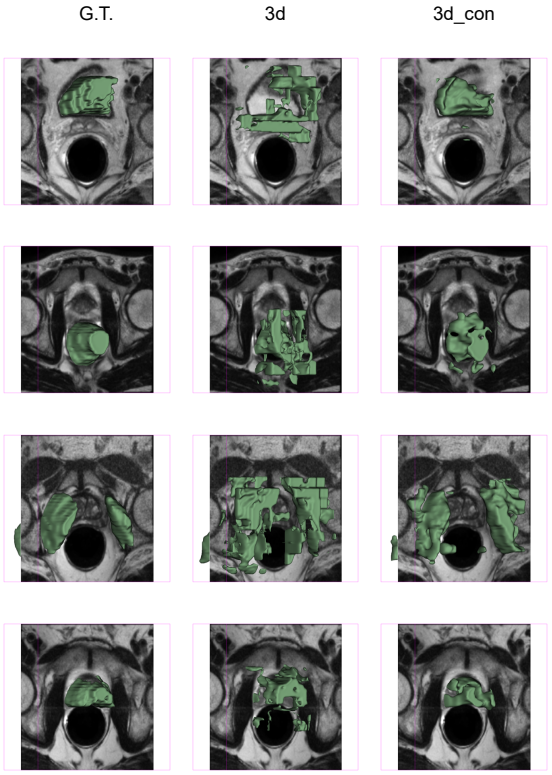



Figure 5.10: Qualitative results achieved by '3d' and '3d_con' on the same support-query pairs. '3d_con' resulted in more compact segmentation with smoother boundaries compared with '3d'

Adding support mask conditioning module alone ('3d_con') led 3.76%, 3.25% and 6.76% absolute increase in Dice score compared to '3d', when support im-

ages came from all, base and novel institutions, respectively. Qualitatively, it predicted more "compact" segmentation with smoother boundaries as shown in Figure 5.10. However, it also resulted in a higher $\Delta$ - a greater improvement was achieved when support and query came from the same institution, possibly because of its sensitivity to support-query (mis)alignment.

This was mitigated by the spatial registration mechanism which not only further improved the Dice score by 6.81% but also reduced $\Delta$ by 9.00%. Figure 10 shows an example where the target structure of the query and support were misaligned, resulting in segmentation failure. Aligning the support towards the query, the spatial registration mechanism considerably improved the performance.

Interestingly, when spatial alignment mechanism was available, the support mask conditioning module led to a further improvement - 9.36%, 9.35% and 9.41% absolute increase in Dice score from '3d_align' to '3d_con_align' when support images came from all, base and novel institutions, respectively.

It is also important to report that the proposed 3D network '3d' without the support mask conditioning module and the spatial registration mechanism contained 5.7 million parameters. It achieved comparable performance to the '2d' baseline with 23.5 million parameters, resulting in around 75% reduction in number of parameters. For reference, the complete version of our proposed method '3d_con_align' contained 27.3 million parameters (16% more parameters compared with '2d'), and had achieved 28.9% of relative Dice improvement over '2d'.

Furthermore, Table 5.5 and Table 5.6 report the mean Dice score achieved by '2d' and '3d_con_align' from different support-query institution combinations when institution 3 is the novel institution. Better performance was

Table 5.5: Mean Dice score (%) achieved at different (s_ins, q_ins) combinations by '2d' when Institution 3 was adopted as the novel institution, where s_ins and q_ins respectively refer to the institution from which the support and query were sampled. The best performance for each query institution was bolded. p-values were derived from paired t-tests performed between the Dice scores where the support institution equals the query institution and the maximum Dice scores achieved when the support institution differs from the query institution.

| | | s_ins | | | | | | | std | mean | p-value |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ins1 | ins2 | ins3 | ins4 | ins5 | ins6 | ins7 | | | |
| q_ins | ins1 | **48.41** | 44.67 | 39.78 | 33.23 | 43.89 | 42.28 | 29.84 | 6.11 | 40.30 | 8.14e-16 |
| | ins2 | 46.42 | **49.15** | 37.66 | 27.39 | 42.35 | 43.05 | 29.09 | 7.74 | 39.30 | 5.38e-04 |
| | ins3 | 41.60 | 38.36 | **44.56** | 37.14 | 40.34 | 37.30 | 36.37 | 2.73 | 39.38 | 1.62e-16 |
| | ins4 | 34.39 | 32.12 | 35.46 | **39.00** | 31.11 | 29.77 | 35.28 | 2.89 | 33.88 | 0.71 |
| | ins5 | 43.91 | 44.63 | 32.26 | 35.96 | **53.49** | 46.44 | 27.39 | 8.39 | 40.58 | 0.87 |
| | ins6 | 41.83 | 44.59 | 39.03 | 29.64 | **46.39** | 44.33 | 30.79 | 6.27 | 39.51 | 4.51e-03 |
| | ins7 | 32.75 | 30.87 | 33.62 | **35.53** | 28.12 | 24.28 | 34.52 | 3.69 | 31.38 | 0.09 |

Table 5.6: Mean Dice score (%) achieved at different (s_ins, q_ins) combinations by '3d_con_align' when Institution 3 was adopted as the novel institution, where s_ins and q_ins respectively refer to the institution from which the support and query were sampled. The best performance for each query institution was bolded. p-values were derived from paired t-tests performed between the Dice scores where the support institution equals the query institution and the maximum Dice scores achieved when the support institution differs from the query institution.

| | | s_ins | | | | | | | std | mean | p-value |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ins1 | ins2 | ins3 | ins4 | ins5 | ins6 | ins7 | | | |
| q_ins | ins1 | **53.76** | 53.31 | 50.63 | 46.46 | 50.77 | 47.78 | 47.83 | 2.62 | 50.08 | 1.06e-15 |
| | ins2 | 54.26 | **57.74** | 50.07 | 44.42 | 51.97 | 52.05 | 48.62 | 3.91 | 51.30 | 0.08 |
| | ins3 | 51.11 | 50.17 | **53.51** | 49.41 | 52.25 | 47.25 | 51.67 | 1.90 | 50.77 | 2.04e-16 |
| | ins4 | 48.23 | 44.70 | 49.95 | 50.73 | 47.09 | 44.02 | **51.47** | 2.70 | 48.03 | 1.40e-05 |
| | ins5 | 54.70 | 52.88 | 54.38 | 54.14 | **61.14** | 50.56 | 54.77 | 2.98 | 54.65 | 0.15 |
| | ins6 | 52.10 | **55.89** | 51.69 | 44.23 | 52.70 | 52.02 | 51.61 | 3.26 | 51.46 | 9.47e-03 |
| | ins7 | 46.46 | 45.81 | 48.77 | 48.78 | 45.45 | 41.57 | **49.85** | 2.60 | 46.67 | 0.20 |

often achieved by both methods when support images came from the query institution. A two-tailed t-test is performed per query institution between the Dice scores where the support institution equals the query institution and the maximum Dice scores achieved when the support institution differs

from the query institution. Such observation is consistent with the hypothesis that the domain shift is smaller between support-query pairs within the same institution, leading to better performance.

Table 5.7: Dice score (%) achieved when query and support both came from base institutions while institution 3 was adopted as the novel institution. 'support ins' here denote if the support and query come from the same institution, different institutions or all base institutions are counted.

| model | support ins | Dice(%) query institution | | | | | | | #param (million) |
|---|---|---|---|---|---|---|---|---|---|
| | | ins1 | ins2 | ins4 | ins5 | ins6 | ins7 | mean | |
| BiGRU | same | 41.67 | 45.03 | 34.16 | 38.25 | 36.63 | 27.43 | 37.20 | 38.85 |
| | diff | 35.40 | 34.00 | 31.60 | 34.09 | 36.77 | 25.76 | 32.94 | |
| | all | 36.18 | 35.38 | 31.92 | 34.61 | 36.75 | 25.97 | 33.47 | |
| | Δ | 15.06% | 24.49% | 7.49% | 10.88% | -0.37% | 6.09% | 11.45% | |
| 2d | same | 48.41 | 49.15 | 39.00 | 53.49 | 44.33 | 34.52 | 44.82 | 23.63 |
| | diff | 39.14 | 37.89 | 33.15 | 38.74 | 38.83 | 30.94 | 36.45 | |
| | all | 40.30 | 39.30 | 33.88 | 40.58 | 39.51 | 31.38 | 37.49 | |
| | Δ | 19.15% | 22.91% | 15.02% | 27.58% | 12.42% | 10.38% | 18.68% | |
| 3d_con_align | same | 53.76 | 57.74 | 50.73 | 61.14 | 52.02 | 49.85 | 54.21 | 27.35 |
| | diff | 49.55 | 50.39 | 47.64 | 53.73 | 51.38 | 46.22 | 49.82 | |
| | all | 50.08 | 51.30 | 48.03 | 54.65 | 51.46 | 46.67 | 50.37 | |
| | Δ | 7.83% | 12.73% | 6.08% | 12.12% | 1.23% | 7.29% | 8.10% | |
| localnet (unsupervised) | same | 36.42 | 48.05 | 27.66 | 39.53 | 26.51 | 19.94 | 33.02 | 5.94 |
| | diff | 27.50 | 30.88 | 21.77 | 20.28 | 29.00 | 20.88 | 25.05 | |
| | all | 28.62 | 33.03 | 22.50 | 22.69 | 28.69 | 20.76 | 26.05 | |
| | Δ | 24.49% | 35.73% | 21.30% | 48.69% | -9.37% | -4.68% | 24.13% | |
| localnet (supervised) | same | 73.05 | 64.65 | 41.41 | 73.38 | 68.87 | 40.44 | 60.30 | 5.94 |
| | diff | 62.48 | 58.76 | 50.72 | 61.92 | 60.79 | 51.28 | 57.66 | |
| | all | 63.80 | 59.50 | 49.56 | 63.35 | 61.80 | 49.92 | 57.99 | |
| | Δ | 14.47% | 9.11% | -22.49% | 15.62% | 11.74% | -26.81% | 4.38% | |
| 3d_supervised | N/A | 86.77 | 84.08 | 80.75 | 86.36 | 80.16 | 80.69 | 83.14 | 5.75 |

Table 5.7 reports the performance of the proposed model ('3d_con_align') and baseline methods on novel classes when both query and support come from base institutions while institution 3 is used as the novel institution. We report results when support comes from the same institution as the query, a different institution from the query, and overall base institutions. The proposed method outperformed all baseline methods in few-shot settings, including two state-of-the-art methods - BiGRU [371] by 17.01%, 16.88% and

16.9% and LSNet [258] (which is adopted as our '2d' baseline) by 9.39%, 13.37% and 12.88% absolute Dice, when support comes from the same institution as query, different institution from query and all base institutions respectively, proving its efficacy even when no novel institution is involved during evaluation. Moreover, better performance was achieved when query and support came from the same institution than different institutions, with the performance gap reported as $\Delta$. The proposed method achieved a smaller mean $\Delta$ compared to baseline few-shot methods, suggesting its ability to mitigate domain shifts from cross-institution query and support data.

Table 5.8: Dice score (%) and 95% Hausdorff distance achieved by '3d_con_align' with various training data availability when Institution 3 was adopted as the novel institution. 'support ins' refers to the institution from which the support was sampled. $\Delta$ refers to the percentage difference when the support image comes from the base and novel institutions.

| training data | s_ins | Dice (%) | | | | | 95% Hausdorff distance (mm) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | fold1 | fold2 | fold3 | fold4 | mean | fold1 | fold2 | fold3 | fold4 | mean |
| whole | all | 59.36 | 60.38 | 45.73 | 37.60 | 50.77 | 22.66 | 29.63 | 27.92 | 26.41 | 26.66 |
| | base | 59.11 | 59.80 | 45.06 | 37.28 | 50.31 | 22.69 | 30.30 | 28.38 | 26.89 | 27.06 |
| | novel | 60.86 | 63.87 | 49.77 | 39.54 | 53.51 | 22.49 | 25.63 | 25.19 | 23.54 | 24.21 |
| | $\Delta$ | 2.89% | 6.37% | 9.47% | 5.70% | 5.98% | -0.91% | -18.21% | -12.65% | -14.26% | -11.79% |
| half | all | 55.38 | 57.16 | 42.54 | 34.14 | 47.30 | 25.98 | 29.37 | 31.65 | 29.00 | 29.00 |
| | base | 55.08 | 55.49 | 41.82 | 33.75 | 46.54 | 26.04 | 30.13 | 32.17 | 29.37 | 29.43 |
| | novel | 57.16 | 67.17 | 46.88 | 36.46 | 51.91 | 25.56 | 24.81 | 28.52 | 26.84 | 26.43 |
| | $\Delta$ | 3.63% | 17.38% | 10.79% | 7.41% | 10.36% | -1.91% | -21.45% | -12.82% | -9.43% | -11.35% |
| half_single_ins | all | 52.85 | 57.46 | 31.31 | 20.39 | 40.50 | 26.84 | 29.38 | 35.41 | 33.89 | 31.38 |
| | base | 52.17 | 56.06 | 30.55 | 19.63 | 39.60 | 26.93 | 29.93 | 36.04 | 34.41 | 31.83 |
| | novel | 56.90 | 65.87 | 35.85 | 24.93 | 45.89 | 26.29 | 26.10 | 31.65 | 30.75 | 28.70 |
| | $\Delta$ | 8.32% | 14.89% | 14.79% | 21.24% | 13.69% | -2.44% | -14.63% | -13.88% | -11.89% | -10.90% |
| quarter | all | 50.90 | 55.87 | 44.78 | 31.98 | 45.88 | 27.90 | 29.61 | 28.87 | 29.48 | 28.96 |
| | base | 50.35 | 54.54 | 43.64 | 31.61 | 45.03 | 27.90 | 30.22 | 29.43 | 29.84 | 29.35 |
| | novel | 54.24 | 63.86 | 51.61 | 34.17 | 50.97 | 27.88 | 25.92 | 25.50 | 27.27 | 26.64 |
| | $\Delta$ | 7.17% | 14.60% | 15.45% | 7.49% | 11.65% | -0.05% | -16.60% | -15.41% | -9.43% | -10.15% |

Table 5.8 reports the performance achieved by the proposed method ('3d_con_align') as the training set varies. The performance dropped as the number of images inside the training set reduced. Notably, the model trained on 'half_single_ins' performed worse than the model trained on 'quarter', which had only half the size of 'half_single_ins'. This suggests that the cross-institution few-shot task can be sensitive to the number of institutions available in the training

set. To quantify this sensitivity is an interesting future research question.

Table 5.9: Dice score achieved by '3d_con_align' with various number of shots when Institution 3 was adopted as the novel institution. The 'support ins' refers to the institution from which the support images were sampled.

| shot | # of shot | support ins | Dice (%) | | | | | 95% Hausdorff distance (mm) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | fold1 | fold2 | fold3 | fold4 | mean | fold1 | fold2 | fold3 | fold4 | mean |
| 1 | 3d_finetune | all | 48.07 | 44.78 | 43.92 | 29.25 | 41.50 | 61.47 | 69.78 | 60.33 | 57.02 | 62.15 |
| | | base | 45.74 | 39.81 | 41.69 | 26.02 | 38.32 | 63.22 | 73.42 | 62.33 | 59.24 | 64.55 |
| | | novel | 62.03 | 74.56 | 57.31 | 48.63 | 60.63 | 50.98 | 47.96 | 48.32 | 43.72 | 47.74 |
| | | Δ | 26.26% | 46.60% | 27.25% | 46.48% | 36.80% | -24.02% | -53.09% | -28.99% | -35.51% | -35.21% |
| | 3d_con_align | all | 59.36 | 60.38 | 45.73 | 37.60 | 50.77 | 22.66 | 29.63 | 27.92 | 26.41 | 26.66 |
| | | base | 59.11 | 59.80 | 45.06 | 37.28 | 50.31 | 22.69 | 30.30 | 28.38 | 26.89 | 27.06 |
| | | novel | 60.86 | 63.87 | 49.77 | 39.54 | 53.51 | 22.49 | 25.63 | 25.19 | 23.54 | 24.21 |
| | | Δ | 2.89% | 6.37% | 9.47% | 5.70% | 5.98% | -0.91% | -18.21% | -12.65% | -14.26% | -11.79% |
| | localnet (supervised) | all | 72.70 | 72.14 | 52.84 | 42.65 | 60.08 | 19.36 | 22.81 | 21.78 | 24.00 | 21.99 |
| | | base | 71.80 | 70.11 | 51.21 | 41.65 | 58.69 | 20.29 | 24.69 | 23.07 | 25.11 | 23.29 |
| | | novel | 78.11 | 84.33 | 62.61 | 48.61 | 68.42 | 13.77 | 11.57 | 14.03 | 17.29 | 14.17 |
| | | Δ | 8.09% | 16.86% | 18.20% | 14.31% | 14.21% | -47.33% | -113.31% | -64.37% | -45.26% | -64.39% |
| 2 | 3d_finetune | all | 52.58 | 57.19 | 45.20 | 35.41 | 47.59 | 52.90 | 55.99 | 44.63 | 49.53 | 50.77 |
| | | base | 49.94 | 52.74 | 43.50 | 32.35 | 44.63 | 55.69 | 60.55 | 47.35 | 53.64 | 54.31 |
| | | novel | 68.42 | 83.86 | 55.43 | 53.81 | 65.38 | 36.18 | 28.60 | 28.30 | 24.88 | 29.49 |
| | | Δ | 27.01% | 37.11% | 21.53% | 39.89% | 31.74% | -53.91% | -111.71% | -67.33% | -115.65% | -84.16% |
| | 3d_con_align | all | 60.64 | 60.41 | 47.82 | 38.91 | 51.95 | 21.63 | 29.19 | 27.08 | 25.66 | 25.89 |
| | | base | 60.03 | 59.82 | 47.02 | 38.54 | 51.35 | 21.80 | 29.89 | 27.55 | 26.07 | 26.33 |
| | | novel | 64.28 | 63.94 | 52.66 | 41.12 | 55.50 | 20.59 | 24.99 | 24.27 | 23.22 | 23.27 |
| | | Δ | 6.61% | 6.44% | 10.71% | 6.27% | 7.47% | -5.88% | -19.59% | -13.52% | -12.26% | -13.15% |
| | localnet (supervised) | all | 71.93 | 71.47 | 50.99 | 43.23 | 59.40 | 10.93 | 21.92 | 15.91 | 16.36 | 16.28 |
| | | base | 70.64 | 68.98 | 48.68 | 41.90 | 57.55 | 11.42 | 24.53 | 16.81 | 17.05 | 17.45 |
| | | novel | 79.66 | 86.41 | 64.84 | 51.21 | 70.53 | 7.99 | 6.25 | 10.48 | 12.23 | 9.24 |
| | | Δ | 11.32% | 20.18% | 24.92% | 18.19% | 18.41% | -42.99% | -292.69% | -60.36% | -39.35% | -88.93% |
| 3 | 3d_finetune | all | 53.92 | 60.16 | 47.94 | 36.24 | 49.57 | 52.14 | 52.74 | 40.64 | 49.66 | 48.80 |
| | | base | 51.61 | 56.12 | 46.51 | 33.30 | 46.88 | 54.27 | 57.69 | 43.23 | 52.82 | 52.00 |
| | | novel | 67.80 | 84.38 | 56.56 | 53.90 | 65.66 | 39.34 | 23.04 | 25.12 | 30.69 | 29.54 |
| | | Δ | 23.88% | 33.50% | 17.77% | 38.23% | 28.60% | -37.96% | -150.45% | -72.11% | -72.12% | -76.02% |
| | 3d_con_align | all | 61.36 | 60.93 | 48.53 | 38.55 | 52.34 | 21.59 | 28.81 | 26.66 | 25.66 | 25.68 |
| | | base | 61.06 | 60.27 | 47.99 | 38.25 | 51.89 | 21.64 | 29.42 | 26.87 | 26.06 | 26.00 |
| | | novel | 63.13 | 64.88 | 51.80 | 40.33 | 55.03 | 21.33 | 25.15 | 25.42 | 23.28 | 23.80 |
| | | Δ | 3.28% | 7.10% | 7.35% | 5.16% | 5.71% | -1.43% | -16.96% | -5.68% | -11.93% | -9.24% |
| | localnet (supervised) | all | 78.07 | 74.86 | 56.98 | 47.94 | 64.46 | 12.73 | 16.10 | 16.97 | 19.12 | 16.23 |
| | | base | 77.57 | 73.08 | 55.73 | 47.09 | 63.37 | 13.50 | 17.41 | 17.89 | 19.98 | 17.19 |
| | | novel | 81.09 | 85.53 | 64.48 | 53.06 | 71.04 | 8.11 | 8.23 | 11.46 | 13.98 | 10.45 |
| | | Δ | 4.35% | 14.55% | 13.57% | 11.24% | 10.80% | -66.36% | -111.55% | -56.08% | -42.92% | -64.60% |
| 4 | 3d_finetune | all | 55.53 | 61.34 | 48.17 | 36.85 | 50.47 | 50.70 | 49.19 | 39.65 | 47.11 | 46.66 |
| | | base | 52.93 | 56.80 | 46.32 | 33.40 | 47.36 | 53.41 | 54.70 | 43.12 | 51.48 | 50.68 |
| | | novel | 68.50 | 84.07 | 57.37 | 54.06 | 66.00 | 37.17 | 21.62 | 22.29 | 25.30 | 26.60 |
| | | Δ | 22.73% | 32.44% | 19.25% | 38.21% | 28.24% | -43.66% | -153.05% | -93.42% | -103.48% | -90.54% |
| | 3d_con_align | all | 61.33 | 60.63 | 49.52 | 38.16 | 52.41 | 21.46 | 29.59 | 26.14 | 26.32 | 25.88 |
| | | base | 60.69 | 59.91 | 49.01 | 37.56 | 51.79 | 21.64 | 30.34 | 26.64 | 27.06 | 26.42 |
| | | novel | 64.52 | 64.23 | 52.08 | 41.19 | 55.50 | 20.57 | 25.80 | 23.64 | 22.63 | 23.16 |
| | | Δ | 5.94% | 6.73% | 5.90% | 8.81% | 6.69% | -5.19% | -17.63% | -12.69% | -19.57% | -14.08% |
| | localnet (supervised) | all | 77.64 | 81.89 | 60.06 | 51.96 | 67.89 | 8.95 | 9.62 | 12.77 | 13.38 | 11.18 |
| | | base | 76.73 | 81.04 | 58.84 | 51.44 | 67.01 | 9.50 | 10.28 | 13.32 | 13.75 | 11.71 |
| | | novel | 82.14 | 86.15 | 66.18 | 54.57 | 72.26 | 6.23 | 6.33 | 10.02 | 11.54 | 8.53 |
| | | Δ | 6.58% | 5.94% | 11.09% | 5.73% | 7.26% | -52.50% | -62.41% | -32.94% | -19.22% | -37.34% |

Table 5.9 reports the performance of the proposed method '3d_con_align'

and '3d_finetune' baseline, using 1 to 4 support examples (denoted as # shot). Performance improved for both methods, as the number of support examples increased. While '3d_finetune' is more sensitive to the number of support examples, the proposed method still outperformed '3d_finetune' using 1 to 4 support examples.

## 5.7 Discussion

While both the spatial registration mechanism and the support mask conditioning were motivated by the observation on multiple structure types found in the multi-institution data set used in this study, they may be promising to be beneficial for wider image types and anatomical regions as similar challenges were found in the few-shot segmentation of other types of non-medical data.

The labels used in this study were annotated by a mixture of clinicians and experienced medical imaging researchers. The estimated time for completing this task was more than one thousand observer-hours, a practically challenging task for most local hospitals if an alternative supervised learning was adopted for adopting or validating a segmentation tool. This further justifies the clinical relevance for the proposed few-shot segmentation approach.

It is noteworthy that the reported segmentation performance results were based on as few as 1 - 4 labelled training examples of the regions of interest, which had not been labelled in the model training stage.

Though challenging, cross-institutional few-shot segmentation could benefit situations when limited number of annotated data are available. Potential applications, although not investigated in this study due to relevant data

availability, include specific pathology detection and segmentation with rare instances without previously observed occurrence in training institutions and longitudinal analysis with available within-subject data from individual patients.

The proposed spatial alignment mechanism is guided by base class segmentation predicted by the segmentation head, necessitating annotations of structures from the same field-of-view. Future research could explore the potential of utilising a supervoxel algorithm for guidance to alleviate the annotation requirement.

While this study validated and mitigated the adverse effects of spatial misalignment between support and query images acquired from different institutions, there are many other cross-institution challenges yet to be resolved. Analysing the independent impact and exploring solutions for various perspectives of variations across institutions on few-shot segmentation performance would be valuable to conduct if data is available.

## 5.8 Conclusion

This work described the first 3D prototypical learning algorithm for medical image segmentation, applied on multiple structures on pelvic MR images from different institutes. Substantial validation was based on clinical data from 589 patients, with full segmentation of eight anatomical classes made available to the scientific community. The demonstrated novelty, efficacy, and clinical applicability of the proposed algorithm suggested an interesting direction for addressing the cost of expert labelling and cross-institute generalisation of current deep learning-based segmentation applications.

# Chapter 6

# Semi-weakly-supervised neural networks for medical image registration

This chapter focuses on another clinically valuable task - registration.

For training registration networks, weak supervision from segmented corresponding regions-of-interest (ROIs) have been proven effective for *a)* supplementing unsupervised methods, and *b)* being used independently in registration tasks in which unsupervised losses are unavailable or ineffective. This correspondence-informing supervision entails cost in annotation that requires significant specialised effort. This chapter describes a semi-weakly-supervised registration pipeline that improves the model performance, when only a small corresponding-ROI-labelled dataset is available, by exploiting unlabelled image pairs. We examine two types of augmentation methods by perturbation on network weights and image resampling, such that consistency-based unsupervised losses can be applied on unlabelled data. The novel WarpDDF

and RegCut approaches are proposed to allow commutative perturbation between an image pair and the predicted spatial transformation (i.e. respective input and output of registration networks), distinct from existing perturbation methods for classification or segmentation. Experiments using 589 male pelvic MR images, labelled with eight anatomical ROIs, show the improvement in registration performance and the ablated contributions from the individual strategies. Furthermore, this study attempts to construct one of the first computational atlases for pelvic structures, enabled by registering inter-subject MRs, and quantifies the significant differences due to the proposed semi-weak supervision with a discussion on the potential clinical use of example atlas-derived statistics.

## 6.1 Introduction

### 6.1.1 Learning-based medical image registration

Pairwise image registration is the process of establishing a plausible 2D or 3D spatial correspondence between two images, often termed moving and fixed images. The estimated spatial-correspondence-representing transformation may be used to align the moving image to the fixed image.

Such correspondence can be useful in many clinical applications. Inter-subject image registration is the basis of the atlas-based segmentation algorithms and population studies that are important for epidemiology research [144]. Intra-subject longitudinal image registration can track suspicious pathological regions for individuals by aligning images acquired from different time points [145]. Registering images from different modalities can combine complementary diagnostic information from different imaging

sources to make informed clinical decisions or guide surgery and interventions [146]. Stemmed from computer vision algorithms such as optical flow and pose estimation, proving its usefulness in medical image analysis, registration algorithms have been developed and adopted by many multidisciplinary researches.

The recent development in deep-learning has promoted the investigation towards *learning-based registration* methods [373, 374, 375]. Fully supervised training of a registration network requires a number of pairs of moving and fixed images, with corresponding ground-truth transformation. Ground-truth transformations, especially deformable transformations, e.g. with higher degrees of freedom than those in rigid or affine transformations, are scarce and prohibitively infeasible to be reliably obtained from many clinical applications. Therefore many methods adopt unsupervised approaches, which rely on the availability of effective and robust image similarity measures. Improved performance could be achieved by introducing weak supervision from auxiliary information - manual tracing of hippocampus head and body for intra-patient hippocampus registration [72], lower pelvic landmarks for intra-patient and multimodel registration [73, 74, 72] and left and right ventricle cavity and myocardium for intra-patient cardiac registration [194]. In cases, these segmentations of corresponding structures may be sufficient or even beneficial without image similarity being used as part of the loss functions - the cardinal form of weakly-supervised registration [73, 74].

### 6.1.2 Semi-weakly-supervised registration

First, it is important to clarify "weakly-supervised" and "semi-supervised" approaches for the purpose of this study, where the supervision comes from the available segmentation of corresponding regions of interest (ROIs), between

moving and fixed images. As discussed in the original papers [73, 74, 194, 72], these segmented ROIs represent a different type and "weaker" labels of correspondence at the ROI level, whilst the objective of registration network training is to learn (and therefore to predict at inference) dense correspondence at voxel level. Semi-supervised learning has also been used to refer to the same class of registration network training methodologies [1], perhaps emphasising the required voxel-level supervision are only partially available at a subset of all voxel locations (i.e. segmented ROIs). This is particularly judicious when the segmented ROIs represent point landmarks. However, to avoid terminological confusion, we hereinafter refer this type of segmentation-driven methods to as weakly-supervised training, whilst the "semi-weakly-supervised" (or semi-supervised in the context of this work) is used to indicate the partially available segmentation labels, that is a proper subset of training images (or subjects) with labelled corresponding ROIs.

Second, it is also useful to point out that many clinical applications lack a valid and robust image intensity-based similarity measure that can be used for the unsupervised loss between two unlabelled images [376]. It is this challenge in many multi-modal registration problems that the weakly-supervised approaches [73, 74, 72] intend to address. In these applications, general training approaches (beyond the existing intensity-similarity-based unsupervised losses) for utilising images without segmented corresponding ROIs are desirable. Although considered outside of the scope of this work, the proposed semi-weakly-supervised approaches, which utilise the "unlabelled image pairs", could also be useful in combination with these intensity-similarity-based unsupervised losses, which utilise "unlabelled voxels".

In general, these segmentation labels for assisting registration require expert knowledge of radiological anatomy and pathology as well as a significant

amount of time. It is therefore valuable to develop methods that require less annotation effort with minimum sacrifice of performance.

This work investigates semi-weakly-supervised registration approaches, which reduce cost in labelling by training on the combination of a small labelled dataset and a large unlabelled dataset, while minimising the registration performance loss.

It is interesting to postulate that, at the time of writing, recent foundation models [377] have yet to be applied in segmentation-supervised registration tasks may be due to the lack of unsupervised (or self-supervised) training methods for such registration. Utilising a large amount of images without labels has been proven effective or even necessary in foundation model development [378, 379, 380], and may resort to methodologies developed in this work for constructing and/or adapting foundation models for registration tasks.

### 6.1.3 Commutative property for transformation prediction

Semi-supervised learning based on consistency regularisation has proven its efficacy on image segmentation - another task requires voxel-level prediction. Such consistency is enforced between predictions with and without various perturbations applied to input images, network weights and architectures. Image perturbation applies augmentations such as affine transformation, CutOut, CutMix and ClassMix to the input image, and penalises the difference between predictions when the augmentation is applied before and after the inference [381]. Specifically, given a randomly sampled augmentation $A$, an image $I$ could be augmented to $\tilde{I} = A(I)$. The model $g$ is

trained to minimise the distance between the prediction from the augmented image $M^{\text{pre}} = g(\tilde{I}) = g(A(I))$ and the prediction from the original image after applying the same augmentation $M^{\text{post}} = A(g(I))$.

In other words, image perturbation for semi-supervised segmentation assumes an intuitive functional commutative property, between the image transformation (for augmentation) $A$ and segmentation prediction $g$, and minimises the difference between $g(A(I))$ and $A(g(I))$.

However, learning-based registration networks have different types of model input and output, i.e. predicting the corresponding transformation parameters $U = g(I^{\text{mov}}, I^{\text{fix}})$, given a pair of moving and fixed images $(I^{\text{mov}}, I^{\text{fix}})$, and they are not in the same domain. Therefore, the commutative relationship, between the transformation prediction $g$ and existing augmentation methods $A$ used in segmentation (or other task types), does not apply directly.

This motivates our development of a class of commutative perturbation methods, such that the perturbed model output (predicted transformation) is expectedly equivalent to the model output based on a perturbed input (an image pair to be registered), for consistent losses to be calculated between the two transformations. This work proposes two cost-efficient image augmentation techniques and the corresponding consistency rule for semi-weakly-supervised registration. Specifically, for each augmentation $A$ applied to the image pairs $(I^{\text{mov}}, I^{\text{fix}})$ to generated the augmented image pairs $(\tilde{I}^{\text{mov}}, \tilde{I}^{\text{fix}}) = A(I^{\text{mov}}, I^{\text{fix}})$, a corresponding augmentation $\tilde{A}$ in the transformation domain is provided such that consistency could be enforced by penalising the difference between $U^{pre} = g(A(I^{\text{mov}}, I^{\text{fix}}))$ and $U^{post} = \tilde{A}(g(I^{\text{mov}}, I^{\text{fix}}))$.

### 6.1.4 Summary of contributions

In this work, we report experimental results using an example application of inter-subject registration between male lower pelvic MR images from prostate cancer patients. The clinical significance of this application is discussed with one of the first atlas construction studies in this area (Sec. 6.5.3).

- We first adopt semi-supervision approaches and established a baseline for comparison, in an inter-subject lower-pelvic MR image registration task.
- We propose two novel commutative image perturbation methods and corresponding consistency rules designed specifically for image registration.
- We carry out experiments on inter-patient multi-structural lower-pelvic MR registration which proved the efficacy of the proposed method.
- We present extensive ablation studies on the effect of image and weight perturbation, with an attempt to demonstrate the value of semi-weak supervision in building a first-of-its-kind lower-pelvic MRI atlas.

## 6.2 Literature Review

### 6.2.1 Medical image registration

Early algorithms [153, 154, 159, 155, 156, 152, 157, 158, 147] formulate registration as an iterative optimisation to maximise similarity between fixed image and transformed moving image.

Nowadays, most of state-of-the-art methods adopt deep neural networks

trained large training sets and allow single-step inference. The training process entails minimising the registration loss, which measures the dissimilarity between the predicted and ground-truth transformation or between the fixed and transformed moving image. Additionally, regularisation terms are incorporated to impose constraints on the estimated displacement field, promoting qualities like smoothness or continuity. These regularisation terms play a crucial role in preventing unrealistic transformations during the registration process. Due to the scarcity of ground-truth transformation, especially for deformations, most deep-learning approaches adopted unsupervised or weakly-supervised training which maximises the similarity between fixed image and moving image warped by the predicted transformation. Compare with unsupervised methods that solely depend on image similarity metrics [186, 191, 184, 189, 382], weakly-supervised approaches guide registration with auxiliary information like segmentation maps and landmark achieved better and more robust performance. Hu et al. [73, 74] improved multimodal registration performance with weak-supervision from full gland segmentation, lower-pelvic landmarks and patient-specific point landmarks. Hering et al. [194] utilised annotations for left and right ventricle cavity and myocardium of a clinical expert for better intra-patient cardiac registration. Zhu et al. [72] validated the efficacy of guidance from manual tracing of hippocampus and prostate segmentation respectively for intra-patient hippocampus and prostate registration.

While weakly supervised transformation estimation methods mitigate issues related to the reliability and cost of ground truth labels, they still necessitate a considerable amount of manually annotated data. Research on weakly-supervised registration using a minimal number of annotated examples remains an unresolved challenge.

### 6.2.2 Semi-supervised learning

Semi-supervised learning improves model performance without increasing labelling expanse by incorporating a large amount of unlabelled data during training. It is now commonly adopted in image segmentation - another task requires pixel-level prediction.

Early methods utilised GANs either to introduce auxiliary tasks to discriminate real and unreal images generated by GAN [383], or to train the model to fool a discriminator which differentiates predictions and ground truth [384]. State-of-the-art methods mostly adopt consistency regularisation, which enforces consistency across predictions with various perturbations. Weight perturbation methods [385, 386, 291, 387] enforced consistency between predictions from multiple networks with the same architecture but different initialisation. Image perturbation methods [282, 388, 389, 387] augmented input images with techniques and enforced consistency of predictions between augmented images. Commonly adopted spatial augmentation techniques include classical transformations such as rotation, scaling and translation, as well as more strongly augmentation methods such as CutOut, CutMix, ClassMix and so on.

These augmentations can be directly applied to both the image and its predicted segmentation to create matched pairs for segmentation tasks. However, these approaches are limited to tasks like semi-supervised tasks where both the input and output are within the image domain. In contrast, for the registration task, where the input consists of a pair of moving and fixed images from the image domain and the output is a predicted transformation in the transformation domain, this direct consistency is not preserved. This work proposed DDF-based transformation which could be directly applied

to the registration output and RegCut, a strong augmentation designed for registration.

While registration has been utilised to generate pseudo-labels for unlabelled images in some semi-supervised segmentation algorithms [390, 391, 392], from our knowledge, no prior work adopted semi-supervision directly on the segmentation-supervised registration tasks.

## 6.3 Task Definition

Given a pair of moving $I^{\text{mov}} \in \mathbb{R}^{W \times H \times D}$ and fixed $I^{\text{fix}} \in \mathbb{R}^{W \times H \times D}$ images (3D MRI), pairwise registration spatially aligns the moving image towards the fixed image. $W$, $H$ and $D$ indicating the width, height and depth of the 3D volumetric images [1], respectively. The components of the images [2], $i^{\text{mov}}_{xyz}$ and $i^{\text{fix}}_{xyz}$, are indexed by their voxel coordinates $x = 1, ..., W$, $y = 1, ..., H$ and $z = 1, ..., D$, representing image intensity at individual voxels, respectively. Registration networks predict a dense displacement field (DDF), $U \in \mathbb{R}^{3 \times W \times H \times D}$ with its components $u_{dxyz}$, where $d \in \{1, 2, 3\}$, and $u_{1xyz}$, $u_{2xyz}$ and $u_{3xyz}$ indicating the $x$-, $y$- and $z$-displacement at voxel location $(x, y, z)$. A registration network $g(\cdot; \theta)$ with parameters $\theta$, i.e. network weights, aims to predict a spatial transformation, represented by DDF, from the given image pair:

$$U = g(I^{\text{mov}}, I^{\text{fix}}; \theta) \tag{6.1}$$

For weakly-supervised training, a labelled training set $\mathcal{D}_{(\text{lab})}$ is available,

[1] The same size is assumed for both images for notional brevity, without losing its generality in the proposed methods or discussion in this study.

[2] In this work, uppercase $I^{\text{mov}}$ and $I^{\text{fix}}$ indicate tensor-valued random variables, whilst lowercase $i^{\text{mov}}_{xyz}$ and $i^{\text{fix}}_{xyz}$ are scalar-valued tensor components with the subscripts $x$, $y$ and $z$.

from which moving-fixed image pairs, $I^{\text{mov}}_{(\text{lab})} \in \mathbb{R}^{W\times H\times D}$ and $I^{\text{fix}}_{(\text{lab})}$, together with a set of $C$-class segmentation masks, $M^{\text{mov}}_{(c)}$ and $M^{\text{fix}}_{(c)}$ of the same image size $\mathbb{R}^{W\times H\times D}$, can be sampled, $(I^{\text{mov}}_{(\text{lab})}, I^{\text{fix}}_{(\text{lab})}, \{M^{\text{mov}}_{(c)}\}, \{M^{\text{fix}}_{(c)}\}) \sim \mathcal{D}_{(\text{lab})}$, where $C$ indicates the number of classes (here, types of anatomical ROIs) thus $c = 1, ..., C$. The components of $M^{\text{mov}}_{(c)}$ and $M^{\text{fix}}_{(c)}$ are indexed as $m^{\text{mov}}_{(c)xyz}$ and $m^{\text{fix}}_{(c)xyz}$, respectively. Both are assigned with a value of 1, if voxel $(x, y, z)$ belongs to the ROI class $c$, 0 otherwise, i.e. $[m^{\text{mov}}_{(1)xyz}, ..., m^{\text{mov}}_{(C)xyz}]^\top$ be a voxel-wise one-hot vector.

Semi-weakly-supervised training has access to an additional unlabelled training set $\mathcal{D}_{(\text{unl})}$, such that the samples of unlabelled moving-fixed image pairs, of the same image size, can be drawn, $(I^{\text{mov}}_{(\text{unl})}, I^{\text{fix}}_{(\text{unl})}) \sim \mathcal{D}_{(\text{unl})}$ without segmentation masks. The following sections describe our proposed methods to utilise both $\mathcal{D}_{(\text{lab})}$ and $\mathcal{D}_{(\text{unl})}$ for training the registration network in Eq. 6.1.

For a matrix $Z \in \mathbb{R}^{ch\times W\times H\times D}$, define $l(Z, x, y, z)$ as a function interpolating a vector $\in \mathbb{R}^{ch}$ at coordinates $(x, y, z)$ at the last three dimensions, $T$ refers to the resampling spatial transformation such that:

$$l(T(X, U), x, y, z) = l(X, x', y', z') \tag{6.2}$$

where

$$x' = x + U_{\mu xyz}$$
$$y' = y + U_{\nu xyz}$$
$$z' = z + U_{\xi xyz}$$

and $U_{\mu xyz}$, $U_{\nu xyz}$ and $U_{\xi xyz}$ respectively denotes the displacement along x-, y- and z-direction at location $(x, y, z)$

## 6.4 Method

### 6.4.1 Weak supervision loss on labelled data

A registration network $U_{(\text{lab})} = g(I^{\text{mov}}_{(\text{lab})}, I^{\text{fix}}_{(\text{lab})}; \theta)$ can be trained with the labelled set $D_{(\text{lab})}$, by minimising a loss function $L_{\text{weak-sup}}$ between the $U_{(\text{lab})}$-warped moving segmentation masks $M^{\text{warp}}_{(c)}$ and the fixed masks $M^{\text{fix}}_{(c))}$, where

$$M^{\text{warp}}_{(c)} = M^{\text{mov}}_{(c)} \circ U_{(\text{lab})} \tag{6.3}$$

where, the image resampling operation is denoted by "$\circ$", resampling the left image-sized mask (or image) $M^{\text{mov}}_{(c)}$ using the right DDF $U_{(\text{lab})}$, such that the indexed values $m^{\text{warp}}_{(c)xyz}$ at its $(x, y, z)$ resampled voxel locations are spatially-interpolated from the $(x+u_{1xyz}, y+u_{2xyz}, z+u_{3xyz})$ locations from the original image $m^{\text{mov}}_{(c)xyz}$ coordinates. In practice, the inverse fixed-to-moving DDF is implemented, for resampling at regular fixed image grid locations avoiding numerical inverting. Substituting the predicted DDF $U_{(\text{lab})}$, we have:

$$L_{\text{weak-sup}} = \mathbb{E}_c[L_{\text{Dice}}(M^{\text{mov}}_{(c)} \circ g(I^{\text{mov}}_{(\text{lab})}, I^{\text{fix}}_{(\text{lab})}; \theta), M^{\text{fix}}_{(c)})] \tag{6.4}$$

where $\mathbb{E}_c(\cdot)$ denotes statistical expectation over $C$ classes. In this work, the Dice loss is used:

$$\begin{aligned} &L_{\text{Dice}}(M^{\text{warp}}_{(c)}, M^{\text{mov}}_{(c)}) \\ &\quad = -\frac{2 \cdot \sum_{x,y,z} m^{\text{warp}}_{(c)xyz} m^{\text{mov}}_{(c)xyz}}{\sum_{x,y,z} m^{\text{warp}}_{(c)xyz} + \sum_{x,y,z} m^{\text{mov}}_{(c)xyz}} \end{aligned} \tag{6.5}$$

In the experiments described in Sec. 6.5, our weak supervision loss is tested independently without using unlabelled training set as a baseline. It is also

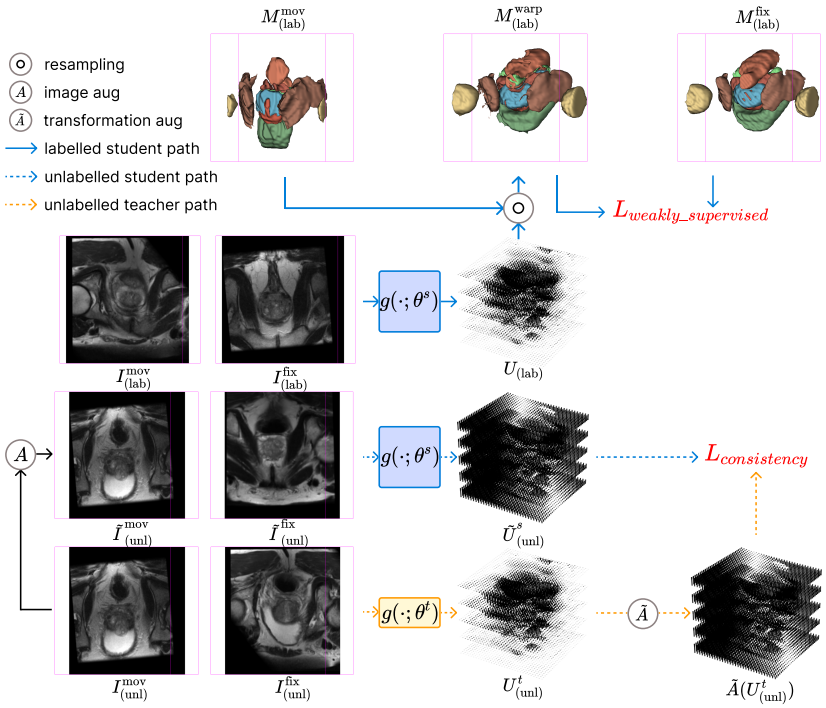



Figure 6.1: The proposed semi-weakly-supervised registration pipeline with $L_{weakly_supervised}$ calculated from the labelled pair and $L_{\text{cons}}$ calculated from the unlabelled pair.

used in conjunction of the consistency losses introduced in the following Secs. 6.4.2 and 6.4.3 for the proposed semi-weakly supervision.

### 6.4.2 Consistency following weight perturbation

For the unlabelled training set $\mathcal{D}_{(\text{unl})}$, network weight perturbation is introduced by adopting the student-teacher paradigm, which includes a student model and a teacher model with the same network architecture but separate weights, as illustrated in Figure 6.1. During training, the weights of teacher are a temporal ensemble from the weights of student. Denote the student and teacher weights as $\theta$ and $\theta_t$, respectively. The teacher weights are updated through an exponential moving average (EMA) scheme based on the current

student weights:

$$\theta^t = \gamma \times \theta^t + (1 - \gamma) \times \theta \tag{6.6}$$

where $\gamma \in (0, 1)$ is a hyperparameter controlling the transfer weight between epochs.

Given an pair of unlabelled image $(I^{\text{mov}}_{(\text{unl})}, I^{\text{fix}}_{(\text{unl})})$, the student and teacher model respectively predicts $U^s_{(\text{unl})} = g(I^{\text{mov}}_{(\text{unl})}, I^{\text{fix}}_{(\text{unl})}; \theta)$ and $U^t_{(\text{unl})} = g(I^{\text{mov}}_{(\text{unl})}, I^{\text{fix}}_{(\text{unl})}; \theta^t)$. Weight perturbation enforces consistency between the two DDFs by a consistency loss function $L_{\text{cons}}$:

$$L_{\text{cons}} = f_{\text{mse}}(U^t_{(\text{unl})}, U^s_{(\text{unl})}) \tag{6.7}$$

where $f_{\text{mse}}$ is a mean-square-error loss defined using the same tensor component notation as in Sec. 6.3:

$$f_{\text{mse}}(U^t_{(\text{unl})}, U^s_{(\text{unl})}) = \mathbb{E}_{d,x,y,z}[(u^t_{(\text{unl})dxyz} - u^s_{(\text{unl})dxyz})^2] \tag{6.8}$$

The student is trained through backpropagation to minimise the weak supervision loss (Eq. 6.5) calculated on the labelled pairs and the consistency loss (Eq. 6.7) calculated on the unlabelled pairs, sampled from $\mathcal{D}_{(\text{lab})}$ and $\mathcal{D}_{(\text{unl})}$, respectively:

$$L = L_{\text{weak-sup}} + \alpha L_{\text{cons}} \tag{6.9}$$

where $\alpha$ is the consistency loss coefficient. This loss function is considered task-agnostic, adapted from previous semi-supervised classification or segmentation (Sec. 6.2.2).

### 6.4.3 Image perturbation for registration

Image perturbation generates an augmented image pair $(\tilde{I}^{\text{mov}}_{(\text{unl})}, \tilde{I}^{\text{fix}}_{(\text{unl})}) = A(I^{\text{mov}}_{(\text{unl})}, I^{\text{fix}}_{(\text{unl})})$ from the pair of unlabelled image $(I^{\text{mov}}_{(\text{unl})}, I^{\text{fix}}_{(\text{unl})})$. During the training, teacher model predicts a DDF for the unaugmented pair $U^t_{(\text{unl})} = g(I^{\text{mov}}_{(\text{unl})}, I^{\text{fix}}_{(\text{unl})}; \theta^t)$ while student model predicts the DDF for the augmented pair $\tilde{U}^s_{(\text{unl})} = g(\tilde{I}^{\text{mov}}_{(\text{unl})}, \tilde{I}^{\text{fix}}_{(\text{unl})}; \theta)$. Whilst $A$ is an augmentation function on the input image pairs, $\tilde{A}$ is an augmentation function applied on the output DDFs $\tilde{A}(U^t_{(\text{unl})})$, such that the two DDFs $\tilde{U}^s_{(\text{unl})}$ and $\tilde{A}(U^t_{(\text{unl})})$ can be compared and the consistency loss may be enforced between the two:

$$L_{\text{cons}} = f_{\text{mse}}(\tilde{A}(U^t_{(\text{unl})}), \tilde{U}^s_{(\text{unl})}) \tag{6.10}$$

In the following subsections, two types of (input) image-domain augmentation $A$ are described, each with its corresponding $\tilde{A}$ augmentation in the (output) transformation-domain, for the proposed semi-weakly supervised training.

#### 6.4.1 WarpDDF

As shown in Figure 6.3, consider a randomly sampled spatial transformation, represented by a DDF $U^{\text{aug}} \in \mathbb{R}^{3\times W\times H\times D}$, an image-domain augmentation $A$ transforms the fixed image $I^{\text{fix}}_{(\text{unl})}$ using $U^{\text{aug}}$, while the moving image $I^{\text{mov}}_{(\text{unl})}$ remains unchanged, that is:

$$\begin{aligned} \tilde{I}^{\text{mov}}_{(\text{unl})} &= I^{\text{mov}}_{(\text{unl})} \\ \tilde{I}^{\text{fix}}_{(\text{unl})} &= I^{\text{fix}}_{(\text{unl})} \circ U^{\text{aug}} \end{aligned} \tag{6.11}$$

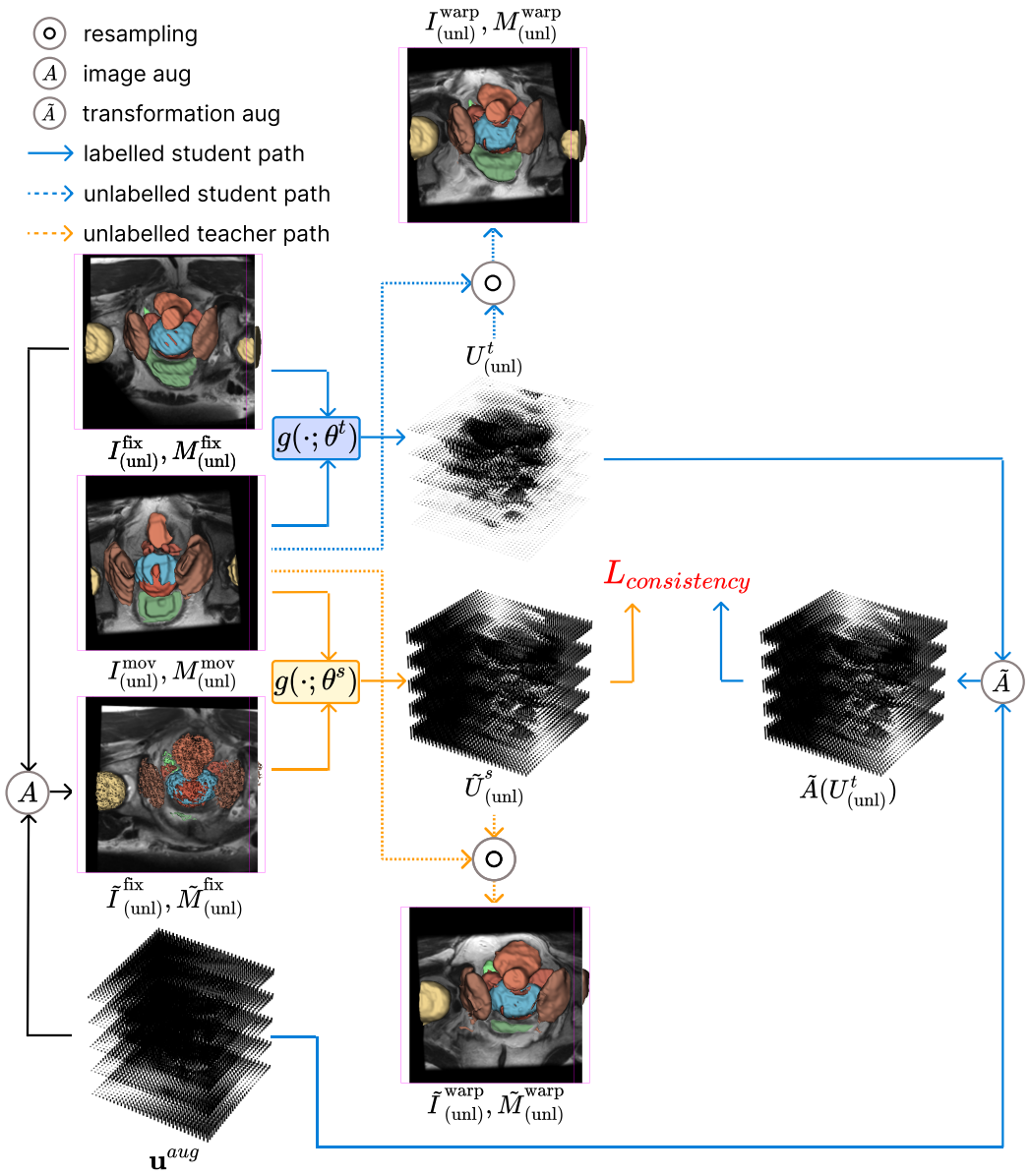



Figure 6.2: DDF-based augmentation transforms the fixed image $I^{\text{fix}}_{(\text{unl})}$ by an augmentation dense displacement field $\approx^{\text{aug}}$ and seek consistency between DDFs predicted from the original and augmented pairs. Note segmentation masks are not available during training, they are included here for illustration purpose.

The corresponding transformation-domain augmentation $\tilde{A}$ becomes:

$$\tilde{A}(U^t_{(\text{unl})}) = U^{\text{aug}} + U^t_{(\text{unl})} \circ U^{\text{aug}} \tag{6.12}$$

where the $U^t_{(\text{unl})} \circ U^{\text{aug}}$ operation generalizes the image resampling (described in Sec. 6.4.1) to the left higher-order-tensor-represented DDF $U^t_{(\text{unl})}$, by independently interpolating each $x$, $y$ and $z$ displacement components (i.e. when $d = 1, 2, 3$, respectively) of $U^t_{(\text{unl})}$, using the right DDF $U^{\text{aug}}$.

We provide the following derivation to show that the two DDFs, $\tilde{U}^s_{(\text{unl})}$ and $\tilde{A}(U^t_{(\text{unl})})$, are expected to be consistent.

$$\begin{aligned} \tilde{I}^{\text{mov}}_{(\text{unl})} \circ \tilde{A}(U^t_{(\text{unl})}) &= \tilde{I}^{\text{mov}}_{(\text{unl})} \circ (U^{\text{aug}} + U^t_{(\text{unl})} \circ U^{\text{aug}}) \\ &= I^{\text{mov}}_{(\text{unl})} \circ U^t_{(\text{unl})} \circ U^{\text{aug}} \\ &= I^{\text{warp}}_{(\text{unl})} \circ U^{\text{aug}} \end{aligned} \tag{6.13}$$

A key insight is that Eq. 6.12 represents a so-called "transformation composing" between $U^t_{(\text{unl})}$ and $U^{\text{aug}}$, which is decomposed and applied to the unchanged moving image $I^{\text{mov}}_{(\text{unl})}$ in Eq. 6.13. This results in applying $U^{\text{aug}}$ on the warped moving image $I^{\text{warp}}_{(\text{unl})}$, which is expected to be aligned with the augmented fixed image $\tilde{I}^{\text{fix}}_{(\text{unl})} = I^{\text{fix}}_{(\text{unl})} \circ U^{\text{aug}}$.

The proposed WarpDDF data augmentation pair $A$ and $\tilde{A}$ is general for representing different types of spatial transformation, including rigid, affine and nonrigid variants. This study adopts random rotation, translation and scaling as an example of the proposed WarpDDF in the experiments reported in Sec. 6.5, as plausible higher-order nonrigid transformation for inter-subject registration remains an open research question.

### 6.4.2 RegCut

RegCut augments the training data by mixing unlabelled moving images with unlabelled fixed images, using a random binary mask $M^{\text{aug}} \in \{1, 0\}^{W \times H \times D}$, indexed with $m^{\text{aug}}_{x,y,z}$. In each training iteration, a rectangular cuboid $B = (r_x, r_y, r_z, r_w, r_h, r_d)$ is sampled, inside which the mask voxels $m^{\text{aug}}_{xyz}$ are labelled as 1, i.e. if $x \in [r_x, r_x + r_w)$, $y \in [r_y, r_y + r_h)$ and $x \in [r_z, r_z + r_d)$, 0 otherwise.

The mask $M^{\text{aug}}$ is used to generate an augmented moving image $\tilde{I}^{\text{mov}}_{(\text{unl})}$, by replacing intensity values with those from the fixed image $I^{\text{fix}}_{(\text{unl})}$ at voxel locations defined by $M^{\text{aug}}$, while, in this case, the fixed image remains unchanged:

$$\begin{aligned} \tilde{I}^{\text{mov}}_{(\text{unl})} &= M^{\text{aug}} \odot I^{\text{fix}}_{(\text{unl})} + (\mathbf{1} - M^{\text{aug}}) \odot I^{\text{mov}}_{(\text{unl})} \\ \tilde{I}^{\text{fix}}_{(\text{unl})} &= I^{\text{fix}}_{(\text{unl})} \end{aligned} \tag{6.14}$$

where $\odot$ is component (element)-wise multiplication and $\mathbf{1}$ is a constant-valued tensor of the same image size.

The corresponding transformation-domain augmentation $\tilde{A}(U^t_{(\text{unl})})$ is defined as follows:

$$\tilde{A}(U^t_{(\text{unl})}) = (\mathbf{1} - M^{\text{aug}}) \odot U^t_{(\text{unl})} \tag{6.15}$$

To show the consistency expected between the augmentation pair $A$ and $\tilde{A}$ (defined in Eqs. 6.14 and 6.15) in the proposed CutReg, we compare the augmented moving image warped by the augmented transformation $\tilde{I}^{\text{mov}}_{(\text{unl})} \circ$

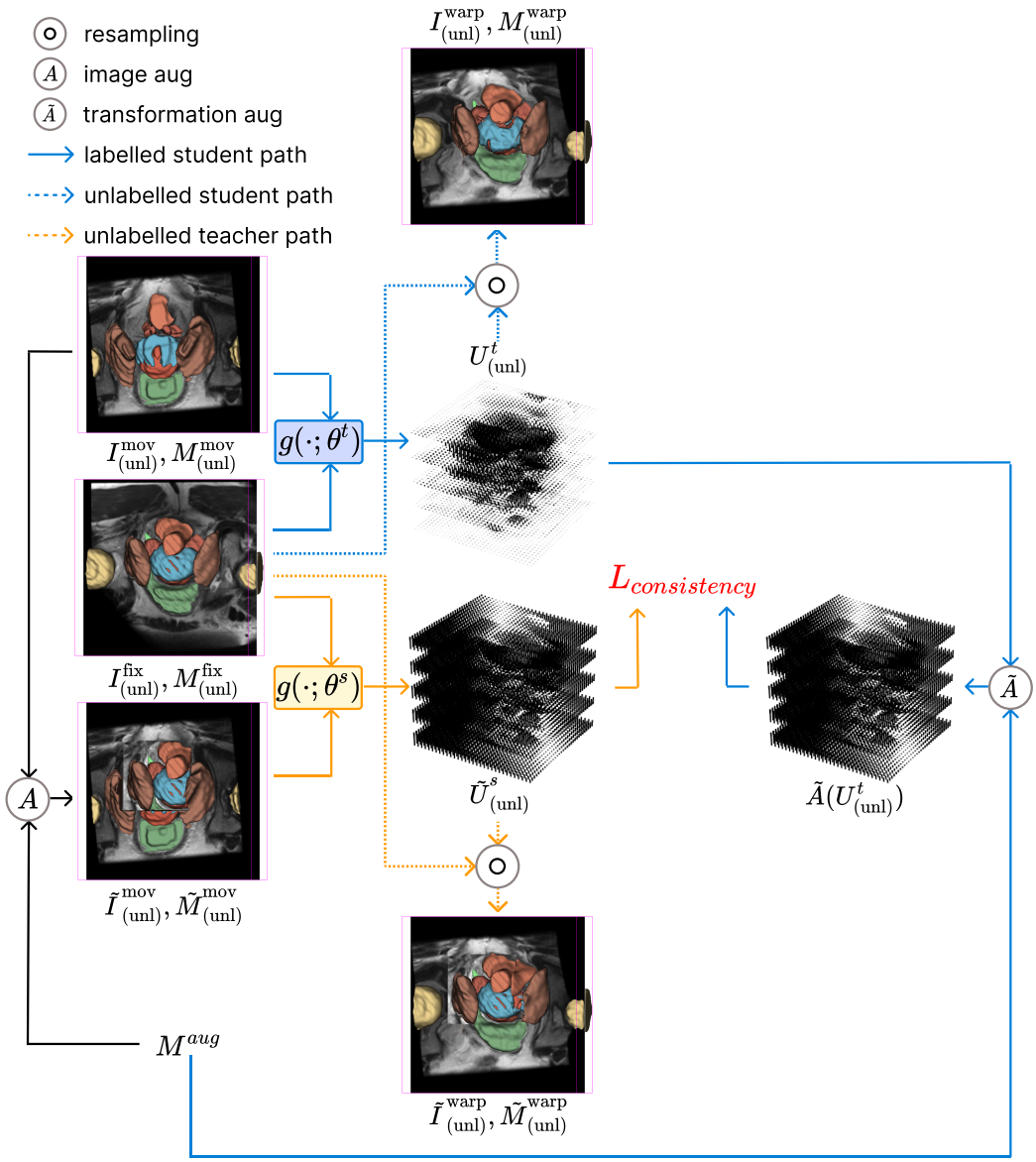



Figure 6.3: RegCut mix unlabelled moving image with unlabelled fixed image following a randomly sampled mask and seek consistency between DDFs predicted from the original and augmented pairs. Note segmentation masks are not available during training, they are included here for illustration purpose.

$\tilde{A}(U^t_{(\text{unl})})$ and the augmented fixed image $\tilde{I}^{\text{fix}}_{(\text{unl})} = I^{\text{fix}}_{(\text{unl})}$:

$$\begin{aligned}
&\tilde{I}^{\text{mov}}_{(\text{unl})} \circ \tilde{A}(U^t_{(\text{unl})}) \\
=&(M^{\text{aug}} \odot I^{\text{fix}}_{(\text{unl})} + (\mathbf{1} - M^{\text{aug}}) \odot I^{\text{mov}}_{(\text{unl})}) \circ (\mathbf{1} - M^{\text{aug}}) \odot U^t_{(\text{unl})} \\
=&M^{\text{aug}} \odot I^{\text{fix}}_{(\text{unl})} + (\mathbf{1} - M^{\text{aug}}) \odot (I^{\text{mov}}_{(\text{unl})} \circ U^t_{(\text{unl})}) \\
=&M^{\text{aug}} \odot I^{\text{fix}}_{(\text{unl})} + (\mathbf{1} - M^{\text{aug}}) \odot I^{\text{warp}}_{(\text{unl})}
\end{aligned} \tag{6.16}$$

where $I^{\text{warp}}_{(\text{unl})} = I^{\text{mov}}_{(\text{unl})} \circ U^t_{(\text{unl})}$ indicates the warped moving image. Eq. 6.16 shows that both regions inside and outside of the mask, $M^{\text{aug}}$ and $\mathbf{1} - M^{\text{aug}}$, are expected to be consistent between the image- and transformation-domain augmentation.

### 6.4.3 Combining WarpDDF and RegCut

By design, the two proposed WarpDDF and RegCut augmentation $A$ can readily be combined together in the input image domain:

$$\begin{aligned}
\tilde{I}^{\text{fix}}_{(\text{unl})} &= T(I^{\text{fix}}_{(\text{unl})}, u^{\text{aug}}) \\
\tilde{I}^{\text{mov}}_{(\text{unl})} &= M^{\text{aug}} \odot \tilde{I}^{\text{fix}}_{(\text{unl})} + (\mathbf{1} - M^{\text{aug}}) \odot I^{\text{mov}}_{(\text{unl})}
\end{aligned} \tag{6.17}$$

Using the similar derivation in Eqs. 6.13 and 6.16, the corresponding output transformation-domain augmentation thus is

$$\tilde{A}(U^t_{(\text{unl})}) = (\mathbf{1} - M) \odot (U^{\text{aug}} + U^t_{(\text{unl})} \circ U^{\text{aug}}) \tag{6.18}$$

## 6.5 Experiments

### 6.5.1 Dataset and Implementation Details

The evaluation is performed on a male lower-pelvic MRI dataset [4] which includes 589 T2-weighted images acquired from the same number of patients. For each image, the dataset provides label for eight anatomical structures of planning interest, including bladder, bone, central gland, neurovascular bundle, obturator internus, rectum, seminal vesicle, transition zone. The images are randomly sampled into training and testing subsets in a 3:1 ratio, resulting in a training set with 442 images and a test set with 147 images.

All images were normalised, resampled and centre-cropped to an image size of $256 \times 256 \times 48$, with a voxel dimension of $0.75 \times 0.75 \times 2.5$ during pre-processing. LocalNet [73] was adopted as the architecture for both student and teacher model with implementation from the MONAI repository [1]. Image perturbation augmented fixed image through random rotation between -5 to 5 degree, scaling between 0.75 to 1.25 and translation between -20 to 20 voxels. The models were trained for 4000 epochs with the first 2000 epochs warmed up on the labelled dataset using only the weakly-supervised loss before adding the weight and image consistency loss for semi-weakly supervision. Training adopts an Adam optimiser starting at a learning rate of 0.00005 with a minibatch size of 1. The training takes around one month to fully converge. The implementation code has been released at `https://github.com/kate-sann5100/LowerPelvicReg`.

Experiments are carried out at various labelled-unlabelled ratios $r$, such that $442 \times r$ images are provided with labels and $442 * (1 - r)$ images are unlabelled. To investigate the effect of utilising unlabelled images, both results with

only labelled images and with all training images are reported. Classical intensity-based registration algorithms, 'NiftyReg' [393], are also compared as a non-learning baseline method.

### 6.5.2 Ablation Studies on Registration Performance

To investigate the effectiveness of the proposed image perturbation methods, we report Dice and the $95^{th}$ percentile Hausdorff distance ($HD^{95}$) results on all eight available anatomical ROIs, after the following variants of registration training strategies,

- 'NoAug': No image perturbation is applied such that teacher and student receive the same image pair.
- 'WarpDDF': Only WarpDDF (Sec. 6.4.1) is applied to the unlabelled fixed image.
- 'RegCut': Only RegCut (Sec. 6.4.2) is used to generate the augmented unlabelled moving image.
- 'WarpDDF+RegCut': Both WarpDDF and RegCut are applied (Sec. 6.4.3).

### 6.5.3 Computational Atlas for Pelvic MR Images

To demonstrate the potential clinical relevance of the semi-weakly supervised registration networks, a lower-pelvic atlas is constructed by registering clinical image samples, from the test set, using the different registration networks. For a set of samples $\{I^i\}_i^N$, an atlas is initialised by choosing the sample that is the most similar to the average of all samples, based on the binary Dice score on their segmentation masks. The final atlas is updated as the average of samples after registering each sample to the initialised atlas as illustrated

in Algorithm 4.

**Algorithm 4** Atlas generation

**input :** Neural network $g$ with parameters $\theta$.
A set of sample $\{I^i\}_i^N$.
**output:** Atlas $A$.
**for** $i \in 1, ..., N$ **do**
  Compute similarity between the sample and the average: $S^i = \text{Sim}(I^i, I^{avg})$
**end**
Choose the sample with highest similarity: $c = \arg\max_i \text{Sim}(I^i, I^{avg})$
Initialise atlas: $A_0 = I^c$

**for** $i \in 1, ..., N$ **do**
  Predict DDF that register the sample to the current atlas: $u_t^i = g(I^i, A_0; \theta)$
  Warp the sample with the predicted DDF: $I_t^i = T(I^i, u_t^i)$
**end**
Update atlas: $A = \frac{1}{N} \sum_{i=0}^{N} I^i$

In this experiment, we define a population diversity $\sigma_{pop}^2$ based on the computational atlas, as an example measure that can be used for subsequent clinical quantitative analysis. The adopted population diversity $\sigma_{pop}^2$ indicate the variance of the established correspondence, represented by the registration-produced displacement field, over all available subject samples. It is calculated as the variance of paired distance between voxels registered to the same atlas locations, averaged over spatial voxel locations, $\sigma_{pop}^2 = \mathbb{E}_{x,y,z}[\text{Var}(\{\|[u^i_{(1)xyz}, u^i_{(2)xyz}, u^i_{(3)xyz}]^\top - [u^j_{(1)xyz}, u^j_{(2)xyz}, u^j_{(3)xyz}]^\top\|^2\}_{i<j})]$ With $u_{dxyz}$ denotes the registration-produced displacement. This measure is an example indicator of morphological and imaging features, which can potentially be predicting or confounding factors for clinical conditions. For example, the benign prostate enlargement is positively indicated by the size of the prostate gland, while the prostate cancer is negatively correlated with

| labelled ratio (%) | method | Dice | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Bladder | Bone | Obturator Internus | Peripheral Zone | Central Gland | Rectum | Seminal Vesicle | Neurovascular Bundle | Mean |
| 0.0 | NifityReg | 9.87±4.53 | 7.61±3.65 | 17.53±4.03 | 6.66±3.39 | 7.58±3.98 | 43.26±5.27 | 3.51±3.21 | 3.74±2.87 | 12.47 |
| 10.0 | sup only | 73.90±4.32 | 83.20±3.28 | 74.95±2.33 | 57.59±4.07 | 75.84±3.58 | 83.07±2.88 | 44.70±4.31 | 37.82±3.88 | 66.38 |
| | NoAug | 74.36±4.64 | 87.46±3.15 | 79.44±2.11 | 63.20±3.95 | 78.87±3.72 | 84.57±3.17 | 49.68±4.74 | 43.68±4.29 | 70.16 |
| | warp | 78.50±4.30 | 89.49±3.03 | 80.73±2.16 | 65.86±4.03 | 81.05±3.26 | 86.86±2.62 | 53.71±4.73 | 47.16±4.31 | 72.92 |
| | RegCut | 77.02±4.68 | 88.43±3.53 | 80.21±2.17 | 65.65±3.89 | 80.04±3.52 | 86.07±2.75 | 55.60±4.67 | 45.56±4.22 | 72.32 |
| | warp+RegCut | 78.04±4.57 | 89.56±3.26 | 81.23±2.11 | 66.05±4.00 | 79.58±3.50 | 87.50±2.79 | 55.10±4.76 | 46.99±4.44 | 73.01 |
| 20.0 | sup only | 77.20±4.42 | 83.71±3.28 | 76.19±2.16 | 61.65±3.73 | 77.34±3.53 | 83.80±2.94 | 50.97±4.19 | 43.26±3.86 | 69.26 |
| | warp+RegCut | 75.68±4.47 | 86.31±3.15 | 78.63±2.25 | 65.98±3.49 | 78.89±3.77 | 85.77±2.62 | 54.71±4.69 | 49.44±3.70 | 71.93 |
| 50.0 | sup only | 85.15±3.06 | 83.89±3.42 | 76.07±2.05 | 68.84±2.76 | 83.38±2.67 | 86.88±2.20 | 58.69±3.83 | 48.98±3.48 | 73.99 |
| | warp+RegCut | 82.55±3.88 | 85.14±3.36 | 77.87±1.96 | 69.45±2.79 | 84.49±2.41 | 87.12±2.30 | 59.73±3.99 | 50.68±3.78 | 74.63 |
| 100.0 | sup only | 83.32±3.81 | 85.96±3.38 | 79.56±1.94 | 69.65±2.90 | 84.17±2.52 | 87.47±2.27 | 60.68±4.20 | 51.49±3.81 | 75.29 |

Table 6.1: Dice(%) of the proposed algorithms and its variants at when labels are available for 10%, 20%, 50% and all training images.

gland size. Statistics about gland volume have been found reliable to measure and straightforward to apply for positioning individual subjects in different (e.g. healthy and pathological) patient cohort populations, in other areas such as preclinical and human neuroimaging studies. The construction of the computational atlas should facilitate such population studies in pelvic organs. In this study, we present experimental results of the above-defined population diversity to quantify their differences due to semi-weakly-supervised registration, where specific hypothesis generation is made and discussed in Sec. 6.6, when interesting results are observed.

## 6.6 Results

Table 6.1 and 6.2 respectively summarised Dice score and 95% Hausdorff distance achieved by the proposed algorithms and its variants at when labels are available for 10%, 20%, 50% and all training images. The proposed method with both strong augmentations achieved an improvement of 6.63%, 2.67% and 0.66% in Dice score with p-value being 8.91e-37, 7.43e-12 and 2.81e-3 as well as a reduction of 1.37 mm, 0.61 mm and 0.24 mm in 95% Hausdorff distance when labels are available for 10%, 20%, 50% of the training images.

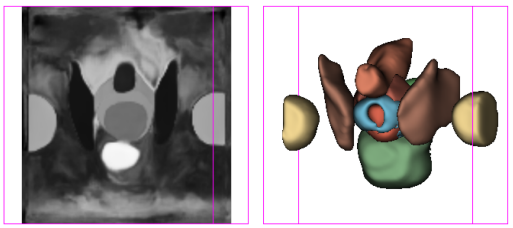
| labelled ratio (%) | method | 95%HD | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Bladder | Bone | Obturator Internus | Peripheral Zone | Central Gland | Rectum | Seminal Vesicle | Neurovascular Bundle | Mean |
| 0.0 | NifityReg | 9.48±4.27 | 42.77±6.52 | 29.95±3.46 | 30.95±4.34 | 17.84±4.45 | 24.79±3.28 | 14.82±4.42 | 37.44±3.95 | 26.01 |
| 10.0 | sup only | 24.92±3.06 | 24.81±3.15 | 20.63±2.72 | 19.98±2.87 | 21.84±3.05 | 23.96±2.79 | 23.79±2.89 | 24.18±2.89 | 23.01 |
| | NoAug | 23.68±2.96 | 23.62±3.10 | 19.55±2.66 | 19.42±2.93 | 20.92±2.93 | 25.20±2.78 | 21.70±2.86 | 23.72±2.99 | 22.23 |
| | warp | 23.62±2.96 | 24.22±3.31 | 19.74±2.61 | 19.03±2.80 | 21.21±2.95 | 22.84±2.67 | 21.99±2.95 | 22.42±2.90 | 21.89 |
| | RegCut | 27.33±3.33 | 25.64±3.36 | 20.55±2.69 | 18.53±2.84 | 20.72±3.00 | 24.03±2.74 | 22.56±3.05 | 23.00±2.93 | 22.79 |
| | warp+RegCut | 23.63±2.98 | 23.96±3.26 | 20.03±2.64 | 18.54±2.79 | 20.52±2.90 | 22.56±2.66 | 22.29±3.00 | 21.52±2.96 | 21.63 |
| 20.0 | sup only | 23.79±3.02 | 25.35±3.24 | 20.48±2.71 | 19.87±2.84 | 22.82±3.06 | 24.32±2.60 | 23.22±2.93 | 23.67±2.79 | 22.94 |
| | warp+RegCut | 24.10±3.02 | 24.35±3.11 | 19.69±2.63 | 19.60±2.91 | 21.89±3.11 | 23.96±2.64 | 22.34±2.99 | 23.23±2.76 | 22.40 |
| 50.0 | sup only | 23.76±3.12 | 25.32±3.22 | 20.86±2.70 | 19.81±2.79 | 21.53±2.99 | 22.85±2.74 | 22.49±2.95 | 22.47±2.72 | 22.39 |
| | warp+RegCut | 23.36±3.09 | 25.02±3.23 | 20.31±2.65 | 19.40±2.82 | 21.43±2.99 | 23.91±2.86 | 22.75±3.01 | 21.71±2.83 | 22.24 |
| 100.0 | sup only | 23.41±3.07 | 25.30±3.25 | 20.21±2.65 | 19.21±2.78 | 21.36±3.02 | 23.50±2.85 | 22.00±2.93 | 21.37±2.78 | 22.05 |

Table 6.2: 95% Hausdorff Distance (mm) of the proposed algorithms and its variants at when labels are available for 10%, 20%, 50% and all training images.

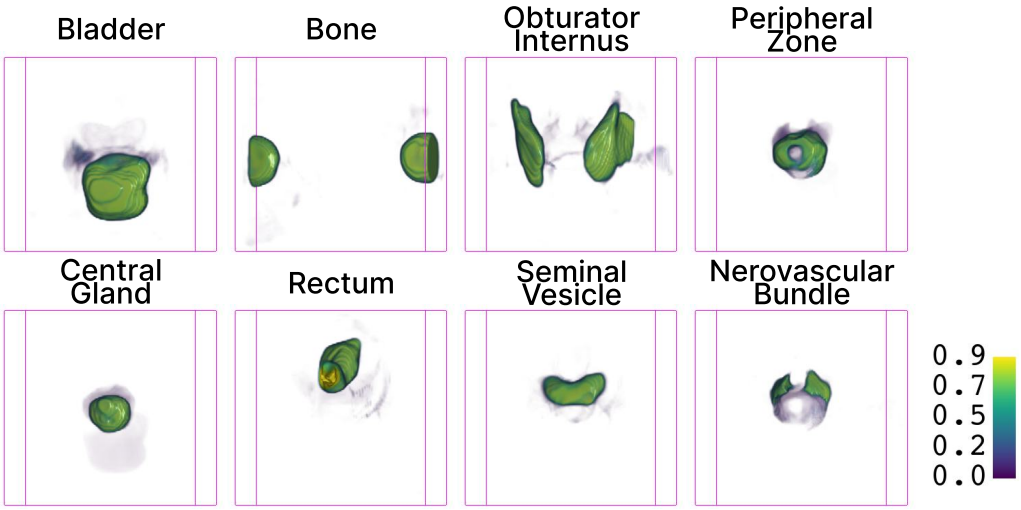
Intuitively the improvement decreases as the labelled ratio increases. Notably, the proposed method trained with a labelled ratio 10% outperformed the fully supervised method trained on the same amount of labelled images, and has a difference of only 2.28% in Dice score to the fully supervised baseline.

Trained on a training set with 10% of the images labelled, the semi-supervision without image perturbation with the proposed augmentations achieved an improvement of 2.78% in Dice score. Adding the warp augmentation, RegCut augmentation and both augmentations respectively achieved a further improvement of 2.76%, 2.16% and 2.85% with p-values being 1.93e-11, 9.47e-9 and 2.37e-10 (paired t-tests at a significance level of $\alpha$=0.05), respectively, proving the efficacy of both proposed augmentations.

Figure 6.4 visualises the obtained atlas as described in Sec. 6.5.3. Table 6.3 clusters the computed population diversity based on their gland zonal volumes (a negative predictor of clinically significant cancer). Use the population diversity difference at the bottom row as a reference, between the top and bottom 20% of the cases (bottom row), when supervised registration



(a)




(b)

Figure 6.4: (a) The resulting atlas and associated segmentation masks. (b) The probability maska of the atlas for the eight structures.

| labelled ratio (%) | method | CG | | |
|---|---|---|---|---|
| | | top 20% | bottom 20% | ratio |
| 10.0 | sup only | 22.67 | 37.52 | 0.60 |
| | NoAug | 22.94 | 40.59 | 0.57 |
| | warp | 20.78 | 28.65 | 0.73 |
| | RegCut | 30.05 | 49.45 | 0.61 |
| | warp+RegCut | 17.68 | 23.93 | 0.74 |
| 20.0 | sup only | 25.76 | 41.42 | 0.62 |
| | warp+RegCut | 20.85 | 33.55 | 0.62 |
| 50.0 | sup only | 25.03 | 35.88 | 0.70 |
| | warp+RegCut | 21.86 | 31.00 | 0.71 |
| 100.0 | sup only | 19.06 | 24.00 | 0.79 |

Table 6.3: Population diversity across the top and bottom 20% cohorts measured by central gland volume of the proposed algorithms and its variants at when labels are available for 10%, 20%, 50% and all training images. The ratio between top and bottom 20% population diversity is reported in the last column.

trained on 100% labelled data. This difference, also indicated by the ratio of 0.79, decreases using fewer labelled data. Compared to alternative methods using a small amount (10%) of labelled data, this difference was indeed increased by the combined DDF+RegCut, reflecting the proposed approaches' ability to recover the difference due to reduction in labelled data.

## 6.7 Conclusion and Discussion

This work proposed semi-weakly supervised methodologies with novel weight and image perturbation to reduce annotation required for segmentation-supervised registration. Experiments demonstrated the efficacy of using unlabelled data with the proposed approaches and augmentations on inter-subject multi-structural lower-pelvic MR registration. We further demonstrated that, using the proposed semi-weakly-supervised registration algorithms, a lower-pelvic anatomical atlas can be constructed and potentially benefits from the added unlabelled data.

Future work can investigate the performance of the lower-pelvic anatomical

atlas across various downstream applications, including aiding in the interpretation of radiological images, intra-patient motion modeling, and anatomical research. For example, benign prostate enlargement is positively indicated by the size of the prostate gland, while prostate cancer is negatively correlated with gland size. Changes in bladder volume may be suggestive of urinary function conditions. Additionally, the shape and content of the rectum correspond to distortions found in diffusion imaging, which is an important sequence in multi-parametric MR imaging for various urological diseases. Metrics for evaluating the atlas could be explored, and better methods for improving atlas generation using registration techniques can be investigated.

# Chapter 7

# Conclusion

This thesis delves into label-efficient learning approaches across various challenges and tasks. It begins with a focus on nature imaging and the few-shot semantic segmentation task. Subsequent research shifts towards medical imaging, where arguably greater challenges in data availability are encountered. Specifically, the second work introduces the novel task of few-shot cross-institution segmentation, requiring algorithms to adapt to not only novel classes but also novel institutions. The third work transitions to the task of registration and proposes to integrate semi-supervision into weakly-supervised registration.

## 7.1 Few-shot Semantic Segmentation

In the first contribution (as described in Chapter 4) of this thesis, we identified and quantitatively validated the deficiency of state-of-the-art methods' performance in segmenting query images with high complexity, particularly with additional classes. We hypothesised that the performance drop stems from the models' inability to effectively distinguish between different fore-

ground classes.

To address this, we proposed a novel self-supervised task that leverages random pseudo-classes generated from the background of query images, thereby utilising extra training data that would otherwise be unavailable when predicting individual target classes.

The experimental results presented in the same chapter demonstrated the effectiveness of the proposed methods across various backbones, architectures, and datasets. Moreover, the majority of the improvement was observed in challenging cases involving the presence of additional classes and previously under-performing classes.

In this study, we employ a simple and effective superpixel-based approach for pseudo-supervision generation. While more advanced deep-learning generation methods, such as the segment anything model (SAM) [347], are available, we refrain from their use to prevent unfair comparisons. This decision is made to avoid information leakage since the training data used for deep-learning generation methods may overlap with the evaluation dataset. However, in medical image applications where such data leakage concerns are not present, future research could explore the potential of using deep-learning models to generate pseudo-classes.

Furthermore, in addition to enhancing inter-class discrimination, future research could explore improving model predictivity by adapting to intra-class variations, such as variations in shape and colour.

## 7.2 Few-shot Cross-institution Segmentation

The second contribution (as described in Chapter 5) extends upon few-shot learning to address the generalisation challenge across institutions caused by domain shifts resulting from discrepancies in image acquisition protocols. We introduce a fully 3D prototypical few-shot segmentation algorithm to train networks capable of effective adaptation to clinically relevant structures absent in training data, using only a few labelled images from a different institute. Specifically, a spatial registration mechanism and a support mask conditioning module are proposed to compensate for spatial variability and leverage available annotations from the support images.

A multi-structure male lower-pelvic dataset, comprising eight anatomical structures of planning interests, has been curated from 589 T2-weighted MRIs. This dataset addresses the need for a sizable dataset with lower-pelvic multi-structure annotations.

The proposed components resulted in performance improvements on this dataset, both independently and collectively. Furthermore, we quantitatively validated the performance degradation when support and query data are acquired from different institutions and demonstrated the efficacy of the proposed spatial alignment mechanism in mitigating this negative effect.

While the proposed spatial alignment mechanism relies on the segmentation of base classes, future work could investigate the potential use of correspondence between supervoxels to alleviate the need for base class structure annotations from the same field-of-view.

Despite the observed improvement in performance, a significant gap remains between the performance of the proposed method and that of supervised

approaches. The research question of achieving accuracy substantial enough for clinical adoption remains unresolved.

Images acquired from different institutions can vary in several ways. While the proposed spatial registration mechanism and support mask conditioning modules address spatial variability across institutions, other types of data variance such as differences in scanners, imaging protocols, and intensity also exist. Future work can independently evaluate the impact of each type of variation on performance and explore solutions to compensate for these other variations with adverse effects.

## 7.3 Semi-weakly Supervised Registration

Weak supervision from auxiliary information such as segmentation and landmarks has been proven effective in improving registration performance. However, the requirement for expensive annotations limits its application in many real-world scenarios. The third contribution of this thesis (as described in Chapter 6) integrates semi-supervision to alleviate the need for extensive annotations by utilising a small labelled dataset alongside a large unlabelled dataset.

We propose a semi-weakly-supervised registration pipeline based on consistency regularisation over both image and weight perturbations to improve model performance. The weight perturbation is implemented by directly adopting the student-teacher paradigm. While image perturbations used in semi-supervised segmentation typically involve both input and output within the image domain, this approach is not directly applicable to the registration task, which involves inputs from the image domain and outputs from the transformation domain. Therefore, we introduce two image augmentations

specifically designed for registration tasks to enable commutative perturbation between an image pair and the predicted spatial transformation.

The proposed perturbations resulted in improved performance both independently and collectively on the multi-structure male lower-pelvic dataset. Moreover, to the best of our knowledge, we made the first attempt at generating a male lower-pelvic atlas, including eight anatomical structures, through registration.

Future work can investigate the potential downstream tasks of the lower-pelvic atlas in image interpretation. For instance, benign prostate enlargement and prostate cancer are positively and negatively correlated with prostate gland size, respectively. Variations in bladder volume can indicate urinary function conditions. Furthermore, the shape and content of the rectum are linked to distortions observed in diffusion-weighted imaging (DWI). Metrics to evaluate the atlas can be explored, and better methods for atlas generation can be investigated.

## 7.4 Conclusion

This thesis has been motivated by the technical intricacy in basic research in computer vision and machine learning, as well as the challenges in multidisciplinary engineering, be it applying existing or innovating solutions to medical imaging problems for real-world patients. The thesis may therefore be concluded by the above discussed future work in both aspects.

# Bibliography

[1] M. J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murrey, A. Myronenko, C. Zhao, D. Yang, *et al.*, "Monai: An open-source framework for deep learning in healthcare," *arXiv preprint arXiv:2211.02701*, 2022.

[2] Y. Li, G. W. P. Data, Y. Fu, Y. Hu, and V. A. Prisacariu, "Few-shot semantic segmentation with self-supervision from pseudo-classes," *arXiv preprint arXiv:2110.11742*, 2021.

[3] Y. Li, Y. Fu, Q. Yang, Z. Min, W. Yan, H. Huisman, D. Barratt, V. A. Prisacariu, and Y. Hu, "Few-shot image segmentation for cross-institution male pelvic organs using registration-assisted prototypical learning," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–5.

[4] Y. Li, Y. Fu, I. J. Gayo, Q. Yang, Z. Min, S. U. Saeed, W. Yan, Y. Wang, J. A. Noble, M. Emberton, *et al.*, "Prototypical few-shot segmentation for cross-institution male pelvic structures with spatial registration," *Medical Image Analysis*, vol. 90, p. 102935, 2023.

[5] A. Heidenreich, P. J. Bastian, J. Bellmunt, M. Bolla, S. Joniau, M. Mason, V. Matveev, N. Mottet, T. Van der Kwast, T. Wiegel, *et al.*, "Guidelines on prostate cancer," *European association of urology*, p. 45, 2012.

[6] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, M. Thun, *et al.*, "Cancer statistics," *Ca Cancer J Clin*, vol. 59, no. 4, 2009.

[7] J. Mohler, R. R. Bahnson, B. Boston, J. E. Busby, A. D'Amico, J. A. Eastham, C. A. Enke, D. George, E. M. Horwitz, R. P. Huben, *et al.*, "Prostate cancer," *Journal of the National Comprehensive Cancer Network*, vol. 8, no. 2, pp. 162–200, 2010.

[8] H. U. Ahmed, A. E.-S. Bosaily, L. C. Brown, R. Gabe, R. Kaplan, M. K. Parmar, Y. Collaco-Moraes, K. Ward, R. G. Hindley, A. Free-

man, *et al.*, "Diagnostic accuracy of multi-parametric mri and trus biopsy in prostate cancer (promis): a paired validating confirmatory study," *The Lancet*, vol. 389, no. 10071, pp. 815–822, 2017.

[9] T. J. Wilt, M. K. Brawer, K. M. Jones, M. J. Barry, W. J. Aronson, S. Fox, J. R. Gingrich, J. T. Wei, P. Gilhooly, B. M. Grob, *et al.*, "Radical prostatectomy versus observation for localized prostate cancer," *N Engl J Med*, vol. 367, pp. 203–213, 2012.

[10] Y. H. Choi, M. Y. Kang, H. H. Sung, H. G. Jeon, B. C. Jeong, S. I. Seo, S. S. Jeon, C. K. Kim, B. K. Park, and H. M. Lee, "Comparison of cancer detection rates between trus-guided biopsy and mri-targeted biopsy according to psa level in biopsy-naive patients: a propensity score matching analysis," *Clinical genitourinary cancer*, vol. 17, no. 1, pp. e19–e25, 2019.

[11] S. Ghafoor, I. A. Burger, and A. H. Vargas, "Multimodality imaging of prostate cancer," *Journal of Nuclear Medicine*, vol. 60, no. 10, pp. 1350–1358, 2019.

[12] G. Segall, D. Delbeke, M. G. Stabin, E. Even-Sapir, J. Fair, R. Sajdak, and G. T. Smith, "Snm practice guideline for sodium 18f-fluoride pet/ct bone scans 1.0," *Journal of Nuclear Medicine*, vol. 51, no. 11, pp. 1813–1820, 2010.

[13] M. Hodolic, "Role of f-choline pet/ct in evaluation of patients with prostate carcinoma," *Radiology and oncology*, vol. 45, no. 1, pp. 17–21, 2011.

[14] S. W. Oh and G. J. Cheon, "Prostate-specific membrane antigen pet imaging in prostate cancer: opportunities and challenges," *Korean journal of radiology*, vol. 19, no. 5, pp. 819–831, 2018.

[15] J. Scheidler, H. Hricak, D. B. Vigneron, K. K. Yu, D. L. Sokolov, L. R. Huang, C. J. Zaloudek, S. J. Nelson, P. R. Carroll, and J. Kurhanewicz, "Prostate cancer: localization with three-dimensional proton mr spectroscopic imaging-clinicopathologic study," *Radiology*, vol. 213, no. 2, pp. 473–480, 1999.

[16] J. C. Weinreb, J. O. Barentsz, P. L. Choyke, F. Cornud, M. A. Haider, K. J. Macura, D. Margolis, M. D. Schnall, F. Shtern, C. M. Tempany, *et al.*, "Pi-rads prostate imaging–reporting and data system: 2015, version 2," *European urology*, vol. 69, no. 1, pp. 16–40, 2016.

[17] V. Kasivisvanathan, A. S. Rannikko, M. Borghi, V. Panebianco, L. A. Mynderse, M. H. Vaarala, A. Briganti, L. Budäus, G. Hellawell, R. G. Hindley, *et al.*, "Mri-targeted or standard biopsy for prostate-cancer diagnosis," *New England Journal of Medicine*, vol. 378, no. 19, pp. 1767–1777, 2018.

[18] H. Hricak, S. White, D. Vigneron, J. Kurhanewicz, A. Kosco, D. Levin, J. Weiss, P. Narayan, and P. R. Carroll, "Carcinoma of the prostate gland: Mr imaging with pelvic phased-array coils versus integrated endorectal–pelvic phased-array coils." *Radiology*, vol. 193, no. 3, pp. 703–709, 1994.

[19] M. de Rooij, E. H. Hamoen, J. A. Witjes, J. O. Barentsz, and M. M. Rovers, "Accuracy of magnetic resonance imaging for local staging of prostate cancer: a diagnostic meta-analysis," *European urology*, vol. 70, no. 2, pp. 233–245, 2016.

[20] N. Mottet, R. C. van den Bergh, E. Briers, T. Van den Broeck, M. G. Cumberbatch, M. De Santis, S. Fanti, N. Fossati, G. Gandaglia, S. Gillessen, *et al.*, "Eau-eanm-estro-esur-siog guidelines on prostate cancer-2020 update. part 1: screening, diagnosis, and local treatment with curative intent," *European urology*, vol. 79, no. 2, pp. 243–262, 2021.

[21] L. Han, T. Au-Yong, W. Tong, K. Chu, L. Szeto, and C. Wong, "Comparison of bone single-photon emission tomography and planar imaging in the detection of vertebral metastases in patients with back pain," *European journal of nuclear medicine*, vol. 25, pp. 635–638, 1998.

[22] G. Shen, H. Deng, S. Hu, and Z. Jia, "Comparison of choline-pet/ct, mri, spect, and bone scintigraphy in the diagnosis of bone metastases in patients with prostate cancer: a meta-analysis," *Skeletal radiology*, vol. 43, pp. 1503–1513, 2014.

[23] N. Fossati, P.-P. M. Willemse, T. Van den Broeck, R. C. van den Bergh, C. Y. Yuan, E. Briers, J. Bellmunt, M. Bolla, P. Cornford, M. De Santis, *et al.*, "The benefits and harms of different extents of lymph node dissection during radical prostatectomy for prostate cancer: a systematic review," *European urology*, vol. 72, no. 1, pp. 84–109, 2017.

[24] T. A. Hope, M. Eiber, W. R. Armstrong, R. Juarez, V. Murthy, C. Lawhn-Heath, S. C. Behr, L. Zhang, F. Barbato, F. Ceci, *et al.*, "Diagnostic accuracy of 68ga-psma-11 pet for pelvic nodal metastasis detection prior to radical prostatectomy and pelvic lymph node dissec-

tion: a multicenter prospective phase 3 imaging trial," *JAMA oncology*, vol. 7, no. 11, pp. 1635–1642, 2021.

[25] H. Hricak, L. Wang, D. C. Wei, F. V. Coakley, O. Akin, V. E. Reuter, M. Gonen, M. W. Kattan, C. N. Onyebuchi, and P. T. Scardino, "The role of preoperative endorectal magnetic resonance imaging in the decision regarding whether to preserve or resect neurovascular bundles during radical retropubic prostatectomy," *Cancer*, vol. 100, no. 12, pp. 2655–2663, 2004.

[26] F. V. Coakley, S. Eberhardt, D. C. Wei, E. S. Wasserman, S. B. Heinze, P. T. Scardino, and H. Hricak, "Blood loss during radical retropubic prostatectomy: relationship to morphologic features on preoperative endorectal magnetic resonance imaging," *Urology*, vol. 59, no. 6, pp. 884–888, 2002.

[27] R. E. Peschel and J. W. Colberg, "Surgery, brachytherapy, and external-beam radiotherapy for early prostate cancer," *The lancet oncology*, vol. 4, no. 4, pp. 233–241, 2003.

[28] J. De la Rosette, H. Ahmed, J. Barentsz, T. B. Johansen, M. Brausi, M. Emberton, F. Frauscher, D. Greene, M. Harisinghani, K. Haustermans, *et al.*, "Focal therapy in prostate cancer-report from a consensus panel," *Journal of Endourology*, vol. 24, no. 5, pp. 775–780, 2010.

[29] B. G. Muller, J. J. Fütterer, R. T. Gupta, A. Katz, A. Kirkham, J. Kurhanewicz, J. W. Moul, P. A. Pinto, A. R. Rastinehad, C. Robertson, *et al.*, "The role of magnetic resonance imaging (mri) in focal therapy for prostate cancer: recommendations from a consensus panel," *BJU international*, vol. 113, no. 2, p. 218, 2014.

[30] J. Tacke, G. Adam, P. Haage, B. Sellhaus, S. Großkortenhaus, and R. W. Günther, "Mr-guided percutaneous cryotherapy of the liver: In vivo evaluation with histologic correlation in an animal model," *Journal of magnetic resonance imaging*, vol. 13, no. 1, pp. 50–56, 2001.

[31] A. Blana, S. Rogenhofer, R. Ganzer, J.-C. Lunz, M. Schostak, W. F. Wieland, and B. Walter, "Eight years' experience with high-intensity focused ultrasonography for treatment of localized prostate cancer," *Urology*, vol. 72, no. 6, pp. 1329–1333, 2008.

[32] A. Gelet, J. Chapelon, R. Bouvier, O. Rouviere, Y. Lasne, D. Lyonnet, and J. Dubernard, "Transrectal high-intensity focused ultrasound: minimally invasive therapy of localized prostate cancer," *Journal of Endourology*, vol. 14, no. 6, pp. 519–528, 2000.

[33] A. Gelet, J. Chapelon, R. Bouvier, O. Rouviere, D. Lyonnet, and J. Dubernard, "Transrectal high intensity focused ultrasound for the treatment of localized prostate cancer: factors influencing the outcome," *European urology*, vol. 40, no. 2, pp. 124–129, 2001.

[34] D. Bonekamp, S. Kohl, M. Wiesenfarth, P. Schelb, J. P. Radtke, M. Götz, P. Kickingereder, K. Yaqubi, B. Hitthaler, N. Gählert, *et al.*, "Radiomic machine learning for characterization of prostate lesions with mri: comparison to adc values," *Radiology*, vol. 289, no. 1, pp. 128–137, 2018.

[35] A. Chaddad, M. J. Kucharczyk, and T. Niazi, "Multimodal radiomic features for the predicting gleason score of prostate cancer," *Cancers*, vol. 10, no. 8, p. 249, 2018.

[36] A. Algohary, R. Shiradkar, S. Pahwa, A. Purysko, S. Verma, D. Moses, R. Shnier, A.-M. Haynes, W. Delprado, J. Thompson, *et al.*, "Combination of peri-tumoral and intra-tumoral radiomic features on biparametric mri accurately stratifies prostate cancer risk: a multi-site study," *Cancers*, vol. 12, no. 8, p. 2200, 2020.

[37] Y. Sumathipala, N. Lay, B. Turkbey, C. Smith, P. L. Choyke, and R. M. Summers, "Prostate cancer detection from multi-institution multiparametric mris using deep convolutional neural networks," *Journal of medical imaging*, vol. 5, no. 4, pp. 044 507–044 507, 2018.

[38] R. Cao, A. M. Bajgiran, S. A. Mirak, S. Shakeri, X. Zhong, D. Enzmann, S. Raman, and K. Sung, "Joint prostate cancer detection and gleason score prediction in mp-mri via focalnet," *IEEE transactions on medical imaging*, vol. 38, no. 11, pp. 2496–2506, 2019.

[39] I. Bhattacharya, A. Seetharaman, W. Shao, R. Sood, C. A. Kunder, R. E. Fan, S. J. C. Soerensen, J. B. Wang, P. Ghanouni, N. C. Teslovich, *et al.*, "Corrsignet: Learning correlated prostate cancer signatures from radiology and pathology images for improved computer aided diagnosis," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*. Springer, 2020, pp. 315–325.

[40] J. Sanyal, I. Banerjee, L. Hahn, and D. Rubin, "An automated two-step pipeline for aggressive prostate lesion detection from multi-parametric mr sequence," *AMIA Summits on Translational Science Proceedings*, vol. 2020, p. 552, 2020.

[41] X. Yu, B. Lou, D. Zhang, D. Winkel, N. Arrahmane, M. Diallo, T. Meng, H. von Busch, R. Grimm, B. Kiefer, *et al.*, "Deep attentive panoptic model for prostate cancer detection using biparametric mri scans," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*. Springer, 2020, pp. 594–604.

[42] A. Saha, M. Hosseinzadeh, and H. Huisman, "End-to-end prostate cancer detection in bpmri via 3d cnns: effects of attention mechanisms, clinical priori and decoupled false positive reduction," *Medical image analysis*, vol. 73, p. 102155, 2021.

[43] I. Bhattacharya, A. Seetharaman, C. Kunder, W. Shao, L. C. Chen, S. J. Soerensen, J. B. Wang, N. C. Teslovich, R. E. Fan, P. Ghanouni, *et al.*, "Selective identification and localization of indolent and aggressive prostate cancers via corrsignia: an mri-pathology correlation and deep learning framework," *Medical Image Analysis*, vol. 75, p. 102288, 2022.

[44] I. Bhattacharya, W. Shao, S. J. Soerensen, R. E. Fan, J. B. Wang, C. Kunder, P. Ghanouni, G. A. Sonn, and M. Rusu, "Integrating zonal priors and pathomic mri biomarkers for improved aggressive prostate cancer detection on mri," in *Medical imaging 2022: computer-aided diagnosis*, vol. 12033. SPIE, 2022, pp. 192–198.

[45] A. Seetharaman, I. Bhattacharya, L. C. Chen, C. A. Kunder, W. Shao, S. J. Soerensen, J. B. Wang, N. C. Teslovich, R. E. Fan, P. Ghanouni, *et al.*, "Automated detection of aggressive and indolent prostate cancer on magnetic resonance imaging," *Medical Physics*, vol. 48, no. 6, pp. 2960–2972, 2021.

[46] M. Hosseinzadeh, A. Saha, P. Brand, I. Slootweg, M. de Rooij, and H. Huisman, "Deep learning–assisted prostate cancer detection on biparametric mri: minimum training data size requirements and effect of prior knowledge," *European Radiology*, pp. 1–11, 2022.

[47] S. D. McGarry, J. D. Bukowy, K. A. Iczkowski, J. G. Unteriner, P. Duvnjak, A. K. Lowman, K. Jacobsohn, M. Hohenwalter, M. O. Griffin, A. W. Barrington, *et al.*, "Gleason probability maps: a radiomics tool for mapping prostate cancer likelihood in mri space," *Tomography*, vol. 5, no. 1, pp. 127–134, 2019.

[48] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature medicine*, vol. 25, no. 8, pp. 1301–1309, 2019.

[49] W. Bulten, M. Balkenhol, J.-J. A. Belinga, A. Brilhante, A. Çakır, L. Egevad, M. Eklund, X. Farré, K. Geronatsiou, V. Molinié, *et al.*, "Artificial intelligence assistance significantly improves gleason grading of prostate biopsies by pathologists," *Modern Pathology*, vol. 34, no. 3, pp. 660–671, 2021.

[50] H. Pinckaers, W. Bulten, J. van der Laak, and G. Litjens, "Detection of prostate cancer in whole-slide images through end-to-end training with image-level labels," *IEEE Transactions on Medical Imaging*, vol. 40, no. 7, pp. 1817–1826, 2021.

[51] H. S. Ryu, M.-S. Jin, J. H. Park, S. Lee, J. Cho, S. Oh, T.-Y. Kwak, J. I. Woo, Y. Mun, S. W. Kim, *et al.*, "Automated gleason scoring and tumor quantification in prostate core needle biopsy images using deep neural networks and its comparison with pathologist-based assessment," *Cancers*, vol. 11, no. 12, p. 1860, 2019.

[52] M. Lucas, I. Jansen, C. D. Savci-Heijink, S. L. Meijer, O. J. de Boer, T. G. van Leeuwen, D. M. de Bruin, and H. A. Marquering, "Deep learning for automatic gleason pattern classification for grade group determination of prostate biopsies," *Virchows Archiv*, vol. 475, pp. 77–83, 2019.

[53] F. Marginean, I. Arvidsson, A. Simoulis, N. C. Overgaard, K. Åström, A. Heyden, A. Bjartell, and A. Krzyzanowska, "An artificial intelligence–based support tool for automation and standardisation of gleason grading in prostate biopsies," *European Urology Focus*, vol. 7, no. 5, pp. 995–1001, 2021.

[54] P. Ström, K. Kartasalo, H. Olsson, L. Solorzano, B. Delahunt, D. M. Berney, D. G. Bostwick, A. J. Evans, D. J. Grignon, P. A. Humphrey, *et al.*, "Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study," *The Lancet Oncology*, vol. 21, no. 2, pp. 222–232, 2020.

[55] R. Narayanan, J. Kurhanewicz, K. Shinohara, E. D. Crawford, A. Simoneau, and J. S. Suri, "Mri-ultrasound registration for targeted

prostate biopsy," in *2009 IEEE International symposium on biomedical imaging: from nano to macro*. IEEE, 2009, pp. 991–994.

[56] A. Ouzzane, R. Renard-Penna, F. Marliere, P. Mozer, J. Olivier, J. Barkatz, P. Puech, and A. Villers, "Magnetic resonance imaging targeted biopsy improves selection of patients considered for active surveillance for clinically low risk prostate cancer based on systematic biopsies," *The Journal of urology*, vol. 194, no. 2, pp. 350–356, 2015.

[57] J. A. Onofrey, L. H. Staib, S. Sarkar, R. Venkataraman, C. B. Nawaf, P. C. Sprenkle, and X. Papademetris, "Learning non-rigid deformations for robust, constrained point-based registration in image-guided mr-trus prostate intervention," *Medical image analysis*, vol. 39, pp. 29–43, 2017.

[58] Y. Lei, S. Tian, X. He, T. Wang, B. Wang, P. Patel, A. B. Jani, H. Mao, W. J. Curran, T. Liu, *et al.*, "Ultrasound prostate segmentation based on multidirectional deeply supervised v-net," *Medical physics*, vol. 46, no. 7, pp. 3194–3206, 2019.

[59] M. Bardis, R. Houshyar, C. Chantaduly, K. Tran-Harding, A. Ushinsky, C. Chahine, M. Rupasinghe, D. Chow, and P. Chang, "Segmentation of the prostate transition zone and peripheral zone on mr images with deep learning," *Radiology: Imaging Cancer*, vol. 3, no. 3, p. e200024, 2021.

[60] L. Stamatakis, M. M. Siddiqui, J. W. Nix, J. Logan, S. Rais-Bahrami, A. Walton-Diaz, A. N. Hoang, S. Vourganti, H. Truong, B. Shuch, *et al.*, "Accuracy of multiparametric magnetic resonance imaging in confirming eligibility for active surveillance for men with prostate cancer," *Cancer*, vol. 119, no. 18, pp. 3359–3366, 2013.

[61] A. W. Diaz, N. A. Shakir, A. K. George, S. Rais-Bahrami, B. Turkbey, J. T. Rothwax, L. Stamatakis, C. W. Hong, M. M. Siddiqui, C. Okoro, *et al.*, "Use of serial multiparametric magnetic resonance imaging in the management of patients with prostate cancer on active surveillance," in *Urologic oncology: seminars and original investigations*, vol. 33, no. 5. Elsevier, 2015, pp. 202–e1.

[62] C. Nguyen, A.-R. Sharif-Afshar, Z. Fan, Y. Xie, S. Wilson, X. Bi, L. Payor, R. Saouaf, H. Kim, and D. Li, "3 d high-resolution diffusion-weighted mri at 3 t: Preliminary application in prostate cancer patients undergoing active surveillance protocol for low-risk prostate cancer," *Magnetic resonance in medicine*, vol. 75, no. 2, pp. 616–626, 2016.

[63] W. S. Lai, J. B. Gordetsky, J. V. Thomas, J. W. Nix, and S. Rais-Bahrami, "Factors predicting prostate cancer upgrading on magnetic resonance imaging–targeted biopsy in an active surveillance population," *Cancer*, vol. 123, no. 11, pp. 1941–1948, 2017.

[64] M. Moazzezi, B. Rose, K. Kisling, K. L. Moore, and X. Ray, "Prospects for daily online adaptive radiotherapy via ethos for prostate cancer patients without nodal involvement using unedited cbct autosegmentation," *Journal of applied clinical medical physics*, vol. 22, no. 10, pp. 82–93, 2021.

[65] C. M. Moore, N. L. Robertson, F. Jichi, A. Damola, G. Ambler, F. Giganti, A. J. Ridout, S. R. Bott, M. Winkler, H. U. Ahmed, *et al.*, "The effect of dutasteride on magnetic resonance imaging defined prostate cancer: Mapped-a randomized, placebo controlled, double-blind clinical trial," *The Journal of urology*, vol. 197, no. 4, pp. 1006–1013, 2017.

[66] T. Matsugasumi, E. Baco, S. Palmer, M. Aron, Y. Sato, N. Fukuda, E. Süer, J.-C. Bernhard, H. Nakagawa, R. A. Azhar, *et al.*, "Prostate cancer volume estimation by combining magnetic resonance imaging and targeted biopsy proven cancer core length: correlation with cancer volume," *The Journal of urology*, vol. 194, no. 4, pp. 957–965, 2015.

[67] J. Thompson, P. Van Leeuwen, D. Moses, R. Shnier, P. Brenner, W. Delprado, M. Pulbrook, M. Böhm, A. M. Haynes, A. Hayen, *et al.*, "The diagnostic performance of multiparametric magnetic resonance imaging to detect significant prostate cancer," *The Journal of urology*, vol. 195, no. 5, pp. 1428–1435, 2016.

[68] F. Khalvati, A. Salmanpour, S. Rahnamayan, M. A. Haider, and H. R. Tizhoosh, "Sequential registration-based segmentation of the prostate gland in mr image volumes," *Journal of digital imaging*, vol. 29, pp. 254–263, 2016.

[69] G. Litjens, R. Toth, W. Van De Ven, C. Hoeks, S. Kerkstra, B. Van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang, *et al.*, "Evaluation of prostate segmentation algorithms for mri: the promise12 challenge," *Medical image analysis*, vol. 18, no. 2, pp. 359–373, 2014.

[70] "Nci-isbi 2013 challenge:automated segmentation of prostate structures," *the cancer imaging archive*, 2015.

[71] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, *et al.*, "A large annotated medical image dataset for the de-

velopment and evaluation of segmentation algorithms," *arXiv preprint arXiv:1902.09063*, 2019.

[72] W. Zhu, A. Myronenko, Z. Xu, W. Li, H. Roth, Y. Huang, F. Milletari, and D. Xu, "Neurreg: Neural registration and its application to image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 3617–3626.

[73] Y. Hu, M. Modat, E. Gibson, W. Li, N. Ghavami, E. Bonmati, G. Wang, S. Bandula, C. M. Moore, M. Emberton, *et al.*, "Weakly-supervised convolutional neural networks for multimodal image registration," *Medical image analysis*, vol. 49, pp. 1–13, 2018.

[74] Y. Hu, M. Modat, E. Gibson, N. Ghavami, E. Bonmati, C. M. Moore, M. Emberton, J. A. Noble, D. C. Barratt, and T. Vercauteren, "Label-driven weakly-supervised learning for multimodal deformable image registration," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 1070–1074.

[75] E. Gibson, Y. Hu, N. Ghavami, *et al.*, "Inter-site variability in prostate segmentation accuracy using deep learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 506–514.

[76] "Data protection act 2018," https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted, 2018.

[77] "Eu gdpr overview," https://gdpr-info.eu/, 2018.

[78] The Common Law Duty of Confidentiality, "Confidentiality," https://www.health-ni.gov.uk/articles/common-law-duty-confidentiality.

[79] W. Yang, A. Liu, J. Wu, and M. Niu, "Prostatic stromal sarcoma: a case report and literature review," *Medicine*, vol. 97, no. 18, p. e0495, 2018.

[80] J. Q. Trinh, S. M. Lele, and B. A. Teply, "A case of metastatic adenoid cystic (basal cell) carcinoma of the prostate: Systemic therapy for a rare disease," *The Prostate*, vol. 83, no. 8, pp. 814–819, 2023.

[81] L. Dickinson, Y. Hu, H. U. Ahmed, *et al.*, "Image-directed, tissue-preserving focal therapy of prostate cancer: a feasibility study of a novel deformable magnetic resonance-ultrasound (mr-us) registration system," *BJU international*, vol. 112, no. 5, p. 594, 2013.

[82] Z. Ren, F. Fang, N. Yan, and Y. Wu, "State of the art in defect detection based on machine vision," *International Journal of Precision Engineering and Manufacturing-Green Technology*, vol. 9, no. 2, pp. 661–691, 2022.

[83] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 5108–5115.

[84] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2020.

[85] K. Men, J. Dai, and Y. Li, "Automatic segmentation of the clinical target volume and organs at risk in the planning ct for rectal cancer using deep dilated convolutional neural networks," *Medical physics*, vol. 44, no. 12, pp. 6377–6389, 2017.

[86] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[87] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.

[88] S. F. Gull and J. Skilling, "Maximum entropy method in image processing," in *Iee proceedings f (communications, radar and signal processing)*, vol. 131, no. 6. IET, 1984, pp. 646–659.

[89] X. Xu, S. Xu, L. Jin, and E. Song, "Characteristic analysis of otsu threshold and its applications," *Pattern recognition letters*, vol. 32, no. 7, pp. 956–961, 2011.

[90] P. Panwar, G. Gopal, and R. Kumar, "Image segmentation using k-means clustering and thresholding," *Image*, vol. 3, no. 05, pp. 1787–1793, 2016.

[91] J. Wu, S. Poehlman, M. D. Noseworthy, and M. V. Kamath, "Texture feature based automated seeded region growing in abdominal mri seg-

mentation," in *2008 international conference on biomedical engineering and informatics*, vol. 2. IEEE, 2008, pp. 263–267.

[92] L. G. Ugarriza, E. Saber, S. R. Vantaram, V. Amuso, M. Shaw, and R. Bhaskar, "Automatic image segmentation by dynamic region growth and multiresolution merging," *IEEE transactions on image processing*, vol. 18, no. 10, pp. 2275–2288, 2009.

[93] A. Thakur and R. S. Anand, "A local statistics based region growing segmentation method for ultrasound medical images," *International Journal of Medical and Health Sciences*, vol. 1, no. 10, pp. 564–569, 2007.

[94] W. Bieniecki, "Oversegmentation avoidance in watershed-based algorithms for color images," in *Proceedings of the International Conference Modern Problems of Radio Engineering, Telecommunications and Computer Science, 2004.* IEEE, 2004, pp. 169–172.

[95] V. Grau, A. Mewes, M. Alcaniz, R. Kikinis, and S. K. Warfield, "Improved watershed transform for medical image segmentation using prior information," *IEEE transactions on medical imaging*, vol. 23, no. 4, pp. 447–458, 2004.

[96] A. Hanbury, "Image segmentation by region based and watershed algorithms," *Wiley Encyclopedia of Computer Science and Engineering*, pp. 1543–1552, 2007.

[97] L. J. Belaid and W. Mourou, "Image segmentation: a watershed transformation algorithm," *Image Analysis & Stereology*, vol. 28, no. 2, pp. 93–102, 2009.

[98] M. N. Ahmed, S. M. Yamany, N. Mohamed, A. A. Farag, and T. Moriarty, "A modified fuzzy c-means algorithm for bias field estimation and segmentation of mri data," *IEEE transactions on medical imaging*, vol. 21, no. 3, pp. 193–199, 2002.

[99] D.-Q. Zhang and S.-C. Chen, "A novel kernelized fuzzy c-means algorithm with application in medical image segmentation," *Artificial intelligence in medicine*, vol. 32, no. 1, pp. 37–50, 2004.

[100] W. Cai, S. Chen, and D. Zhang, "Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation," *Pattern recognition*, vol. 40, no. 3, pp. 825–838, 2007.

[101] D. Aneja and T. K. Rawat, "Fuzzy clustering algorithms for effective medical image segmentation," *International Journal of Intelligent Systems and Applications*, vol. 5, no. 11, pp. 55–61, 2013.

[102] A. Bose and K. Mali, "Fuzzy-based artificial bee colony optimization for gray image segmentation," *Signal, Image and Video Processing*, vol. 10, no. 6, pp. 1089–1096, 2016.

[103] Y. Zhao, Z. Huang, H. Che, F. Xie, M. Liu, M. Wang, D. Sun, *et al.*, "Segmentation of brain tissues from mri images using multitask fuzzy clustering algorithm," *Journal of Healthcare Engineering*, vol. 2023, 2023.

[104] C. Lopez-Molina, B. De Baets, H. Bustince, J. Sanz, and E. Barrenechea, "Multiscale edge detection based on gaussian smoothing and edge tracking," *Knowledge-Based Systems*, vol. 44, pp. 101–111, 2013.

[105] J.-H. Cho, S. B. Cho, *et al.*, "An architecture for real-time hardware co-simulation of edge detection in image processing using prewitt edge operator," in *2014 International Conference on Electronics, Information and Communications (ICEIC)*. IEEE, 2014, pp. 1–2.

[106] O. R. Vincent, O. Folorunso, *et al.*, "A descriptive algorithm for sobel image edge detection," in *Proceedings of informing science & IT education conference (InSITE)*, vol. 40, 2009, pp. 97–107.

[107] N. Salman, "Image segmentation based on watershed and edge detection techniques." *Int. Arab J. Inf. Technol.*, vol. 3, no. 2, pp. 104–110, 2006.

[108] X. Wang, "Laplacian operator-based edge detectors," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 5, pp. 886–890, 2007.

[109] K. Held, E. R. Kops, B. J. Krause, W. M. Wells, R. Kikinis, and H.-W. Muller-Gartner, "Markov random field segmentation of brain mr images," *IEEE transactions on medical imaging*, vol. 16, no. 6, pp. 878–886, 1997.

[110] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm," *IEEE transactions on medical imaging*, vol. 20, no. 1, pp. 45–57, 2001.

[111] Y. Li, Z. Chi, *et al.*, "Mr brain image segmentation based on self-organizing map network," *International Journal of Information Technology*, vol. 11, no. 8, pp. 45–53, 2005.

[112] M. B. Cuadra, C. Pollo, A. Bardera, O. Cuisenaire, J.-G. Villemure, and J.-P. Thiran, "Atlas-based segmentation of pathological mr brain images using a model of lesion growth," *IEEE transactions on medical imaging*, vol. 23, no. 10, pp. 1301–1314, 2004.

[113] H. Park, P. H. Bland, and C. R. Meyer, "Construction of an abdominal probabilistic atlas and its application in segmentation," *IEEE Transactions on medical imaging*, vol. 22, no. 4, pp. 483–492, 2003.

[114] I. Isgum, M. Staring, A. Rutten, M. Prokop, M. A. Viergever, and B. Van Ginneken, "Multi atlas based segmentation with local decision fusion application to cardiac and aortic segmentation in ct scans," *IEEE transactions on medical imaging*, vol. 28, no. 7, pp. 1000–1010, 2009.

[115] D. M. Greig, B. T. Porteous, and A. H. Seheult, "Exact maximum a posteriori estimation for binary images," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 51, no. 2, pp. 271–279, 1989.

[116] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[117] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1520–1528.

[118] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[119] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[120] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, "High-resolution representations for labeling pixels and regions," *arXiv preprint arXiv:1904.04514*, 2019.

[121] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[122] ——, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[123] G. Ghiasi and C. C. Fowlkes, "Laplacian pyramid reconstruction and refinement for semantic segmentation," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 519–534.

[124] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[125] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[126] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3640–3649.

[127] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," *arXiv preprint arXiv:1805.10180*, 2018.

[128] J. R. Petrella, R. E. Coleman, and P. M. Doraiswamy, "Neuroimaging and early diagnosis of alzheimer disease: a look to the future," *Radiology*, vol. 226, no. 2, pp. 315–336, 2003.

[129] M. Hutchinson and U. Raff, "Structural changes of the substantia nigra in parkinson's disease as revealed by mr imaging," *American journal of neuroradiology*, vol. 21, no. 4, pp. 697–701, 2000.

[130] X. Han and B. Fischl, "Atlas renormalization for improved brain mr image segmentation across scanner platforms," *IEEE transactions on medical imaging*, vol. 26, no. 4, pp. 479–486, 2007.

[131] R. D. Howe and Y. Matsuoka, "Robotics for surgery," *Annual review of biomedical engineering*, vol. 1, no. 1, pp. 211–240, 1999.

[132] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer, 2016, pp. 424–432.

[133] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.

[134] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*. Springer, 2018, pp. 3–11.

[135] X. Xiao, S. Lian, Z. Luo, and S. Li, "Weighted res-unet for high-quality retina vessel segmentation," in *2018 9th international conference on information technology in medicine and education (ITME)*. IEEE, 2018, pp. 327–331.

[136] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[137] Z. Wang, N. Zou, D. Shen, and S. Ji, "Non-local u-nets for biomedical image segmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 04, 2020, pp. 6315–6322.

[138] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, *et al.*, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?" *IEEE transactions on medical imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[139] P. F. Christ, M. E. A. Elshaer, F. Ettlinger, S. Tatavarty, M. Bickel, P. Bilic, M. Rempfler, M. Armbruster, F. Hofmann, M. D'Anastasi, *et al.*, "Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields,"

in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer, 2016, pp. 415–423.

[140] Z. Khan, N. Yahya, K. Alsaih, M. I. Al-Hiyali, and F. Meriaudeau, "Recent automatic segmentation algorithms of mri prostate regions: a review," *IEEE Access*, vol. 9, pp. 97 878–97 905, 2021.

[141] C. Wu, S. Montagne, D. Hamzaoui, N. Ayache, H. Delingette, and R. Renard-Penna, "Automatic segmentation of prostate zonal anatomy on mri: a systematic review of the literature," *Insights into Imaging*, vol. 13, no. 1, p. 202, 2022.

[142] S. Bhandary, D. Kuhn, Z. Babaiee, T. Fechter, M. Benndorf, C. Zamboglou, A.-L. Grosu, and R. Grosu, "Investigation and benchmarking of u-nets on prostate segmentation tasks," *Computerized Medical Imaging and Graphics*, p. 102241, 2023.

[143] A. Elnakib, G. Gimel'farb, J. S. Suri, and A. El-Baz, "Medical image segmentation: a brief survey," *Multi modality state-of-the-art medical image segmentation and registration methodologies*, pp. 1–39, 2011.

[144] A. Klein, J. Andersson, B. A. Ardekani, J. Ashburner, B. Avants, M.-C. Chiang, G. E. Christensen, D. L. Collins, J. Gee, P. Hellier, *et al.*, "Evaluation of 14 nonlinear deformation algorithms applied to human brain mri registration," *Neuroimage*, vol. 46, no. 3, pp. 786–802, 2009.

[145] Q. Yang, Y. Fu, F. Giganti, N. Ghavami, Q. Chen, J. A. Noble, T. Vercauteren, D. Barratt, and Y. Hu, "Longitudinal image registration with temporal-order and subject-specificity discrimination," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 243–252.

[146] Y. Hu, H. U. Ahmed, Z. Taylor, C. Allen, M. Emberton, D. Hawkes, and D. Barratt, "Mr to ultrasound registration for image-guided prostate interventions," *Medical image analysis*, vol. 16, no. 3, pp. 687–703, 2012.

[147] S. Klein, *Optimisation methods for medical image registration*. Utrecht University, 2008.

[148] H. Livyatan, Z. Yaniv, and L. Joskowicz, "Gradient-based 2-d/3-d rigid registration of fluoroscopic x-ray to ct," *IEEE transactions on medical imaging*, vol. 22, no. 11, pp. 1395–1406, 2003.

[149] M. Andreetto, G. M. Cortelazzo, and L. Lucchese, "Frequency domain registration of computer tomography data," in *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004.* IEEE, 2004, pp. 550–557.

[150] S. Heger, F. Portheine, J. A. Ohnsorge, E. Schkommodau, and K. Radermacher, "User-interactive registration of bone with a-mode ultrasound," *IEEE engineering in medicine and biology magazine*, vol. 24, no. 2, pp. 85–95, 2005.

[151] L. Tang, G. Hamarneh, and A. Celler, "Co-registration of bone ct and spect images using mutual information," in *2006 IEEE International Symposium on Signal Processing and Information Technology.* IEEE, 2006, pp. 116–121.

[152] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: application to breast mr images," *IEEE transactions on medical imaging*, vol. 18, no. 8, pp. 712–721, 1999.

[153] G. E. Christensen, R. D. Rabbitt, and M. I. Miller, "3d brain mapping using a deformable neuroanatomy," *Physics in Medicine & Biology*, vol. 39, no. 3, p. 609, 1994.

[154] R. Sprengel, K. Rohr, and H. S. Stiehl, "Thin-plate spline approximation for image registration," in *Proceedings of 18th annual international conference of the IEEE Engineering in Medicine and Biology Society*, vol. 3. IEEE, 1996, pp. 1190–1191.

[155] J.-P. Thirion, "Image matching as a diffusion process: an analogy with maxwell's demons," *Medical image analysis*, vol. 2, no. 3, pp. 243–260, 1998.

[156] J. Ashburner and K. J. Friston, "Nonlinear spatial normalization using basis functions," *Human brain mapping*, vol. 7, no. 4, pp. 254–266, 1999.

[157] S. Gefen, O. Tretiak, and J. Nissanov, "Elastic 3-d alignment of rat brain histological images," *IEEE transactions on medical imaging*, vol. 22, no. 11, pp. 1480–1489, 2003.

[158] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain," *Medical image analysis*, vol. 12, no. 1, pp. 26–41, 2008.

[159] P. Viola and W. M. Wells III, "Alignment by maximization of mutual information," *International journal of computer vision*, vol. 24, no. 2, pp. 137–154, 1997.

[160] J. Lötjönen and T. Mäkelä, "Elastic matching using a deformation sphere," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2001: 4th International Conference Utrecht, The Netherlands, October 14–17, 2001 Proceedings 4*. Springer Berlin Heidelberg, 2001, pp. 541–548.

[161] G. Wu, M. Kim, Q. Wang, Y. Gao, S. Liao, and D. Shen, "Unsupervised deep feature learning for deformable registration of mr brain images," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II 16*. Springer, 2013, pp. 649–656.

[162] M. Simonovsky, B. Gutiérrez-Becker, D. Mateus, N. Navab, and N. Komodakis, "A deep metric for multimodal registration," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part III 19*. Springer, 2016, pp. 10–18.

[163] G. Haskins, J. Kruecker, U. Kruger, S. Xu, P. A. Pinto, B. J. Wood, and P. Yan, "Learning deep similarity metric for 3d mr–trus image registration," *International journal of computer assisted radiology and surgery*, vol. 14, pp. 417–425, 2019.

[164] A. Sedghi, J. Luo, A. Mehrtash, S. Pieper, C. M. Tempany, T. Kapur, P. Mousavi, and W. M. Wells III, "Semi-supervised deep metrics for image registration," *arXiv preprint arXiv:1804.01565*, 2018.

[165] R. Liao, S. Miao, P. de Tournemire, S. Grbic, A. Kamen, T. Mansi, and D. Comaniciu, "An artificial agent for robust image registration," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.

[166] K. Ma, J. Wang, V. Singh, B. Tamersoy, Y.-J. Chang, A. Wimmer, and T. Chen, "Multimodal image registration with deep context reinforcement learning," in *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*. Springer, 2017, pp. 240–248.

[167] S. Miao, S. Piat, P. Fischer, A. Tuysuzoglu, P. Mewes, T. Mansi, and R. Liao, "Dilated fcn for multi-agent 2d/3d medical image registration," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[168] J. Krebs, T. Mansi, H. Delingette, L. Zhang, F. C. Ghesu, S. Miao, A. K. Maier, N. Ayache, R. Liao, and A. Kamen, "Robust non-rigid registration through agent-based action learning," in *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*. Springer, 2017, pp. 344–352.

[169] S. Miao, Z. J. Wang, and R. Liao, "A cnn regression approach for real-time 2d/3d registration," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1352–1363, 2016.

[170] X. Yang, R. Kwitt, and M. Niethammer, "Fast predictive image registration," in *Deep Learning and Data Labeling for Medical Applications: First International Workshop, LABELS 2016, and Second International Workshop, DLMIA 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Proceedings 1*. Springer, 2016, pp. 48–57.

[171] M.-M. Rohé, M. Datar, T. Heimann, M. Sermesant, and X. Pennec, "Svf-net: learning deformable image registration using shape matching," in *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*. Springer, 2017, pp. 266–274.

[172] X. Cao, J. Yang, J. Zhang, D. Nie, M. Kim, Q. Wang, and D. Shen, "Deformable image registration based on similarity-steered cnn regression," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*. Springer, 2017, pp. 300–308.

[173] S. Miao, Z. J. Wang, Y. Zheng, and R. Liao, "Real-time 2d/3d registration via cnn regression," in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2016, pp. 1430–1434.

[174] H. Sokooti, B. De Vos, F. Berendsen, B. P. Lelieveldt, I. Išgum, and M. Staring, "Nonrigid image registration using multi-scale 3d convolutional neural networks," in *Medical Image Computing and Computer*

*Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*. Springer, 2017, pp. 232–239.

[175] H. Uzunova, M. Wilms, H. Handels, and J. Ehrhardt, "Training cnns for image registration from few samples with model-based data augmentation," in *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*. Springer, 2017, pp. 223–231.

[176] E. Chee and Z. Wu, "Airnet: Self-supervised affine registration for 3d medical images using neural networks," *arXiv preprint arXiv:1810.02583*, 2018.

[177] K. A. Eppenhof and J. P. Pluim, "Pulmonary ct registration through supervised learning with convolutional neural networks," *IEEE transactions on medical imaging*, vol. 38, no. 5, pp. 1097–1105, 2018.

[178] M. Ito and F. Ino, "An automated method for generating training sets for deep learning based image registration." in *Bioimaging*, 2018, pp. 140–147.

[179] S. S. M. Salehi, S. Khan, D. Erdogmus, and A. Gholipour, "Real-time deep registration with geodesic loss," *arXiv preprint arXiv:1803.05982*, 2018.

[180] J. M. Sloan, K. A. Goatman, and J. P. Siebert, "Learning rigid image registration-utilizing convolutional neural networks for medical image registration," 2018.

[181] J. Zheng, S. Miao, Z. Jane Wang, and R. Liao, "Pairwise domain adaptation module for cnn-based 2-d/3-d registration," *Journal of Medical Imaging*, vol. 5, no. 2, pp. 021 204–021 204, 2018.

[182] M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes, "Computing large deformation metric mappings via geodesic flows of diffeomorphisms," *International journal of computer vision*, vol. 61, pp. 139–157, 2005.

[183] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, "Diffeomorphic demons: Efficient non-parametric image registration," *NeuroImage*, vol. 45, no. 1, pp. S61–S72, 2009.

[184] H. Li and Y. Fan, "Non-rigid image registration using fully convolutional networks with deep self-supervision," *arXiv preprint arXiv:1709.00799*, 2017.

[185] D. Kuang and T. Schmah, "Faim–a convnet method for unsupervised 3d medical image registration," in *Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10*. Springer, 2019, pp. 646–654.

[186] B. D. De Vos, F. F. Berendsen, M. A. Viergever, M. Staring, and I. Išgum, "End-to-end unsupervised deformable image registration with a convolutional neural network," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*. Springer, 2017, pp. 204–212.

[187] E. Ferrante, O. Oktay, B. Glocker, and D. H. Milone, "On the adaptability of unsupervised cnn-based deformable image registration to unseen image domains," in *Machine Learning in Medical Imaging: 9th International Workshop, MLMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 9*. Springer, 2018, pp. 294–302.

[188] C. Shu, X. Chen, Q. Xie, and H. Han, "An unsupervised network for fast microscopic image registration," in *Medical Imaging 2018: Digital Pathology*, vol. 10581. SPIE, 2018, pp. 363–370.

[189] C. Stergios, S. Mihir, V. Maria, C. Guillaume, R. Marie-Pierre, M. Stavroula, and P. Nikos, "Linear and deformable image registration with 3d convolutional neural networks," in *Image Analysis for Moving Organ, Breast, and Thoracic Images: Third International Workshop, RAMBO 2018, Fourth International Workshop, BIA 2018, and First International Workshop, TIA 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16 and 20, 2018, Proceedings 3*. Springer, 2018, pp. 13–22.

[190] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, "Unsupervised learning for fast probabilistic diffeomorphic registration," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*. Springer, 2018, pp. 729–738.

[191] S. Ghosal and N. Ray, "Deep deformable registration: enhancing accuracy by fully convolutional neural net," *Pattern Recognition Letters*,

vol. 94, pp. 81–86, 2017.

[192] A. Sheikhjafari, M. Noga, K. Punithakumar, and N. Ray, "Unsupervised deformable image registration with fully connected generative neural network," in *Medical imaging with deep learning*, 2022.

[193] X. Ying, X. Li, and M. C. Chuah, "Weakly-supervised object representation learning for few-shot semantic segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1497–1506.

[194] A. Hering, S. Kuckertz, S. Heldmann, and M. P. Heinrich, "Enhancing label-driven deep deformable image registration with local distance metrics for state-of-the-art cardiac motion tracking," in *Bildverarbeitung für die Medizin 2019: Algorithmen–Systeme–Anwendungen. Proceedings des Workshops vom 17. bis 19. März 2019 in Lübeck.* Springer, 2019, pp. 309–314.

[195] 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.

[196] L. Fe-Fei *et al.*, "A bayesian approach to unsupervised one-shot learning of object categories," in *proceedings ninth IEEE international conference on computer vision.* IEEE, 2003, pp. 1134–1141.

[197] M. Fink, "Object classification from a single example utilizing class relevance metrics," *Advances in neural information processing systems*, vol. 17, 2004.

[198] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 4, pp. 594–611, 2006.

[199] R. Kwitt, S. Hegenbart, and M. Niethammer, "One-shot learning of scene locations via feature trajectory transfer," in *Proceedings of The IEEE conference on computer vision and pattern recognition*, 2016, pp. 78–86.

[200] B. Hariharan and R. Girshick, "Low-shot visual recognition by shrinking and hallucinating features," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3018–3027.

[201] B. Liu, X. Wang, M. Dixit, R. Kwitt, and N. Vasconcelos, "Feature space transfer for data augmentation," in *Proceedings of the IEEE con-*

*ference on computer vision and pattern recognition*, 2018, pp. 9090–9098.

[202] E. Schwartz, L. Karlinsky, J. Shtok, S. Harary, M. Marder, A. Kumar, R. Feris, R. Giryes, and A. Bronstein, "Delta-encoder: an effective sample synthesis method for few-shot object recognition," *Advances in neural information processing systems*, vol. 31, 2018.

[203] T. Pfister, J. Charles, and A. Zisserman, "Domain-adaptive discriminative one-shot learning of gestures," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*. Springer, 2014, pp. 814–829.

[204] M. Douze, A. Szlam, B. Hariharan, and H. Jégou, "Low-shot learning with large-scale diffusion," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3349–3358.

[205] Y.-H. H. Tsai and R. Salakhutdinov, "Improving one-shot learning through fusing side information," *arXiv preprint arXiv:1710.08347*, 2017.

[206] H. Gao, Z. Shou, A. Zareian, H. Zhang, and S.-F. Chang, "Low-shot learning via covariance-preserving adversarial augmentation networks," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[207] Y. Zhang, H. Tang, and K. Jia, "Fine-grained visual categorization using meta-learning optimization with sample selection of auxiliary data," in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 233–248.

[208] Z. Hu, X. Li, C. Tu, Z. Liu, and M. Sun, "Few-shot charge prediction with discriminative legal attributes," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 487–498.

[209] S. Motiian, Q. Jones, S. Iranmanesh, and G. Doretto, "Few-shot adversarial domain adaptation," *Advances in neural information processing systems*, vol. 30, 2017.

[210] S. Benaim and L. Wolf, "One-shot unsupervised cross domain translation," *advances in neural information processing systems*, vol. 31, 2018.

[211] W. Yan, J. Yap, and G. Mori, "Multi-task transfer methods to improve one-shot learning for multimedia event detection." in *BMVC*, 2015, pp. 37–1.

[212] Z. Luo, Y. Zou, J. Hoffman, and L. F. Fei-Fei, "Label efficient learning of transferable representations acrosss domains and tasks," *Advances in neural information processing systems*, vol. 30, 2017.

[213] G. Koch, R. Zemel, R. Salakhutdinov, *et al.*, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2, no. 1. Lille, 2015.

[214] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," *arXiv preprint arXiv:1606.04080*, 2016.

[215] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.

[216] A. Miller, A. Fisch, J. Dodge, A.-H. Karimi, A. Bordes, and J. Weston, "Key-value memory networks for directly reading documents," *arXiv preprint arXiv:1606.03126*, 2016.

[217] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.

[218] V. Joshi, M. Peters, and M. Hopkins, "Extending a parser to distant domains using a few dozen partially annotated examples," *arXiv preprint arXiv:1805.06556*, 2018.

[219] R. Salakhutdinov, J. Tenenbaum, and A. Torralba, "One-shot learning with a hierarchical nonparametric bayesian model," in *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*. JMLR Workshop and Conference Proceedings, 2012, pp. 195–206.

[220] R. Salakhutdinov, J. B. Tenenbaum, and A. Torralba, "Learning to learn with compound hd models." in *NIPS*, 2011, pp. 2061–2069.

[221] C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt, "Advances in variational inference," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 2008–2026, 2018.

[222] H. Edwards and A. Storkey, "Towards a neural statistician," *arXiv preprint arXiv:1606.02185*, 2016.

[223] L. B. Hewitt, M. I. Nye, A. Gane, T. Jaakkola, and J. B. Tenenbaum, "The variational homoencoder: Learning to learn high capacity gen-

erative models from few examples," *arXiv preprint arXiv:1807.08919*, 2018.

[224] D. Rezende, I. Danihelka, K. Gregor, D. Wierstra, *et al.*, "One-shot generalization in deep generative models," in *International conference on machine learning*. PMLR, 2016, pp. 1521–1529.

[225] S. Reed, Y. Chen, T. Paine, A. v. d. Oord, S. Eslami, D. Rezende, O. Vinyals, and N. de Freitas, "Few-shot autoregressive density estimation: Towards learning to learn distributions," *arXiv preprint arXiv:1710.10304*, 2017.

[226] R. Zhang, T. Che, Z. Ghahramani, Y. Bengio, and Y. Song, "Metagan: An adversarial approach to few-shot learning," *Advances in neural information processing systems*, vol. 31, 2018.

[227] J. Gordon, J. Bronskill, M. Bauer, S. Nowozin, and R. E. Turner, "Meta-learning probabilistic inference for prediction," *arXiv preprint arXiv:1805.09921*, 2018.

[228] Y.-X. Wang and M. Hebert, "Learning from small sample sets by combining unsupervised meta-training with cnns," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[229] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, "One-shot video object segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 221–230.

[230] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," *Advances in neural information processing systems*, vol. 31, 2018.

[231] R. Keshari, M. Vatsa, R. Singh, and A. Noore, "Learning structure and strength of cnn filters for small sample size training," in *proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9349–9358.

[232] J. Kozerawski and M. Turk, "Clear: Cumulative learning for one-shot one-class image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3446–3455.

[233] D. Yoo, H. Fan, V. Boddeti, and K. Kitani, "Efficient k-shot learning with regularized deep networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[234] E. Bart and S. Ullman, "Cross-generalization: Learning novel classes from a single example by feature replacement," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 672–679.

[235] J. Hoffman, E. Tzeng, J. Donahue, Y. Jia, K. Saenko, and T. Darrell, "One-shot adaptation of supervised deep convolutional models," *arXiv preprint arXiv:1312.6204*, 2013.

[236] S. Azadi, M. Fisher, V. G. Kim, Z. Wang, E. Shechtman, and T. Darrell, "Multi-content gan for few-shot font style transfer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7564–7573.

[237] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[238] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *arXiv preprint arXiv:1803.02999*, 2018.

[239] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, "One-shot learning for semantic segmentation," *arXiv preprint arXiv:1709.03410*, 2017.

[240] N. Dong and E. P. Xing, "Few-shot semantic segmentation with prototype learning." in *BMVC*, vol. 3, no. 4, 2018.

[241] Y. Liu, X. Zhang, S. Zhang, and X. He, "Part-aware prototype network for few-shot semantic segmentation," in *European Conference on Computer Vision*. Springer, 2020, pp. 142–158.

[242] B. Yang, C. Liu, B. Li, J. Jiao, and Q. Ye, "Prototype mixture models for few-shot semantic segmentation," in *European Conference on Computer Vision*. Springer, 2020, pp. 763–778.

[243] C. Zhang, G. Lin, F. Liu, J. Guo, Q. Wu, and R. Yao, "Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[244] X. Cao and X. Zhen, "Few-shot semantic segmentation with democratic attention networks," 2020.

[245] Y. Zhao, B. Price, S. Cohen, and D. Gurari, "Objectness-aware one-shot semantic segmentation," *arXiv preprint arXiv:2004.02945*, 2020.

[246] K. Nguyen and S. Todorovic, "Feature weighting and boosting for few-shot segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[247] C. Zhang, G. Lin, F. Liu, R. Yao, and C. Shen, "Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5217–5226.

[248] Z. Tian, H. Zhao, M. Shu, Z. Yang, R. Li, and J. Jia, "Prior guided feature enrichment network for few-shot segmentation," *IEEE Annals of the History of Computing*, 2020.

[249] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "Panet: Few-shot image semantic segmentation with prototype alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9197–9206.

[250] A. G. Roy, S. Siddiqui, S. Pölsterl, *et al.*, "'squeeze & excite'guided few-shot segmentation of volumetric images," *Medical image analysis*, vol. 59, p. 101587, 2020.

[251] S. Guo, L. Xu, C. Feng, H. Xiong, Z. Gao, and H. Zhang, "Multi-level semantic adaptation for few-shot segmentation on cardiac image sequences," *Medical Image Analysis*, vol. 73, p. 102170, 2021.

[252] H. Tang, X. Liu, S. Sun, X. Yan, and X. Xie, "Recurrent mask refinement for few-shot medical image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3918–3928.

[253] S. Hansen, S. Gautam, R. Jenssen, and M. Kampffmeyer, "Anomaly detection-inspired few-shot medical image segmentation through self-supervision with supervoxels," *Medical Image Analysis*, vol. 78, p. 102385, 2022.

[254] R. Wang, Q. Zhou, and G. Zheng, "Few-shot medical image segmentation regularized with self-reference and contrastive learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 514–523.

[255] Q. Shen, Y. Li, J. Jin, and B. Liu, "Q-net: Query-informed few-shot medical image segmentation," *arXiv preprint arXiv:2208.11451*, 2022.

[256] S. Huang, T. Xu, N. Shen, F. Mu, and J. Li, "Rethinking few-shot medical segmentation: A vector quantization view," in *Proceedings of the*

*IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3072–3081.

[257] H. Wu, F. Xiao, and C. Liang, "Dual contrastive learning with anatomical auxiliary supervision for few-shot medical image segmentation," in *European Conference on Computer Vision*. Springer, 2022, pp. 417–434.

[258] Q. Yu, K. Dang, N. Tajbakhsh, *et al.*, "A location-sensitive local prototype network for few-shot medical image segmentation," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 262–266.

[259] C. Ouyang, C. Biffi, C. Chen, T. Kart, H. Qiu, and D. Rueckert, "Self-supervised learning for few-shot medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 41, no. 7, pp. 1837–1848, 2022.

[260] L. Sun, C. Li, X. Ding, Y. Huang, Z. Chen, G. Wang, Y. Yu, and J. Paisley, "Few-shot medical image segmentation using a global correlation network with discriminative embedding," *Computers in biology and medicine*, vol. 140, p. 105067, 2022.

[261] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[262] J. T. Springenberg, "Unsupervised and semi-supervised learning with categorical generative adversarial networks," *arXiv preprint arXiv:1511.06390*, 2015.

[263] A. Odena, "Semi-supervised learning with generative adversarial networks," *arXiv preprint arXiv:1606.01583*, 2016.

[264] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, 2016.

[265] E. Denton, S. Gross, and R. Fergus, "Semi-supervised learning with context-conditional generative adversarial networks," *arXiv preprint arXiv:1611.06430*, 2016.

[266] Z. Dai, Z. Yang, F. Yang, W. W. Cohen, and R. R. Salakhutdinov, "Good semi-supervised learning that requires a bad gan," *Advances in neural information processing systems*, vol. 30, 2017.

[267] G.-J. Qi, L. Zhang, H. Hu, M. Edraki, J. Wang, and X.-S. Hua, "Global versus localized generative adversarial nets," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1517–1525.

[268] A. Kumar, P. Sattigeri, and T. Fletcher, "Semi-supervised learning with gans: Manifold invariance with improved inference," *Advances in neural information processing systems*, vol. 30, 2017.

[269] Z. Deng, H. Zhang, X. Liang, L. Yang, S. Xu, J. Zhu, and E. P. Xing, "Structured generative adversarial networks," *Advances in neural information processing systems*, vol. 30, 2017.

[270] Y. Liu, G. Deng, X. Zeng, S. Wu, Z. Yu, and H.-S. Wong, "Regularizing discriminative capability of cgans for semi-supervised generative learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5720–5729.

[271] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[272] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, "Semi-supervised learning with deep generative models," *Advances in neural information processing systems*, vol. 27, 2014.

[273] L. Maaløe, C. K. Sønderby, S. K. Sønderby, and O. Winther, "Auxiliary deep generative models," in *International conference on machine learning*. PMLR, 2016, pp. 1445–1453.

[274] T. Joy, S. M. Schmon, P. H. Torr, N. Siddharth, and T. Rainforth, "Capturing label characteristics in vaes," *arXiv preprint arXiv:2006.10102*, 2020.

[275] M. Ehsan Abbasnejad, A. Dick, and A. van den Hengel, "Infinite variational autoencoder for semi-supervised learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5888–5897.

[276] B. Paige, J.-W. van de Meent, A. Desmaison, N. Goodman, P. Kohli, F. Wood, P. Torr, *et al.*, "Learning disentangled representations with semi-supervised deep generative models," *Advances in neural information processing systems*, vol. 30, 2017.

[277] Y. Li, Q. Pan, S. Wang, H. Peng, T. Yang, and E. Cambria, "Disentangled variational auto-encoder for semi-supervised learning," *Information Sciences*, vol. 482, pp. 73–85, 2019.

[278] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," *Advances in neural information processing systems*, vol. 28, 2015.

[279] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[280] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," *Advances in neural information processing systems*, vol. 33, pp. 6256–6268, 2020.

[281] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.

[282] G. French, S. Laine, T. Aila, M. Mackiewicz, and G. Finlayson, "Semi-supervised semantic segmentation needs strong, varied perturbations," *arXiv preprint arXiv:1906.01916*, 2019.

[283] Y. Shu, H. Li, B. Xiao, X. Bi, and W. Li, "Cross-mix monitoring for medical image segmentation with limited supervision," *IEEE Transactions on Multimedia*, 2022.

[284] H. Basak, R. Bhattacharya, R. Hussain, and A. Chatterjee, "An embarrassingly simple consistency regularization method for semi-supervised medical image segmentation," *arXiv preprint arXiv:2202.00677*, 2022.

[285] W. Huang, C. Chen, Z. Xiong, Y. Zhang, X. Chen, X. Sun, and F. Wu, "Semi-supervised neuron segmentation via reinforced consistency learning," *IEEE Transactions on Medical Imaging*, vol. 41, no. 11, pp. 3016–3028, 2022.

[286] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson, "Classmix: Segmentation-based data augmentation for semi-supervised learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1369–1378.

[287] Y. Chen, X. Ouyang, K. Zhu, and G. Agam, "Complexmix: Semi-supervised semantic segmentation via mask-based data augmentation," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 2264–2268.

[288] L. Zhang and G.-J. Qi, "Wcp: Worst-case perturbations for semi-supervised deep learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3912–3921.

[289] S. An, H. Zhu, J. Zhang, J. Ye, S. Wang, J. Yin, and H. Zhang, "Deep tri-training for semi-supervised image segmentation," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 097–10 104, 2022.

[290] X. Zeng, R. Huang, Y. Zhong, D. Sun, C. Han, D. Lin, D. Ni, and Y. Wang, "Reciprocal learning for semi-supervised segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*. Springer, 2021, pp. 352–361.

[291] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2613–2622.

[292] S. Park, J. Park, S.-J. Shin, and I.-C. Moon, "Adversarial dropout for supervised and semi-supervised learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[293] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer, 2019, pp. 605–613.

[294] Z. Xie, E. Tu, H. Zheng, Y. Gu, and J. Yang, "Semi-supervised skin lesion segmentation with learning model confidence," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 1135–1139.

[295] X. Luo, G. Wang, W. Liao, J. Chen, T. Song, Y. Chen, S. Zhang, D. N. Metaxas, and S. Zhang, "Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency," *Medical Image Analysis*, vol. 80, p. 102517, 2022.

[296] Z. Fang, J. Bai, X. Guo, X. Wang, F. Gao, H.-Y. Yang, B. Kong, Y. Hou, K. Cao, Q. Song, *et al.*, "Annotation-efficient covid-19 pneumonia lesion segmentation using error-aware unified semisupervised and

active learning," *IEEE Transactions on Artificial Intelligence*, vol. 4, no. 2, pp. 255–267, 2022.

[297] C. Chen, K. Zhou, Z. Wang, and R. Xiao, "Generative consistency for semi-supervised cerebrovascular segmentation from tof-mra," *IEEE Transactions on Medical Imaging*, vol. 42, no. 2, pp. 346–353, 2022.

[298] A. Xu, S. Wang, S. Ye, J. Fan, X. Shi, and X. Xia, "Ca-mt: A self-ensembling model for semi-supervised cardiac segmentation with elliptical descriptor based contour-aware," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–5.

[299] G. Chen, J. Ru, Y. Zhou, I. Rekik, Z. Pan, X. Liu, Y. Lin, B. Lu, and J. Shi, "Mtans: Multi-scale mean teacher combined adversarial network with shape-aware embedding for semi-supervised brain lesion segmentation," *NeuroImage*, vol. 244, p. 118568, 2021.

[300] Y. Xie, J. Zhang, Z. Liao, J. Verjans, C. Shen, and Y. Xia, "Intra-and inter-pair consistency for semi-supervised gland segmentation," *IEEE Transactions on Image Processing*, vol. 31, pp. 894–905, 2021.

[301] L. Hu, J. Li, X. Peng, J. Xiao, B. Zhan, C. Zu, X. Wu, J. Zhou, and Y. Wang, "Semi-supervised npc segmentation with uncertainty and attention guided consistency," *Knowledge-Based Systems*, vol. 239, p. 108021, 2022.

[302] Y. Liu, W. Wang, G. Luo, K. Wang, and S. Li, "A contrastive consistency semi-supervised left atrium segmentation model," *Computerized Medical Imaging and Graphics*, vol. 99, p. 102092, 2022.

[303] Z. Xu, Y. Wang, D. Lu, L. Yu, J. Yan, J. Luo, K. Ma, Y. Zheng, and R. K.-y. Tong, "All-around real label supervision: Cyclic prototype consistency learning for semi-supervised medical image segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 7, pp. 3174–3184, 2022.

[304] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2, 2013, p. 896.

[305] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory*, 1998, pp. 92–100.

[306] W. Dong-DongChen and Z. WeiGao, "Tri-net for semi-supervised deep learning," in *Proceedings of twenty-seventh international joint conference on artificial intelligence*, 2018, pp. 2014–2020.

[307] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 687–10 698.

[308] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S4l: Self-supervised semi-supervised learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1476–1485.

[309] H. Pham, Z. Dai, Q. Xie, and Q. V. Le, "Meta pseudo labels," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 557–11 568.

[310] X. Wang, D. Kihara, J. Luo, and G.-J. Qi, "Enaet: A self-trained framework for semi-supervised and supervised learning with ensemble transformations," *IEEE Transactions on Image Processing*, vol. 30, pp. 1639–1647, 2020.

[311] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," *Advances in neural information processing systems*, vol. 33, pp. 22 243–22 255, 2020.

[312] H. Yao, X. Hu, and X. Li, "Enhancing pseudo label quality for semi-supervised domain-generalized medical image segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 3099–3107.

[313] Y. Shi, J. Zhang, T. Ling, J. Lu, Y. Zheng, Q. Yu, L. Qi, and Y. Gao, "Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation," *IEEE transactions on medical imaging*, vol. 41, no. 3, pp. 608–620, 2021.

[314] X. Wang, Y. Yuan, D. Guo, X. Huang, Y. Cui, M. Xia, Z. Wang, C. Bai, and S. Chen, "Ssa-net: Spatial self-attention network for covid-19 pneumonia infection segmentation with semi-supervised few-shot learning," *Medical Image Analysis*, vol. 79, p. 102459, 2022.

[315] C. Li, L. Dong, Q. Dou, F. Lin, K. Zhang, Z. Feng, W. Si, X. Deng, Z. Deng, and P.-A. Heng, "Self-ensembling co-training framework for

semi-supervised covid-19 ct segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 11, pp. 4140–4151, 2021.

[316] B. H. Thompson, G. Di Caterina, and J. P. Voisey, "Pseudo-label refinement using superpixels for semi-supervised brain tumour segmentation," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–5.

[317] K. Han, L. Liu, Y. Song, Y. Liu, C. Qiu, Y. Tang, Q. Teng, and Z. Liu, "An effective semi-supervised approach for liver ct image segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 3999–4007, 2022.

[318] R. Wang, Y. Wu, H. Chen, L. Wang, and D. Meng, "Neighbor matching for semi-supervised learning," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*. Springer, 2021, pp. 439–449.

[319] C. M. Seibold, S. Reiß, J. Kleesiek, and R. Stiefelhagen, "Reference-guided pseudo-label generation for medical semantic segmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 2, 2022, pp. 2171–2179.

[320] H. Zheng, L. Lin, H. Hu, Q. Zhang, Q. Chen, Y. Iwamoto, X. Han, Y.-W. Chen, R. Tong, and J. Wu, "Semi-supervised segmentation of liver using adversarial learning with deep atlas prior," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*. Springer, 2019, pp. 148–156.

[321] H. Huang, Q. Chen, L. Lin, M. Cai, Q. Zhang, Y. Iwamoto, X. Han, A. Furukawa, S. Kanasaki, Y.-W. Chen, *et al.*, "Mtl-abs 3 net: Atlas-based semi-supervised organ segmentation network with multi-task learning for medical images," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 3988–3998, 2022.

[322] Y. He, G. Yang, J. Yang, Y. Chen, Y. Kong, J. Wu, L. Tang, X. Zhu, J.-L. Dillenseger, P. Shao, *et al.*, "Dense biased networks with deep priori anatomy and hard region adaptation: Semi-supervised learning for fine renal artery segmentation," *Medical image analysis*, vol. 63, p. 101722, 2020.

[323] J. Yang, Y. Tao, Q. Xu, Y. Zhang, X. Ma, S. Yuan, and Q. Chen, "Self-supervised sequence recovery for semi-supervised retinal layer segmen-

tation," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 3872–3883, 2022.

[324] B. Zhang, J. Xiao, and T. Qin, "Self-guided and cross-guided learning for few-shot segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8312–8321.

[325] J. Peng, P. Wang, C. Desrosiers, and M. Pedersoli, "Self-paced contrastive learning for semi-supervised medical image segmentation with meta-labels," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 686–16 699, 2021.

[326] H. Wu, Z. Wang, Y. Song, L. Yang, and J. Qin, "Cross-patch dense contrastive learning for semi-supervised segmentation of cellular nuclei in histopathologic images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 666–11 675.

[327] Z. Zhao, J. Hu, Z. Zeng, X. Yang, P. Qian, B. Veeravalli, and C. Guan, "Mmgl: Multi-scale multi-view global-local contrastive learning for semi-supervised cardiac image segmentation," in *2022 IEEE international conference on image processing (ICIP)*. IEEE, 2022, pp. 401–405.

[328] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[329] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[330] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[331] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[332] J.-J. Hwang, S. X. Yu, J. Shi, M. D. Collins, T.-J. Yang, X. Zhang, and L.-C. Chen, "Segsort: Segmentation by discriminative sorting of

segments," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7334–7344.

[333] B. Mičušĺík and J. Košecká, "Semantic segmentation of street scenes by superpixel co-occurrence and 3d geometry," in *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*. IEEE, 2009, pp. 625–632.

[334] C. Ouyang, C. Biffi, C. Chen, T. Kart, H. Qiu, and D. Rueckert, "Self-supervision with superpixels: Training few-shot medical image segmentation without annotation," in *European Conference on Computer Vision*. Springer, 2020, pp. 762–780.

[335] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International journal of computer vision*, vol. 59, no. 2, pp. 167–181, 2004.

[336] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *European Conference on Computer Vision*. Springer, 2014, pp. 297–312.

[337] K. Rakelly, E. Shelhamer, T. Darrell, A. Efros, and S. Levine, "Conditional networks for few-shot semantic segmentation," 2018.

[338] M. Siam, B. N. Oreshkin, and M. Jagersand, "Amp: Adaptive masked proxies for few-shot segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[339] X. Zhang, Y. Wei, Y. Yang, and T. S. Huang, "Sg-one: Similarity guidance network for one-shot semantic segmentation," *IEEE Transactions on Cybernetics*, 2020.

[340] W. Liu, C. Zhang, G. Lin, and F. Liu, "Crnet: Cross-reference networks for few-shot segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4165–4173.

[341] X. Ying, X. Li, and M. C. Chuah, "Weakly-supervised object representation learning for few-shot semantic segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2021, pp. 1497–1506.

[342] G. Li, V. Jampani, L. Sevilla-Lara, D. Sun, J. Kim, and J. Kim, "Adaptive prototype learning and allocation for few-shot segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8334–8343.

[343] M. Boudiaf, H. Kervadec, Z. I. Masud, P. Piantanida, I. Ben Ayed, and J. Dolz, "Few-shot segmentation without meta-learning: A good transductive inference is all you need?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 979–13 988.

[344] M. LJPvd and G. Hinton, "Visualizing high-dimensional data using t-sne," *J Mach Learn Res*, vol. 9, pp. 2579–2605, 2008.

[345] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[346] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1395–1403.

[347] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[348] L. Henschel, S. Conjeti, S. Estrada, K. Diers, B. Fischl, and M. Reuter, "Fastsurfer-a fast and accurate deep learning based neuroimaging pipeline," *NeuroImage*, vol. 219, p. 117012, 2020.

[349] A. D. Weston, P. Korfiatis, T. L. Kline, K. A. Philbrick, P. Kostandy, T. Sakinis, M. Sugimoto, N. Takahashi, and B. J. Erickson, "Automated abdominal segmentation of ct scans for body composition analysis using deep learning," *Radiology*, vol. 290, no. 3, pp. 669–679, 2019.

[350] C. Fiorino, M. Reni, A. Bolognesi, G. M. Cattaneo, and R. Calandrino, "Intra-and inter-observer variability in contouring prostate and seminal vesicles: implications for conformal treatment planning," *Radiotherapy and oncology*, vol. 47, no. 3, pp. 285–292, 1998.

[351] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1199–1208.

[352] J. Ren, I. Hacihaliloglu, E. A. Singer, D. J. Foran, and X. Qi, "Adversarial domain adaptation for classification of prostate histopathology whole-slide images," in *International conference on medical image*

*computing and computer-assisted intervention*. Springer, 2018, pp. 201–209.

[353] A. Meyer, A. Mehrtash, M. Rak, O. Bashkanov, B. Langbein, A. Ziaei, A. S. Kibel, C. M. Tempany, C. Hansen, and J. Tokuda, "Domain adaptation for segmentation of critical structures for prostate cancer therapy," *Scientific reports*, vol. 11, no. 1, pp. 1–14, 2021.

[354] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "Fedbn: Federated learning on non-iid features via local batch normalization," *arXiv preprint arXiv:2102.07623*, 2021.

[355] A. R. Feyjie, R. Azad, M. Pedersoli, C. Kauffman, I. B. Ayed, and J. Dolz, "Semi-supervised few-shot learning for medical image segmentation," *arXiv preprint arXiv:2003.08462*, 2020.

[356] M. Abdel-Basset, V. Chang, H. Hawash, R. K. Chakrabortty, and M. Ryan, "Fss-2019-ncov: A deep learning architecture for semi-supervised few-shot segmentation of covid-19 infection," *Knowledge-Based Systems*, vol. 212, p. 106647, 2021.

[357] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2017, pp. 240–248.

[358] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein, *et al.*, "The future of digital health with federated learning," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–7, 2020.

[359] E. Hosseini-Asl, R. Keynton, and A. El-Baz, "Alzheimer's disease diagnostics by adaptation of 3d convolutional network," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 126–130.

[360] C. S. Perone, P. Ballester, R. C. Barros, and J. Cohen-Adad, "Unsupervised domain adaptation for medical imaging segmentation with self-ensembling," *NeuroImage*, vol. 194, pp. 1–11, 2019.

[361] Y. Fu, M. R. Robu, B. Koo, C. Schneider, S. v. Laarhoven, D. Stoyanov, B. Davidson, M. J. Clarkson, and Y. Hu, "More unlabelled data or label more data? a study on semi-supervised laparoscopic image segmentation," in *Domain Adaptation and Representation Transfer and Medical*

*Image Learning with Less Labels and Imperfect Data*. Springer, 2019, pp. 173–180.

[362] Y. Xia, D. Yang, Z. Yu, F. Liu, J. Cai, L. Yu, Z. Zhu, D. Xu, A. Yuille, and H. Roth, "Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation," *Medical Image Analysis*, vol. 65, p. 101766, 2020.

[363] A. Li, S. Wang, W. Li, S. Liu, and S. Zhang, "Predicting human mobility with federated learning," in *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, 2020, pp. 441–444.

[364] D. Tomar, M. Lortkipanidze, G. Vray, B. Bozorgtabar, and J.-P. Thiran, "Self-attentive spatial adaptive normalization for cross-modality domain adaptation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 10, pp. 2926–2938, 2021.

[365] L. Dickinson, H. U. Ahmed, A. Kirkham, C. Allen, A. Freeman, J. Barber, R. G. Hindley, T. Leslie, C. Ogden, R. Persad, *et al.*, "A multicentre prospective development study evaluating focal therapy using high intensity focused ultrasound for localised prostate cancer: the index study," *Contemporary clinical trials*, vol. 36, no. 1, pp. 68–80, 2013.

[366] S. Hamid, I. A. Donaldson, Y. Hu, *et al.*, "The smarttarget biopsy trial: a prospective, within-person randomised, blinded trial comparing the accuracy of visual-registration and magnetic resonance imaging/ultrasound image-fusion targeted biopsies for prostate cancer risk stratification," *European urology*, vol. 75, no. 5, pp. 733–740, 2019.

[367] L. A. Simmons, H. U. Ahmed, C. M. Moore, S. Punwani, A. Freeman, Y. Hu, D. Barratt, S. C. Charman, J. Van der Meulen, and M. Emberton, "The picture study-prostate imaging (multi-parametric mri and prostate histoscanning) compared to transperineal ultrasound guided biopsy for significant prostate cancer risk evaluation," *Contemporary clinical trials*, vol. 37, no. 1, pp. 69–83, 2014.

[368] G. Litjens, J. Futterer, and H. Huisman, "Data from prostate-3t," *The Cancer Imaging Archive*, 2015.

[369] B. N. Bloch, A. Jain, and C. C. Jaffe, "Data from prostate-diagnosis," *The Cancer Imaging Archive*, 2015.

[370] P. Choyke, B. Turkbey, P. Pinto, M. Merino, and B. Wood, "Data from prostate-mri," *The Cancer Imaging Archive*, 2016.

[371] S. Kim, S. An, P. Chikontwe, and S. H. Park, "Bidirectional rnn-based few shot learning for 3d medical image segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 1808–1816.

[372] W. Yan, Q. Yang, T. Syer, Z. Min, S. Punwani, M. Emberton, D. Barratt, B. Chiu, and Y. Hu, "The impact of using voxel-level segmentation metrics on evaluating multifocal prostate cancer localisation," in *Applications of Medical Artificial Intelligence: First International Workshop, AMAI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings*. Springer, 2022, pp. 128–138.

[373] G. Haskins, U. Kruger, and P. Yan, "Deep learning in medical image registration: a survey," *Machine Vision and Applications*, vol. 31, pp. 1–18, 2020.

[374] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang, "Deep learning in medical image registration: a review," *Physics in Medicine & Biology*, vol. 65, no. 20, p. 20TR01, 2020.

[375] N. J. Tustison, B. B. Avants, and J. C. Gee, "Learning image-based spatial transformations via convolutional neural networks: A review," *Magnetic resonance imaging*, vol. 64, pp. 142–153, 2019.

[376] Y. Hu, "Registration of magnetic resonance and ultrasound images for guiding prostate cancer interventions," Ph.D. dissertation, UCL (University College London), 2013.

[377] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[378] B. Azad, R. Azad, S. Eskandari, A. Bozorgpour, A. Kazerouni, I. Rekik, and D. Merhof, "Foundational models in medical imaging: A comprehensive survey and future vision," *arXiv preprint arXiv:2310.18689*, 2023.

[379] Y. Tang and M. Zephyr. (2023) Visual foundation models for medical image analysis. [Online]. Available: https://developer.nvidia.com/blog/visual-foundation-models-for-medical-image-analysis/

[380] Y. Zhou, M. A. Chia, S. K. Wagner, M. S. Ayhan, D. J. Williamson, R. R. Struyven, T. Liu, M. Xu, M. G. Lozano, P. Woodward-Court,

*et al.*, "A foundation model for generalizable disease detection from retinal images," *Nature*, pp. 1–8, 2023.

[381] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[382] A. Sheikhjafari, M. Noga, K. Punithakumar, and N. Ray, "Unsupervised deformable image registration with fully connected generative neural network," 2018.

[383] N. Souly, C. Spampinato, and M. Shah, "Semi supervised semantic segmentation using generative adversarial network," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5688–5696.

[384] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," *arXiv preprint arXiv:1802.07934*, 2018.

[385] Z. Feng, Q. Zhou, G. Cheng, X. Tan, J. Shi, and L. Ma, "Semi-supervised semantic segmentation via dynamic self-training and class-balanced curriculum," *arXiv preprint arXiv:2004.08514*, vol. 1, no. 2, p. 5, 2020.

[386] Z. Ke, D. Qiu, K. Li, Q. Yan, and R. W. Lau, "Guided collaborative training for pixel-wise semi-supervised learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*. Springer, 2020, pp. 429–445.

[387] Y. Liu, Y. Tian, Y. Chen, F. Liu, V. Belagiannis, and G. Carneiro, "Perturbed and strict mean teachers for semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4258–4267.

[388] J. Kim, J. Jang, H. Park, and S. Jeong, "Structured consistency loss for semi-supervised semantic segmentation," *arXiv preprint arXiv:2001.04647*, 2020.

[389] Y. Zou, Z. Zhang, H. Zhang, C.-L. Li, X. Bian, J.-B. Huang, and T. Pfister, "Pseudoseg: Designing pseudo labels for semantic segmentation," *arXiv preprint arXiv:2010.09713*, 2020.

[390] R. Ito, K. Nakae, J. Hata, H. Okano, and S. Ishii, "Semi-supervised deep learning of brain tissue segmentation," *Neural Networks*, vol. 116, pp. 25–34, 2019.

[391] Z. Xu and M. Niethammer, "Deepatlas: Joint semi-supervised learning of image registration and segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer, 2019, pp. 420–429.

[392] Z. Li, Z. Li, R. Liu, Z. Luo, and X. Fan, "Coupling deep deformable registration with contextual refinement for semi-supervised medical image segmentation," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–5.

[393] M. Modat, D. M. Cash, P. Daga, G. P. Winston, J. S. Duncan, and S. Ourselin, "Global image registration using a symmetric block-matching approach," *Journal of medical imaging*, vol. 1, no. 2, pp. 024 003–024 003, 2014.