CULTURAL CARE WARNING NOTICE

Cultural advice
This item may contain culturally sensitive information.

All users are advised that this item may contain images, voices and/or names of people who have died.

Indigenous Cultural and Intellectual Property
This item may contain Indigenous Cultural and Intellectual Property. Please consult with the relevant communities if you wish to use any of the content in this item.

Copyright
This item is protected by copyright, and must be used in accordance with the

provisions of the Copyright Act 1968 (Cth).

Deep Learning-based Radiomics Framework for Multi-Modality PET-CT Images

THE UNIVERSITY OF SYDNEY

Faculty of Engineering

The University of Sydney

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Yige Peng

March 2023

© Copyright by Yige Peng 2023

All Rights Reserved

Statement of Originality

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Yige Peng

Date: 31/03/23

Authorship Attribution Statement

Chapter 3 of this thesis is based on a publication currently in preparation to submit as Y. Peng, L. Bi, D. Feng, and J. Kim, 'Multi-Modal PET/CT Tumor Segmentation via Self-supervised False Positive and False Negative Reduction Network', *IEEE J. Biomed. Health Inform. (JBHI)*, 2023. I designed the algorithm, conducted the experiments, analysed the data and drafted the manuscript.

Chapter 4 of this is based on the publications Y. Peng, L. Bi, Y. Guo, D. Feng, M. Fulham, and J. Kim, 'Deep multi-modality collaborative learning for distant metastases predication in PET-CT soft-tissue sarcoma studies', in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2019, pp. 3658–3688, and as Y. Peng, L. Bi, A. Kumar, M. Fulham, D. Feng, and J. Kim, 'Predicting distant metastases in soft-tissue sarcomas from PET-CT scans using constrained hierarchical multi-modality feature learning', *Phys. Med. Biol.*, vol. 66, no. 24, p. 245004, 2021. I designed the algorithm, conducted the experiments, analysed the data and drafted the manuscript.

Chapter 5 of this thesis is based on the publication Y. Peng, L. Bi, M. Fulham, D. Feng, and J. Kim, 'Multi-modality information fusion for radiomics-based neural architecture search', in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VII 23*, Springer, 2020, pp. 763–771. I designed the entire algorithm, conducted all of the experiments,

analysed the data and wrote the drafts of the manuscript.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Yige Peng Signature: Date: 31/03/23

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Professor Jinman Kim Signature: Date: 31/03/23

Abstract

Multimodal positron emission tomography - computed tomography (PET-CT) imaging is widely regarded as the imaging modality of choice for cancer management. This is because PET-CT combines the high sensitivity of PET in detecting regions of abnormal functions and the specificity of CT in depicting the underlying anatomy of where the abnormal functions are occurring.

Radiomics is an emerging research field that enables the extraction and analysis of quantitative features from medical images, providing valuable insights into the underlying pathophysiology that cannot be discerned by the naked eyes. This information is capable of assisting decision-making in clinical practice, leading to better personalised treatment planning, patient outcome prediction, and therapy response assessment.

The aim of this thesis is to propose a new deep learning-based radiomics framework for multimodal PET-CT images. The proposed framework comprises of three methods: 1) a tumour segmentation method via a self-supervision enabled false positive and false negative reduction network; 2) a constrained hierarchical multi-modality feature learning is constructed to predict the patient outcome with multimodal PET-CT images; 3) an automatic neural architecture search method to automatically find the optimal network architecture for both patient outcome prediction and tumour segmentation.

Extensive experiments have been conducted on three datasets, including one public soft-tissue sarcomas dataset, one public challenge dataset, and one in-house lung cancer data. The results demonstrated that the proposed methods obtained better performance in all tasks when compared to the state-of-the-art methods.

Acknowledgements

I would like to acknowledge that my PhD journey was both the most rewarding and most challenging period of my thirty years. It was a life-changing experience during which I had the opportunity to work with inspiring supervisors, fellow colleagues, and friends.

Foremost, I express my immense gratitude to my supervisor, Professor, David Dagan Feng, for his unwavering support, guidance, and encouragement throughout my PhD study. His insights and vision have played an instrumental role in shaping both my research work and my life values. His profound and insightful life lessons have taught me how to prepare for future challenges. It is my great honour and pleasure to have received his supervision during my PhD.

I am also sincerely grateful to my supervisor, Professor, Jinman Kim, for his invaluable inspiration, constructive feedback, and incredible patience. Without his comprehensive support and patient guidance, I would not have been able to work on such interesting research projects and submit this thesis. He has set the benchmark for me in both work and life, with his focused work attitude and rigorous and conscientious style.

I would like to deeply thank my supervisor, Associate Professor, Lei Bi, for his steadfast backing, invaluable suggestions, and support in every possible way. His thorough mentorship helped me gain a comprehensive understanding of the research process, from conceptualisation to implementation and analysis. I am also extremely grateful to Lei Bi for answering all my trivial questions.

I express my special appreciation to, Professor, Michael Fulham at the Department of

Molecular Imaging, Royal Prince Alfred (RPA) hospital for his invaluable feedback and clinical insights on my research projects.

I would also like to extend my appreciation to Dr Ashnil Kumar, Dr Yuyu Guo, Dr Euijoon Ahn, Dr Younhyun Jung, Tian Xia, Hoijoon Jung, Dr Ke Yan, Robin Huang, Dr Xiaohang Fu, Mingjian Li, Ge Jin, Zimo Huang, Mingyuan Meng, Shijia Zhou, Yuxin Xue, and other former and current colleagues in the Biomedical Data Analysis and Visualisation (BDAV) research group. Thanks for their assistance and advice throughout my PhD study.

Last but not the least, I cannot express my gratitude enough to my parents for their love, support and understanding, and for standing by me during all the ups and downs of my PhD journey. Their faith in me and their belief in my abilities has been my constant source of strength and motivation. I would also like to thank my relatives and friends, who have always been a major source of help whenever I needed it.

Thank you all for being part of this incredible journey and for making it a memorable one.

List of Publications

The following publications were produced during the course of my PhD candidature. Most of these publications were based on the work that is presented in this thesis, including 2 first author journals and 2 first author conference papers. Publications marked with * were either directly or have major contribution to this thesis. Other publications were related research, where I have made contributions.

Published or Accepted:

1. Y. Xue, **Y. Peng**, L. Bi, D. Feng, J. Kim, 'CG-3DSRGAN: A classification guided 3D generative adversarial network for image quality recovery from low-dose PET images', *(EMBC)*, IEEE, 2023,
2. M. Meng, **Y. Peng**, L. Bi, and J. Kim, 'Multi-task deep learning for joint tumor segmentation and outcome prediction in head and neck cancer', in *Head and Neck Tumor Segmentation and Outcome Prediction: Second Challenge, HECKTOR 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings*, Springer, 2022, pp. 160–167.
3. ***Y. Peng**, L. Bi, A. Kumar, M. Fulham, D. Feng, and J. Kim, 'Predicting distant metastases in soft-tissue sarcomas from PET-CT scans using constrained hierarchical multi-modality feature learning', *Phys. Med. Biol.*, vol. 66, no. 24, p. 245004, 2021.
4. ***Y. Peng**, L. Bi, M. Fulham, D. Feng, and J. Kim, 'Multi-modality information fusion for radiomics-based neural architecture search', in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru,*

October 4–8, 2020, Proceedings, Part VII 23, Springer, 2020, pp. 763–771 **(Student Travel Award)**.

5. ***Y. Peng**, L. Bi, Y. Guo, D. Feng, M. Fulham, and J. Kim, 'Deep multi-modality collaborative learning for distant metastases predication in PET-CT soft-tissue sarcoma studies', in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2019, pp. 3658–3688 **(Oral)**.

Under Preparation, Review or Revision:

6. ***Y. Peng**, J. Kim, D. Feng, and L. Bi, 'Multi-Modal PET-CT Tumor Segmentation via Self-supervised Enabled False Positive and False Negative Reduction Network', *Pattern Recognition.*, 2023 (Under Review).
7. Y. Xue, L. Bi, **Y. Peng**, M. Fulham, D. Feng, J. Kim, 'PET Synthesis via Self-supervised Adaptive Residual Estimation Generative Adversarial Network', *IEEE Transactions on Radiation and Plasma Medical Sciences*, 2023 (Major Revision).

Table of Contents

List of Figures

List of Tables

List of Abbreviations

PET-CT	-	Positron Emission Tomography - Computed Tomography
MRI	-	Magnetic Resonance Imaging
US	-	Ultrasound
FDG	-	^{18}F-Fluorodeoxyglucose
STSs	-	Soft Tissue Sarcomas
RPA	-	Royal Prince Alfred
BDAV	-	Biomedical Data Analysis and Visualisation
ROIs	-	Regions of Interest
SVM	-	Support Vector Machines
CNN	-	Convolutional Neural Network
NAS	-	Neural Architecture Search
SFRN	-	Self-supervision enabled False positive and false negative Reduction Network
CHMFL	-	Constrained Hierarchical Multimodal Feature Learning
HMFL	-	Hierarchical Multimodal Feature Learning
CFL	-	Constrained Feature Learning
MM-NAS	-	Multimodal Neural Architecture Search
HU	-	Hounsfield Unit
SNR	-	Signal-to-Noise Ratio

SUV	-	Standard Uptake Value
MRF	-	Markov Random Field
MSAM	-	Multimodal Spatial Attention Module
DM	-	Distant Metastasis
NSCLC	-	Non-Small Cell Lung Cancer
KNN	-	K-Nearest Neighbour
RNN	-	Recurrent Neural Network
GBM	-	Glioblastoma Multiforme
LASSO	-	Least Absolute Shrinkage and Selection Operator
IDH1	-	Isocitrate DeHydrogenase 1
LSTM	-	Long-Short-Term-Memory
SSL	-	Self-Supervised Learning
GSM	-	Global Segmentation Module
LRM	-	Local Refinement Module
T2FS	-	T2-weighted with fat-suppression
AdamW	-	Adaptive-moment estimation with decoupled Weight decay
MosNet	-	Modality-specific segmentation network
Acc	-	accuracy
ROC	-	receiver operating characteristic

AUC	-	Area Under the ROC Curve
Pre.	-	precision
Sen.	-	sensitivity
Spe.	-	specificity
IoU	-	intersection over union
GT	-	ground truth
FP	-	False Positive
FN	-	False Negative
FPR	-	False Positive Rate
FNR	-	False Negative Rate
Adam	-	Adaptive-moment-estimation
HC	-	Hand-Crafted
GLCM	-	Grey-Level Co-occurrence Matrix
GLRLM	-	Grey-Level Run-Length Matrix
GLSZM	-	Grey-Level Size Zone Matrix
NGTDM	-	Neighbourhood Grey-Tone Difference Matrix
RF	-	Random Forest
MOR	-	Many-Objective Radiomics
t-SNE	-	t-distributed stochastic neighbourhood embedding

Chapter 1. Introduction

1.1 Background

Medical imaging plays an important role in modern healthcare and is indispensable to numerous clinical applications, such as disease diagnosis, surgical planning, and therapeutic procedures evaluation [1]. The wide range of medical imaging modalities includes digital radiography, magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography combined (PET), X-ray, Ultrasound (US), as well as combined imaging such as PET-CT and PET-MRI; these modalities provide anatomical and functional information about human body's structure and physiology. Among these modalities, the multimodal PET-CT, using radiopharmaceutical ^{18}F-Fluorodeoxyglucose (FDG) PET, is widely considered as the imaging modality of choice for the diagnosis, staging, and evaluation of treatment response in many cancers, including lung cancer, lymphoma, and soft-tissue sarcomas (STSs) [2]. This is attributed to the fact that PET-CT combines the high sensitivity of PET in detecting regions of abnormal functions and the specificity of CT in depicting the underlying anatomy of where the abnormal functions are occurring. With PET, sites of the disease usually display greater FDG uptake (glucose metabolism) than normal structures. The spatial extent of the disease within a particular structure, however, cannot be accurately

determined due to tumour heterogeneity, partial volume effect, and the inherent low resolution of PET, especially when compared to CT and MRI [3]. Complementarily, CT provides the anatomical localization of sites of abnormal FDG uptake in PET as an aid in image interpretation [4].

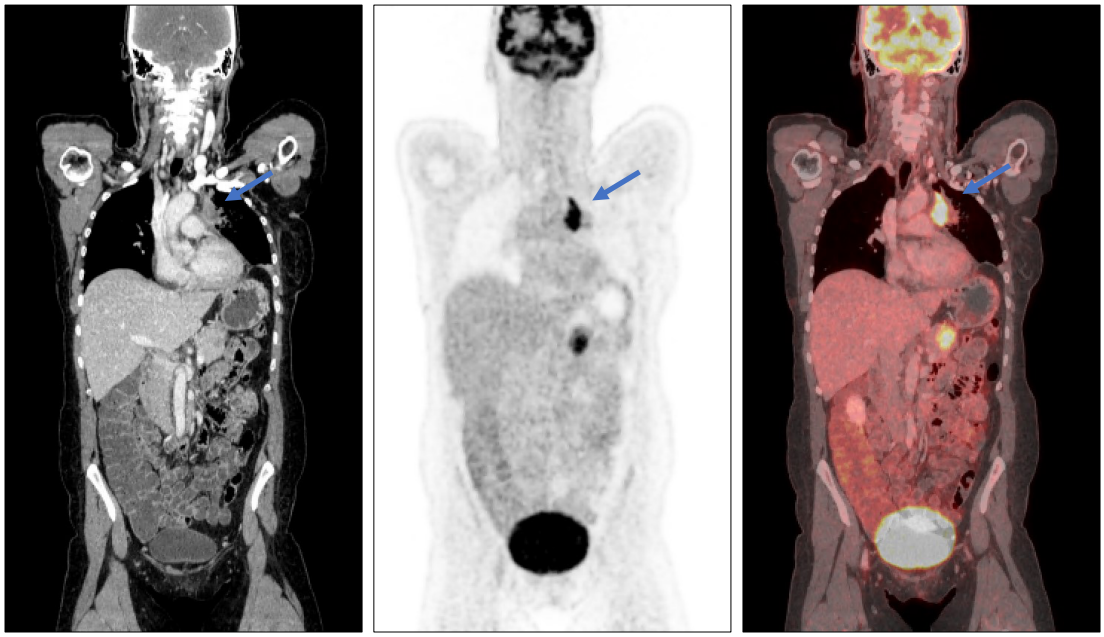


Figure 1.1. One example of lung cancer diagnosed with different imaging modalities. The three images from left to right are CT, PET and fused PET-CT images. These images are commonly used in cancer assessment. The blue arrows point to the region of lung cancer. These PET-CT images are from the public challenge dataset AutoPET [5].

Most cancers/tumours are not homogeneous, instead, they are generally made up of multiple clonal subpopulations of cancer cells. In solid cancerous tumours such as STSs and lung cancers, the extent of heterogeneous characteristics is expressed at multiple levels, such as genes, proteins and anatomical landmarks within tumours, and they exhibit considerable spatial and temporal variations that could potentially provide valuable information about

tumour aggressiveness for patient risk assessment [6]. However, studying tumour heterogeneity using histopathological samples from biopsies is very difficult, it is an invasive and time-consuming procedure that involves tumour tissue samples collection using biopsy needles or during surgical resection, and tissue preparation and staining for histopathological examination under a microscope. In addition, the information obtained may vary depending on which part of a tumour is sampled, which may not be representative of the entire tumour and can miss important areas of heterogeneity [7]. Therefore, the treatment planning, therapy response assessment and other prognosis based on histopathological samples may not fully account for the complexity and diversity of tumours. This challenge is addressed by an emerging translational research field, "radiomics". The underlying hypothesis of radiomics is that the heterogeneous characteristics of tumours could be translated into heterogeneous lesion metabolism and anatomy in medical images, such that metabolic and anatomical information can be derived and captured via imaging feature extraction and analysis. Unlike biopsies, radiomics can capture the heterogeneity information of the entire tumour volume, providing a more comprehensive picture of the tumour and its response to treatment, and other prognostic information. Thus, radiomics allows for better personalised treatment planning, patient outcome prediction, and therapy response assessment [8].

Radiomics works by extracting and analysing innumerable features from medical images. Through quantitative feature extraction and machine learning modelling, radiomics is capable of leveraging much more information from medical images than what can be discerned by the naked eye, and such information can provide valuable insights into the underlying biological and physiological processes of heterogeneous tumours so as to assist decision-making in clinical practice [9]. Radiomics not only visually interpret these images, but also capture information that is relevant to diagnosis and prognosis [10]. With the predictive power for personalised treatment planning and patient outcomes prediction (e.g.,

distant metastases prediction, survival analysis), radiomics exhibits promising potential in cancer-related research [11]. Conventional radiomics methods consist of four main steps (see Figure 1.2): (i) image acquisition/reconstruction, (ii) manual segmentation and annotation of regions of interest (ROIs, e.g., tumours), (iii) extracting hand-crafted radiomics features (e.g., intensity, texture, shape) and, (iv) building predictive models, such as support vector machines (SVM), regression models, etc., to correlate extracted features with the clinical outcomes (e.g., distant metastases, overall survival, etc). Although conventional radiomics has been used in many different applications with various medical images [6], [7], [12]–[14], their performance relies on a prior skillset in accurate tumour delineation, radiomics features definition and extraction, and manual tuning of numerous predictive model parameters. As a result, these studies could introduce human bias and lack the capacity to comprehend high-level semantic information.

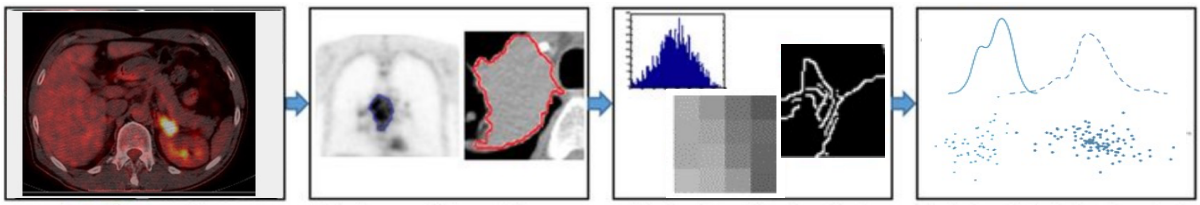


Figure 1.2. The main steps of conventional radiomics. (a) image acquisition / reconstruction, (b) region of interest segmentation, (c) radiomic features extraction, and (d) statistical analysis and modelling. The PET-CT images are from the public challenge dataset AutoPET [5].

The increase in radiomics performance is closely associated with recent advances in deep learning that have enabled data-driven approaches for automated image feature quantification. The convolutional neural network (CNN) is one of the most widely used techniques in medical image analysis and has inspired a shift from conventional radiomics

analysis toward CNN-based radiomics analysis. This is attributed to the CNN's ability to extract high-level semantic information in an end-to-end manner that is meaningful to the specific task (e.g., tumour or organ segmentation, patient outcome prediction, etc.), which reduces the need for prior knowledge in hand-crafted radiomics features definition and other manual input [15], [16]. Several studies have shown that deep radiomics features extracted from CNN achieved promising performance in outcome prediction for head-and-neck cancer patients, neoadjuvant chemoradiation response prediction for locally advanced rectal cancer patients [17], preoperative meningiomas grading [18], and glioblastoma multiforme survival prediction [19]. However, these approaches were designed for single-modality medical images, and they fail to fully exploit the potential of CNN to capture complementary information from multiple combined imaging modalities. Although there is a limited number of CNN-based radiomics studies where the multimodal PET-CT images were utilized, they are hybrid methods combining conventional radiomics components and deep learning-based radiomics [20]–[22], and their focus is to utilise both conventional radiomics features and deep radiomics features extracted from CNNs for clinical applications, rather than advancing the technical usage of CNN for radiomics. As a result, these methods suffer from similar limitations of conventional radiomics methods in prior knowledge about feature definition and extraction, and these methods fail to fully utilise the potential of CNNs in capturing complementary information from multimodal medical images.

Another challenge in the field of radiomics is the scarcity of manual annotation of ROIs, especially for multimodal PET-CT images, which is an important prerequisite to support patient outcome prediction in radiomics [23]. Conventional radiomics studies rely on tumour annotations to quantify imaging features for specific diseases. Similarly, most of the existing CNN-based radiomics models utilised the annotated primary tumour regions as the input to reduce interference from non-relevant background information. However, tumour annotation

in PET-CT images remains a challenging task due to the large cost and complicated acquisition procedures. Specifically, one of the main difficulties in using PET images is determining the spatial extent of the tumour due to the lower resolution and signal-to-noise ratio of PET in comparison to CT. Furthermore, increased FDG uptake can also be observed in normal structures, such as the heart, and kidneys, and sites of inflammation (e.g., pneumonitis) [24], which can lead to over-segmentation of tumour region with false positive segmentation errors. Additionally, some well-differentiated tumours may not exhibit high FDG uptake, making them difficult to detect in PET images alone. Moreover, the task of automated PET-CT tumour segmentation requires consideration of complementary features from both modalities. The optimal extraction and utilization of data from multimodal PET-CT images in radiomics have not been fully explored in comparison to unimodal imaging problems [25].

Although CNN-based methods have become state-of-the-art in radiomics, existing CNNs for patient outcome prediction and its important prerequisite automated tumour segmentation are still heavily dependent on human expertise to design dataset-specific CNN architectures, such as the number of convolutional layers and the structure of convolutional blocks. CNN architecture design and optimisation necessitate a large amount of domain knowledge, such as in validating the architecture performance and tuning the hyperparameters. To address this limitation, neural architecture search (NAS) has recently been proposed to simplify the challenges in architecture design by automatically searching for an optimal network architecture based on a given dataset. Hence, NAS is capable of minimising manual input and reliance on prior knowledge in architecture design and network finetuning for different tasks and datasets [26]. While a small number of investigators have attempted to apply NAS for medical image analysis with single modality imaging data [27], [28], the usage of NAS regarding multimodal medical images adds a higher level of

complexity and has not yet been explored, and the task complexity is further compounded when multiple imaging modalities (e.g., PET-CT), are included in the analysis.

1.2 Contributions of this Thesis

The overall aim of this thesis is to introduce a new deep learning enabled radiomics framework for multimodal PET-CT images. In this thesis, new CNNs have been designed and implemented to automatically learn the complementary characteristics from different image modalities for two important tasks in deep learning-based radiomics: (1) predictive model construction, and (2) tumour segmentation to support the patient outcome prediction. When compared with existing studies, the following are the innovative methods and contributions made in the thesis:

1) A new Self-supervision enabled False positive and false negative Reduction Network (SFRN) is proposed for tumour segmentation in multimodal PET-CT images. Initially, a global segmentation module was employed to coarsely delineate candidate tumour regions. Then, the candidate tumour regions were refined at the pixel level by removing both false positives and false negatives using a local refinement module. A classification branch is further incorporated to enhance the ability of our network to distinguish the tumour regions from healthy regions in multimodal PET-CT images, thus further improving the ability to alleviate the two types of segmentation errors introduced by the global segmentation module. The SFRN outperformed the state-of-the-art segmentation methods on two multimodal PET-CT datasets (one public STS dataset and one in-house lung cancer data).

2) A new constrained hierarchical multimodal feature learning method, hereby denoted as CHMFL, is proposed for patient outcome prediction in radiomics with multimodal

PET-CT images. In this method, a constrained feature learning (CFL) module was used to spatially guide the learning process to focus on the important semantic regions (e.g., tumours). The formulation of this module means that it can target the functional hot spots with high FDG uptake in PET within the anatomical context of CT. A hierarchical multimodal feature learning (HMFL) module is also designed to derive optimal radiomics features by integrating complementary features across modalities at different scales. The formulation of HMFL combined multimodal features from different scales in an iterative manner and enabled a more comprehensive and flexible fusion of PET and CT features, e.g., low-level PET texture features from a shallow layer with semantic CT features from a deeper layer. The proposed method was evaluated in predicting the development of distant metastases on a well-established benchmark STSs PET-CT dataset, and the experimental results showed that our method achieved overall better performance when compared to the state-of-the-art methods.

3) A new Multimodal NAS (MM-NAS) method is proposed to search for a multimodal CNN architecture for use in PET-CT radiomics studies. The proposed method was able to find various fusion modules e.g., fusion via different network operations (e.g., convolution, pooling, etc.) at different stages of the network. These searched fusion modules provided more flexible options for integrating the complementary PET and CT data without prior knowledge in architecture design and human input for hyperparameter tuning. The proposed method was capable of building an optimal, fully automated radiomics CNN architecture and enabled an optimal fusion of multimodal PET-CT images for radiomics. The proposed MM-NAS was evaluated for its ability to predict distant metastases of STSs as well as to segment STSs and lung cancer. The experimental results showed that the MM-NAS obtained overall

better performance when compared to the state-of-the-art methods.

1.3 Thesis Organizations

The rest of this thesis is organized as follows. The related works of radiomics are discussed in Chapter 2. The proposed methods are described in the following three chapters. A self-supervision enabled false positive and false negative reduction network is introduced in Chapter 3. The constrained hierarchical multimodal feature learning method is introduced in Chapter 4. Chapter 5 presents the multimodal NAS method of automatically searching for a multimodal CNN architecture in the PET-CT radiomics framework. Finally, the contributions of this thesis are summarised and the directions for future work are presented in Chapter 6.

Chapter 2. Related Works

In this chapter, the related works relevant to this thesis are discussed. This chapter begins with the introduction of various medical imaging modalities that are used in this thesis. This is followed by different radiomics methods and corresponding challenges in this research field. The gaps in existing studies will also be summarised. In the methodology Chapters 3-5, where appropriate, specific related works will be highlighted, by referring to Chapter 2.

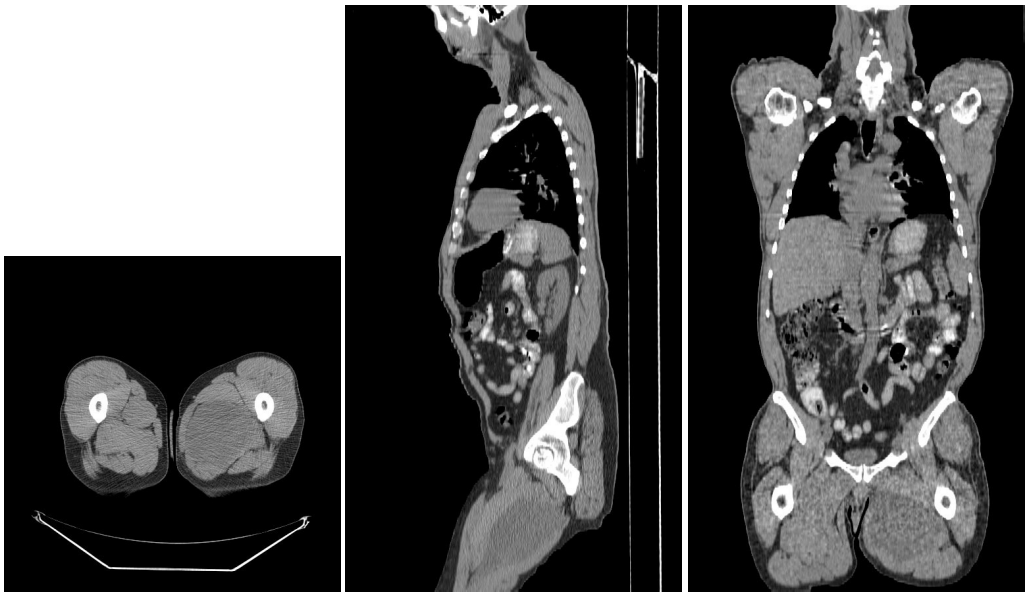
2.1 Medical Imaging

A variety of different medical imaging modalities are routinely used as part of clinical patient management, including digital radiography, magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography combined (PET), X-ray, Ultrasound (US), as well as combined imaging such as PET-CT and PET-MRI. These modalities provide different insights into the condition of diseases, such as anatomical and functional information about human body's structure and physiology. In this thesis, the imaging data of CT, PET, and multimodal PET-CT are mainly used, which are further explained as followed.

2.1.1 Computed Tomography

Computed tomography (CT) is a medical imaging technique that utilises a computerised x-ray imaging procedure to generate detailed images of all parts of the body including the bones, muscles, fat, organs and blood vessels. This technique employs a series of narrow beams of x-rays that pass through the body from different angles and are detected by highly sensitive detectors [29]. Then the acquired 2D slices or tomograms are combined to form a 3D CT imaging volume. CT allows physicians to visualise internal organs in a non-invasive manner. Unlike traditional x-ray imaging, the images produced by CT scanners do not superimpose structures on each other. CT scanners are also capable of capturing images with high spatial resolutions, potentially less than 1mm per dimension.

Three slices from different planes of the same CT volume with soft-tissue sarcoma (STS) are shown in Figure 2.1. The differences in voxel intensities can be clearly seen by the high-intensity values of the bones (in white) in all slices, the low intensity of the air within the lungs in (b) and (c) of Figure 2.1, and the intensity of the soft tissues (STS, liver, etc.) in all slices.



(a) Axial View (b) Sagittal View (c) Coronal View

Figure 2.1. Slices from different view planes of the same CT volume. These CT images are from the public challenge dataset AutoPET [5].

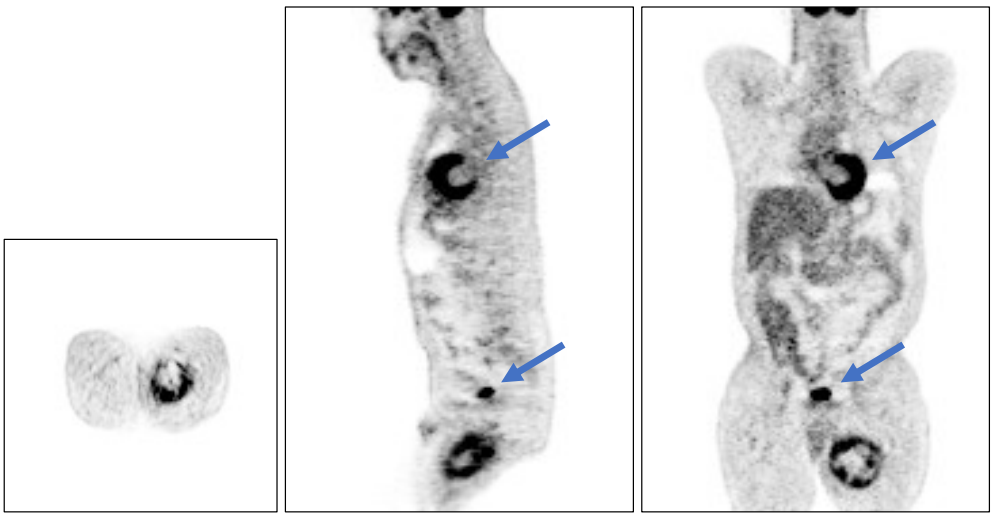
As detailed parts of the body including the bones, muscles, fat, organs and blood vessels can be generated in the CT images via a computerised x-ray imaging procedure with high spatial resolutions, CT images are primarily utilised in the identification and examination of anatomical conditions in clinical practice, such as the assessment of airway obstruction in patients with obstructive sleep apnoea [30] or the evaluation of emphysema [31]. Additionally, CT images are useful for measuring tumour growth [32] and the determination of lung nodule growth rates [33]. Unfortunately, benign and malignant tumours typically present comparable Hounsfield Unit (HU) values with the soft tissues on CT images. Therefore, the detection and visualization of tumours utilising solely CT images present a significant challenge, and a different procedure, such as biopsy or alternative imaging modalities such as the positron emission tomography (PET, refer to Section 2.1.2) is usually required.

2.1.2 Positron Emission Tomography

Positron emission tomography (PET) is a nuclear medicine imaging technique that produces functional images by detecting the gamma rays emitted from a positron-emitting radiotracer that has been introduced into the subject's body [34]. This technique constructs greyscale volumetric images of the target tissue or organ, providing valuable insights into metabolic activity, blood flow, and other physiological processes. In this thesis, only studies with PET images using ^{18}F-Fluorodeoxyglucose (FDG) as the radiotracer are investigated, which is widely used for cancer patients [35].

Voxel intensity in an FDG-PET image indicates glucose metabolism at the corresponding location in the body, and this is instrumental in characterising the nature of lesions, where malignant tumours have abnormally high intensities [36]. Hence, FDG-PET images exhibit a diagnostic and prognostic accuracy between 80-90%, and they outperform anatomical imaging

techniques like CT in detecting malignant cancers [37]. Nonetheless, FDG-PET images are in lower spatial resolution and lower signal-to-noise ratio (SNR) when compared to modalities like CT [38]. As a result, noise and limited anatomical information in PET hinders the precise localisation of lesions [37].



(a) Axial View (b) Sagittal View (c) Coronal View

Figure 2.2. Slices from different planes of the same PET volume. Apart from the STS region, all other high-intensity regions correspond to areas with naturally occurring high glucose metabolism regions, e.g., the heart (see the top middle within (b) and (c)), and bladder (see the small black region above STS within (b) and (c)). These PET images are from the public challenge dataset AutoPET [5].

Figure 2.2 shows three slices from different planes of the same FDG-PET volume with an STS patient. The black region of high intensity in the right leg is the primary STS tumour which can be seen in all three slices. All other high-intensity regions correspond to areas with naturally occurring high glucose metabolism regions, e.g., the heart (see the top middle within (b) and (c) in Figure 2.2), and bladder (see the small black region above STS within (b) and (c) in Figure 2.2), the normal regions are pointed by the blue arrows in Figure 2.2.

Moreover, it is important to note that the voxel values in PET images do not inherently correspond to any physical characteristics, unlike CT voxel values, which relates to the X-ray absorption of different materials, PET voxel values are derived from the measurement of the rate of positron annihilation events within a voxel, which is influenced by factors, such as the radioactive decay of the injected tracer FDG, the tracer distribution in the body, and the scanner characteristics. As a result, this makes the comparison of PET values across different scans problematic, even if the scans are of the same patient [39]. The standard uptake value (SUV) compensates for this limitation by providing a measure to evaluate the uptake of PET radiotracers via normalising the original voxel intensities based on the subject's body measurements information, such as mass or weight, and PET acquisition parameters like dose and time. Furthermore, regions with an SUV value higher than a specified value (called the 'threshold') are identified as regions of interest (ROIs) which are often likely to be tumours [40]. SUV thresholding of PET images is one of the fundamental techniques to detect abnormal FDG uptake sites before and after treatment [41].

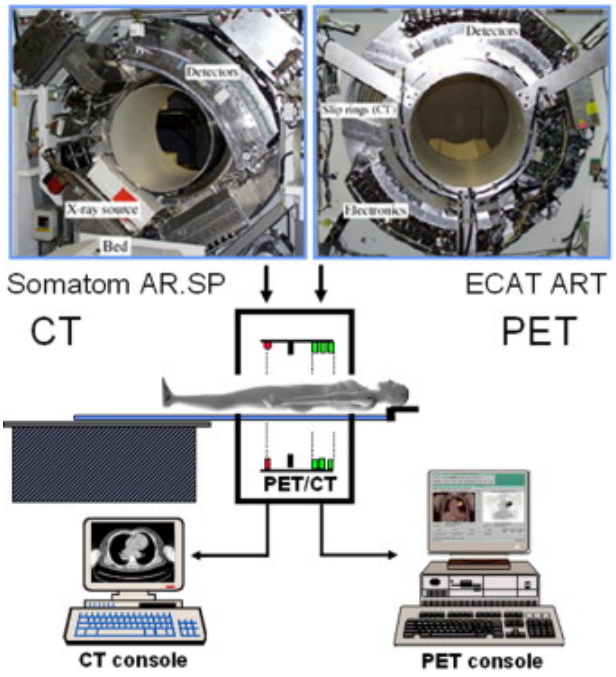
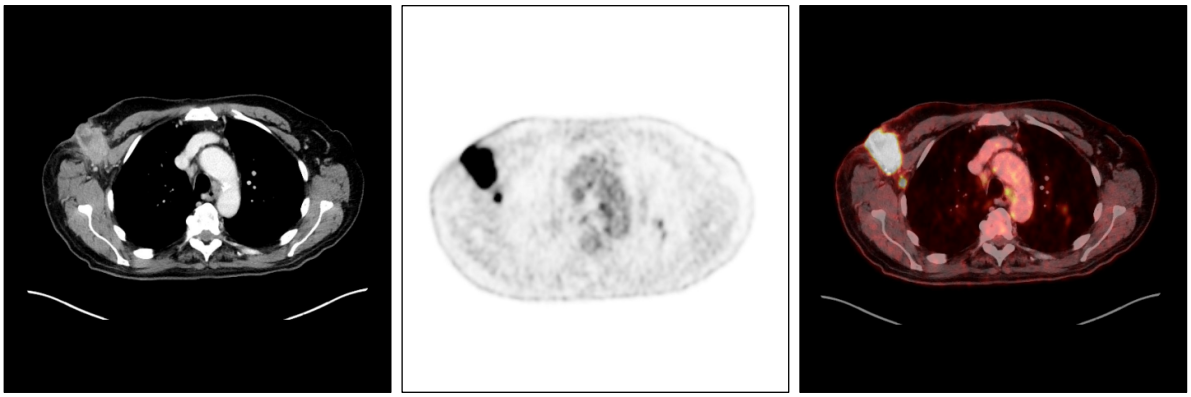


Figure 2.3. The first combined PET-CT prototype at the University of Pittsburgh [42].

2.1.3 Multimodal PET-CT Imaging

Both PET and CT scans are important for diagnosis and follow-up in clinical oncology [43]. While these scans can be carried out on the same device, local misregistration is observed when simply integrating PET and CT imaging [44]. To address this concern and enhance the overall image quality, the University of Pittsburgh in 1998 has developed the first combined dual-modality PET-CT to align two image sets acquired independently (Figure 2.3) [42].

Multimodal PET-CT refers to the sequential acquisition of CT and PET volumes in the same scanner during the same imaging session [45], [46]. The respective volumes acquired by the scanner have different pixel, contrast, and spatial resolutions. Figure 2.4 shows the axial images acquired from a combined PET-CT scanner. Figures 2.4(a) and 2.4(b) are the CT and PET images, respectively. Figure 2.4(c) depicts the fused image produced after applying scanner parameters to transform them to the same coordinate space.



(a) CT (b) PET (c) Fused PET-CT

Figure 2.4. A Multimodal PET-CT image. These PET-CT images are from the public challenge dataset AutoPET [5].

Combined PET-CT scanners offer numerous advantages over their single modality

counterparts. The total time for image acquisition is significantly shorter, allowing better instrument utilisation and a higher patient throughput [47]. Furthermore, studies have shown that PET-CT is more sensitive than either PET or CT conducted alone, and that combining CT with PET improves the sensitivity of PET images [48]. Overall, PET-CT provides improved tumour diagnosis, localisation, and staging, compared to single modality PET or CT [47], [48]. The clinical importance of PET-CT and the increasing trend of PET adoption indicate that in the very near future, all PET studies will be in the form of PET-CT images [37], [49].

The value of PET-CT arises from its ability to present complementary anatomical (from CT) and functional (from PET) information. This is accomplished through the scanner's ability to merge the spatial co-alignment of the two modalities, thereby allowing physicians to observe the relationship between the anatomical and functional details. Such integration can reveal important insights, such as determining whether a lung tumour has invaded an adjacent structure. Consequently, an essential challenge for researchers in the field of PET-CT image processing is to effectively leverage and maximize the benefits found in these complementary features and spatial relationships.

2.2 Radiomics

Radiomics has emerged as a translational research field in response to the increasing availability of medical imaging data and the growing need for personalised, data-driven approaches to modern healthcare. Radiomics refers to the extraction and analysis of innumerable features from non-invasive medical images. Through quantitative feature extraction and predictive model construction, radiomics is capable of capturing and analysing additional information from medical images than what can be discerned by the naked eye. Such information can provide valuable insights into the underlying biological and physiological processes of heterogeneous diseases so as to assist diagnosis (e.g., tumour

grading), and prognosis (e.g., distant metastases prediction, survival analysis) in various cancers, allowing for more personalised and precise patient care. As the field continues to evolve, it is likely that radiomics will play an increasingly important role in integrating image-derived information to assist clinical decision-making.

The existing radiomics studies can be categorised into two main categories: (1) conventional radiomics methods – where the statistical machine learning methods (e.g., random forest, support vector machine, etc.) are utilised to associate patient outcomes (e.g., distant metastases, overall survival, etc.) with the hand-crafted features (e.g., intensity, texture, shape, etc.) that are extracted from the regions of interest (ROI, e.g., tumours) in medical images; (2) deep learning-based radiomics methods – where a deep neural network is adopted as the predictive model to directly predict the patient outcomes from the medical images in an end-to-end manner. Moreover, the annotation or segmentation of the ROIs is an important prerequisite to support the model construction and outcome prediction. The following Chapters 3-5 includes further technical details related to the radiomics literature.

2.3 Automated Medical Image Segmentation Methods

The automated segmentation in radiomics studies with multimodal PET-CT images have focused on the delineation of ROIs, which are usually the tumour regions. Various strategies for automatic tumour segmentation in PET-CT images have been proposed. Overall, they can be categorized into two main types of traditional methods and deep learning counterparts.

2.3.1 Traditional Segmentation Methods

Thresholding is a classical traditional method used to distinguish tumours from the background based on differences in standardized uptake value (SUV) [50]–[52]. The selection of an appropriate SUV threshold is crucial in clinical practice, and various thresholds have been

employed, ranging from an SUV of >2.5, to 41%-90% of the maximum SUV value in the tumour to identify a region of interest [53]. However, the accuracy of thresholding can be compromised by normal physiological processes and benign conditions, such as pneumonia, exhibiting high FDG uptake while on contrary, some primary tumours may have SUV that is less than 2.5. Such SUV variations can lead to both false negative and positive segmentation errors. Furthermore, several factors can affect the SUV, such as the type of scanner, the time between the FDG injection and data acquisition, the image reconstruction method, the calculation of the SUV by the scanner vendor, image noise, etc. [54]. Therefore, the selection of a suitable threshold requires specialised domain knowledge of PET-CT imaging [53]. In recent years, thresholding-based methods have been generally replaced by machine learning counterparts [50]. Various machine learning strategies for tumour segmentation have been explored, including the fusion of modality-specific features or complementary information from PET and CT, such as graph-based methods [4], [55]–[59]. For instance, Bagci et al. proposed a random walk method for co-segmentation of multiple objects in PET, PET-CT, PET- magnetic resonance imaging (MRI), and fused PET-MRI-CT images via a hyper-graph [4]. Similarly, Han et al. formulated the tumour segmentation problem as a graph-based Markov Random Field (MRF) with an energy function that leveraged the advantageous characteristics of each modality and penalized the segmentation difference between PET and CT images [57]. Furthermore, some researchers used one modality to guide tumour localization in another modality. Wojak et al. proposed a joint variational segmentation method using PET intensities to provide local constraints to adjust the segmentations on CT [60]. Bagci et al. proposed a random walk segmentation method that employs FDG uptake value thresholds in PET to automatically initialize foreground and background seeds, and then found corresponding boundaries in the CT image [56]. However, these methods that utilize PET only to drive segmentation are highly dependent on the PET SUVs, hence they are inherently limited

in the presence of normal high-uptake activity, which can result in false positive segmentation results.

2.3.2 Deep Learning-based Segmentation Methods

In recent years, deep learning methods based on CNNs have made great progress in automatic medical image analysis, and supervised CNN-based methods are regarded as state-of-the-art in PET-CT tumour segmentation. The success of supervised deep learning methods is mainly attributed to their ability to automatically extract features from images that are optimized for individual tasks with ground truth [25]. Guo et al. investigated the performance of CNNs on multimodal medical images, including PET-CT by using multiple CNNs, one for each imaging modality, then assembling the CNNs output results to produce the tumour segmentation results [61]. Jin et al. proposed a two-stream chained network for gross tumour volume segmentation in PET-CT images, utilising early fusion and late fusion to fuse the complementary information from the two modalities [62]. Kumar et al. proposed a co-learning approach to segment multiple objects with PET-CT images, inclusive of tumours and normal organs, where a two-branched U-Net was implemented to extract and fuse PET and CT features across multiple CNN blocks [63]. Li et al. proposed a variational method that used a 3D fully convolutional network to generate a tumour probability map from the CT images and integrated it with the PET images to segment the tumours via a fuzzy variational model [64]. Bi et al. proposed a recurrent fusion network for tumour segmentation in PET-CT images, using cascaded CNNs to combine the multimodal imaging features with the estimated intermediary segmentation results from multiple recurrent fusion phases [65]. Fu et al proposed a multimodal spatial attention module (MSAM) that automatically learned to emphasize spatial regions related to tumours and suppress normal regions with high FDG uptake in PET images, then the spatial attention maps were subsequently employed in a CNN to guide the tumour segmentation on

CT images [66]. Xue et al. proposed a co-learning framework for improving the segmentation of tumour lesions in PET-CT images, utilizing a shared down-sampling block and hierarchical feature co-learning module to detect salient and complementary features of the lesions during the fusion process [67]. Xiang et al proposed a modality-specific segmentation network for tumour segmentation in PET-CT images, this network used two separate branches to simultaneously learn the PET and CT imaging features, and meanwhile, an adversarial task was incorporated to minimise the modality discrepancy and preserve modality-common representation [68]. The aforementioned approaches employed one stage CNN model trained in an end-to-end manner, which could fail to capture fine details or heterogeneous structures of tumour regions, especially when multiple complex imaging modalities, such as PET-CT, are involved. Both the texture and semantic information of small tumours could be lost after several down-sampling convolutional layers, especially within the one-stage CNN model, resulting in false negative segmentation errors.

2.4 Conventional Predictive Model Construction Methods

Conventional radiomics methods can be divided into five main steps, as illustrated in Figure 2.5: (i) image acquisition/reconstruction; (ii) regions of interest (ROIs) segmentation; (iii) hand-crafted features (e.g., intensity, texture, shape, etc.) extraction; (iv) predictive model construction. The first step involves routine clinical image acquisition and reconstruction. With the acquired medical images, the ROIs (e.g., tumours) needs to be manually annotated by at least one experienced radiologist or oncologist. Quantitative hand-crafted radiomics features, which contain information about tumour heterogeneity, texture patterns, and biomarkers, are then extracted from the segmented ROIs from the medical images. Afterwards, statistical machine learning methods (e.g., random forest, logistic regression, SVM, etc.) are implemented to correlate these features with patient's future outcomes or

diagnostic results. In general, conventional radiomics studies extract hand-crafted radiomics features to quantify various tumour phenotypes on medical images and further utilise these features as predictors of genetics and clinical outcomes [69]. As the hand-crafted radiomics features are defined and extracted based on many different factors, such as the disease types, imaging modalities, etc., the existing conventional radiomics methods have been categorised according to the predictive models comprising of regression models, random forest, support vector machine, k-nearest neighbour.

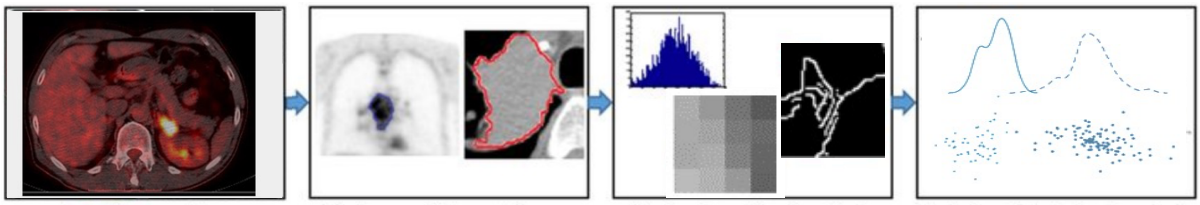


Figure 2.5. The pipeline of conventional radiomics methods. (a) image acquisition / reconstruction, (b) region of interest segmentation, (c) radiomic features extraction, and (d) statistical analysis and modelling. The PET-CT images are from the public challenge dataset AutoPET [5].

2.4.1 Cox Proportional Hazards Regression

Cox Proportional Hazards Regression, also known as the Cox regression model, is a statistical methodology used to investigate the relationship between a set of independent variables and survival time [70]. The Cox regression model provides an estimate of the hazard function, which is the probability that an event (such as death, failure, or relapse) will occur at a certain time given that an individual has survived up to that time. It also helps identify the factors that affect the hazard rate with the help of variables that are known to affect the outcome, such as age, gender, and initial diagnosis. The hazard function is defined as follows:

$$\lambda(t|X_i) = \lambda_0(t)exp\left(\beta_1 X_{i1} + \cdots + \beta_p X_{ip}\right) \quad (2.1)$$

where $X_i = (X_{i1}, \ldots, X_{ip})$ is the realized values of the covariates for subject i. β_i are the effect parameter(s). Note that between subjects, the baseline hazard $\lambda_0(t)$ is identical (has no dependency on i).

Cox regression has been commonly used in radiomics studies because it is capable of analysing time-to-event data and identifying imaging features that are associated with specific clinical outcomes. For example, Spraker et al implemented a Cox regression model to demonstrate the prognostic value of radiomics features extracted from MR images on overall survival prediction tasks for patients with STSs [71]. Similarly, Coroller et al. used the Cox regression model with univariate and multivariate analysis to prove that at least 35 radiomics features from pre-treatment CT images are prognostic for distant metastasis (DM) in lung adenocarcinoma [72]. Although these studies proved that radiomics features were capable of capturing detailed information about the tumour phenotype and could be used as a prognostic biomarker for clinical patient outcomes, only single modality images were utilised in this study. More recently, Lv et al. proposed a multi-level fusion strategy for multimodal PET-CT images to predict the survival outcomes of patients with head-and-neck cancer; they demonstrated that Cox regression models with fusion of radiomics features consistently outperformed those using single modality images in patient outcome prediction [73].

2.4.2 Logistic Regression

Logistic regression is a statistical method used to analyse the relationship between a dependent variable and one or more independent variables. It is a type of regression analysis commonly used to predict the probability of a binary outcome, such as yes/no or true/false [74].

In logistic regression, the dependent variable is a categorical variable with two possible outcomes, and the independent variables can be either continuous or categorical. The model estimates the probability of the dependent variable belonging to one of the two possible outcomes based on the independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1, and a logit transformation is applied to the probability - that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function is represented by the following equations:

$$Logit(p_i) = \frac{1}{1+exp(-p_i)} \tag{2.2}$$

$$\ln \frac{p_i}{1-p_i} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \, . \tag{2.3}$$

In this logistic regression equation, $Logit(p_i)$ is the dependent variable and X_k is the independent variable. The β_k parameters, or coefficients, in this model, are commonly estimated via maximum likelihood estimation. This method tests different values of $\beta_k$through multiple iterations to optimize for the best fit of log odds. All of these iterations produce the log-likelihood function, and logistic regression seeks to maximize this function to find the best parameter estimate. Once the optimal coefficient (or coefficients if there is more than one independent variable) is found, the conditional probabilities for each observation can be calculated, logged, and summed together to yield a predicted probability.

Logistic regression has been widely used in radiomics, where hand-crafted radiomics features extracted from medical images are defined as independent variables and patient outcomes or diagnostic results as the dependent variable. These models can be used to identify which radiomics features are most strongly associated with the clinical outcome of interest and

can be used to develop more accurate and personalized diagnostic and prognostic tools for various cancers. For example, Xiong et al. evaluated the prognostic value of FDG PET radiomics features in predicting local control in oesophageal cancer after treatment with concurrent chemoradiotherapy [75]. In addition, Vallieres et al built a joint FDG-PET and MRI texture-based model to evaluate the lung metastasis risk in STS at an early stage [7]. They implemented multivariable and univariate analysis to select radiomics features that were extracted from different modality scans, then the selected features were fed into a logistic regression model to predict lung metastases.

2.4.3 Random Forest

Random forest is a type of ensemble learning algorithm that combines multiple decision trees to create a more robust and accurate predictive model, and it can be used for both classification and regression problems [76].

In radiomics, random forest models are often used to identify the most important radiomics features for clinical outcomes prediction. The algorithm works by randomly selecting subsets of features from the input data and building multiple decision trees based on these subsets. The algorithm then combines the predictions of these decision trees to create a final prediction. By using multiple decision trees and random feature selection, random forest models can reduce overfitting and improve the generalization of the model to new data. Vallieres et al extracted 1615 hand-crafted radiomics features from pre-treatment FDG-PET and CT images to assess the risk of locoregional recurrences and DM for patients with head-and-neck cancer, random forest was used as the predictive model in this study [12]. Peeken et al proved the prognostic value of CT-based radiomics features by using the random forest in patients with STSs treated with neoadjuvant radiation therapy, these features can be used to predict tumour grading, and systemic and local progression. [14] While random forest provides

flexibility and reduced risk of overfitting, it requires more computational resources and usually takes a longer time to build the forest from the decision trees.

2.4.4 Support Vector Machine

Support vector machine (SVM) is a supervised machine learning algorithm that can be used for both classification and regression tasks. SVM works by identifying the best possible hyperplane, which separates the different classes of data points in a high-dimensional feature space, in order to achieve maximum margin between the classes [77].

Specifically, given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. SVM maps training examples to points in space so as to maximise the width of the gap between the two categories. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. SVM works by identifying the best possible hyperplane, which separates the different classes of data points in a high-dimensional feature space, in order to achieve maximum margin between the classes. This makes the classification more accurate and less susceptible to overfitting. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

In radiomics, SVMs are often utilised to classify medical images based on the radiomics features extracted from them or to predict clinical outcomes associated with certain cancers. Hao et al defined a new hand-crafted shell feature from PET images alone to predict distant failure in non-small cell lung cancer (NSCLC) and cervix cancer patients, the shell feature showed better predictive performance with an SVM classifier in both cohorts when compared to other matrix-based hand-crafted features extracted from PET [6]. Juntu et al have proved

that MRI-based radiomics features were able to discriminate the malignancy grade of STSs by leveraging SVM [78].

2.4.5 K-Nearest Neighbour

K-nearest neighbour (KNN) is a non-parametric, supervised learning method, which uses proximity to make classifications or predictions about the grouping of an individual data point [79]. While it can be used for either regression or classification problems, it is typically used as a classification algorithm in radiomics, working off the assumption that similar points can be found near one another.

For classification problems, KNN works by finding the k-number of the nearest data points in the training set to the new data point and then assigning the class of the majority of those k-nearest neighbours to the new data point. The value of k is chosen by the user and influences the accuracy of the model. Corino et al have proved that MRI-based radiomics features were able to discriminate the malignancy grade of STSs by leveraging the KNN classifier respectively [13]. Although KNN is easy to implement and has few hyperparameters, KNN does not perform well with high-dimensional data inputs and is more prone to overfitting.

2.4.6 Challenges to Conventional Radiomics Methods

Despite the promising clinical potential of radiomics, there are still some limitations in current conventional radiomics studies.

Firstly, while there is an increasing interest in utilising multimodal medical imaging data in radiomics [80], all these conventional radiomics studies still heavily depend on the hand-crafted radiomics features and high demand of prior knowledge (e.g., ROI annotation and manual tuning of many parameters for different predictive models).

Secondly, not all hand-crafted features are recommended for use as they may be sensitive

to the imaging data acquisition modes and corresponding reconstruction parameters [69]. Specifically, Galavis et al assessed the variability of 50 PET radiomics features based on different acquisition modes, reconstruction algorithms, iteration numbers, and other factors, and among them, forty features were shown to have substantial variability with a relative difference of more than 30% [81].

Thirdly, conventional radiomics studies rely on an ad-hoc definition of "traditional" hand-crafted image features that have been used across a wide range of generic object recognition tasks. Thus, such features, which are primarily based on textures, shape, intensity, etc., are not optimised for specific imaging modalities and specific disease types, e.g., features required in PET lung cancer detection and characterisation differ greatly from MRI brain degradation analysis.

In addition, these low-level features are usually extracted in the small lesion region, and hence, they cannot fully describe the image that comprises high-level semantics. Therefore, it will be challenging to make use of these relatively simple image features to improve the prognosis of the disease with the increasing amount of high-dimensional information obtained from medical images.

Nevertheless, due to the wide variations among different imaging modalities, the differences resulting from manual annotation and the complicated feature selection and analysis processes, the conventional radiomics methods tend to be time-consuming and error-prone with human bias.

2.5 Deep Learning-Based Predictive Model Construction Methods

In recent years, deep learning-based approaches have made significant advances in the field of medical image analysis and have been applied to various tasks such as classification, segmentation, and detection [82]. The success of these approaches can be largely attributed to

their ability to train computational models consisting of multiple processing layers, which enables the learning of features that correspond to both shallow and deep semantics of the images [83].

Despite the demonstrated success of deep learning in medical image analysis, there are still gaps between radiomics and deep learning that have not been properly addressed. Most deep learning methods in medical image analysis were designed to detect or classify diseases, and they did not investigate their potential applications in predicting prognostic characteristics of the disease. In this chapter, a comprehensive overview of the progress and challenges in supervised deep learning-based radiomics methods is provided. Specifically, the existing radiomics studies will be categorised and reviewed based on commonly used deep learning techniques, including recurrent neural networks (RNNs) and convolutional neural networks (CNNs) [84].

2.5.1 Convolutional Neural Network (CNN)

CNN is one of the most widely used techniques in deep learning and has become the state-of-the-art in radiomics. This is attributed to the CNN's ability to directly extract high-level semantic information in an end-to-end manner that is meaningful to the patient outcomes, which reduces the need for prior knowledge in hand-crafted radiomics features definition and other manual input [15], [16]. The fundamentals of CNN are: 1) convolutional layers to learn adjustable weights (i.e., filters) that can be used to extract features from the input (see Figure 2.8); 2) activation functions (e.g., sigmoid, ReLU, also known as rectified linear units, etc.) to introduce non-linearity into the network, allowing the network to learn more complex relationships between input and output; 3) pooling layers (e.g., average- or max- pooling) to reduce the spatial dimensions (i.e., width and height) of the input volume (down-sampling the data in the process), while preserving its depth and aggregating features. The output of a CNN

with a single layer can be represented as follows:

$$f_s(\boldsymbol{X};\boldsymbol{W},\boldsymbol{b}) = pool_p\big(\sigma(\boldsymbol{W} *_s \boldsymbol{X} + \boldsymbol{b})\big) \tag{2.4}$$

where $\boldsymbol{X}$ is the input feature map, $\sigma(\cdot)$ is the pointwise non-linear function, and $\theta = \{\boldsymbol{W},\boldsymbol{b}\}$ are the set of parameters (i.e., weights and biases). The $pool_p(\cdot)$ function denotes a pooling operation and p is the size of the pooling region. The symbol $*_s$ represents convolution operation with stride s. As a result, the resolution of the output feature map f_s is downsampled by a factor of s.

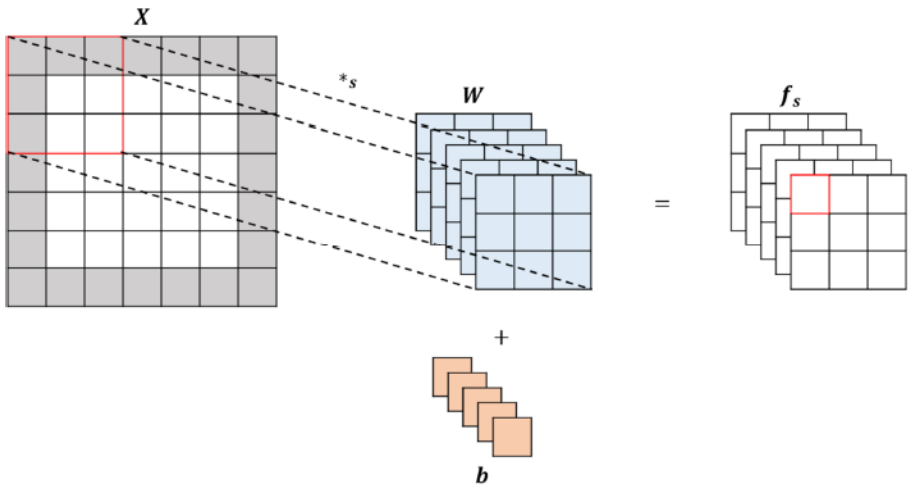


Figure 2.6. An example convolutional layer. The input feature map has a dimension of 5×5. The convolutional layer has 4 kernels (in blue) of 3×3, padding (in grey) is 1 and the stride s is 2.

There are many well-known CNN backbones employed for different tasks, such as U-Net for object segmentation [85], ResNet for classification [86], etc. A large proportion of recent CNN-based methods are derived from those well-established backbones. As for the CNN in radiomics, at an earlier stage, Oakden-Rayner et al tried to predict patient longevity

and overall individual disease status with cross-sectional CT images by leveraging both deep learning and conventional radiomics methods [87]. Their proposed method provided a conceptual framework for the use of radiomics techniques to identify tissue-wide pathological changes (imaging biomarkers) to quantify the latent health of the patient. Their experimental results demonstrated that modern deep learning techniques could be used within a radiomics framework, and were comparable to conventional radiomics and clinical manual methods for longevity prediction. Subsequent studies have demonstrated that CNN can better capture and understand radiological information from medical images [44]. Motivated by this success, many studies have attempted to adopt CNN for radiomics. For instance, Lao et al investigated the potential of transfer learning-based deep learning features in generating similar radiomics signatures for overall survival prediction in Glioblastoma Multiforme (GBM) [19]. Their method involved extracting conventional radiomics features and deep learning features from the segmented tumour region where CNN is utilised as a feature extractor for deep learning features. The optimal feature set was selected using the least absolute shrinkage and selection operator (LASSO) Cox regression model. The proposed method achieved better performance than traditional risk factors for overall survival prediction, and the ensembled model further improved predictive performance. However, this study still utilised a complicated and error-prone method similar to conventional radiomics methods and focused only on single-modality MR images. In contrast, Li et al implemented an end-to-end deep learning-based radiomics model to predict the mutation status of isocitrate dehydrogenase 1 (IDH1) for patients with low-grade GBM [88]. Instead of calculating hand-crafted radiomics features only from segmented regions in medical images, all the image features were completely obtained by CNN without extra calculation and extraction operations. Similarly, Kumar et al presented a CNN-based discovery radiomics framework for lung cancer grade classification using CT imaging data. aiming to discover customised radiomics sequencers tailored for better representing lung

cancer characteristics [89]. Additionally, deep learning networks were used for mortality risk stratification with standard-of-care CT images from NSCLC patients by Hosny et al. [16]. Diamant et al. utilised a CNN to extract deep learning-based radiomics features from pre-treatment CT images to predict treatment outcomes of patients with head and neck squamous cell carcinoma, outperforming conventional radiomics methods [90]. Furthermore, Zhu et al and Hermessi et al developed a deep learning diagnosis model based on MR images to classify meningioma and STS respectively [18], [91]. More recently, Fu et al compared the performance of hand-crafted radiomics features and deep learning-based radiomics features in predicting the neoadjuvant chemoradiation response for locally advanced rectal cancer patients with pre-treatment diffusion-weighted MR images [17], and found that deep learning achieved significantly better prediction performance. However, most of these deep learning-based radiomics methods were designed for a single imaging modality or a particular body region, which limits their ability to derive complementary radiomics features that fully represent tumours.

Few studies have reported on CNN-based radiomics approaches for multiple combined imaging modalities. Chen et al. proposed a hybrid predictive model consisting of a many-objective radiomics model and a 3D CNN to predict lymph node metastases using PET-CT images from patients with head-and-neck cancer [21]. The underlying assumption was that CNN's abstract level features and hand-crafted texture features were complementary [92], so combining them could provide more accurate results. However, using a 3D CNN alone performed sub-optimally, and incorporating conventional radiomics components, such as hand-crafted features, improved performance. This is because the deep learning features extracted from the last convolutional layer of the CNN only contained high-level semantic information, while the texture information from shallow convolutional layers was neglected. Therefore, existing CNN-based radiomics studies suffered from the limitations of traditional

radiomics methods and failed to fully utilise the potential of CNNs in capturing complementary information from multimodal medical images.

2.5.2 Recurrent Neural Network (RNN)

RNN is a type of neural network that is commonly used for natural language processing and sequential data analysis. Unlike CNNs, RNNs are designed to process sequences of data, where the current input depends on the previous inputs and their associated hidden states. This makes RNNs particularly useful for tasks where the meaning and context of a sequence of data can only be understood by considering the sequence as a whole. The key feature of RNNs is that they have a "memory" that allows them to retain information about previous inputs, making them well-suited for processing sequential data [93].

In radiomics, RNNs can be used to analyse medical images that have a time component, such as MRI images taken over a period of time to track the progression of a disease. RNNs can also be used for other tasks such as predicting the survival time of a patient, where the features extracted from the images are processed in a time-dependent manner. For example, Azizi et al. adopted Long-Short-Term-Memory (LSTM) for the classification of benign and malignant prostate cancer based on sequences of US images [94]. Although the experimental results showed that they achieved higher predictive accuracy with image sequences, the generalizability of their method was strictly limited to sequential medical imaging data with a temporal component. When compared to CNNs, RNNs require a large number of computational resources due to their recurrent nature and the need to process each input in a sequence. Additionally, RNNs are not well-suited for processing spatial information in medical images while CNNs are generally better at capturing spatial information in volumetric medical imaging data.

2.6 Summary of Gaps

In this Chapter, a comprehensive review of radiomics studies has been conducted, encompassing conventional and deep learning-based radiomics methods. The analysis of the literature has identified advancements in the field and the challenges that still exist. In summary, the gaps identified can be summarised as follows:

1. **Image Annotations.** Both conventional and deep learning-based radiomics methods rely on lesion labels to define the ROIs for feature extraction. Manual annotations are considered the gold standard, but the scarcity of manually annotated medical images due to high cost and complicated acquisition procedures remains a challenge. There is an urgent need for algorithms that are less dependent on labels, such as with semi- and self-supervised methods, as well as methods designed to work with small amounts of labelled data e.g., transfer learning and domain adaptation.
2. **Image Segmentation.** Despite great advances in single modality segmentation methods, the task of automated PET-CT lesion segmentation poses a unique challenge due to the need to consider complementary features from both modalities. Therefore, robust methodologies are anticipated in the field of radiomics, particularly for multimodal PET-CT images.
3. **Multi-modal Imaging.** Most existing deep learning-based radiomics methods are designed for a single modality image (e.g., CT, MRI, etc.), which limits their ability to derive complementary radiomics features that represent tumours. Although there is a limited number of CNN-based radiomics studies where multimodal PET-CT images were used, their focus is to utilise both conventional radiomics features and deep radiomics features extracted from CNNs for clinical applications, rather than advancing the technical usage of CNN for radiomics. As a result, these methods suffer from similar limitations of conventional radiomics methods in prior knowledge about feature

definition and extraction, and these methods fail to fully utilise the potential of CNNs in capturing complementary information from multimodal medical images. Therefore, there is a need for deep learning methods in radiomics that enable optimal extraction and analysis of information from multimodal PET and CT images.

4. **Manual Architecture Design.** Although CNN-based deep learning-based radiomics methods have become the state-of-the-art, existing CNNs heavily rely on human expertise to design dataset-specific deep learning architectures, including number of convolutional layers, and structure of convolutional blocks. Architecture design and optimization require a significant amount of domain knowledge, such as to validate the architecture performance and tuning the hyperparameters. This complexity is further compounded when multiple imaging modalities, such as PET-CT, are used in the radiomics analysis. Therefore, automated architecture search can ease the subsequent manual designs, such that final architecture of CNN can be achieved more efficiently.

Chapter 3. Tumour Segmentation with Self-supervision Enabled False Positive and False Negative Reduction Network

In this chapter, an automated tumour segmentation method, termed Self-supervision enabled False positive and false negative Reduction Network (SFRN), is introduced for multimodal PET-CT images. SFRN includes a self-supervised pre-training strategy to improve the feature representation ability of the CNNs within SFRN in characterizing tumour regions, contributing to better generalizability across different tumours in multimodal PET-CT images. Moreover, a multi-stage network is built with the pre-trained weights obtained from the self-supervised pre-training strategy, which consists of a global segmentation module (GSM) to coarsely locate the tumour regions, followed by a local refinement module (LRM) with a hybrid loss to iteratively eliminate the false positive and false negative errors introduced from the GSM. Furthermore, a classification branch is further incorporated to enhance the ability of our network to distinguish the tumour regions from healthy regions in

multimodal PET-CT images, which allows the further elimination of false positive regions. Experimental results with three multimodal PET-CT datasets (one public challenge dataset, one public STS dataset and one in-house lung cancer data) show that the SFRN achieved consistently better segmentation results when compared to the existing state-of-the-art methods. In addition, the preliminary version of the SFRN (GSM with false positive reduction network only) achieved the leading performance in Dice score and ranked 2nd place in the final ranking at the 2022 MICCAI AutoPET Challenge [95], [96]. In this chapter, following contributions are further introduced: (i) an LRM to remove false negative segmentation and, (ii) to incorporate a classification branch to improve the ability of tumour segmentation.

3.1 Contributions

The main contributions of this chapter are as follows:

1) A self-supervised pre-training strategy is proposed to improve the feature representation ability of tumour regions in the CNNs, contributing to better generalizability across different diseases and PET-CT datasets where tumours could be located at any part of the body with better segmentation results. When compared to the existing self-supervised learning (SSL) methods for multimodal medical images, the strategy can be applied to different multi-modal imaging data without specific characteristics required.
2) A local refinement module (LRM) is introduced to refine the candidate tumour regions generated from the GSM. A hybrid loss function is formulated to simultaneously minimize both false positive and false negative errors in multimodal PET-CT images. When compared to the existing one-stage methods, the SFRN is capable of segmenting tumours of various sizes, whereas existing methods tend to segment relatively large tumours only.

3) A classification branch is incorporated into the SFRN to further enhance the ability of the CNNs to distinguish the tumour regions from healthy regions in multimodal PET-CT images, which allows the elimination of false positive regions.

This chapter's contributions address the challenge of existing automated tumour segmentation methods for multimodal images in the field of radiomics mentioned in Chapter 1. It also aligns to the gap in Chapter 2 that refers to challenges and limitations in radiomics using multimodal PET-CT images. This chapter also expands on the literature in Section 2.3 by including detailed descriptions to the state-of-the-art comparative methods. The challenge dataset and STS dataset is from public resources and has been cited, the other lung cancer dataset is private in-house data.

3.2 Materials and Methods

3.2.1 Materials and Pre-processing

Non-small cell lung cancer (NSCLC) and Soft-tissue sarcomas (STSs) datasets were used in the evaluation. Although the AutoPET challenge provided a large training dataset, the testing data was not available, and less than half of the PET-CT studies had different types of diseases [96]. Thus, the challenge data was only used for self-supervised pre-training. All three datasets were pathologically confirmed (see Figure 3.1 for three examples of PET-CT images from different datasets), their details were described below and shown in Table 3.1.

For the AutoPET challenge data (denoted as AutoPET dataset), the given training dataset consisted of 1,014 PET-CT scans derived from 900 patients acquired at the University Hospital Tübingen, Germany [96]. All images were in NifTI format. There were 513 scans without lesions, and 188, 168, and 145 scans were histologically proven with malignant melanoma, lung cancer, and lymphoma, respectively. In addition, all patients had clinical reports including cancer diagnosis, sex, and age. A separate testing dataset was not released

to the public and was only used for online evaluation. The testing dataset had a preliminary testing set of 5 studies for self-evaluation and a final testing set of 200 studies for final ranking. The preliminary testing set was part of the final testing set where 100 studies were from the same hospital as the training database (University Hospital Tübingen) and the other 100 scans were acquired from the University Hospital of the LMU in Munich with a similar acquisition protocol. The tumour regions for all the training and testing datasets were annotated by two radiologists with more than 5 years of experience in Hybrid Imaging and experience in machine learning research.

Table 3.1. Details of the Datasets

Datasets	**AutoPET**	**STS**	**FD**
Number of Studies	501	51	117
Lesions	malignant lymphoma, melanoma, NSCLC	STSs	NSCLC
PET Resolution (pixels)	400×400	128×128	168×168
PET Spacing (mm^2)	2.04	3.91-5.47	4.06
CT Resolution (pixels)	512×512	512×512	512×512
CT Spacing (mm^2)	0.98	0.98	1.37
Slice Thickness (mm)	2-3 (CT) 3 (PET)	3.27	3 or 5
Scanner	Siemens Biograph mCT	GE Discovery ST	Siemens Biograph TruePoint

The STSs dataset (denoted as the STS dataset) is publicly available at the Cancer Imaging Archive and was acquired from McGill University Health Centre, Quebec, Canada [7], [97]. This dataset has 51 patients with histologically proven, extremity primary STS. Each patient had 4 imaging modalities FDG PET, CT and T1-weighted and T2-weighted with fat-suppression (T2FS) MR scans. The gross tumour volume was manually annotated slice-by-slice on T2FS MR scans by an expert radiation oncologist and then registered to PET and CT images via a rigid registration algorithm.

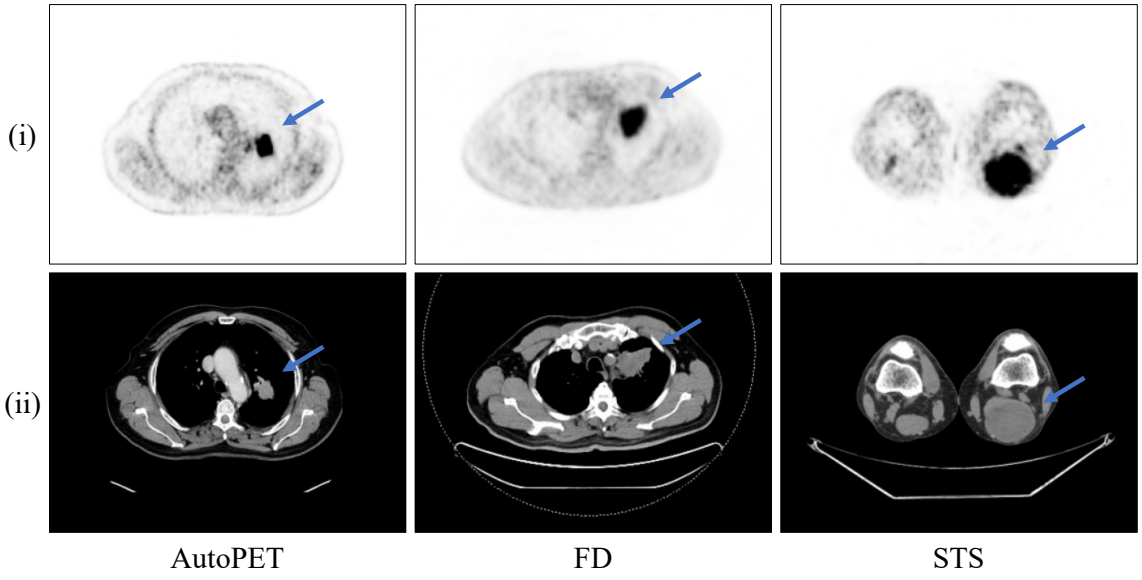


Figure 3.1. Three examples of PET-CT images from different datasets used in this chapter. The top row (i) are PET images, and the bottom row (ii) are CT images. The blue arrows point to the tumour regions.

The NSCLS dataset was acquired from the Department of Nuclear Medicine at Fudan University Shanghai Cancer Centre, Shanghai, China (denoted as the FD dataset). As shown in Table 3.1, the slice thickness was either 3mm or 5mm, and the specific number of corresponding PET-CT studies was 77 and 40 respectively. All the data were analysed anonymously, two radiologists annotated the tumour regions from the axial plane of the CT

with the ITK-SNAP software (V3.6). PET images were used to assist in excluding conditions, such as pneumonia.

Multiple pre-processing steps were applied to all the imaging datasets. Firstly, Linear up-sampling was applied to both PET and CT images based on their spacing and slice thickness to ensure that the spatial dimensions were the same in all directions. Secondly, to compress the usage of GPU memory, all the PET-CT image volumes were cropped into a patch size of 224×224 in the axial plane. Then the images were set to the SUV range of [0, 14.25] for PET and HU range of [-800, 400] for CT, and further mapped to [0, 1] via min-max normalization. Finally, for PET images, the input slices were normalized with the mean and standard deviation values of the entire training dataset, such that to adjust all the regions of interest (ROIs) to a notionally common scale based on the metabolism intensity of tumour regions. For the CT images, the input slices were normalized with the mean and standard deviation values of the individual patient.

3.2.2 Overview of the Proposed Method

The SFRN is outlined in Figure 3.2. The SFRN takes multimodal PET-CT images as input and consists of two main modules - global segmentation module (GSM) and local refinement module (LRM). The encoders within the SFRN are first pre-trained via self-supervised learning for the enhanced ability to characterize the tumour regions. Then the GSM is used to coarsely locate the candidate tumour regions in PET-CT images, which will be refined by the LRM to eliminate both false positive and false negative segmentation errors.

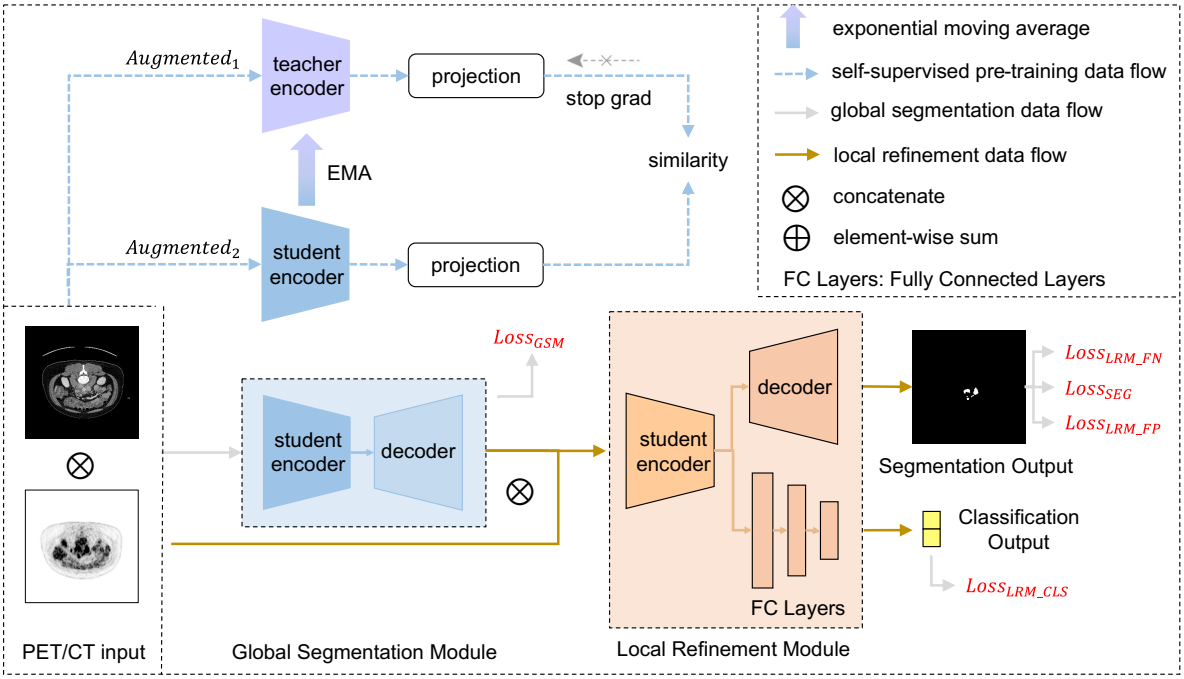


Figure 3.2. The overview of the SFRN. The arrows in different colours indicate different steps which are taken sequentially. The self-supervised pre-training improves the representation ability of the encoders in our model to characterize the tumour regions in PET-CT images; this is followed by the global segmentation which uses the pre-trained ResNet50 encoder to coarsely delineate the candidate regions. Afterward, the LRM removes the false-positive and false-negative errors using the output of the GSM that is concatenated with the paired PET-CT images as input.

3.2.3 Self-supervised Pre-training

Inspired by the existing self-supervised learning methods for natural images [98]–[100], a self-supervised pre-training framework was introduced, which was a combination of contrastive learning and knowledge distillation. This framework was illustrated in the top-left module of Figure 3.2 and Algorithm 3.1, the pseudo-code implementation.

Although this pre-training framework shared a similar overall structure as recent self-supervised approaches [98], [101]–[103], our method discarded the predictor block and incorporated knowledge distillation. Knowledge distillation is a learning paradigm that aims to

transfer knowledge from a complicated model to a simpler model by training the simpler model with the SoftMax outputs of the complicated model [104]. Given a fixed teacher network f_{ϑ_t}, the student network would be optimized by minimizing the cross-entropy loss w.r.t the parameters of the student network ϑ_s, and the equations are shown below:

$$Loss = \min_{\vartheta_s}[-P_t(\boldsymbol{x}) \log P_s(\boldsymbol{x})] \tag{3.1}$$

$$P_s(\boldsymbol{x})^{(i)} = \frac{\exp\left(f_{\vartheta_s}(\boldsymbol{x})^i/\tau_s\right)}{\sum_{k=1}^{K} \exp\left(f_{\vartheta_s}(\boldsymbol{x})^k/\tau_s\right)} \tag{3.2}$$

Where $\boldsymbol{x}$ is an input image, the output probability distribution of both networks over K dimensions are denoted by P_s and P_t respectively, and the probability P is obtained by normalizing the output of the network with a SoftMax function. The τ_s is a temperature parameter that controls the sharpness of the output distribution and is usually larger than zero. However, unlike typical knowledge distillation, our teacher network is built from past iterations of the student network without any real labels using an exponential moving average (EMA, as shown in Figure 3.1). As there is no predictor in our framework, the teacher and student networks share the identical architecture. Thus, there will be no issues with the EMA in our framework.

Furthermore, dimensional collapse is a common issue where the model maps all input to the same constant vector [105], and there are trivial solutions such as an extra predictor [102], or an additional clustering step [106]. To avoid model collapse, a centring and sharpening of the momentum teacher outputs (illustrated in Algorithm 1 as well) is further implemented, according to [100], centring prevents one dimension to dominate but encourages collapse to the uniform distribution, while the sharpening has the opposite effect. Applying both operations

balanced their effects which is sufficient to avoid collapse in presence of a momentum teacher.

Algorithm 3.1 Self-supervised pre-training

```
% fs, ft: student and teacher branches
% C: center (K)
% temps, tempt: student and teacher temperatures
% momn, momc: network and center momentum rates
ft.params = fs.params
for x in dataloader do     % load a minibatch x with n samples
    x1, x2 = augment(x), augment(x)     % random augment views

    s1, s2 = fs(x1), fs(x2)     % student output n-by-k
    t1, t2 = ft(x1), ft(x2)     % teacher output n-by-k

    s1, s2 = softmax(s1/temps , dim=1), softmax(s2 / temps, dim=1)
    t1, t2 = t1.detach(), t2.detach()     % stop gradient
    % center + sharpen
    t1 = softmax((t1 - C) / tempt , dim=1)
    t2 = softmax((t2 - C) / tempt , dim=1)
    ce1 = -(t1 * log(s2)).sum(dim=1).mean()
    ce2 = -(t2 * log(s1)).sum(dim=1).mean()

    loss = (ce1 + ce2) / 2
    loss.backward()

    % student, teacher and center updates
    update(fs)
    ft.params = 1 * ft.params + (1 - momn) * fs.params
    C = momc * C + (1 - momc) * cat([t1, t2]).mean(dim=0)

end for
```

3.2.4 False-positive and false-negative Reduction Network

The false-positive and false-negative network consisted of two main modules, as shown in Figure 3.2: a global segmentation module (GSM) and a local refinement module (LRM). The encoders in both modules were pre-trained on the AutoPET training dataset only via the self-supervised learning method demonstrated in Section 3.2.3.

For the GSM, a ResNet50 encoder was combined with a U-Net based decoder [85], [86], where the concatenated 3-channel PET-CT images were used as the input. Two channels of the input images were set to be PET while the rest of 1 channel was assigned to CT. Then a combined loss $Loss_{GSM}$ was used for training. $Loss_{GSM}$ consisted of a dice loss and a cross-entropy loss [107], [108], which was defined as:

$$Loss_{GSM} = -\frac{2\sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2} - \frac{1}{N}\sum_{i=1}^N [g_i \log p_i + (1 - g_i)\log(1 - p_i)] \quad (3.3)$$

Where $p_i \in [0, 1]$ is the probability map of each pixel within the predicted tumour region, $g_i \in [0, 1]$ is the probability map of each pixel within the ground truth tumour mask (label), and the sums run over all available N pixels of the segmentation.

The GSM can coarsely annotate the tumour lesion regions with a probability map at a threshold of 0.5. There would be segmentation errors, including false positives and false negatives, just like the other existing CNNs methods. These errors would be further reduced with our LRM.

The backbone of our LRM is a 2D U-Net, combined with a classification branch. The original PET-CT images were fed into the LRM along with the output from the GSM, making the input data 5 channels (i.e., paired PET-CT images, global tumour probability map, and global binary segmentation prediction). To effectively remove the false positives, and false negatives as well as keep the accurate tumour segmentation, four losses were designed to optimize the LRM. A standard dice and cross-entropy loss were employed as the fundamental segmentation loss (denoted as $Loss_{SEG}$), which was mathematically the same as the $Loss_{GSM}$. As the tumours usually possessed a small proportion of the entire input images, the false positives and false negatives would be much less, resulting in an imbalanced segmentation problem. Therefore, the focal loss was used to optimize the training process of reducing false positives and false negatives [109], the loss function of reducing false positive errors was defined as:

$$Loss_{LRM\_FP} = \min\left\{\frac{1}{N}\sum_{i=1}^N [-\alpha(1 - p_i)^\gamma \log(p_i)]\right\} \quad (3.4)$$

Where $p_i \in [0, 1]$ was the predicted probability of false positive pixels, $\alpha \in [0, 1]$ was a weighting factor s to balance the sample, γ was a focusing parameter smoothly adjusted the rate at which easy examples were down-weighted, then the sums ran over all available N pixels of the segmentation. In this work, α and γ were set to 0.25 and 2 respectively. A similar formula was held for $Loss_{LRM\_FN}$.

For the classification branch, there were two steps to produce the final classification results: (1) the feature maps at each scale of the convolutional block in the pre-trained encoder were first fed into a convolutional layer with a kernel size of 1 × 1 × 1, followed by batch normalization, ReLU activation, and a global average pooling layer. (2) all the features obtained from the last step were in the same dimension, and would be concatenated into two fully connected layers. ReLU layers and dropout layers with a probability of 0.5 were added after the first connected layer to reduce overfitting. A weighted cross-entropy loss was used for the training process and denoted as $Loss_{LRM\_CLS}$. As there were four losses utilized within the LRM, the combined losses of our LRM were defined as followed:

$$Loss_{LRM} = Loss_{SEG} + Loss_{LRM\_CLS} - Loss_{LRM\_FP} + Loss_{LRM\_FN} \tag{3.5}$$

3.2.5 Implementation Details

Our method was implemented with PyTorch [110] framework using one NVIDIA GeForce GTX 2080Ti GPU. Our model was initialized using the approach presented by He et al [111], and an adaptive-moment estimation with decoupled weight decay (AdamW) [112] was used for network optimization. During the training phase, the batch size was set to 8 and the learning rate was set to 0.0001 using a cosine annealing schedule. Data augmentation techniques were in real-time to avoid overfitting. The used data augmentation techniques are random rotation

(90°, 180°, or 270°) in the axial axis and randomly flipping in one of the two axes (sagittal and coronal).

Furthermore, only PET-CT slices with lesions was used for the GSM, while an equal number of PET-CT slices without tumours were sampled and added to the training of the LRM. All the training was terminated when no further change was in the total loss. In our method, the total loss was generally stable after 160 epochs.

3.3 Experiments and Results

3.3.1 Experimental Setup

The following experiments were conducted to evaluate the effectiveness of the proposed method. The proposed method was firstly compared with the state-of-the-art PET-CT segmentation methods including: (1) U-Net: a widely used baseline in medical image segmentation with an encoder-decoder architecture [85]; (2) Co-learning: a two-branched U-Net was designed to extract and fuse PET and CT information with spatial context [63]; (3) MSAM: multi-modal spatial attention module, which used PET images as an attention map to guide the tumour segmentation on CT images [66]; (4) MosNet: a modality-specific segmentation network, which used two separate branches to simultaneously learn the PET and CT imaging features along with a modality discriminator [68]; (5) nnUNet [113]: a U-Net based self-configuring CNN for biomedical image segmentation which demonstrated good generalizability across 23 public datasets. nnUNet achieved 1st and 3rd place at the AutoPET challenge.

An ablation study was also conducted to investigate all the components of our SFRN, including the SSL strategy, GSM, the false positive/negative reductions, and the classification branch.

All the experiments were conducted on two datasets: each of the datasets was divided into a 70/10/20 (training/validation/testing) split. For example, with the FD dataset, 81 patients were

used for training, 12 patients for validation, and 24 patients for testing. Images acquired from a patient can either be within training, validation or testing dataset only.

3.3.2 Evaluation Metrics

For all experimental comparisons, the *p*-value with an unpaired student's *t*-test was computed. Four established segmentation evaluation metrics were adopted: Dice score, precision (Pre.), sensitivity (Sen.), and specificity (Spe.), defined as below:

$$Dice = \frac{2|\boldsymbol{GT} \cap \boldsymbol{PS}|}{|\boldsymbol{GT}| + |\boldsymbol{PS}|} \tag{3.6}$$

$$Pre. = \frac{|TP|}{|TP| + |FP|} \tag{3.7}$$

$$Sen. = \frac{|TP|}{|TP| + |FN|} \tag{3.8}$$

$$Spe. = \frac{|TN|}{|TN| + |FP|} \tag{3.9}$$

Where $\boldsymbol{GT}$ denotes the ground truth, $\boldsymbol{PS}$ is the algorithm predicted segmentation result, TPs are the true positive pixels (ROIs), TNs are the true negative pixels (background), FPs are the false positive pixels and FNs are the false negative pixels.

Using the four metrics in combination provides a comprehensive assessment of the performance of a tumour segmentation algorithm. The Dice score evaluates the spatial overlap, while precision and sensitivity measure the trade-off between correctly identifying tumour regions and minimizing false positives and false negatives. Specificity complements these metrics by focusing on true negatives. In the medical field, where patient care and treatment decisions are at

stake, it is essential to strike a balance between these metrics to ensure accurate and clinically meaningful tumour segmentations.

3.3.3 Results

3.3.3.1 Comparison to the State-of-the-art Methods

A comparison of our SFRN method against the state-of-the-art methods is presented in Table 3.2. The results indicate that our SFRN obtained better performance with the best Dice score on STS (67.25) and the FD (74.20) datasets. With the FD dataset, our SFRN also achieved the best sensitivity (78.04) and the second-best precision (75.43).

Table 3.2. Classification Performance Comparison with Existing Radiomics Methods

	Methods	Evaluation Metrics				
		Dice	Pre.	Sen.	Spe.	*p*-value
STS	U-Net [85]	59.63±28.79	64.50±30.39	64.49±28.74	99.65±1.42	1.07×10^{-6}
	Co-learning [63]	60.53±25.72	58.41±27.06	70.53±27.01	99.45±0.67	8.83×10^{-28}
	MSAM [66]	61.62±32.12	58.30±32.86	71.75±34.16	99.59±0.34	1.72×10^{-15}
	MosNet [68]	63.27±26.83	60.18±28.00	**76.07**±29.73	99.32±0.49	1.57×10^{-11}
	nnUNet [113]	66.57±27.63	**87.11±19.82**	61.63±33.31	**99.92±0.25**	3.82×10^{-26}
	BDAV (ours)	65.43±32.51	66.41±34.15	69.38±34.46	99.71±0.49	3.15×10^{-3}
	SFRN (ours)	**67.35±25.57**	67.41±26.79	72.97±**23.39**	99.72±0.41	-
FD	MosNet [68]	53.82±36.80	64.55±39.84	51.64±37.67	99.84±0.29	9.33×10^{-19}
	U-Net [85]	63.97±24.92	71.95±26.29	63.80±26.90	99.74±0.36	0.031
	Co-learning [63]	64.25±27.47	70.27±27.19	67.18±30.71	99.70±0.28	8.86×10^{-35}
	nnUNet [113]	65.70±33.03	65.44±33.67	70.17±36.28	**99.98±0.22**	2.52×10^{-14}
	MSAM [66]	68.13±30.28	70.01±30.66	71.50±32.88	99.71±0.35	5.01×10^{-59}
	BDAV (Ours)	73.95±23.28	**76.35**±24.82	76.59±25.62	99.75±0.28	1.9×10^{-3}
	SFRN (ours)	**74.20±22.94**	75.43±**24.43**	**78.04±25.18**	99.73±0.29	-

The bold numbers represent the best results, and they are presented in the form of 'mean value ± standard deviation'.

Segmentation results from the comparison methods for the FD dataset are presented in

Figure 3.3. The CT images from the FD dataset were normalized to a lung window for better visualization of lung cancer regions, and the PET images were normalized to the median value of max SUV values of all the patients and presented in the inverted grayscale colormap. Different colours were painted on the segmentation outputs to present differences among comparison methods (i.e., green for true positive regions, black for true negative regions, red for false positive regions, and yellow for false negative regions). It shows that our SFRN was able to consistently outperform nnU-Net, MosNet, MSAM, and Co-learning, even with small tumours, as shown in Figure 3.3.

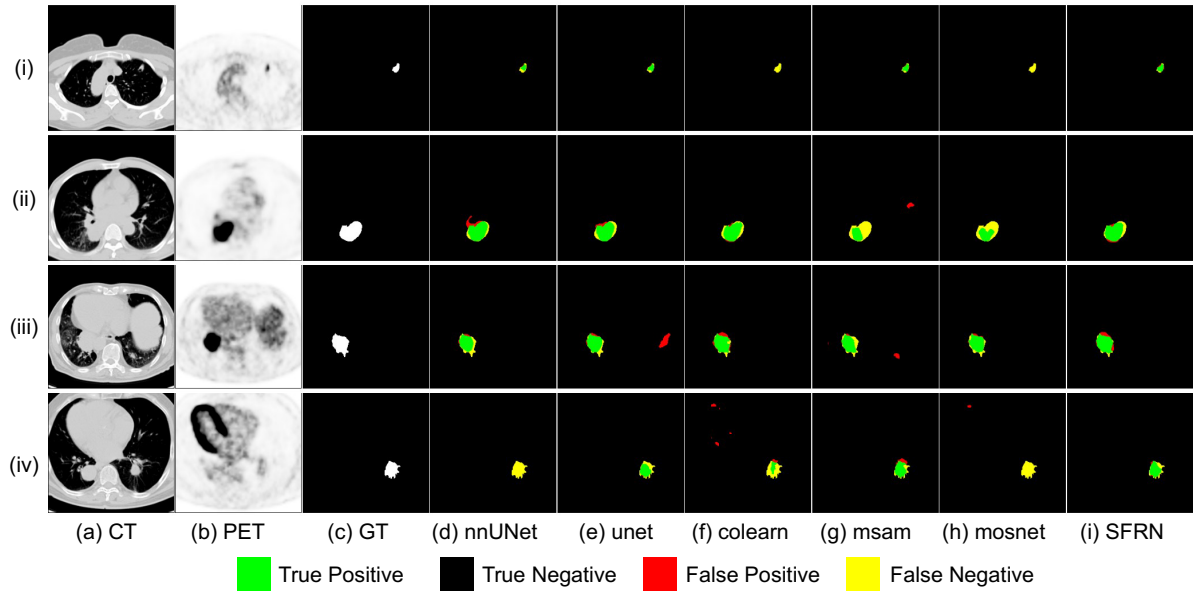


Figure 3.3. Four example PET-CT studies of lung cancer (in axial slices) with (a) CT in the first column, (b) PET in the second column, and (c) ground truth (GT) in the third column. The segmentation results from different methods are presented in columns (d) to (i). Note that co-learning (column f) and MosNet (h) failed to segment the tumour on the first example (row i), and nnUNet (column d) and MosNet (h) failed to segment the tumour on the fourth example (row iv).

Figure 3.4 presents the results using STS datasets with the CT normalized to soft tissue

window. Our SFRN consistently produced better segmentation results with fewer false positive and false negative predictions. All the comparison methods suffered from false positive errors due to normal physiological regions being picked up that are in close proximity to the tumour regions (see the red segmentations in Figure 3.4).

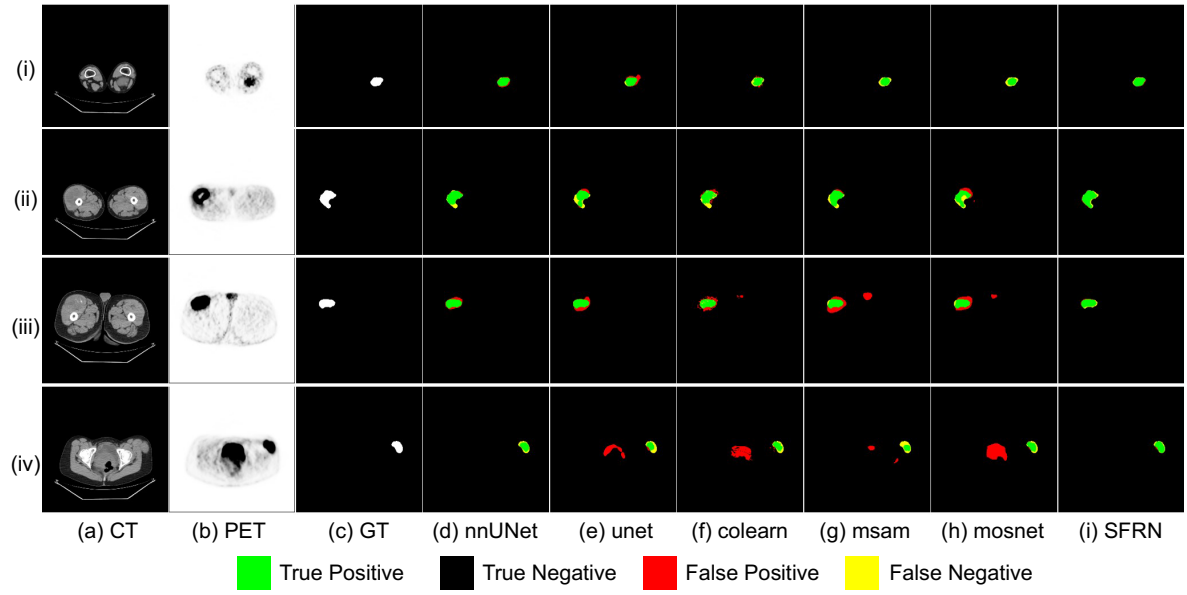


Figure 3.4. Four example PET-CT studies of STSs shown on axial image slices with (a) CT in the first column, (b) PET in the second column, and (c) ground truth (GT) in the third column. The segmentation results from different methods are presented in columns (d) to (i).

Table 3.3. Part of the MICCAI 2022 AutoPET challenge final results on the hidden testing dataset.

Teams (Backbones)	Evaluation Metrics		
	Dice↑	False Negative Volume ↓	False Positive Volume ↓
Blackbean (nnUNet)	**62.26**	0.5445	**2.8372**
BDAV (GSM+LRM_FP)	62.08	0.7518	3.6111
FightTumor (nnUNet)	60.04	**0.4681**	5.1026
UIH-FL (nnUNet)	60.96	0.8316	4.8533
Heiligerl (nnUNet + SwinUNETR)	60.52	0.6287	5.8741

The bold numbers represent the best results.

The MICCAI 2022 AutoPET challenge results are presented in Table 3.3 and our algorithm was referred to as BDAV. Since the testing dataset is inaccessible, we, therefore, were unable to evaluate our updated SFRN method on the testing dataset. All other top-5 teams used nnUNet as their backbone with different training strategies and post-processing techniques, such as a different combination of loss functions, assembling multiple trained models, and more data augmentation transformations, which were heavily dependent on their experience and prior knowledge about the tumour and imaging characteristics. I suggest that nnUNet, used for the comparisons in Table 3.2, could be regarded as a fair comparison between the top five challenge methods and our SFRN. In the challenge, our GSM + LRM_FP eventually achieved overall 2nd place over all the evaluation metrics. Specifically, our Dice score (62.08) and false positive volumes (3.6111) were ranked 2nd place with a difference of 0.18 and 0.77 respectively. As shown in Table 1, the BDAV result was further improved by our SFRN across all the evaluation metrics, and this is further detailed in the ablation study.

3.3.3.2 SFRN Ablation Study

To evaluate the effectiveness of our self-supervised pre-training strategy, six sets of comparative experiments were conducted on U-Net, the most widely used segmentation baseline, along with our SFRN and its components on both STS and FD datasets. Each comparative experiment contained two training strategies - with/without pre-training. The results are shown in Table 3.4. The results demonstrated that our SSL is capable of constantly improving the segmentation performance in Dice scores across all the comparison methods on both STS and FD datasets.

Different components of our SFRN, i.e., GSM, LRM_FP, LRM_FN, and LRM_CLS have been evaluated on both STS and FD datasets with the results shown in Table 3.5. The results

indicated that each component from our SFRN methods consistently contributed to a better performance e.g., segmentation accuracy. Our SFRN method achieved the best Dice score (67.35), precision (67.41), and specificity (99.72) with minimum standard deviations on the STS dataset. It also obtained the best Dice score (74.20) on the FD dataset with a second-best precision (75.43) and specificity (99.73), and a third-best sensitivity (78.04).

To quantitatively evaluate the effectiveness of different components in reducing segmentation errors, both the false positive rate (FPR) and false negative rate (FNR) were counted for each method and shown in Table 3.5. They are calculated by dividing the number of false positive/negative pixels by the input image size, and ‰ was used as the unit for better comparison. When compared to the GSM baseline, the LRM_FP and LRM_FN successfully decreased the false positive segmentation and false negative segmentation respectively on STS and FD datasets. Moreover, our SFRN with the LRM_CLS obtained the least segmentation errors on both datasets (i.e., 3.65‰ FPR and 1.05‰ FNR on the STS dataset, 3.24‰ FPR and 5.52‰ FPR on the FD dataset).

Segmentation predictions from the different components, for example of tumours of lung cancer and STS, are presented overlaid on top of PET or CT images in Figure 3.5 and Figure 3.6. CT images in both Figure 3.4 and Figure 3.5 were normalized to soft tissue window (see column (b)) for clarity. For the lung cancers from the FD dataset, the segmentation results were shown on CT images that are normalized to lung window (see column (c) – (g) in Figure 3.5), which could exhibit more details within the lung region than the CT images in the soft-tissue window. For the STS dataset, as the contour of STS can be hard to see on the CT images, the segmentation results were shown on top of the PET images (see column (c) – (g) in Figure 3.5). In addition, as the variations of false positive and negative pixels among the ablation studies are relatively small compared to the tumour size, the visualization of segmentation errors in different colours can be hard to differentiate, all the ground truth of tumour labels were outlined in red while the

segmentation results were in blue on the medical images.

As expected, the methods with the LRM_FP component tended to have fewer false segmentation errors while those with LRM_FN had fewer false negative errors. But our SFRN consistently produced more accurate segmentations which were aligned with the ground truth (see column (g) in Figure 3.5 and Figure 3.6.

Table 3.4. Evaluation of Our Self-supervised Pre-training Strategy on Two Datasets

	Methods	SSL	Evaluation Metrics			
			Dice	Pre.	Sen.	Spe.
STS	U-Net [85]		59.63±**28.79**	**64.50**±**30.39**	64.49±**28.74**	**99.65**±**1.42**
		✓	**62.75**±29.86	60.42±30.83	**73.86**±32.36	99.24±1.49
	GSM		62.32±**27.42**	**60.08**±**30.48**	75.20±**26.50**	**99.30**±**1.12**
		✓	**64.31**±30.96	57.55±31.43	**83.30**±33.15	99.07±1.54
	GSM+ LRM_FN		62.70 ±28.24	**62.87**±30.72	69.97±31.83	**99.66**±**0.42**
		✓	**65.23**±**26.88**	62.14±**28.8**6	**77.75**±**27.29**	99.35±0.99
	GSM+ LRM_FP		62.88±**28.92**	65.93±**30.39**	67.77±**31.64**	99.63±0.49
		✓	**65.43**±32.51	**66.41**±34.15	**69.38**±34.46	**99.71**±**0.49**
	GSM+ LRM_FP+ LRM_FN		63.73±**26.89**	**64.26**±29.85	71.75±**28.38**	**99.56**±**0.59**
		✓	**66.30**±27.54	63.37±**28.82**	**77.34**±28.77	99.50±0.67
	SFRN		64.04±28.48	65.92±29.64	72.14±32.83	99.60±0.57
		✓	**67.35**±**25.57**	**67.41**±**26.79**	**72.97**±**23.39**	**99.72**±**0.41**
FD	U-Net [85]		**63.97**±**24.92**	71.95±**26.29**	**63.80**±**26.90**	99.74±**0.36**
		✓	67.47±27.51	**69.68**±28.59	72.18±30.64	**99.63**±0.45
	GSM		72.36±**23.50**	70.89±**26.05**	**78.97**±24.36	99.57±0.52
		✓	**73.83**±23.04	**74.68**±24.50	78.31±**25.21**	**99.71**±**0.29**
	GSM+ LRM_FN		72.63±24.16	**76.77**±26.78	73.34±25.37	**99.75**±**0.29**
		✓	**73.99**±**22.95**	74.92±**24.27**	**78.36**±**25.14**	99.72±0.29
	GSM+ LRM_FP		72.59±24.32	**77.20**±26.44	72.56±25.71	**99.77**±0.32
		✓	**73.95**±**23.28**	76.35±**24.82**	**76.59**±**25.62**	99.75±**0.28**
	GSM+		73.43±24.30	**77.37**±27.31	74.82±**25.07**	**99.77**±0.42

LRM_FP+ LRM_FN	✓	**74.04**±**23.18**	75.34±**24.77**	**77.69**±25.31	99.72±**0.28**
SFRN		73.66±24.09	**77.02**±27.50	75.38±**24.54**	**99.76**±0.47
	✓	**74.20**±**22.94**	75.43±**24.43**	**78.04**±25.18	99.73±**0.29**

The bold numbers represent the best results, and they are presented in the form of 'mean value ± standard deviation'.

Table 3.5. Results of SFRN Ablation Study on Two Datasets (Part 1).

	GSM	LRM_FN	LRM_FP	LRM_CLS	Dice	Pre.	Sen.	Spe.
STS	✓				64.31±30.96	57.55±31.43	**83.30**±33.15	99.07±1.54
	✓		✓		65.43±32.51	66.41±34.15	69.38±34.46	99.71±0.49
	✓	✓			65.23±26.88	62.14±28.86	77.75±27.29	99.35±0. 99
	✓	✓	✓		66.30±27.54	63.37±28.82	77.34±28.77	99.50±0.67
	✓	✓	✓	✓	**67.35**±**25.57**	**67.41**±**26.79**	72.97±**23.39**	**99.72**±**0.41**
FD	✓				73.83±23.04	74.68±24.50	78.31±25.21	99.71±0.29
	✓		✓		73.95±23.28	**76.35**±24.82	76.59±25.62	**99.75**±**0.28**
	✓	✓			73.99±22.95	74.92±**24.27**	**78.36**±**25.14**	99.72±0.29
	✓	✓	✓		74.04±23.18	75.34±24.77	77.69±25.31	99.72±0.28
	✓	✓	✓	✓	**74.20**±**22.94**	75.43±24.43	78.04±25.18	99.73±0.29

The bold numbers represent the best results, and they are presented in the form of 'mean value ± standard deviation'.

Table 3.6. Results of SFRN Ablation Study on Two Datasets (Part 2).

	GSM	LRM_FN	LRM_FP	LRM_CLS	FPR‰	FNR‰	*p*-value
STS	✓				11.99±19.9	1.71±91.81	1.63×10^{-4}
	✓		✓		3.72±6.34	2.03±1.77	3.15×10^{-3}
	✓	✓			8.33±12.86	1.51±1.46	1.28×10^{-3}
	✓	✓	✓		6.45±8.66	1.44±1.28	0.036
	✓	✓	✓	✓	**3.65**±**5.29**	**1.05**±1.**08**	-

FD	✓				3.63±3.77	5.68±15.04	1.28×10^{-5}
	✓		✓		3.49±3.78	5.72±**14.89**	1.9×10^{-3}
	✓	✓			3.59±3.78	5.59±14.92	3×10^{-3}
	✓	✓	✓		3.41±3.67	5.57±14.93	0.087
	✓	✓	✓	✓	**3.24±3.67**	**5.52**±14.90	-

The bold numbers represent the best results, and they are presented in the form of 'mean value ± standard deviation'.

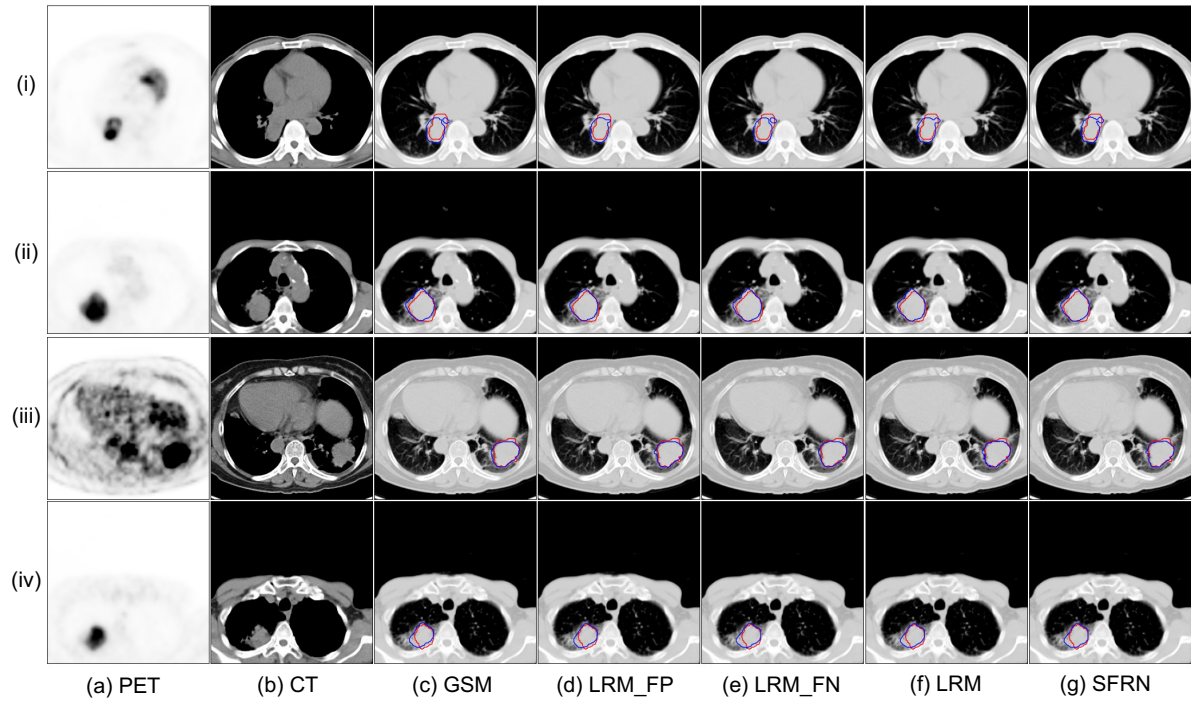


Figure 3.5. Four PET-CT studies of lung cancer patients with (a) PET in the first column and (b) CT in the second column. The segmentation results from methods using different components of the SFRN, including the GSM, the false positive (FP) / false negative (FN) reduction within the LRM, and the classification branch, are shown in columns (c) to (g). The red contour outlines the 'ground truth' segmentation and the blue contour outlines the results from the comparison methods.

3.4 Discussions

The main findings are that: (i) the SFRN consistently outperformed the state-of-the-art methods

across different datasets; (ii) the self-supervised pre-training strategy improved the segmentation performance across each of the component within our SFRN on both STS and FD datasets; (iii) the LRM was superior in removing both false positive and false negative regions and thus improving the overall segmentation accuracy; (iv) the classification branch effectively decreased segmentation errors via the ability to avoid false positive segmentations on healthy regions; and (v) the GSM achieved competitive performance by using the self-supervised training when compared to the existing methods on both STS and FD datasets.

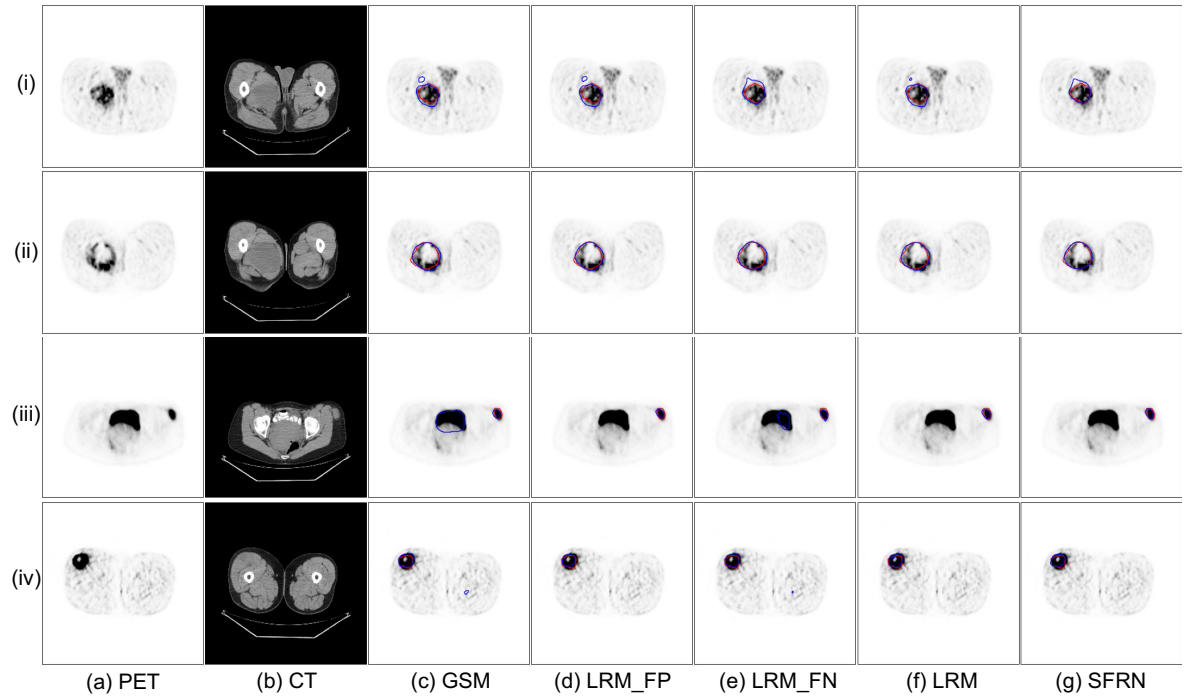


Figure 3.6. Four PET-CT studies of STSs shown on axial image slices with (a) PET in the first column, and (b) CT in the second column. The segmentation results of methods using different components are shown in columns (c) to (g), where the red contour outlines the 'ground truth' segmentation and the blue contour outlines the results from the comparison methods.

3.4.1 Comparison to the Existing PET-CT Tumour Segmentation Methods

With the MICCAI 2022 AutoPET challenge, a competitive segmentation performance was

achieved with the best-performing team (see Table 3.3). The method of the 1st ranked team (Blackbean [114]) used larger input image size, and longer training time compared to the SFRN, which required more computational resources. Although the 3rd team (FightTumor [115]) obtained the least false negative volume, their method utilised extra post-processing steps based on the number of connected voxels, and the HU value in the corresponding CT image of their generated segmentation outputs, and ensembled 13 trained models to obtain the final segmentation results. The reliance on ensemble limits the generalizability of their model and require prior knowledge in medical imaging and parameter tuning. Similarly, the 4th team (UIH-FL [116]) implemented different post-processing steps based on based on the number of connected voxels in the generated segmentation outputs.

In the Dice metric, the nnUNet outperformed the BDAV on the STS dataset by 1%, but it was surpassed on the FD dataset by the BDAV with a margin of 8% (see Table 3.2). The nnUNet achieved the best precision (87.11) and specificity (99.92) with the minimum standard deviations on the STS dataset (i.e., 19.82 and 0.25 respectively, see Table 3.2), while the sensitivity of 61.63 was the lowest among all the comparison methods. This was attributed to the fact that nnUNet tends to under-segment the tumour regions, which resulted in fewer false-negative errors. As shown in Figure 3.3 and Figure 3.4, the segmentation results of nnUNet (column (d)) generally produced additional false negative regions (shown in yellow) and fewer false positive regions (shown in red) when compared to other comparison methods. When segmenting small tumours, such as lung cancers from the FD dataset, the tendency of under-segmenting resulted in failure to detect tumours, especially when the tumour region was close to normal organs with much higher SUV (e.g., heart, see row (iv) in Figure 3.3). This led to a decreased performance over all the evaluation metrics on the FD dataset apart from the consistent best specificity of 99.98 with the minimum stand deviation 0.22.

Although MosNet achieved the highest sensitivity of 76.07 on the STS dataset, however, it

this was likely attributed to the fact that MosNet is dominant by the modality-specific representation features from PET during the experiments, such that tends to overfit to segment all the regions with higher SUV values, including normal high uptake regions. Figure 3.4.(h) shows example segmentation results, where MosNet over-segmented the bladder region to be part of tumour regions. As for lung cancer studies, tumours can be small or share similar anatomy with their surroundings, and it is challenging to determine based on the SUV if the pixels in the PET image correspond to disease or a benign process (e.g., pneumonia). Such inability resulted in missing tumours that were close to normal high-uptake activities (see column (h) in Figure 3.3), leading to the poor performance of MosNet on the FD dataset (see Table 3.2).

There was a consistent trend among U-Net, Co-learning, and MSAM methods on both the FD and the STS datasets (see Table 3.2) where MSAM outperformed Co-learning and Co-learning outperformed U-Net. Co-learning improved U-Net via the fusion of anatomical and functional visual features from PET-CT images, but the fusion approach from Co-learning was not designed to capture nuanced morphological details that were more critical in tumour segmentation, which resulted in limited performance improvement of less than 1% increase from U-Net in dice score on both datasets. Further improvement of MSAM over Co-learning was likely attributed to the multimodal spatial attention module, which allowed the network to leverage the PET images as an attention map to guide the tumour segmentation on the CT images. Although the precision of MSAM was slightly dropped, the dice score increased from 64.25 to 68.13 on the FD dataset and from 60.53 to 61.62 on the STS dataset. Nevertheless, our SFRN consistently outperformed the state-of-the-art methods across various datasets by introducing the local refinement module to remove both the false positive and negative errors. The p-values of comparison methods which are smaller than 0.05 indicated that the segmentation outputs of our SFRN were significantly different.

Despite the inability to benchmark the competition results to the SFRN, its performance

comparison to BDAV on different datasets can be used to infer SFRN's performance compared to the competition results. The SFRN makes large improvements over the BDAV across all the evaluation metrics on two datasets except for the precision value on the FD dataset, where the BDAV and SFRN obtained the top two precision scores. The SFRN was also capable of accurately delineating both small-sized tumours as well as avoiding including false positive regions with the normal physiological process.

3.4.2 SFRN Ablation Study

Within the SFRN, the ResNet-50 was used as the encoder of the GSM combined with a U-Net decoder, while the LRM utilized a standard 2D U-Net as the backbone. When compared to the existing supervised methods in Table 3.2, the GSM obtained competitive performance with the second-highest Dice score on the STS dataset and the highest Dice score on the FD dataset. It can be attributed to the pre-training model via self-supervised contrastive learning on the MICCAI 2022 AutoPET challenge data. However, the GSM was better at tumour detection and exhibited a tendency for over-segmenting the tumour regions (see column (c) in Figure 3.5 and Figure 3.5, blue contours tend to cover high SUV regions), which resulted in the highest sensitivity of 83.3 and the second highest sensitivity of 78.31 on the STS and FD dataset respectively

The LRM was further divided into three parts: false negative reduction (LRM_FN), false positive reduction (LRM_FP), and classification branch (LRM_CLS). As expected, the use of LRM_FP and LRM_FN could effectively reduce the corresponding segmentation errors on both datasets. Although the improvement in Dice score was relatively small from the LRM_FP, the decreased false positive segmentations contributed to a 9% increase in precision on the STS dataset, making both the specificity (99.71) and precision (66.41) of LRM_FP the 2nd best on the STS dataset. Similarly, the LRM_FP achieved the best precision of 76.35 and the best specificity of 99.75 on the FD dataset. In addition, the LRM_FN constrained the CNN to emphasize the

neglected cancerous regions, removing the false negative segmentations from the predicted output. This led to the highest sensitivity of 78.36 on the FD dataset and the second-best sensitivity of 77.75 on the STS dataset. The overall performance could be further improved by integrating the LRM_FP and LRM_FN, which contributed a 2% improvement in Dice score along with minor enhancement in precision and specificity on both datasets.

As for the self-supervised pre-training strategy, the pre-trained U-Net achieved an improvement of more than 3% in Dice on both datasets (see Table 3). Besides, our strategy constantly improved the overall segmentation performance across all the components, i.e., GSM, LRM_FP, LRM_FN, and LRM_CLS, on both STS and FD datasets. Although the precision and specificity slightly dropped in most of the pre-trained comparison methods in Table 3.4, it was expected that the models would tend to over-segment the tumour regions with a better perceptiveness obtained from the self-supervised pre-training. Additionally, Our SFRN without pre-training obtained Dice scores of 64.04 and 73.66 on the STS and FD datasets respectively (see Table 3), which are also competitive with the existing PET-CT tumour segmentation methods. It was only outperformed by nnUNet on the STS dataset with a 1.39% in Dice and beat all other existing methods on the FD dataset (see Table 3.2).

Overall, the SFRN obtained the highest Dice score of 67.35 and 74.20 on the STS and FD datasets, respectively, by incorporating a classification branch into the LRM, and also achieved the highest specificity (99.72) and precision (67.41) on the STS dataset. This was attributed to the ability to distinguish benign or cancerous regions in the PET-CT slices, which reduced the false positive segmentation on the entire slice, thereby boosting the performance in precision and specificity. For instance, Fig 5 row (i) shows the false positive segmentations above the tumour which were reduced by using LRM_FP (column (d)), and the false negative segmentations were refined into over-segmentation by using LRM_FN (column (e)), then the integration of these two components produced a balanced segmentation output that was shown in column (f), which could

be further refined by our SFRN with the classification branch. To quantitatively evaluate the performance of our SFRN and different components, the number of false positive and false negative pixels in each PET-CT slice was calculated for methods using different components. Our SFRN obtained the least number of false positive pixels (183.43 on STS, 162.63 on FD) and false negative pixels (53.14 on STS, 277.40 on FD) on both datasets (see Table 3.3). All the methods contained the LRM_FP segmented fewer false positive pixels in their outputs when compared to those without false positive reduction, while those with LRM_FN presented fewer false negative pixels in their segmentation results. Therefore, the LRM was consistently superior at learning to suppress non-tumour regions (including benign pixels of high intensity) and highlighting tumour regions in PET-CT, contributing to overall better segmentation results.

3.5 Summary

In this chapter, A CNN-based approach is proposed to improve the performance of segmenting tumours from multimodal PET/-T images. The SFRN method consists of two main modules, namely the global segmentation module and the local refinement module. The global segmentation module was designed to coarsely delineate the candidate tumour regions, then the candidate tumour regions were refined by removing both false positive and false negative segmentation errors via the local refinement module. The encoders in both modules were pre-trained via self-supervised learning for better feature representation ability of tumours. The SFRN surpassed the existing methods for tumour segmentation with PET-CT images on two different datasets.

Chapter 4. Patient Outcome Prediction with Constrained Hierarchical Multimodal Feature Learning

In this chapter, a deep learning-based radiomics method, named the Constrained Hierarchical Multimodal Feature Learning (CHMFL), is introduced that is designed for radiomics with multimodal PET-CT images. This new radiomics method is capable of integrating functional imaging (PET) features with anatomical imaging (CT) features, at different scales, in an end-to-end iterative manner. In contrast to existing radiomics methods, the CHMFL removes the reliance on manual input and prior knowledge in medical images, such as tumour annotations and feature selection needed in conventional radiomics. Further, the CHMFL leverages the complementary information across different modalities to automatically focus on semantically important regions, i.e., tumours. The CHMFL method was evaluated in predicting the development of distant metastases (DM) using imaging data before the DM developed on a well-established benchmark soft-tissue sarcomas (STSs) PET-CT dataset. The experimental results demonstrate that CHMFL achieved overall better performance when

compared to the state-of-the-art methods.

4.1 Contributions

The main contributions of this chapter are as follows:

1) A constrained feature learning (CFL) module is introduced to spatially guide the network training process to focus on the important semantic regions (e.g., tumours). CFL formulation enables the targeting of functional 'hot spots' in PET which refers to pixels with high FDG uptake within the anatomical context of CT. The CFL allows the CNN to automatically detect and focus on the tumour, while conventional radiomics and other CNN-based radiomics methods with either a single or a multimodal imaging data require manual annotations as the input to constrain the feature extraction process within the tumour region.
2) A hierarchical multi-modality feature learning (HMFL) module is proposed that derives optimal radiomics features by integrating complementary features across modalities at different scales. The formulation of the module combines multimodal features from different scales in an iterative manner. In comparison, existing multimodal radiomics methods extract imaging features separately from individual imaging modalities and fuse the features later or integrate the multi-modal images at an earlier stage. The hierarchical combination of features enables a more complex and flexible fusion of PET and CT features, e.g., low-level PET texture features from a shallow layer with semantic CT features from a deeper layer.

This chapter's contributions address the challenge of existing multimodal radiomics studies mentioned in Chapter 1 and aligns to the gap that there is a need for deep learning methods that enable optimal extraction and analysis of information from multimodal PET and CT images in

Chapter 2. The chapter also further expand on the literature in Chapter 2.4 by including detailed comparisons to the state-of-the-art comparative methods.

4.2 Materials and Methods

4.2.1 Materials and Pre-processing

A well-benchmarked public PET-CT STSs dataset from the Cancer Imaging Archive is used for method evaluation [7], [97] (see Chapter 3.2.1 Materials and Pre-processing for details).

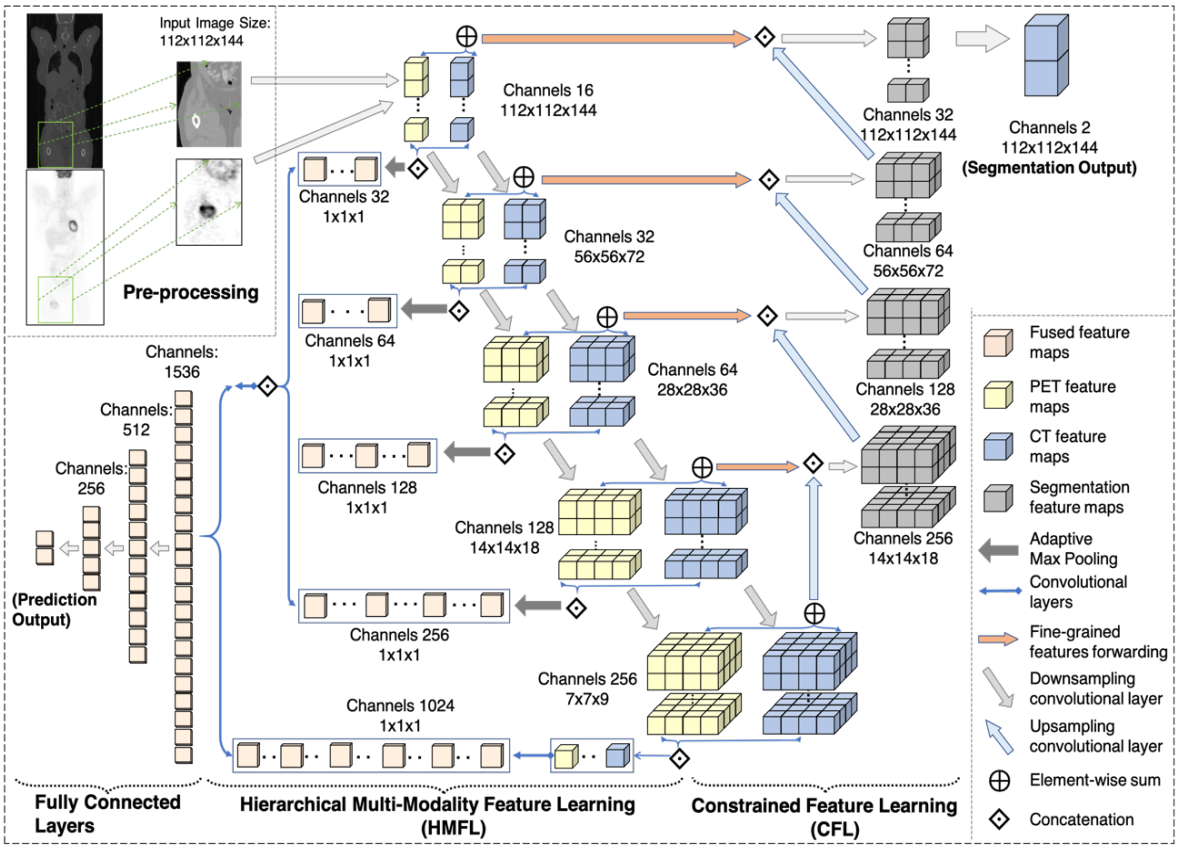


Figure 4.1. The CHMFL architecture.

4.2.2 Overview of the Proposed Method

In Figure 4.1, the CHMFL architecture was outlined. The volumetric PET and CT images were pre-processed and then fed separately into two identical branches. Each branch has multiple

down-sampling convolutional layers for feature extraction (as shown within the yellow PET and blue CT feature maps in Figure 4.1). The feature maps derived after each convolutional layer were adaptively pooled and then concatenated into a single feature vector to facilitate hierarchical multimodal feature learning (HMFL). The constrained feature learning (CFL) module used several up-sampling convolutional layers to guide the network to focus on the important regions (e.g., the tumour). This process also incorporated the fine-grained features forwarded from the HMFL module at each level. Finally, the derived multimodal PET-CT features (as shown in the left lower part of Figure 4.1) were fed into three fully connected layers for distant metastases (DM) prediction.

4.2.3 Constrained Feature Learning (CFL) Module

The CFL module was designed to guide the learning process to focus on semantically important regions at both the training and inference stages. This was achieved by gathering and assembling the complementary information from multimodal PET-CT images to obtain a 2-channel volumetric segmentation output. 4 transposed convolutional blocks were used to expand the spatial support from the feature maps at a lower scale for up-sampling. These up-sampling blocks at different levels shared similar structures (see CFL module in Table 4.1 for details). Meanwhile, the multimodal PET-CT features extracted from the HMFL module were forwarded to the up-sampling blocks by horizontal connections (see Figure 4.1). In this way, the network gathered fine-grained detail for tumour contour prediction that would be otherwise lost in the down-sampling path. In turn, tumour regions were emphasized in the HMFL module by the backpropagation process. Moreover, in order to avoid the vanishing gradient problem with network deepening, a residual learning was formulated after the concatenation of forwarded PET-CT features and the corresponding up-sampled feature maps at each level: the concatenated feature map was processed through several convolutional layers and non-

linearities, then added to the output of the last non-linearity within the residual learning.

Table 4.1. Network Architecture Used in the HMFL and CFL Module

Layers	Details (kernel size, stride, padding, …)	Output Size (batch size, channel number, …)
HMFL Module		
Input Transition	Conv3d (5×5×5, 1, 2) BatchNorm; ELU	1×16×112×112×144
Down_Conv_1	Conv3d (2×2×2, 2, 0) BatchNorm; ELU	1×32×56×56×72
Down_Conv_2	Conv3d (2×2×2, 2, 0) BatchNorm; ELU	1×64×28×28×36
Down_Conv_3	Conv3d (2×2×2, 2, 0) BatchNorm; ELU	1×128×14×14×18
Down_Conv_4	Conv3d (2×2×2, 2, 0) BatchNorm; ELU	1×256×7×7×9
CFL Module		
Up_Conv_1	ConvTranspose3d (2×2×2, 2, 0) BatchNorm; ELU	1×128×14×14×18
Up_Conv_2	ConvTranspose3d (2×2×2, 2, 0) BatchNorm; ELU	1×64×28×28×36
Up_Conv_3	ConvTranspose3d (2×2×2, 2, 0) BatchNorm; ELU	1×32×56×56×72
Up_Conv_4	ConvTranspose3d (2×2×2, 2, 0) BatchNorm; ELU	1×16×112×112×144
Output Transition	Conv3d (5×5×5, 1, 2) BatchNorm; ELU Conv3d (1×1×1, 1, 0); SoftMax	1×2×112×112×144

During the training stage, two loss functions were employed for different tasks. A pixel-wise cross-entropy loss was used to compare the predicted segmentation output with the ground truth tumour annotation. Another cross-entropy loss was used for DM prediction. Given a weight w for our CFL module $0 \leq w \leq 1$, the total loss L was defined as follows:

$$L = -(1-w) * \sum_{m=1}^{M=2} p_{1,m} \, log \, q_{1,m} - w * \sum_{n=1}^{N} p_{2,n} \, log \, q_{2,n} \qquad (4.1)$$

where $p_{1,m}$ represents the target probability of DM, $q_{1,m}$ (the output of this network)

represents the predicted probability of developing DM, and M denotes the number of output neurons generated by the last fully connected layer in this network, $q_{2,n} \in Q$ is the predicted binary segmentation volume, $p_{2,n} \in P$ is the ground-truth binary annotation image and N denotes the total number of image voxels. w_i is a weight to balance the two losses.

4.2.4 Hierarchical Multimodal Feature Learning (HMFL) Module

Five convolutional blocks were used for multimodal image feature extraction (more details of the HMFL module are provided in Table 4.1). PET and CT images were processed separately by the identical PET and CT branches. Within each convolutional block, the output feature map of the 3D convolutional layer was defined as:

$$F = \boldsymbol{W} * \boldsymbol{X} + \boldsymbol{b} \tag{4.2}$$

where $\boldsymbol{X}$ is the input to the convolution layer, $*$ is the convolution operation, $\boldsymbol{W}$ denotes for the learned weights, and $\boldsymbol{b}$ is the learned bias. A batch normalization layer and a non-linear activation function ELU were also added. By performing a 3D convolution with a kernel size of (i, j, k), the value at the location (x, y, z) of the feature map F was determined from its neighbourhood:

$$F(x, y, z) = \sum_i \sum_j \sum_k \boldsymbol{W}(\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}) * \boldsymbol{X}(\boldsymbol{x} + \boldsymbol{i}, \boldsymbol{y} + \boldsymbol{j}, \boldsymbol{z} + \boldsymbol{k}) \tag{4.3}$$

with $-\left\lfloor \frac{I}{2} \right\rfloor \leq i \leq \left\lfloor \frac{I}{2} \right\rfloor$, $-\left\lfloor \frac{J}{2} \right\rfloor \leq j \leq \left\lfloor \frac{J}{2} \right\rfloor$, $-\left\lfloor \frac{K}{2} \right\rfloor \leq k \leq \left\lfloor \frac{K}{2} \right\rfloor$.

For hierarchical multimodal feature learning (HMFL), PET and CT feature maps were firstly concatenated to include multimodal context information at each scale of the convolutional layers. After concatenation, an adaptive pooling layer was used to project the

fused feature map into a single vector. This combination of feature maps from different scales could obtain both diverse texture details from shallow layers and high-level semantic layers, which can be defined as:

$$F_{fusion} = F(\bigcup_{l=1}^{L=4} APL(\boldsymbol{F_{pet}^{l}} \otimes \boldsymbol{F_{ct}^{l}})) \otimes F(\boldsymbol{F_{pet}^{5}} \otimes \boldsymbol{F_{ct}^{5}}) \quad (4.4)$$

Where APL denotes the adaptive max pooling layer, and L is the number of convolutional layers for multi-modal PET-CT feature extraction, and ($\otimes$) represents the concatenation operation.

Multimodal feature maps at each scale were concatenated into a single fully connected layer and processed with additional two fully connected layers. ReLU layers and dropout layers with a probability of 0.5 were added after each fully connected layer to reduce overfitting.

4.2.5 Implementation Details

Our method was implemented with PyTorch [110] and ran on an 11GB NVIDIA GeForce GTX 1080Ti GPU. The learning rate was set to 0.0001 and the batch size was set to 1. Our model was initialized using the approach presented in He et al. [111], and adaptive-moment-estimation (Adam) [117] was used for network optimization. I have further conducted an experiment where data augmentation techniques were adopted at the training stage, i.e., randomly rotation (90◦, 180◦, or 270◦) in the axial axis and randomly flip in one of all three axes (axial, sagittal and coronal), and this comparison experiment was named as CHMFL_Agumented. The training was terminated when no further changes in the total loss. In our method, the total loss was generally stable after 200 epochs and our CNN model took approximately five hours to fine-tune with. In addition, our model took around 10 seconds to inference 8 patients; this time is similar to the existing 3D based CNN models.

4.3 Experiments and Results

4.3.1 Experimental Setup

The following experiments were carried out compared our proposed method to:

1) state-of-the-art radiomics methods that were separated into 3 categories:
    a. Traditional radiomics: The method proposed in [12] was used, Multiresolution auto-correlation handcrafted (HC) clinical texture features were extracted. and included: the grey-level co-occurrence matrix (GLCM), grey-level run-length matrix (GLRLM), grey-level size zone matrix (GLSZM), and neighbourhood grey-tone difference matrix (NGTDM) features. A stepwise forward feature selection scheme was performed with multivariable analysis. The optimal conventional feature set contained 25 different radiomics features. A random forest (RF) classifier was trained with these texture features. This method was referred as HC+RF;
    b. CNN-based radiomics method. The method of [90] was reimplemented based on the technical details from their paper that used 2D image slices as the input data. This method has four main operational layers: 2D convolutional, non-linearity (PReLU), max-pooling, and fully connected layers for classification. This method was referred as CNLPC.
    c. Hybrid methods that combine CNNs with traditional radiomic components:
        i. The 3DMCL method proposed by [20], which had two branches to separately extract features from PET and CT volumes, and then deep features were combined with hand-crafted radiomics features and fed into fully connected layers to make a final prediction.

ii. A hybrid predictive model, comprised of a many-objective radiomics (MOR) model and a 3D CNN, which used spatial contextual information from PET-CT images that was designed by [21]. The output of the 2 components was fused through an evidential reasoning approach to predict lymph node metastases in head-and-neck cancers. This method was referred as MOR+3D CNNs.

2) Different imaging modalities and different CNN dimensions: All the PET-CT image slices containing tumour regions were used as the input for the 2D CNNs-based comparison methods. The PET-CT volumes were used, based on the bounding box, as the input for the 3D CNNs-based comparison methods. The PET-CT input slices were directly obtained from the 3D volumes that went through the same pre-processing steps in Section 2.1. The 2D and 3D CNNs used a similar architecture to the 3DMCL method that was suggested by [20] and the state-of-the-art method in this DM prediction problem for STS patients.
3) Individual components of the proposed method:
    a. CFL - our proposed method without HMFL module, used PET and CT images as input.
    b. Mask + HMFL - our proposed method without CFL module, used PET, CT, and tumour label images as input (2-channel PET-label image and 2-channel CT-label image).

The state-of-the-art methods used in our comparisons were those mentioned in the related works, and they can be divided into three categories: A holdout 6-fold cross-validation approach was used for our method and the comparison methods. The 48 PET-CT data were randomly divided into 6 equal-sized subsets and each subset had 8 PET-CT images. For each fold, 5 subsets were used to train the network and the remaining subset was used for testing. This process was repeated 6 times to assess the 48 PET-CT images. The results presented in Section III are the mean

value across all 6 folds.

4.3.2 Evaluation Metrics

The same evaluation metrics were used for comparison as in the previous chapter (Section 3.3.1), Six established evaluation metrics were adopted, including accuracy (Acc.), sensitivity (Sen.), specificity (Spe.), precision (Prec.), F1 score (F1), and area under the receiver-operating characteristic curve (AUC). The definitions of sensitivity, specificity, and precision are defined as in the previous chapter (see Equations 3.7-3.9 in Chapter 3.3.2). For all experimental comparisons with our proposed CHMFL method, the *p*-value with an unpaired student's *t*-test was computed. The accuracy and F1 score are defined as below:

$$Acc. = \frac{|TP|+|TN|}{|TP|+|FP|+|TN|+|FN|} \tag{4.5}$$

$$F1 = \frac{2\times \text{Prec.}\times \text{Sen.}}{(\text{Prec.}+\text{Sen.})} \tag{4.5}$$

Where TPs are the true positive pixels (ROIs), TNs are the true negative pixels (background), FPs are the false positive pixels and FNs are the false negative pixels.

4.3.3 Results

The CHMFL achieved the overall best DM prediction performance with the highest accuracy (0.854) and AUC (0.873) - see Table 4.2 and Figure 4.2; Our CHMFL's F1 score (0.857), specificity (0.833) and precision (0.840) ranked at the second place, and sensitivity (0.875) is the third best. In addition, our CFL module alone obtained the highest sensitivity (0.958), and our Mask + HMFL obtained the highest specificity (0.875) and precision (0.857).

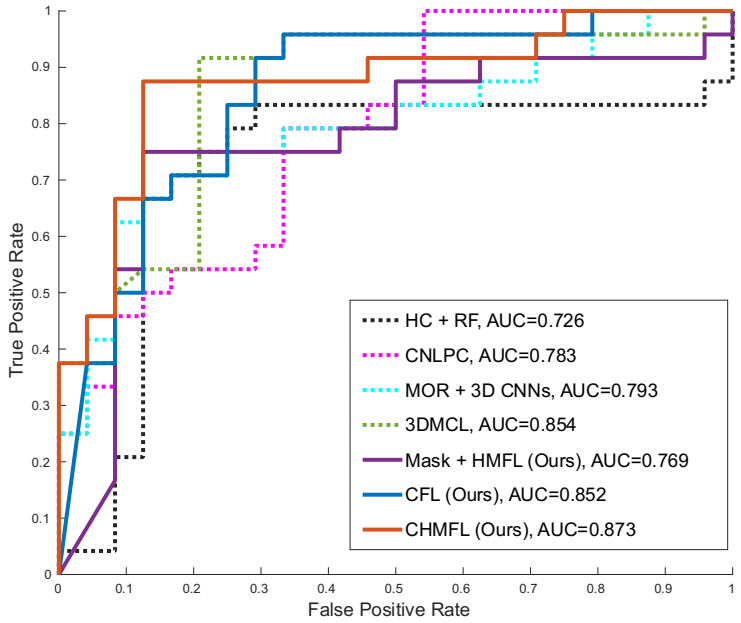


Figure 4.2. Classification performance (measured in receiver operating characteristic (ROC) curve) of our CHMFL in comparison to other existing radiomics methods.

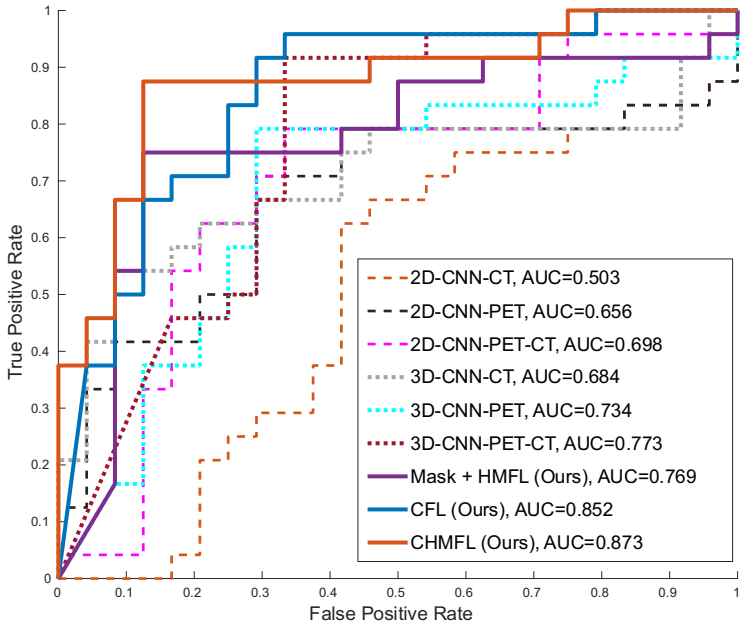


Figure 4.3. Classification performance (measured in ROC) of our CHMFL in comparison to methods using different modal image and convolutional layers.

When compared with methods using 2D or 3D CNNs with different modality imaging data

(e.g., PET, CT, PET-CT) – see Table 4.3 and Figure 4.3, our CHMFL method outperformed all the comparison CNNs based methods regardless of imaging modality and the kernel dimension of CNN. 3D CNNs performed better than those using 2D CNNs. Methods based on PET images outperformed methods based on CT images.

Table 4.2. Classification Performance Comparisons with Existing Radiomics Methods

Methods	Evaluation Metrics					
	Acc.	Sen.	Spe.	Prec.	F1	AUC
HC + RF [12]	0.750 (0.102) *	0.792 (0.125) *	(**0.105**) *	0.731 (0.094) *	0.760 (0.102) *	0.726 (0.126) *
CNLPC [90]	0.729 (0.094) *	0.792 (0.188) *	0.667 (0.130) *	0.703 (0.083) *	0.745 (0.109) *	0.783 (0.187) *
MOR + 3D CNNs [21]	0.729 (0.111) *	0.750 (0.224) *	0.780 (0.204) *	0.720 (0.112) *	0.750 (0.129) *	0.793 (0.174) *
3DMCL [20]	0.854 (0.085) *	0.917 (0.188) *	0.792 (0.187) *	0.815 (0.168) *	0.863 (0.094) *	0.854 (0.148) *
Our Methods						
Mask + HMFL	0.813 (0.094)	0.750 (0.129)	**0.875** (0.224)	**0.857** (0.174)	0.800 (0.094)	0.769 (0.139)
CFL	0.813 (0.102)	**0.958** (0.102)	0.667 (0.209)	0.742 (0.145)	0.836 (0.066)	0.852 (0.155)
CHMFL	0.854 (**0.051**)	0.875 (**0.102**)	0.833 (0.129)	0.840 (**0.074**)	0.857 (**0.034**)	0.873 (0.184)
CHMFL_ Augment	**0.896** (0.094)	**0.958** **(0.102)**	0.833 (0.204)	0.852 (0.142)	**0.902** (0.080)	**0.903** **(0.112)**

Table 4.3. Classification Performance Comparisons with Methods Using Different Modal Image and Convolutional Layers

Methods	Evaluation Metrics					
	Acc.	Sen.	Spe.	Prec.	F1	AUC
2D-CNN-CT	0.583 (0.102) *	0.708 (0.292) *	0.458 (0.292) *	0.567 (0.094) *	0.630 (0.155) *	0.503 (0.254) *
2D-CNN-PET	0.729 (0.051) *	0.542 (0.129) *	0.917 (0.209) *	**0.867** (0.174) *	0.667 (0.051) *	0.656 (0.139) *
2D-CNN-PET-CT	0.729 (0.094) *	0.792 (0.188) *	0.667 (0.130) *	0.703 (0.083) *	0.745 (0.109) *	0.698 (0.187) *
3D-CNN-CT	0.667 (0.085) *	0.667 (0.213) *	0.667 (0.258) *	0.667 (0.112) *	0.667 (0.131) *	0.684 (0.203) *
3D-CNN-PET	0.771 (0.094) *	0.750 (0.188) *	0.792 (0.224) *	0.783 (0.168) *	0.766 (0.094) *	0.734 (0.156) *
3D-CNN-PET-CT	0.792 (0.105) *	0.792 (0.209) *	0.792 (0.204) *	0.792 (0.143) *	0.792 (0.111) *	0.773 (0.148) *
Our Methods						
Mask + HMFL	0.813 (0.094)	0.750 (0.129)	**0.875** (0.224)	**0.857** (0.174)	0.800 (0.094)	0.769 (0.139)
CFL	0.813 (0.102)	**0.958** (0.102)	0.667 (0.209)	0.742 (0.145)	0.836 (0.066)	0.852 (0.155)
CHMFL	0.854 (**0.051**)	0.875 (**0.102**)	0.833 (0.129)	0.840 (**0.074**)	0.857 (**0.034**)	0.873 (0.184)
CHMFL_ Augment	**0.896** (0.094)	**0.958** **(0.102)**	0.833 (0.204)	0.852 (0.142)	**0.902** (0.080)	**0.903** **(0.112)**

*: $p < 0.05$, in comparison to our proposed CHMFL method derived from an unpaired student's t-test.
The results are presented in the form of 'mean value (standard deviation)'.

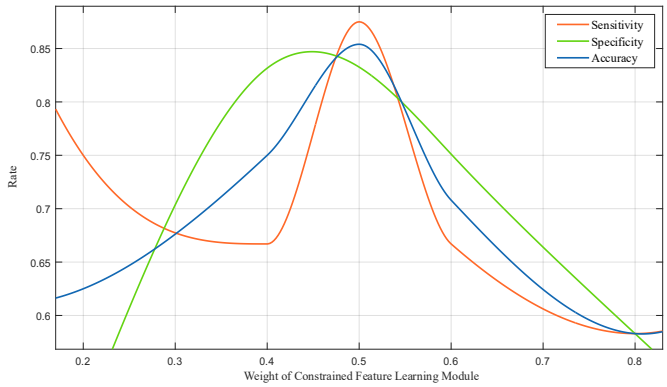


Figure 4.4. Analysis of the weight used in the CFL module.

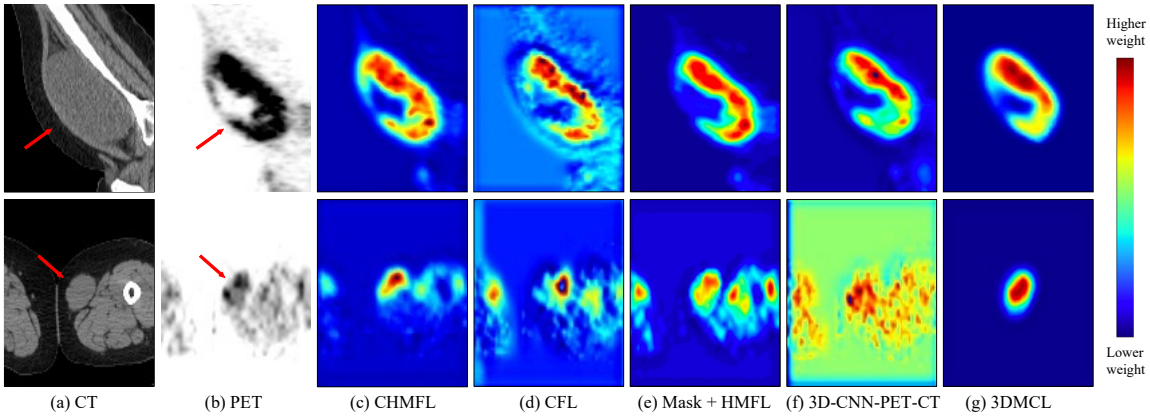


Figure 4.5. The original CT (a) CT and PET (b) PET images of STSs in the calf (top row) and thigh (bottom row) and the feature map visualizations from our approach (CHMFL) and the other approaches. Images were cropped to the tumour ROI and the red arrows indicate tumour regions. Blue in the feature map visualizations indicates low weight, whereas yellow and red indicate higher weights.

I evaluated how the CFL module's weight w affected the performance of CHMFL over three key evaluation metrics (e.g., accuracy, sensitivity and specificity). The result in Figure 4.4 suggests that the best performance was achieved when the weight was 0.5.

Two example PET-CT studies are shown in Figure 4.5 with corresponding visualization results of the extracted feature map with respect to three existing radiomics methods that outperformed other comparison methods except for our CFL and CHMFL methods. In Figure 4.5,

class activation maps [118] were generated to visualize the predicted class scores on any given image. Global average pooling layers were implemented to capture the spatial average of the feature map of each unit at the last convolutional layer. A channel-wise sum of these feature maps was used to generate the final output (i.e., class activation maps). The discriminative object parts detected by the CNNs were highlighted.

The discriminative ability of both the 3DMCL and our CFL and CHMFL methods is depicted in Figure 4.6, via t-distributed stochastic neighbourhood embedding (t-SNE) [119] visualization. t-SNE is an unsupervised, non-linear technique primarily used for visualizing high-dimensional image features in a two or three-dimensional space, which allows for exploring the relationship of the extracted features.

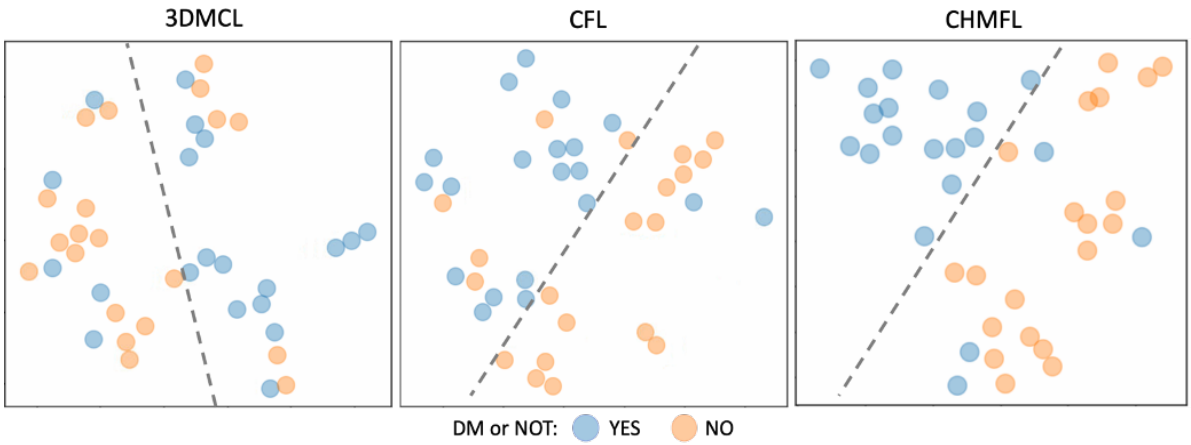


Figure 4.6 t-SNE visualization result of 3DMCL, CFL and CHMFL methods. A dashed line is added to demonstrate how the features are separated.

4.4 Discussions

The main findings are that: (i) our CFL module automatically identified the tumour and deep features learned from tumour segmentation could be useful for patient outcome prediction; (ii) our HMFL module derived multimodal PET-CT image features; (iii) our method improved upon current single-modality methods and, (iv) our CHMFL outperformed state-of-the-art radiomics

methods.

4.4.1 CFL Module Analysis

The CFL automatically focused on ROIs that were semantically more important to the STSs DM predictions. When compared to 3D-CNN-PET-CT (Figure 4.5 (f)), both our CFL and CHMFL methods with the CFL module were able to correctly focus on tumour regions in the derived feature maps (Figure 4.5(c) and Figure 4.5(d)). In contrast, 3D-CNN-PET-CT falsely concentrated on many normal uptake regions, which resulted in a >10% decrease in AUC value and > 4% decrease in both accuracy and F1 score (shown in Table II). Moreover, when compared to existing methods that segmented tumour region before feature extraction and outcome prediction, such as 3DMCL (the state-of-the-art method for STS DM prediction, Figure 4.5 (e)), both our CHMFL and CFL methods could accurately identify the entire sarcoma with more details. Although Mask + HMFL was forced to focus on the tumour region by incorporating an extra channel of tumour label image as input, there were still some false positive regions, e.g., as in the bottom case of Figure 4.5 (e) when there were similar tissues around the tumour. Without our CFL module, the Mask + HMFL method had a tendency of not predicting DM due to concentrating on more regions other than tumours. Although this contributed to a 4% increase in specificity and 1% increase in precision, there are >10% decrease in both AUC score and sensitivity when compared with our proposed CHMFL (shown in Table 4.2). It was also noted that automatically constraining the learning process to extract feature maps only from the tumour region can obtain more information that can potentially reflect underlying pathophysiology, such as the heterogeneity of STS, which is an important prognostic factor of DM development [120]. In addition, such an automated process removes the reliance on accurate manual tumour delineation during the inference stage while obtaining better overall performance.

4.4.2 HMFL Module Analysis

The inclusion of HMFL in our method further improved the performance of the CFL. Most existing CNNs based radiomics methods, including 3DMCL, CNLPC and MOR+3DCNNs, only leverage high-level features extracted from the last convolutional layer in their model, and therefore inherently disregarded the complementary PET and CT image features at the lower level of the network. In contrast, our method iteratively and hierarchically fused the multimodal PET and CT image features across the different image scales, which enabled more flexible and complex multimodal information fusion. As an example, the feature map derived from our CHMFL method (Figure 4.5 (c)) captured more details inside the tumour better predicted the tumour contour when compared with our CFL method (Figure 4.5 (d)).

4.4.3 Evaluation of CNN-Based Methods with Different Image Modalities and Different Convolutional Layers

There was a marked difference in performance between PET-CT CNNs and CNNs with PET alone or CT alone. Further, PET-based methods outperformed CT-based methods. This was expected since PET images provided metabolism information of tumours, while CT can only provide the anatomical information, and tumour regions are not always visible in CT. The relatively lower performance of 2D CNNs, when compared to 3D CNNs counterparts is attributed to the fact that volumetric image features derived from 3D CNNs are better to discriminate the spatial information within the tumour that is associated with the DM development, e.g., volumetric tumour shape and size [16]. In contrast, 2D CNNs based methods (e.g., 2D-CNN-CT and 2D-CNN-PET) have limited representation capability of tumour characteristics in two dimensions with few axial slices. Therefore, it would be better to incorporate 3D CNNs with multimodal imaging data when the computational power is available, which allows achieving better performance (as shown in Table 4.2.

4.4.4 Comparison of CHMFL with Existing Methods

Our CHMFL method obtained the best overall performance when compared with the existing radiomics methods. HC+RF method achieved competitive performance over all the evaluation metrics except the AUC score when compared with CNLPC and MOR+3DCNNs. Unfortunately, the performance of HC+RF was reliant on effective feature handcrafting and tuning a large number of parameters, which may limit its generalizability to different datasets. The performance improvement from 3DMCL to CNLPC and MOR+3DCNNs was likely due to the use of multimodal PET and CT images providing complementary information. When compared to the second-best performing method 3DMCL, our method achieved much higher specificity (as shown in Table I). 3DMCL is reliant on using single-level image features for prediction, which results in 3DMCL overfits to the positive prediction of DM. which were unable to discriminate the tumours. As exemplified in Figure 4.6 our CHMFL had greater separability between the patients with/without DM than both 3DMCL and 3D-CNN-PET-CT, where only a few cases were not properly separated.

4.4.5 Limitations and Future Work

Our focus in the current study was to investigate the prediction of distant tumour spread (metastatic disease) in patients with STSs from PET-CT images. Predicting the presence of distant metastases (DM) as a binary classification is an abstraction of a time to event prediction problem (i.e., estimating the point at which an event occurs). The time to event problem is a more complicated modelling challenge than binary classification and may require different methodological approaches. In the public dataset used in this chapter, all the patients have a 7-year follow-up period for outcome observation and DM was generally confirmed within 4 years after diagnosis of primary STS; this was appropriate for binary classification. The public dataset

is small (n=51) and thus there was no separate held-out data used only for testing. Only the mean results across all validation experiments were reported. The results may be different with a held-out cohort in a much larger dataset. I have been actively working on characterizing and annotating a much larger soft tissue sarcoma dataset with my colleagues in the research group. Moreover, the results have not been generalized to other tumour types or where other imaging modalities are employed. In future work I intend to evaluate our approach in non-small cell lung cancer and lymphomas, using PET-CT, and also include other parameters such as local tumour recurrences and long-term survival. In lymphomas there are generally multiple sites of disease and disease recurrence occurs unpredictably and so analysing multiple lesions will be necessary to attempt to predict where the disease will occur. I would like to adapt our approach to such a situation, and this will require multiple bounding boxes.

4.5 Summary

In this chapter, a constrained hierarchical multimodal feature learning method was proposed for radiomics with multimodal PET-CT images. The proposed method was evaluated in predicting the development of distant metastases. The experimental results on a well-established public dataset of STSs showed that our method was capable of better identifying PET-CT radiomics features in primary tumours that were associated with the development of DM, when compared to the state-of-the-art radiomics methods.

Chapter 5. Automated Multimodal Information Fusion for Radiomics via Neural Architecture Search

In this chapter, a multimodal neural architecture search (MM-NAS) method is introduced for multimodal PET-CT information fusion for radiomics analysis. Radiomics methods based on convolutional neural networks (CNN) are regarded as the state-of-the-art because they can learn high-level semantic image information in an end-to-end fashion. However, majority of existing CNN-based radiomics methods were designed for single-modality images such as CT and MRI [18], [89], [90]. For the few methods that attempted to fuse multimodal images, the focus was on fusing the image features that were separately extracted from the individual modalities [19], [20], [88]. In addition, these methods required human expertise to design the dataset specific architectures e.g., the number of convolutional layers, the layer to fuse multimodal image features. Architecture design and optimization require a large amount of domain knowledge such as in validating the architecture performance and tuning the hyperparameters. NAS has recently been proposed to simplify the challenges in architecture

design by automatically searching for an optimal network architecture based on a given dataset; the outputs from the NAS can then be further optimized where necessary. The NAS thus enables reduced manual input and reliance on prior knowledge [26]. Investigators have attempted to apply the NAS for single-modal medical image related tasks, with the main focus on image segmentation [27], [28].

In this chapter, a new NAS is designed to automatically derive optimal multimodal image features for radiomics studies, including tumour segmentation and patient outcome prediction. In contrast to existing radiomics methods, the MM-NAS simplifies the challenges in architecture design by automatically searching for an optimal network architecture based on a given dataset. The MM-NAS method was designed and evaluated in two well-established applications: (i) prediction of distant metastases (DM) development; and (ii) tumour segmentation, using a well-established benchmark soft-tissue sarcomas (STSs) PET-CT dataset. The experimental results demonstrate that MM-NAS achieved overall better performance when compared to the state-of-the-art NAS methods.

5.1 Contributions

The main contributions of this chapter are as follows:

1) An iterative bi-level optimisation strategy is proposed to automatically search for a suitable CNN architecture for multimodal PET-CT images for radiomics studies. In contrast to existing radiomics methods using CNNs that are manually pre-designed with fixed architectures, the MM-NAS enables reduced manual input and reliance on prior knowledge by automatically building an optimal radiomics CNN architecture based on a given dataset.
2) Different computational cells are introduced as the basic unit that can be stacked multiple times to form a CNN. The computational cells share similar functionalities

with convolutional blocks in CNNs, where the cells used for down-sampling are referred to as reduction cells, while the normal cells do not change the input size. When compared to existing NAS methods in medical image analysis that directly search for the entire network architecture, searching for the best cell structure is found to be much more efficient and the cell itself is more likely to generalise to other problems, such as tumour segmentation and patient outcome prediction.

3) The MM-NAS also enables optimal fusion of PET-CT images for radiomics by searching for various fusion modules via different network operations (e.g., convolution, pooling, etc.) at different stages of the network. These searched fusion modules provide greater flexibility for integrating complementary PET and CT data.

This chapter's contributions address the challenge in existing deep learning-based methods that are reliant on human expertise to design dataset-specific CNN architectures. This challenge is referred to in Chapter 1 and aligns with the gap that automated CNN with minimum manual input and specialised skillsets are necessary for radiomics with PET-CT images as defined in Chapter 2. It further expands on the literature in Chapter 2.4 by including detailed comparisons to the state-of-the-art comparative methods.

5.2 Materials and Methods

5.2.1 Materials and Pre-processing

A well-benchmarked public PET-CT STSs dataset from the Cancer Imaging Archive is used for method evaluation [7], [97] (see Chapter 3.2.1 Materials and Pre-processing for details).

5.2.2 Overview of the Proposed Method

The MM-NAS was inspired by several existing NAS methods [121]–[123], which are designed

to enhance the computational efficiency of CNNs by searching for optimal different computational cells (e.g., normal cells and reduction cells). Computational cells are the basic building units of CNNs, which can be stacked multiple times to form a powerful architecture. To achieve this, the MM-NAS workflow is as follows: (i) searching for the optimal cell structure of the encoder based on the given training dataset, and (ii) training the searched CNN on the training dataset and then evaluating it on the testing dataset.

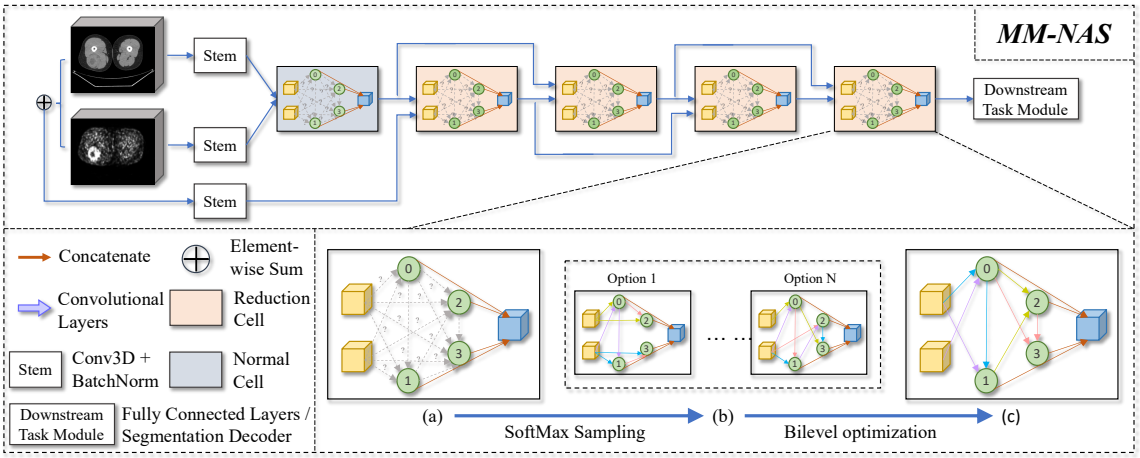


Figure 5.1. MM-NAS overview – the CNN architecture has multiple different cells (normal, reduction); each cell is a directed acyclic graph as the basic unit; directed arrows indicate the forward path: (a) initial operations on the edges of each cell are unknown; (b) continuous production of alternative cells by SoftMax sampling; and (c) optimal cell architecture after iterative bi-level optimization.

5.2.3 Search Space

In the MM-NAS (as shown in Figure 5.1), every cell is viewed as a directed acyclic graph composed of two inputs. The input and output of a normal cell have the same dimensions. However, the reduction cell doubles the channel number and reduces the input feature map by half. The reduction cell and normal cell are the main types of cells used in the MM-NAS.

Within each cell, there are several directional graph nodes with one output node. Additionally, a stem block that consists of a 3D convolutional layer and a batch normalisation layer was incorporated to facilitate input image transitions. In the proposed method, the outputs of PET and CT stem blocks are separately fed into the first normal cell to facilitate the fusion process. The output feature maps of the first normal cell then flow into the first reduction cell with the sum of PET and CT images, which is also processed by one stem block. The remaining reduction cells use the output feature maps from the previous two layers as input.

Each intermediate node n^i inside a cell is a latent representation (e.g., a feature map in CNNs). The searched operations on edge (i,j) are represented using the vector $\boldsymbol{x}^{(i,j)}(\boldsymbol{n}^j) = \left\{x_\sigma^{(i,j)} \middle| \sigma \in O\right\}$ and the vector of all optional operations as $\mathbf{O}^{(i,j)} = \left\{\sigma\left(n^i; \vartheta_\sigma^{(i,j)}\right) \middle| \sigma \in O\right\}$, where O denotes the set of optional operations, $\vartheta_\sigma^{(i,j)}$ denotes the parameters of the operation σ on edge (i,j). Then the intermediate nodes can be computed by the sum of all their predecessors:

$$n^j = \langle \boldsymbol{x}^{(i,j)}(\boldsymbol{n}^j), \mathbf{O}^{(i,j)} \rangle = \sum_{i<j, \sigma \in O} \boldsymbol{x}^{(i,j)}(\boldsymbol{n}^j)\sigma\left(n^i; \vartheta_\sigma^{(i,j)}\right) \tag{5.1}$$

A special zero operation is also included to indicate a lack of connection between two nodes. The task of learning the cell, therefore, reduces to learning the operations on its edges.

5.2.4 Optimization Strategy

As all the possible operations are mixed through a SoftMax function, this makes the search space continuous:

$$\boldsymbol{x}^{(i,j)}(\boldsymbol{n}^j) = \sum_{\sigma \in O} \frac{\exp\left(\alpha_\sigma^{(i,j)}\right)}{\sum_{\sigma' \in O} \exp\left(\alpha_{\sigma'}^{(i,j)}\right)} \boldsymbol{x}_{\boldsymbol{\sigma}}^{(i,j)} \tag{5.2}$$

Where $\alpha_{\sigma}^{(i,j)}$ denotes a probability distribution over the operation set O. With the continuous search space, the searching goal is to jointly learn the architecture α and the weights θ within all the mixed operations (e.g., weights of the convolution filters). Similar to architecture search using reinforcement learning [121], [123] or evolution algorithm [124] where the performance of validation dataset is treated as the reward or fitness, the MM-NAS aims to optimise the validation loss, but using gradient descent.

Denote by L_{train} and L_{val} the training and the validation loss. Because both losses are determined not only by the architecture α , but also by the weights $\boldsymbol{\theta}$ in the network, where $\theta = \{(\vartheta^{(i,j)})|(i,j) \in C\}$, C is the computational cell. The aim of searching for the best architecture is to find a proper α that minimizes the validation loss $L_{val}(\theta^*(\alpha), \alpha)$, where the weights θ associated with the architecture are obtained by minimizing the training loss:

$$\min_{\alpha} L_{val}(\boldsymbol{\theta}^*(\boldsymbol{\alpha}), \alpha) \tag{5.3}$$

$$s.t.\ \boldsymbol{\theta}^*(\boldsymbol{\alpha}) = \operatorname{argmin}_{\theta} L_{train}(\boldsymbol{\theta}, \alpha) \tag{5.4}$$

The nested formulation also emerges in hyperparameter optimization based on gradient [125], [126]. This is related in the sense that the architecture α can be seen as a distinctive type of hyperparameter, despite its dimension being significantly higher than scalar-valued hyperparameters, such as the learning rate. As a result, it is more challenging to optimise.

Evaluating the architecture gradient exactly can be prohibitive due to the expensive inner optimization. We, therefore, propose a simple approximation scheme as follows:

$$\nabla_{\alpha} L_{val}(\boldsymbol{\theta}^{*}(\boldsymbol{\alpha}), \alpha) \approx \nabla_{\alpha} L_{val}(\boldsymbol{\theta} - \delta \nabla_{\theta} L_{train}(\boldsymbol{\theta}, \alpha), \alpha) \tag{5.5}$$

where $\boldsymbol{\theta}$ denotes the current weights maintained by the algorithm, and δ is the learning rate for a step of inner optimization. The idea is to approximate $\boldsymbol{\theta}^{*}(\boldsymbol{\alpha})$ by adapting θ using only a single training step, without solving the inner optimization (equation 5.4) completely by training until convergence. Related techniques have been used in gradient-based hyperparameter tuning [127] and unrolled generative adversarial networks [128]. Note equation 5.5 will reduce to $\nabla_{\alpha} L_{val}(\theta, \alpha)$ if θ is already a local optimum for the inner optimization and thus $\nabla_{\alpha} L_{train}(\theta, \alpha) = 0$.

5.2.5 Implementation Details

The MM-NAS was implemented using the PyTorch framework [110]. The input image size was fixed to 112 × 112 × 144. The operation set $\mathbf{O}^{(i,j)}$ for each cell included 3D standard convolutions, 3D separable convolutions, 3D dilated convolutions, 3D max pooling, 3D average pooling, skip connections and zero operations. All operations were of stride one (if applicable), and the kernel size of pooling operations was 3, while the kernel size for the convolutional operations was either set to 3 or 5. During the architecture search step, the cross-entropy loss was used for training optimization. The parameters of each cell were optimized by Adam optimizer with a learning rate of 0.0005 while the weight in the whole network was optimized by SGD with a learning rate of 0.0001, and the batch size was set to 1.

In order to evaluate the MM-NAS in both predicting DM and segmenting tumours of STSs, there were modifications for task-specific implementations. For DM prediction, the output feature maps of the last reduction cell were fed into two convolutional layers and one fully connected layer for classification. As for the task of tumour segmentation, the common encoder-decoder structure for segmentation tasks was utilised where only the encoder

architecture was searched while the decoder was directly taken from a standard U-Net. The encoder architecture consisted of three branches, each of which shared the exact same structure as shown in Figure 5.1. These branches were designed to be automatically searched for learning and utilising the PET, CT and concatenated PET-CT images, respectively.

With the 40 PET-CT volumetric training images searching the architecture for the task of DM prediction, it took approximately 3 minutes to process one epoch, and the best architecture was obtained at epoch 70 out of the total 200 epochs. As for training the searched architecture, cross-entropy loss with Adam was used for training optimization in the second step. The learning rate was set to 0.001 and batch size was set to 1, and it took around 2 minutes to train one epoch, the best model was obtained at approximately epoch 80 out of 200 epochs.

As for the task of tumour segmentation using the same volumetric data, it took approximately 8 minutes to run one epoch, and the best architecture was obtained at epoch 120 out of the total 200 epochs. As for training the searched architecture, weighted cross-entropy loss with Adam was used for training optimization in the second step. The learning rate was set to 0.001 and batch size was set to 1, and it took around 4 minutes to train one epoch, the best model was obtained at approximate epoch 70 out of 200 epochs. All the experiments were conducted on an 11GB NVIDIA GeForce GTX 2080Ti GPU.

5.3 Experiments and Results

5.3.1 Experimental Setup

The following comparison experiments were conducted for DM prediction:

1) a comparison with the state-of-the-art radiomics methods:
    a. HC+RF – I followed the conventional radiomics method and used hand-crafted (HC) features (e.g. intensity solidity, skewness, grey-level co-occurrence matrix

features, etc.) extracted from tumour region with random forest (RF) as the classifier for prediction [12];

b. DLHN – a deep learning based radiomics method using a 2D CNN to predict head & neck cancer outcomes (e.g., DM, loco-regional failure, and overall survival) [90];

c. 3DMCL – a deep learning based multimodal collaborative learning method using 3D CNN for distant metastases prediction with PET-CT images [20].

2) compared the performance of using multimodal CNNs to single-modality CNNs.

3) compared the performance of using 2D CNNs with 3D CNNs for radiomics.

Similarly, the following comparison experiments were conducted for tumour segmentation:

1) a comparison with the state-of-the-art radiomics methods:

a. Co-learning – a two-branched U-Net was designed to extract and fuse PET and CT information with spatial context across multiple CNN blocks [63];

b. NAS-Unet – a U-like backbone network with a differential architecture strategy, which contained three types of primitive operation set on search space to automatically find two cell architecture DownSC and UpSC for semantic image segmentation [129];

c. V-NAS – a differentiable neural architecture search method for volumetric medical image segmentation, the network itself chose between 2D, 3D or Pseudo-3D convolutions at each layer. [130].

d. MM-MRI-NAS – A brain tumour segmentation method designed for volumetric multimodal MRI images, patching strategies were utilised for the input data, and there are two types of cells to be automatically searched (downward cell and upward cell) [131].

2) compared the performance of using multimodal CNNs to single-modality CNNs.

3) compared the performance of using 2D CNNs with 3D CNNs for tumour segmentation.

A 6-fold cross-validation approach was used for the MM-NAS and the comparison methods. In each-fold cross-validation, 40 PET-CT images were used for training and the remaining 8 images were used for testing.

5.3.2 Evaluation Metrics

The same evaluation metrics for comparison were used as in the previous chapter (Chapter 4.3.2), including accuracy (Acc.), sensitivity (Sen.), specificity (Spe.), precision (Prec.), F1 score (F1), and area under the receiver-operating characteristic curve (AUC).

Moreover, five commonly used evaluation metrics were adopted for the tumour segmentation, including Dice score, intersection over union (IoU), sensitivity (Sen.), specificity (Spe.) and accuracy (Acc.), most of which are also introduced in Chapter 3.3.2. The IoU is defined as follows:

$$IoU = \frac{|GT \cap PS|}{|GT \cup PS|} \tag{5.6}$$

Where GT denotes the ground truth, PS is the algorithm predicted segmentation result.

5.3.3 Results

The receiver-operating characteristic (ROC) curve is shown in Figure 5.2. It shows that our 2D MM-NAS achieved better performance in DM prediction when compared with 2D CNN-based methods. Our 3D MM-NAS outperformed other 3D CNN-based comparison methods and achieved the overall best performance.

Table 5.1. Comparisons with Existing Radiomics Methods on DM Prediction

Methods	Evaluation Metrics					
	Acc.	Sen.	Spe.	Prec.	F1	AUC
HC+RF [12]	0.750	0.792	0.708	0.731	0.760	0.726
DLHN [90]	0.729	0.792	0.667	0.703	0.745	0.698
3DMCL [20]	0.854	0.917	0.792	0.815	0.863	0.854
2D MM-NAS (Ours)	0.750	0.833	0.667	0.714	0.769	0.711
3D MM-NAS (Ours)	**0.896**	**0.917**	**0.875**	**0.880**	**0.898**	**0.896**

Table 5.2. Comparison of Methods using Different Imaging Modalities with Convolutional Kernels for DM Prediction

Methods	Evaluation Metrics					
	Acc.	Sen.	Spe.	Prec.	F1	AUC
2D CT CNN	0.583	0.708	0.458	0.567	0.630	0.503
2D PET CNN	0.729	0.542	**0.917**	**0.867**	0.667	0.656
2D PET-CT CNN	0.729	0.792	0.667	0.703	0.745	0.698
2D MM-NAS (Ours)	**0.750**	**0.833**	0.667	0.714	**0.769**	**0.711**
3D CT CNN	0.667	0.667	0.667	0.667	0.667	0.684
3D PET CNN	0.771	0.750	0.792	0.783	0.766	0.734
3D PET-CT CNN	0.792	0.792	0.792	0.792	0.792	0.773
3D MM-NAS (Ours)	**0.896**	**0.917**	**0.875**	**0.880**	**0.898**	**0.896**

Table 5.1 and Table 5.2 present results of 3D MM-NAS achieving the best outcomes for DM prediction in all measures with an AUC value of 0.896, an accuracy of 0.896, a sensitivity of 0.917, a specificity of 0.875, a precision of 0.880, and an F1 score of 0. 898.

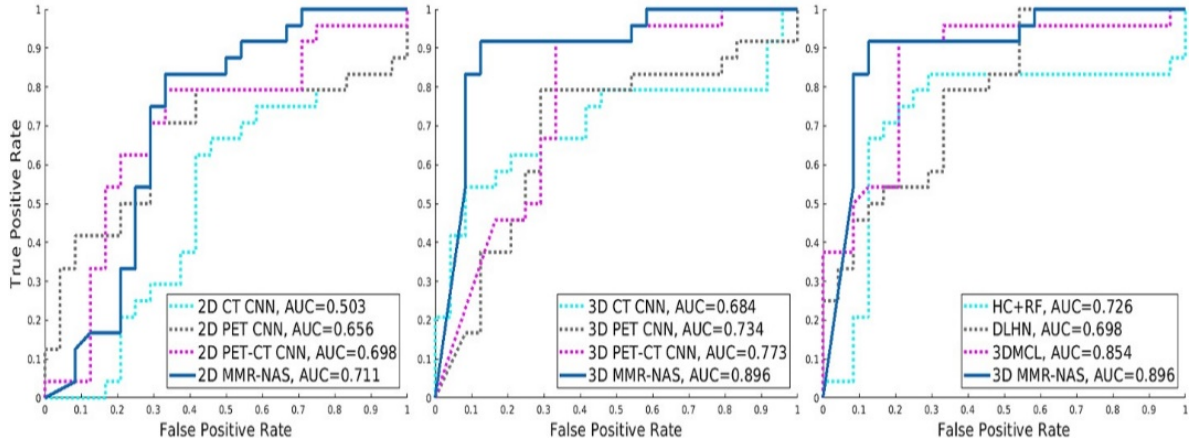


Figure 5.2. ROC curves of ours and comparative radiomics methods.

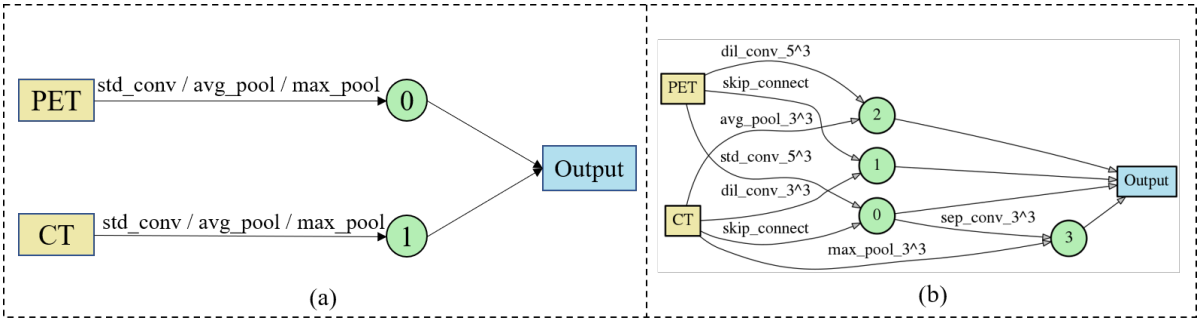


Figure 5.3. The comparison between (a) the simplified fusion approach of the DLHN and the 3DMCL for DM prediction; (b) the learned normal cell of the MM-NAS for PET-CT fusion.

Table 5.3. Comparison of the state-of-the-art methods on STSs segmentation

Methods	Evaluation Metrics				
	Dice	IoU	Sen.	Spe.	Acc.
Co-learning [63]	0.616	0.501	0.697	**0.996**	**0.992**
NAS-Unet [129]	0.532	0.409	0.670	0.994	0.989
V-NAS [130]	0.529	0.390	0.678	0.992	0.990
MM-MRI-NAS [131]	0.526	0.372	**0.856**	0.990	0.988
2D MM-NAS (Ours)	**0.621**	**0.531**	0.746	0.994	0.991

Table 5.3 and Table 5.4 presents results of 2D MM-NAS which resulted in the best performance for tumour segmentation in all measures with the highest dice score of 0.621, and the highest IoU of 0.531, accuracy (0.991), sensitivity (0.746), and specificity (0.994) are second-

best among the comparison methods. Figure 5.3 presents two examples of PET-CT studies with STS which shows that the MM-NAS delineated the tumour region with a better result when compared to other existing NAS methods.

Figure 5.3 is a comparison between the simplified fusion approach of common radiomics methods and one example of the normal cell learned via MM-NAS for multimodal PET-CT fusion, which enables flexible fusion with more options for different operations.

Table 5.4. Comparison of Methods using Different Imaging Modalities with Convolutional Kernels for STSs segmentation.

Methods	Evaluation Metrics				
	Dice	IoU	Sen.	Spe.	Acc.
2D CT V-Net [107]	0.376	0.269	0.568	0.983	0.978
2D CT U-Net [85]	0.465	0.364	0.554	0.993	0.990
2D PET U-Net	0.607	0.492	**0.757**	0.994	0.990
2D PET-concat-CT U-Net	0.608	0.493	0.707	**0.996**	**0.992**
2D PET-sum-CT U-Net	0.603	0.484	0.748	0.994	0.991
2D MM-NAS (Ours)	**0.621**	**0.531**	0.746	0.994	0.991
3D CT V-Net [107]	0.416	0.268	0.514	0.980	0.970
3D CT U-Net [85]	0.452	0.310	0.485	0.986	0.976
3D PET U-Net	0.546	0.400	0.626	0.986	0.980
3D PET-concat-CT U-Net	0.521	0.392	0.680	0.994	0.992
3D PET-sum-CT U-Net	0.555	0.408	0.606	0.988	0.980
3D MM-NAS (Ours)	**0.604**	**0.457**	**0.628**	**0.988**	**0.982**

5.4 Discussions

The main findings are that the MM-NAS: (i) performed better than the commonly used radiomics methods for DM prediction and tumour segmentation, (ii) derived optimal multimodal radiomics features from PET-CT images and, (iii) removed the reliance on prior knowledge when building

the optimal CNN architecture.

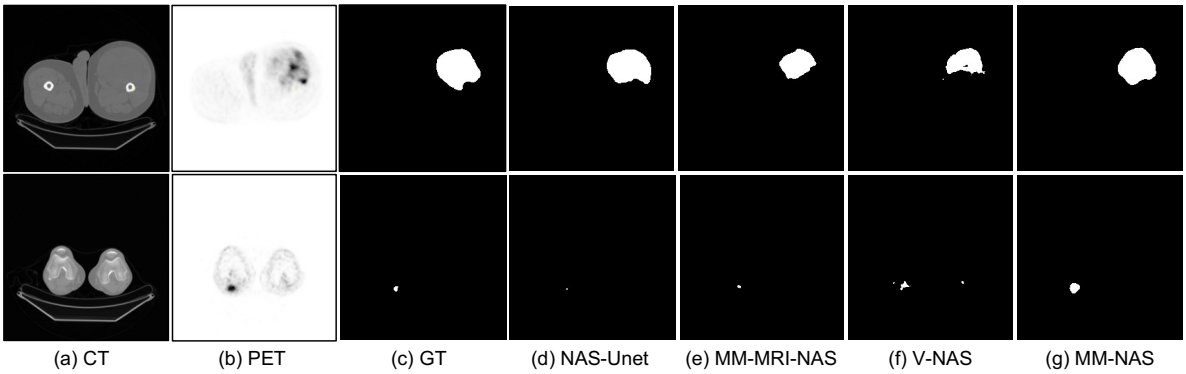


Figure 5.4. Two examples of PET-CT studies with STS are shown on axial image slices with (a) CT in the first column, (b) PET in the second column, and (c) ground truth (GT) in the third column. The segmentation results from the different state-of-the-art NAS methods are presented in columns (d) to (f), and our MM-NAS is (g).

5.4.1 Comparison to Existing Methods

The improved performance of the MM-NAS is attributed to the search for the optimal computation cells, within the NAS, that allowed for fusing multimodal image features at different stages of the network. Existing approaches often choose to fuse the separately extracted feature maps (after several convolutional / pooling layers). The MM-NAS derived cell structure offers more freedom to integrate multimodal images via various operations and connections (see Figure 5.4), thus producing the optimal radiomics features to predict distant disease. The state-of-the-art method of 3DMCL outperformed HC+RF and DLHN due to the collaborative learning of both pre-defined radiomics features and deep features, whereas the MM-NAS obtained better performance over all the evaluation metrics without feature handcrafting. Thus, the elimination of prior knowledge could contribute to better generalizability for applications in other radiomics studies.

For the segmentation task, the NAS-Unet and V-NAS are designed for single-modality

medical images only, which are not suitable for multimodal PET-CT images. This resulted in poorer performance than the MM-NAS and Co-learning (see Table 5.3). Co-learning improved over the single-modality NAS methods via the fusion of anatomical and functional visual features from PET-CT images, but the fusion approach from Co-learning was not designed to capture fine details that are more critical in tumour segmentation, which results in false negative segmentation errors. Although the MM-MRI-NAS was designed for multi-modal medical imaging data, the patch strategy loses global and spatial information of the tumour region regarding the comprehensive image, especially when some STSs are larger than the patch size that is designed for the brain tumours in MRI images. This also resulted in the MM-MRI-NAS to under-segment the tumour regions without sufficient spatial information (see Figure 5.4).

5.4.2 Comparison among Single-Modality and Multimodal PET-CT Images

The differences between PET-CT CNN and CNN with PET or CT alone show the advantage of incorporating multimodal information. Across the single-modality CNNs, PET-based methods outperformed CT-based methods (see Table 5.2). This was ascribed to the functional features, which can better characterize the tumour, when compared to anatomical features from CT that rely on changes in size which are often a later development. Such features from PET could potentially uncover functional information that relates to the biological behaviour of tumours [2].

Similar performance is also presented on the task of STSs segmentation (see Table 5.4), methods using PET outperform those using CT images only, and the performance is consistent within comparison methods using 2D and 3D CNNs. Moreover, the methods using concatenated PET-CT images or element-wise fused PET-CT images show better performance than all the other methods only using single modality data.

5.4.3 Analysis of Methods using 2D or 3D CNNs

When compared to 3D CNNs, the relatively poor performance of 2D CNNs in DM prediction is expected (see Table 5.2). This is attributed to the fact that volumetric image features derived from 3D CNNs are better able to derive spatial information e.g., volumetric tumour shape and size. Spatial information has strong correlations to DM predictions [16].

However, the 3D CNNs obtained slightly worse results on the task of STSs segmentation, this is due to the fact that, given the same computational resources as the patient outcome prediction, fewer nodes are designed in each cell and fewer operations are allowed within each node for searching the optimal architecture for STSs segmentation. The Dice score of 3D MM-NAS is only 0.018 lower than that of 2D MM-NAS (see Table 5.4). If the same search space is provided with more computational resources, the 3D MM-NAS has great potential in outperforming the 2D version.

5.4.4 Limitations and Future Work

The focus in the current study was to investigate the automated ways of fusing multimodal PET-CT information in radiomics. But only one small public dataset of 51 patients was used, and the generalisability of the MM-NAS in different cancers has not been investigated, which could be for future research direction. In future work, if there will be more computational resources, I would like to improve the segmentation results of the 3D MM-NAS, further upgrading the NAS methods with more flexibility and robustness. Moreover, I intend to evaluate the MM-NAS in non-small cell lung cancer and lymphomas, using multimodal PET-CT images, and also attempt to predict various outcomes of cancer patients, such as local tumour recurrences and where the disease will occur.

5.5 Summary

In this chapter, a multimodal neural architecture search method (MM-NAS) was proposed for

multimodal PET-CT images in radiomics studies. The MM-NAS method automatically searches for a CNN architecture which can then be used to fuse and derive optimal PET-CT image features for multimodal radiomics studies. This enabled reduced prior knowledge and minimum manual input for existing CNN-based methods. The experimental results on a well-established public dataset of STSs showed that the automatically generated PET-CT image features are the most relevant for DM prediction and are capable of accurately segmenting the tumour regions.

Chapter 6. Conclusions and Future Work

6.1 Conclusions

Radiomics has become an important prognostic tool in cancer management within the realm of modern healthcare, and deep learning-based methods are regarded as the state-of-the-art in this field. This thesis has addressed several challenges and limitations associated with the application of deep learning techniques to multimodal PET-CT images, and has proposed solutions to enhance its practicality and effectiveness. The following innovative methods were presented in this thesis:

1. A new Self-supervised enabled False positive and False negative Reduction Network (SFRN) for tumour segmentation in multimodal PET-CT images (Chapter 3). This addresses the challenge of automated tumour segmentation in radiomics where robust methodologies are anticipated for multimodal PET-CT images. This also resolves the need for algorithms that are less dependent on labels, such as with semi- and self-supervised methods. Experimental results with three multimodal PET-CT datasets (one public challenge dataset, one public soft-tissue sarcomas (STSs) dataset and one in-house lung cancer data) show that the SFRN achieved consistently better segmentation results when compared to the state-of-the-art methods.

2. A new constrained hierarchical multimodal feature learning (CHMFL) method for patient outcome prediction with multimodal PET-CT images (Chapter 4). This addresses the problem of optimal extraction and analysis of prognostic information from multimodal PET-CT images in the field of radiomics, where complementary features from both modalities are incorporated. The CHMFL method is evaluated in predicting the development of distant metastases (DM) using imaging data before the DM developed on a well-established benchmark PET-CT STS dataset. The experimental results demonstrate that CHMFL achieved overall better performance when compared to the state-of-the-art methods.

3. A new multimodal NAS (MM-NAS) method to automatically search for a multimodal CNN architecture for use in PET-CT radiomics studies (Chapter 5). This addresses the limitation of reliance on human expertise to design dataset-specific and task-specific CNN architectures in the field of radiomics, easing the subsequent manual designs. Thus, the final architecture of CNN in multimodal radiomics can be achieved more efficiently. The MM-NAS method is designed and evaluated in two well-established applications: (i) prediction of DM development; and (ii) tumour segmentation, using a well-established benchmark PET-CT STS dataset. The experimental results demonstrate that MM-NAS achieved overall better performance when compared to the state-of-the-art NAS methods.

6.2 Future Work

There are several interesting areas for advancing deep learning-based radiomics studies in multimodal PET-CT images.

While this thesis has primarily focused on the application of multimodal PET-CT in cancer risk assessment, it's important to recognize that the medical field encompasses a wide variety of data modalities, including MRI, patient clinical reports, and laboratory results, among others. An

exciting direction in radiomics involves the integration of these diverse data modalities alongside PET-CT, aiming to move beyond reliance solely on PET-CT data. Such integration holds the promise of providing complementary insights from various modalities, potentially enhancing disease diagnosis and treatment planning [132]. However, this endeavour calls for the development of intricate and innovative deep learning frameworks and network architectures capable of handling "omni-modality" data. This research direction becomes even more compelling when combined with cutting-edge deep learning techniques, such as transformers inspired by natural language processing [133].

In addition, predicting the presence of DM as a binary classification is an abstraction of a time-to-event prediction problem (i.e., estimating the time point at which an event occurs). The time-to-event problem poses a more complicated modelling challenge than binary classification and may require different methodological approaches. The public STS dataset utilized in this thesis features patients with a 7-year follow-up period for outcome observation, with DM generally confirmed within 4 years after primary STS diagnosis, making it appropriate for binary classification, but not feasible for time point prediction. The dataset's limited size (n=51) also precluded the use of separate held-out data for testing. A more extensive STS radiomics dataset, if available, would facilitate further evaluation and exploration.

While this thesis has demonstrated DM prediction in STS, the method's generalization to other tumour types, such as NSCLC and lymphomas, represents a promising direction. Additionally, exploring alternative outcomes, such as local tumour recurrences and long-term survival, could expand the applicability of the radiomics framework. In lymphomas, where multiple disease sites are common, and disease recurrence occurs unpredictably, analysing multiple lesions will be essential to predict the disease's location. My radiomics framework can be adapted to such a situation by designing and modifying the input data and network architecture, and the suitable input data will require bounding boxes for all lesions on the images.

Furthermore, it is worth noting that the current exploration of CNN-based methods primarily revolves around supervised deep learning. These methods, although effective, are reliant on the availability of annotated training data. Chapter 3 of this thesis has partially validated the efficacy of a self-supervised learning (SSL) strategy for pre-training, demonstrating improved performance through enhanced representation capabilities of tumour regions. However, in the context of multimodal medical imaging, most existing SSL methods lack the capacity to harness cross-modality complementary information. Prior research efforts have primarily concentrated on domain adaptation or cross-modality image registration, with a focus on MRI and CT images [134]–[136]. These methods depend on shared anatomical information across different modalities, such as T1-weighted and T2-weighted MRI, and CT images, to generate supervision information for the SSL network. This requirement fundamentally differs from the needs of multimodal PET-CT images, which offer distinct yet complementary functional (from PET) and anatomical (from CT) information. Therefore, existing multimodal SSL methods are not directly applicable to multimodal PET-CT images, highlighting the pressing need for the development of accurate and robust SSL methodologies tailored to this specific domain.

The demand for such methodologies extends beyond the scope of this thesis and has the potential to impact various facets of PET-CT image analysis, including radiomics, registration, and detection. Future research in this direction holds great promise and can substantially enhance the understanding and utilization of multimodal PET-CT data in clinical practice, such as treatment planning, therapy response assessment, etc.

REFERENCES

[1] I. Bankman, *Handbook of medical image processing and analysis*. Elsevier, 2008.

[2] M. Hatt, F. Tixier, L. Pierce, P. E. Kinahan, C. C. Le Rest, and D. Visvikis, 'Characterization of PET/CT images using texture analysis: the past, the present… any future?', *Eur. J. Nucl. Med. Mol. Imaging*, vol. 44, no. 1, pp. 151–165, 2017, doi: 10.1007/s00259-016-3427-0.

[3] R. L. Wahl, H. Jacene, Y. Kasamon, and M. A. Lodge, 'From RECIST to PERCIST: Evolving Considerations for PET Response Criteria in Solid Tumors', *J Nucl Med*, vol. 50, no. Suppl 1, pp. 122-150 , 2009.

[4] U. Bagci *et al.*, 'Joint segmentation of anatomical and functional images: Applications in quantification of lesions from PET, PET-CT, MRI-PET, and MRI-PET-CT images', *Med. Image Anal.*, vol. 17, no. 8, pp. 929–945, 2013.

[5] S. Gatidis and T. Kuestner, 'A whole-body FDG-PET/CT dataset with manually annotated tumor lesions'. The Cancer Imaging Archive, 2022. doi: 10.7937/GKR0-XV29.

[6] H. Hao *et al.*, 'Shell feature: a new radiomics descriptor for predicting distant failure after radiotherapy in non-small cell lung cancer and cervix cancer', *Phys. Med. Biol.*, vol. 63, no. 9, p. 095007, 2018.

[7] M. Vallières, C. R. Freeman, S. R. Skamene, and I. El Naqa, 'A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the

extremities', *Phys. Med. Biol.*, vol. 60, no. 14, p. 5471, 2015.

[8] K. S. M. Van der Geest *et al.*, 'Diagnostic value of [18F] FDG-PET/CT for treatment monitoring in large vessel vasculitis: a systematic review and meta-analysis', *Eur. J. Nucl. Med. Mol. Imaging*, vol. 48, no. 12, pp. 3886–3902, 2021.

[9] M. Hatt, C. C. Le Rest, F. Tixier, B. Badic, U. Schick, and D. Visvikis, 'Radiomics: data are also images', *J. Nucl. Med.*, vol. 60, no. Supplement 2, pp. 38S-44S, 2019.

[10] R. J. Gillies, P. E. Kinahan, and H. Hricak, 'Radiomics: images are more than pictures, they are data', *Radiology*, vol. 278, no. 2, pp. 563–577, 2016.

[11] M. Vallières, A. Zwanenburg, B. Badic, C. C. Le Rest, D. Visvikis, and M. Hatt, 'Responsible radiomics research for faster clinical translation', *Journal of Nuclear Medicine*, vol. 59, no. 2. Soc Nuclear Med, pp. 189–193, 2018.

[12] M. Vallieres *et al.*, 'Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer', *Sci. Rep.*, vol. 7, no. 1, pp. 1–14, 2017.

[13] V. D. Corino *et al.*, 'Radiomic analysis of soft tissues sarcomas can distinguish intermediate from high-grade lesions', *J. Magn. Reson. Imaging*, vol. 47, no. 3, pp. 829–840, 2018.

[14] J. C. Peeken *et al.*, 'CT-based radiomic features predict tumor grading and have prognostic value in patients with soft tissue sarcomas treated with neoadjuvant radiation therapy', *Radiother. Oncol.*, vol. 135, pp. 187–196, 2019.

[15] P. Afshar, A. Mohammadi, P. K. N, O. A, and H. Benali, 'From handcrafted to deep-learning-based cancer radiomics: challenges and opportunities', *IEEE Signal Process Mag*, vol. 36, pp. 132–60, 2019.

[16] A. Hosny *et al.*, 'Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study', *PLoS Med.*, vol. 15, no. 11, p. e1002711, 2018.

[17] J. Fu *et al.*, 'Deep learning-based radiomic features for improving neoadjuvant chemoradiation response prediction in locally advanced rectal cancer Phys', *Med Biol*, vol. 65 75001, 2020.

[18] Y. Zhu *et al.*, 'A deep learning radiomics model for preoperative grading in meningioma Eur', *J Radiol*, vol. 116, pp. 128–34, 2019.

[19] J. Lao *et al.*, 'A deep learning-based radiomics model for prediction of survival in glioblastoma

multiforme Sci', *Rep*, vol. 7 10353, 2017.

[20] Y. Peng, L. Bi, Y. Guo, D. Feng, M. Fulham, and J. Kim, 'Deep multi-modality collaborative learning for distant metastases predication in PET-CT soft-tissue sarcoma studies', in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2019, pp. 3658–3688.

[21] L. Chen *et al.*, 'Combining many-objective radiomics and 3D convolutional neural network through evidential reasoning to predict lymph node metastasis in head and neck cancer', *Phys. Med. Biol.*, vol. 64, no. 7, p. 75011, 2019.

[22] M. Meng, B. Gu, L. Bi, S. Song, D. D. Feng, and J. Kim, 'DeepMTS: Deep multi-task learning for survival prediction in patients with advanced nasopharyngeal carcinoma using pretreatment PET/CT', *IEEE J. Biomed. Health Inform.*, vol. 26, no. 9, pp. 4497–4507, 2022.

[23] T. Zhou, S. Ruan, and S. Canu, 'A review: Deep learning for medical image segmentation using multi-modality fusion', *Array*, vol. 3, p. 100004, 2019.

[24] D. Taïeb *et al.*, 'European association of nuclear medicine practice guideline/society of nuclear medicine and molecular imaging procedure standard 2019 for radionuclide imaging of phaeochromocytoma and paraganglioma', *Eur. J. Nucl. Med. Mol. Imaging*, vol. 46, no. 10, pp. 2112-2137, 2019.

[25] I. Domingues, G. Pereira, P. Martins, H. Duarte, J. Santos, and P. H. Abreu, 'Using deep learning techniques in medical imaging: a systematic review of applications on CT and PET', *Artif. Intell. Rev.*, vol. 53, no. 6, pp. 4093–4160, 2020.

[26] T. Elsken, J. H. Metzen, and F. Hutter, 'Neural architecture search: A survey', *J. Mach. Learn. Res.*, vol. 20, no. 1, pp. 1997–2017, 2019.

[27] W. Bae, S. Lee, Y. Lee, B. Park, M. Chung, and K.-H. Jung, 'Resource Optimized Neural Architecture Search for 3D Medical Image Segmentation', in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2019, pp. 228–236.

[28] N. Dong, M. Xu, X. Liang, Y. Jiang, W. Dai, and E. Xing, 'Neural Architecture Search for Adversarial Medical Image Segmentation', in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2019, pp. 828–836.

[29] G. T. Herman, 'Fundamentals of Computerized Tomography: Image Reconstruction from Projections', in *Advances in Computer Vision and Pattern Recognition, Springer-Verlag London*, 2nd ed.2009.

[30] Z. Abramson, S. Susarla, M. August, M. Troulis, and L. Kaban, 'Three-dimensional computed tomographic analysis of airway anatomy in patients with obstructive sleep apnea', *J. Oral Maxillofac. Surg.*, vol. 68, no. 2, pp. 354-362, 2010.

[31] A. Madani, C. Keyzer, and P. Gevenois, 'Quantitative computed tomography assessment of lung structure and function in pulmonary emphysema', *Eur. Respir. J.*, vol. 18, no. 4, pp. 720-730, 2001.

[32] M. Hasegawa *et al.*, 'Growth rate of small lung cancers detected on mass CT screening', *Br. J. Radiol.*, vol. 73, no. 876, pp. 1252-1259, 2000.

[33] D. F. Yankelevitz, A. P. Reeves, W. J. Kostis, B. Zhao, and C. I. Henschke, 'Small pulmonary nodules: Volumetrically determined growth rates based on CT evaluation', *Radiology*, vol. 217, no. 1, pp. 251-256, 2000.

[34] P. E. Valk, D. Delbeke, and D. L. Bailey, *Positron Emission Tomography: Clinical Practice*. Springer-Verlag London, 2006.

[35] H. Jadvar, 'Imaging evaluation of prostate cancer with 18 F-fluorodeoxyglucose PET/CT: utility and limitations', *Eur. J. Nucl. Med. Mol. Imaging*, vol. 40, pp. 5–10, 2013.

[36] S. Belhassen and H. Zaidi, 'A novel fuzzy c-means algorithm for unsupervised heterogeneous tumor quantification in pet', *Med. Phys.*, vol. 37, no. 3, pp. 1309-1324, 2010.

[37] J. Czernin, M. Dahlbom, O. Ratib, and C. Schiepers, *Atlas of PET/CT Imaging in Oncology*. Berlin, Heidelberg: Springer-Verlag, 2004.

[38] H. Zaidi and I. E. Naqa, 'PET-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques', *Eur. J. Nucl. Med. Mol. Imaging*, vol. 37, pp. 2165-2187, 2010.

[39] J. A. Thie, 'Understanding the standardized uptake value, its methods, and implications for usage', *J. Nucl. Med.*, vol. 45, no. 9, pp. 1431-1434, 2004.

[40] K. Hirata *et al.*, 'A semi-automated technique determining the liver standardized uptake value reference for tumor delineation in FDG PET-CT', *PloS One*, vol. 9, no. 8, p. 105682, 2014.

[41] H. Yu, C. Caldwell, K. Mah, and D. Mozeg, 'Coregistered FDG PET/CTbased textural

characterization of head and neck cancer for radiation treatment planning', *IEEE Trans. Med. Imaging*, vol. 28, no. 3, pp. 374-383, 2009.

[42] D. W. Townsend, 'Combined positron emission tomography–computed tomography: the historical perspective', in *Seminars in Ultrasound, CT and MRI*, Elsevier, 2008, pp. 232–235.

[43] T. Beyer, G. Antoch, T. Blodgett, L. F. Freudenberg, T. Akhurst, and SMueller, 'Dual-modality PET/CT imaging: the effect of respiratory motion on combined image quality in clinical oncology', *Eur. J. Nucl. Med. Mol. Imaging*, vol. 30, no. 4, pp. 588-596, 2003.

[44] T. Beyer, D. Townsend, and T. Blodgett, 'Dual-modality PET/CT tomography for clinical oncology', *Q. J. Nucl. Med. Mol. Imaging*, vol. 46, no. 1, p. 24, 2002.

[45] D. W. Townsend and T. Beyer, 'A combined PET/CT scanner: the path to true image fusion', *Br. J. Radiol.*, vol. 75, no. Supplement 9, pp. 24-30, 2002.

[46] D. W. Townsend, T. Beyer, and T. M. Blodgett, 'PET/CT scanners: A hardware approach to image fusion', *Semin. Nucl. Med.*, vol. 33, no. 3, pp. 193-204, 2003.

[47] T. M. Blodgett, C. C. Meltzer, and D. W. Townsend, 'PET/CT: Form and function', *Radiology*, vol. 242, no. 2, pp. 360-385, 2007.

[48] G. K. Schulthess, H. C. Steinert, and T. F. Hany, 'Integrated PET/CT: Current applications and future directions', *Radiology*, vol. 238, no. 2, pp. 405-422, 2006.

[49] G. W. Goerres, G. K. Schulthess, and H. C. Steinert, 'Why most PET of lung and head-and-neck cancer will be PET/CT', *J. Nucl. Med.*, vol. 45, no. Supplement 1, pp. 66-71 , 2004.

[50] L. Bi, J. Kim, D. Feng, and M. Fulham, 'Multi-stage thresholded region classification for whole-body PET-CT lymphoma studies', in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part I 17*, Springer, 2014, pp. 569–576.

[51] B. Foster, U. Bagci, A. Mansoor, Z. Xu, and D. J. Mollura, 'A review on segmentation of positron emission tomography images', *Comput. Biol. Med.*, vol. 50, pp. 76–96, 2014.

[52] M. Moussallem, P.-J. Valette, A. Traverse-Glehen, C. Houzard, C. Jegou, and F. Giammarile, 'New strategy for automatic tumor segmentation by adaptive thresholding on PET/CT images', *J. Appl. Clin. Med. Phys.*, vol. 13, no. 5, pp. 236–251, 2012.

[53] R. L. Wahl, H. Jacene, Y. Kasamon, and M. A. Lodge, 'From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors', *J. Nucl. Med.*, vol. 50, no. Suppl 1, pp. 122S-150S, 2009.

[54] P. Bennett, A. Mintz, B. Perry, A. Trout, and P. Vergara-Wentland, *Specialty Imaging: PET-E-Book*. Elsevier Health Sciences, 2017.

[55] H. Cui *et al.*, 'Topology polymorphism graph for lung tumor segmentation in PET-CT images', *Phys. Med. Biol.*, vol. 60, no. 12, p. 4893, 2015.

[56] U. Bagci, J. K. Udupa, J. Yao, and D. J. Mollura, 'Co-segmentation of functional and anatomical images', in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012: 15th International Conference, Nice, France, October 1-5, 2012, Proceedings, Part III 15*, Springer, 2012, pp. 459–467.

[57] D. Han *et al.*, 'Globally optimal tumor segmentation in PET-CT images: a graph-based co-segmentation method', in *Information Processing in Medical Imaging: 22nd International Conference, IPMI 2011, Kloster Irsee, Germany, July 3-8, 2011. Proceedings 22*, Springer, 2011, pp. 245–256.

[58] W. Ju, D. Xiang, B. Zhang, L. Wang, I. Kopriva, and X. Chen, 'Random walk and graph cut for co-segmentation of lung tumor on PET-CT images', *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5854–5867, 2015.

[59] Q. Song *et al.*, 'Optimal co-segmentation of tumor in PET-CT images with context information', *IEEE Trans. Med. Imaging*, vol. 32, no. 9, pp. 1685–1697, 2013.

[60] J. Wojak, E. D. Angelini, and I. Bloch, 'Joint variational segmentation of CT-PET data for tumoral lesions', in *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, IEEE, 2010, pp. 217–220.

[61] Z. Guo, X. Li, H. Huang, N. Guo, and Q. Li, 'Deep learning-based image segmentation on multimodal medical imaging', *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 3, no. 2, pp. 162–169, 2019.

[62] D. Jin *et al.*, 'Accurate esophageal gross tumor volume segmentation in PET/CT using two-stream chained 3D deep network fusion', in *International Conference on Medical Image Computing and*

Computer-Assisted Intervention, Springer, 2019, pp. 182–191.

[63] A. Kumar, M. Fulham, D. Feng, and J. Kim, 'Co-learning feature fusion maps from PET-CT images of lung cancer', *IEEE Trans. Med. Imaging*, vol. 39, no. 1, pp. 204–217, 2019.

[64] L. Li, X. Zhao, W. Lu, and S. Tan, 'Deep learning for variational multimodality tumor segmentation in PET/CT', *Neurocomputing*, vol. 392, pp. 277–295, 2020.

[65] L. Bi *et al.*, 'Recurrent feature fusion learning for multi-modality pet-ct tumor segmentation', *Comput. Methods Programs Biomed.*, vol. 203, p. 106043, 2021.

[66] X. Fu, L. Bi, A. Kumar, M. Fulham, and J. Kim, 'Multimodal spatial attention module for targeting multimodal PET-CT lung tumor segmentation', *IEEE J. Biomed. Health Inform.*, vol. 25, no. 9, pp. 3507–3516, 2021.

[67] Z. Xue *et al.*, 'Multi-Modal Co-Learning for Liver Lesion Segmentation on PET-CT Images', *IEEE Trans. Med. Imaging*, vol. 40, no. 12, pp. 3531–3542, 2021.

[68] D. Xiang, B. Zhang, Y. Lu, and S. Deng, 'Modality-Specific Segmentation Network for Lung Tumor Segmentation in PET-CT Images', *IEEE J. Biomed. Health Inform.*, 2022.

[69] S. S. F. Yip and H. J. W. L. Aerts, 'Applications and limitations of radiomics', *Phys Med Biol*, vol. 61, no. 13, p. 150, 2016.

[70] D. R. Cox, 'Regression models and life-tables', *J. R. Stat. Soc. Ser. B Methodol.*, vol. 34, no. 2, pp. 187–202, 1972.

[71] M. B. Spraker, 'MRI radiomic features are independently associated with overall survival in soft tissue sarcoma', *Adv Radiat Oncol*, vol. 4, no. 2, pp. 413-421, 2019.

[72] T. P. Coroller *et al.*, 'CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma', *Radiother. Oncol.*, vol. 114, no. 3, pp. 345-350, 2015.

[73] W. Lv, S. Ashrafinia, J. Ma, L. Lu, and A. Rahmim, 'Multi-level multi-modality fusion radiomics: application to PET and CT imaging for prognostication of head and neck cancer', *IEEE J. Biomed. Health Inform.*, vol. 24, no. 8, pp. 2268–2277, 2019.

[74] J. Tolles and W. J. Meurer, 'Logistic regression: relating patient characteristics to outcomes', *Jama*, vol. 316, no. 5, pp. 533–534, 2016.

[75] J. Xiong, W. Yu, J. Ma, Y. Ren, X. Fu, and J. Zhao, 'The role of PET-based radiomic features in

predicting local control of esophageal cancer treated with concurrent chemoradiotherapy', *Sci. Rep.*, vol. 8, no. 1, p. 9902, 2018.

[76] L. Breiman, 'Random forests', *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[77] W. S. Noble, 'What is a support vector machine?', *Nat. Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, 2006.

[78] J. Juntu, J. Sijbers, S. Backer, J. Rajan, and D. Dyck, 'Machine learning study of several classifiers trained with texture analysis features to differentiate benign from malignant soft-tissue tumors in T1-MRI images', *J Magn Reson Imaging*, vol. 31, no. 3, pp. 680-689, Mar. 2010.

[79] P. Cunningham and S. J. Delany, 'k-Nearest neighbour classifiers-A Tutorial', *ACM Comput. Surv. CSUR*, vol. 54, no. 6, pp. 1–25, 2021.

[80] J. E. Van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi, and B. Baessler, 'Radiomics in medical imaging—"how-to" guide and critical reflection', *Insights Imaging*, vol. 11, no. 1, pp. 1–16, 2020.

[81] P. E. Galavis, C. Hollensen, N. Jallow, B. Paliwal, and R. Jeraj, 'Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters', *Acta Oncol.*, vol. 49, no. 7, pp. 1012–1016, 2010.

[82] G. Litjens *et al.*, 'A survey on deep learning in medical image analysis', *Med. Image Anal.*, vol. 42, p. 6088, 2017.

[83] H. Greenspan, B. Ginneken, and R. M. Summers, 'Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique', *IEEE Trans. Med. Imaging*, vol. 35, no. 5, p. 11531159, 2016.

[84] M. Avanzo *et al.*, 'Machine and deep learning methods for radiomics', *Med. Phys.*, vol. 47, no. 5, pp. e185–e202, 2020.

[85] O. Ronneberger, P. Fischer, and T. Brox, 'U-net: Convolutional networks for biomedical image segmentation', in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, pp. 234–241.

[86] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep residual learning for image recognition', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[87] L. Oakden-Rayner, G. Carneiro, T. Bessen, J. C. Nascimento, A. P. Bradley, and L. J. Palmer, 'Precision radiology: predicting longevity using feature engineering and deep learning methods in a radiomics framework', *Sci Rep*, vol. 7, no. 1, p. 1648, 2017.

[88] Z. Li, Y. Wang, J. Yu, Y. Guo, and W. Cao, 'Deep Learning based Radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma', *Sci Rep*, vol. 7, no. 1, p. 5467, 2017.

[89] D. Kumar, A. G. Chung, M. J. Shaifee, F. Khalvati, M. A. Haider, and A. Wong, 'Discovery radiomics for pathologically-proven computed tomography lung cancer prediction', in *Image Analysis and Recognition: 14th International Conference, ICIAR 2017, Montreal, QC, Canada, July 5–7, 2017, Proceedings 14*, Springer, 2017, pp. 54–62.

[90] A. Diamant, A. Chatterjee, M. Vallières, G. Shenouda, and J. Seuntjens, 'Deep learning in head & neck cancer outcome prediction', *Sci. Rep.*, vol. 9, no. 1, p. 2764, 2019.

[91] H. Hermessi, O. Mourali, and E. Zagrouba, 'Deep feature learning for soft tissue sarcoma classification in MR images via transfer learning', *Expert Syst Appl*, vol. 120, pp. 116-127, 2019.

[92] S. Wang, Y. Hou, Z. Li, J. Dong, and C. Tang, 'Combining convnets with hand-crafted features for action recognition based on an HMM-SVM classifier', *Multimed Tools Appl*, vol. 77, no. 15, pp. 18983-18998, 2018.

[93] I. Sutskever, J. Martens, and G. E. Hinton, 'Generating text with recurrent neural networks', in *Proceedings of the 28th international conference on machine learning (ICML-11*, 2011, pp. 1017–1024.

[94] S. Azizi, 'Deep recurrent neural networks for prostate cancer detection: analysis of temporal enhanced ultrasound', *IEEE Trans. Med. Imaging*, vol. 37, no. 12, pp. 2695–2703, 2018.

[95] Y. Peng, J. Kim, D. Feng, and L. Bi, 'Automatic Tumor Segmentation via False Positive Reduction Network for Whole-Body Multi-Modal PET/CT Images', *ArXiv Prepr. ArXiv220907705*, 2022.

[96] S. Gatidis, T. Küstner, M. Ingrisch, M. Fabritius, and C. Cyran, 'Automated Lesion Segmentation in Whole-Body FDG-PET/CT', Mar. 2022, doi: 10.5281/ZENODO.6362493.

[97] K. Clark *et al.*, 'The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository', *J. Digit. Imaging*, vol. 26, no. 6, pp. 1045–1057, 2013.

[98] J.-B. Grill *et al.*, 'Bootstrap your own latent-a new approach to self-supervised learning', *Adv.*

Neural Inf. Process. Syst., vol. 33, pp. 21271–21284, 2020.

[99] A. Tarvainen and H. Valpola, 'Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results', *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[100] M. Caron *et al.*, 'Emerging properties in self-supervised vision transformers', in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9650–9660.

[101] X. Chen, S. Xie, and K. He, 'An empirical study of training self-supervised vision transformers', in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9640–9649.

[102] X. Chen and K. He, 'Exploring simple siamese representation learning', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15750–15758.

[103] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, 'A simple framework for contrastive learning of visual representations', in *International conference on machine learning*, PMLR, 2020, pp. 1597–1607.

[104] J. Gou, B. Yu, S. J. Maybank, and D. Tao, 'Knowledge distillation: A survey', *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, 2021.

[105] L. Jing, P. Vincent, Y. LeCun, and Y. Tian, 'Understanding dimensional collapse in contrastive self-supervised learning', *ArXiv Prepr. ArXiv211009348*, 2021.

[106] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, 'Unsupervised learning of visual features by contrasting cluster assignments', *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 9912–9924, 2020.

[107] 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.

[108] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4. Springer, 2006.

[109] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, 'Focal loss for dense object detection', in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[110] A. Paszke *et al.*, ‘Automatic differentiation in pytorch’, 2017.

[111] K. He, X. Zhang, S. Ren, and J. Sun, ‘Delving deep into rectifiers: Surpassing human-level performance on imagenet classification’, in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[112] I. Loshchilov and F. Hutter, ‘Decoupled Weight Decay Regularization’, presented at the International Conference on Learning Representations, 2018.

[113] 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’, *Nat. Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[114] J. Ye *et al.*, ‘Exploring Vanilla U-Net for Lesion Segmentation from Whole-body FDG-PET/CT Scans’, *ArXiv Prepr. ArXiv221007490*, 2022.

[115] ‘GitHub - JunMa11/PETCTSeg: Automatic segmentation models for PET and CT scans’, *GitHub*. https://github.com/JunMa11/PETCTSeg (accessed Sep. 06, 2023).

[116] J. Zhang, Y. Huang, Z. Zhang, and Y. Shi, ‘Whole-Body Lesion Segmentation in 18F-FDG PET/CT’, *ArXiv Prepr. ArXiv220907851*, 2022.

[117] D. Kingma and J. Ba, ‘Adam: A Method for Stochastic Optimization’, in *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015*, 2015.

[118] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, ‘Learning deep features for discriminative localization’, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[119] L. van der Maaten and G. Hinton, ‘Visualizing data using t-SNE’, *J. Mach. Learn. Res.*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[120] J. F. Eary, F. O’Sullivan, J. O’Sullivan, and E. U. Conrad, ‘Spatial heterogeneity in sarcoma 18F-FDG uptake as a predictor of patient outcome’, *J. Nucl. Med.*, vol. 49, no. 12, pp. 1973–1979, 2008.

[121] B. Zoph, V. Vasudevan, J. Shlens, and Q. V Le, ‘Learning transferable architectures for scalable image recognition’, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8697–8710.

[122] H. Liu, K. Simonyan, and Y. Yang, 'Darts: Differentiable architecture search', *ArXiv Prepr. ArXiv180609055*, 2018.

[123] H. Pham, M. Y. Guan, B. Zoph, Q. V Le, and J. Dean, 'Efficient neural architecture search via parameter sharing', *ArXiv Prepr. ArXiv180203268*, 2018.

[124] H. Liu, K. Simonyan, O. Vinyals, C. Fernando, and K. Kavukcuoglu, 'Hierarchical representations for efficient architecture search', *ArXiv Prepr. ArXiv171100436*, 2017.

[125] F. Pedregosa, 'Hyperparameter optimization with approximate gradient', in *International conference on machine learning*, PMLR, 2016, pp. 737–746.

[126] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil, 'Bilevel programming for hyperparameter optimization and meta-learning', in *International Conference on Machine Learning*, PMLR, 2018, pp. 1568–1577.

[127] J. Luketina, M. Berglund, K. Greff, and T. Raiko, 'Scalable gradient-based tuning of continuous regularization hyperparameters', in *International conference on machine learning*, PMLR, 2016, pp. 2952–2960.

[128] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein, 'Unrolled Generative Adversarial Networks', in *International Conference on Learning Representations*,

[129] Y. Weng, T. Zhou, Y. Li, and X. Qiu, 'Nas-unet: Neural architecture search for medical image segmentation', *IEEE Access*, vol. 7, pp. 44247–44257, 2019.

[130] Z. Zhu, C. Liu, D. Yang, A. Yuille, and D. Xu, 'V-NAS: Neural architecture search for volumetric medical image segmentation', in *2019 International conference on 3d vision (3DV)*, IEEE, 2019, pp. 240–248.

[131] F. Wang, 'Neural architecture search for gliomas segmentation on multimodal magnetic resonance imaging', *ArXiv Prepr. ArXiv200506338*, 2020.

[132] K. M. Boehm, P. Khosravi, R. Vanguri, J. Gao, and S. P. Shah, 'Harnessing multimodal data integration to advance precision oncology', *Nat. Rev. Cancer*, vol. 22, no. 2, pp. 114–126, 2022.

[133] A. Dosovitskiy *et al.*, 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale', in *International Conference on Learning Representations*, 2020.

[134] X. Du and Y. Liu, 'Constraint-based unsupervised domain adaptation network for multi-

modality cardiac image segmentation', *IEEE J. Biomed. Health Inform.*, vol. 26, no. 1, pp. 67–78, 2021.

[135] F. Fang, Y. Yao, T. Zhou, G. Xie, and J. Lu, 'Self-supervised multi-modal hybrid fusion network for brain tumor segmentation', *IEEE J. Biomed. Health Inform.*, vol. 26, no. 11, pp. 5310–5320, 2021.

[136] H. Siebert, L. Hansen, and M. P. Heinrich, 'Learning a Metric for Multimodal Medical Image Registration without Supervision Based on Cycle Constraints', *Sensors*, vol. 22, no. 3, p. 1107, 2022.