Localisation in 3D Images Using Cross-features Correlation Learning

Majedaldein I. Almahasneh
809508

Submitted to Swansea University in fulfilment
of the requirements for the Degree of Doctor of Philosophy

Swansea University
Prifysgol Abertawe

Department of Computer Science
Swansea University

April 17, 2022

Declaration

This work has not been previously accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed ..... (candidate)

Date 17 April 2022

Statement 1

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed ..... (candidate)

Date 17 April 2022

Statement 2

I hereby give my consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ..... (candidate)

Date 17 April 2022

إقْرَأْ

I dedicate this work to my parents.
Thank you mom and dad.

Abstract

Object detection and segmentation have evolved drastically over the past two decades thanks to the continuous advancement in the field of deep learning. Substantial research efforts have been dedicated towards integrating object detection techniques into a wide range of real-world problems. Most existing methods take advantage of the successful application and representational ability of convolutional neural networks (CNNs). Generally, these methods target mainstream applications that are typically based on 2D imaging scenarios. Additionally, driven by the strong correlation between the quality of the feature embedding and the performance in CNNs, most works focus on design characteristics of CNNs, e.g., depth and width, to enhance their modelling capacity and discriminative ability. Limited research was directed towards exploiting feature-level dependencies, which can be feasibly used to enhance the performance of CNNs. Moreover, directly adopting such approaches into more complex imaging domains that target data of higher dimensions (e.g., 3D multi-modal and volumetric images) is not straightforwardly applicable due to the different nature and complexity of the problem. In this thesis, we explore the possibility of incorporating feature-level correspondence and correlations into object detection and segmentation contexts that target the localisation of 3D objects from 3D multi-modal and volumetric image data. Accordingly, we first explore the detection problem of 3D solar active regions in multi-spectral solar imagery where different imaging bands correspond to different 2D layers (altitudes) in the 3D solar atmosphere.

We propose a joint analysis approach in which information from different imaging bands is first individually analysed using band-specific network branches to extract inter-band features that are then dynamically cross-integrated and jointly analysed to investigate spatial correspondence and co-dependencies between the different bands. The aggregated embeddings are further analysed using band-specific detection network branches to predict separate sets of results (one for each band). Throughout our study, we evaluate different types of feature fusion, using convolutional embeddings of different semantic levels, as well as the impact of using different numbers of image bands inputs to perform the joint analysis. We test the proposed approach over different

multi-modal datasets (multi-modal solar images and brain MRI) and applications. The proposed joint analysis based framework consistently improves the CNN's performance when detecting target regions in contrast to single band based baseline methods.

We then generalise our cross-band joint analysis detection scheme into the 3D segmentation problem using multi-modal images. We adopt the joint analysis principles into a segmentation framework where cross-band information is dynamically analysed and cross-integrated at various semantic levels. The proposed segmentation network also takes advantage of band-specific skip connections to maximise the inter-band information and assist the network in capturing fine details using embeddings of different spatial scales. Furthermore, a recursive training strategy, based on weak labels (e.g., bounding boxes), is proposed to overcome the difficulty of producing dense labels to train the segmentation network. We evaluate the proposed segmentation approach using different feature fusion approaches, over different datasets (multi-modal solar images, brain MRI, and cloud satellite imagery), and using different levels of supervisions. Promising results were achieved and demonstrate an improved performance in contrast to single band based analysis and state-of-the-art segmentation methods.

Additionally, we investigate the possibility of explicitly modelling objective driven feature-level correlations, in a localised manner, within 3D medical imaging scenarios (3D CT pulmonary imaging) to enhance the effectiveness of the feature extraction process in CNNs and subsequently the detection performance. Particularly, we present a framework to perform the 3D detection of pulmonary nodules as an ensemble of two stages, candidate proposal and a false positive reduction. We propose a 3D channel attention block in which cross-channel information is incorporated to infer channel-wise feature importance with respect to the target objective. Unlike common attention approaches that rely on heavy dimensionality reduction and computationally expensive multi-layer perceptron networks, the proposed approach utilises fully convolutional networks to allow directly exploiting rich 3D descriptors and performing the attention in an efficient manner. We also propose a fully convolutional 3D spatial attention approach that elevates cross-sectional information to infer spatial attention. We demonstrate the effectiveness of the proposed attention approaches against a number of popular channel and spatial attention mechanisms. Furthermore, for the False positive reduction stage, in addition to attention, we adopt a joint analysis based approach that takes into account the variable nodule morphology by aggregating spatial information from different contextual levels. We also propose a Zoom-in convolutional path that incorporates semantic information of different spatial scales to assist the network in capturing fine details. The proposed detection approach demonstrates considerable

gains in performance in contrast to state-of-the-art lung nodule detection methods.

We further explore the possibility of incorporating long-range dependencies between arbitrary positions in the input features using Transformer networks to infer self-attention, in the context of 3D pulmonary nodule detection, in contrast to localised (convolutional based) attention . We present a hybrid 3D detection approach that takes advantage of both, the Transformers ability in modelling global context and correlations and the spatial representational characteristics of convolutional neural networks, providing complementary information and subsequently improving the discriminative ability of the detection model. We propose two hybrid Transformer CNN variants where we investigate the impact of exploiting a deeper Transformer design –in which more Transformer layers and trainable parameters are incorporated– is used along with high-level convolutional feature inputs of a single spatial resolution, in contrast to a shallower Transformer design –of less Transformer layers and trainable parameters– while exploiting convolutional embeddings of different semantic levels and relatively higher resolution.

Extensive quantitative and qualitative analyses are presented for the proposed methods in this thesis and demonstrate the feasibility of exploiting feature-level relations, either implicitly or explicitly, in different detection and segmentation problems.

Acknowledgements

First and foremost I would like to extend my sincere thanks to my supervisors, Prof. Xinghua Xie and Dr. Adeline Paiement for their consistent support, generous advice, and patience throughout the stages of my PhD. Their invaluable knowledge, passion for science, and willingness to share, have always inspired me during my research journey, and on a personal level, and I am very grateful. I would also like to thank Dr. Jean Aboudarham, Ms. Caryl Richards, Dr. Ali Alqahtani, Dr. Mohammed Ali, and my peers in the PhD lab and all the amazing members of the Computer Vision group at Swansea University, as well as Supercomputing Wales for their support and guidance. I would like to especially express my gratitude to Dr. Jingjing Deng, Dr. Jay Morgan, Dr. Olga Vesela, and Ms. Kasia Szymaniak who took time to provide help and advice whenever I needed it, thank you. I must also extend my sincere gratitude to my parents, Prof. Ihssan Almahasneh and Ms. Jamila Almahasneh, for providing their unconditional and endless support throughout my life, for guiding and influencing me to be the version of myself that I am. Not only for being a great father and mother, but also for being my best teachers and friends. Without them, I would not have been able to carry out my PhD, and I am grateful for that. I also thank my family members, Majed Almahasneh, Mais Almahasneh, Aula Almahasneh, and Khetam Alfaraj, as well as my friends Dr. Ali Telmisani, Dr. Amer Sahouri, Mr. Rashid Mashagbeh, and Dr. Bashar Juraideh, who had to put up with my stresses throughout the years of my study. My PhD would have been impossible without their constant encouragement, positive attitude, and support.

Contents

List of Publication

The following is a list of published papers as a result of the work in this thesis.

1. M. Almahasneh, A. Paiement, X. Xie, J. Aboudarham, Active region detection in multi-spectral solar images. International Conference on Pattern Recognition Applications and Methods, 2021.

2. M. Almahasneh, A. Paiement, X. Xie, J. Aboudarham, MSMT-CNN for solar active region detection with multi-spectral analysis. Springer Nature Computer Science, 2022.

3. M. Almahasneh, A. Paiement, X. Xie, J. Aboudarham, MLMT-CNN for object detection and segmentation in multi-layer and multi-spectral images. Machine Vision and Applications, 2021.

4. M. Almahasneh, X. Xie, A. Paiement, AttentNet for Pulmonary Lung Nodule Detection Using 3D Attention. Medical image analysis, 2022, (under review).

List of Acronyms

AdaBoost Adaptive Boosting.

ARs Active Regions.

BraTS Brain Tumor Segmentation.

CA Channel Attention.

CAM Class Activation Map.

CBAM Convolutional Block Attention Modules.

CNN Convolutional Neural Networks.

CT Computed Tomography.

DL Deep Learning.

DNN Deep Neural Networks.

EIT Extreme ultraviolet Imaging Telescope.

EUV Extreme Ultraviolet.

FCN Fully Convolutional Neural Network.

FROC Free Receiver Operating Characteristic.

GPU Graphical Processing Units.

Haar Haar Wavelets.

HFC Heliophysics Feature Catalogue.

HOG Histogram of Oriented Gradients.

IoU Intersection over Union.

LAD Lower Atmosphere Dataset.

LBP Local Binary Pattern.

LUNA16 Lung Nodule Analysis.

MDI Michelson Doppler Imager.

MLMT Multi-layer Multi-tasking.

MLP Multi-layer Perceptron.

MLSC Multi-level Spatial Context.

MLSC-Z Multi-level Spatial Context with Zoom-in paths.

MRI Magnetic resonance imaging.

NMS Non Maximum Suppression.

NN Neural Networks.

NOAA National Oceanic and Atmospheric Administration.

PET Positron Emission Tomography.

PM Paris Meudon.

RCNN Regional Convolutional Neural Networks.

ReLU Rectified Linear Unit.

RoI Region of Interest.

RPN Region Proposal Network.

SA Spatial Attention.

SDO Solar Dynamics Observatory.

SGD Stochastic Gradient Descent.

SH Spectroheliograph.

SIFT Scale Invariant Feature Transform.

SOHO Solar and Heliospheric Observatory.

SPOCA Spatial Possibilistic Clustering Algorithm.

SURF Speeded Up Robust Features.

SVM Support Vector Machines.

UAD Upper Atmosphere Dataset.

List of Tables

List of Figures

Chapter 1

Introduction

Contents

1.1 Motivation

Localising and identifying objects is a simple task for the human brain. However, adopting these tasks into computers is not trivial. Computer vision is the branch of artificial intelligence that focuses on understanding, automating, and incorporating tasks of the human visual system into computers. Object detection and segmentation are two fundamental visual recognition tasks of which many computer vision applications are based on. Advances in imaging technologies and hardware attracted interest of the research community towards developing novel localisation methodologies that target different real-world problems (e.g., computer aided diagnosis [4], text detection [5], face recognition [6], pedestrian detection [7]).

Early approaches utilised hand-crafted features, such as Histogram of Oriented Gradients (HOG) [8], Haar Wavelets (Haar) features [9], Local Binary Pattern (LBP) [10], Scale Invariant Feature Transform (SIFT) [11], and Speeded Up Robust Features (SURF) [12], in combination with classifiers, such as Support Vector Machines (SVM) [13], Adaptive Boosting (AdaBoost) [14], Random Forests [15], and Cascaded classifiers [16], to perform the object detection task in a sliding window fashion. However, these methods are based low level feature descriptors which limits their ability in capturing task semantics, leading to low discriminative ability particularly when applied to more complex tasks. On the other hand, early segmentation methods involved simple unsupervised methods such as clustering (e.g., K-means [17] and Spectral clustering [18]), region growing [19], and graph based strategies [20]. These methods also rely on manually engineered features, and are hyper-parameter and pre- and post processing dependant, making them difficult to adapt to new applications.

Recent advances in Deep Learning (DL) techniques marked a milestone in the field of computer vision, particularly, the rise of Convolutional Neural Networks (CNN) in which objective driven visual features and high level semantic information are learned automatically avoiding the need of manual feature engineering. Starting from LeNet [21] where CNNs where first optimised using backpropagation and Stochastic Gradient Descent (SGD), to AlexNet [22] in which deeper CNNs were utilised with the help of Graphical Processing Units (GPU) to achieve state-of-the-art performance in visual recognition tasks at the time. Since then, and due to their exceptional representational powers, CNNs have attracted a lot of research efforts in which extensions and variations where proposed to enhance their performance, and have therefore become the de facto approach to computer vision problems [22, 23], including object detection, e.g., RCNN [24], SSD [25], and [26], and segmentation, e.g., Fully Convolutional Neural Network (FCN) [27],

U-Net [28], and SegNet [29].

Generally, CNN based localisation involves two main steps, feature extraction, in which a CNN is utilised as a backbone to extract representative high level feature maps from pixel level inputs, and a localisation stage, in which positions of target objects are predicted based on the extracted feature maps. Driven by the immediate correlation between the quality of the extracted information and the inferred performance of CNNs [22, 30–32], considerable amount of research was directed towards investigating CNN design characteristics –such as depth (VGG [33] and ResNet [34]), and width (GoogLeNet [35] and Wide ResNet [36])– to improve their performance. While this has proven feasibility on various mainstream applications that typically target RGB images, it is still challenging to adopt these concepts to more complex applications that target volumetric and multi-modal data due to different nature and complexity as well as the increased computational overhead associated with such tasks [37].

More recently, some works demonstrated the significance of incorporation of feature-level correlation modelling mechanisms to enhance the representational power of CNNs. Such methods aim at capturing cross-channel (e.g., [1, 38]) and inter-spatial (e.g., [1]) dependencies, as well as global context and long range dependencies (e.g., [2, 39]), to infer feature importance (i.e., attention) and allow more effective feature extraction by promoting the network to focus on meaningful embeddings. In the same line, other methods utilised cross-feature spatial correspondence and feature-level correlations indirectly by dynamically analysing subsets of the input and cross integrating this information at different semantic levels (e.g., low- and high-level features [40, 41], decision-making level [42], and transfer learning [43]) to provoke the network in capturing dependencies within the different parts of the input. These methods demonstrate great potential in different computer vision tasks, however, these methods in general target 2D imaging scenes, limited research was dedicated towards incorporating feature correlation learning for 3D imaging scenarios.

In this work, motivated by the continuing advancements in deep learning based object detection and segmentation , as well as the growing interest in understanding the impact of feature relations and dependencies on the the effectiveness and adaptability of convolutional neural networks in real-world applications, we investigate different 3D object detection and segmentation problems and evaluate the feasibility of directly utilising feature-level dependencies and correlations within these contexts:

- **Detection and Segmentation 3D objects in multi-dimensional imagery.**
  Several deep learning based methods were proposed to solve different mainstream object

detection and segmentation tasks, these typically target 2D scenes based on RGB imaging. Other methods that are designed for 3D objects typically target volumetric images or point cloud data. The problem of detecting 3D objects from 3D images that capture sparse layers of a 3D scene (e.g., multi-spectral solar imaging) is widely overlooked. Furthermore, methods that target volumetric data generally formulate the localisation problem as a pixel-wise classification task, or as perform the detection based on 2D slices along the depth axis. While such approaches partially solves the problem, training segmentation Deep Neural Networks (DNN) requires significant amounts of labelled samples which is not typically available for this type of data. Incorporating weak-supervision methods may be an opportune solution to reduce the complexity of such tasks. On the other hand, the 2D based detection approach neglects the 3D aspect of the data, which may be directly used to enrich the information used to perform the detection. Limited research was proposed towards handling the problem as a 3D bounding box prediction task from 3D data (e.g., medical imaging), which may also be used as a prior to reduce the segmentation complexity.

- **Exploiting and modelling feature-level correlations.**
  Driven by the direct impact of the learned embeddings in CNNs on the quality of their final performance, research efforts focused on exploring several engineering aspects of Neural Networks (NN). These involved depth, width, cardinality, receptive field, and capacity of detection networks, which has consistently proven sufficient within many computer vision domains. However, limited research was directed towards understanding the influence and explicitly incorporating feature-level correspondence and correlations into the CNNs, particularly within 3D imaging contexts. Feature-level correlation learning targets capturing inter-spatial or cross-feature (e.g., cross-channel) relations, which may be observed either in a localised or a global context, to provide complementary information to the analysis and improve the quality of the learning process. This can be done either explicitly by directly modelling these relations, or by implicitly provoking DNNs to capture feature-level relations by dynamically and jointly analysing different subsets of the features. Such analysis may be effectively utilised with minimal additional resource requirement which is typically limited in contexts where target 3D image data (e.g., volumetric data or multi-modal data).

- **Objective driven inference of feature importance.**

In the same line of feature-level correlation learning, feature importance can be utilised to enhance the performance of DNNs. This is typically addressed using attention mechanisms, in which feature extraction networks are promoted to selectively focus on features with high relevance and ignore less important features with respect to a given objective, in a learnable manner. The recent emergence and notable success of convolutional attention as well as self-attention methods has drawn a lot of interest towards this problem. However, most existing methods are designed to handle 2D images or sequence-to-sequence applications (e.g., natural language processing). Generalising these methods to more complex data of higher dimensions (e.g., 3D medical data) is not straightforward due to the increased computational cost associated with such tasks. Moreover, convolutional based methods that target 2D imaging applications neglect the inherent 3D aspect of the data when directly adopted into such domains. Incorporating attention strategies with DNNs is an applicable option in which the performance of DNNs may be improved.

In this work, we explore different object detection and segmentation problems in the context of 3D imaging data. Particularly, we look into the detection and segmentation of 3D objects in multi-spectral data that observes sparse 2D layers of a 3D scene. We also look into the detection of 3D objects in volumetric images using 3D bounding boxes. We explore different feature-level correlation learning approaches as well as attention techniques to increase the discriminative ability and performance of deep learning models. In Sections 1.2 and 1.3, we present an overview and briefly discuss the overall contribution of this work, respectively. Section 1.4 presents the outline of this thesis.

1.2 Overview

Existing object localisation methods rely on CNNs and focus on exploring design characteristics to improve their discriminative ability. Such methods aim at mainstream, typically 2D imaging based, applications. Generalising such methods into more complex contexts that target data of higher dimensions (e.g., 3D and multi-modal images) is not straightforwardly applicable and therefore requires the design of specialised methods. The aim of this work is to investigate the possibility of incorporating feature-level correspondence and correlation learning schemes within detection and segmentation contexts that target such scenarios. Particularly, we explore the impact of incorporating cross-channel, inter-spatial, as well as global correlation learning schemes, that which may be inferred either implicitly, by promoting CNNs to capture, and dy-

namically integrate, feature-level relations using design characteristics and objectives, or explicitly by directly modelling such relations in a learnable manner. In this study, we explore these hypotheses within multi-modal, as well as, 3D imaging based localisation contexts. We explore a joint analysis approach in which cross-band information is dynamically and gradually utilised to perform localisation of 3D objects in multi-spectral images that observe different layers of a 3D scene. Specifically, we consider the detection and segmentation of 3D solar Active Regions (ARs) in multi-spectral atmospheric imagery as a case-study to evaluate the benefits of such analysis. We also investigate the influence of directly modelling cross-channel, inter-spatial, and global correlations to infer feature importance, within volumetric imaging contexts. Particularly, we consider 3D pulmonary nodule computed tomography imaging scenarios as a case-study to evaluate the proposed correlation learning schemes. In the remainder of this section, we briefly highlight the contributions of this thesis with respect to the motivations and rational discussed in Section 1.1.

1.3 Contributions

1.3.1 3D object detection in multi-layer multi-spectral images based on cross-band dependencies

We present a multi-tasking deep learning framework that targets the detection of 3D objects using multi-spectral imagery that observe sparse 2D layers of a 3D scene (e.g., multi-spectral solar images). The proposed method performs the detection in a two stage pipeline based on regional convolutional networks. We exploit the dependencies between the different imaged layers (imaging bands) to produce 3D detections where different image bands (layers) have their own set of results. The proposed method analyses band-specific information and dynamically aggregates and jointly analyses cross-band features at different semantic levels throughout the detection network to capture spatial correspondence and feature-level correlations with respect to the detection objective. A training strategy is presented to optimise the weights of the proposed multi-tasking detector more effectively with respect to each of the individual tasks in contrast the the classical training approach in which all tasks are optimised simultaneously, with no additional computational overhead. Extensive analysis demonstrate the effectiveness of the proposed paradigm using stat-of-the-art CNN backbones and different data types and numbers of imaging modalities.

1.3.2 3D object segmentation in multi-layer multi-spectral images based on cross-band dependencies

We investigate the possibility of generalising the proposed joint analysis based detection paradigm (discussed in Section 1.3.1) into the segmentation task of 3D objects in multi-layer multi-spectral images (solar images). We apply the cross-band analysis principles into a multi-tasking encoder-decoder convolutional network to dynamically capture inter- and cross-band relations. The proposed segmentation network also aggregates band-specific spatial information from different resolutions and semantic levels to assist the network in recovering finer details. To overcome the difficulty in designing 3D (i.e., multi-layer) pixel-wise annotations, we propose a weakly-supervised iterative learning strategy in which segmentations are recursively refined starting from bounding box priors. The proposed approach demonstrates feasibility against state-of-the-art methods, over different datasets, and using different levels of supervision.

1.3.3 Object detection using cross-channel and inter-spatial correlations based attention

We present a 3D object detection framework from volumetric images (medical computed tomography). The proposed framework performs the detection as an ensemble of two stages, candidate proposal, and a false positive reduction stage to reduce the number of false alarms. We present a 3D fully convolutional attention block in which cross-channel correlations are explicitly modelled and used to infer feature importance. We also present a 3D spatial attention block based on cross-sectional features (axial, coronal, and sagittal) to infer inter-spatial correlations. We incorporate a joint analysis approach that exploits different levels of spatial contextual information simultaneously to reduce the number of false alarms. In the same line, we present a 3D zoom-in convolutional path to assist the network in effectively capturing spatial information at different scales and semantic levels in a learnable manner. We provide extensive analysis on a public pulmonary nodule detection dataset in which we evaluate and compare different spatial and cross-channel attention strategies, and combinations of both. We demonstrate that both channel and spatial attention techniques can enhance the overall network performance, with channel attention showing more performance gains in contrast to spatial attention, or the combination of both attention strategies. The proposed method demonstrates effectiveness against stat-of-the-art nodule detection and attention methods.

1.3.4 Object detection using global correlations based attention

Convolutional neural network demonstrate superior spatial representational abilities, However, they suffer in modelling long-range relations due to their inherent locality. On the other hand, Transformer networks (self-attention based networks) [39] demonstrate a great ability when modelling global context and long-range correlations between arbitrary positions. Nonetheless, the tokenisation of image inputs in Transformers degrades the spatial locality information. Moreover, Transformers are computationally expensive due to their quadratic cost with respect to the input size, which makes them difficult to directly adopt for computer vision tasks. In this work, we explore the possibility of exploiting global context to infer self-attention (i.e., Transformers) to perform the detection task from volumetric medical images (pulmonary computed tomography). We present a hybrid Transformer CNN architecture in which both, the representational power of CNNs and the Transformer ability in modelling long-range correlations, are simultaneously used to effectively perform the detection task in an end-to-end 3D manner. We propose two hybrid Transformer CNN variations in which we evaluate the trade-off between incorporating a deep Transformer design using 3D features of relatively small resolution, against using information from multiple –and relatively larger– spatial scales with a shallower –less computation demanding– Transformer networks. We also investigate the possibility of jointly exploiting localised attention mechanisms (convolutional based cross-channel and inter-spatial attention) with global context based attention (self-attention) to further enhance the detection performance. An extensive analysis, including an ablation study in which different Transformer configurations are evaluated, as well as a comparison against state-of-the-art nodule detection and attention methods, is provided. To the best of our knowledge, we are the first incorporate Transformers into the 3D pulmonary nodule detection problem.

1.4 Outline

The remainder of this thesis is outlined as follows:

Chapter 2 – *Machine Learning and Object Localisation Background*: presents required background materials related to the work conducted in this thesis, including basic concepts of neural networks, convolutional neural networks, and deep learning based object detection and segmentation methods. This chapter also discusses state-of-the-art multi-spectral and

volumetric object localisation deep learning based methods, as well as an overview of deep learning based attention mechanisms.

Chapter 3 – *Application Related Background*: provides background information and existing works associated with problems investigated in this thesis. Particularly, we consider two applications of which we investigate the possibility of incorporating cross-feature correspondence and correlation learning schemes within. These include multi-spectral imaging scenarios that observe sparse 2D layers in a 3D scene, in this case we look into multi-spectral imaging of solar active regions in the solar atmosphere. We also consider volumetric medical imaging scenarios, specifically, we look into 3D computed tomography imaging of pulmonary nodules. This chapter identifies gaps and challenges and discusses potential improvements associated with existing solutions that target these applications.

Chapter 4 – *MLMT-CNN for Object Detection in Multi-layer and Multi-spectral Images*: presents a Multi-layer Multi-tasking (MLMT) framework, MLMT-CNN, to tackle the 3D solar active region detection problem from multi-spectral images that observe different layers of the 3D solar atmosphere. A novel joint analysis approach in which inter-band and cross-band spatial correspondence and correlations are dynamically analysed to perform the detection is proposed. Extensive analysis is presented in which the proposed framework is compared to state-of-the-art detection methods, using different cross-band feature aggregation strategies, and over different datasets.

Chapter 5 – *MLMT-CNN for Object Segmentation in Multi-layer and Multi-spectral Images*: Extends our joint analysis approach, MLMT-CNN, to handle the pixel-wise classification task from multi-layer multi-spectral solar images. A weakly-supervised recursive training approach is proposed to overcome the difficulty in producing dense segmentation annotations. Comparative analysis against state-of-the-art segmentation methods and using different levels of supervision and for different applications, is provided.

Chapter 6 – *AttentNet for Pulmonary Lung Nodule Detection Using 3D Cross-channel and inter-spatial Convolutional Attention*: presents AttentNet, an automated, two stage, 3D pulmonary nodule detection framework from CT images. Two 3D convolutional attention approaches are proposed to capture cross-channel and inter-spatial correlations and subsequently assist the network focusing on learning effective features and produce a more accurate prediction. Extended analysis is carried out in which the proposed methods are

evaluated on a public dataset and compared against state-of-the-art nodule detection and attention methods.

Chapter 7 – *TransCNN for Pulmonary Lung Nodule Detection Using Self-attention*: presents a hybrid CNN Transformer pulmonary nodule detection framework in which 3D inputs are leveraged to perform the detection task, taking advantage of the spatial representational powers of CNNs and the Transformer ability in modelling global context. An extended experiment and comparison against localised attention mechanisms as well as existing state-of-the-art nodule detection methods, is presented.

Chapter 8 – *Conclusions and Future Work*: concludes the work presented in this thesis and discusses potential future directions and improvements.

Chapter 2

Machine Learning and Object Localisation Background

Contents

2.1 Introduction

In this chapter, we provide the background materials required for this thesis. We start by discussing the basics of neural networks and convolutional neural networks in Sections 2.2 and 2.3. Object localisation (detection and segmentation) from multi-modal data as well as from 3D volumetric imagery form the majority of this thesis. Therefore, we present an overview of deep learning based object detection and segmentation and compare different state-of-the-art methods in Sections 2.5.2 and 2.5, respectively. In Section 2.5.3, we discuss methods of weakly-supervised object segmentation of which we adopt in this thesis. Moreover, we overview object detection and segmentation methods that specifically target multi-modal imagery in Section 2.6. In Section 2.7, we investigate localisation methods designed to handle data of 3D nature. Finally, in Section 2.8, we provide an overview of deep learning based attention mechanisms that are used in this thesis. Particularly, we focus on convolutional based localised attention (cross-channel and spatial attention) in 2.8.1. We also discuss global context based attention in 2.8.2. In Section 2.9, we summarise this chapter.

2.2 Neural Networks

Digital neural networks (NNs) are algorithms designed to emulate the biological neural network in the human brain. They are modelled as acyclic graphs that consist of a number of connected nodes known as neurons, where each connection is weighted by a value that is found through an optimisation process that is governed by a pre-defined objective [44]. In a neural network, each node represents a linear combination of the output values from the previous connected nodes (see Fig. 2.2. Non-linearity is imposed through some activation function e.g., Sigmoid, Rectified Linear Unit (ReLU) (see Fig. 2.1), at the firing end of each node. Accordingly, a neuron can be viewed as a mapping function between an input and an output space, in which information from the input are embedded. Typically, neurons are organised in the form of groups known as layers. Deeper models can therefore be created by stacking a number of layers, such that all neurons in subsequent layers are pair-wise fully connected, while inter-layer neurons share no connections.

A network can include any number of layers, and individual layers may contain any number of neurons. However, the choice of number of layers and neurons must be carefully engineered with respect to the complexity of the task and data.

A network is trained using a gradient descent optimisation process that involves a forward-

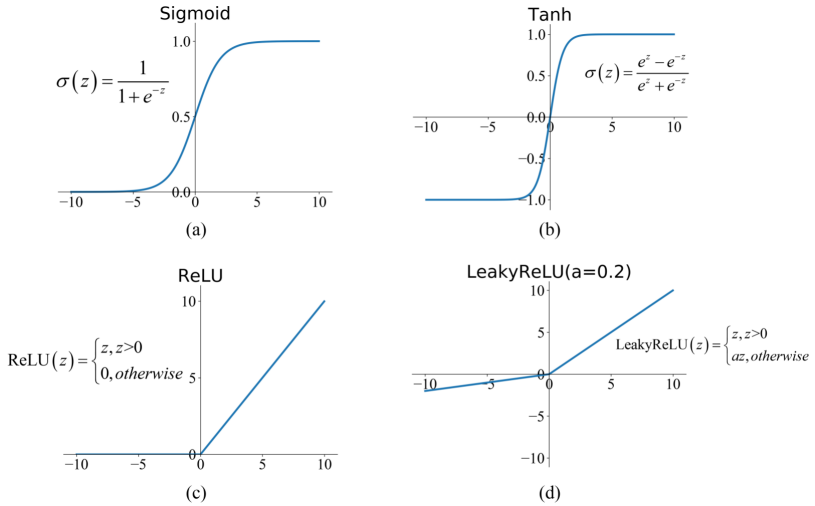

Figure 2.1: Visualisation of different activation functions, Sigmoid (a), TanH (b), ReLU (c), and LeakyReLU (d). Adopted from [45].

pass and back-propagation step. Initially, all network parameters are assigned small random real values. Inputs are fed into the first layer of the network, and are sequentially passed throughout the network's layers up to the last layer, where the final output is produced. This is known as the forward-pass. The quality of the prediction (output) is then quantified using some objective function (loss function), in which the error between the input the target output is computed. The gradient of the loss function is then computed with respect to all weights –individually, using the chain rule– in the neural network, and is used to update (tweak) the network's parameters such that the overall error is decreased. This is known as back-propagation. Completing a single forward pass and a back-propagation step represents a single training iteration. Training the network with all training samples represents a single training epoch. This process is repeated until the model being trained converges to a desired (or a minimum) loss.

Common loss functions may include, e.g., mean squared error and mean absolute error losses (regression problems), binary cross entropy and hinge loss (binary classification problems), and categorical cross entropy loss (multi-class classification).

The process of minimising (or maximising) the objective function of which the network is trained with, is known as the optimisation problem. Different optimisation algorithms control different settings such as the learning rate when computing the model's parameters update, these may include, e.g., Gradient Descent (GD), Stochastic GD (SGD) [46], Mini-batch GD [47], and Adam [48].

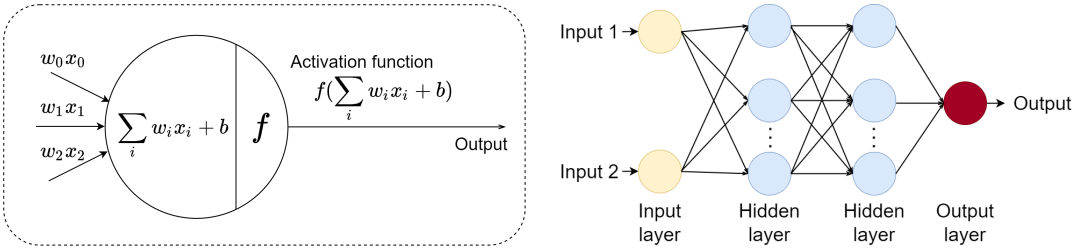

Figure 2.2: An overview of a single perceptron (left) and a multi-layer neural network (right). A perceptron computes the weighted sum of the input values (x) added to a bias value (b). The resulting value is then non-linearly mapped using some activation function. Accordingly, each layer of the neural network contains a number of perceptrons that bear learnable weights and biases.

Using Gradient Descent (GD), a local minimum of a differentiable function (in this case, a loss function) is found by iteratively taking steps towards the opposite direction of the function's gradient with respect to the coefficients of the tested model (the steepest descent), computed using the entire training dataset. This approach however is computationally expensive particularly for large datasets. SGD and mini-batch GD were proposed extending on GD, to overcome the high computational cost associated with it, by using a single data point (i.e., SGD) or a subset of data points (i.e., mini-batch GD) at a time to compute the derivative of the loss function. Several other variations of GD were proposed to enhance the quality of the optimisation process. Momentum based optimisers use the moving weighted average of past gradients to compute the parameter update values aiming to accelerate the convergence towards the relevant direction (i.e., reduces gradient's fluctuation). Adam adds to this by using the squared gradients to adopt the value of the learning rate used to update the networks parameters.

Neural networks are able to learn complex models and perform non-trivial tasks (e.g., classification) robustly, however, a major drawback of a neural networks is that the number of parameters can grow notably particularly for large input (e.g., images).

2.3 Convolutional Neural Networks

A convolutional neural network (CNN) is a generalised form of neural networks that is designed by combining neural networks with image processing concepts. In a standard neural network, 2D images inputs are flattened into 1-dimensional vectors to enable processing them by the network layers. This however, degrades the inherent spatial correlation in imagery data. It also leads to a significant increase in the number of trainable parameters, making it more resource demanding.

On the other hand, CNNs exploits spatial connectivity by incorporating kernels of weights in a sliding-window fashion to capture spatial features from input space, in this case, images (see Fig. 2.3) [44]. Weights (i.e., kernels) are shared in the different positions of an input image, reducing the parameter overhead by significant margin. The essence of this process is that if an important pattern –with respect to a given objective– exists in the investigated image , a kernels learns to capture it in any possible position in the image.

A convolutional neural network consists of several hidden layers, such as convolutional layers, pooling, activation, regularisation layers, and fully connected layers. Pooling layers (e.g., maximum and average pooling) perform dimensionality reduction by filtering feature maps extracted by previous convolutional layers using some statistical operation. Regularisation layers (e.g., dropout and batch normalization layers) promote the generalisability of the network and therefore reduces the risk of over-fitting. Dropout [49] layers achieves that by randomly omitting neuron values during training to prevent complex co-adaptations of the network on training samples aiming and promote a favorable regularisation effect. Batch normalisation [50] standardises the mean and variance of the input data of to a layer promoting faster learning, more independant learning of each layer, makes the network less prone to the choice of weight initialisation, and a regularisation effect. Using combinations of these layers allows the network to dynamically embed information from different spatial scales and semantic levels [51].

CNNs have become the default method of choice as the backbone of most computer vision tasks (e.g., object detection [25, 26, 52], and object segmentation [27, 28]) due to their spatial representational powers and ability of learning task related features without human intervention (e.g., hand-crafted features). Nonetheless, designing CNNs requires careful engineering and hyper-parameter optimisation that is typically guided by trial and error. CNNs also lack the ability of encoding positional information, and are prone to spatial variance. Additionally, training CNNs requires considerable amounts of annotated data and computational resource.

2.4 Object Detection

Object detection involves two main tasks, localisation, in which the coordinates of an object are determined (e.g., bounding box), and classification, in which the category of the localised object is predicted. In the past two decades, object detection has evolved drastically, from sliding window and hand-crafted features (e.g., Haar features [16], and HOG [8]) based detection to the more advanced deep learning based detectors (e.g., [25, 26, 52–54]). Generally, DL based

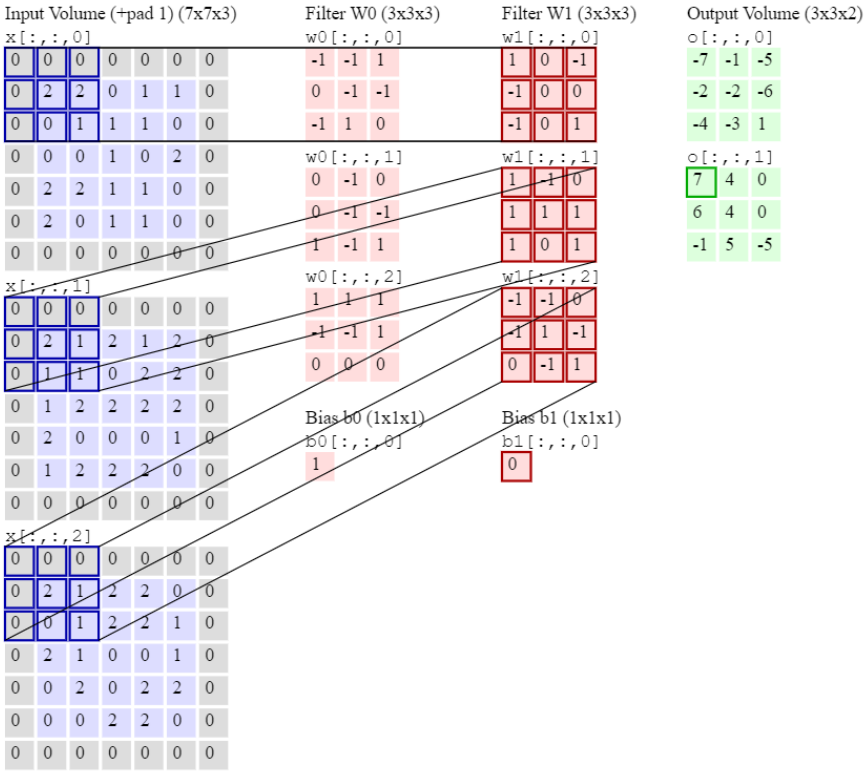

Figure 2.3: Visualisation of the convolutional operation using two convolutional kernels of size 3x3x3 (each), and stride of size 1.

detectors exploit on convolutional neural networks to analyse images. These may be split into two main categories, 1) two stage detection, in which images are analysed in two steps, region proposal (generate a set of suspicious locations) and a final classification stage, and 2) one stage detection, where a DNN learns to regress object locations and classes in a single step. See Fig. 2.4.

2.4.1 One-stage object detection

Liu et al. [25] proposed single shot detectors (SSD), to perform object detection in a single stage fashion. In their approach, unlike sliding window based approaches, an input image is divided into a grid in which each sub-region (grid cell) is analysed to determine the presence of objects within that particular cell. To allow the detection of multiple objects and dynamically detect objects of different size, SSD assigns multiple boxes of different aspect ratios and scales to each of the grid cells, these are known as anchor boxes, and are commonly used in CNN based object

detectors. For each object in the image, the anchor with the highest overlap is assigned the class of that object. Input images are first analysed by a feature extraction CNN (e.g VGG [33] or ResNet [34]). The resulting feature maps are then passed into an SSD detection network, the detection network consists of 5 consecutive convolutional layers, in which the feature map is further down-sampled, so that each layer bears a particular receptive field. The idea is to allow the network detect objects of different scales and improve the compactness of the predicted bounding boxes. As per, the earlier layers realise smaller receptive field and can therefore better represent smaller objects, and vice-versa. The detection layers are trained to regress offsets that are used to tweak the location and size of the pre-defined anchor with respect to the ground truth box. They also predict a class probability for each of these anchors. Redundant detections are finally eliminated using non-maximum suppression (NMS).

A similar approach known as YOLO (you only look once) was proposed in [26]. YOLO follows the same grid approach, however, YOLO consists of a single detection layer that in contrast the multiple detection layers used in SSD. Both SSD and YOLO demonstrate a great potential in object detection. YOLO tend to outperform SSD, when considering the speed of the detection, however, both detectors suffer when detecting small objects as well as neighboring objects, due to the grid strategy followed in both approaches [53, 55, 56].

A more recent single stage detector, CornerNet, was proposed in [53]. Unlike SSD and YOLO, CornerNet detects object in the form of a pair of keypoints, the top-left and the bottom-right corners of objects, avoiding the need of engineering anchors or dividing the image into a grid. An encoder-decoder CNN is trained to predict two heatmaps that embed the locations of the top-left and bottom-right corners, respectively. Additionally, the network predicts offsets that are used to tweak the predicted corner locations. The final bounding boxes are therefore derived using the two corner locations. ConrnerNet shows better performance comparing to both single stage detectors, SSD and YOLO, but lower results comparing to other two stage detectors [53].

2.4.2 Two-stage object detection

Girshick et al. [24] proposed a two stage detector, Regional Convolutional Neural Networks (RCNN) , to solve the detection problem in two steps, region proposal and detection. In the region proposal stage, suspicious location in which objects may exist are predicted, in a class agnostic manner. These are then passed into the detection stage to predict for each location the class it belongs to. In RCNN, a selective search method [57] –based on hierarchical clustering of pixels based on their morphological characteristics, color, and texture– is used to perform the

region proposal task. Proposed regions are then extracted, directly from the image (i.e pixel-level), and are passed, one by one, into a feature extraction CNN (e.g., VGG or ResNet) to extract deep features. The extracted feature maps are passed into a fully connected classification layer to perform the final class prediction. This approach however, demanded extensive time and computational resources.

Fast RCNN [58] fixed this problem by extracting the feature maps for the entire input image first, followed by pooling the suspected regions (region proposals) directly from the convolutional feature maps. Since the down-sampling rates are known, the relative locations of the proposed regions can simply be determined. This operation is known as Region of Interest (RoI) pooling and has gained popularity in object detection for its simplicity and effectiveness. These simple modifications have drastically improved both the speed and the accuracy on the initial version of RCNN.

Building on Fast RCNN, Faster RCNN [52] further enhanced the detection strategy and performance by replacing the selective search region proposal approach with a fully convolutional Region Proposal Network (RPN), see Fig. 2.5. RPN takes feature maps produced by a feature extraction CNN as an input. Anchors of different sizes and scales are assigned at every possible location (pixel) on the extracted feature map, enhancing the networks ability in separating neighbouring objects. The RPN is then trained to predict offsets to tweak each anchor with respect to their associated ground truth boxes. Additionally, RPN predicts the abjectness (i.e., *object* vs. *not object* probability) for each anchor, all in a fully convolutional manner. The proposed regions are then pooled –using RoI pooling– and are passed into a detection network. The detection network –a fully connected network– in turn, predicts offsets to improve the compactness of the initially proposed bounding box (proposed by the RPN), as well as the class of which a detected object belongs to. The use of RoI pooling and fully connected layers, as well as the fact that Faster RCNN consists of two stages, increases the overall time overhead. However, Faster RCNN has repeatedly demonstrated outstanding performance, in terms of accuracy, in contrast to single stage detectors [59, 60].

A variation of Faster RCNN, R-FCN (region-based fully convolutional network) was proposed in [54]. The idea is to improve the speed of Faster RCNN by reducing the detection into a fully convolutional single stage, while retaining high detection accuracy using the anchors strategy followed in Faster RCNN. R-FCN exploits the RPN concept from Faster RCNN and modifies it by replacing the fully connected classification layer by a convolutional layer that is trained to predict a positive-sensitive score map. This map acts as a scoring map in which re-

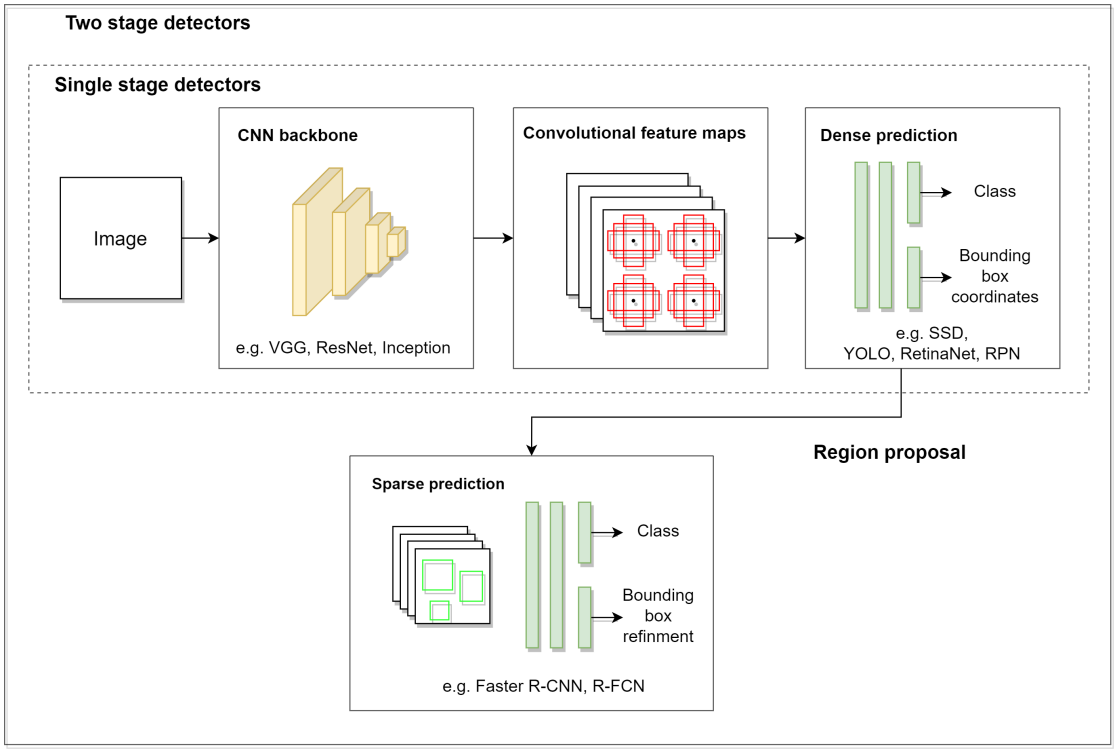

Figure 2.4: Overview of popular object detection approaches.

gions of interests (i.e., proposals by the region proposal layer) are each pooled from. The bins of the pooled map are used to form the final classification score by a simple averaging operation. Results show that while R-FCN outperforms Faster RCNN in terms of speed, Faster RCNN still yields higher detection accuracy.

Generally, while one stage detectors (e.g., SSD) can achieve higher detection speed in contrast to two stage detectors, two stage detectors (e.g., Faster RCNN) can achieve higher accuracy in contrast to single stage detectors [53, 54, 59, 60], see Fig. 2.6. Nonetheless, such methods target the detection from 2D images, and are not directly applicable for 3D data that observes sparse 2D layers of a 3D object (e.g., multi-spectral solar data) or volumetric images. In this thesis, we explore the possibility of incorporating and generalising DL based detection concepts into multi-modal imaging scenarios that observe different layers of 3D objects, as well as 3D volumetric imaging scenarios.

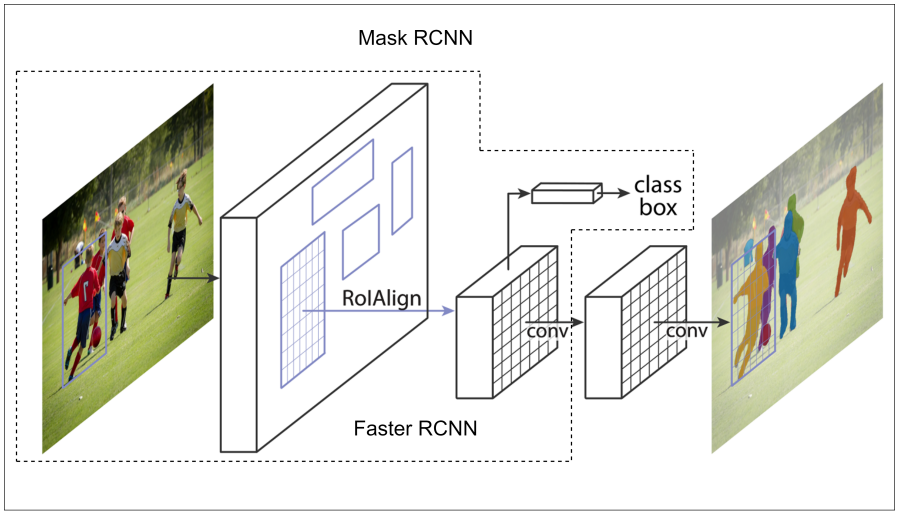

Figure 2.5: An overview of Faster RCNN [52] and Mask RCNN [61]. Input images are first processed by convolutional layers for feature extraction. These are then passed into the region proposal network to predict suspicious locations that may include objects. The proposed locations are then pooled from the feature maps and are passed into parallel detection and segmentation networks to classify bounding boxes and predict segmentation masks for each suspected location. Faster RCNN [52] is equivalent to Mask RCNN when discarding the mask prediction network from the architecture. Adapted from [61].

2.5 Object Segmentation

Object segmentation is the process of assigning a class label into the individual pixels of a given image [62]. Early unsupervised approaches involved simple thresholding, clustering methods (e.g., K-means and Fuzzy C-means), graph based methods, and region growing methods. Typically, these methods rely on hand-crafted features and heavy image processing. More recently, object segmentation has evolved dramatically thanks to the emergence of convolutional neural networks and their exceptional representational power. See Fig. 2.7. Generally, CNN based segmentation may be split into two categories, semantic segmentation, in which all objects of a unique class are treated as a single entity, and instance segmentation, where different objects (instances) of the same class are detected as individual entities.

2.5.1 Semantic segmentation

Starting from common CNN classification architectures (e.g., VGG [33]), where convolutional feature maps are flattened into 1-dimensional vectors and are passed into fully connected layers to predict image level –also 1-dimensional– labels, Shelhamer et al. proposed generalising this concept by substituting the fully connected layers at the end of the network (prediction heads)

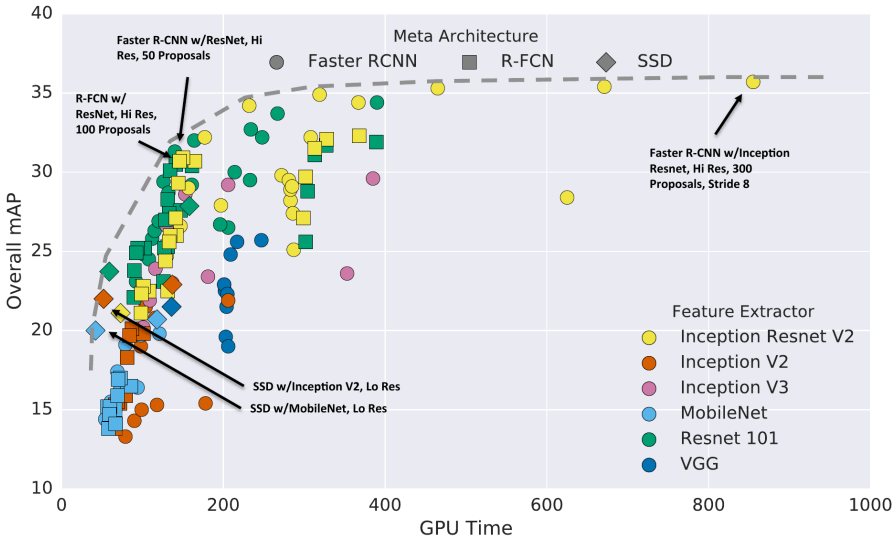

Figure 2.6: Performance comparison of popular object detectors, Faster RCNN [52], R-FCN [54], and SSD [25], in terms of detection accuracy and inference time. Different marker styles indicate different detectors, different colors represent different feature extraction CNN backbones. Each detection framework and CNN backbone pair may be indicated by multiple points where each point represent changes in hyperparameters (e.g., input size). Adopted from [60].

with convolutional layers to predict 2D segmentation masks directly from 2D feature maps. Using only convolutional layers promotes faster training by reducing the total number of trainable parameters. Moreover, unlike fully connected layers, convolutional layers are only locally connected, and can therefore straightforwardly process arbitrary sizes of images. The proposed fully convolutional network (FCN) consists of a standard CNN encoder network, in which a stack of 5 convolutional layers is used to extract features from the input image. Each convolutional layers is followed by a max-pooling layer used to reduce the dimensions of the feature maps. A deconvolutional layer is then used to up-sample the intermediate feature map into the same dimensions of the original image, allowing the prediction of per-pixel probabilities. Finally, a segmentation network of 1x1 convolutional layers is used to predict the final 2D segmentation map. At the input level, images are padded to allow the production of larger feature maps at the end of the feature extracted CNN.

FCNs demonstrated a great potential in image segmentation and attracted a lot of research efforts in which variations and improvement were proposed upon the concept of fully convolutional segmentation. Chen et al. [63] proposed Deeplab, by incorporating a number of modifications to the standard FCN approach in an attempt to overcome some limitations in the original design. Particularly, the repeated use of dimensionality reduction layers (max-pooling) in FCN

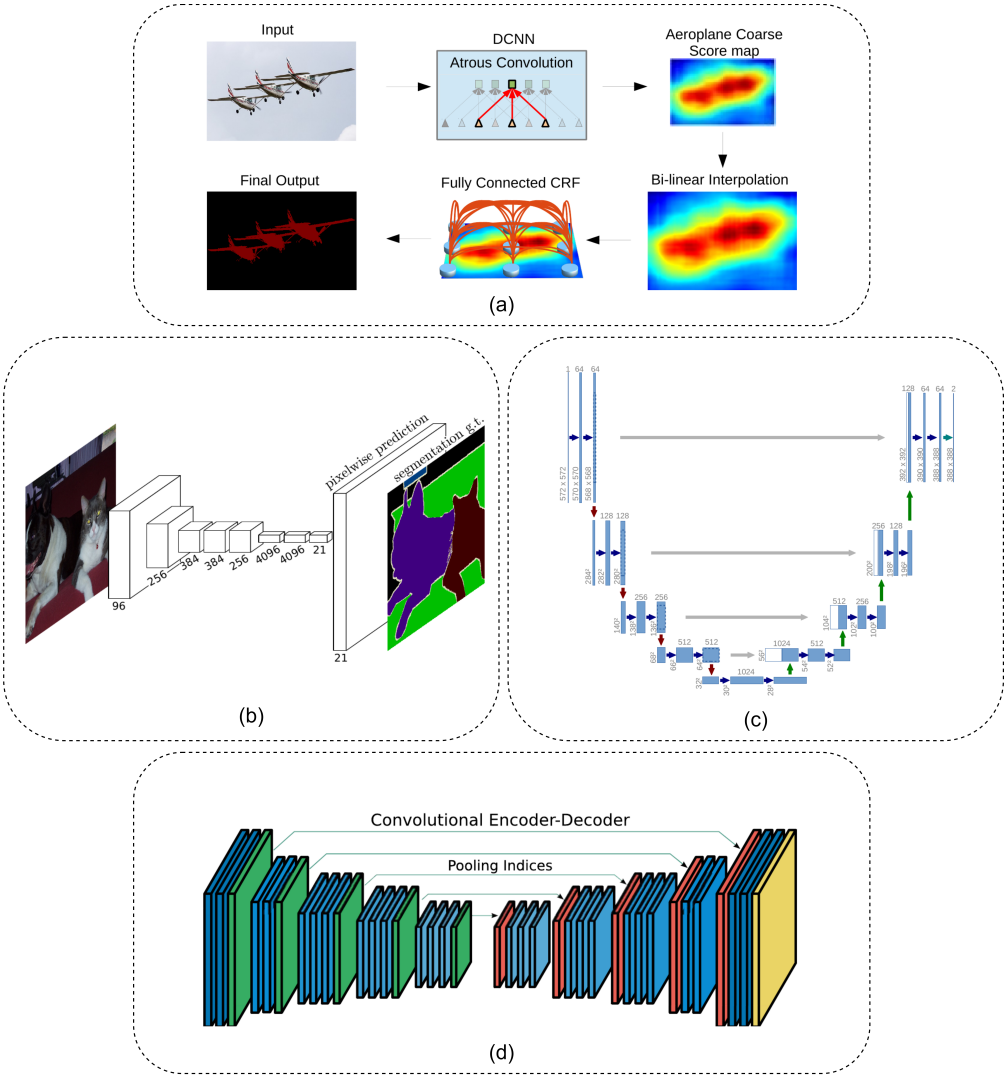

Figure 2.7: Overview of popular object segmentation networks. Adopted from DeepLab [63] (a), FCN [27] (b), U-Net [28] (c), and SegNet [29] (d). The figure is best viewed with close-up and using color.

degrades the low-level spatial features which are important in the context of segmentation. This leads to poor details in the up-sampled feature map and hence, course segmentations. [63] addressed these limitations by modifying the stride value in the last two pooling layers such that the dimensions of their input feature maps is not reduced (i.e., stride=1), and discarding the input padding step used in the original FCN. This allows having larger feature maps as an input to the prediction layer, however, it also increases the computation overhead of the network. To overcome this issue, authors exploited dilated convolutional layers, in which the receptive field is efficiently expanded allowing the extraction of multi-scale spatial embeddings.

Another segmentation approach was introduced in [29], SegNet. Building on FCNs, SegNet incorporates an encoder-decoder design to perform the segmentation. Unlike FCN, where deconvolutions are used to up-sample feature maps, the decoder in SegNet gradually reconstructs the semgentation mask, non-linearly, using pooling index maps from the the correspondent encoder layers to indicate the new pixel values in the up-sampled feature map. Each up-sampling layer is followed by a convolutional layer to learn new dense feature maps. Finally, a 1x1 convolutional layer is used to perform the final prediction. The idea is to avoid having to learn the up-sampling task by non-linearly up-sampling using the aforementioned un-pooling operation. While [29] demonstrated that their approach improves the segmentation of low-level features (e.g., edges of objects), however, the up-sampling procedure followed in SegNet disregards the adjacency information in the processed feature maps and can therefore impact the final quality of the segmentation.

A popular approach known as U-Net was first proposed in [28] for medical images segmentation. U-Net has a similar encoder-decoder architecture to that in SegNet, however, U-Net exploits deconvolutional layers to perform up-sampling in the decoder part in a learnable fashion. The novelty in U-Net relies in the incorporation of skip connections that connect convolutional blocks from the encoder to their correspondents in the decoder part (see Fig. 2.7). This allows the aggregation of spatial features from different semantic levels (low- and high-level features) and spatial scales, assisting the network in recovering fine details that are lost due to the repeated down-sampling operations [64]. U-Net has demonstrated great potential in object segmentation and has attracted considerable research efforts in which variations and extensions were proposed to handle different types of data and applications [65–67].

2.5.2 Instance segmentation

He et al. proposed Mask RCNN to handle the instance segmentation problem in two stages, object detection followed by segmentation [61]. Building on Faster RCNN (see section), a fully convolutional segmentation branch was integrated into Faster RCNN's detection framework. First, feature maps of detected regions are sampled from the convolutional features extracted by the detection CNN backbone, these are then directly used to predict the final instance masks, see Fig. 2.5. Mask RCNN demonstrated a great performance over various datasets and applications.

In the same line, Dai et al. proposed performing instance segmentation as a cascade of three tasks [68]. A shared CNN is first used to extract feature maps that are then used to predict instance locations in the form of bounding boxes. The predicted locations (regions of interest) as well as the convolutional features are then used as an input into a class-agnostic segmentation network that realises foreground and background locations. The resulting masks are used to mask out background locations from the convolutional features. The masked feature maps then used as an input into the final classification network where instance classes are predicted.

Gao et al. proposed SSAP, single-shot instance segmentation using affinity pyramid [69]. Their approach performs instance segmentation based on a multi-tasking U-shaped network. The network is trained to predict an semantic segmentation mask, as well as an pixel-pair affinity pyramid in which the probability of two pixels belonging to the same instance is predicted over multiple feature resolutions, and at every possible location. The resulting semantic mask and affinity pyramid are then processed using a cascaded graph partition module to divide the predicted semantic mask into instances. Their method achieved promising results on the Cityscapes dataset.

Wang et al. proposed SOLO (segmenting objects by locations) to solve the instance segmentation problem in a single stage fashion [70]. Images are first divided into a uniform S x S grid. A cell is associated with a particular object instance if the center of that object falls within that cell. A CNN is then trained to predict cell-wise semantic classes forming a classification map of shape C x S x S (where C is the number of classes). Additionally, S x S class-agnostic instance masks (each of size H x W) are predicted such that each mask corresponds to a single cell (instance) in the S x S grid. Finally, the semantic class probability mask is used to determine the class of each of the predicted instance masks.

Generally, all cited approaches show promising results various applications. Single stage instance segmentation methods are characterised by their computational efficiency in contrast to two-stage approaches, however, two stage methods show higher accuracy particularly when

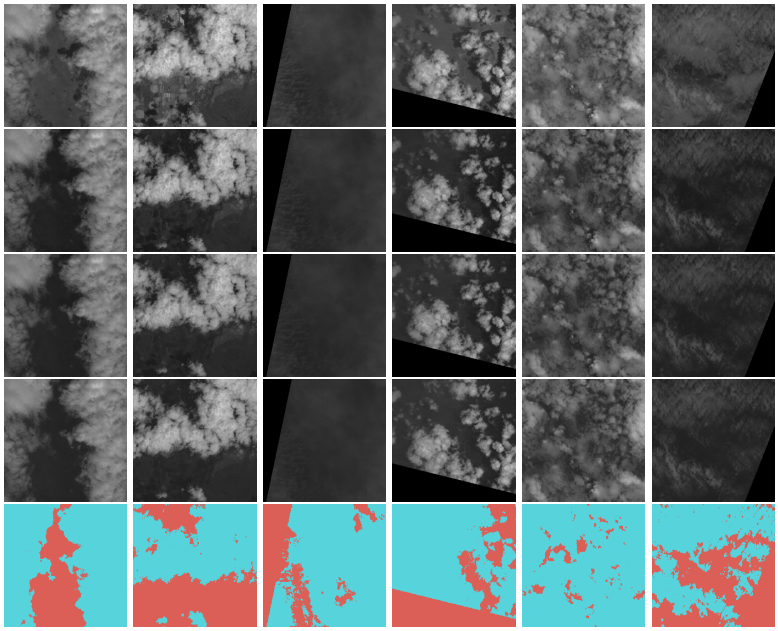

Figure 2.8: Cloud-38 [71] multi-spectral cloud segmentation dataset. Each column represents a single multi-spectral sample. Rows, from top to bottom, show the bands Near Infrared, Red, Green, and Blue, respectively. The last row represents ground-truth masks bearing two main classes, cloud (blue) and background (red).

handling small objects.

2.5.3 Weakly-Supervised Object Segmentation

DL provides powerful solutions for object segmentation. Nevertheless, training segmentation CNNs requires large numbers of densely labelled samples. Creating such dataset requires extended amounts of labour and time. Weakly-supervised learning is commonly used to address these issues, where ill- or partially-labelled datasets are used to provide supervision signals during the training process. [72] proposed two CNN based approaches for segmenting multi-spectral Landsat-8 images using U-Net from weak labels. Image bands were combined by finding the median pixel values across 7 Landsat-8 bands, and divided into 50x50 patches to be used as an input for a U-Net model. Available 78 classes were reduced to two classes (*crop land* and *non crop land*) for convenience. In the first approach, while segmenting the whole image, the U-Net was provided with the label of a single pixel chosen randomly within the pixels of each class, to compute the training loss using that pixel only. On the other hand, the second approach was based on image-level label, where the U-Net was converted to a classifier by the addition

of an extra dense layer. Segmentation was obtained by thresholding a Class Activation Map (CAM) computed as a weighted sum over the last convolutional layer filters, using the weights from the final dense layer. Results demonstrate that the pixel-level based approach achieves closer performance to the fully supervised U-Net baseline in contrast to the image-level label based approach.

Additionally, [73] proposed a CNN iterative training technique to perform segmentation using weak labels estimated by finding the agreement between GrabCut [74] and a segment proposal method known as Multi-scale Combinatorial Grouping (MCG) [75]. The resulting masks were then used to initialise an iterative training of a CNN, such that training label is provided by the previous network's prediction, and is therefore refined as the network gets better over the iterations. They evaluate this approach on the Cityscapes dataset [76], scoring 96.9% of the fully supervised performance. Moreover, similarly to [72], they demonstrate that using pixel-level weak label outperforms image-level label.

This iterative training approach is commonly used to tackle the segmentation problem from weak labels. Both [77] and [78] demonstrated that iteratively training a segmentation network from weak labels estimated based on bounding boxes priors can achieve close performance to full supervision. Both works show that even when naively converting bounding boxes into rectangle masks, the segmentation network was still able to recover segmentation masks gradually throughout the iterative training. These findings suggest that this iterative approach may be an opportune solution to perform segmentation when densely annotated data is not straightforwardly applicable. In this work, we further explore DL solutions to perform the segmentation task from multi-spectral imagery using weak annotations derived from bounding box priors (Chapter 5).

2.6 Multi-modal Images Based Object Detection and Segmentation

Generally, existing methods that target multi-modal imaging scenarios are designed under the assumption that different imaging modalities observe different aspects (components) of the same sensed, typically 2D, scene. Some of these methods process multi-modal images by stacking the different modalities into multi-channel images. Accordingly, a feature map is jointly extracted from these bands with all bands contributing equally towards the feature map, and a single localisation result is produced for the composite image. In [71, 79], this strategy was used to segment clouds from RGB and NIR (near infrared) images from the multi-spectral dataset

Cloud-38 (examples of the Cloud-38 multi-spectral dataset are presented Fig. 2.8), while in [80, 81] it allowed detecting power plants from respectively 3 and 7 image bands. The use of all 7 channels allowed [81] outperforming [80] on the same dataset and demonstrated the potential for DNNs to improve localisation results by exploiting more image bands.

A fusion strategy at decision-making level was proposed in [42] to combine independent analyses of multiple bands, each by a band-specific YOLO model. This work aimed at object detection from RGB, NIR (near infrared), MIR (mid infrared), and FIR (far infrared) images for autonomous vehicles. The detections generated from individual images were combined to form the final set of detections using NMS (non-maximum suppression) to ensure the reduction of duplicates. A strong limitation of this approach is that it fails to directly exploit the inter-dependencies between the bands.

Another feature map-fusion approach was proposed in [41] for pedestrians detection using visible and thermal images. It used the ACF+T+THOG detector [82] for region proposal, which employs HOG features extracted separately from the RGB and thermal images. A DNN classifier then performed the final detection from the fused multi-spectral inputs. Two fusion architectures (early and late) were explored. The early fusion approach corresponds to the previously described strategy of a multi-channel input. For late fusion, the two bands were processed individually in separate subnetworks and their respective feature maps were concatenated before performing the final analysis by a fully connected layer. This strategy obtained better results, authors suggest this is due to the band-specific feature optimisation promoted by the late fusion design. Moreover, they reckon that small misalignments may be overcome as spatial information gets less relevant in late network stages.

However, when comparing image-level, feature-level fusion, and fusion at decision-making level, [40] found on the contrary that image fusion worked best when segmenting soft tissue sarcomas in multi-modal medical images, i.e., Positron Emission Tomography (PET), Computed Tomography (CT), and Magnetic resonance imaging (MRI). These different observations suggest that there is no universal best fusion strategy, it needs to be adapted with respect to the considered task.

In [83, 84], a slow fusion approach was utilised for handling video based classification task. A single input sample consists of a number of consecutive RGB frames, where each frame represents a time step in the image sequence. A CNN is then used to extract spatial features by performing the convolution process over the time axis (i.e., using 3D convolutional filters). No padding was applied to the inputs prior to the convolutional process, accordingly, the time axes

is gradually reduced by applying further convolutions to the resulting feature maps. The resulting features are then passed into a 2D CNN to further extract spatial features and finally perform the classification task. From a feature fusion perspective, the intuition behind this approach is to allow the convolutional filters integrate (fuse) spatial information across the different video frames (temporal axis) by applying the convolution in a 3D manner (using 3D convolution). Both [83, 84] demonstrate promising results in their target tasks. Such methods are not straightforwardly applicable for multi-modal based tasks due to the limited number of frames (i.e., modalities) available for such problems in contrast to video data.

In [43], a transfer-learning based feature fusion approach was proposed to segment coronal holes in Solar Dynamics Observatory (SDO) 7 Extreme Ultraviolet (EUV) bands and line-of-sight magnetogram. A CNN is initially trained with weak labels and single band images, the learned network is then fine-tuned over the other bands to progressively integrate information across the different bands. A final unique prediction is produced by another CNN that takes as input the combined feature maps from all band specialised CNNs. Results were evaluated subjectively by the authors. This approach however fails to directly exploit cross-band relations.

In general, cross-band feature fusion seem to be an effective strategy for handling multi-modal data. Nonetheless, existing works are designed to handle multi-modal images that observe different compositions of a 2D scene, in this thesis, we focus on the problem of detecting 3D objects in multi-modal imagery that observe sparse layers of a 3D scene (Chapters 4 and 5).

2.7 Multi-dimensional Object Detection and Segmentation

Recent advances in imaging technologies, computing hardware, as well as in deep learning attracted research efforts towards exploring modern, and application specific, object localisation solutions. In this section, we review object detection methods that target the detection of 3D objects using 3D bounding boxes from multi-dimensional images (e.g., RGB-D, 2.5D, 3D volumetric, point cloud, and point cloud+RGB data).

Volumetric data (e.g., 3D medical images) is commonly handled in one of three different approaches. The first approach is by slicing through the a 3D image such that individual 2D slices are handled one at a time [85–87]. The second approach is based on fusing location information from different image cross-sections –typically, but not exclusively, axial, coronal, and sagittal– [88–90]. This approach is known as 2.5D (or pseudo 3D). The first and second approach may utilise any 2D CNN to perform the detection. To produce the final 3D bounding box, results

are aggregated from the different input slices using a pre-defined criteria (e.g., voting). Other works use probability feature maps from the different slices to form a dense 3D probability maps. The final 3D detection is then inferred using areas in which the detection probabilities are high [91, 92]. The third approach focuses on explicitly performing the detection in a 3D manner, [93] utilised a two stage 3D CNN based on Faster RCNN to detect 3D bounding boxes directly from medical volumetric images. In the same line, [94] used a single stage detection 3D CNN to predict organ locations, and [95] used a 3D classification CNN to predict 3D bounding boxes in a convolutional fashion (i.e., sliding window).

Other methods were proposed to handle the segmentation task directly from 3D volumes. Motivated by the success of U-Net [28] for 2D image segmentation, [96] proposed a 3D CNN that adopts the key principles of U-Net. The network consists of an encoder-decoder CNN that incorporates 3D convolutions and utilises skip connections between the correspondent encoder and decoder layers. In the same line, V-Net was proposed in [97]. While using a similar architecture of the 3D U-Net, V-Net proposed augmenting the convolutional blocks by residual connections to allow deeper learning, they also exploited dice loss in contrast to the cross entropy used in U-Net. Both U-Net and V-Net demonstrated outstanding performance over several datasets.

Other works focused on handling the 3D detection task for point cloud sensors and applications, [98] proposed a detection method for self-driving cars using Lidar (light detection and ranging) data. Based on point cloud maps, a 2D bird's eye view projection –that embeds height density and intensity information– was generated and passed to a 2D based single stage YOLO [99] detector to predict locations of interest in the form of 2D bounding boxes. Point cloud maps are then used to infer 3D boxes using the 2D bounding box priors. Another Lidar based detector was proposed in [100] based on the two stage Faster RCNN [52] detection approach. First, 3D bounding box proposals are generated using a 3D CNN that takes an input point cloud images that are segmented into foreground only images to reduce the complexity of the task. The resulting proposals are then passed into a detection network for refinement and classification. On the other hand, [101] proposed a Multi-view 3D detector based on fusing features from both RGB and Lidar sensors. The point cloud images are projected to bird's eye view and front view images. The bird's eye view images are then used to find 3D region proposal boxes. These are then projected to the front view and the RGB images. Following, feature maps are extracted –using a CNN– by fusing all three images using an adaptive RoI pooling strategy and are used to produce the final 3D box predictions. All [98, 100, 101] demonstrated promising

results on the self-driving target detection task.

Generally, while 2D based approaches require less computational resources and are simple to design, 3D based approaches attain a superior performance when handling 3D data [102–104]. Nonetheless, most approaches focus on the detection from 3D volumetric or point cloud data, the detection from 3D imagery that observe sparse 2D layers of a 3D object (e.g., multi-spectral solar data) is generally overlooked. Additionally, while most works focus on proposing approaches to handle different types of data, exploring different levels of feature dimensionality, or investigating different deep learning architectures and detection strategies, limited research was dedicated towards directly exploiting data correlations and feature importance to improve the quality of the learning and therefore the detection. We explore these matters in this thesis.

2.8 Attention Mechanisms

Attention in DL may be split into two main categories, local (e.g., CNN based attention) and global attention. In local attention, attention is inferred in a localised manner such that each part of the attention output is computed using a subset of the input. On the other hand, global attention targets long range context in which the entire input is incorporated to compute every part of the attention output. In this section, we review both local and global state-of-the-art attention techniques.

2.8.1 Local Attention

Local attention can be divided into two main categories, channel-wise and spatial-wise attention. The objective of channel-wise attention is refine inter-channel embeddings by directly modelling the correlations between the different channels. Hu et al. proposed squeeze and excitation units, Multi-layer Perceptron (MLP) networks that generate attention maps using a dimensionality reduced descriptor in which cross-channel spatial information is aggregated. The resulting attention maps are then used to recalibrate the channels of the convolutional feature [38].

$$\mathrm{A}_c = \sigma(MLP(AvgPool(\mathrm{F}))^{\mathbb{R}\in C\mathrm{x}1\mathrm{x}1}) \tag{2.1}$$

where A_c is the resulting Channel Attention (CA) map. F represents an intermediate convolutional feature map of size C x H x W. *AvgPool* represents adaptive average pooling. MLP is a multi-layer perception network and $\sigma(\cdot)$ is a sigmoid activation function. Subsequently, a refined feature map F' is computed using element-wise multiplication such that $\mathrm{F}' = \mathrm{F} \otimes \mathrm{A}_c$.

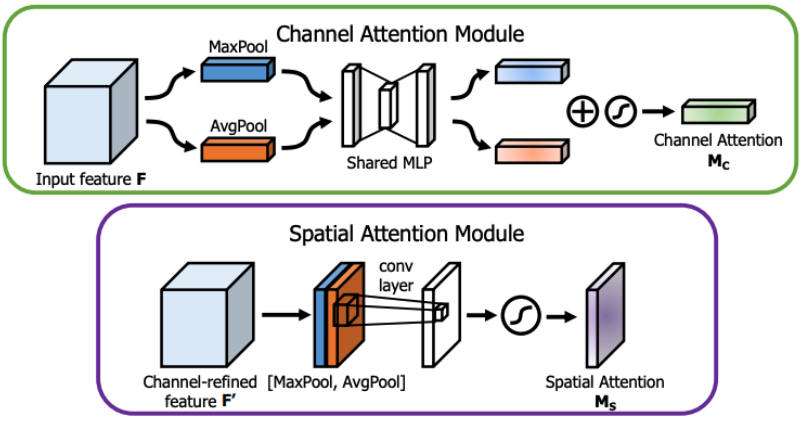

Figure 2.9: An overview of channel attention (top) and spatial attention (bottom) as proposed in [1]. Max and average poolings are used to reduce the dimensionality of the input feature maps. Note that in the spatial attention case, pooling is preformed along the channel axis. The resulting features are passed into a shared fully connected layer (in the case of channel attention), or a convolutional layer (in the case of spatial attention). The resulting embeddings are then passed into a Sigmoid activation function in which outputs are used to infer the final attention maps. Adapted from [1].

Thereafter, Woo et al. proposed Convolutional Block Attention Modules (CBAM), in which they extend the concept of squeeze and excitation by adding a subsequent Spatial Attention (SA) gate [1]. Spatial attention utilises the inter-spatial relationship within the convolutional features to assist the CNN learning *where* to attend within a feature map. This was achieved by projecting the channels of the convolutional feature into a 2D map embedding using channel-wise pooling operators, the resulting map was then passed through a convolutional layer followed by a sigmoid function to generate the final spatial attention map. Moreover, in contrast to squeeze and excitation units, [1] demonstrate that incorporating max pooling in addition to average pooling can provide complementary clues that can enhance the overall attention performance. Accordingly, cross-channel and spatial attention can be described as follows:

$$\mathrm{A}_c = \sigma(MLP(AvgPool(\mathrm{F})) + MLP(MaxPool(\mathrm{F})))^{\mathbb{R}\in C\mathrm{x}1\mathrm{x}1} \tag{2.2}$$

$$\mathrm{A}_s = \sigma(Conv2D_{3\mathrm{x}3}(AvgPool(\mathrm{F}) + MaxPool(\mathrm{F})))^{\mathbb{R}\in C\mathrm{x}H\mathrm{x}W} \tag{2.3}$$

where A_c and A_s are the channel and spatial attention maps, respectively. F represents an intermediate convolutional feature map of size C x H x W. *AvgPool* and *MaxPool* are adaptive average and max pooling layers. MLP is a multi-layer perception network. $Conv2D_{3\mathrm{x}3}$ is a convolutional layer with kernel size of 3 x 3. $\sigma(\cdot)$ is a sigmoid activation function. Note that in the case of spatial attention, pooling operations are performed along the channel axis. Consequently,

a refined feature map F' is computed using element-wise multiplication of the intermediate feature map F and the resulting attention map M, where $M \in \{A_c, A_s\}$. See Fig. 2.9. Moreover, [1] shows that combining both attention approaches by applying them in a consecutive order (i.e., channel-wise followed by spatial attention) can further enhance the overall performance in contrast to using either of the attention approaches individually, or using other combination schemes (e.g., spatial followed by channel-wise attention, or by applying both approaches in parallel). Incorporating channel and spatial attention in CBAM can improve the representational capabilities of CNNs, and therefore enhances their performance [1]. This is in line with the findings of Sun et al. [105] (pulmonary nodule classification in CT images), Lu et al. [106] (pulmonary tuberculosis detection in CT images), Nawshad et al. [107] (COVID-19 detection in X-ray images), Sangeroki et al. [108] (thoracic disease detection in chest X-rays), and Park et al. [109] (generic object detection).

Generally, these methods target 2D images and rely on heavy dimensionality reduction and expensive MLP networks to infer attention, limited research was dedicated to handle attention for 3D data by explicitly exploiting 3D information to infer attention. We further explore these limitations in this thesis (Chapters 6 and 7).

2.8.2 Global Attention

Global attention aims to model long-range correlations and inter-dependencies between arbitrary positions. Vaswani et al. proposed Transformer networks based on multi-headed self-attention for sequence-to-sequence tasks. An input vector is used in three ways, *query*, *key*, and *value*. Accordingly, attention is expressed as a mapping between a *query* and a set of *key* and *value* pairs, to an output vector, by finding the weighted sum of the *values* using weights that are found by a compatibility function of the query and the corresponding key [39]. More formally, the over all attention process is described as follows:

$$\begin{aligned} &\text{MultiHead}(Q,K,V) = \text{Concat}(head_1, .., head_i)W^O \\ &\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \\ &\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \end{aligned} \tag{2.4}$$

where Q, K, and V represent *queries*, *keys*, and *values*, and their correspondent learnable parameters W^Q, W^K, and W^Q, respectively. W^O represents a learnable linear projection process.

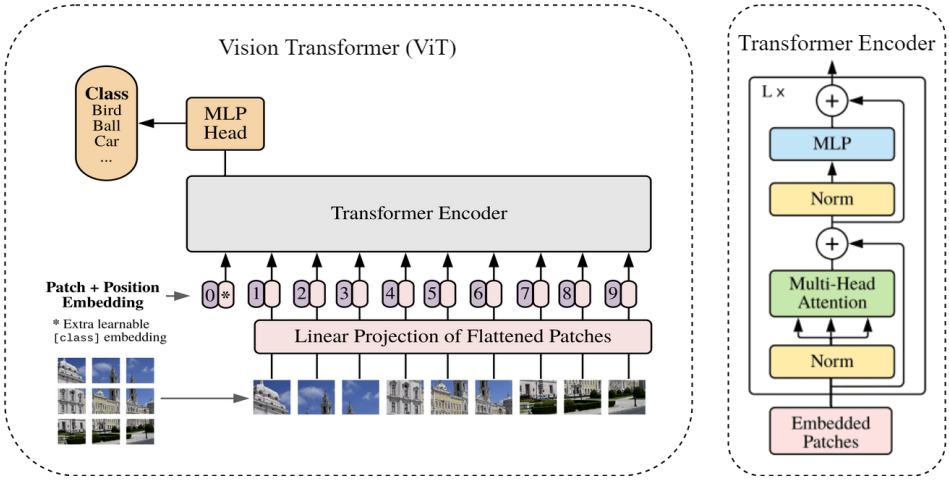

Figure 2.10: An overview of Vision Transformer as proposed in [2]. An input image is split into multiple fixed size patches that are then linearly projected into a new feature space. These are then used along with positional embeddings as an input for a Transformer encoder layer where multi-headed self-attention is performed. Adapted from [2].

Dosovitskiy et al. [2] generalised this concept to computer vision tasks by splitting an image (RGB images) into a sequence of vectorised patches that can then be managed by a pure Transformer (e.g., [39]). see Fig. 2.10.

While Transformers demonstrate a great potential in computer vision tasks, they rely on heavy pre-training and are difficult to scale to large inputs due to their computational cost. [2, 110, 111]. Moreover, vision transformers suffer when modelling local structures due to the tokenisation of input images [112].

A segmentation –Transformer based– network was proposed in [114], Segmenter, in which RGB images are directly, patch-wise, fed into a Transformer network to capture global context. Image patches are first flattened and used as the sequence input to a standard Transformer encoder. The resulting features, along with class embeddings, are then passed into a transformer decoder network that is trained to map patch-level embeddings to patch-level class scores. The final prediction is found by up-sampling the resulting patch-wise class scores into pixel-level probability maps using bi-linear interpolation. Results demonstrate that the performance of this approach is highly dependant on both, pre-training, as well as the patch size used as an input to the transformer. Smaller patch size lead to better performance but at higher computational cost, and vice versa. This observation is in line with the understanding that the image tokenisation in vision Transformers degrades the spatial information leading to a decreased performance.

To tackle these limitations, a number of studies proposed using a hybrid architecture that

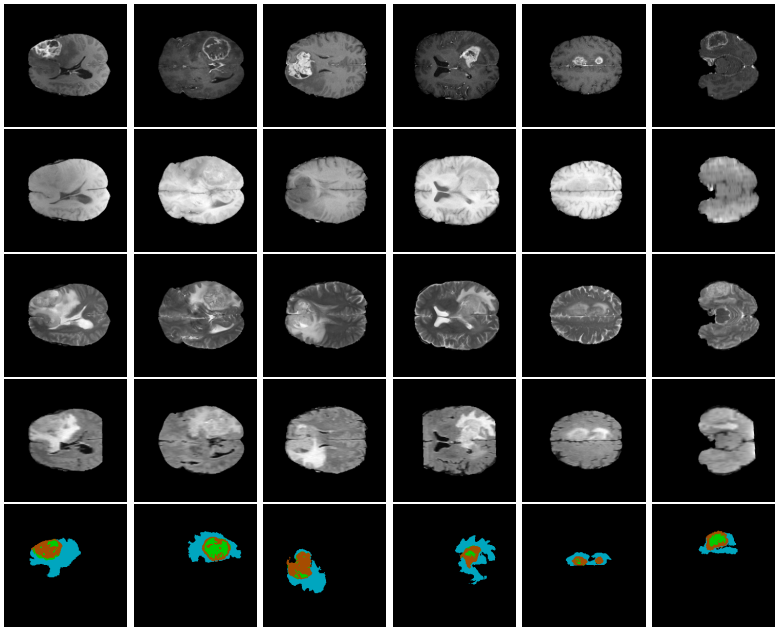

Figure 2.11: BraTS [113] multi-modal MRI brain tumour segmentation dataset. Each column represents a single multi-modal sample. Rows, from top to bottom, show the modalities T1Gd, T1N, T2N, and Flair, respectively. The last row is the ground-truth mask showing 3 main classes, enhancing tumour (blue), the peritumoral edema (brown), and the necrotic and non-enhancing tumour core (green).

combines both, CNNs for their spatial representation power and their relatively low computational cost, along with Transformers for their ability in modelling long-range dependencies.

Carion et al. proposed DETR [115], to handle the detection problem in RGB images using Transformer networks. A 2D CNN encoder was incorporated to embed and down-sample spatial features that are then fed into a Transformer network to capture global context and long-term correlations. The resulting spatial are then flattened, aggregated with positional encodings, and are then used as the input sequence to a Transformer network in which global context and long-range correlations are investigated. Subsequently, the Transformer network performs the detection as a prediction task of a as set –of which size is fixed– of bounding boxes, and using a bipartite matching loss. Authors demonstrate promising results, however, they find that such approach struggles when dealing datasets that target the detection of small objects.

Wang et al. [110] proposed TransBTS, where a CNN encoder was incorporated to extract and down-sample embeddings of 3D MRI (magnetic resonance imaging) images from the multi-modal Brain Tumor Segmentation (BraTS) dataset [113] (examples of the BraTS MRI images are presented in Fig. 2.11). The down-sampled features were then fed into a pure Transformer unit to capture long-range relations and perform attention, the Transformers output was passed

into a decoder CNN to perform object segmentation. This approach takes advantage of both, the representational power of CNNs as well as the Transforrmer's ability in modelling long range correlations. Other works used a similar approach to perform segmentation in 2D images and demonstrated promising results and a great potential of this hybrid approach for different applications, e.g., [111, 116–118].

In this thesis, due to the recent success and the increasing interest in vision transformers, we explore the possibility of incorporating Transformers to handle the detection task of 3D medical images (Chapter 7) and compare the influence of global context based attention in contrast to localised attention mechanisms.

2.9 Summary

In this chapter, required background knowledge has been presented in preparation for proposed methods in the following chapters. We have discussed basic concepts of neural networks and convolutional neural networks as well as deep learning based object detection and segmentation approaches. We have also provided an overview of multi-modal and 3D object localisation, deep learning based, state-of-the-art methods. Additionally, we have discussed concepts of deep learning based attention mechanisms that can be exploited to improve the performance of neural networks.

In the remainder of this thesis, we discuss background information and existing works associated with problems investigated in this thesis, in Chapter 3. Following, we explore the problem of detecting and segmenting 3D objects in multi-modal data that observe sparse 2D layers of a 3D scene. We focus on incorporating inter-band and cross-band information in a joint analysis based approach to perform the detection and segmentation of solar active regions in Chapters 4 and 5, respectively. We then investigate the problem of detecting pulmonary nodules in 3D medical images in Chapters 6 and 7, where we explore the possibility of generalising the cross-band joint analysis approach and reformulate the problem into an attention problem and explore the possibility of explicitly modelling inter-channel (spatial), cross-channel, and global correlations in a learnable manner.

Chapter 3

Application Related Background

Contents

3.1 Introduction

In Chapter 2, we presented background materials and knowledge associated with deep learning and object detection and segmentation methods that will be used over the course of this thesis.

In this chapter, we provide an overview of existing methods that are relevant to the work conducted in this thesis. Particularly, we focus on the detection and segmentation of solar active region from multi-spectral images in Section 3.2, as well as the detection of pulmonary nodule form 3D computed tomography images in in Section 3.3.

In each section, we discuss and compare state-of-the-art methods proposed for each task. We also identify limitations and possible improvements on these methods. Finally, we conclude and summarise this chapter in Section 3.4.

3.2 Solar Active Regions Localisation

Solar active regions (ARs) are 3D objects that span the solar atmosphere and are characterised by their highly dynamic and strong magnetic fields. Active regions generate different forms of solar activity such as coronal ejections and flares. The development of other solar features such as sunspots, coronal loops, and prominences is associated with the development of these active regions. The precise localisation of solar ARs is therefore crucial to study and understand solar activities and phenomenons. Such analysis is enabled by remotely observing the solar atmosphere using multi-spectral ground- and space-based sensors.

The solar atmosphere consists of different layers spanning from the solar surface towards the chromosphere, transition region, and the solar corona. It consists of different elements, each of these elements emits light of particular wavelengths at certain a temperature within the temperature gradient that spans the solar atmosphere. Therefore, when observing the solar disk, different wavelengths (bands) reveal different, and sparse, 2D layers (distinct altitudes that correspond to particular temperatures) of the 3D solar atmosphere. See Table 3.1, and Figs. 3.1 and 3.2. This is different from common multi-spectral imaging scenarios, such as Earth sensing from space or RGB-D imaging, where different bands show different compositions of the observed scene. For convenience, we refer to this special multi-spectral scenario as *multi-layer*.

Most AR localisation existing methods are based on unsupervised learning and morphological analysis. [121] proposed a method for single-band images from PM/SH (Paris Meudon's Spectroheliograph) and SOHO's (Solar and Heliospheric Observatory) Extreme ultraviolet

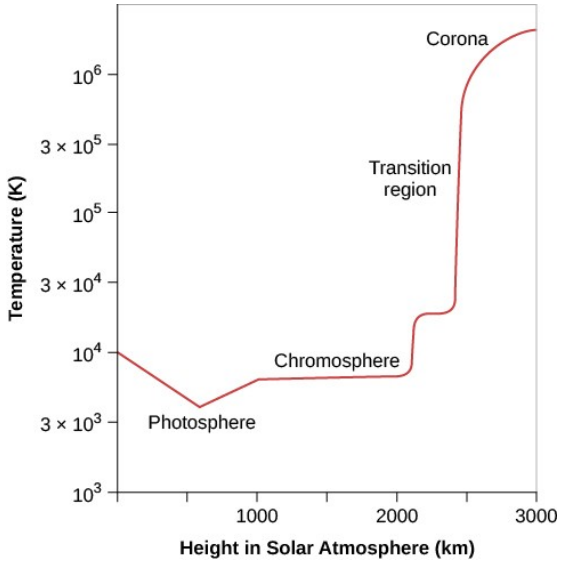

Figure 3.1: The temperature gradient as observed over different heights in the solar atmosphere. Adopted from [119].

Wavelength	Ion	Peak temperature	Observed layer
304 Å	He II	8.0×10^4 K	Chromosphere
171 Å	Fe IX-X	1.3×10^6 K	Transition region
195 Å	Fe XII	1.6×10^6 K	Lower Corona
284 Å	Fe XV	2.0×10^6 K	Higher Corona

Table 3.1: SOHO EIT solar imaging bands and their correspondence to distinct temperatures in the solar atmosphere. Using different imaging bands, SOHO EIT is able to isolate emissions from narrow temperature ranges and correspondingly observe distinct, and sparse, solar layers (altitudes). Adopted from [120].

Imaging Telescope (EIT), based on local thresholding and morphological operations followed by region growing. The method was evaluated against manual detections (synoptic maps) produced at PM/SH and National Oceanic and Atmospheric Administration (NOAA), and detected similar numbers of solar ARs as PM, and about $\sim 50\%$ over detection than NOAA.

In [122], AR segmentation was performed by computing the pixel-wise fractal dimension (a measure of non-linear growth that reflects the degree of irregularity over multiple scales) in a convolutional fashion, and feeding the resulting feature map to a Fuzzy C-means [123] process to produce the final segmentation. This method processes a single band at a time, and was subjectively evaluated on the SOHO/EIT 171 Å, 304 Å, and 284 Å bands. The use of individual bands was justified by the fact that each imaging band provides information from a

different solar altitude, authors showed how solar ARs span different areas in different imaging bands. This however, neglects the inter-dependencies between the bands, which can be exploited for increased performance. We address this in Chapters 4 and 5, where we design a multi-tasking framework to predict AR locations individually in the different bands while simultaneously incorporation (correspondence) between the different solar layers by dynamically aggregating cross-band information on multiple semantic levels.

Additionally, [124] proposed SMART for AR extraction from Solar and Heliospheric Observatory (SOHO) Michelson Doppler Imager (MDI) magnetograms. Their approach is based on thresholding two consecutive images to identify candidate AR areas. This is followed by discarding candidates that are not present in both images. The resulting mask is then dilated to include decaying areas around the detected ARs. Performance was evaluated by comparing the number of detections to those in NOAA over a solar cycle (1997-2008), where SMART shows a lower number 72% of the time. The authors believe that this is due to the fact that SMART tends to merge nearby spots in one detection. SMART was used to extract solar ARs for the Heliophysics Integrated Observatory (HELIO)[1].

Another method known as Spatial Possibilistic Clustering Algorithm (SPOCA) [125] used clustering to extract (pixel-wise) solar ARs and coronal holes from SOHO/EIT 171 Å and 195 Å combined images. SPOCA's segmentation is based on Fuzzy C-means and Possibilistic C-means [126], followed by post-processing with morphological operations. Both SPOCA and [122] suggest that using fuzzy logic based approaches can assist in overcoming the uncertainty in defining AR boundaries. The quality of results was subjectively evaluated on 112 observations. SPOCA is now used in the Heliophysics Feature Catalogue (HFC) online catalogue.

A comparison between SMART and SPOCA, was presented in [127], where both areas and number of detections were investigated over a $\sim$6 weeks period, in which solar activity level was high (May–June 2003), and compared to those in NOAA. Results show that both SMART and SPOCA detect more regions than NOAA, with SMART being the highest. These results, although apparently in contradiction with those of [124], may indicate that SMART is good at detecting solar ARs in periods of high activity, but under-performs at medium and low activity.

Both SMART and SPOCA detect comparable numbers of active regions [127]. Authors explain that the differences noted in their study may be caused by the different AR definition used in each method. Moreover, SMART and SPOCA use different types of solar imagery (photospheric, and coronal, respectively). This may indicate that detecting solar ARs using

[1] `http://hfe.helio-vo.eu/Helio/`

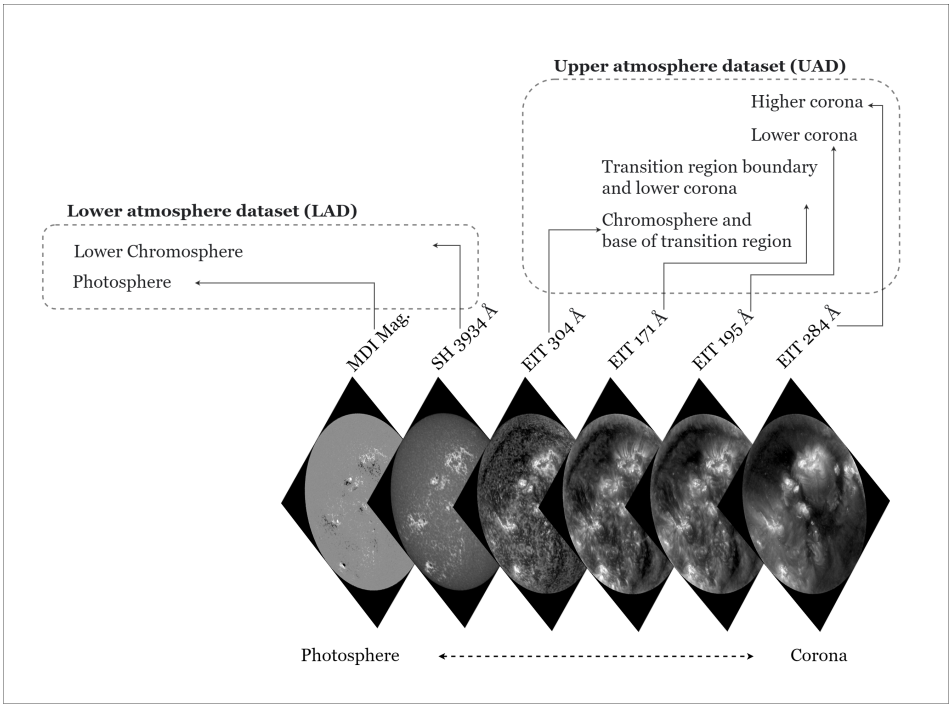

Figure 3.2: Illustration of the solar disk as it appears in SOHO/MDI (Magnetograms) images, PM/SH (3934 Å), and SOHO/EIT (304 Å, 171 Å, 195 Å, and 284 Å) multi-spectral imagery, where each image band corresponds to a certain altitude in the solar atmosphere.

single band images may be an under-constrained problem.

To the best of our knowledge, DL methods has not yet been explored for the problem of AR localisation at the time conducting this work. Indeed, the lack of labelled AR datasests limits the opportunity of involving DL methods. Most cited works are therefore based on unsupervised learning approaches, making them difficult generalize to different imaging sensors (e.g., ground- and space-based) and bands or adapt to cyclic variations and activity levels of the solar atmosphere. In Chapters 4 and 5, we address these limitations by investigating DL solutions to handle the problem of localising solar ARs, we also address the data limitations associated with both, the solar AR detection and the segmentation tasks.

Moreover, most cited works are designed to process individual image bands at a time (e.g., [121, 122, 124]) and therefore fail to exploit the multi-spectral aspect of the data. Other methods that incorporate information from multiple imaging bands (e.g., [125]) neglect the multi-layer aspect of the data by producing a single set of 2D predictions for all layers (i.e., image bands) of the 3D solar ARs. Over the course of Chapters 4 and 5, we investigate the possibility of exploiting inter-band and cross-channel (i.e., cross-band) correspondence and correlations in a

multi-tasking joint analysis based approach in which localisation is carried out in each imaging band (i.e., solar layer) individually, but based on multi-spectral information.

3.3 Pulmonary Nodule Localisation

The early detection of pulmonary cancer plays a crucial role in the treatment. Manual screening of 3D CT scans is time consuming and may be impacted by multiple factors, including the experience and the well-being of radiologists. The increasing amounts of pulmonary data collections along with the advances in convolutional neural networks have attracted research interest towards automating the pulmonary nodule localisation and classification task. Generally, most existing works deploy two stage CNN based detectors and investigate different levels of feature dimensionality, e.g, 2D, pseudo 3D (cross-sectional 2D planes), and 3D [3, 128]. Berens et al. [129] proposed a two stage detector based on a 2D U-Net [28] for region proposal, followed by a false alarm reduction CNN that takes three orthogonal 2D slices as an input. They evaluate their approach on the pulmonary nodule detection dataset Lung Nodule Analysis (LUNA16) [3], and conclude that directly incorporating 3D information may be a good direction to improve on the performance of their approach. Indeed, this was demonstrated in [130], where the authors compare 2D, pseudo 3D, and 3D CNNs for the lung nodule detection task and find that using 3D kernels significantly improves the performance. Similarly, Riquelme et al. demonstrate in their extended survey on DL for lung nodule detection [128], that the best performing methods are the ones that incorporate 3D information in their approaches. This shows that 2D based approaches fail to fully exploit the inherent 3D nature of the nodule structure.

Liao et al. proposed a 3D region proposal network based on Faster RCNN [52] for the candidate proposal stage. The top five suspicious proposals are passed into a subsequent CNN in which a modified *or-gate* [131] is employed to predict the final score [132]. Zhu et al. proposed a similar 3D CNN based on dual path networks (DPN) [133] for nodule detection along with a gradient boosting machine (GBM) [134] for the final classification stage [135]. They deploy an encoder-decoder design for their detection network to allow learning nodule features from different semantic levels. They show that by using grouped convolutions [136] and dense convolutional connections [137] in their DPN, the network was able to detect more nodules while decreasing the computational overhead.

Building on [132] and [135], Li et al. adopted squeeze and excitation paths (i.e., [38]) to their detection CNN to assist the network learn inter-dependencies within the extracted features [138].

Other works also exploited an encoder-decoder design, e.g., [139–141], and have demonstrated great potential within the lung nodule detection task. Additionally, both [140] and [138] adopt focal loss [142] to tackle the class imbalance in the pulmonary images and have empirically demonstrated an enhanced performance in the nodule detection task.

Generally, most existing works explore different feature dimensionality, semantic level, and spatial scale, as well as different objective functions to perform the nodule detection task. Limited research has been dedicated towards incorporating feature importance and inter-spatial and cross-channel correlations. In Chapter 6, building on state-of-the-art methods, we investigate DL techniques study the possibility of explicitly modelling and inferring feature importance and spatial and cross-channel correlation on a deep feature level. We also study the possibility of incorporating information from different feature dimensionality (2D, pseudo 3D, and 3D) simultaneously, in an end-to-end, effective manner.

For the false positive reduction stage, [143] proposed using an ensemble of five CNNs to analyse nodules using different levels of context. Each CNN was trained using a unique crop size (i.e., level of context) and the final score was defined as a function of all the predictions. In the same line, [105] proposed a more efficient approach by jointly analysing nodules on different contextual levels using a single CNN. Both experiments demonstrate that incorporating contextual information leads to an enhanced prediction. This is expected since pulmonary nodules have highly variable morphology and size (more details in Chapter 6). Thus, in Chapter 6, we design a false positive reduction network in which we employ a joint analysis approach that aggregates spatial features from multiple contextual levels, improving the overall detection performance.

3.4 Summary

In this chapter, we have discussed background information and existing works associated with problems investigated in this thesis. We identified gaps and challenges and discussed potential improvements associated with existing solutions.

More specifically, we first provided an overview on the detection and segmentation of solar active region from multi-spectral and multi-layer images. Generally, most existing works are based on unsupervised learning and rely on fixed pre- and post-processing steps making them difficult to generalise to new solar imaging domains and solar cyclic changes. The 3D (multi-layer) aspect of multi-spectral solar data has been commonly neglected in available solutions.

We then provided an overview on the problem of pulmonary nodule detection from volumet-

ric CT scans and discussed related state-of-the-art deep learning based solutions. Most existing works focused on investigating different levels of feature dimensionality (2D, pseudo 3D, and 3D), and semantic and spatial levels of extracted embeddings. However, limited research has been dedicated towards exploring feature importance, spatial and cross-channel, and global correlations, which can be exploited for more accurate detection.

In the rest of this thesis, we explore a deep learning based joint analysis strategy in which inter-band and cross-band (cross-channel) correspondence and correlation is incorporated to perform 3D (i.e., multi-layer) detection and segmentation of solar active regions in Chapter 4 and Chapter 5 respectively. In Chapter 6 and 7, we further investigate the possibility of explicitly modelling inter- and cross-channel correlations, as well as long range correlations (global context) to infer feature importance, and consequently more effective learning, to solve the problem of pulmonary nodule detection.

Chapter 4

MLMT-CNN for Object Detection in Multi-layer and Multi-spectral Images

Contents

4.1 Introduction

Solar active regions (ARs) are areas in the solar atmosphere that observe strong and dynamic magnetic fields, the detection these regions is essential in studying solar weather and behaviours. Such analysis is possible by remotely sensing the solar atmosphere using multiple wavelengths captured from different ground- and space-based sensors (e.g., see Figs. 4.1 and 4.2). However, unlike traditional multi-spectral scenarios such as Earth imaging from space, e.g., [41,42,71,79–82], where multiple imaging bands reveal different aspects or compositions of the same scene, in solar physics, different bands capture the solar atmosphere at different temperatures that are associated with different altitudes across the solar atmosphere (i.e., distinct 2D layers of the 3D solar atmosphere) [122]. Thus, we refer to this special multi-spectral scenario as *multi-layer*.

Very few solutions were presented to the AR localisation problem. Most of these methods exploited single image bands only, e.g., [121, 122]. Authors justified this by the fact that each band provides information from a different solar altitude, they show how areas of solar ARs differ from band to band [122]. We, however, argue that inter-dependencies exist between bands, which can be exploited for enhanced performance.

The SPOCA method [125] used Fuzzy C-means and Possibilistic C-means [126] clustering to segment solar ARs and coronal holes from the SOHO/EIT (solar and heliospheric observatory's extreme ultraviolet imaging telescope) 171 Å and 195 Å combined images, assuming that they should yield identical detection. This approximation may result in a poor analysis of at least one of these bands. The use of fuzzy logic in SPOCA aims to address the uncertainty in defining AR boundaries [125]. Generally, these methods are mainly based on clustering and morphological operations, thus are pre- and post-processing dependant, which makes them difficult to adapt to new image domains and hyperparameter-dependant.

Deep learning (DL) has dramatically improved object detection in the last few years. Generally, most existing methods target 2D images or 3D volumes (e.g., [25,26,52–54]), however, the sparse 3D nature of the multi-spectral solar data requires designing a specialized DL framework. The term *sparse* here indicates the nature of the multi-spectral images, where each wavelength is centered at elements that emit light of particular wavelengths at a particular temperature within the temperature gradient spanning the solar atmosphere, and therefore observing distinct physical locations (i.e., different 2D cuts that lie at different altitudes in the solar atmosphere). See Table 3.1, and Figs. 3.1 and 3.2. This is different from common multi-spectral images that show different compositions of the same physical location.

DL based methods that target multi-spectral data commonly treat multi-spectral images in a similar fashion to RGB images, by stacking different bands into multi-channel images [40, 79–81, 144]. These methods are designed under the assumption that the different image bands capture different aspects of the same scene, which makes it ill-suited for our multi-layer case, where spatial positioning indeed differs from band to band. Another common approach is to aggregate information from different bands at different levels (e.g., feature level and image level) [40–42, 82, 145–147]. This feature fusion strategy demonstrates potential for DNNs to improve localisation by exploiting the multi-spectral aspect of the data. Some works found that feature level fusion assists CNNs in producing a more consistent detection than using image level fusion for pedestrian detection from RGB and thermal images [41]. Contrary, image fusion worked best when segmenting soft tissue sarcomas in multi-modal medical images [40]. This suggests that there is no universal best fusion strategy. Thus, we investigate different types of fusion and different stages to apply fusion. Another feature fusion strategy was used to segment coronal holes from 7 EUV bands and line-of-sight magnetogram in [43]. The method relies on training a CNN, using weak labels, to segment coronal holes from a single band, followed by fine-tuning the learned CNN over the other bands consecutively. Finally, embeddings of the band-specialised models are aggregated and are passed into a final segmentation CNN, resulting in a unique final prediction. This unique localisation result for all multi-spectral images is a common limitation to all cited works for our multi-layer scenario, which we address in this study with a multi-task network.

In this chapter, we introduce a novel MultiLayer MultiTask CNN (MLMT-CNN), a multi-tasking DNN framework, as a solution for the solar AR detection problem by taking into consideration the multi-layer aspect of the data and the 3-dimensional spatial dependencies between image bands. Our approach exploits both band specific and cross-band information by extracting and aggregating features on different semantic levels, allowing the network to focus on effective inter-band details while simultaneously learning correspondence and inter-dependencies between the different bands. The proposed detection paradigm performs the detection of solar ARs, simultaneously, in the different bands with respect to their correspondent band-specific ground-truth, and is therefore multi-tasking.

The 3D nature of our multi-spectral and multi-layer imaging scenario, which differs from other multi-spectral cases such as Earth observations, requires a new benchmark. Therefore, we introduce two annotated datasets comprised of images of the solar atmosphere from both ground and space-based sensors. They cover evenly all phases of solar activity, which follows

an 11-year cycle. To the best of our knowledge, no localisation ground-truth is readily available for such data. We therefore design a labelling tool that takes into account multi-spectral and temporal information to assist the manual labelling process of the solar images.

Furthermore, we propose a training approach that accounts to the different objectives of the individual MLMT components using their correspondent losses, in contrast to the classical training in which all components are deemed to reach an optimal solution simultaneously according to their overall loss.

Our contributions may be summarised as:

1. We present a paradigm to handle multi-spectral solar images that show several layers of a 3D object that span the solar atmosphere (i.e., multi-layer). We demonstrate the effectiveness of our approach in MLMT, a multi-task DL framework for solar AR detection based on feature aggregation and joint analysis of multi-spectral and multi-layer information. We demonstrate the potential of our proposed paradigm by implementing it with different state-of-the-art CNN backbones, as well as handling different data types and arbitrary number of bands.

2. We propose a training strategy for MLMT that optimises the DNN weights more effectively for each objective than the classical training strategy.

3. We introduce two balanced and annotated datasets of multi-layer images of the solar atmosphere for AR detection, from both ground- and space-based data.

4. We design and release a multi-spectral and multi-layer image annotation tool that facilitates bounding box labelling using temporal and spectral information.

5. We further validate our approach on an artificially created dataset of multi-modal medical images of similar spatial configurations to the multi-layer solar images.

In the reminder of this chapter, we present the details of the proposed method in Section 4.2. In Section 4.3, we present extended experiments along with the details of the solar data and our annotation tool. Finally, Section 4.4 summarises this chapter and provides key findings and conclusions.

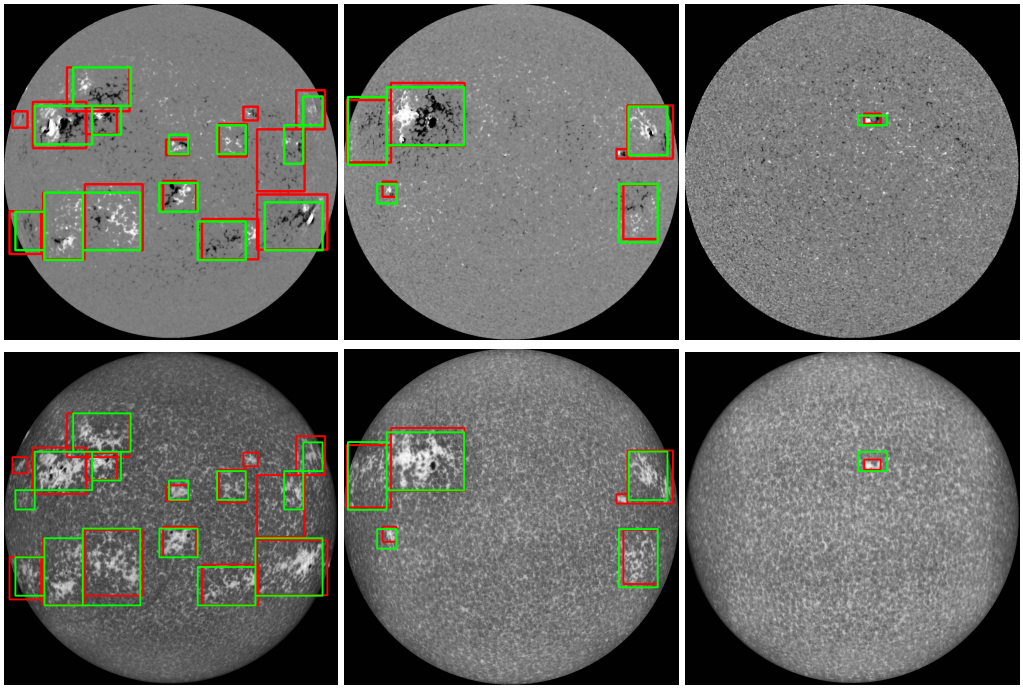

Figure 4.1: Ground-truth (red) and MLMT-CNN's (green) detection of solar ARs at three levels of solar activity (left to right: high, medium, low) in randomly selected images from (top to bottom) SOHO/MDI Magnetogram and PM/SH 3934 Å.

4.2 Proposed Method

Our framework exploits several time-matched multi-layer images (see Section 4.3.1.1) in parallel, to predict separate, although related, localisation results for each image band. Our localisation involves two stages: detection, in the form of bounding box around an object and its classification of object type, followed by a segmentation stage (Chapter 5) to produce a pixel-wise classification map enclosed in the predicted bounding box.

For both stages, we deploy a multi-layer and multi-task DL framework that analyses information from neighbouring layers (image bands). The network learns band-specific features that are then fused at multiple levels in the network, inducing the network to learn effective inter-band features while also capturing cross-band correspondence and inter-dependencies. Finally, the resulting embeddings from neighbouring layers are jointly analysed to produce their separate but related results. Note that the term multi-tasking here refers to the multi-layer nature of the localisation problem, in which band-specific (i.e., layer-specific) detections are predicted with respect to band-specific ground-truth, simultaneously for all input image bands.

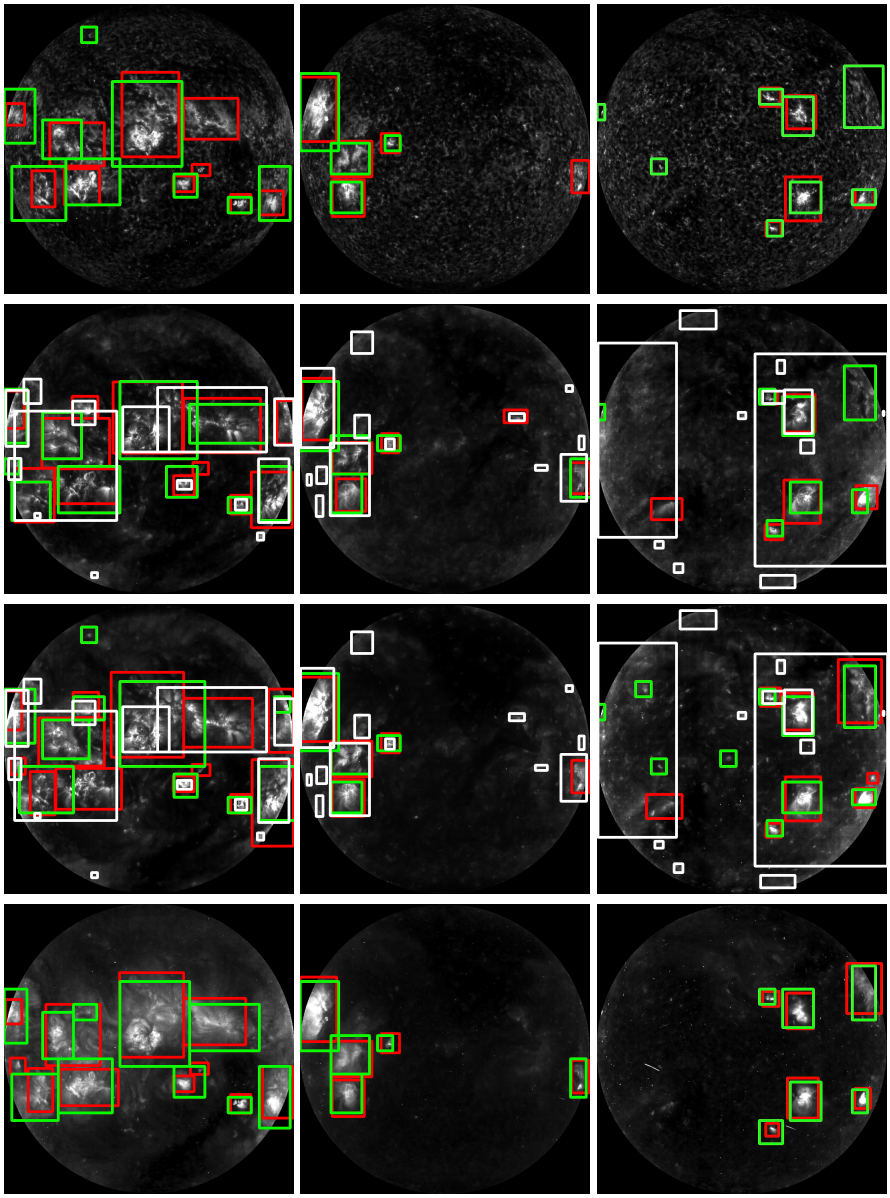

Figure 4.2: Ground-truth (red) and MLMT-CNN (green) and SPOCA's (white) detection of solar ARs at three levels of solar activity (left to right: high, medium, low) in randomly selected images from (top to bottom) SOHO/EIT 304 Å, 171 Å, 195 Å, and 284 Å.

In this section, we introduce the main concepts of the MLMT-CNN framework in Section 4.2.1, the backbone network used in our framework in Section 4.2.2, and the details of our detection approach in Section 4.2.3.

4.2.1 MultiLayer-MultiTask (MLMT) Framework

While some existing works were developed for analysing multi-spectral images, to our best knowledge, the problem of detecting objects over multi-layer imagery, which is a sparse 3D multi-spectral case in which different bands show different scenes (i.e., layers), was not yet addressed. The proposed multi-layer and multi-task (MLMT) framework addresses this scenario by incorporating three key principles:

1. Extracting features from different image bands individually using parallel feature extraction branches. This allows the network to learn independent features from each band according to their band-specific objective.

2. Aggregating the learned features from the different branches using feature fusion operations. In this work, we test fusion by addition and concatenation at different feature levels (i.e., early and late feature fusion). This assists the network to jointly analyse the extracted features from different bands and thus learn inter-dependencies between the image bands.

3. Generating a set of results per image band, based on a multi-task loss, allowing the detection of different sections or layers of 3D objects in the different bands in a multi-tasking manner.

Points 1 and 3 are motivated by the nature of the multi-layer data, where different bands capture different locations in a 3D scene, each providing some unique information. Our framework aims at obtaining specialised results for each image band, in contrast to most existing works where focus is on producing a single set of predictions to all image bands. This is crucial since the localisation information may differ from one band to another in cases of multi-layer images (e.g., solar images). Yet, all bands are spatially correlated, which motivates point 2. Our framework exploits the inter-dependencies between the different bands by its joint analysis strategy, enhancing its performance in individual bands. Furthermore, our framework emulates how experts manually detect solar ARs, where a suspected region's correlation with other bands is evaluated prior to its final classification. This demonstrates the usefulness and importance of accounting for (spatially and temporally) neighbouring slices in detecting ARs.

Our framework design is flexible and can accommodate any number of available image bands (layers) and perform different tasks (e.g., detection and segmentation). Additionally, as suggested by existing works, since different scenarios may require different fusion strategies, the modularity of our framework allows it to be easily adapted to different cases. We demonstrate

this by applying our framework to different applications in Section 4.3 (e.g., solar ARs, BraTS-prime datasets) , where we examine different types and levels of feature fusion such as early and late addition and concatenation.

4.2.2 Backbone Networks

The 3 key principles are applicable to different backbones since they are not architecture dependent. We demonstrate this by applying these principles into different backbone networks and tasks, particularly, AR detection as we show in this chapter, and segmentation in Chapter 5.

We adopt the Faster RCNN architecture as the backbone of our detection framework. Faster RCNN is a DL-based detector that may be trained to detect and classify a number of objects from a –usually RGB– image. It consists of three main parts: 1) convolutional layers to extract features from the input image. From these features, 2) a region proposal network (RPN) proposes suspicious locations that might contain an object, and 3) a detection network predicts the object class of each proposed locations. We apply our framework to the three stages detection strategy of Faster RCNN generalising it to jointly analyse multiple images that span different locations or layers of a 3D scene.

Comparing to other state-of-the-art architectures (e.g., YOLO and SSD), the multi-stage design of Faster RCNN allows aggregating information from different bands at different levels, namely low level (i.e., feature extraction stage) and high level information (i.e., region proposals). Additionally, Faster RCNN has scored the highest accuracy in [60].

4.2.3 MLMT-CNN: Detection Stage

Our detection DNN is presented in Fig. 4.3. It takes the pre-processed multi-layer image as input. A CNN (ResNet50 or VGG16 in our experiments) is first used as a feature extraction network. Parallel branches (subnetworks) produce a feature map per image band, following the late (or feature map) fusion strategy. Since individual bands provide different information, this allows the subnetworks' filters to be optimised for their input bands individually, and therefore focus on learning effective inter-band embeddings.

The feature maps from the different bands are then aggregated by a concatenation operation, aiming the assist the network in learning cross-band inter-dependencies. The combined feature map is jointly analysed by one parallel network per image band that performs region proposal (RPN). The RPN stage uses three aspect ratios ([1:1], [1:2], [2:1]) and four sizes of anchor (32,

64, 128, and 256 pixel width). We found empirically that these match well the typical size and shape of solar ARs. One specialised RPN per image band is trained.

During training time, for each band, the correspondent region proposals along with the combined feature map are passed into a detection network to perform the final prediction for the band. However, at testing time, the band-specialised detector modules use the region proposals from all bands. This combination of region proposals helps finding potential AR locations (region proposals) in solar layers (imaging bands) where they are more difficult to identify using clues from neighbouring layers. This also aids the network in learning the inter-dependencies between the different bands more dynamically, benefiting from information from different bands simultaneously while having band-specialised region proposal and detection models.

It is worth noting that during training, the RPN proposals for a band are filtered (i.e., labelled as positive or negative) with respect to their overlap with the band's own ground-truth. Hence, combining them in the training time would mean implicitly inheriting the ground-truth of a band to another, in contradiction with the band-specific ground-truth used for training the detector module. Indeed, different bands show distinct cuts of a 3D object in which each cut must have its own ground-truth. Combining ground-truths of different bands at training time may hinder the learning of both the RPN and detector modules. Therefore, region proposals are only combined at testing time to ensure a better learning of the final detection modules.

Using the combined feature map aids the network to learn the relationship between the image bands, in both region proposal and classification stages, hence providing an enhanced performance in line with the nature of the data. This prediction is still band-specialised thanks to the different ground-truths being used for each band at training time. The essence of our approach is in line with concepts of attention, where the network is provoked to focus on each of the input bands individually using the parallel band-specific CNN branches and prediction heads, while simultaneously incorporating cross-band inter-dependencies by sharing and aggregating information from the different bands at different semantic levels. We demonstrate in Section 4.3 that this is particularly helpful in cases where an AR is difficult to detect in a single band.

We train our MLMT framework using all input bands and branches according to a combined loss function:

$$L = \sum_{b} \left(\frac{1}{N_{cls}} \sum_{i} L_{cls}(p_{b_i}, p^*_{b_i}) + \lambda \frac{1}{N_{reg}} \sum_{i} p^*_{b_i} L_{reg}(t_{b_i}, t^*_{b_i}) \right) \tag{4.1}$$

where b and i refer to the image band and the index of the bounding box being processed, respectively. p and $p^\star$ are the predicted anchor's class probability and its actual label, respectively. Lastly, t and $t^\star$ represent the predicted bounding box coordinates and the ground-truth coordi-

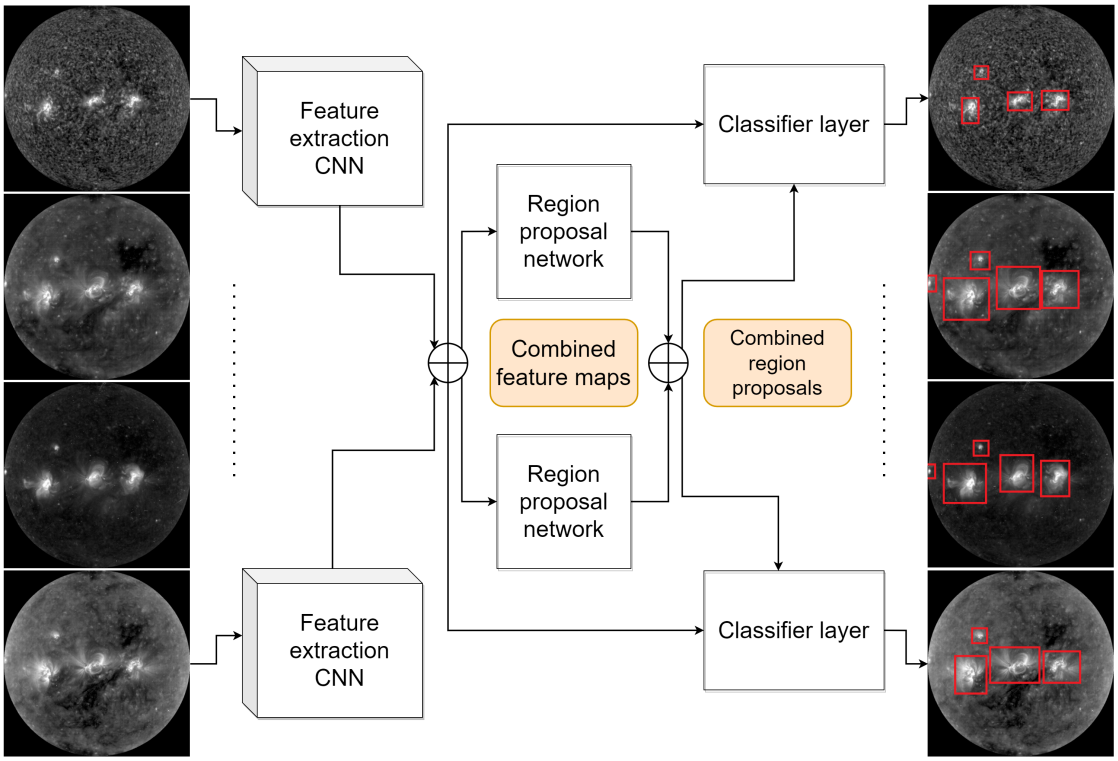

Figure 4.3: MLMT for detection using the Faster-RCNN backbone following the late feature fusion approach. 'Plus' sign denotes concatenation of the feature maps, or of the lists of region proposals (at testing time). Each image band is analysed independently using a band-specific convolutional branch to extract band-specific features. These are then fused and are jointly analysed by band-specific RPN modules such that each RPN produces region proposal for its correspondent band. The fusion process assists the network in learning inter-dependencies between the different bands. Region proposals from different RPNs are then aggregated and passed onto band-specific detection heads, where each band gets a separate (but related) set of predictions.

nates, respectively. The terms L_{cls} and L_{reg} are the bounding-box classification loss and the bounding-box regression loss, respectively, as defined in [52]. More specifically, the term L_{cls} indicates the log loss:

$$L_{cls}(p, p^*) = -\log(p_t) \tag{4.2}$$

where $p_t = p$ if the class label is 1, and $p_t = 1 - p$ otherwise. Moreover, the term L_{reg} represents the smooth_{L_1} defined over the parameterised target and the predicted bounding boxes, $t^* = (t_x^*, t_y^*, t_w^*, t_h^*)$ and $t = (t_x, t_y, t_w, t_h)$, respectively. Accordingly, the regression loss is com-

puted as follows:

$$L_{reg}(t,t^*) = \sum_{i \in x,y,w,h} smooth_{L_1}(t_i - t_i^*)$$

where:

$$\text{smooth}_{L_1}(d) = \begin{cases} 0.5x^2 & if\ |d| < 1 \\ |d| - 0.5 & otherwise \end{cases} \tag{4.3}$$

The values of the parameterised predicted and target bounding box coordinates (i.e., t and t^*, respectively) are computed with respect to the anchor boxes as follows:

$$\begin{aligned} t_x = (x - x_a)/w_a, \quad t_y = (y - y_a)/h_a, \\ t_w = \log(w/w_a), \quad t_h = \log(h/h_a), \\ t_x^* = (x^* - x_a)/w_a, \quad t_y^* = (y^* - y_a)/h_a, \\ t_w^* = \log(w^*/w_a), \quad t_h^* = \log(h^*/h_a) \end{aligned} \tag{4.4}$$

where $x, y, w,$ and h represent the center coordinates, width, and height of a predicted bounding box, respectively. Similarly x^*, y^*, w^* and h^* and $x_a, y_a, w_a,$ and h_a represent the coordinates of target bounding boxes and the anchors, respectively. The parameter λ balances the classification and the regression losses (we set λ to 10 as suggested in [52]). N_{cls} and N_{reg} represent the size of the mini batch being processed and the number of anchors, respectively. It is worth noting that our proposed framework is not limited to using Faster RCNN's loss and may be trained with using other task-suitable loss functions.

During training, the weights of each stage (i.e., feature extraction, region proposal, and detection) are stored independently whenever the correspondent loss decreases. At testing time, the best performing set of weights is retrieved per stage. We refer to this practice as 'Multi-Objective Optimisation' (MOO). The improved performance that we observe in Section 4.3 may be explained by each stage having a different objective to optimise, which may be reached at different times.

In this study, we experiment with a 2, 3, and 4-band pipeline. However, the approach may generalise straightforwardly to n bands and new imaging modalities.

4.3 Experiments

All experiments were implemented using Tensorflow with an NVIDIA GeForce GTX 1080 Ti GPU. Our detection framework was trained for 3000 epochs ($\sim$4 days), respectively, using Adam

optimiser [48] with a starting learning rate of 2e-5.

Generally, the performance of detection methods may be evaluated in different ways, e.g., precision, recall, F1-score, mean average precision, and accuracy. We evaluate our detection stage using both precision and recall since they provide an insight on, respectively, the specificity and the sensitivity of the tested model. Additionally, we use F1-score to study the harmonic mean of both specificity and sensitivity. Additionally, we compute the average precision score to evaluate the performance of the proposed approach in contrast to single band based detectors.

Solar ARs are dynamic structures that are constantly changing (e.g., merging and splitting, from and to multiple regions) during their lifetime [125, 148–150]. Using common similarity criterion to determine the quality of detections (i.e., categorising predictions into true positive, false positive, and false negative detections) such as Intersection over Union (IoU) fails to capture two prevalent scenarios when handling solar active region structures: 1) when a solar AR area is detected by multiple neighbouring bounding boxes, and 2) when a cluster of neighboring solar ARs is detected in a single bounding box.

In the first scenario (considering an IoU based criteria with a threshold of 0.5), a small predicted box that observes an $\frac{\text{intersection}}{\text{predicted area}} >= 0.5$ of a larger ground-truth box, but has an IoU < 0.5, is deemed false positives when evaluated against the IoU threshold. However, in effect, the predicted area represents $\frac{\text{intersection}}{\text{predicted area}} > 0.5$ of an active region. Using IoU in such cases does not take such scenarios into consideration. A similar phenomena is observed when a big detection encloses smaller ground-truth boxes of which $\frac{\text{intersection}}{\text{ground-truth area}} >= 0.5$, but has an IoU $<$ 0.5 (i.e., scenario 2). Such detection is also deemed false positive (and the ground-truth boxes are subsequently deemed false negative). However, effectively, the enclosed area within the predicted box represents an active region. See figure 4.4.

Therefore, to address the two aforementioned prevalent scenarios, we design a dynamic criterion that accounts for the variable characteristics associated with solar ARs when evaluating the performance of the proposed approach. Accordingly, a detection is considered a true positive if its intersection with a ground-truth box is greater or equal to 50% of either the predicted or ground-truth area, and is of an area that lies within the AR area distribution of the annotated dataset, otherwise, a detection is deemed false positive. Ground-truth bounding boxes that fail to associate with any prediction according to the proposed true positive criteria are subsequently deemed false negatives. We empirically found that this provides a good trade-off of precision over recall within our target application.

Additionally, to further demonstrate the aforementioned scenarios associated with the IoU

based evaluation criterion, we provide results of the proposed detection method when using an IoU based criterion to indicate the quality of detections using IoU thresholds of 0.25 and 0.5. In this case, any detection associated with an IoU greater than or equal to the threshold value is deemed true positives, otherwise it is considered a false positive. It is worth noting that a threshold value of 0.5, is commonly –but not exclusively– used to evaluate object detection methods [151, 152]. The use of a lower threshold (i.e., 0.25 IoU) here is merely to showcase the aforementioned predominant scenarios of which IoU criteria fails to account for, even at lower a threshold.

It is worth noting that both our annotation and evaluation processes were validated by a solar physics expert. Non Maximum Suppression (NMS) is used to discard any redundant detections.

4.3.1 Data

4.3.1.1 Labelled AR Datasets

Solar data has been continuously acquired using space- and ground-based sensors over the past decades with increasing amount of data. Ground-based observations benefit from significantly easier setups, but tend to suffer from a lower quality due to their dependency on weather conditions and atmospheric obstacles. It is therefore highly desirable to consider both modalities when designing observation analysis methods. The Solar and Heliospheric Observatory (SOHO) spacecraft and Paris-Meudon (PM) observatory benefit from a large period of data, starting in 1996 and 1909 respectively, in comparison to more recent observatories such as the Solar Dynamics Observatory (SDO) which was launched in 2010. Hence, this study works on these images, available respectively from the SOHO archive[1] and the BASS2000 online portal[2].

Multi-layer solar images comprise of measurements at different ultraviolet and X-ray wavelengths (denoted as *bands*). Imaging telescopes divide the electromagnetic spectrum into bands of varying but narrow widths and centred on the emission wavelengths of ionised elements of interest. Since these ionised elements exist at given temperatures, they allow imaging different altitude regions of the solar atmosphere, following its temperature gradient.

Solar ARs are areas of strong magnetic field. Therefore, the multi-spectral images may be complemented by information about the intensity and polarity of the magnetic field, obtained from polarised light in the form of magnetograms. With current technologies, magnetograms are mainly available for the photosphere. We use those provided by SOHO/MDI.

[1] `https://sohowww.nascom.nasa.gov/data/`
[2] `http://bass2000.obspm.fr`

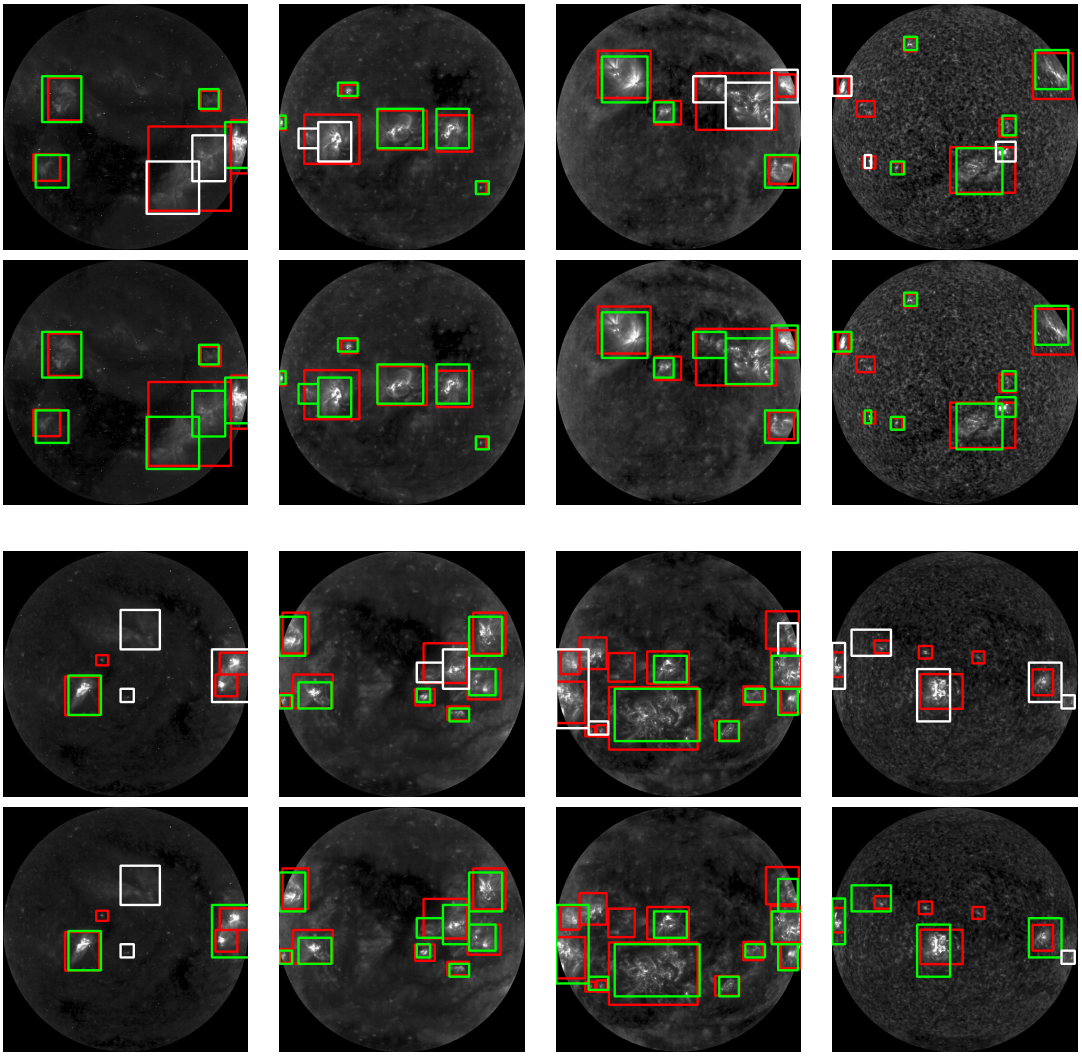

Figure 4.4: Detections of solar active regions visualized when an IoU based criterion is applied during evaluation (odd rows) against the proposed criterion (even rows). Ground-truth (red), and MLMT-CNN detections (true positive in green and false positive in white) of solar ARs in images from SOHO/EIT (top to bottom) 284 Å, 171 Å, 195 Å, and 304 Å. We observe that using IoU as an evaluation criterion causes some detected ARs to be regarded as false positives. Particularly, when an AR is detected by multiple boxes, or when multiple neighboring ARs are detected as a single structure. The proposed criterion on the other hand accounts for such cases.

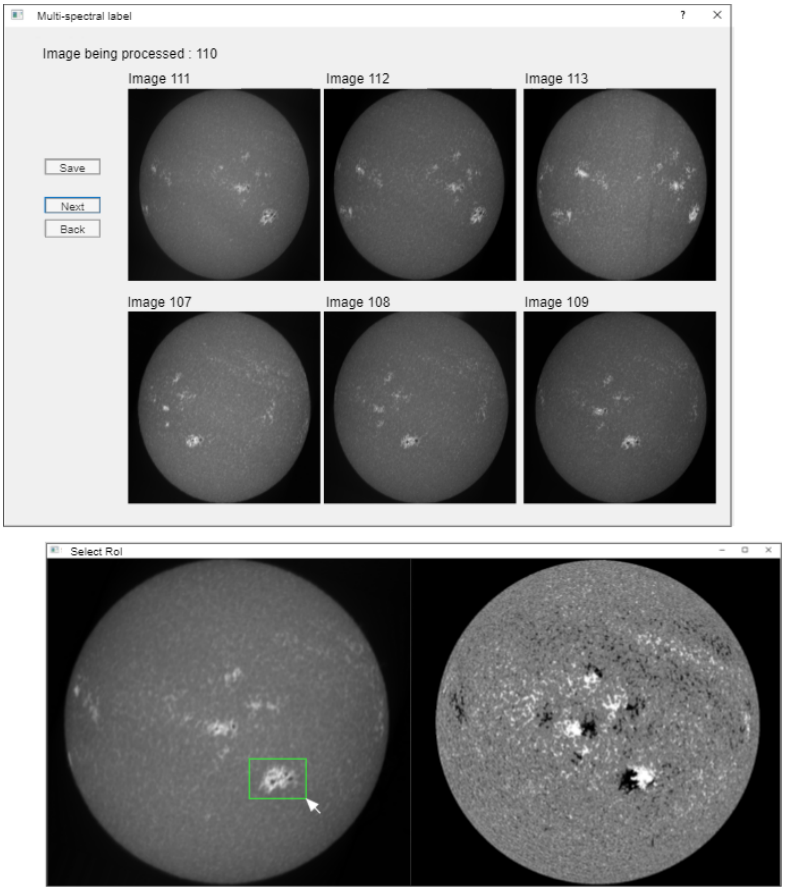

Figure 4.5: Our multi-spectral labelling tool used to annotate multi-spectral solar observations used in this study. In the top window, the first and the second rows, respectively, show 3 subsequent and 3 previous time steps in relation to the observation being annotated. The left side of the bottom window is the observation being annotated, while the right side shows an auxiliary image or spectrum.

The images of this study were acquired in the 171 Å, 195 Å, 284 Å, and 304 Å bands (SOHO/EIT imager, Fe-IX/X, Fe-XII, Fe-XV, and He II emissions lines respectively), 3934 Å band (Paris Meudon (PM) Spectroheliograph (SH) imager, Ca II K3 emission line), and the magnetogram images (SOHO/MDI imager, line-of-sight) as illustrated in Figs. 4.1 and 4.2. These correspond to observing the photosphere (magnetogram), chromosphere (3934 Å), chromosphere and base of the transition region (304 Å), transition region (171 Å and 195 Å), and corona (284 Å). Solar observations are acquired frequently to study the evolution of solar features and events over time. Table 4.1 summarises the observation frequencies and other data properties.

Our detection framework requires ground-truth annotations of solar ARs in the form of bounding boxes. Available solar datasets such as [153], contain pre-processed SDO imagery

Table 4.1: Technical summary of the two annotated datasets. Note that the values in brackets indicate the number of samples of which the SPOCA subset is formed.

Dataset	Modality	Image resolution	Obs. frequency	Activity level	# images	# BBoxes	Seg. masks availability
UAD	SOHO/EIT 284 Å	1024x1024	12 min	High	84 (5)	801 (41)	Yes
				Medium	93 (9)	610 (53)	
				Low	146 (12)	868 (72)	
				All	323	2279 (166)	
	SOHO/EIT 171 Å	1024x1024	12 min	High	84 (5)	873 (46)	
				Medium	93 (9)	635 (66)	
				Low	146 (12)	673 (56)	
				All	323	2181 (168)	
	SOHO/EIT 195 Å	1024x1024	12 min	High	84 (5)	875 (51)	
				Medium	93 (9)	678 (61)	
				Low	146 (12)	1118 (101)	
				All	323	2671 (213)	
	SOHO/EIT 304 Å	1024x1024	12 min	High	84 (5)	807 (46)	
				Medium	93 (9)	614 (56)	
				Low	146 (12)	1071 (79)	
				All	323	2492 (181)	
LAD	PM/SH 3934 Å	1500x1340	$\sim$ 1 day	High	47	618	No
				Medium	91	697	
				Low	128	471	
				All	266	1786	
	SOHO/MDI Magnetogram	1024x1024	96 min	High	47	618	
				Medium	91	697	
				Low	128	471	
				All	266	1786	

that is homogenised spatially and temporally, suitable for ML problems. However, these do not include any form of localisation ground truth. To the best of our knowledge, no such annotated dataset is currently publicly available. Therefore, we publish two new datasets with localisation annotations, which we refer to as the Lower Atmosphere Dataset (LAD), and upper atmosphere dataset (UAD). Upper Atmosphere Dataset (UAD), see Fig. 3.2. Both datasets include bounding box annotations produced using our multi-spectral labelling tool which displays, side by side, images from an auxiliary modality and from a sequence of 3 previous and 3 subsequent time steps (see Fig. 4.5). Additionally, the UAD dataset includes weak segmentation labels (details in Chapter 5). All annotations were validated by a solar physics expert.

A solar cycle lasts approx. 11 years, during which the magnetic flux frequency, i.e., rate of solar ARs appearing on the solar disc, varies. It may be broken down into three main periods of high, medium, and low activity level [154]. Accordingly, we select images evenly from each activity level, with years 2002-03 for high activity, 2004-05 for medium activity, and 2008-

10 for low activity. A random selection with a minimum 24 hrs (and average 21 days) gap avoided introducing any bias from consecutive observations. Table 4.1 presents an overview of annotated images and ARs over the three solar activity levels for both datasets. The numbers naturally reflect the fact that ARs appear more (resp. less) frequently in high (resp. low) activity periods, as seen in Figs. 4.1 and 4.2.

We split the datasets into training and testing sets in the following proportions. For LAD, we use 213 images (1380 bounding box) for training, and 53 images (406 bounding box) for testing. For UAD, we use 283 images for training, and 40 images for testing. This amounts to 2205, 1919, 2341, and 2016 training bounding boxes in the 304 Å, 171 Å, 195 Å, and 284 Å bands respectively, and 287, 262, 330, and 263 testing bounding boxes. Furthermore, in order to compare against the localisation of SPOCA, we consider a subset of the UAD testing set for which SPOCA detection results are available in HFC: the SPOCA subset. It consists of 26 testing images (181, 168, 213, and 166 bounding boxes in the 304 Å, 171 Å, 195 Å and 284 Å images respectively). Both datasets are augmented using north-south mirroring, east-west mirroring, and a combination of the two. Augmentation with arbitrary rotations of the images is a popular way of augmenting astronomy datasets. However, such rotations are ruled out from our study because solar ARs tend to appear predominantly alongside the solar equator. All annotations were validated by a solar physics expert.

Pre-processing

In this section, we describe the pre-processing applied to the annotated solar AR datasets, of which we use in both our detection and segmentation (Chapter 5) experiments. Our system takes as input time-matched observations, possibly acquired by different instruments or at different orientations of the same instrument. As such, they need to be spatially aligned prior to analysis. We therefore harmonise the radius and centre location of the solar disk. This is done either using SOHO/EIT image preparation routines (for EIT images), or by thresholding the solar disc –using Otsu thresholding [155]– , finding the minimum enclosing circle fitting, and re-projecting the enclosed disc into a unified centre and radius (for the PM/SH images). Orientation is normalised by SOHO/EIT and PM routines to a vertical north-south solar axis. Although this process does not perfectly normalise the solar coordinates due to a possible small time difference and resulting east-west rotation of the Sun between the two acquisitions, it ensures a sufficient alignment for our purpose of AR detection from spatially and temporally correspondent solar disks. We then eliminate any prominences or solar eruptions that may appear near the solar limb of SOHO/EIT images by masking out all areas outside the solar disk.

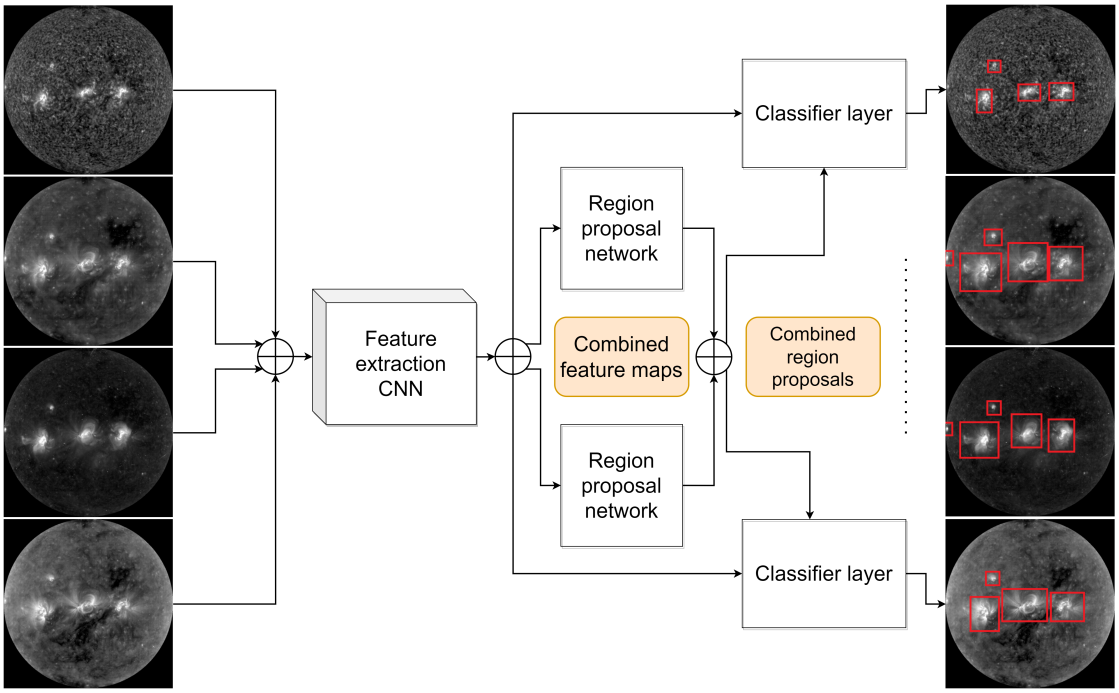

Figure 4.6: MLMT for detection using the Faster-RCNN backbone following the early feature fusion approach. 'Plus' sign denotes concatenation of the feature maps, or of the lists of region proposals (at testing time).

4.3.1.2 BraTS-prime

To further demonstrate the benefits of our joint analysis based approach, we create a synthetic dataset from the BraTS (Brain Tumour Segmentation) multi-modal dataset [113] of similar spatial configurations to the solar imaging bands. BraTS comes with manual labels that can be used to train a detection or segmentation DNN, it consists of full 3D MRI image volumes of brain in 4 modalities (T1GD, T1, T2, and Flair) and 3 classes: enhancing tumour (ET), necrotic and non-enhancing tumour core (NCR/NET), and peritumoural edema (ED). We create the synthetic dataset by selecting one 2D slice of each image modality (T1GD, followed by T1, T2, and Flair), each separated by a spatial gap of size 1 voxel. This emulates the solar images scenario where each band shows ARs in a different solar altitude. Although such spatial gap may seem much lower than for solar images, they are justified by the speed of change of the imaged brain from one slice to another neighbouring one being much larger than for the generally smoother solar ARs. For each modality, we use a total of 11,533 and 190 training and testing images, respectively. From this point on, we refer to this dataset as BraTS-prime.

4.3.2 Detection Stage Evaluation

All tested CNNs were initialised with pre-trained ImageNet [156] weights. Indeed, [157] demonstrated that CNNs pre-trained on RGB images may fine-tune and adapt well to other modalities such as depth images, provided that the image's gain and contrast are suitably enhanced to match those of the pre-training RGB images. A single-channel solar image was repeated along the depth axis resulting in a 3-channel image matching the pre-trained CNN's input depth. Its worth noting that each of the prediction heads (i.e., the band specific region proposal and detection networks) adopt a similar hyper-parameter configuration to that proposed in Faster RCNN [52].

HFC's SPOCA detections were obtained from 171 Å and 195 Å images only, combined as two channels of an RGB image, and SPOCA produces a single detection for both bands. We compare this detection against the ground-truth detections of each of the bands, individually. SPOCA may only combine image bands that are located close to each other in the solar atmosphere and for which it makes sense to produce a common set of detection results. Thus, HFC's SPOCA results are only available for bands of the transition region (171 Å) and low corona (195 Å), and no images from the chromosphere (304 Å) or the high corona (284 Å) were used. However, to prove the versatility of our detector, we also experiment with a combination of chromosphere, transition region, and corona bands on the SPOCA subset in addition to the whole UAD.

4.3.2.1 Independent Detection on Single Image Bands

We first compare detection results produced by Faster RCNN over individual image bands, different DL-based feature extraction networks are tested, namely, ResNet50 and VGG16. We also evaluate the performance of a classical sliding window detection approach based on an SVM [158] classifier and HOG [8] features. Results are presneted in Table 4.2. This serves as baseline to assess our proposed framework.

Generally, we find that both CNN based detectors significantly outperform the classical SVM HOG detector confirming the superiority of CNNs in extracting effective features and learning semantic patterns associated with a given objective in contrast to the crisp hand-crafted features such as HOG. Moreover, we find that ResNet50 produces better results comparing to VGG16, over both UAD and SPOCA datasets. This demonstrates the importance of the residual learning concept in ResNet50 allowing learning deeper features and reducing the risk of vanishing gradi-

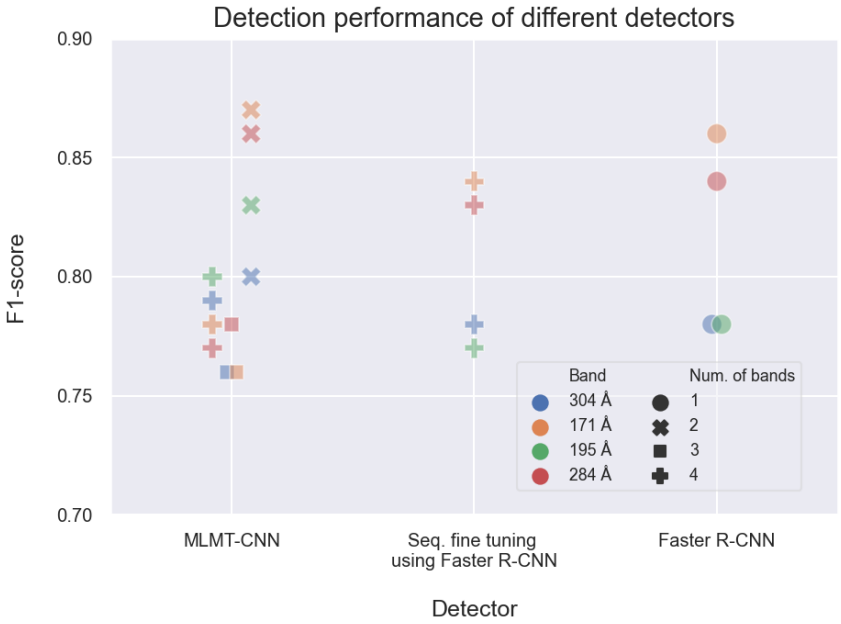

Figure 4.7: Comparison of the detection performance over UAD dataset using different detectors, when using different numbers of image band inputs to perform the analysis.

ant. Accordingly, we continue using ResNet50 as the network of choice for the remaining part of our experiment.

When comparing the detection results per image band, we notice that 304 Å images are repeatedly amongst the most difficult to analyse in UAD, having the lowest F1-scores in all tests. On the other hand, 171 Å shows the highest results of all UAD bands, followed by 284 Å and 195 Å, respectively. This may be explained by ARs having a denser or less ambiguous appearance in 171 Å, 195 Å, and 284 Å image bands than in 304 Å since they are higher in the corona. A similar observation can be made in the LAD dataset when comparing the Magnetogram results to PM/SH 3934 Å, where Magnetograms observe a lower altitude than PM/SH 3934 Å. This demonstrates that the these bands are not equal in how difficult they may be analysed, even though they were acquired at the same time with same size and resolution. These observations suggest that detecting solar ARs using information provided by a single band may be an under-constrained problem.

4.3.2.2 Joint Detection on Multiple Image Bands

We now present the results of our framework when detecting solar ARs over the UAD bands jointly. We experiment with different types of feature fusion and different combinations of

Table 4.2: Detection performance of the single image band detectors. For each band, the highest scores are highlighted in bold.

Detector	Dataset	Band	Precision	Recall	F1
Faster RCNN (ResNet50)	LAD	3934 Å	**0.93**	**0.82**	**0.87**
	LAD	Magn.	**0.89**	**0.78**	**0.83**
	UAD	304 Å	**0.73**	**0.83**	**0.78**
	UAD	171 Å	**0.84**	**0.89**	**0.86**
	UAD	195 Å	**0.81**	**0.75**	**0.78**
	UAD	284 Å	**0.86**	**0.82**	**0.84**
	SPOCA	304 Å	**0.72**	**0.82**	**0.77**
	SPOCA	171 Å	**0.87**	**0.87**	**0.87**
	SPOCA	195 Å	**0.82**	**0.73**	**0.77**
	SPOCA	284 Å	**0.86**	**0.82**	**0.84**
Faster RCNN (VGG16)	UAD	304 Å	0.67	0.78	0.72
	UAD	171 Å	**0.84**	0.81	0.82
	UAD	195 Å	0.79	0.73	0.76
	UAD	284 Å	0.83	0.81	0.82
	SPOCA	304 Å	0.68	0.80	0.74
	SPOCA	171 Å	0.85	0.80	0.82
	SPOCA	195 Å	0.78	0.72	0.75
	SPOCA	284 Å	0.84	**0.82**	0.83
SVM (HOG)	UAD	304 Å	0.55	0.48	0.52
	UAD	171 Å	0.63	0.58	0.61
	UAD	195 Å	0.62	0.58	0.60
	UAD	284 Å	0.84	0.55	0.67

bands. We compare against the state-of-the-art AR detector HFC's SPOCA [125]. We further compare against a sequential fine-tuning method derived from [43] through adapting the first stage of their approach to Faster RCNN by sequentially fine tuning it over the neighbouring image bands. We evaluate this approach on UAD. Moreover, we compare against Faster RCNN on single bands to demonstrate the benefit of jointly processing the image bands, taking into account their inter-dependencies for enhanced individual detections.

In our first experiment, we compare early fusion, i.e., pixel level concatenation, (see Fig. 4.6) against late fusion (feature level concatenation or addition) , on the LAD dataset. Overall, the three approaches show an enhanced performance in contrast to single band based detection. However, we find that late fusion with concatenation shows higher performance than early fusion, having 0.90 F1-score versus 0.88 for magnetograms, while both scored 0.89 over 3934 Å. We further test late fusion using element wise addition and observe a decrease of 1% and 3%

Table 4.3: Detection performance of the MLMT-CNN detectors. For each band, the highest scores are highlighted in bold.

Detector	Fusion	Dataset	Bands	Prec.	Recall	F1
MLMT-CNN (ResNet50 – MOO)	Early – concat.	LAD	3934 Å	0.96	**0.82**	**0.89**
			Magn.	0.95	0.82	0.88
	Late – concat.		3934 Å	**0.97**	**0.82**	**0.89**
			Magn.	**0.96**	**0.85**	**0.90**
	Late – addition		3934 Å	0.95	**0.82**	0.88
			Magn.	0.94	0.80	0.87
MLMT-CNN (ResNet50)	Late – concat.	UAD	171 Å	0.92	0.77	0.84
			284 Å	0.90	0.81	0.85
			171 Å	0.82	0.85	0.83
			195 Å	0.86	0.72	0.78
			195 Å	0.88	0.67	0.77
			284 Å	0.84	0.78	0.81
			304 Å	0.82	0.79	**0.80**
			195 Å	0.87	0.75	0.80
MLMT-CNN (ResNet50 – MOO)	Late – concat.	UAD	171 Å	0.90	0.83	**0.87**
			284 Å	**0.93**	0.80	**0.86**
		SPOCA	171 Å	**0.89**	0.83	**0.86**
			284 Å	**0.92**	**0.80**	**0.86**
		UAD	171 Å	0.86	0.77	0.82
			195 Å	0.89	0.75	0.81
		SPOCA	171 Å	0.83	0.77	0.80
			195 Å	0.86	0.73	0.79
		UAD	195 Å	0.88	0.68	0.77
			284 Å	0.84	0.78	0.81
		SPOCA	195 Å	**0.87**	0.67	0.75
			284 Å	0.81	0.78	0.80
		UAD	304 Å	0.82	0.78	**0.80**
			195 Å	0.88	**0.78**	**0.83**
		SPOCA	304 Å	**0.79**	**0.78**	**0.79**
			195 Å	0.85	0.77	**0.81**
		UAD	304 Å	0.78	0.74	0.76
			171 Å	0.76	0.76	0.76
			284 Å	0.79	0.78	0.78
		UAD	304 Å	**0.93**	0.69	0.79
			171 Å	**0.94**	0.66	0.78
			195 Å	**0.91**	0.72	0.80
			284 Å	**0.93**	0.66	0.77
SPOCA	Early – concat.	SPOCA	171 Å	0.54	**0.93**	0.68
			195 Å	0.58	**0.82**	0.68
[43] using Faster RCNN (ResNet50)	Sequential fine-tuning	UAD	304 Å	0.73	**0.83**	0.78
			171 Å	0.80	**0.90**	0.84
			195 Å	0.83	0.72	0.77
			284 Å	0.86	0.80	0.83

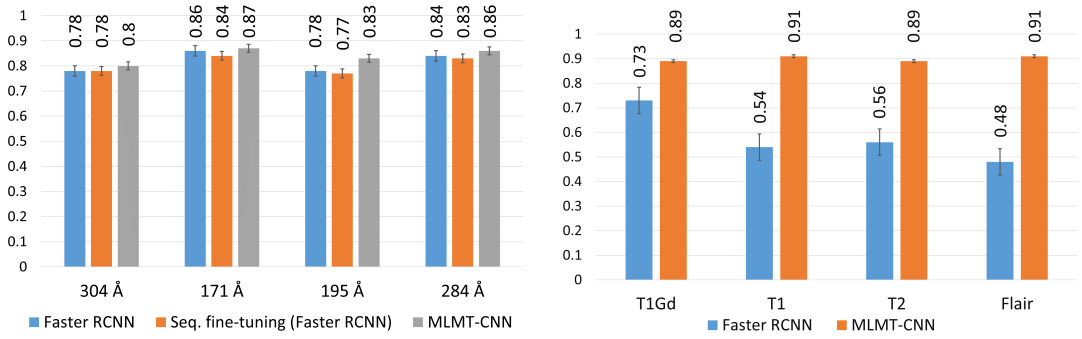

Figure 4.8: Comparison of the detection results over UAD (left) and BraTS-prime (right) datasets. Each group of bars represents the F1-score achieved on an imaging modality. Different colors represent different methods.

in the F1-score over 3934 Å and Magnetogram, respectively. Late fusion is thus adopted for all following experiments.

We also evaluate the benefit of our MOO strategy using our 2-band based architecture on the UAD dataset. As seen in Table 4.3, this approach generally improves the F1-scores in most bands comparing to the non-MOO architectures. This behaviour may indicate that the two feature extraction stages were indeed more effectively optimised for their different tasks at different epochs. Thus we use this MOO approach for all other experiments.

On the UAD dataset, with various combinations of 2 bands, we notice a general improvement over single band detections. In addition, the performance varies in correspondence to the bands being used. Combining bands that are difficult to analyse (304 Å or 195 Å that have lowest F1-scores in the single band analyses) with easier bands (171 Å and 284 Å) unsurprisingly enhances their respective performance. More interestingly, combining the difficult 304 Å and 195 Å bands together also improve on their individual performance. Similarly, when combining bands that are easier to analyse (171 Å and 284 Å), performances are also improved over their individual analyses. Following these settings, our 2-band based approach was able to record higher or similar F1-scores in contrast to the best performing single-band detector. This supports our hypothesis that joint detection may provide performance gains through learning the inter-dependencies between the image bands. Moreover, the most dramatic improvement in F1-scores across both LAD and UAD datasets is for the 3934 Å images when magnetograms are added to the analysis. This is in line with the current understanding of AR having strong magnetic signatures.

Generally, in the UAD dataset, we find that using a combination of 2 bands produces the best

Table 4.4: AR detection performance of MLMT-CNN and baseline detectors when an IoU (intersection over union) based evaluation criterion is used. Results presented in the table are found using two IoU threshold values, 0.5 and 0.25 (indicated within brackets). The rows with grey background indicate results produced using single-band based detectors. For each band, the highest scores are highlighted in bold.

Detector	Dataset	Bands	Prec.	Recall	F1
			IoU 0.5 (0.25)		
Faster-RCNN (ResNet50)	UAD	304 Å	**0.56** (0.70)	**0.65** (**0.82**)	**0.60** (**0.75**)
		171 Å	**0.65** (0.78)	0.68 (0.82)	**0.66** (0.80)
		195 Å	0.61 (0.76)	0.52 (0.57)	0.56 (0.65)
		284 Å	0.66 (0.82)	**0.63** (**0.79**)	0.64 (**0.81**)
	SPCOA	304 Å	**0.55** (0.70)	**0.65** (**0.82**)	**0.59** (**0.75**)
		171 Å	0.66 (0.81)	0.67 (**0.83**)	0.67 (**0.82**)
		195 Å	0.60 (0.76)	0.54 (0.69)	0.57 (0.73)
		284 Å	0.68 (0.83)	**0.65** (**0.80**)	**0.66** (0.82)
MLMT-CNN (Late concat. – ResNet50)	UAD	304 Å	0.51 (**0.77**)	0.48 (0.73)	0.49 (**0.75**)
		171 Å	0.69 (**0.86**)	0.61 (0.76)	0.65 (**0.81**)
		195 Å	**0.67** (**0.83**)	**0.59** (**0.73**)	**0.62** (**0.77**)
		284 Å	**0.69** (**0.88**)	0.59 (0.76)	0.64 (**0.81**)
	SPOCA	304 Å	0.52 (**0.75**)	0.50 (0.73)	0.51 (0.74)
		171 Å	**0.70** (**0.84**)	0.64 (0.76)	0.66 (0.80)
		195 Å	**0.64** (**0.80**)	**0.58** (**0.73**)	**0.61** (**0.76**)
		284 Å	**0.71** (**0.88**)	0.62 (0.78)	**0.66** (**0.83**)
SPOCA	SPOCA	171 Å	0.16 (0.34)	0.26 (0.56)	0.19 (0.42)
		195 Å	0.16 (0.33)	0.20 (0.43)	0.18 (0.37)
[43] using Faster-RCNN (ResNet50)	UAD	304 Å	**0.56** (0.70)	**0.65** (**0.82**)	**0.60** (**0.75**)
		171 Å	0.62 (0.74)	**0.70** (**0.84**)	**0.66** (0.79)
		195 Å	0.62 (0.79)	0.55 (0.70)	0.58 (0.74)
		284 Å	0.68 (0.83)	0.62 (0.75)	**0.65** (0.79)
	SPOCA	304 Å	**0.55** (0.70)	**0.65** (**0.82**)	**0.59** (**0.75**)
		171 Å	0.66 (0.77)	**0.71** (0.82)	**0.68** (0.79)
		195 Å	0.62 (0.76)	0.54 (0.67)	0.58 (0.71)
		284 Å	0.69 (0.83)	0.63 (0.77)	**0.66** (0.80)

F1 scores in comparison to using 3 or 4 bands in the analysis, see Table 4.3 and Fig. 4.7. This may be caused by the fact that optimising the network for multiple tasks (2, 3, or 4 detection tasks) simultaneously increases the complexity of the problem. While the network successfully learned to produce better detections in the case of 2 bands, it was difficult to find a generalised yet optimal model for 3 or 4 bands at the same time. Thus, for 4 bands, the model obtains the best precision but at the expense of a poor recall. Visual results are presented in Fig. 4.9.

Table 4.5: F1-scores of single image band based detectors against MLMT-CNN with different fusion strategies over BraTS-prime (with 1 slice gap). All detectors are based on ResNet50. For each band, the highest scores are highlighted in bold.

Bands	Faster RCNN	MLMLT-CNN (Early - addition)	MLMLT-CNN (Early - concat.)	MLMLT-CNN (Late - concat.)
T1Gd	0.73	0.74	0.83	**0.89**
T1	0.54	0.78	0.89	**0.91**
T2	0.56	0.76	0.86	**0.89**
Flair	0.48	0.75	0.86	**0.91**

We further compute the average precision (AP) score over the UAD dataset for both, the best performing MLMT-CNN (i.e., using late concatenation feature fusion based on 2 image band inputs) and Faster RCNN using single band based analysis. We observe that MLMT-CNN consistently obtains higher average precision scores of 0.88, 0.96, 0.92, and 0.97, in contrast to Faster RCNN with AP of 0.80, 0.90, 0.89, and 0.92 over the SOHO/EIT bands 304 Å, 171 Å, 195 Å, and 284 Å, respectively. This confirms the advantage of the joint analysis approach when detecting solar ARs. On the other hand, the single band based analysis using Faster RCNN requires less inference time of ~ 0.5 seconds (GPU time) per image band input, in contrast to ~ 0.8 seconds when using the 2 bands based joint analysis approach.

On the SPOCA subset, over the bands 171 Å and 195 Å for which it is originally designed, the SPOCA method obtains the poorest performance, in terms of F1 score, of all multi-band and single-band experiments. It is worth noting that this method relies on manually tuned parameters according to the developers' own definition and interpretation of AR boundaries, which may differ from the ones we used when annotating the dataset. While supervised DL-based methods could integrate this definition during training, SPOCA could not perform such adaptation. This may have had a negative impact on its scores. Furthermore, visual inspection shows a poor performance for SPOCA on low solar activity images, see Fig. 4.2. This may be due to the use of clustering in SPOCA, since in low activity periods the number of AR pixels (if any) is significantly smaller than solar background pixels, which makes it difficult to identify clusters.

Moreover, the sequential fine tuning approach similar to [43] shows a close performance to single band detection using Faster RCNN with an identical precision, recall and F1-score over the band 304 Å and a slight decrease over the other 3 bands, See Table 4.3 and Fig. 4.8. This may be due to the fact that its transfer learning does not incorporate the bands' inter-dependencies when analysing the different bands. Moreover, the method was designed in [43] to produce a single prediction for the different bands, this differs from our usage where we predict a different

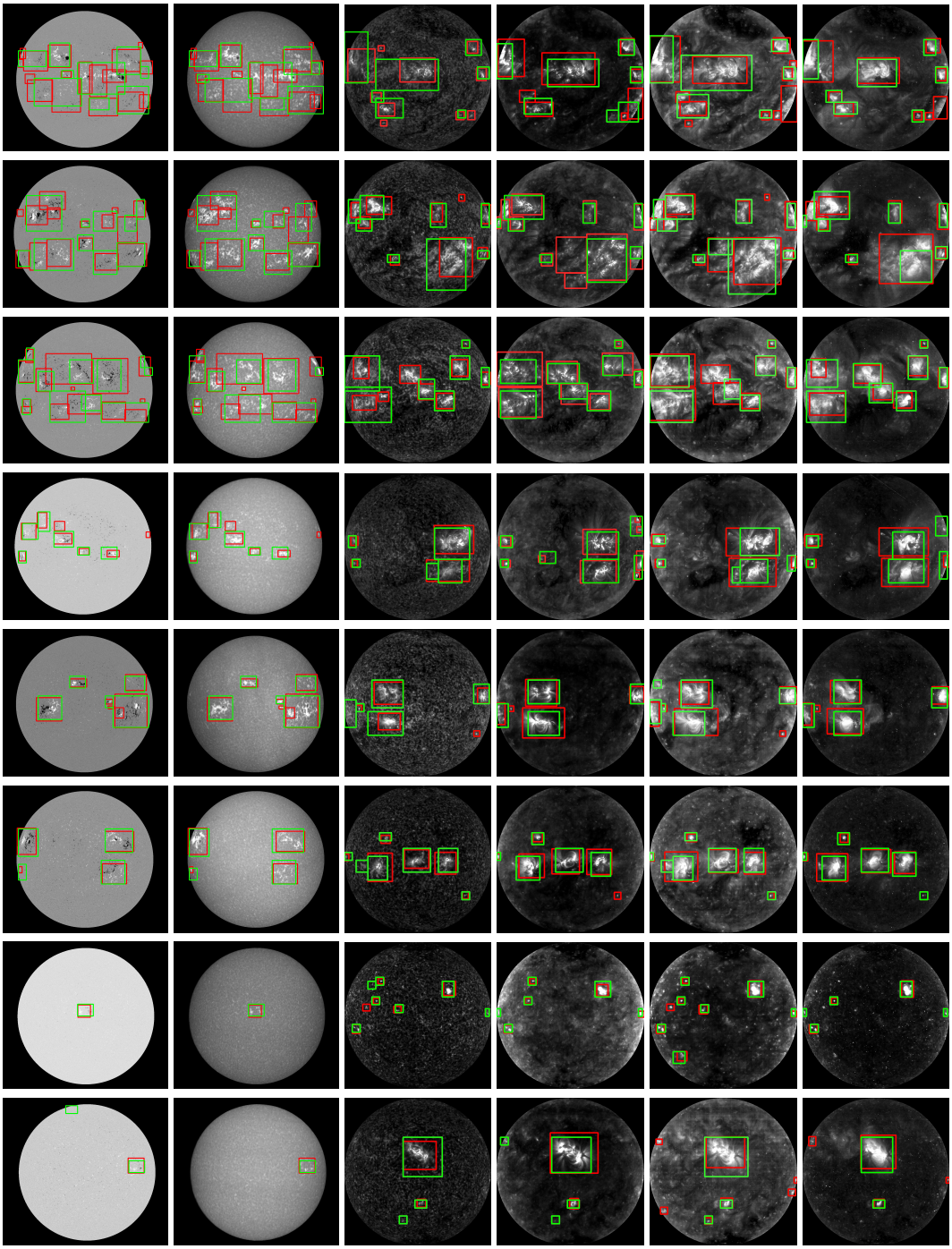

Figure 4.9: Ground-truth (red) and MLMT-CNN's (green) detection of solar ARs in randomly selected images from (left to right) SOHO/MDI Magnetogram and PM/SH 3934 Å, SOHO/EIT 304 Å, 171 Å, 195 Å, and 284 Å. Contrast has been increased for convenience of visualisation.

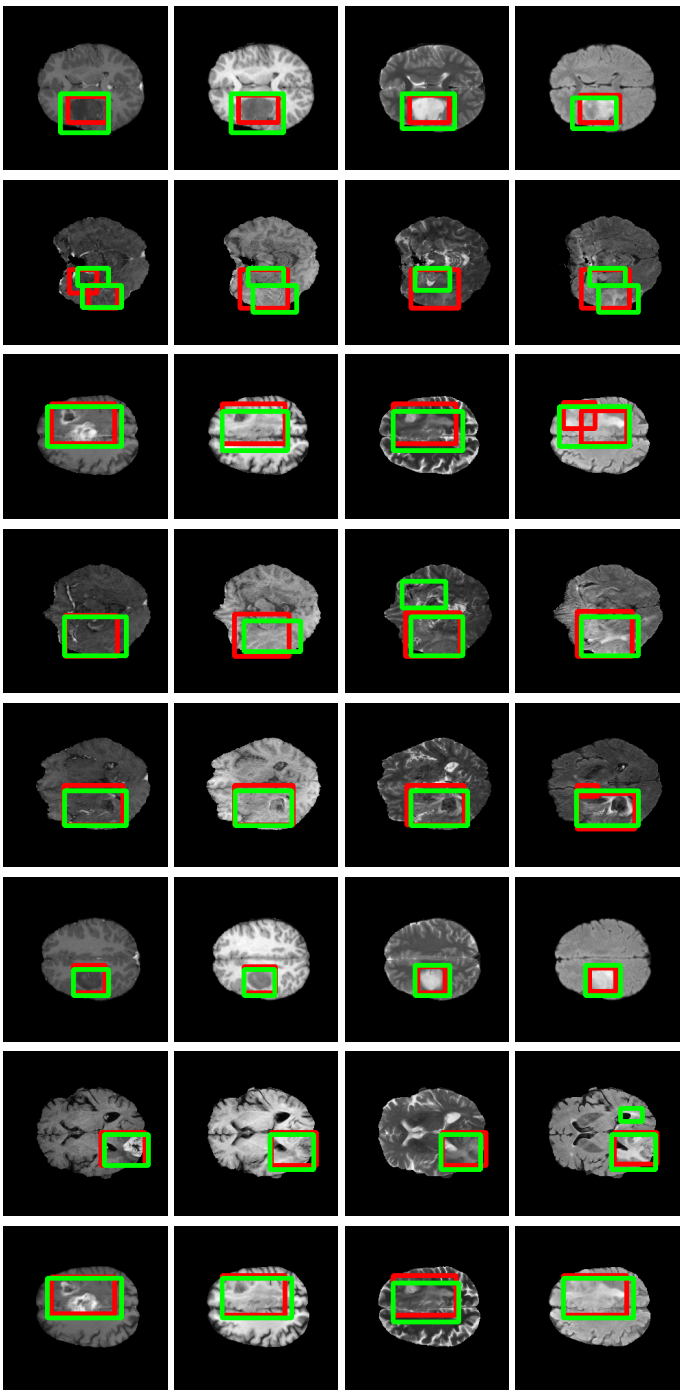

Figure 4.10: Detection results of MLMT-CNN (green) over the BraTS-prime dataset (red) in randomly selected images. Each row indicates a single multi-modal image sample. Columns indicate the modalities T1Gd, T1N, T2N, and Flair, respectively.

set of detections per band.

Additionally, we evaluate our method using IoU and compare our results to both, single band based detection by Faster-RCNN and joint detection from SPOCA and the fine tuned network of [43] (see Table 4.4). When using an IoU threshold of 50%, we find that our method produces the highest F1 score over the 195 Å band amongst all methods on both UAD and SPOCA datasets, with a comparable performance over 171 Å and 284 Å. On the other hand, our method shows a drop in performance performance on 304 Å comparing to Faster RCNN and the sequentially fine-tuned network. Generally, all methods show a significant decrease in the F1 scores when using IoU based criterion, with SPOCA being the lowest amongst all methods. A similar pattern is observed when using a less strict IoU threshold of 25%. During our visual inspection, we notice that in some cases, detected solar ARs are regarded false positives when evaluated against the IoU criterion. Particularly, when a cluster of neighbouring active region areas (i.e., neighbouring solar ARs) is detected as a single AR structure, or when an AR is detected by multiple neighbouring boxes. See Fig. 4.4. These observations suggest that an IoU based criterion does not perfectly capture the dynamic characteristics of solar ARs, even when incorporating an IoU threshold as low as 25%. Unlike generic object detection tasks, where object morphology and boundaries are well defined, solar ARs are dynamic structures that are continuously evolving (e.g., merging and splitting, emerging and dying out) [125, 148–150]. Therefore, when designing our evaluation criterion (Section 4.3), we take into account the aforementioned scenarios associated with solar ARs.

We further evaluate our detection approach with different fusions, over the 4 bands of BraTS-prime dataset, and compare it against single band based detection. All fusion strategies significantly outperform single band detectors, with late concatenation fusion being the highest, showing an average F1-score increase of 39% across all modalities. See Table 4.5 and Fig. 4.8. This confirms our hypothesis that exploiting inter-dependencies between the image bands by the joint analysis may provide a superior performance in contrast to single band based detection. Visual results over the BraTS-prime dataset are presented in Fig. 4.10

4.4 Summary

Accurate detection of solar ARs is an important step in studying solar weather and phenomenons. Contrary to typical multi-spectral imaging scenarios, where the different image bands show different compositions of an observed scene, in the solar scene, different bands image different

altitudes in the solar atmosphere, showing different layers of 3D objects (e.g., solar ARs) that span the solar atmosphere.

In this chapter, we presented MLMT-CNN, a multi-layer and multi-tasking framework to tackle the 3D solar AR detection problem from multi-spectral images that observe different layers of the 3D solar atmosphere. The proposed approach dynamically analyses information from neighbouring layers (imaging bands) by learning band-specific features that are aggregated at different semantic levels and analysed jointly to capture cross-band correlations. The resulting embeddings are used to produce separate, yet related, results for each of the neighbouring layers. To address the lack of labelled solar AR datasets, we design a multi-spectral labelling tool, in which we use to create two, deep learning suitable, ground- and space-based, multi-spectral AR detection datasets. All annotated images were validated by a solar physics expert. Additionally, we create a synthetic dataset of similar spatial configurations to that in the solar images, from a multi-modal brain imaging dataset, and use it to further evaluate the proposed approach.

In our experiments, we find that by fusing information from different image bands at different feature levels, CNNs were able to detect objects more consistently across the different layers. Furthermore, our study suggests that different imaging scenarios may require different types of feature fusion strategies. We also show that the number of bands used in the analysis might affect the performance and must be optimised to each imaging scenario. Generally, MLMT-CNN showed competitive results against both baseline and state-of-the-art detection and methods.

The proposed framework is versatile and may use different CNN backbones or tasks, it may also be straightforwardly generalised to any number or modalities of images. To demonstrate this, in Chapter 5, we extend our framework to perform multi-layer solar AR segmentation by adapting the main concepts of our proposed detection approach. Moreover, in Chapter 6, we generalize and reformulate our cross-band joint analysis approach to handling volumetric 3D imaging scenarios, where we investigate the influence of incorporating inter- and cross-channel correlation learning in 3D medical imaging localisation and classification tasks. In Chapter 7, we further study the impact of exploiting long range correlations and global context and compare it to the inter- and cross-channel correlation modelling approach. The work in this chapter has been presented in the following publications:

- M. Almahasneh, A. Paiement, X. Xie, J. Aboudarham, Active region detection in multi-spectral solar images. International Conference on Pattern Recognition Applications and Methods, 2021.

- M. Almahasneh, A. Paiement, X. Xie, J. Aboudarham, MSMT-CNN for solar active region detection with multi-spectral analysis. Springer Nature Computer Science, 2022.
- M. Almahasneh, A. Paiement, X. Xie, J. Aboudarham, MLMT-CNN for object detection and segmentation in multi-layer and multi-spectral images. Machine Vision and Applications, 2021.

Chapter 5

MLMT-CNN for Object Segmentation in Multi-layer and Multi-spectral Images

Contents

5.1 Introduction

We investigated the possibility of exploiting DL methods to detect solar ARs in the form of 3D bounding boxes in multi-spectral images that observe different altitudes in the solar atmosphere, in Chapter 4. We proposed a multi-layer multi-tasking (MLMT-CNN) joint analysis framework that incorporates inter-band information and cross-band spatial correspondence and feature-level correlations to extract effective features and produce separate, yet related, detections across the individual imaging bands (layers). In this chapter, extending on our previous work, we investigate the possibility of generalising the proposed approach to perform AR segmentation from the multi-layer solar images. We design a multi-tasking segmentation approach (i.e., performs the AR segmentation in the different bands with respect to their band-specific ground-truth simultaneously), that combines the concepts of our joint analysis approach and the encoder-decoder design of U-Net [28]. The proposed approach takes predicted locations from the detection stage in the form of bounding boxes (i.e., Chapter 4) and exploits cross-band information, in addition to spatial and semantic information of different levels to produce more accurate segmentations. Moreover, to address the difficulty of producing accurate and detailed annotations for AR segmentation, we propose a recursive training approach based on weak labels (i.e., bounding boxes).

Developing segmentation DNNs requires a big number of pixel-wise labels. Designing such label can be time consuming and may require extensive labour, particularly when dealing with multi-layer data, where 3D objects (e.g., solar ARs) must be labelled in each band separately. Additionally, from the solar data perspective, the fuzzy nature of AR boundaries makes preparing such dataset more challenging. Therefore, we investigate weakly-supervised learning methods to overcome these issues. Weak-supervision is a branch of machine learning where noisy or partially-labelled data samples are used to provide supervision signals in a supervised learning set up. Both [72] and [43] show that CNNs can be directly trained to segment multi-spectral images directly using weak labels. [72] demonstrates that the network was able to learn generalised representations from the weak annotations. In [73, 77, 78], an iterative training strategy was used to train segmentation CNNs from initial weak dense annotations that are approximated from bounding boxes. By updating the initial training labels using the predictions from previous training iterations, CNNs were able to gradually refine their predictions and result in a close performance to models trained with full supervision. These findings suggest that this iterative approach may be opportune for our solar data case, hence, we utilise an iterative training strategy for our segmentation task.

In this chapter, our contribution may be summarised as follows: 1) we extend our multi-layer multi-tasking (MLMT-CNN) detection framework proposed in chapter 4 to perform pixel-wise classification in multi-spectral images that observe different layers of a 3D object, using our joint analysis strategy to exploit inter-band and cross-band spatial correspondence and feature-level correlations. 2) To address the difficulty of producing accurate and detailed annotations for AR segmentation, we propose a recursive training approach based on weak labels (i.e., bounding boxes). 3) We demonstrate the proposed segmentation approach using state-of-the-art backbone CNNs and evaluate its performance over different datasets (solar ARs, BraTS-prime, and a multi-spectral satellite cloud imaging dataset), different levels of supervision, and different combinations of input bands.

The rest of this chapter presents the details of the proposed method in Section 5.2, Section 5.3 presents the dataset details and carries out extensive analysis and experimental results. Lastly, Section 5.4 discusses key conclusions and provides an overall summary for this chapter.

5.2 MLMT-CNN: Segmentation Stage

Similar to our detection framework (i.e., Chapter 6), our segmentation approach takes as an input time matched multi-band images, in this case, solar ARs that are detected by the detection stage. The network design aims to dynamically embed information from the neighbouring solar layers (image bands) by focusing on learning band-specific features using parallel feature extraction branches, while simultaneously learning cross-band relations by jointly analysing the extracted features from different bands through cross-band feature aggregation. Finally, based on a multi-tasking objective, band specific networks are used to predict segmentation masks per image band.

This framework is versatile and may be used with various DNN backbones. In Chapter 6, we experiment with Faster RCNN as the backbone of our detection stage, here, we experiment with U-Net as a backbone to perform the segmentation task, demonstrating the benefits of our joint analysis scheme in learning the inter-dependencies between the different image bands in both stages. We also evaluate our approach for different applications (solar ARs, BraTS-prime, and Cloud-38-prime datasets), using different feature aggregation types and levels (addition and concatenation, early and late).

In this section, we introduce the backbone network used in our segmentation framework in Section 5.2.1, and the details of our segmentation strategy in Sections 5.2.2.

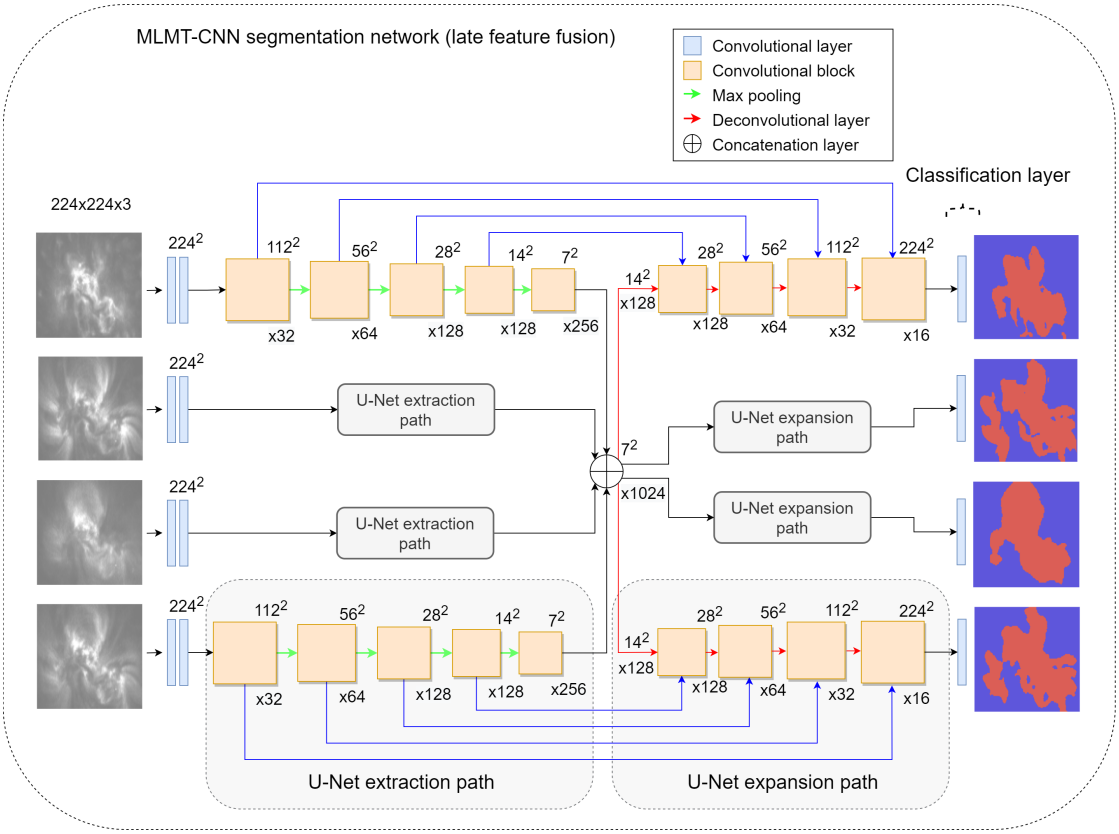

Figure 5.1: MLMT-CNN segmentation architecture using the U-Net backbone, following the late feature fusion approach. Inputs from different bands are first analysed using band-specific feature extraction CNN branches, these are then cross-integrated and jointly analysed to allow the network capture spatial correspondence and dependencies between the different bands. Blue arrows are skip connections, applied to the appropriate channel of the joint feature map for each branch. Finally, the network predicts band-specific masks based on the aggregated features, driven by a multi-tasking objective.

5.2.1 Backbone Networks

The key principles of our framework (see Section 4.2.1) may straightforwardly generalise to different backbones since they are architecture agnostic. For our segmentation task, we adopt the principles of U-Net [28] to design our network. Nevertheless, other competing networks can also be used, and we also experimented with FCN8 [27] in early tests. U-Net is a fully convolutional network that consists of 3 main parts: 1) contraction path, 2) bottleneck, and 3) expansion path. We dynamically use these 3 components to design our network such that individual band segmentations are predicted. We also take advantage of U-Net's skip connections to allow combining features from different semantic and spatial levels within the same band, this maximises the learned information within individual bands. These features are fused at the bottleneck of

U-Net to allow capture cross-band spatial correspondence and cross-feature correlations.

5.2.2 Segmentation Framework

Our segmentation framework is presented in Figs. 5.1 and 5.2. It consists of 3 parts: 1- band-specific feature extraction, 2- cross-band feature fusion, and 3- band-specific mask reconstruction. The network takes as input the AR detections (patches) produced by the detection stage. Each detection is cropped from all image bands, and resized into 224x224 pixel before entering the segmentation network.

The feature extraction part consists of parallel U-Net contracting paths (one per band), each specialised to extract a feature map from its band individually. The resulting feature maps are then combined in the latent space (i.e., late fusion). It is worth noting that different feature fusion operations may be used. In this work, we experiment with addition and concatenation. The combined feature map is passed to the mask reconstruction part where parallel U-Net expensive paths (a specialised path per band) perform the final prediction. Skip connections are utilised between each band's contracting path and its correspondent expensive path to preserve fine details learned in early layers of that band (blue arrows in Fig. 5.1).

To overcome the lack of dense AR annotation, we use weak labels to train our segmentation network along with a recursive training approach. In the first round of iterations, weak annotations are used to guide the training. Once the network converges, the training is repeated from random weights using the new labels predicted by the model from the previous round. This process is repeated until validation loss stops decreasing, or starts to increase. The idea is inspired by [72, 73, 77, 78], where authors demonstrate that iteratively training segmentation CNNs with weak labels can achieve results close to fully supervised.

Our weak label was carefully designed to provide a conservative representation of solar ARs, favouring a high precision over recall, to accelerate the first training round (as detailed in 5.3). Recursive training allows the network to learn a more generalised representation in a self-supervised recursive manner that aims to limit the bias that may be introduced by the initial weak label. In other words, given a carefully engineered (generally representative) weak label, the essence of our approach is to gradually (recursively) converge towards a more generalised decision boundary (model) in which the predicted segments are more meaningful in contrast to the previous constrained training iteration of which model is biased towards ill defined labels. This is in line with the discovery that sampling as little as 4% of the pixels to compute the training loss enables CNNs to achieve a close performance to fully supervised, caused by the

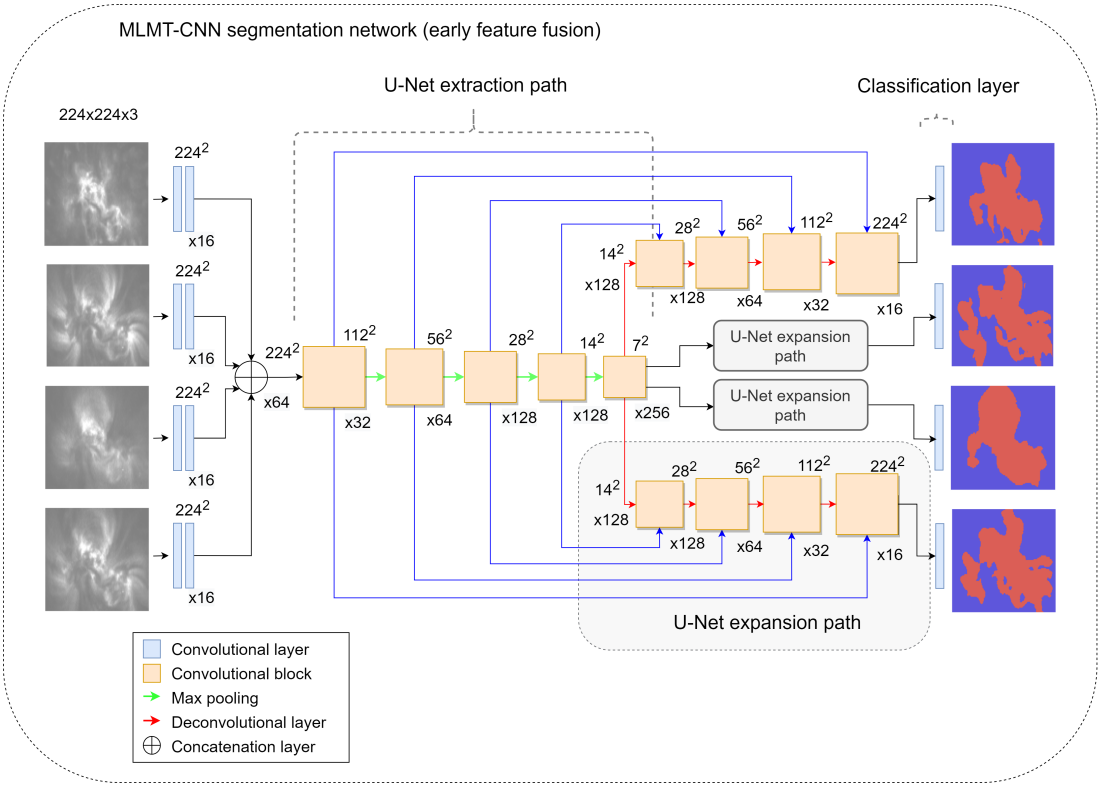

Figure 5.2: MLMT-CNN segmentation architecture using the U-Net backbone, following the early feature fusion approach.

strong correlation within the training data of a pixel-level task [159]. It is worth noting that in the case of using incorrect weak labels (e.g., limited task representation) to initialise the training procedure, the learning process of the target task may be hindered and may indeed lead to the divergence of the model. Therefore, care must be taken when designing the weak label to maximise the representation of the target task. Including a reference manually labelled validation subset may form a tool to better interpret the performance of the learned model. However, in the case of limited availability of such solution, careful qualitative examination may be followed to interpret the predictions of the learned model. The results of our recursive approach were validated by a solar physics expert, and will be further discussed in Section 5.3.

Moreover, the solar data suffers from a class imbalance by nature, since most of the solar disk is covered by quite sun (solar background). The use of AR crops (patches from previous detection) helps in reducing this imbalance significantly, yet it does not solve the matter completely. Hence, we train our model using a weighted categorical cross entropy loss that combines

information from all image bands as follows:

$$L(y,\hat{y}) = -\sum_{b=0}\sum_{c=0}\omega_c \sum_{i=0} y_{icb} \ * \ log\ (\hat{y}_{icb}) \tag{5.1}$$

where y and $\hat{y}$ are the actual and the predicted classes, respectively, ω_c is the weight of the c^{th} class, and i and b denote the pixel and the band being processed, respectively. We use the values 2, 1, and 2 as the weights for the three AR, solar background (quite sun), and image background classes, respectively. These weights were found to be best performing by experimenting with different values based on prior computed class ratios. Adding the weighting term to the combined loss prevents any bias that might be caused by the dominating solar background class.

5.3 Experiments

Our segmentation experiment was carried out using Tensorflow on an NVIDIA GeForce GTX 1080 Ti GPU. The segmentation stage was trained for 250 epochs ($\sim$0.41 day) using Adam optimiser [48] with a starting learning rate of 4e-3. All methods under comparison were evaluated using the IoU (intersection over union) criteria. It is worth noting that other popular similarity measures (e.g., Dice score and Hausdorff distance) may also be used to evaluate the segmentation performance. Dice score and IoU are positively correlated, they both range between 0 (no similarity) and 1 (highest similarity), using either of these metrics may therefore suffice the purpose of comparing the performance of different models. On the other hand, Hausdorff distance [160] may be used to evaluate how close the predicted object boundaries are with respect to the boundaries defined in the ground-truth. However, since our segmentation approach targets data of weak labels (solar ARs in this case), such evaluation is not straightforwardly applicable and is therefore avoided.

5.3.1 Data

5.3.1.1 Weak AR Segmentation Labels

The vanishing nature of AR borders makes preparing dense AR labels difficult, subjective, and time-consuming [125]. To address this, we adopt a weak labelling procedure, along with a recursive training technique that will be detailed in Section 5.3.2, to perform the AR segmentation. Starting from the bounding box, we produce an imperfect but quick weak annotation for the iterative training procedure. Its worth noting that weak labels are only produced for the UAD datset, see Table 4.1.

Table 5.1: Performance of single image segmentation over BraTS-prime (fully supervised detectors) and Weak-BraTS-prime (weakly-supervised detectors). For each class, the highest scores are highlighted in bold.

Architecture	Supervision	Bands	IoU score per class			Mean IoU
			NCR/NET	ED	ET	
FCN8	Fully supervised	T1Gd	0.54	0.43	0.70	0.56
		T1	0.08	0.33	0.0	0.14
		T2	0.49	0.48	0.23	0.40
		Flair	0.43	0.51	0.19	0.38
U-Net	Fully supervised	T1Gd	**0.69**	0.52	**0.80**	**0.67**
		T1	0.56	0.50	0.19	**0.42**
		T2	0.63	0.56	0.36	**0.52**
		Flair	0.50	**0.59**	0.29	**0.46**
U-Net	Weakly-supervised	T1Gd	0.66	0.33	0.53	0.51
		T1	0.58	0.39	0.0	0.32
		T2	0.58	0.43	0.1	0.37
		Flair	0.44	0.49	0.0	0.31

The weak annotation is generated by the following procedure. A local contrast stretching is applied to each AR bounding box (i.e., detection label) by stretching its intensities between the 65^{th} and the 98^{th} percentiles (picked empirically), forcing a gap between the two quiet Sun (i.e., background, lower range of values) and AR (higher range of values) classes. This is followed by an Otsu binarisation that minimises within-class variance and maximises between-class variance [155]. A mild morphological dilation follows to suppress any small holes in the initial mask, done using a 3x3 elliptic structuring element. Lastly, any remaining components with area under 10 pixels are discarded, resulting in the weak AR mask.

This segmentation label approximation approach aims at only labelling pixels that we are certain about as AR considering their evident activity, building on prior knowledge on the pixel intensity (i.e., solar ARs are the brightest regions in the solar disk). This is motivated by the discovery in [159] that training data of a pixel-level task has a strong between-sample correlation, and that randomly sampling as little as 4% of the pixels to train a CNN can achieve about the same performance as full supervision.

5.3.1.2 Weak-BraTS-prime

In order to evaluate our recursive training approach, we create weak labels for our synthetic dataset BraTS-prime presented in Section 4.3.1.2. This is achieved by introducing a morpho-

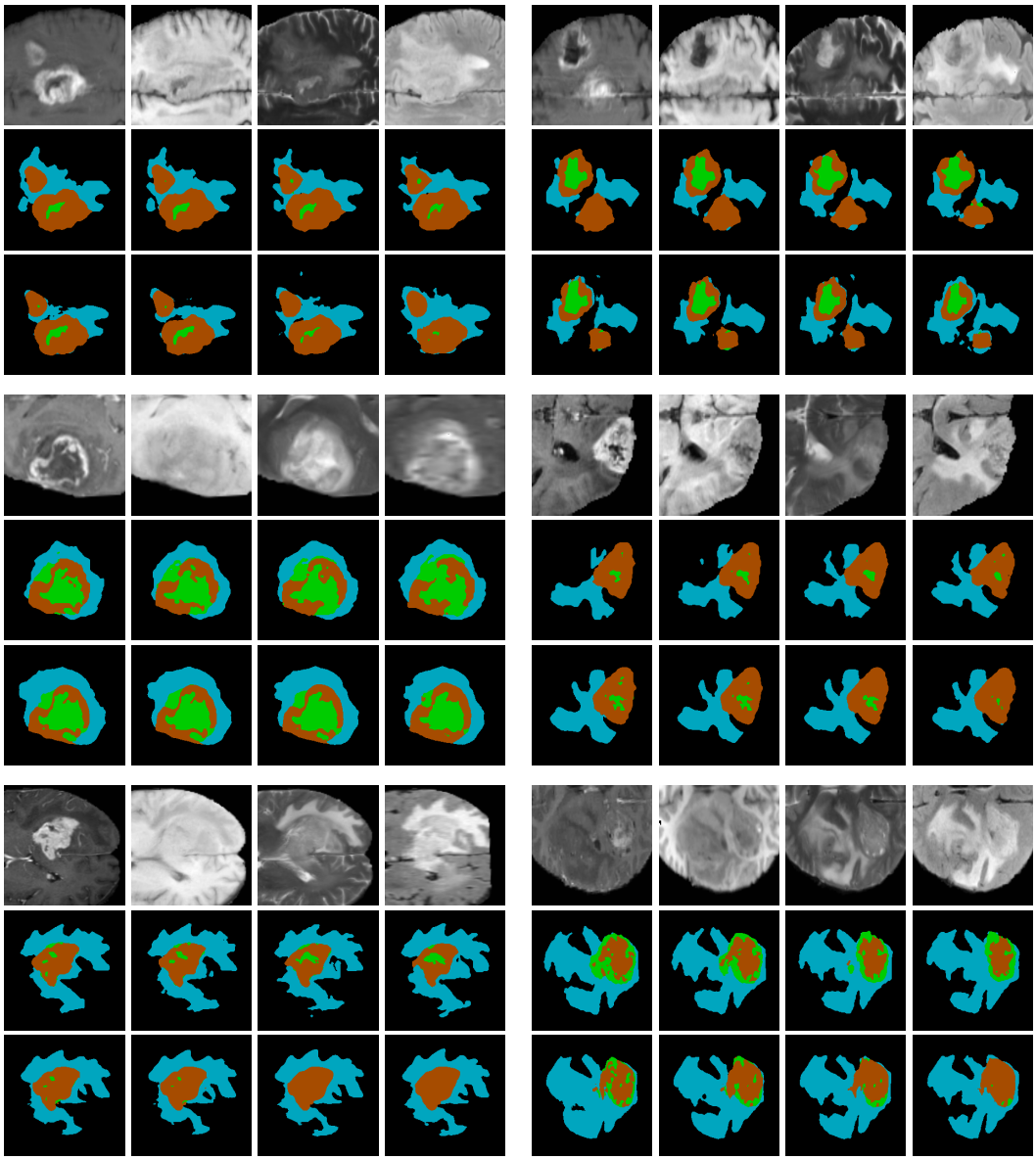

Figure 5.3: MLMT-CNN segmentation performance over BraTS-prime dataset. Each group of images represents a single multi-modal sample. Rows in each group show, from top to bottom, images, ground-truth masks, and predicted masks, respectively. Columns show the modalities T1Gd, T1N, T2N, and Flair, respectively. Masks show 3 main classes, enhancing tumour (blue), the peritumoral edema (brown), and the necrotic and non-enhancing tumour core (green).

Table 5.2: Segmentation performance of MLMT-CNN (U-Net) with full supervision over BraTS-prime for different numbers of modalities and feature fusions. For each class, the highest scores are highlighted in bold.

Architecture	Fusion	Slice gap	Bands	IoU score per class			Mean IoU
				NCR/NET	ED	ET	
MLMT-CNN (U-Net)	Early - concat.	1	T1	0.60	0.56	0.41	0.52
			T2	0.59	0.59	0.39	0.52
			T1	0.62	0.59	0.35	0.52
			T2	0.63	0.61	0.36	0.53
			Flair	0.63	0.63	0.39	0.55
			T1Gd	**0.75**	**0.66**	0.78	**0.73**
			T1	**0.75**	**0.69**	**0.76**	**0.73**
			T2	**0.74**	**0.71**	**0.70**	**0.72**
			Flair	**0.73**	**0.68**	0.62	**0.68**
	Early - addition	1	T1Gd	0.74	0.64	0.78	0.72
			T1	0.73	0.67	0.74	0.71
			T2	0.69	0.66	0.65	0.67
			Flair	0.68	0.67	0.61	0.65
	Late - concat.	1	T1Gd	0.71	0.63	0.81	0.72
			T1	0.73	0.65	0.74	0.71
			T2	0.71	0.68	**0.70**	0.70
			Flair	0.69	0.67	**0.64**	0.67
	Late - addition	1	T1Gd	0.70	0.60	0.81	0.70
			T1	0.71	0.63	0.73	0.69
			T2	0.67	0.66	0.67	0.67
			Flair	0.68	0.66	0.60	0.65
	Early - concat.	2	T1Gd	0.68	0.55	0.76	0.66
			T1	0.67	0.61	0.66	0.65
			T2	0.64	0.65	0.55	0.61
			Flair	0.57	0.62	0.46	0.55
		3	T1Gd	0.63	0.51	0.73	0.62
			T1	0.63	0.60	0.62	0.62
			T2	0.57	0.64	0.43	0.55
			Flair	0.59	0.61	0.41	0.54
[43] using U-Net	Sequential fine-tuning	1	T1Gd	0.71	0.55	**0.82**	0.69
			T1	0.56	0.50	0.19	0.42
			T2	0.65	0.58	0.26	0.50
			Flair	0.57	0.61	0.33	0.50

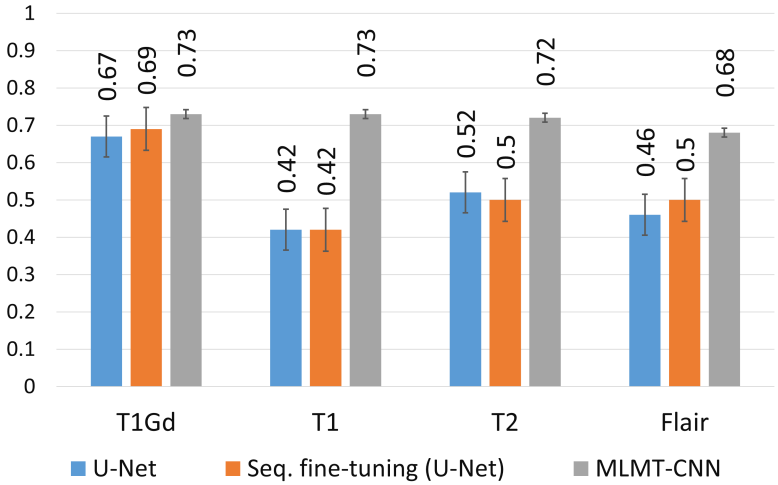

Figure 5.4: Comparison of the segmentation results over BraTS-prime dataset. Each group of bars represents the IoU score achieved on an imaging modality. Different colors represent different methods.

logical erosion (using a 4x4 elliptic kernel) to the original manual annotation. This process was applied to each class separately, and the final weak label was produced by aggregating the eroded masks of all classes. This approach differs from our approach in labelling solar ARs, due to the different complexity and number of classes between the two datasets, and their different appearances of images which call for different unsupervised segmentation methods. Nonetheless, it is in line with the idea of including only pixels that we are confident about in the weak label, which we follow when labelling solar ARs. In this study, we refer this dataset as Weak-BraTS-prime. As discussed in Section 4.3.1.2, the synthetic BraTS-prime dataset is created by selecting one 2D slice of each image modality separated by a spatial gap of size $g = 1$ to emulate the solar images scenario where each band shows ARs in a different solar altitude. Here, we further experiment with g being either 1, 2, or 3 voxels, to instigate the influence of the image modalities having different levels of spatial correspondence on the segmentation.

5.3.1.3 Weak-Cloud-38

We further evaluate our recursive training approach on a third weakly labelled dataset derived from the Cloud-38 [71] multi-spectral dataset. Cloud-38 consists of Landsat-8 observations in 4 bands (Near Infrared, Red, Green, and Blue), and their pixel-wise ground truth for cloud segmentation. The variety of cloud shapes, sizes, and densities (compactness), in addition to the complexity of the segmentation problem (i.e., number of classes), makes this dataset closer to our AR scenario than Weak-BraTS-prime, and hence, a fairer comparison. However, it is worth

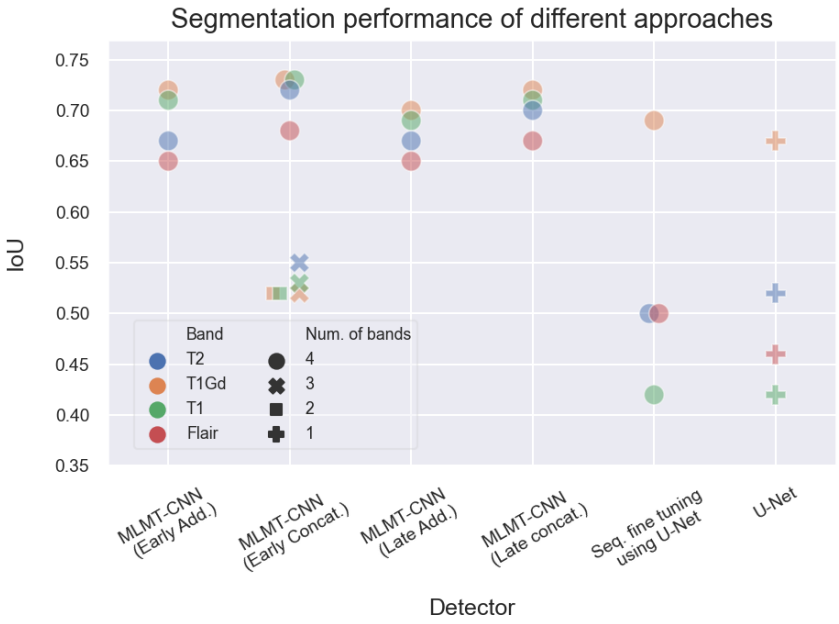

Figure 5.5: Comparison of the segmentation performance over BraTS-prime multi-modal dataset using different feature fusion strategies and different numbers of image band inputs.

noting that the 3D aspect is missing from this dataset, as the different bands image the same (∼2D) scene. To prepare the weakly labelled dataset, we select informative patches with more than 80% nonzero pixels, avoiding black margins in Landsat-8 images. The remaining images are then initially segmented using a threshold value of 120 (picked empirically). Any connected components with area smaller than 10 pixel are discarded. Finally, we apply a morphological dilation using an elliptic structuring element of size 7x7 pixel. The resulting dataset consists of 2,502 images per band. We split the dataset into 2,382 and 120 training and testing images, respectively. To enhance the training set, we augment the images using 3 types of mirroring: horizontal, vertical, and a combination of both. This is consistent with the augmentation approach applied to the solar images. In this work, we refer to this dataset as weak-Cloud-38.

5.3.2 Segmentation Stage Evaluation

Our AR segmentation results were all qualitatively assessed and validated by a solar physics expert. We also visually compare the results against SPOCA and a sequentially fine-tuned U-Net model (similar to the first stage of [43]).

Additionally, to quantitatively demonstrate the benefit of the joint analysis, and due to the lack of manual AR pixel-wise ground-truth, we evaluate our approach using the BraTS-prime

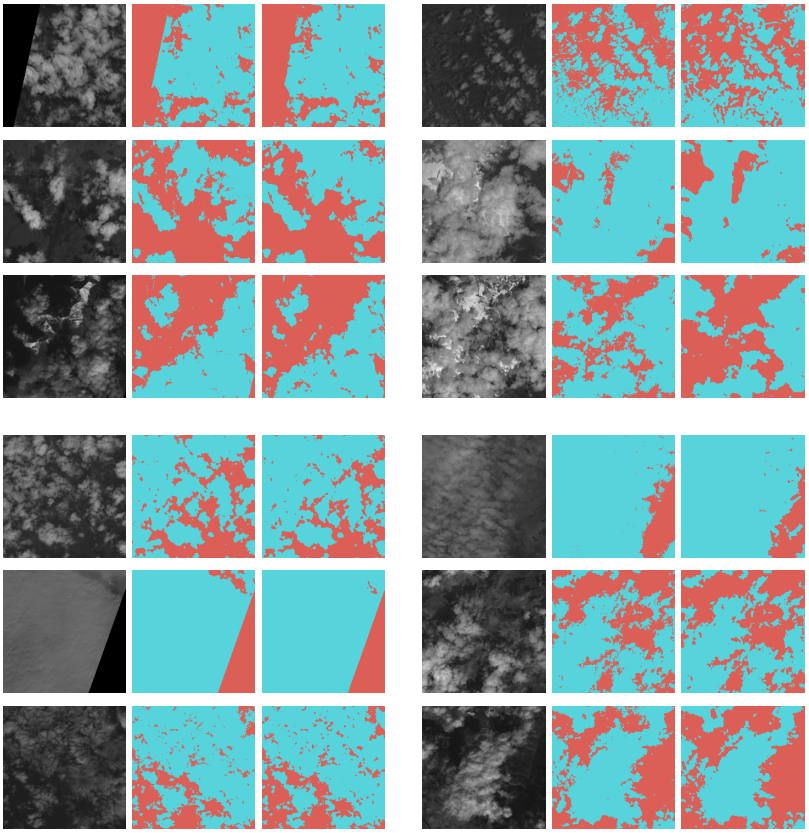

Figure 5.6: Visualisation of MLMT-CNN segmentation results over the Cloud-38 datasets. Each group of images corresponds to a particular image band, Red (top left), Green (top right), Blue (bottom left), and near infrared (bottom right). Each sample (i.e., a single row in a group) consists of, left to right, the input image, ground-truth label, and the predicted mask, respectively. Masks show two main classes, cloud (in blue color) and background (in red color).

synthetic dataset. Weak-Cloud-38 may not be used for this purpose because of its different bands capturing the same scene, rather than different layers of a 3D object. It is worth noting that we do not aim to achieve state-of-the-art performance in tumour segmentation, but rather to confirm the benefit of the joint analysis in scenarios similar to our solar case, where different modalities show different cuts of a 3D object. Since ground-truth is available for this dataset, we follow the classical fully-supervised training procedure. Furthermore, we use Weak-BraTS-prime and Weak-Cloud-38 to evaluate our iterative training strategy from weak labels against full supervision. Its worth noting that the segmentation subnetworks adopt the same layers configuration of their correspondent blocks in U-Net [28].

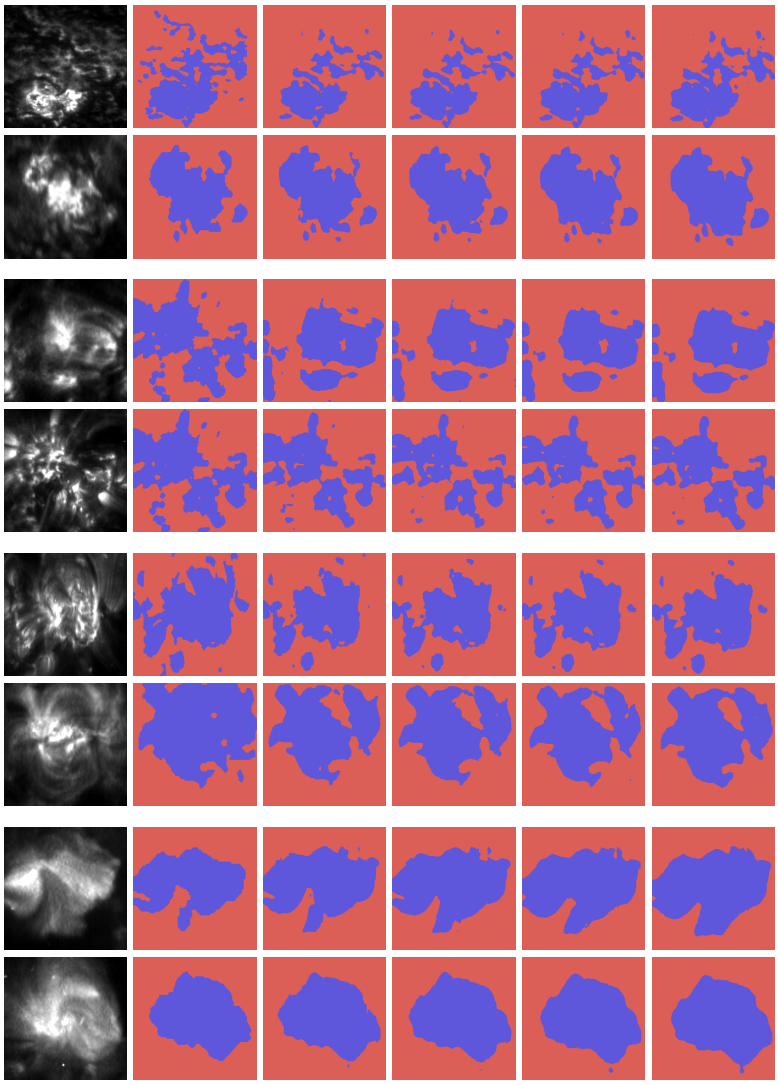

Figure 5.7: MLMT-CNN's recursive segmentation results of solar ARs visualised progressively over different training iterations in randomly selected multi-spectral solar image samples. Columns show, from left to right, input image band, initial weak label, and the predicted masks of recursive training iterations 1, 2, 3, and 4, respectively. Each group of two rows indicates a particular SOHO/EIT imaging band, from top to bottom, 304 Å, 171 Å, 195 Å, and 284 Å.

Table 5.3: Evaluation of weakly-supervised MLMT-CNN (U-Net) on Weak-BraTS-prime. For each class, the highest scores are highlighted in bold.

# train. stages	Bands	IoU score per class			Mean IoU
		NCR/NET	ED	ET	
1	T1Gd	0.67	0.40	0.38	0.48
	T1	0.66	0.41	**0.40**	0.49
	T2	0.62	0.45	0.39	0.49
	Flair	0.64	0.46	0.38	0.49
2	T1Gd	**0.69**	0.43	**0.40**	**0.51**
	T1	**0.69**	0.41	**0.40**	**0.50**
	T2	0.66	0.45	0.38	**0.50**
	Flair	0.67	**0.47**	0.38	**0.51**
3	T1Gd	0.67	0.40	0.37	0.48
	T1	0.67	0.40	0.37	0.48
	T2	0.64	0.42	0.36	0.47
	Flair	0.64	0.45	0.34	0.48

Table 5.4: Comparison of full- and weak-supervision for MLMT-CNN (U-Net) over weak-Cloud-38. For each band, the highest scores of the weakly-supervised models are highlighted in bold.

Super-vision	# train. stages	IoU score per band				Mean IoU
		Red	Green	Blue	NIR	
Fully	NA	0.95	0.95	0.95	0.95	0.95
Weakly	1	0.78	**0.80**	**0.83**	**0.83**	**0.81**
	2	**0.79**	**0.80**	**0.83**	**0.83**	**0.81**
	3	0.78	**0.81**	0.82	**0.83**	**0.81**

5.3.2.1 Independent Segmentation on Single Image Bands

We first compare segmentation results produced by U-Net and FCN8 over the AR and BraTS-prime (Table 5.1) individual image bands, analysed independently, to evaluate different DL-based segmentation networks. These results also serve as baseline to assess our joint analysis based approach in Section 5.3.2.2.

We notice that U-Net produces higher IoU values over all bands for BraTS-prime, as well as smoother AR boundaries, compared to FCN8. This is expected since U-Net utilises skip connections to help retrieving fine details in the mask reconstruction process. Therefore, we use the building blocks of U-Net in our joint segmentation framework.

When comparing the results of U-Net over different modalities, we notice that the T1-Gd modality gets the highest IoU score for the ET class. A similar trend can be seen when comparing the results of the NCR/NET class over different modalities. On the other hand, we find that Flair

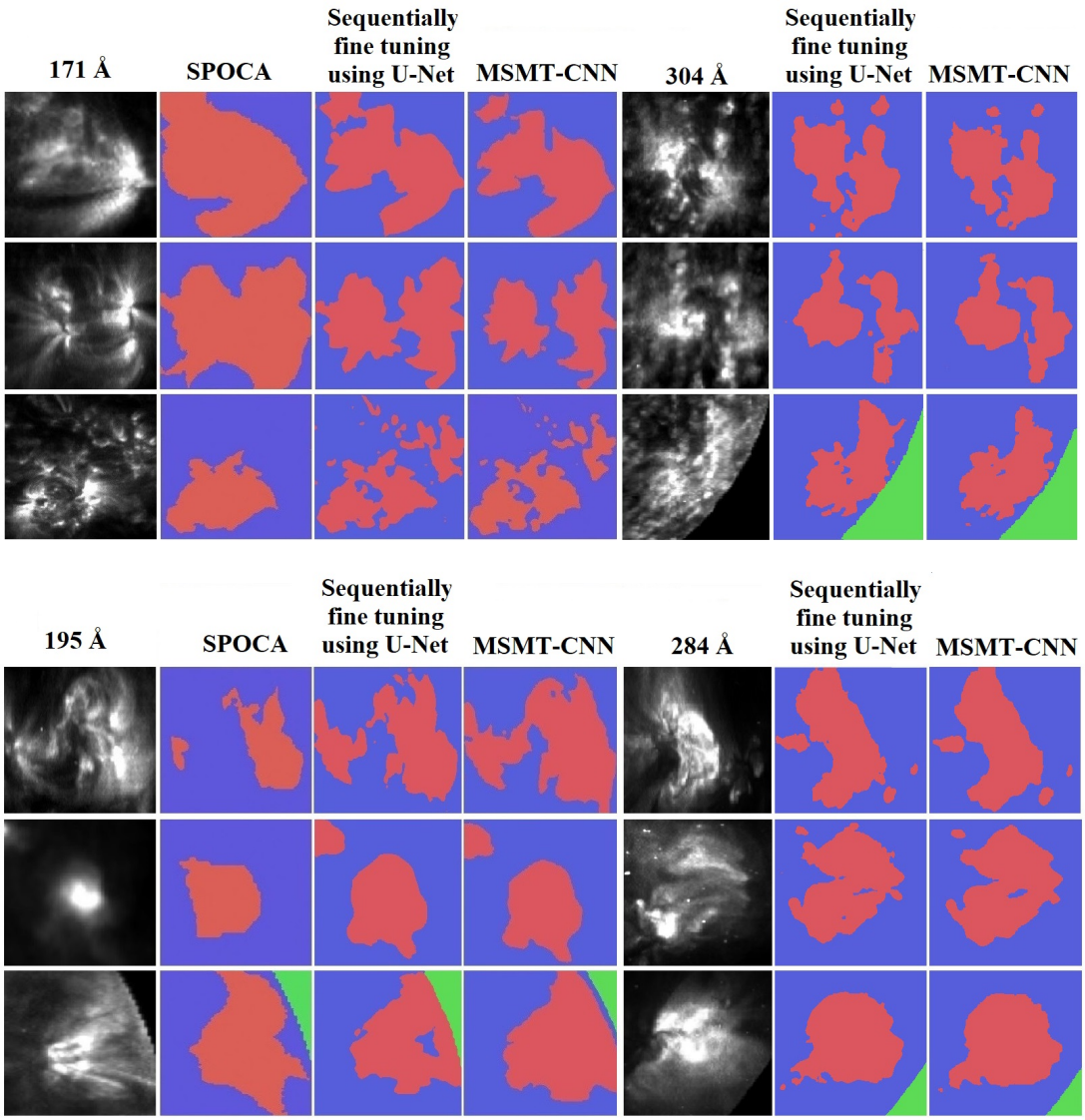

Figure 5.8: AR segmentation comparison between the proposed method, SPOCA, and sequentially fine-tuned DNNs similar to [43], over the SPOCA subset. Red is AR, blue denotes the quite Sun background, and green is outside of the solar disk.

gets the highest IoU for the ED class comparing to the other modalities. This contrast in the IoU scores is in line with the understanding that different modalities provide different information.

5.3.2.2 Joint Segmentation on Multiple Image Bands

Similar to our detection experiment (i.e., Chapter 4), we assess our framework using different combinations of image bands and different types of feature fusion to evaluate their influence on the segmentation performance.

Quantitative results First, we present our BraTS-prime segmentation on combined bands using our joint analysis approach (Table 5.2). We note that all combinations improve on the single-band results, with the best improvement coming from combining all four modalities. All following BraTS-prime experiments use a four-band architecture.

We compared four fusion strategies, namely fusing features after one block of convolution only (early) and at the end of convolutions (late), using addition and concatenation. A visualisation of MLMT-CNN following the early fusion scheme is presented in Fig. 5.2. We find that early fusion with concatenation shows higher results. This differs from our observation in the AR detection experiment, hence confirming that the fusion strategy needs to be adapted to the analysis scenario. Visual results of the proposed early fusion based segmentation scheme is presented in Fig. 5.3. Accordingly, we continue using early fusion with concatenation for all BraTS-prime segmentation experiments. On the other hand, as expected, following the joint analysis early fusion based approach with 4 bands increases the inference time requirement from ~ 0.008 seconds (GPU time) when using U-Net single band based analysis, to ~ 0.029 seconds.

As expected, there is a negative correlation between the IoU score and the width of slice gap, where the overall increase in the IoU was the highest for smaller gaps and higher levels of spatial correspondence (i.e., gap of 1 pixel). This observation, together with the improved results from combining bands, suggest that jointly analysing related multi-modal images in scenarios similar to our solar case may indeed aid the network in learning the inter-dependencies between the different modalities.

We compare against sequentially fine-tuned U-Net models similar to the first stage of [43] in Table 5.2 and Figs. 5.4 and 5.5. They achieved comparable IoU scores to those produced by U-Net on single bands. Hence, they do not benefit from the combination of modalities as our framework does.

Additionally, as a mean to assess our iterative training steps, we use weak-BraTS-prime and weak-Cloud-38 to evaluate this strategy against manual annotations, and compare it to the classical training approach.

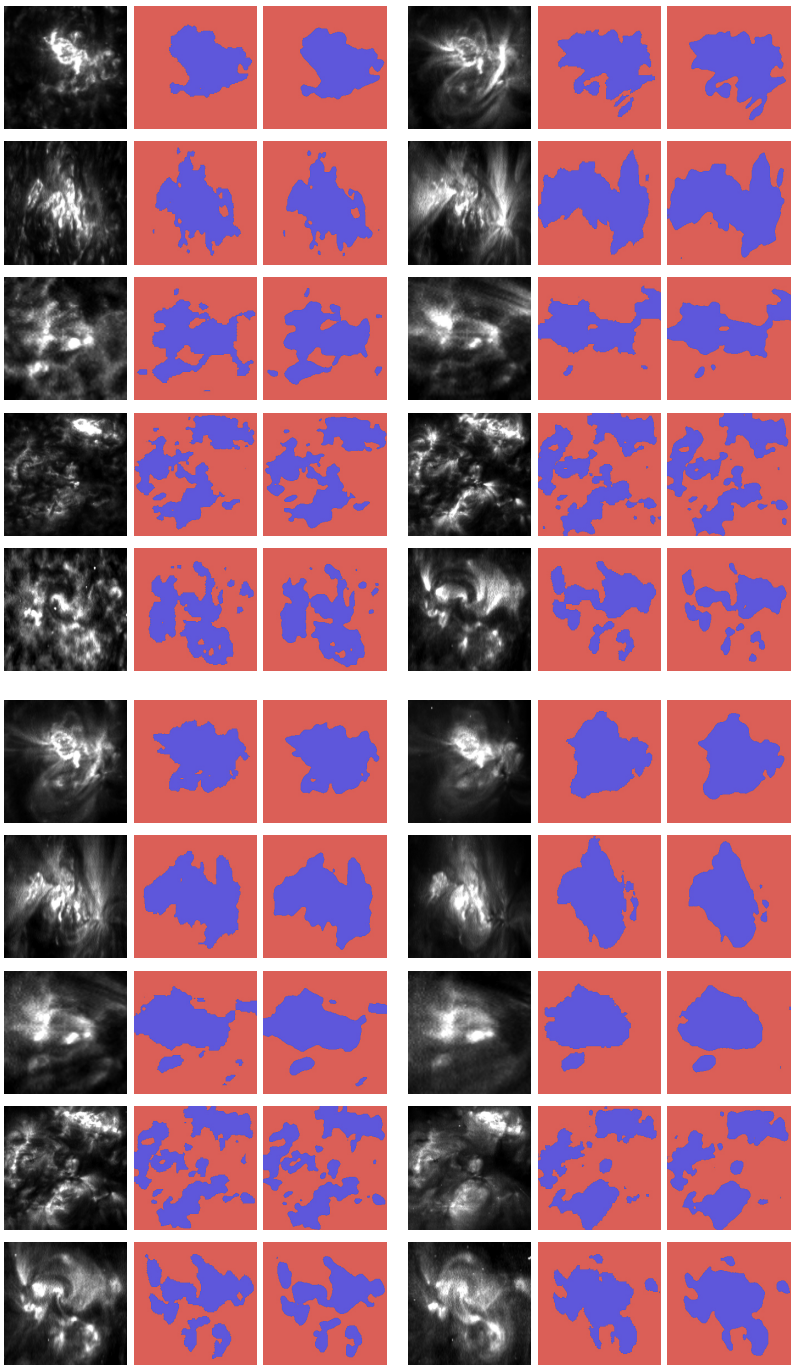

Figure 5.9: Weak label (middle) and MLMT-CNN's (right) segmentation of solar ARs in randomly selected images (left) from SOHO/EIT 304 Å (top left), 171 Å (top right), 195 Å (bottom left), and 284 Å (bottom right).

Table 5.5: Similarity between the proposed architecture and the refined weak labels across different training iterations.

Architecture	# train. stages	# Bands	IoU score per band				Mean IoU
			284 Å	195 Å	171 Å	304 Å	
MLMT-CNN (U-Net)	1	4	0.84	0.81	0.71	0.71	0.77
	2	4	0.93	0.93	0.88	0.88	0.91
	3	4	0.93	**0.94**	**0.91**	**0.89**	**0.92**
	4	4	**0.94**	**0.94**	**0.91**	**0.89**	**0.92**

Table 5.6: Similarity between SPOCA's predictions and our presented architecture.

Dataset	Input	IoU score per band			
		284 Å	195 Å	171 Å	304 Å
SPOCA	Patch	0.41	0.46	0.44	0.33

When evaluating the recursively trained model using weak-BraTS-prime dataset against the fully supervised model on BraTS-prime manual annotations, we notice an increase in the IoU scores after one step of recursion (i.e., 2 stages of training, first using the weak labels, then using the previous predictions as labels), achieving 71% of the fully supervised performance (Table 5.3). Moreover, this iterative training process achieves 85% of the fully supervised approach over the Weak-Cloud-38 dataset, with the best performance also being after one round of recursion, with an increase of 1% over the Red band (Table 5.4). These observations indicate that our recursive training strategy is beneficial in cases where manual annotations are not available, such as solar ARs. Visualisation of MLMT-CNN's segmentation performance over Cloud-38 dataset are presented in Fig. 5.6

Furthermore, in contrast to the single band based segmentation of weak-BraTS-prime (last 4 rows of Table 5.1), we also note that performance still benefits from the joint analysis even when trained – classically or recursively – with weak labels (Table 5.3).

Qualitative results Lastly, in Table 5.5, we evaluate the proposed segmentation approach over the AR class by computing the IoU between the predicted masks and the weak annotations used during training to indicate the agreement between the two. Nevertheless, since weak labels are used to compute the IoU, care must be taken when interpreting the results. Note that the proposed weak label aims to include only pixels that have an evident activity (see Section 5.3.1.1). Generally, we observe a progressive increase in the overall IoU score across the training iterations, with the lowest IoU recorded in the first training stage, in which the initial weak label was

used to train the network, while the highest IoU increase was observed in the second training stage, in which the network was trained using refined annotations produced by the network from the first training iteration. In other words, the highest disagreement was observed between the initial weak label and the predictions from the first training iterations. This is expected, the weak annotations used in the first iterations are designed to only include pixels with evident activity (this may be represent an under-segmentation process), on the other hand, we expect the network to be biased towards learning a generalized representation of solar ARs, such that more AR pixels are included in the predicted mask, causing a low IoU value between the predictions and the weak labels. The following training iterations show comparable performance and less increase in the overall IoU score, this may indicate that the annotation refinement throughout the training iterations is approaching a more stable version. Fig. 5.7 present visual results obtained by different training iterations using the proposed recursive training approach.

We also visually compare our segmentation results on the SPOCA subset, using our proposed architecture, against SPOCA and sequentially fine-tuned DNNs similar to [43] (without their final stage of fusing the CNNs' individual predictions) (Fig. 5.8. The results show that our framework generally finds more detailed solar AR shapes than SPOCA, while being more sensitive to fainter regions of ARs.

Additionally, in Table 5.6, we compare our AR segmentation results to SPOCA by finding the IoU between the predictions produced by the two approaches over the SPOCA subset. This may be used to indicate the agreement between the two methods. We find that both 171 Å and 195 Å achieve a higher agreement of 44% and 46%, respectively, in contrast to 304 Å and 284 Å scoring 33% and 41%, respectively. This is expected since SPOCA was designed to segment solar ARs in 171 Å and 195 Å. Overall, the similarity between our predictions and SPOCA's is relatively low. However, as discussed in our detection experiment (Section 4.3.2.2), SPOCA was manually tuned by the developers according to their own interpretation of AR boundaries which may be different from our interpretation when annotating the dataset. Hence, care must be taken when interpreting the results.

Comparison against sequentially fine-tuned CNNs in the spirit of [43] is fairer, since the DNNs were trained on our data. Segmentation of the sequentially fine-tuned CNNs appears to be of similar quality to ours, although shapes of an AR between neighbouring bands evolve more smoothly with our method. This is an advantage of accounting for the 3D geometry of solar ARs in performing the 2D segmentation. More visual results are presented in Fig. 5.9.

5.4 Summary

In this chapter, we presented a segmentation approach to handle the pixel-wise classification task in multi-spectral images that observe different layers of a 3D object. The proposed method incorporates the joint analysis principles from Chapter 4 to exploit bands-specific and cross-band spatial correspondence and feature-level correlations as well as spatial and semantic information of different levels. Additionally, a recursive training approach based on weak labels (i.e., bounding boxes) was proposed to overcome the difficulty in producing dense AR annotations. The proposed segmentation approach was applied to state-of-the-art backbone CNNs and was evaluated over different applications (brain and cloud segmentation from multi-modal imagery), using different levels of supervision (full- and weak-supervision), and different combinations of input bands.

Our joint analysis approach showed competitive results against both baseline and state-of-the-art segmentation methods, and using different datasets. Inline with our observation in Chapter 4, we find that CNNs can produce enhanced predictions by jointly analysing information from different image bands, at different semantic levels. The proposed approach was also evaluated qualitatively on solar ARs. Results demonstrate that CNNs may show a satisfactory localisation performance when iteratively trained from weak annotations. Both our weak annotations and predictions were validated by a solar physics expert.

Incorporating spatial correspondence and feature-level correlations between different bands has demonstrated effectiveness in within the detection (Chapter 4) and segmentation task of 3D objects in multi-spectral data. In Chapter 6, we explore the possibility of generalising this concept to handling 3D volumetric medical images, where we reformulate the problem as an attention problem in which inter- and cross-channel correlations are explicitly learnt by CNNs. We also explore the possibility of applying the joint analysis into a false alarm reduction task by aggregating information from inputs of different levels of spatial context. In Chapter 7, in addition to cross- and inter-channel attention and correlation analysis, we investigate the influence of incorporating long range correlations and global context on the lung nodule localisation problem and study the possibility of combining attention mechanisms for an improved learning. The work in this chapter has been published in the following journal:

- M. Almahasneh, A. Paiement, X. Xie, J. Aboudarham, MLMT-CNN for object detection and segmentation in multi-layer and multi-spectral images. Machine Vision and Applications, 2021.

Chapter 6

AttentNet for Pulmonary Nodule Detection Using 3D Cross-channel and Inter-spatial Convolutional Attention

Contents

6.1 Introduction

Pulmonary cancer remains the leading cause of cancer related death across genders. With a 5-year-survival rate between 10% to 20%, lung cancer accounted for 18% (1.8 million) of the total cancer deaths worldwide in the year of 2020 [161]. Early detection of pulmonary nodules plays an important role in the success of the treatment, inspecting low-dose computed tomography (CT) scans has proven to be an adequate method for initial diagnoses [161–163]. Pulmonary nodules tend to be isolated and have a spherical shape comparing to the continuous and elongated blood vessels. Doctors rely on these morphological presumption to locate nodules in 3D CT scans. However, considering the increasing amount of CT images that require evaluation, this process becomes time exhaustive, laborious, and prone to human error. In [164], four experienced radiologists were given 1018 CT scans to independently locate suspicious pulmonary lesions, this resulted a set of 2669 different nodules located by at least one radiologist, with 928 locations marked by all four radiologists. The contrast in the radiologists' predictions demonstrates the uncertainty faced when managing lung nodules. Factors such as the experience, the mental and the emotional state of the examiner, can indeed affect the accuracy of the analysis. In fact, diagnostic errors caused by inaccurate radiological reporting are considered of the main causes of patient mortality and permanent injuries [165–168].

Object detection has evolved drastically in line with the recent advances in deep learning, however, typical deep learning object detection methods (e.g., [25, 26, 52–54]) are designed for generic object detection tasks and typically aim at analysing 2D and RGB images. They are not suited to be directly applied to the medical imaging domain (e.g., 3D CT images) due to the different nature of the data and the complexity of the task. The recent introduction of data collections (e.g., LUNA16 dataset [3]) attracted more interest to the pulmonary nodule detection problem. It also made the integration of DL tools possible, however, training DNNs requires huge amounts of labelled samples, which remains a challenge when dealing with pulmonary images. Preparing such datasets is very resource intensive due to the 3D nature of the CT images and the high complexity of the task. Additionally, unlike common computer vision problems, the sparse nature of lung nodules within the lung region, along with the prevailing class imbalance in the pulmonary data, make the detection task more challenging. Therefore, this scenario requires designing a specialised DL approach.

Computer aided detection (CAD) are systems designed specifically to assist radiologists in detecting lung nodules. Incorporating such systems has been clinically proven to decrease

observational oversights (i.e., false negative rate) while significantly reducing the reading time required per scan and retaining a consistent quality [168–170]. Indeed, a number of studies have demonstrated that CAD systems were able to detect nodules that were originally missed by experienced radiologists [168, 171–174]. Generally, CAD systems consists of two consecutive stages, a candidate proposal stage, in which candidate locations are proposed at a high sensitivity and typically on the account of high false positive rates, and a false positive reduction stage to minimise the number of the false alarms and produce the final set of predictions [3].

In [129], a 2D U-Net based candidate proposal stage was deployed along with a CNN classifier that takes three cross-sectional (axial, coronal, and sagittal slices) images as an input. Authors conclude that directly incorporating 3D information may be of potential for an enhanced performance. This was confirmed in [128, 130], where authors compared the impact of exploiting different levels of feature dimensionality (2D, pseudo 3D, and 3D) to perform the detection. Results demonstrate that directly incorporating the 3D aspect can significantly enhance the overall performance, 2D based approaches neglect the inherent 3D nature of the problem. In this line, [132] proposed a two stage 3D detection framework based of Faster RCNN in which an encoder-decoder design was exploited for the region proposal stage. Their approach demonstrated great results in the nodule detection task. Continuing from [132], [135] proposed exploiting grouped convolutions ([136]) and dense convolutional connections ([137]) to improve the performance of the two stage detector. Additionally, [139–141] also demonstrated great potential using a similar encoder-decoder CNN so solve the nodule detection task.

Generally, existing methods investigate different levels of depth, spatial scale and dimensionality of the CNN features. Limited research has been dedicated towards incorporating feature importance and inter-spatial and cross-channel correlations, which can be exploited for more accurate, and effective, detection. This problem is commonly addressed using attention mechanisms. Attention simulates the cognitive process of selective focus on features with high relevance to a task while excluding others. Incorporating such methods can improve CNNs performance by explicitly modelling correlations between the extracted features, in a learnable manner, and therefore focus on important structures associated with a given objective. Note that the term cross-channel here refers to the dependencies between embeddings within a feature channel with respect to embeddings from other feature channels. The term inter-spatial on the other hand refers to the spatial contextual dependencies between the different locations within a feature map.

Indeed, [38] demonstrate using their channel-wise attention method, known as squeeze and

excitation networks (i.e., Eq. 2.1), that by directly learning the correlations between the extracted feature maps (i.e., channels), CNNs are able to yield significant improvements over different tasks and datasets. Similarly, CBAM (i.e., Eqs.2.2 and 2.3) [1] shows that in addition to cross-channel correlations (channel-wise attention), inter-spatial correlations (spatial-wise attention) can also be utilised, either solely or in combination with cross-channel correlations, to enhance the performance of CNNs. In this line, [138] adopted squeeze and excitation paths ([38]) for their pulmonary nodule detection network, they demonstrate an enhanced performance in contrast to the network when no attention is used. However, common attention approaches (e.g., [38] and [1]) are originally designed targeting 2D data, they also rely on heavy dimensionality reduction and expensive multi-layer perceptron networks (see Eqs. [38], 2.2, and 2.3), making them not optimal for handling 3D data. We argue that directly incorporating the 3D aspect of the data to infer attention can yield an enhanced performance.

In this work, we investigate the possibilities of incorporating DL methods to solve the pulmonary detection problem. Specifically, we present a two stage (candidate proposal and false positive reduction) detector based on Faster RCNN [52]. In line with [132,135,138], we adopt a 3D encoder-decoder structure for the backbone network of our candidate proposal stage. Moreover, motivated by the high relevance of nodule morphology and the sparse nature of nodule locations, as well as the success of attention mechanisms in medical imaging related tasks (e.g., detection [105–108,175] and segmentation [110,111,116,176]), we investigate the possibility of explicitly modelling and inferring feature importance and spatial and cross-channel correlations. We also study the possibility of incorporating information from different feature dimensionalities (2D, pseudo 3D, and 3D) simultaneously, in an end-to-end learnable and effective manner. More specifically, we evaluate different state-of-the-art attention mechanisms and propose two 3D fully convolutional –cross-channel and inter-spatial– attention blocks that demonstrate potential within the pulmonary nodule detection task.

Furthermore, we utilise a cross-sectional augmentation approach to battle the limited availability of annotated samples. In line with the objective of the candidate proposal stage, we exploit a testing time augmentation strategy that further enhances the sensitivity of the trained model. To tackle the class imbalance problem present in the lung data, we adopt focal loss [142] and exploit an online hard example mining technique [177].

For our False positive reduction stage, in addition to exploiting attention techniques, and inspired by the variant nodule sizes (See Fig. 6.1), we adapt our joint analysis approach from Chapters 4 and 5, and demonstrate that by aggregating information of multiple levels spatial

context and simultaneously analysing them, the network was able to produce an enhanced performance. This is inline with the findings in [143] (using CNN ensembles) and [105] (using feature aggregation), where they demonstrate that incorporating information of different extents of spatial context can improve the overall detection. Additionally, we propose a zoom-in convolutional path to assist the network in capturing information of different spatial scales in a learnable manner.

Lastly, we propose a modified version of ReLU [178] activation in an attempt to refine it against dying neurons by allowing the generation of small negative outputs for inputs that lie in the flat segment of ReLU. We empirically demonstrate the benefits of our proposed activation within the lung nodule detection task.

More formally, our contributions may be summarised as follows:

1. We present a framework to handle 3D pulmonary nodule detection from CT images. Our framework detects nodules in two stages, candidate proposal and false positive reduction. We evaluate our framework on the publicly available dataset, LUNA16, and demonstrate an outstanding performance against state-of-the-art lung nodule detection methods.

2. We propose two fully convolutional attention blocks in which we incorporate 3D features to infer cross-channel and cross-sectional spatial correlations. We demonstrate the benefits of the proposed attention strategies within the lung nodule detection task. We show that while both attention approaches can enhance the performance of the detection CNN, channel attention shows more significant gains in contrast to spatial attention. We also carry out extensive experiments using different state-of-the-art attention mechanisms and compare their performance against our proposed methods.

3. We deploy a joint analysis based approach that improves the performance of the false positive reduction stage by simultaneously exploiting different levels of spatial contextual information.

4. We present a zoom-in convolutional block that allows the network learning information from different scales and therefore enhance the final prediction.

5. We propose a modification of the ReLU activation function to reduce the risk of the dying ReLU problem and empirically evaluate its influence within the lung nodule detection task.

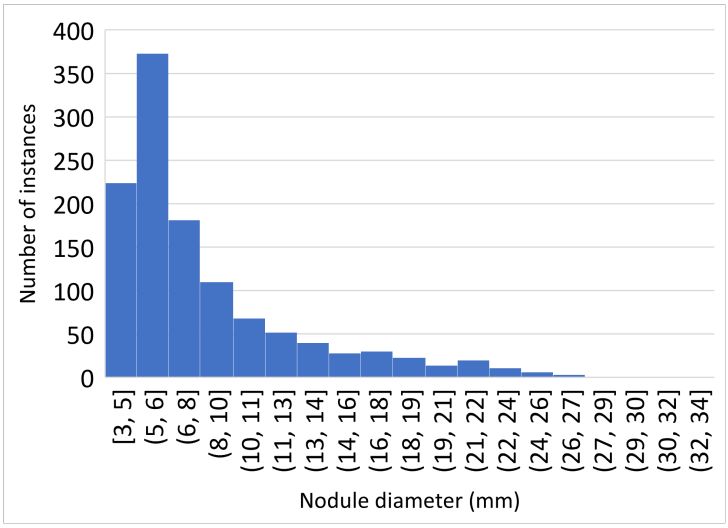

Figure 6.1: Distribution of nodule diameters in LUNA16 [3] dataset. The average nodule diameter is 8.32 mm.

In the remainder of this chapter, Section 6.2 presents the details of our proposed framework, Section 6.3 discusses datasets, pre-processing, experiments and results. Last, Section 6.4 presents a summary of this chapter.

6.2 Proposed Method

Our proposed framework, AttentNet, detects pulmonary nodules in two stages: 1) candidate proposal, in which we adopt a 3D encoder-decoder design to propose suspicious locations of nodules at high sensitivity, 2) and a subsequent false positive reduction stage to reduce the number of false alarms. For the candidate proposal stage, we incorporate fully convolutional attention blocks to efficiently assist the network in focusing on informative features along both, cross-channel and inter-spatial axes. For the false positive reduction stage, we exploit nodule morphology by jointly analysing inputs of different contextual levels. Additionally, we incorporate a zoom-in convolutional path to allow the network pick fine details from different feature scales. During inference time, detections are aggregated from different transformations of the input image. The final classification score is found by ensembling the models of both stages. The two detection stages are optimised separately, each according to its correspondent objective. An overview of our proposed framework is presented in fig. 6.2.

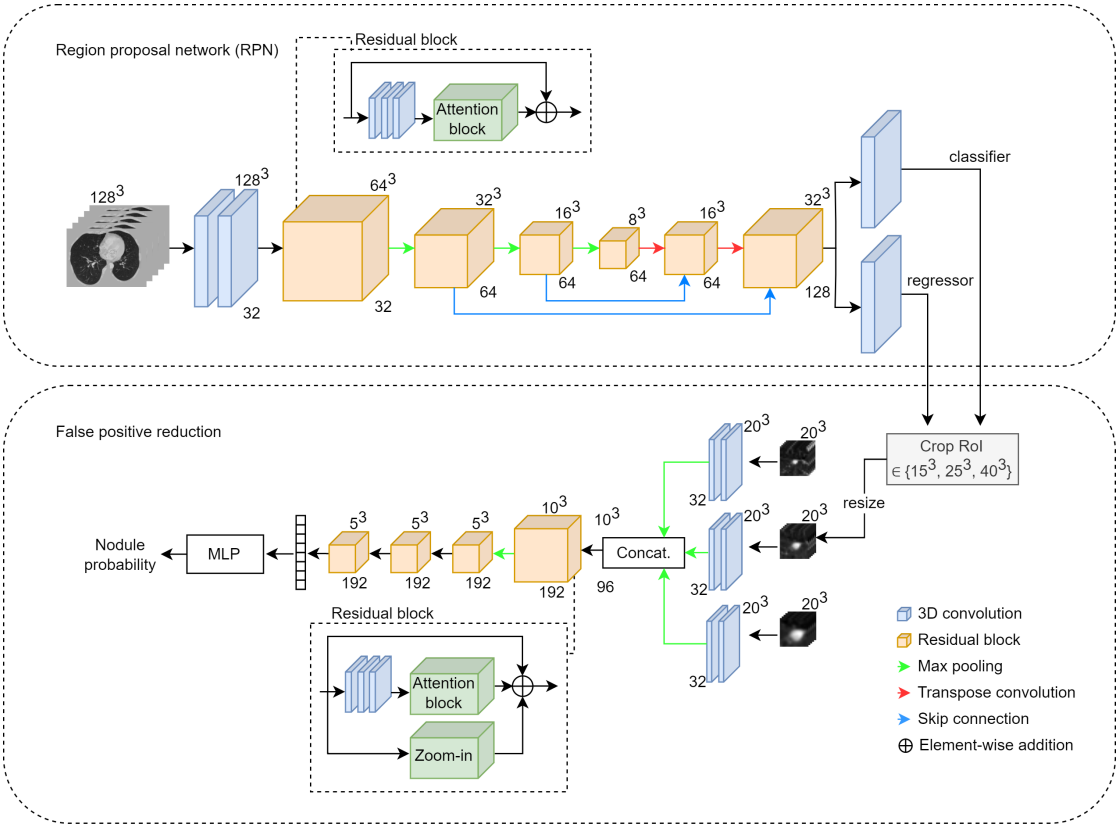

Figure 6.2: The framework of AttentNet. AttentNet performs pulmonary nodule detection in two stages, candidate proposal (i.e., region proposal network), in which we exploit a 3D encoder-decoder network to predict suspicious nodule locations, and a false positive reduction stage in which a 3D CNN is used to extract deep features from the proposed nodules and produce the final prediction. We augment the building blocks of our network with attention units to assist the network in focusing on effective nodule features and therefore produce an enhanced detection.

6.2.1 Candidate proposal stage

We incorporate an encoder-decoder network as the backbone of our candidate proposal stage. An input image is first processed by two consecutive 3D convolutional layers of 32 channels each, using 3 x 3 x 3 kernels, followed by a 3D max pooling layer to reduce the spatial dimension of the resulting feature map by a factor of 2. The resulting volumetric feature is then passed into a sequence of four residual blocks [34] that consist of 2, 3, 3 and 3 residual units, respectively. Each of the residual blocks is followed by a max pooling layer. The resulting embeddings are then up-sampled by two subsequent deconvolutional layers using 2 x 2 x 2 kernels and strides of 2. Each up-sampling layer is followed by a residual block (consisting of 3 residual units each). To avoid overfitting, we regularise the network using 3 dropout layers in the encoder part and

one dropout layer in the last decoder layer. We use 0.3 dropout rate for all dropout layers.

It is worth noting that at an effective image resolution, even a single image can exhaust the GPU memory. To address this constraint and allow processing the 3D inputs at an effective resolution and a representative training batch size, input images are split into 128 x 128 x 128 size patches during training, additionally, we utilise grouped convolutions [136] within the residual units. Grouped convolution allows efficient model parallelism by splitting both, the feature channels, and the convolutional kernels into a number groups. Each group of filters processes their correspondent feature channels and produce a part of the output channels. The resulting feature maps are then projected into a new linear space using a subsequent 1 x 1 x 1 convolution. This allows reducing the computational overhead and the number of parameters while retaining the number of features and a consistent performance [37, 136, 179]. Grouped convolution have repeatedly demonstrated effectiveness within the lung nodule detection task [135,180–182]. Accordingly, a single residual unit consists of three convolutional layers of 1 x 1 x 1, 3 x 3 x 3 , and 1 x 1 x 1 kernel sizes, respectively, with the second layer utilising grouped convolutions of 32 groups [37,136,179]. An overview of our candidate proposal backbone network is presented in Fig. 6.2.

Inspired by the success demonstrated by attention mechanisms in medical imaging related tasks (e.g., [105–108, 175, 176]), as well as the sparsity of pulmonary nodule locations, and the high relevance of nodule morphology for the nodule detection task, we utilise two attention mechanisms within the building blocks of our backbone network to assist the network in focusing at meaningful and effective embeddings. Specifically, we propose 3D fully convolutional cross-channel and spatial attention blocks for the candidate proposal stage.

Given an intermediate 3D convolutional feature map $\mathrm{F} \in \mathbb{R}^{C \mathrm{x} D_F \mathrm{x} H_F \mathrm{x} W_F}$, cross channel attention is computed as follows: First, spatial information is aggregated using adaptive average pooling, resulting an embedding $\mathrm{E} \in \mathbb{R}^{C \mathrm{x} D_E \mathrm{x} H_E \mathrm{x} W_E}$. Here, C represents the channels of the feature map, D_F, H_F, W_F and D_E, H_E, W_E represent the depth, height, and width of F and E, respectively. Note that in our preliminary experiments, we test other types of spatial feature aggregation (e.g., adaptive max pooling) and observe no particular benefit within our nodule detection task, therefore we continue using average pooling. Consequently, E is passed into a convolutional layer with 3D kernels of size D_E x H_E x W_E, and no padding, producing a C x 1 x 1 x 1 descriptor. A sigmoid function is then applied to produce the final channel attention map A_c, in which channel-wise importance scores are predicted. The feature map F is then recalibrated using element-wise multiplication with the attention map A. For our experiment, we set D_E, H_E, W_E

to 3, therefore, the overall attention map computation can be summarised as follows:

$$\begin{aligned} \mathrm{A}_c &= \sigma(\mathrm{W}(\mathrm{E}))^{\mathbb{R}\in C\mathrm{x}1\mathrm{x}1\mathrm{x}1} \\ &= \sigma(Conv3D_{3\mathrm{x}3\mathrm{x}3}(AvgPool3D(\mathrm{F})^{\mathbb{R}\in C\mathrm{x}3\mathrm{x}3\mathrm{x}3})) \end{aligned} \tag{6.1}$$

where *AvgPool3D* represents adaptive 3D average pooling. $Conv3D_{3\mathrm{x}3\mathrm{x}3}$ is a 3D convolutional layer, the subscript represents the size of the kernels used in the convolutional layer. $\sigma(\cdot)$ is a sigmoid activation function. Note that we continue using these notations for the reminder of this chapter. Accordingly, the refined feature map is computed as follows:

$$\mathrm{F}' = \mathrm{A}_c \otimes \mathrm{F} \tag{6.2}$$

where $\otimes$ denotes element-wise multiplication. Our cross-channel attention block is visualised in Fig. 6.3

Our channel-wise attention strategy aims to assist the network in effectively learning to focus on informative feature and ignore (down-weight) irrelevant or less informative ones. Other works that incorporate multi-layer perceptrons rely on heavy dimensionality reduction to decrease the computation overhead of the attention network. Particularly, [1, 38] use an adaptive pooling operation to create a C x 1 x 1 x 1 descriptor in which spatial information of the intermediate features is embedded. This is then passed to a multi-layer perceptron where the feature's dimensionality is further reduced by a pre-defined factor (see Eqs. 2.1 and 2.2). Note that the size of the spatial descriptor is less proportional to the intermediate embedding when dealing with 3D feature maps (e.g., pulmonary nodules) in contrast to that in 2D based analysis (e.g., [1,38]), leading to a limited spatial information. We avoid this by replacing the multi-layer perceptron by a fully convolutional network, allowing an efficient use of rich spatial descriptors with higher dimensionality (i.e., C x 3 x 3 x 3).

To perform spatial attention, we adopt a joint analysis approach that integrates spatial information from different image cross-sections (i.e., axial, coronal, and sagittal). First, an intermediate 3D feature map $\mathrm{F} \in \mathbb{R}^{C\mathrm{x}D_F\mathrm{x}H_F\mathrm{x}W_F}$ is linearly transformed into $\mathrm{E} \in \mathbb{R}^{1\mathrm{x}D_F\mathrm{x}H_F\mathrm{x}W_F}$ using a convolutional layer with kernels of size 1 x 1 x 1:

$$\mathrm{E} = Conv3D_{1\mathrm{x}1\mathrm{x}1}(F^{\mathbb{R}\in C\mathrm{x}D_F\mathrm{x}H_F\mathrm{x}W_F})^{\mathbb{R}\in 1\mathrm{x}D_F\mathrm{x}H_F\mathrm{x}W_F} \tag{6.3}$$

Note that the resulting feature map E is of dimensions 1 x D_F x H_F x W_F, therefore, by reducing the first axis (i.e., D_F x H_F x W_F), E can be processed using 2D convolutions along the D_F axis (this is equivalent to having a 2D input of D_F channels). The feature map E is now projected into

the axial, coronal and sagittal planes. Each of these feature maps is then processed by a unique 2D convolutional network, producing three 2D convolutional feature maps $\{\mathrm{E}_{axi}, \mathrm{E}_{cor}, \mathrm{E}_{sag}\} \in \mathbb{R}^{D_F \mathrm{x} H_F \mathrm{x} W_F}$:

$$\begin{aligned} \mathrm{E}_{axi} &= Conv2D_{3\mathrm{x}3}(Axial(\mathrm{E}))^{\mathbb{R} \in D_F \mathrm{x} H_F \mathrm{x} W_F} \\ \mathrm{E}_{cor} &= Conv2D_{3\mathrm{x}3}(Coronal(\mathrm{E}))^{\mathbb{R} \in D_F \mathrm{x} H_F \mathrm{x} W_F} \\ \mathrm{E}_{sag} &= Conv2D_{3\mathrm{x}3}(Sagittal(\mathrm{E}))^{\mathbb{R} \in D_F \mathrm{x} H_F \mathrm{x} W_F} \end{aligned} \tag{6.4}$$

where $Axial(\cdot)$ $Coronal(\cdot)$ and $Sagittal(\cdot)$ are transformation functions that project inputs from axial plane into coronal and sagittal planes, respectively. This is equivalent to applying 2D convolution with different absolute orientations in the 3D space. The resulting embeddings are then spatially aligned, and are combined (concatenated) along a new axis to form a cross-sectional (3D) feature map $\mathrm{E}_{cs} \in \mathbb{R}^{3 \mathrm{x} D_F \mathrm{x} H_F \mathrm{x} W_F}$:

$$\mathrm{E}_{cs} = Concat(\mathrm{E}_{axi}, \mathrm{E}_{cor}, \mathrm{E}_{sag})^{\mathbb{R} \in 3 \mathrm{x} D_F \mathrm{x} H_F \mathrm{x} W_F} \tag{6.5}$$

Last, E_{cs} is transformed back into $\mathbb{R}^{C \mathrm{x} D_F \mathrm{x} H_F \mathrm{x} W_F}$ using 1 x 1 x 1 convolution. A sigmoid function is then applied to produce the final spatial attention map A_s, in which spatial importance scores are predicted:

$$\mathrm{A}_s = \sigma(Conv3D_{1\mathrm{x}1\mathrm{x}1}(\mathrm{E}_{cs})^{\mathbb{R} \in C_F \mathrm{x} D_F \mathrm{x} H \mathrm{x} W_F}) \tag{6.6}$$

The attention map A_s is used to refine the intermediate feature map using element-wise multiplication:

$$\mathrm{F}' = \mathrm{A}_s \otimes \mathrm{F} \tag{6.7}$$

The overall proposed spatial attention process is visualised in Fig. 6.4

While spatial attention proposed in [1] focuses on capturing spatial correlations within the convolutional features, we argue that the use of element-wise pooling operations along the channel axis, to aggregate spatial information limits the volumentric information when dealing with 3D data. Note that spatial attention in [1] was originally designed for 2D based analysis, see Eq. 2.3. Our cross-sectional spatial attention approach not only leverages 2D inter-spatial relations, but also captures 3D information by jointly analysing the three orthogonal planes of the 3D feature map (axial, coronal, and sagittal). This particularly important when managing volumetric images, where target structures may have different visual appearance when viewed in different cross-sectional planes, e.g., pulmonary nodules. See Fig. 6.5.

Our attention blocks (cross-channel, or spatial attention) can be straightforwardly integrated and jointly trained with any 3D CNN architecture. In line with [1] and [38], we place our

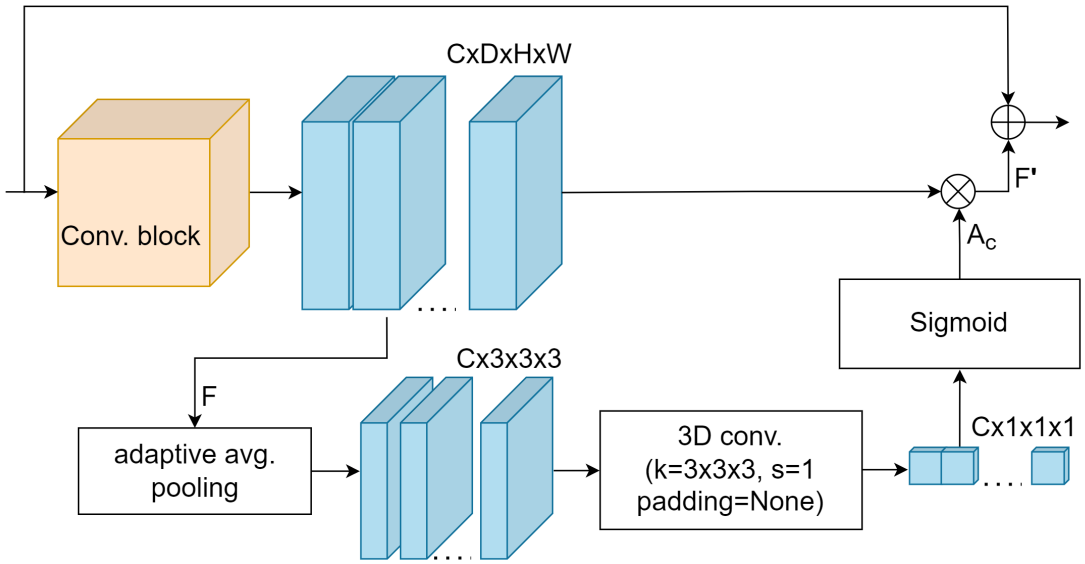

Figure 6.3: An overview of our proposed 3D fully convolutional cross-channel attention unit within a residual block. As illustrated, our channel attention exploits 3D adaptive pooling to embed spatial information from an intermediate convolutional feature map F, these are passed into a 3D convolutional layer in which output is an attention map A_c of size C x 1 x 1 x 1. This will then be used to adaptively refine the intermediate feature maps inferring channel importance and inter-channel correlations (F'). Here, $\otimes$ and $\oplus$ represent element-wise multiplication and addition, respectively. Note that the addition operation represents the residual path in the residual block. The parameter k represents the kernel size used in the convolutional layers.

proposed attention block prior to the residual path of a residual unit. In our experiment, we individually evaluate the benefits of both, channel and spatial attention, within the pulmonary nodule detection task. We further evaluate the impact of incorporating both attention techniques in combination. More details are presented in Section 6.3.

To perform the final detection, the output of the feature extraction network is passed into two parallel convolutional layers, a classification and a regression layer, to predict classes and nodule coordinates for each position in the feature map, respectively. In line with [135,138,141], we train our network with 3 reference anchor sizes, 5, 10, and 20, set based on the nodule size distribution (see Fig. 6.1). An anchor is considered to be positive if it has an intersection over union (IoU) ≥ 0.5, and negative if IoU < 0.2. The network is trained according to the combined loss function:

$$L = \lambda\, L_{cls} \;+\; p^* L_{reg} \tag{6.8}$$

where L_{cls} and L_{reg} are the classification and location regression losses, respectively. λ is a balancing operator and is set to 1 in our experiment. $p^* \in \{1,0\}$ denoting positive and negative

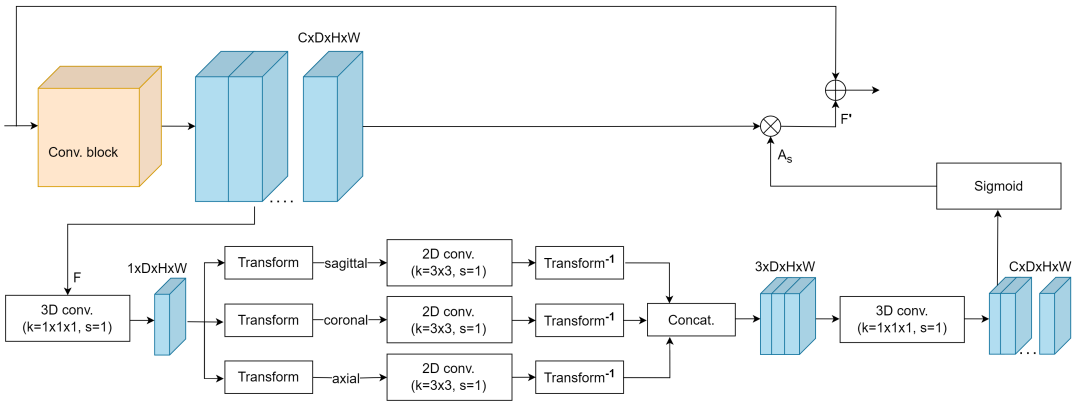

Figure 6.4: An overview of our proposed 3D fully convolutional inter-spatial attention unit within a residual block. Our spatial attention takes as input an intermediate 3D feature map F of C channels, projects it into a 1 channel feature map (using 1 x 1 x 1 convolutions) that is then transformed into three orthogonal planes (axial, coronal, and sagittal). Each of the resulting features is then processed by a unique 2D convolutional layer to learn cross-sectional spatial representations. The resulting feature maps are then spatially aligned, aggregated by concatenation, and are linearly projected back to C 3D channels in which we use to infer cross-sectional spatial attention A_s. Intermediate feature maps are adaptively refined (F') using element-wise multiplication ($\otimes$ in the figure). Here, $\oplus$ represent element-wise addition used in the residual path of the residual block. The parameters k and s represent the kernel size and the stride used in the convolutional layers.

anchors, respectively. Similar to [138, 140], we adopt a focal cross-entropy loss [142] for the nodule classification task:

$$L_{cls}(p_t) = -\alpha(1-p_t)^{\gamma}\log(p_t) \tag{6.9}$$

here, considering an anchor's predicted probability p, $p_t = p$ if the ground-truth label is positive, and $p_t = (1-p)$ otherwise. γ is a modulating parameter. Well classified samples (i.e., $p_t \to 1$) cause the modulating term to approach 0, down-weighting their loss values, and vice versa for hard examples. Note that α acts as a class balancing parameter. We find that setting γ and α to 2 and 0.5, respectively, provides a favorable balanced performance. In line with the essence of attention, focal loss assists the network in focusing on more informative samples and therefore, structures. This is particularly important when managing class imbalanced data, in this case, pulmonary nodule images.

Moreover, we use smooth L1 loss [52] for location regression task :

$$L_{reg}(t^*,t) = \begin{cases} |t^*-t| & \text{if } |t^*-t| > 1 \\ (t^*-t)^2 & \text{otherwise} \end{cases} \tag{6.10}$$

given that t and t^* are vectors representing the relative coordinates of a nodule location in,

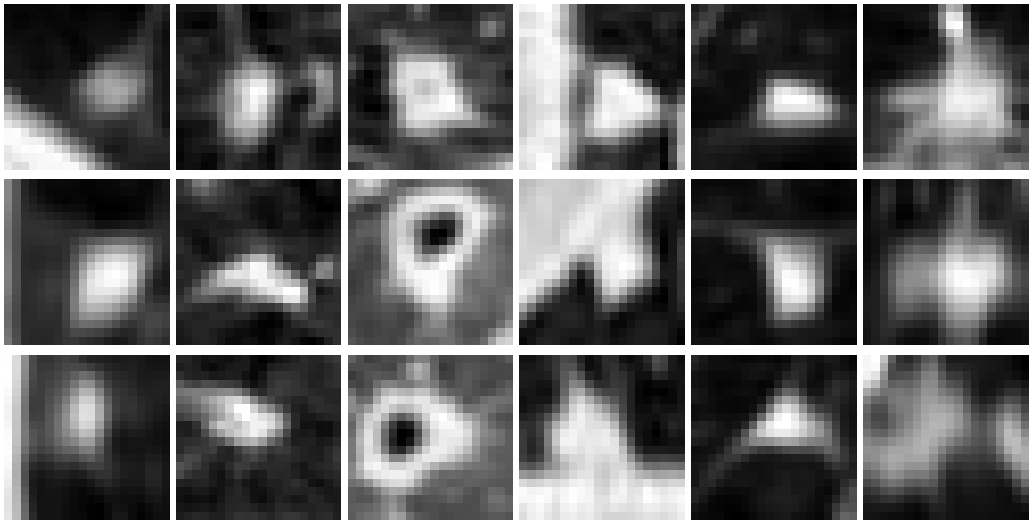

Figure 6.5: Pulmonary nodules viewed in different cross-sectional planes: axial (top), coronal (middle), and sagittal (bottom).

respectively, the prediction space and ground-truth:

$$t_i = (\frac{x-x_a}{d_a}, \frac{y-y_a}{d_a}, \frac{z-z_a}{d_a}, \log\frac{d}{d_a}) \tag{6.11}$$

where (x, y, z, d) and (x_a, y_a, z_a, d_a) are the predicted nodule center and diameter, and the coordinates of an anchor i, respectively. Similarly, the ground-truth relative coordinates are defined as t_i^* using the original nodule coordinates (x^*, y^*, z^*, d^*).

To further reduce the class imbalance impact, we adopt an online hard negative mining strategy [177]. During training, N negative samples with the highest nodule probability are selected to be used in the computation of the loss. Remaining samples are ignored and do not contribute to the loss. Here, we set N to 2. Hard negative mining has repeatedly demonstrated usefulness in the nodule detection task [132, 135, 138–141].

For the network's activation function, we use a modified rectified linear unit (ReLU). ReLU [178] functions have gained popularity due to their simplicity and robustness. This manifests in their ability of preserving the gradient flow in the positive input range. Since the linear portion of ReLU does not saturate (i.e., unbounded), it allows gradients on active neurons to remain proportional to their activation. Moreover, in the negative input range, ReLU promotes network sparsity by setting the activation value to zero [178]. However, this becomes a problem when many inputs have negative values, leading to a degraded gradient flow in the backpropagation process, and therefore, no learning. To avoid this problem, we simply modify ReLU such that it

generates small negative outputs when the input lies in the negative range. Particularly, we incorporate a hyperbolic tangent function for the negative range of ReLU. Given a ReLU function $f(x) = max(0,x)$, our modified ReLU can be formulated as follows:

$$g(x) = \begin{cases} x & \text{if } x > 0 \\ tanh(x) & \text{otherwise} \end{cases} \tag{6.12}$$

where x is the input of the activation function. The idea of this modification is to preserve the linear characteristic of ReLU for positive input values, while improving the gradient flow for negative inputs. Moreover, unlike pure ReLU, the modified ReLU is smooth around the origin (x=0), promoting a faster learning process. A visual comparison of both activations is presented in Fig. 6.6. In Section 6.3, we empirically demonstrate the benefits of our proposed activation and compare it to pure ReLU activation. We also compare the performance of the proposed activation function against commonly used ReLU alternatives, Leaky ReLU [183] and ELU [184], in which the negative gradient flow is enhanced by allowing outputs for input values that lie within the negative range of ReLU. It is worth noting that Leaky ReLU has been widely used within the pulmonary nodule detection task to reduce the risk of the dying neurons [185–189], we demonstrate in our experiment that our modified activation can outperform ReLU, Leaky ReLU, and ELU activations in the nodule detection task.

Furthermore, we employ a testing time augmentation (TTA) strategy in which an input image is orientated along axial, coronal, and sagittal cross-sections. Each of the resulting images is then processed by the network to predict candidate locations. Results from all cross-sections are then aggregated and used to form the final set of predictions It is worth noting that these augmentations are similar to the ones used during training. Testing time augmentation is a simple and an effective way to enhance the performance of DNNs [33, 110, 190–192]. We demonstrate this in our experiment for the nodule detection task.

6.2.2 False positive reduction stage

Pulmonary nodules are highly variable in shape, size, and density, they have similar morphological characteristics to neighbouring organs and non-nodule structures, e.g., blood vessels and airways (see Figs 6.1 and 6.7). This high morphological variability increases the complexity of the detection task leading to high rates of false positive detections [3, 193–198]. To address this issue, we deploy a false positive reduction stage, in which we utilise a joint analysis based approach that incorporates nodule morphology and spatial context information to perform the

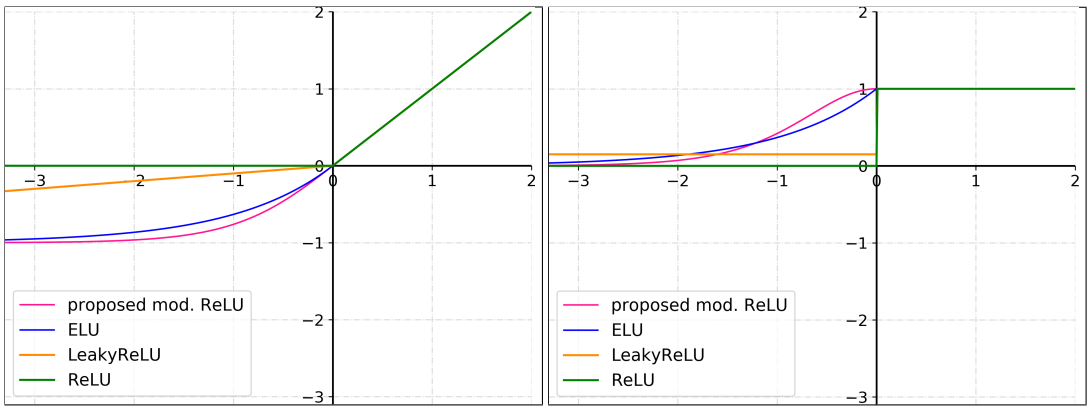

Figure 6.6: Visual comparison of ReLU, Leaky ReLU, ELU, and our proposed modified ReLU activation (left), and their correspondent derivatives (right).

final detection. We further utilise cross-channel attention to assist the network in focusing on important information by modelling correlations between the embedded features. An overview of our false positive reduction stage is presented in Fig. 6.2.

First, a suspected region (i.e., candidate locations proposed by the candidate proposal stage) is extracted in three different scales, such that each patch spans a different level of spatial context (area) around the center of the suspected region. Here, we use small (15 x 15 x 15), medium (25 x 25 x 25), and large (40 x 40 x 40) patches, representing short, medium, and long range spatial context, respectively. Note that these are picked with respect to nodule size distribution. The extracted patches are then resized (individually) into 20 x 20 x 20 using bicubic interpolation, and are fed into three parallel convolutional layers in which each layer is specialised in a particular type of patches (i.e., small, medium, or large). The resulting feature maps are then individually down-sampled using a max pooling operation, aggregated (concatenated), and are jointly analysed by a sequence of four residual blocks [34], in which 2, 3, 3, and 3 residual units are incorporated, respectively. Moreover, the first residual block is followed by subsequent max pooling layer to further reduce the dimensionality of the 3D feature maps. In line with the concept of attention, our design aims at assisting the network in learning contextual information by independently analysing nodules at different extents of spatial information (scales), while modelling correlations between the different contexts by jointly analysing the aggregated feature maps. Fig. 6.8 shows examples of nodule images of different contextual levels and a random selection of their feature maps. We notice that by jointly exploiting inputs of different scales, the network was able to integrate information from multiple levels of spatial context (MLSC).

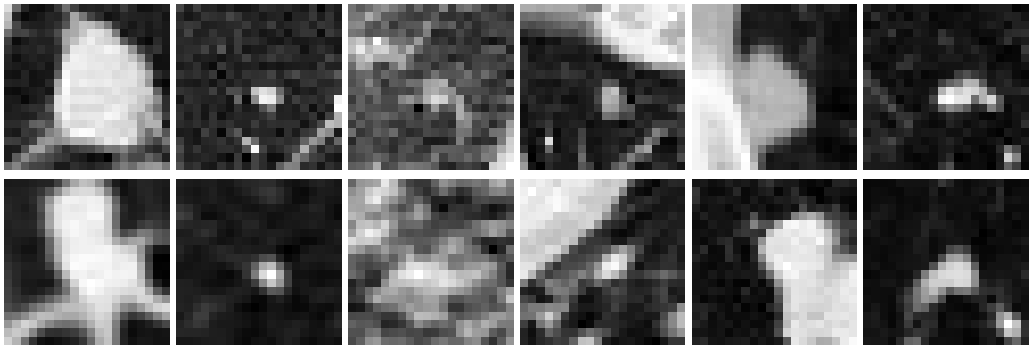

Figure 6.7: A visual comparison between pulmonary nodule (top) and non-nodule (bottom) regions. The two classes share similar appearance and morphology, increasing the complexity of the classification task.

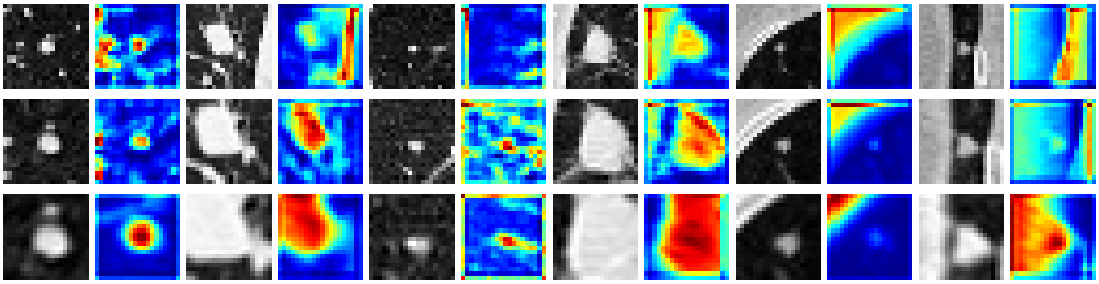

Figure 6.8: Examples of random pulmonary nodules (odd columns) and their correspondent feature maps (even columns) from different levels of spatial context: long range context (top), medium range context (middle), and short range context (bottom). Note that the feature maps are selected randomly from the second convolutional layer in the false positive reduction network. We notice that exploiting information from different input scales (levels of context) assists the network in integrating spatial contextual information of different levels.

In our experiment (Section 6.3), we demonstrate the benefit of the joint analysis and compare it against the performance when single scale inputs are used.

As observed in Fig. 6.1, most nodules are of small size, making the classification task more challenging [193–196, 199]. Moreover, smaller structures are prone to shrinkage due to the use of down-sampling layers, this contributes negatively to the quality of the acquired information, and therefore the accuracy of the detection. Thus, we augment the residual units of our backbone network by zoom-in convolutional paths to assist the network in picking fine details from different feature scales. An intermediate feature map F is first processed by a deconvolutional layer using kernels of size 2 x 2 x 2 and stride of 2, the feature map is therefore up-sampled by a factor of 2. The resulting embedding is then regularised using a batch normalisation layer, followed by a non-linear activation layer, and is finally transformed back into its original di-

Table 6.1: Candidate proposal ablation study: FROC at different numbers of false positives per scan obtained by different methods under comparison on a randomly selected keep-out fold from LUNA16 dataset. Here, CA and SA stand for channel attention and spatial attention, respectively. The highest scores are highlighted in bold.

FROC	Mean	0.125	0.25	0.5	1.0	2.0	4.0	8.0
RPN	0.772	0.576	0.669	0.754	0.790	0.846	0.878	0.890
RPN+CA [38]	0.813	0.638	0.697	0.792	**0.856**	0.877	**0.911**	0.919
RPN+SA+CA [1]	0.743	0.576	0.631	0.687	0.745	0.816	0.862	0.886
RPN+SA [1]	0.719	0.507	0.602	0.688	0.725	0.801	0.843	0.869
RPN+CA [1]	0.779	0.584	0.679	0.753	0.807	0.847	0.884	0.902
RPN+proposed CA+SA	0.782	0.544	0.689	0.765	0.810	0.864	0.886	0.912
RPN+proposed SA	0.784	0.575	0.670	0.750	0.813	0.860	0.900	0.923
RPN+proposed CA	**0.826**	**0.657**	**0.761**	**0.808**	0.834	**0.887**	0.909	**0.929**

mensions using a 3D max pooling operation. The output of the zoom-in layer is incorporated as a skip connection similar to a residual path using element-wise addition. See Fig. 6.9. The intuition behind this strategy is to promote the network into learning to emphasise (magnify) small structures using the up-sampling convolution, such that they are less prone to diminishing due to repeated down-sampling operations and the increasing receptive field. In our experiment (Section 6.3), we demonstrate the effectiveness of our proposed zoom-in path within the false positive reduction task, and evaluate its performance within the candidate proposal task.

Surprisingly, unlike our finding in the candidate proposal stage, we observe no significant improvement when integrating our proposed spatial attention (Eqs. 6.3 to 6.7) within the false positive reduction task. A similar pattern was found when evaluating spatial attention from [1]. Nonetheless, when evaluating channel attention within the false positive reduction task, we observe an enhanced performance using the channel attention approach proposed in [1], while no significant yield was found by integrating our proposed technique (Eqs. 6.1 to 6.2). This may be due to the different complexity of the two tasks. In Section 6.3, we provide an extensive analysis in which we compare different types and combinations of attention techniques. Accordingly, we adopt the cross-channel attention approach from [1] into the building blocks of our false positive reduction network. The over all channel attention process is described in Eq. 2.2. Similar to [1], channel attention is placed prior to the residual path in a residual unit.

The output of the last residual block is then passed into a fully connected notwork of 3 subsequent layers in which the final layer predicts nodule probabilities using a sigmoid function. Similar to the candidate proposal stage, we evaluate focal cross entropy [142] within our false positive reduction task. We observe that setting the modulation parameter γ (Eq. 6.9) to 0 pro-

duces the best overall performance. While using higher values provoke lower false positive rates on the account of lower sensitivity in the nodule class, setting γ to 0 induces a favorable balance in the overall performance. Note that when γ is set to 0, the loss computation is equivalent to using pure cross entropy.

6.2.3 Integration of detection stages

Our framework performs the detection in two stages, candidate proposal and false positive reduction. In line with [3] and [141], we find that ensembling the two stages such that the final prediction is a function of the two models, achieves the best detection performance. First, for each tested image, predictions are then sorted in a descending order with respect to their nodule probability, locations with the highest 300 nodule scores are used as an initial pool of candidates. These are then evaluated against a threshold t, where t is set to 0.3. To eliminate any overlapping predictions, candidates are then processed by a non maximum suppression (NMS) operation with an intersection over union threshold of 0.1.

The remaining nodule candidates are then passed into the false positive reduction stage, where candidate regions are extracted and are processed by the false positive reduction network. For each detection, the final score is defined as the average probability of both stages. Last, detections with probabilities ≥ 0.3 are used to form the final set of detection.

6.3 Experiments

All experiments were implemented using PyTorch DL library with an NVIDIA V100 16GB GPU. Both the candidate proposal and false positive reduction stages were trained (separately) for 250 and 10 epochs ($\sim$2 and $\sim$0.5 days), respectively, using a batch size of 7 and 64, respectively. For both stages, we use Stochastic Gradient Descent (SGD) [200] optimisation with an initial learning rate of 0.01. For the candidate proposal stage, the learning rate is decreased to 0.001, 0.0005, and 0.0001 after 50, 100, and 150 epochs, respectively, while it is conserved at 0.01 for the false positive reduction stage. Network parameters are initialised using He et al. [201].

We train and evaluate both stages following 10-fold cross validation as suggested by LUNA16 [3]. Performance is evaluated using the LUNA16's official metric, Free Receiver Operating Characteristic (FROC) [202], in which sensitivity is computed at 7 predefined false positive rates (i.e., 0.125, 0.25, 0.5, 1, 2, 4, and 8) per scan. Most clinical setups define their

Table 6.2: FROC at different numbers of false positives per scan obtained by the best performing candidate proposal network (proposed RPN with cross-channel attention) on a randomly selected keep-out fold from LUNA16 dataset using different activation functions.

FROC	Mean	0.125	0.25	0.5	1.0	2.0	4.0	8.0
ReLU	0.826	**0.657**	**0.761**	0.808	0.834	0.887	0.909	0.929
Leaky ReLU	0.822	0.602	0.732	**0.826**	0.868	0.897	0.910	0.916
ELU	0.830	0.641	0.743	0.802	0.861	0.895	**0.923**	**0.943**
prop. mod. ReLU	**0.833**	0.642	0.733	0.808	**0.882**	**0.909**	0.922	0.938

effective threshold between 1 and 4 false positives per scan. Including lower false positive rates in the evaluation metric makes the task more challenging [3].

In the remainder of this section, we present data and pre-processing in Sections 6.3.1 and 6.3.2, respectively. Furthermore, to evaluate the contribution of the individual components of our approach, we first perform an ablation study for each of the detection stages (candidate proposal and false positive reduction stage) in Section 6.3.3, then we evaluate the performance of the fully integrated system in Section 6.3.4.

6.3.1 Data

We use the LUNA16[1] (LUng Nodule Analysis 2016) [3] dataset, a subset of LIDC/IDRI dataset [203], to carry out our experiments. The LIDC/IDRI dataset was collected in two stages, a blinded annotation stage, where 4 radiologists were asked to independently mark suspicious locations, and an unblinded stage, where results of all radiologists where anonymised and provided to each of the radiologists to assist them and re-evaluate their initial annotations.

As recommended by [204], [205], and [206], thin slices must be used for pulmonary nodule analysis, therefore, LUNA16 excludes scans with slice thickness > 2.5 mm. Scans with impaired slices or inconsistent slice spacing are also excluded, resulting a total of 888 CT scans. Moreover, only nodules that are ≥ 3 mm and are accepted by a minimum of 3 out 4 radiologists are included in LUNA16, summing up to 1186 nodule labels. Following the lung cancer screening protocols in [207], any remaining annotations are flagged as irrelevant findings.

Furthermore, for the false positive reduction task, LUNA16 provides 754,975 candidate locations acquired using multiple existing methods ([208–212]). These include 1,166 nodules that match the radiologists annotations and 753,809 of non-nodule locations.

[1] available at https://luna16.grand-challenge.org/

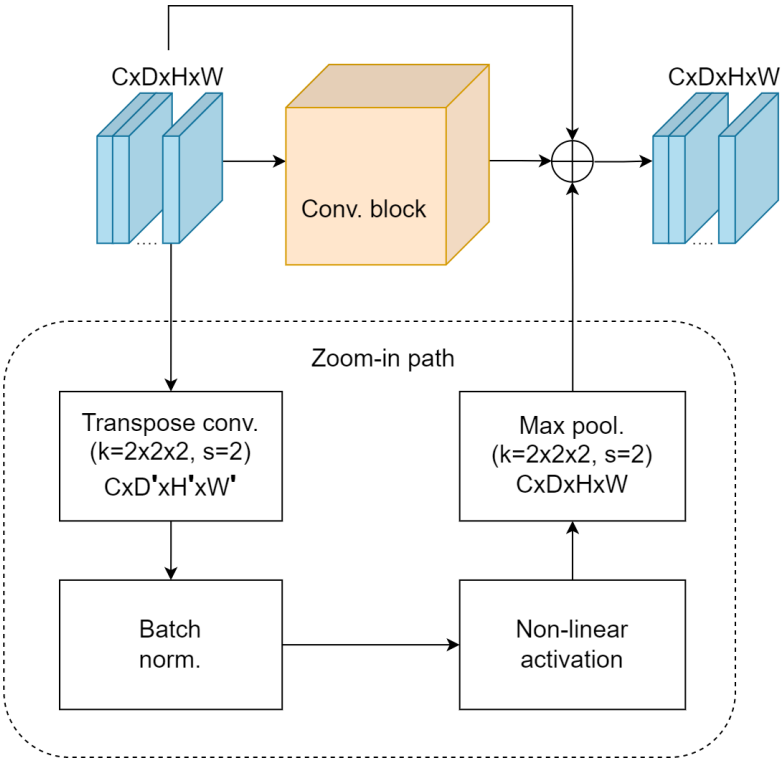

Figure 6.9: An overview of our proposed 3D fully convolutional zoom-in path within a residual block. Our zoom-in path takes intermediate feature maps of size C x D x H x W as an input. These features are up-sampled by a transposed convolutional operation resulting a feature map of size C x D' x H' x W' to assist the network in capturing finer details from different spatial scales. The extracted features are then normalised using a batch normalisation layer, mapped into a non-linear space, and are pooled using a max pooling operator in which the resulting embeddings are of the same dimensions as the original input. These are then aggregated along with the features from the residual block using an element-wise addition process ($\oplus$ in the figure). The parameters k and s represent the kernel size and the stride used in the convolutional and the max pooling layers.

6.3.2 Pre-processing

Similar to [138] and [135], we pre-process LUNA16 images by clipping their intensity values between -1200 and 600 Hounsfield units (HU) followed by rescaling them between 0 and 1. We use the lung masks provided by LUNA16 to isolate the informative lung region and mask out any surrounding organs. Due to GPU limitations, for the candidate proposal stage, images were split into patches of size 128 x 128 x 128 during the training stage. We follow the same approach for the testing stage allowing an overlap of 32 pixels between the cropped patches.

Additionally, we follow a cross-sectional augmentation strategy, where scans, along with their annotations, are transformed from axial plane (original plane), into sagittal and coronal

Table 6.3: False positive reduction ablation study: FROC at different numbers of false positives per scan obtained by different methods under comparison on a randomly selected keep-out fold from LUNA16 dataset. The top section of the table compares results of the false positive reduction network when incorporating spatial information from different contextual levels. The bottom section presents a comparison of different attention mechanisms. Here, CA and SA stand for channel attention and spatial attention, respectively. The highest scores are highlighted in bold.

FROC	Mean	0.125	0.25	0.5	1.0	2.0	4.0	8.0
Short range spatial context	0.623	0.207	0.354	0.538	0.681	0.763	0.877	0.936
Medium range spatial context	0.769	0.503	0.646	0.738	0.822	0.861	0.902	0.910
Long range spatial context	0.715	0.487	0.578	0.681	0.740	0.814	0.854	0.855
Multi-level spatial context (MLSC)	0.792	0.529	0.626	0.738	**0.831**	**0.915**	**0.946**	**0.963**
MLSC+zoom-in (MLSC-Z)	**0.813**	**0.661**	**0.730**	**0.783**	0.822	0.846	0.912	0.938
MLSC-Z+proposed CA+SA	0.678	0.455	0.500	0.575	0.715	0.792	0.844	0.866
MLSC-Z+proposed SA	0.805	0.564	0.661	0.790	0.828	0.881	0.939	**0.972**
MLSC-Z+proposed CA	0.775	0.564	0.640	0.740	0.790	0.849	0.899	0.946
MLSC-Z+CA+SA [1]	0.740	0.433	0.566	0.713	0.801	0.849	0.907	0.909
MLSC-Z+SA [1]	0.778	0.505	0.634	0.744	0.821	0.903	0.919	0.919
MLSC-Z+CA [38]	0.833	0.593	0.711	0.812	**0.883**	**0.926**	**0.950**	0.957
MLSC-Z+CA [1]	**0.848**	**0.702**	**0.745**	**0.815**	0.862	0.906	0.940	0.963

planes.

For the false positive reduction stage, patches are extracted using annotation coordinates provided by LUNA16 for the false positive reduction task. Each patch is cropped in 3 different sizes (levels of spatial context), 15 x 15 x 15, 25 x 25 x 25, and 40 x 40 x 40. This ensures the coverage of 99% of the nodules [213]. All patches are then resized to 20 x 20 x 20 using bicubic interpolation. Furthermore, during training, we augment the nodule class using cross-sectional augmentations, followed by 1-pixel shifts along the z, y, and x axes.

6.3.3 Ablation study

6.3.3.1 Candidate proposal stage

For the candidate proposal task, we compare the impact of different attention mechanisms against the performance of our backbone CNN when no attention is used. Results are presented in Table 6.1 Fig. 6.10. Performing 10-folds cross validation requires extensive amounts of time (e.g., $\sim$2 days per fold), therefore, for the purpose of evaluating the candidate proposal stage, training and testing were performed using a randomly selected keep-out fold. Nonetheless, in Section 6.3.4, we evaluate our fully integrated system using 10-folds cross validation as suggested in LUNA16 [3]. Note that all our experiments are trained using ReLU activation function unless stated otherwise.

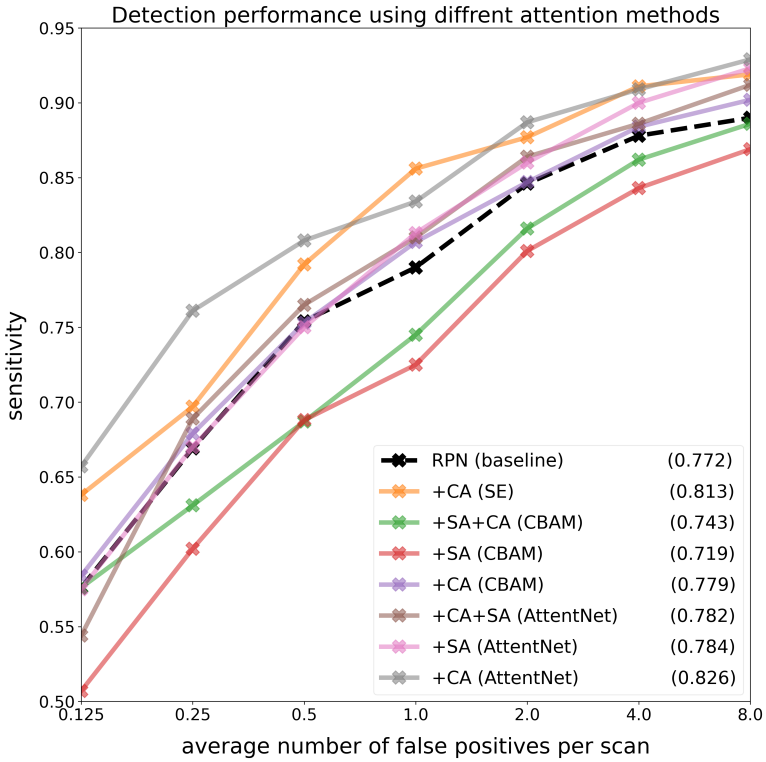

Figure 6.10: FROC of AttentNet using different attention mechanisms under comparison on a randomly selected keep-out fold from LUNA16 dataset. Legend indicates average FROC score across all false positive thresholds for each tested detector.

In our first experiment, we evaluate the performance of our candidate proposal network when no attention is incorporated. This will serve as a baseline for our candidate proposal stage. We find that the network has successfully detected lung nodules at a recall (sensitivity) of 0.878 at 4 false positives per scan, and an FROC score (i.e., average sensitivity over 7 false positive thresholds, 0.125, 0.25, 0.5, 1.0, 2.0, 4.0, 8.0) of 0.772, prior to any false positive reduction. Note that most clinical setups define their effective threshold between 1 and 4 false positives per scan [3]. This demonstrates a compelling potential of CNNs in the pulmonary nodule detection task.

To evaluate the impact of different attention mechanisms, we compare cross-channel and spatial-wise attention when applied individually, and when applied in combination (i.e., subsequently) within the building blocks of the detection network.

We find that all channel attention approaches (i.e., the proposed approach and the ones from [1] and [38]) contribute positively to the overall detection performance, with our proposed

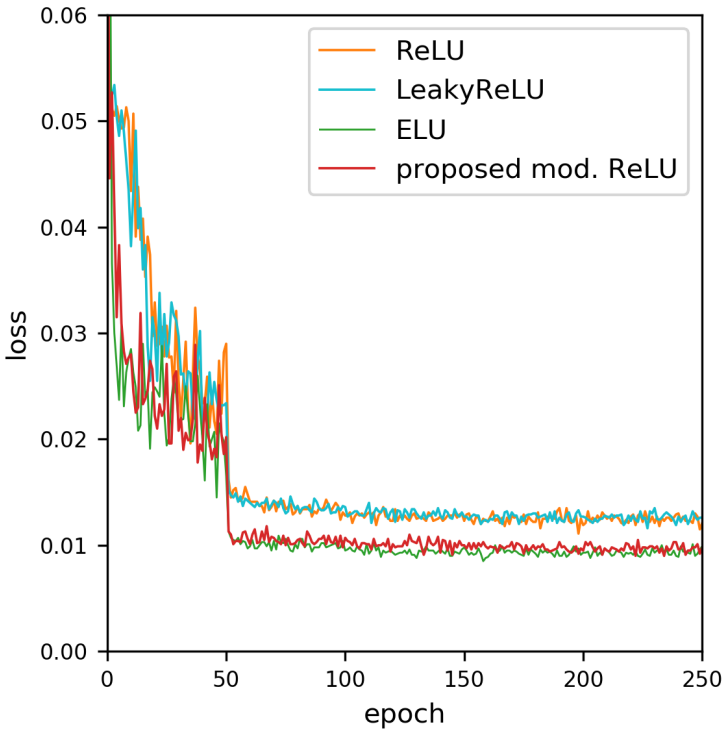

Figure 6.11: Validation loss of the proposed lung nodule detection network using different activation functions. Our proposed modified ReLU activation is smooth around the origin, promoting faster learning in contrast to both, pure ReLU and leaky ReLU activations.

strategy producing the highest FROC score of 0.826, in contrast to 0.779 and 0.813 using the attention approaches from [1] and [38], respectively. This demonstrates the advantage of our fully convolutional design, by replacing the MLP units (i.e., channel attention in [1] and [38]) with convolutional operations, we avoid the need for heavy dimesionality reduction (see Section 6.2.1), and attention can therefore be efficiently performed using spatial embeddings of higher dimensions, leading to an enhanced performance. As demonstrated by the results, this is particularly useful when managing 3D data of high spatial resolution.

Moreover, when evaluating the impact of spatial attention, we find that our proposed spatial attention strategy was able to successfully enhance the network's performance with an FROC score of 0.784, showing an increase of 1.2% comparing to the baseline network when no attention is used. On the other hand, we find that spatial approach from [1] leads to a worse performance comparing to the baseline network. This shows that, in contrast to [1], where spatial descriptors are aggregated using element-wise pooling operations along the channel axis,

Table 6.4: FROC at different numbers of false positives per scan obtained by our fully integrated two stage pulmonary nodule detection system, AttentNet, in contrast to baseline methods using 10-folds cross validation on LUNA16 dataset. Here, TTA indicates the use of testing time augmentation. The highest scores are highlighted in bold.

FROC	Mean	0.125	0.25	0.5	1.0	2.0	4.0	8.0
ZNET [129]	0.811	0.661	0.724	0.779	0.831	0.872	0.892	0.915
3D RCNN [132]	0.834	0.662	0.746	0.815	0.864	0.902	0.918	0.932
DeepLung [135]	0.842	0.692	0.769	0.824	0.865	0.893	0.917	0.933
DeepSeed [138]	0.862	0.739	0.803	0.858	0.888	0.907	0.916	0.920
AttentNet (RPN)	0.842	0.656	0.774	0.831	0.874	0.903	0.923	0.936
AttentNet	0.871	**0.752**	**0.817**	**0.857**	0.885	**0.920**	0.933	0.933
AttentNet+TTA	**0.874**	0.748	0.812	0.856	**0.893**	0.919	**0.942**	**0.945**

neglecting the volumetric aspect of the data, the incorporation of cross-sectional spatial information in our attention strategy can successfully assist the network in learning correlations between the different image cross-sections, and consequently capturing important 3D information within the images.

Generally, while both our spatial and cross-channel attention demonstrate a positive impact within the pulmonary nodule detection task, we observe that the network yields a better performance using channel-wise attention in contrast spatial attention. A similar trend is observed when evaluating channel and spatial attention from [1] and [38]. Moreover, when combining both our channel and spatial attention by applying them in subsequently (as suggested in [1]), we notice an increase in the FROC score comparing to the baseline network when no attention is used, however, we find that the network benefits the most when cross-channel and spatial attention are incorporated individually, with channel attention being the best performing amongst all configurations. In terms of inference time, in contrast to the RPN baseline with no attention, requiring ~ 0.082 seconds (GPU time) per sample (i.e., 128 x 128 x 128 image patch. See Section 6.3.2.), the proposed channel and spatial attention are slower recording an inference time of ~ 0.100 and ~ 0.123, respectively. This is expected since aggregating the network with attention paths increases the computational complexity. Furthermore, when evaluating the performance attention methods from [1], we find that their channel attention produces better results in contrast to both, their spatial attention, and spatial and channel attention when applied in combination. These observations indicate a high importance of inter-channel dependencies in the pulmonary nodule detection task. In fact, we argue that channel attention not only assists the network in modelling inter-channel correlations, but also inherently infers spatial attention

by assisting the network in focusing on important feature maps (channels) in which informative spatial features are embedded. Accordingly, and due to the improved performance demonstrated in our experiment, we continue using our proposed channel attention as the method of choice for the candidate proposal task.

In our preliminary experiments, we evaluate our zoom-in path within the candidate proposal stage. We observe comparable performance to the network when no zoom-in paths are used. This is expected since the candidate proposal network benefits from an encoder-decoder design in which spatial features are extracted from multiple spatial scales. However, the zoom-in path demonstrates benefits when incorporated within the false positive reduction task, where an encoder-decoder design is not straightforwardly applicable due to the relatively narrow input dimensions (details in Section 6.3.3.2).

Finally, we evaluate the impact of our proposed modified ReLU activation function within the nodule detection task by applying it to the best performing network configuration amongst all tested detectors (i.e., RPN with the proposed cross-channel attention), and compare its performance to the network with pure ReLU, Leaky ReLU, and ELU activations. Results are presented in Table 6.2. We find that the modified ReLU activation produces an increased FROC score of 0.833 comparing to 0.826, 0.822, and 0.830 when using pure ReLU, Leaky ReLU, or ELU activations, respectively. More particularly, when comparing to pure ReLU and Leaky ReLU, a significant increase is observed in the sensitivity at false positive per scan thresholds from 1.0 to 8.0, in which an effective clinical threshold (i.e., 1.0 - 4.0) is defined within [3]. In the same line, the proposed modified ReLU activation achieves the highest sensitivity at 1 and 2 false positives per scan. ELU on the other hand obtains higher sensitivity at 4 and 8 false positives per scan, and a comparable average FROC score to the proposed activation. This is expected since ELU and the proposed activation have similar characteristics over the negative segment of the activation function. The performance gain using the proposed activation may be explained by the smoother gradient flow in contrast to that in ELU as demonstrated in Fig. 6.6. Unlike unbounded activations (e.g., Leaky ReLU), our modified ReLU activation, as well as ELU activation, are bounded for negative inputs, promoting network regularisation and reducing the risk of overfitting. These observations demonstrate that allowing small outputs for inputs in the negative range of ReLU can improve the gradient flow and therefore enhances the learning process, see Fig. 6.11. Accordingly, we adopt this modified ReLU activation for our detection network and continue using it for the remainder of our experiment.

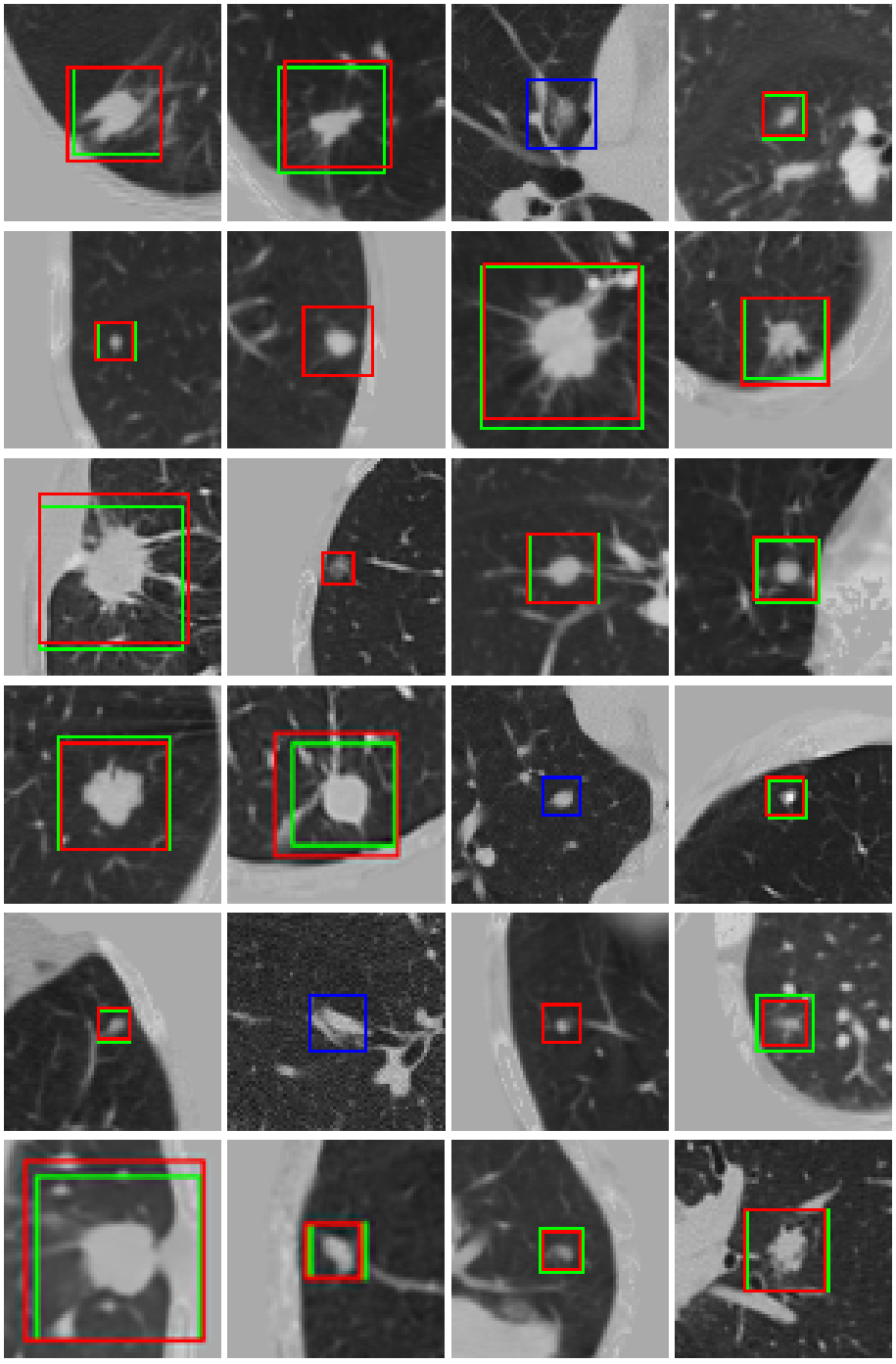

Figure 6.12: Pulmonary nodules detected by AttentNet (green) against ground-truth boxes (red). Blue boxes indicate false positive detections. Red boxes with no overlapping detections indicate false negative samples.

6.3.3.2 False positive reduction stage

For the false positive reduction task, we first evaluate the impact of using different input scales (levels of spatial context) on the detection, and compare it against our proposed joint analysis approach in which we incorporate inputs of multi-level spatial context (MLSC). We then evaluate the performance of our zoom-in path, as well as different attention mechanisms, against the performance of our backbone CNN when no attention is used. Results are presented in Table 6.3. Similar to the candidate proposal stage, we evaluate our false positive reduction network using a randomly selected keep-out fold.

When comparing different input scales for the false positive reduction network, we find that medium range spatial context (25 x 25 x 25) produces the best performance in contrast to short (15 x 15 x 15) and long (40 x 40 x 40) range spatial context. However, we find that jointly analysing Multi-level Spatial Context (MLSC) inputs enhances the overall detection comparing to using any of input scales individually, with an FROC score of 0.792 comparing 0.769 produced using medium range spatial context inputs. This demonstrates the importance incorporating spatial information from different levels when managing objects with the high size variability such as the pulmonary nodules (see Fig. 6.1). Furthermore, Fig. 6.8 shows that our MLSC network was able to integrate spatial contextual information of different levels using our joint analysis approach. We also find that augmenting our Multi-level Spatial Context with Zoom-in paths (MLSC-Z) enhances the performance even further, with an FROC score of 0.813, showing an increase of 2.1% comparing to the performance when the zoom-in path is not used. Note that unlike the candidate proposal stage where the network benefits from an encoder-decoder design in which feature maps are extracted from multiple spatial scales, due to the nature of the false positive reduction task and the relatively narrow input dimensions, an encoder-decoder approach is not straightforwardly applicable. Thus, we design and incorporate our zoom-in path within the building blocks of the false positive reduction network.

When comparing different attention approaches, similar to our observation in the candidate proposal stage (Section 6.3.3.1), channel attention demonstrates the best performance comparing to spatial attention. Indeed, the essence of cross-channel attention is in line with the concept of our joint analysis approach, in which we aim to assist the network in capturing correlations between inputs of different levels of spatial context, this suggests that both, channel attention and the joint analysis approach provide complementary information to one another. Furthermore, contrary to our finding in the candidate proposal task, we notice that channel attention approaches form [1] and [38] show better performance comparing to our proposed channel at-

tention within the false positive reduction task, with the channel attention from [1] producing the highest results, showing an increase of 3.5% in the overall FROC score in contrast to the performance when no attention is used. This may be due the different complexity between the two tasks, this may indicate that the MLP based attention (e.g., [1] and [38]) is more suitable for the false positive reduction task in contrast to our fully convolutional channel-wise attention strategy. Accordingly, we continue using channel attention from [1] as our approach of choice in the false positive reduction stage. On the other hand, we find that our spatial attention produces the highest sensitivity at 8 false positives per scan, and a higher overall performance in contrast to the spatial attention approach from [1]. This demonstrates the benefits of incorporating cross-sectional spatial information in our spatial attention comparing to the channel-wise pooling approach used in [1].

Overall, our results demonstrate the importance of exploiting morphological information when dealing with pulmonary nodules. Combining nodule morphology with attention mechanisms further enhances the network ability in learning effective embeddings and consequently produce more accurate predictions.

6.3.4 Integrated system performance

In this section, we evaluate the performance of our proposed nodule detection system by integrating both detection stages (candidate proposal and false positive reduction stage) in an ensemble model. Note that all experiments in this section are done following 10-folds cross validation as suggested in LUNA16 [3]. Results are presented in Table 6.4.

First, we evaluate the performance of our candidate proposal network (RPN) when no false positive reduction stage is incorporated. We find that that our network produces a better overall performance in contrast to the two stage detector [129] and higher or comparable results to the single stage detectors, [132], [135] and [138]. Additionally, when comparing our network to baselines at false positives per scan ≥ 1, in which an effective clinical threshold is defined [3], we find that our network outperforms [129], [132], and [135] at 1.0, 2.0, 4.0 and 8.0 false positives per scan, as well as [138] at 4.0 and 8.0 false positives per scan with a comparable performance at 1.0 and 2.0 false positives per scan. Note that our proposed network contains 3.1M trainable parameters, in contrast to 17M, 5.4M, 1.4M, and 5.4M parameters in [129], [132], [135], and [138], respectively, making it one of the most compact networks amongst all methods under comparison. This demonstrates the positive impact of our proposed 3D attention strategy, by enabling the network in focusing on important features, our network was able to efficiently

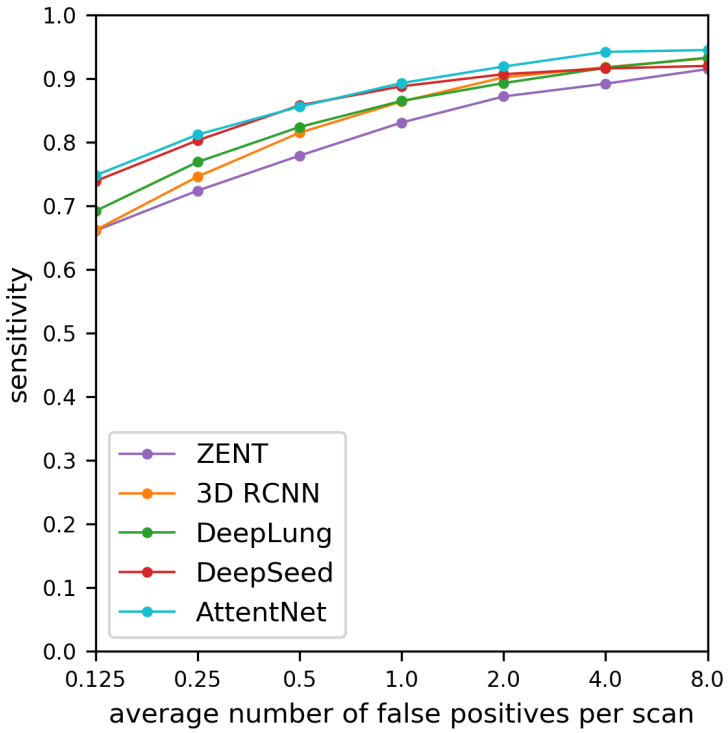

Figure 6.13: FROC of all systems under comparison using 10-folds cross-validation over LUNA16 dataset. Our proposed system produces the highest FROC score (average sensitivity over all false positive thresholds) of 0.874 with a total sensitivity of 95%.

predict nodule locations with an enhanced performance, prior to any false positive reduction.

Furthermore, we find that by integrating both detection stages as an ensemble, our network was able to achieve the highest performance score amongst all methods under comparison. With an FROC score of 0.871, showing a significant increase of 2.9% in contrast to our network prior to the false positive reduction step, and 0.9% comparing to [138], the second highest nodule detector. This indicates the importance of the false positive reduction step, particularly when managing a critical task such as pulmonary nodule detection. By analysing suspected nodule regions and extracting deeper semantic features, the network was able was able to produce an increased performance.

Moreover, to evaluate the impact of testing time augmentation (TTA), we compare the performance of our network with TTA against the network when no TTA is used. We find that TTA produces an enhanced FROC score of 0.874, showing an increase of 0.3% in contrast to the network when no TTA is used, being the best performing detector amongst all network ap-

proaches. This is in line with the finding in [33, 110, 190–192], in which TTA was demonstrated to be a simple and effective way to boost the performance of DL models. This is in line with the essence of cross-feature correlation analysis. It indicates that even the simple aggregation of cross-feature information at the decision level may lead to boosts in the overall detection performance, in this case, the aggregation of locations predicted using input transformations that are spatially correlated, of which the inference model was trained with, was useful. Accordingly, we continue using TTA as a part of our network. Fig. 6.12 compares pulmonary nodules as detected by our network against their correspondent ground-truth locations.

Overall, our proposed detector demonstrates promising results and a great potential within the pulmonary nodule detection task. AttentNet can efficiently achieve state-of-the-art performance with higher FROC score in contrast to state-of-the-art detectors, and a total sensitivity of 95%. See Fig. 6.13. AttentNet significantly outperforms baselines at 1.0, 2.0, and 4.0 false positives per scan, in which an effective clinical threshold is defined [3], and produces the highest sensitivity at the more challenging lower false positive per scan thresholds (i.e., 0.125, 0.25, and 0.5).

6.4 Summary

Accurate detection of lung cancer can significantly impact the success of the treatment, however, such diagnosis requires careful examination of the lung region using 3D CT scans making it time consuming and prone to human error. The use of computer aided detection (CAD) can decrease observational oversights and hence increase the accuracy of the diagnoses. However, pulmonary nodule detection is a challenging task due to the variable nodule morphology and sparsity of nodule locations within the lung region.

In this chapter, we present AttentNet, an automated 3D lung nodule detection framework from CT images. Our framework detects nodules as an ensemble of two stages, candidate proposal, in which a high number of candidates is produced, and a false positive reduction stage to reduce the number of false alarms.

The proposed framework exploits attention mechanisms to assist the network focusing on learning effective features and therefore produce an increased performance. Particularly, we propose a 3D cross-channel attention unit as well as an inter-channel cross-sectional spatial attention unit, in which we demonstrate effectiveness within the lung nodule detection task. We show that by incorporating fully convolutional networks, attention can be efficiently performed

using richer descriptors of higher spatial dimensionality, improving the overall performance in contrast to popular multi-layer perceptron based attention networks. We demonstrate the benefits of both channel and spatial attention within the detection task, we also show that channel attention yields higher performance gains in contrast to spatial attention methods.

Additionally, for the false positive reduction task, we show that by jointly analysing inputs of different spatial scales along with cross-channel attention, the network was able to aggregate information of different contextual levels and produce enhanced predictions. We also present convolutional zoom-in paths to assist the network in capturing spatial information from various semantic and spatial scales, we demonstrate the benefits of the proposed zoom-in path in the false positive reduction task.

In our experiment, we carry out an extensive analysis on LUNA16 dataset and show that AttentNet can outperform state-of-the-art lung nodule detectors by a considerable margin, with an FROC score of 0.874. Generally, our findings are inline with those in Chapters 4 and 5, where we demonstrate the benefits of dynamically exploiting inter- and cross-channel (or band) correlations to perform 3D localisation of solar ARs. In this chapter, we show that explicitly modelling such correlations can indeed be beneficial for the pulmonary nodule detection task. Our findings also confirm the benefits of the joint analysis approach in which information is aggregated from different inputs.

In Chapter 7, we study the feasibility of incorporating global context and long-range correlations within the pulmonary nodule detection task and compare their influence in contrast to inter- and cross-channel correlations on the overall localisation performance.

The work in this chapter has been published in the following journal:

- M. Almahasneh, X. Xie, A. Paiement, AttentNet for Pulmonary Lung Nodule Detection Using 3D Attention. Medical image analysis, 2022, (under review).

Chapter 7

TransCNN for Pulmonary Nodule Detection Using Self-attention

Contents

7.1 Introduction

The strength of CNNs manifests in their rich representational power and ability in embedding inter- and cross-channel spatial information. However, due to their inherent locality, CNNs suffer in modelling long-range relations. Moreover, performance of CNNs may be prone to network factors, such as depth and width [34, 35, 37, 136, 214], and data characteristics such as target structures with high morphological variance (e.g., texture, shape, and size). To combat these limitations , research efforts has been directed towards self-attention mechanisms (e.g., Multi-headed attention [39] and Vision Transformers [2]) in which global context is incorporated to model long-range correlations between arbitrary positions. In chapter 6, we demonstrated the benefits of modelling cross-channel and spatial correlations to perform attention within the lung nodule detection task. In this chapter, we investigate the possibility of exploiting long-range relations and global context to infer attention.

Transformers have received considerable interest due to their simple design and outstanding performance on different tasks. However, due to the tokenisation of inputs, Transformers degrade the local spatial context when dealing with image inputs [112]. Additionally, Transformers suffer from a quadratically growing complexity with respect to the size of the input sequence [2, 215].

To address these limitations, [2] proposed a patch-wise approach in which 2D inputs are split into a grid of cells that are used to model correlations as input sequences to perform image classification, reducing the overall computational overhead of Transformer. [216] proposed a similar approach for object detection in which the Transformer's sequence output was reshaped into a 2D embedding and was directly passed into a convolutional detection network to perform the final prediction. Both [2, 216] demonstrate promising results, however they rely on heavy pretraining which is often not applicable for different applications. Additionally, splitting images into patches can impair the spatial locality information, which is important for object localisation tasks. [117] proposed attaching a CNN decoder to the last Transformer layer to up-sample the output feature maps into a larger scale and perform object segmentation. In this line, [115] proposed using a CNN encoder to down-sample the input images into an effective size before passing them into a pure Transformer to perform the detection. Nonetheless, [117] and [115] shows that while their approaches achieve good detections for large objects, they struggle when detecting smaller object.

To address these issue, [111] proposed a hybrid 2D segmentation architecture in which a

CNN encoder is used to extract and down-sample spatial embeddings, into an size that can be used directly and effectively with a pure Transformer to perform self-attention, the resulting Transformer embedding is then reshaped back into a 2D feature map, and is passed into a CNN decoder to up-sample and perform the segmentation from a large spatial scale. The idea is to avoid the patch-wise approach by down-sampling the input image into a smaller spatial scale and feeding the Transformer high level features to model long-range relations, then up-sampling the resulting features taking advantage from a higher spatial scale to perform the final prediction and allow the localisation of fine details. Such approach leverages both, the CNN spatial representational power, as well as global context using Transformers. This approach demonstrated great performance on the organ segmentation task [111], and was adopted for different segmentation problems (e.g., 3D brain tumour segmentation [110, 217] and multi-organ segmentation [118]).

Inspired by the recent success achieved by self-attention methods, and based on the success of the hybrid Transformer CNN based approaches, in this chapter, we investigate the possibility of incorporating long-range dependencies to solve the pulmonary nodule detection problem. Particularly, based on the principles of [110, 111, 118, 217], we propose TransCNN, a hybrid Transformer CNN framework in which 3D inputs are leveraged to perform the detection task, taking advantage of the representational ability of CNNs and global context modelling in Transformers.

In our study, we explore different Transformer configuration in an ablation study to determine the best performing architecture. We also explore different types of convolutions to evaluate the impact of different feature types when used as an input for Transformer. Additionally, we investigate the possibility of combining convolutional attention mechanisms (e.g., cross-channel and inter-spatial attention) with Transformers to further enhance the detection performance.

Accordingly, we propose two hybrid Transformer CNN variations that demonstrate promising results and potential within the 3D pulmonary nodule detection task in contrast to state-of-the-art nodule detectors. In the first approach, we evaluate a *Deep Transformer* when placed within the bottleneck of an encoder-decoder CNN and using convolutional features of a single spatial scale. In the second approach however, we incorporate a *Multi-scale Transformer*, that is shallower (less Transformer layers) and requires less parameters, but takes inputs of multiple spatial scales and a relatively larger dimensions. The idea is to evaluate the trade-off between exploiting deeper Transformer (i.e., more parameters) design with relatively smaller inputs, against using information from larger scales, and a shallower Transformer. Note that the terms *deep* and *shallow* here are used to indicate the comparative context of the *Deep Transformer* –in which

we use relatively more layers and more trainable parameters– and *Multi-scale Transformer* –in which relatively less layers and less trainable parameters– approaches in relation to one another. To the best of our knowledge, we are the first incorporate Transformers into the 3D object detection task from volumetric images.

It is worth noting that unlike our work in Chapter 6 (AttentNet), where we investigate attention mechanisms in a two stage (i.e., candidate proposal and false positive reduction) detector, in this work we focus on evaluating Transformer in a single stage detector (i.e., candidate proposal task). Nevertheless, in our experiment, we evaluate the proposed approach when combined the false positive reduction stage from Chapter 6, and compare its performance to when no false positive reduction stage is used.

In the remainder part of this chapter, we discuss the details of the two proposed detector variations in Section 7.2. We then present the details of our experiment and results in Section 7.3. Last, we summarise this chapter in Section 7.4.

7.2 Proposed Method

Due to GPU limitations, and due to high computational cost of Transformers, directly using images –at an effective spatial scale– as an input for Transformers is not feasible, this is particularly difficult when dealing with volumetric images (e.g., pulmonary CT imagery). Additionally, splitting the input image into patches degrades the locality information and therefore is not suited for localisation tasks [2, 110]. Our proposed detection, based on [110, 111, 118, 217] , approach takes an input 3D CT images and processes them using a number of convolutional layers to extract, and down-sample, spatial embeddings into an effective size that can be then processed using an image Transformer to capture long-range correlations and perform attention. The output of the Transformer is then up-sampled using convolutional operations and is finally used to perform the detection. The idea is to combines both, the representational power of CNNs, as well as the ability of Transformers in modelling global context, simultaneously, to perform the pulmonary nodule detection task.

Following, we discuss the details of two approaches in which we exploit Transformer to perform the detection, in Sections 7.2.1 and 7.2.2, respectively.

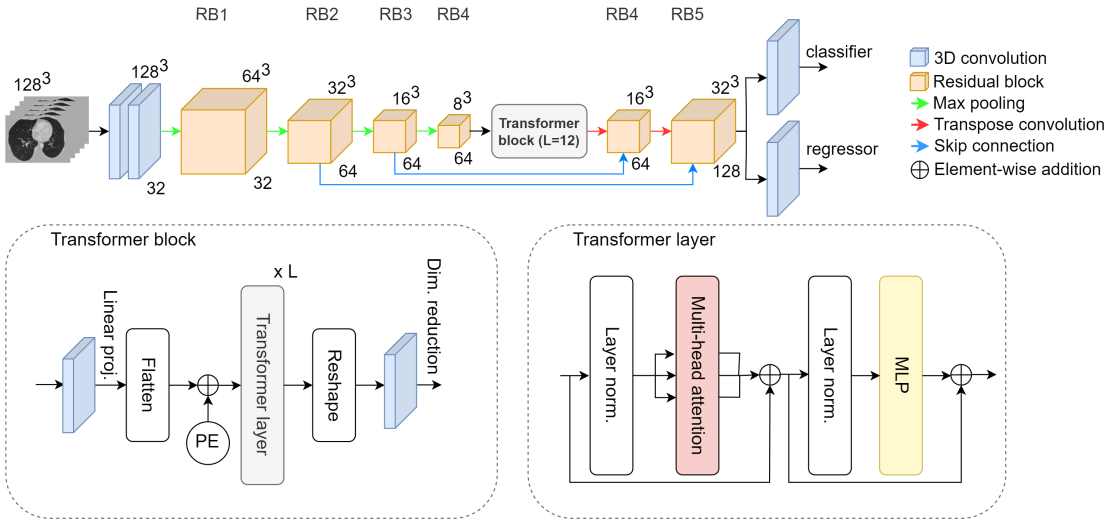

Figure 7.1: The framework of TransCNN using *Deep Transformer* approach. Input images are first analysed using a 3D CNN encoder to gradually down-sample and extract high-level embeddings that can be then effectively used with Transformer. The resulting features are then flattened and are passed as an input sequence into a block of 12 Transformer layers, in which global context and long-range correlations are investigated to perform attention. The resulting embeddings are then reshaped back into 3D feature maps and are up-sampled using the CNN's decoder where the final detection is performed. In the figure, RB denotes a residual block. PE indicates the positional embedding layer. Green, red, and blue arrows, denote down-sampling (max-pooling), up-sampling (deconvolution), and skip connections, respectively.

7.2.1 Deep Transformer

An overview backbone CNN is presented in Fig. 7.1. For the CNN component of the detection network, we utilise a similar encoder-decoder design to that used in AttentNet (Chapter 6). An input image $X \in \mathbb{R}^{CxDxHxW}$ (1 x 128 x 128 x 128 in this case) is processed using two convolutional layers, followed by a max pooling layer (stride = 2), and a sequence of 4 residual blocks [34]. The first, second and third residual blocks are followed by a max pooling layer each, in which the input dimensions are down-sampled by a factor of 2.

Unlike AttentNet (Chapter 6), while grouped convolutions can significantly reduce the number of parameters and therefore the computational overhead, we observe that using feature maps extracted using grouped convolutions can significantly degrade the performance of Transformer in contrast to using standard convolution (more details in Section 7.3). Thus, for this experiment, we continue using standard convolution.

The resulting high level feature F of size N x $\frac{D}{16}$ x $\frac{H}{16}$ x $\frac{W}{16}$ can then be effectively processed using Transformer. Note that N is the number of the 3D feature maps produced by the last convolutional layer in the CNN encoder (64 in this case). Accordingly, F is passed into a convolutional

layer, in which kernels of size 1 x 1 x 1, and a stride of 1, are used to linearly project the feature maps into a new embedding space of size 768 x $\frac{D}{16}$ x $\frac{H}{16}$ x $\frac{W}{16}$. The resulting embedding then flattened into (768,($\frac{D}{16}$ x $\frac{H}{16}$ x $\frac{W}{16}$)), and is passed as the input sequence into a deep Transformer block of 12 pure Transformer layers, in which global context and long-range dependencies are analysed to perform attention. A single Transformer layer consists of, respectively, layer normalisation, multi-headed attention layer [39] (see Eq. 2.4), layer normalisation, and a multi-layer perception.

The output embeddings from the Transformer block is therefore of size (768,($\frac{D}{16}$ x $\frac{H}{16}$ x $\frac{W}{16}$)), this is reshaped back into 768 x $\frac{D}{16}$ x $\frac{H}{16}$ x $\frac{W}{16}$ to enable processing it using convolutional layers. The resulting volumetric feature map is the passed into a convolutional layer in which the dimensionality of the resulting feature map is reduced back into 64 x $\frac{D}{16}$ x $\frac{H}{16}$ x $\frac{W}{16}$ to allow efficiently up-sampling operations in the following step.

The resulting feature maps are then processed by 2 subsequent deconvolutional layers such that each layer performs a dimension up-sampling of factor 2, with respect to its correspondent input. Each of the deconvolutional layers is followed by a residual block. The resulting feature maps are then passed into two parallel convolutional layer, in which a per position classification and nodule coordinate regression is preformed.

Similar to AttentNet (Chapter 6), to decrease the risk of overfitting, we use 3 dropout layers in the encoder part, and 1 layer in the last decoder layer. In the same line, the proposed detector was trained using focal loss [142] and smooth L1 loss [52] for the classification and regression tasks, respectively, as well as online hard negative mining strategy [177], and a similar anchor configuration to that followed in our AttentNet experiment (see Section 6.2.1).

7.2.2 Multi-scale Transformer

Similar to the *Deep Transformer* approach, input images are first processed using an encoder CNN in which spatial embeddings are extracted and down-sampled into an effective size that can then be used directly with a pure Transformer. However, in contrast to the bottleneck Transformer placement adopted in *Deep Transformer*, here we leverage convolutional features of multiple spatial scales as the input sequence to Transformer layers in which both global and local context is dynamically modelled.

Particularly, we place three layers of Transformers, each at a different feature spatial and semantic level of the CNN component, N_1 x $\frac{D}{8}$ x $\frac{H}{8}$ x $\frac{W}{8}$, and N_2 x $\frac{D}{16}$ x $\frac{H}{16}$ x $\frac{W}{16}$ in the third and forth convolutional layer of the CNN's encoder, and N_3 x $\frac{D}{8}$ x $\frac{H}{8}$ x $\frac{W}{8}$ in the first layer of the

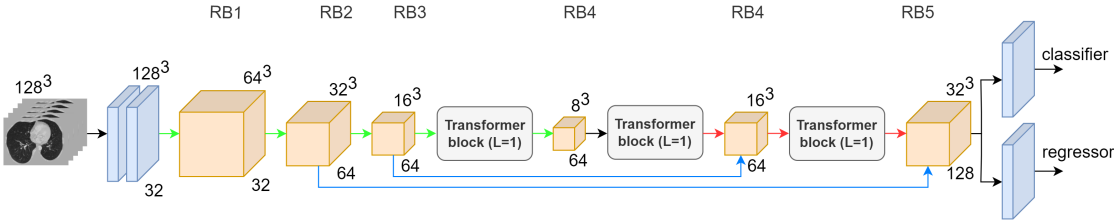

Figure 7.2: The framework of TransCNN using *Multi-scale Transformer* approach. Input images are first analysed using a 3D convolutional layers to gradually down-sample and extract high-level embeddings that can be then effectively used with Transformer. Feature maps of different spatial scales and semantic levels from the encoder-decoder CNN are used as sequence inputs into three individual Transformer blocks –of 1 Transformer layer each– to perform self-attention. Particularly, inputs of size 16^3 and 8^3 in the third and forth (i.e., RB3 and RB4 in the figure) convolutional layer of the CNN's encoder, and 16^3 in the first (i.e., RB5 in the figure) layer of the CNN's decoder layer. Each Transformer layer is then followed by a residual block to further extract convolutional features and re-enforce the feature locality of which Transformers suffer to model. The resulting feature maps are then further up-sampled and used to perform the final detection. In the figure, RB denotes a residual block. Green, red, and blue arrows, denote down-sampling (max-pooling), up-sampling (deconvolution), and skip connections, respectively. Orange and blue cubic blocks represent, residual blocks and 3D convolution, respectively.

CNN's decoder layer. Where N_1, N_2, and N_3 are the numbers of 3D feature channels in the first, second, and third feature maps, respectively. Each Transformer layer is then followed by a residual block (convolution) to further extract spatial features. An overview of the proposed approach is presented in Fig. 7.2.

The idea is to 1) use spatial features of multiple spatial scales and of different semantic levels to investigate for global context and long range relations to perform attention, and 2) re-enforce the local spatial context in the extracted embeddings by applying further convolutions to the Transformer output. In contrast to the *Deep Transformer* approach, this *Multi-scale Transformer* strategy has a shallower design (i.e., less Transformer layers within a Transformer block) and therefore requires less trainable parameters ($\sim$58M vs. $\sim$92M), while having the advantage of inputs of higher dimension, and different scales and semantic levels.

7.3 Experiments

Similar to our AttentNet experiment in Chapter 6, we train and evaluate the proposed detector using LUNA16 [3] dataset, we also use similar pre-processing and data augmentation strategy (see Section 6.3.2). We also use similar optimiser configuration and batch size (see Section 6.3). We set the number of training epochs to 250, equating to $\sim$2 training days using x2 NVIDIA V100 16GB GPUs. We also use FROC score [202] –average sensitivity at 7 predefined false

Table 7.1: Transformer ablation study: FROC scores obtained using different Transformer configurations on a randomly selected keep-out fold from LUNA16 dataset. Note that RB denotes *residual block*, and is used to indicate the residual block of which output is used as an input for the Transformer.

# Transformer layers	Embedding dim.	Transformer position	Grouped conv.	FROC
–	–	–	×	0.818
4	768	RB4	×	0.837
8	768	RB4	×	0.823
12	768	RB4	×	**0.842**
12	576	RB4	×	0.833
12	384	RB4	×	0.839
1	768	RB3, RB4, RB5	×	0.829
12	768	RB4	✓	0.746

positive rates per scan (i.e., 0.125, 0.25, 0.5, 1, 2, 4, and 8)– to evaluate the performance of all tested detectors.

In the remainder of this section, we present an ablation study in which we explore the best Transformer configuration, in Section 7.3.1. Following, in Section 7.3.2, we compare the performance of the proposed Transformer based detector, against popular attention mechanisms. Finally, in Section 7.3.3, we evaluate the performance of the proposed detector against stat-of-the-art lung nodule detectors, as well as AttentNet from Chapter 6.

As discussed in Section 7.1, in our experiment, we focus on evaluating Transformer for the candidate proposal task, however, in Section 7.3.3, we evaluate the proposed detector when combined the false positive reduction stage from AttentNet (Chapter 6) and compare it to the performance when no false positive reduction is used.

7.3.1 Transformer Ablation Study

In this section, we investigate different Transformer layouts (number of Transformer layers and embedding spatial dimensions), as well as the positioning of the Transformer block, and different types of convolutional features (standard convolution vs. grouped convolutions [136]), to find the best performing Transformer configuration. For convenience, we use split the data into a 80% to 20% training and testing sets, respectively, however, in Section 7.3.3, we follow a 10-folds cross validation process to compare the performance of the best performing detector against state-of-the-art baseline detectors. Results are presented in Table 7.1 and Fig. 7.3.

We first evaluate the *Deep Transformer* approach, in which the Transformer block is positioned within the bottleneck of the encoder-decoder CNN, i.e., post residual block 4 (RB4 in

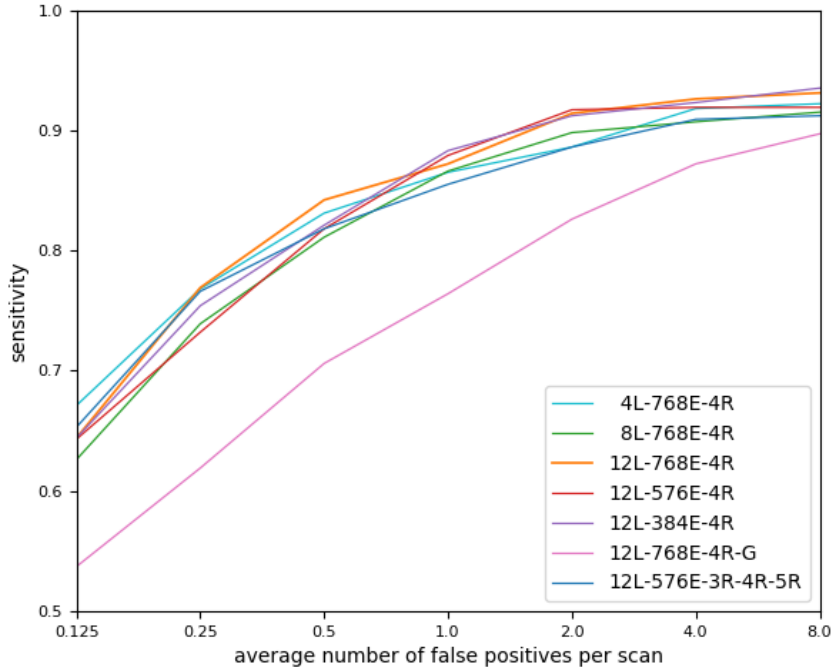

Figure 7.3: Transformer ablation study: FROC scores at different false positive threshold obtained using different Transformer configurations on a randomly selected keep-out fold from LUNA16 dataset. Note that L and E indicate the number of Transformer layers and the feature embedding size used to configure the Transformer block of each detector. R indicates the residual block in which a Transformer block takes as input, and G indicates the use of grouped convolutions to construct the CNN component of the detector.

the Table 7.1) and using standard convolution (i.e., no grouped convolution). We find that all configurations show an enhanced performance against the baseline CNN when no Transformer is used –with an FROC score of 0.818–, with the best performance achieved when using 12 Transformer layers within the Transformer block and an embedding size of 768 –with an FROC score of 0.842–, in contrast to all other combinations (i.e., 4 or 8 layers, and 576 or 384 feature embedding size). This is in line with the suggested configuration followed in [2]. Therefore, we continue using Transformer with 12 layers and a feature embedding of size 768 for the remainder of our experiment.

Moreover, when comparing the performance when using features maps extracted using grouped convolution (i.e., similar to AttentNet in chapter 6) against using standard convolution, we find that standard convolution significantly outperforms grouped convolution. We observe that grouped convolution deteriorates the performance of the detector even when compared to

the detection when no Transformer is used, with an FROC score of 0.746. This may caused by the fact that grouped convolution splits the input feature maps and convolutional filters into a number of groups (32 in this case) and performs the feature extraction from each group individually to reduce the computational overhead. Contrary, standard convolution uses all inputs with each filter to produce all outputs. This may indicate that standard convolution is better when modelling long range relations across the channel axis when compared to grouped convolution, which is in line with the concepts of Transformer, in which long-range correlations are investigated.

Lastly, we evaluate our *Multi-scale Transformer* approach. We find that this approach shows an enhanced performance in contrast to the baseline network when no Transformer is used. Similarly when compared to the *Deep Transformer* approach with 8 Transformer layers using only $\sim$58M parameters, comparing to $\sim$64M parameters, respectively. The *Multi-scale Transformer* approach produces promising results, even when compared to the best performing *Deep Transformer* with 12 layers, requiring only 63% of the overall learnable parameters (i.e., $\sim$58M vs.$\sim$92M).

Overall, we find that *Deep Transformer* performs the best comparing to all tested approaches. We also find that *Multi-scale Transformer* can also achieve promising performance, at a significantly lower computational cost (i.e., number of parameters). In terms of inference time, as expected, due to the increased computational overhead caused by using Transformer networks, when comparing to the RPN baseline –with no attention– requiring ~ 0.093 seconds (GPU time) per sample (i.e., 128 x 128 x 128 image patch), the proposed *Deep Transformer* (with 12 Transformer layers) approach records a slower performance requiring ~ 0.116 seconds inference time, and ~ 0.111 seconds when following the *Multi-scale Transformer* scheme.

7.3.2 Transformer Against Convolutional Attention

In this section, we evaluate the performance of the proposed Transformer based detectors against convolutional based attention methods, to study the impact of incorporating global context comparing to cross-channel and inter-spatial convolution based attention. Similar to Section 7.3.1, we evaluate the performance of all detectors using 80% and 20% training to testing data split. Results are presented in Table 7.2 and Fig. 7.4.

Overall, we notice that most tested attention methods have a positive impact on the detection performance when compared to the baseline CNN (RPN) when no attention is used. Particularly, when comparing channel attention (CA) methods from [38], [1] and AttentNet (Chapter 6), we

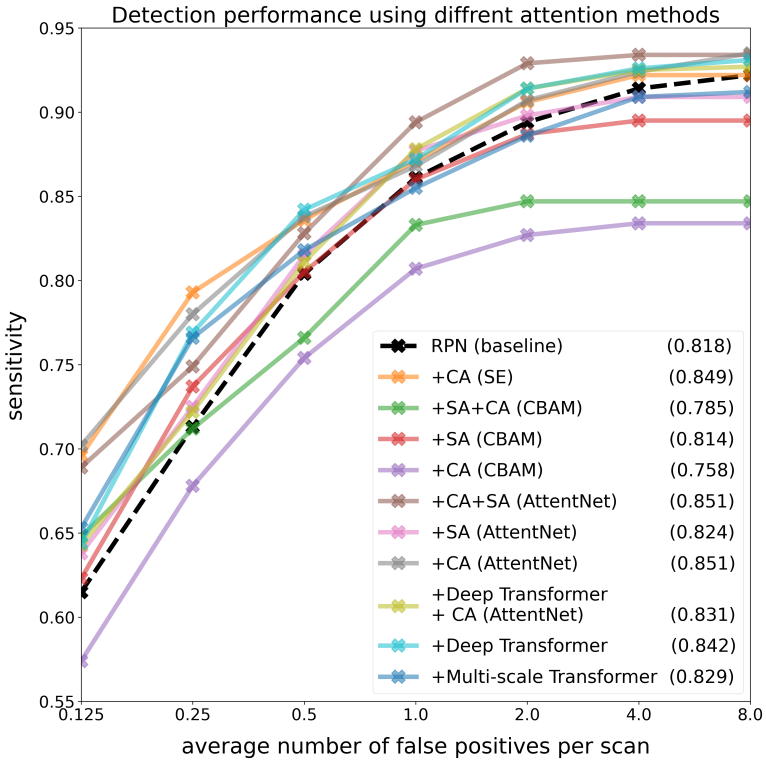

Figure 7.4: FROC of TransCNN using different attention mechanisms under comparison on a randomly selected keep-out fold from LUNA16 dataset. Legend indicates average FROC score across all false positive thresholds for each tested detector.

find that the channel attention approach from AttentNet performs the best, with an overall FROC score of 0.851. A similar observation is found when comparing spatial attention methods from AttentNet and [1]. Additionally, when comparing the best performing channel and spatial attention (i.e., channel and spatial attention from AttentNet), we find that channel attention performs better than Spatial attention from. We also find that combining the best performing channel and spatial attention (i.e., channel and spatial attention from AttentNet) achieve an identical performance to that when using only channel attention. These findings are in line with our findings in Chapter 6. These findings demonstrate the effectiveness of convolutional attention, particularly cross-channel attention, in which cross-channel correlations are exploited to enhance (refine) the quality of the extracted features and therefore produce an increased performance.

When comparing self-attention (i.e., Transformer) based detectors against convolutional attention, we find that *Deep Transformer* produces better results than both spatial and channel attention from [1] and spatial attention from AttentNet, and comparable results to those by chan-

Table 7.2: FROC score at different numbers of false positives per scan obtained by different attention methods under comparison on a randomly selected keep-out fold from LUNA16 dataset. Here, CA and SA stand for channel attention and spatial attention, respectively. The highest scores are highlighted in bold.

FROC	Mean	0.125	0.25	0.5	1.0	2.0	4.0	8.0
RPN (baseline)	0.818	0.615	0.713	0.804	0.861	0.894	0.914	0.922
RPN+CA [38]	0.849	0.696	**0.793**	0.836	0.871	0.906	0.922	0.922
RPN+SA+CA [1]	0.785	0.648	0.712	0.766	0.833	0.847	0.847	0.847
RPN+SA [1]	0.814	0.623	0.737	0.805	0.860	0.887	0.895	0.895
RPN+CA [1]	0.758	0.574	0.678	0.754	0.807	0.827	0.834	0.834
RPN+(CA+SA from AttentNet)	**0.851**	0.689	0.749	0.828	**0.894**	**0.929**	**0.934**	0.934
RPN+(SA from AttentNet)	0.824	0.638	0.725	0.815	0.877	0.898	0.909	0.909
RPN+(CA from AttentNet)	**0.851**	**0.702**	0.780	0.838	0.868	0.907	0.924	**0.935**
RPN+Deep Transformer	0.842	0.644	0.769	**0.842**	0.872	0.914	0.926	0.931
RPN+Deep Transformer+(CA from AttentNet)	0.831	0.643	0.722	0.811	0.878	0.914	0.925	0.927
RPN+Multi-scale Transformer	0.829	0.653	0.766	0.818	0.855	0.886	0.909	0.912

nel attention from [38], channel attention from AttentNet, and channel and spatial attention from AttentNet combined. Specifically, *Deep Transformer* yields 0.842 FROC score, comparing to 0.851 by the best performing detector, and accordingly is the 4^{th} best detector out of all 11 tested detectors. This demonstrates a great potential of Transformers within the pulmonary nodule 3D detection task. This shows that global context can indeed be exploited for an increased performance.

We also evaluate the possibility of combining Transformer with the best performing cross-channel attention from AttentNet. We find no particular benefit from this combination, yet more interestingly, we observe a deterioration in the performance in contrast to both *Deep Transformer* and channel attention from AttentNet when used individually. Intuitively, cross-channel attention aims to capture long-range correlations along the channel axis, in a localised manner (i.e., using convolution), to rank the relevance of the individual feature maps (channels) with respect to a given objective (e.g., nodule detection in this case), which is different in principle to the Transformers, in which long-range relations are captured in a non-localised manner (i.e., between arbitrary positions) aiming to evaluate the importance of global context to the given task. Based on this intuition, and the aforementioned observation, this suggests that incorporating both localised attention (convolution based) and global attention simultaneously may hinder the quality of the feature maps produced by one another, in this case, when detecting pulmonary nodule.

Moreover, when evaluating *Multi-scale Transformer*, we notice that it achieves lower FROC score than *Deep Transformer* However, *Multi-scale Transformer* gets higher FROC score than

Table 7.3: FROC at different numbers of false positives per scan obtained by our proposed Transformer based detector, TransCNN, in contrast to baseline methods using 10-folds cross validation on LUNA16 dataset. Here, FP-red indicates the use of the false positive reduction stage from AttentNet (Chapter 6), TTA indicates the use of testing time augmentation, and CA indicate cross-channel attention. The highest scores are highlighted in bold.

FROC	Mean	0.125	0.25	0.5	1.0	2.0	4.0	8.0
ZNET [129]	0.811	0.661	0.724	0.779	0.831	0.872	0.892	0.915
3D RCNN [132]	0.834	0.662	0.746	0.815	0.864	0.902	0.918	0.932
DeepLung [135]	0.842	0.692	0.769	0.824	0.865	0.893	0.917	0.933
DeepSeed [138]	0.862	0.739	0.803	0.858	0.888	0.907	0.916	0.920
AttentNet (RPN)	0.842	0.656	0.774	0.831	0.874	0.903	0.923	0.936
AttentNet (RPN+FP-red.)	0.871	0.752	0.817	0.857	0.885	**0.920**	0.933	0.933
AttentNet (RPN+FP-red.+TTA)	0.874	0.748	0.812	0.856	**0.893**	0.919	0.942	0.945
TransCNN (RPN)	0.830	0.656	0.737	0.804	0.863	0.901	0.916	0.932
TransCNN (RPN+FP-red.)	0.867	0.745	0.814	0.853	0.884	0.917	0.929	0.929
TransCNN (RPN+FP-red.+TTA)	0.872	0.748	0.807	0.861	0.888	0.915	0.940	0.943
CA from Ch. 6 (RPN)	0.846	0.665	0.775	0.837	0.881	0.909	0.921	0.937
CA from Ch. 6 (RPN+FP-red.)	0.872	**0.764**	**0.819**	0.861	0.882	**0.920**	0.930	0.930
CA from Ch. 6 (RPN+FP-red.+TTA)	**0.876**	0.755	0.818	**0.863**	0.889	0.916	**0.943**	**0.947**

channel and spatial attention from [1], and spatial attention from AttentNet, and is accordingly the 6^{th} best detector out of all 11 tested detectors.

Generally, our experiment demonstrates promising results and a great potential of Transformers within the 3D detection task, in this case pulmonary nodule detection. Results show that exploiting global context and self-attention can indeed achieve results comparable to state-of-the-art performance. We anticipate that the performance of Transformers within the 3D detection can be enhanced even further in the future by undergoing improvements in a similar trend to that observed with CNNs over the past years.

7.3.3 System Performance

In this section, we evaluate the performance of the proposed detector, TransCNN, against state-of-the-art pulmonary nodule detectors, as well as the cross-channel attention based AttentNet from Chapter 6. We also evaluate cross-channel attention from Chapter 6, using the same configuration of the CNN component followed in TransCNN (i.e., using standard convolution, in contrast to grouped convolution used in AttentNet). As suggested in LUNA16 [3], we perform a 10-folds cross validation to evaluate the performance of all detectors under comparison. Note that performing 10-folds cross validation requires extensive amounts of time (e.g., $\sim$2 days per fold), therefore, In this Section, we only evaluate our *Deep Transformer* approach, since it

achieves the best results comparing to the *multi-scale Transformer* approach, as demonstrated in Sections 7.3.1 and 7.3.2. Moreover, while we focus in this work on investigating Transformers in a single stage detector (i.e., equivalent to the candidate proposal stage in two stage detector), in this section, we further evaluate the proposed detector when combined with the false positive reduction network from AttentNet, and compare it to the performance when no false positive reduction stage is incorporated. Similar to AttentNet, the two stages are combined in an ensemble, in which scores of suspicious locations are defined as the average probability predicted by both, the candidate proposal and the false positive reduction stage. Detections with probabilities ≥ 0.3 are used to form the final set of detections. Additionally, and similar to our experiment in Chapter 6, we evaluate the proposed detector using testing time augmentation (TTA), in which predictions are aggregated from different image transformations similar to those used during training to improve the overall detection performance. Results are presented in Table 7.3.

Generally, TransCNN (RPN only) achieves close performance to methods from [129, 132, 135, 138], as well as AttentNet (RPN only), with an overall FRROC score of 0.830. Particularly, TransCNN gets close or higher sensitivity at false positive per scan thresholds from 1.0 to 8.0, in which an effective clinical threshold (i.e., 1.0 - 4.0) is defined within [3], to detectors from [129, 132, 135, 138]. This demonstrates a great potential of incorporating global context (i.e., Transformers) for the nodule detection task. Moreover, we observe a significant increase in the overall performance of TransCNN when incorporating the false positive reduction stage , with an FROC score of 0.867 , outperforming all detectors from [129, 132, 135, 138], and approaching the performance of AttentNet when the false positive reduction stage is used. This is in line with our finding in Chapter 6. This demonstrates the importance of the false positive reduction stage within the pulmonary detection task.

Furthermore, we find that cross-channel attention from Chapter 6 performs better than both TransCNN and AttentNet, when comparing all three detectors using RPN only. In the same line, we observe that when incorporating the false positive reduction stage, cross-channel attention achieves the best FROC score amongst all tested detectors. This is in line with our findings in Chapter 6. This indicates that cross-channel correlations are indeed of high relevance when detecting pulmonary nodules. Note that unlike our channel attention experiment in AttentNet, here, we use standard convolution instead of grouped convolution to build the backbone CNN. Accordingly, both channel attention based detectors, i.e., using standard and grouped convolution, achieve 0.842 and 0.846 FROC score, respectively, and equate to $\sim$8M and $\sim$3M learnable parameters, respectively. This demonstrates that grouped convolution can indeed be exploited to

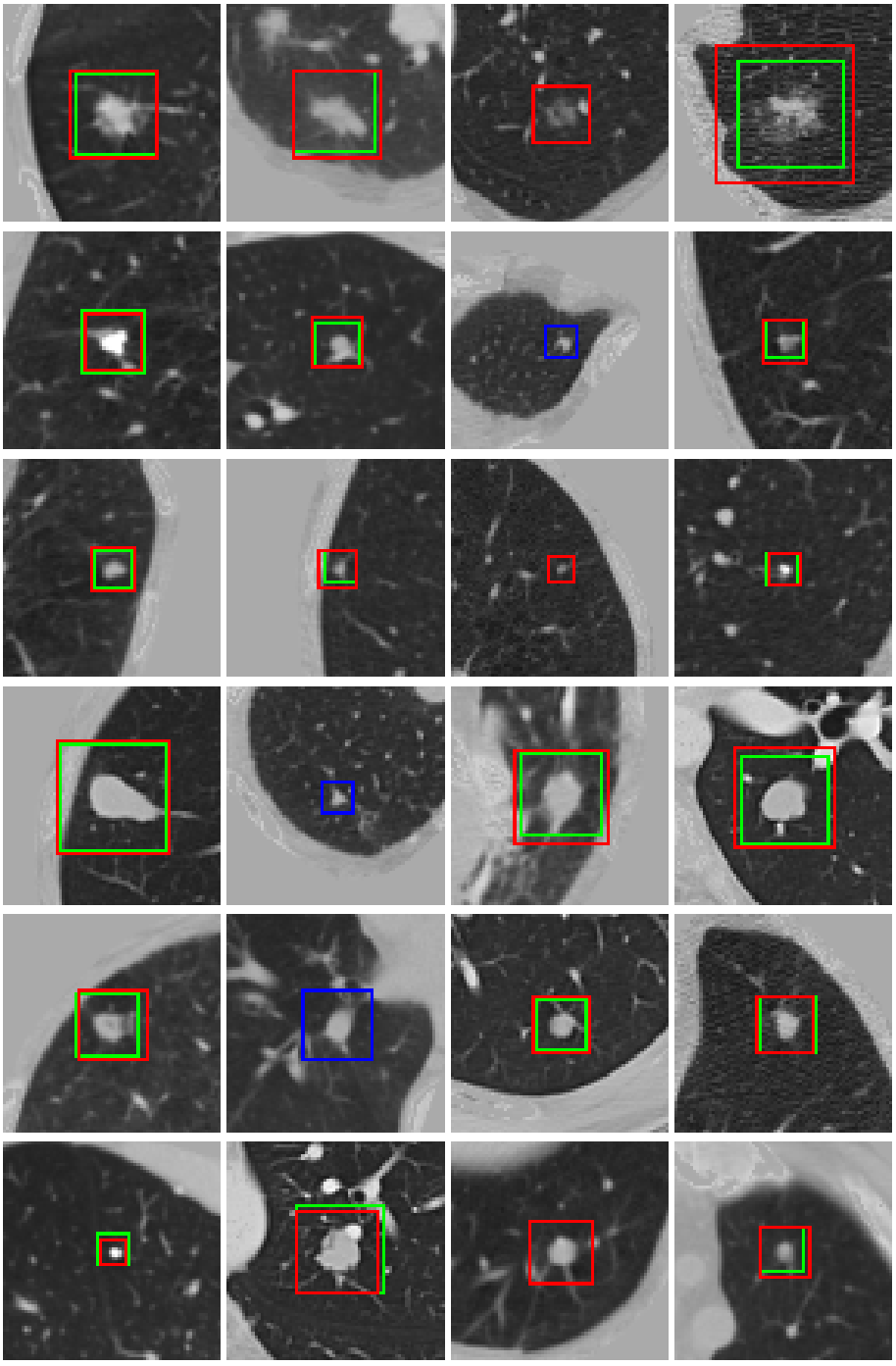

Figure 7.5: Pulmonary nodules detected by TransCNN (green) and their correspondent ground-truth boxes (red). Blue boxes indicate false positive detections. Red boxes with no overlapping detections indicate false negative samples.

effectively perform the 3D nodule detection task while maintaining a relatively low parameter overhead. This also demonstrates the versatility of our cross-channel attention, it can be straightforwardly and effectively incorporated with different CNN backbone and types of convolution.

Similar to our finding in Chapter 6, we observe that exploiting testing time augmentation (TTA) can enhance the overall detection with minimal additional computation cost, increasing the overall FROC score of TransCNN from 0.867 to 0.872, making it the 3rd highest FROC score amongst all 12 tested detectors. This trend is consistent with all detectors in which TTA was exploited. Therefore, we adopt TTA to our proposed detector, TransCNN. Fig. 7.5 visualises nodules as detected by TransCNN with respect to their correspondent ground-truth locations.

Overall, results demonstrate a superior performance of the cross-channel attention based detector comparing to all tested detectors and attention techniques. We also find that incorporating global context and self-attention (i.e., Transformer) can indeed achieve performance close or higher to stat-of-the-art methods within the 3D pulmonary detection task. These promising findings indicate potential for further future improvements on the Transformer performance, and possibly for different 3D detection tasks and applications.

7.4 Summary

We presented TransCNN, a hybrid Transformer CNN based 3D pulmonary nodule detector that exploits long-range correlations and global context to infer attention. The essence of this approach is to combine both, the spatial representational power and locality of CNNs with the Transformer ability in modelling global context.

Accordingly, we propose and evaluate two different approaches, *Deep Transformer* and *Multi-scale Transformer*, in which we explore the impact of exploiting a Transformer of more layers and trainable parameters using high-level convolutional features as inputs (i.e., at the CNN's bottleneck) in contrast to using a shallower Transformer (i.e., less Transformer layers and less trainable parameters) while taking advantage of inputs of different spatial scales (and relatively higher dimensions) and semantic levels. We compare the performance of our proposed approach against state-of-the-art cross-channel and inter-spatial convolutional based attention, as well as our proposed 3D channel and spatial attention based detector from Chapter 6 (AttentNet).

In our extended experiment, we demonstrate the the effectiveness of the two proposed approaches and find that Transformers can indeed achieve comparable or higher performance to

stat-of-the-art detectors and attention methods. We also find that Transformer's performance is sensitive to the convolutional feature type. Particularly, we observe that Transformer's observe a significant drop in performance when using feature maps produced by grouped convolution as an input, in contrast to standard convolution. Therefore, Transformer input features must be carefully selected with respect to the given task.

Additionally, we find that our 3D based cross-channel attention (proposed Chapter 6) achieves the best results out of all tested detectors and attention methods within the nodule detection task, with an overall FROC score of 0.876, comparing to 0.872 achieved by the proposed Transformer based detector.

We also observe that combining both cross-channel attention (the best performing convolutional attention approach) and global context based attention (i.e., Transformer) may lead to a decreased performance in contrast to when each attention approach is applied individually. This suggests that applying localised and global attention simultaneously may impact the quality of the feature maps (i.e., learning) of one another negatively, when detecting pulmonary nodules.

Generally, Transformers demonstrate a great potential within the nodule detection task. We anticipate further Transformer performance enhancement and future improvements to take place in a similar trend to that observed with CNNs in the past decade. To the best of our knowledge, we are the first to evaluate Transformer within the object detection task from volumetric medical images.

Chapter 8

Conclusions and Future Work

Contents

8.1 Conclusion

This study explored the possibility of incorporating feature correspondence and feature-level correlations within deep learning based object detection and segmentation problems that target contexts of three dimensional imaging. We investigated implicit cross-feature spatial correspondence and feature-level correlation learning schemes that aim at inducing CNNs, through design characteristics and objectives, to capture such dependencies. We also explore the possibility of explicitly modelling feature-level dependencies, in a learnable manner, and exploiting these relations to directly infer feature importance and thus enhance the effectiveness of the feature extraction process in CNNs. We evaluate the influence of cross-channel and inter-spatial correlations, when utilised in a localised manner using convolution, as well as long-range dependencies between arbitrary locations to capture global context. In this thesis, we show that both feature-level correlation learning strategies can be feasibly incorporated to address detection and segmentation problems from multi-modal and volumetric imagery.

- **Cross-modal joint analysis based 3D detection.** A method to handle the detection problem of 3D solar active regions in multi-spectral images that observe different 2D layers (altitudes) in the solar atmosphere was developed in Chapter 4. We introduced a joint analysis approach in which cross-band information is individually analysed to extract band-specific features that are dynamically aggregated throughout the detection network, at different semantic levels, and are jointly analysed to capture spatial correspondence and dependencies between the different bands, and subsequently predict active region locations, individually, in each of the imaged solar layers. To address the lack of manually labelled solar active region detection data, we designed a multi-spectral annotation tool that provides cues from different imaging bands as well as temporal information when labeling active regions. Accordingly we designed and presented two manually labelled 3D, multi-spectral, solar active region detection datasets, of which we use to perform our analysis. Throughout our study, we demonstrate the advantage of exploiting cross-band dependencies to perform the detection, over different applications (solar active regions detection, and brain tumour detection from multi-modal MRI images). We demonstrated that the cross-band fusion of high level embeddings leads to an improved performance in contrast to the fusion of low level embeddings or transfer learning based fusion approaches, within the solar active region detection task. We also demonstrated that different imaging scenarios benefit from different feature fusion strategies, or different number of image

input bands, and may accordingly require careful engineering with respect to the target application. The proposed joint analysis approach achieves promising results in contrast to baseline, single band (e.g., Faster R-CNN), and multi-band based approaches. To the best of our knowledge, our work is the first to address the problem of solar active region detection using deep learning methods.

- **Cross-modal joint analysis based 3D segmentation.** In Chapter 5, we further incorporated and demonstrated the effectiveness of the proposed joint analysis scheme within the segmentation problem of 3D solar active regions from multi-spectral images. The proposed segmentation approach integrates the principles of the proposed joint analysis scheme into an encoder-decoder CNN to exploit both cross-band relations, as well as band-specific information of different spatial scales and semantic levels. To overcome the difficulty in preparing accurate and detailed segmentation annotations, we presented a weakly-supervised approach in which the segmentation network was recursively trained to segment active regions based on bounding box priors. Throughout our experiment, we demonstrate that using cross-band spatial dependencies can improve the performance of CNN segmentation models in contrast to popular single band based analysis (e.g., FCN and U-Net). We also explore different feature fusion schemes using semantic information of different levels (early, and late fusion, as well as band-specific skip connections). The proposed joint analysis scheme demonstrated promising performance in different applications and types of data (active regions segmentation in multi-spectral solar images, brain tumour segmentation in multi-modal MRI images, and cloud segmentation in multi-spectral satellite imagery), using different levels of supervision (weak- and full-supervision).

- **Local cross-channel and inter-spatial correlation based attention in 3D detection contexts.** We investigated the possibility of explicitly modelling cross-feature correlations in a localised manner within the 3D detection task of pulmonary nodules from volumetric computed tomography images, in Chapter 6. We presented a two stage (i.e., candidate proposal, and false positive reduction) 3D detection framework, where we introduce two fully convolutional attention schemes, to efficiently capture cross-channel and inter-spatial dependencies based on rich 3D descriptors. The proposed channel attention scheme aims at analysing and modelling correlations between the different feature channels in a localised manner to infer channel-wise importance. The spatial attention scheme on the other hand,

considers cross-sectional spatial relations to infer spatial attention. In our experiment, we demonstrated that while both spatial and cross-channel attention schemes improve the performance within the 3D pulmonary nodule detection context, incorporating channel-wise attention leads to higher gains in performance in contrast to either the spatial-wise attention or the combination of both spatial and channel-wise attention, using less trainable parameters. We also demonstrated that directly using 3D information to infer attention is superior in 3D contexts comparing to popular methods that rely on heavy dimensionality reduction strategies, including CBAM and Squeeze and Excitation networks. This was feasible thanks to the fully convolutional scheme followed in the proposed attention mechanism. Our study demonstrates that feature-level correlations can be explicitly and efficiently utilised in an objective driven manner leading to significant gains in the discriminative ability of CNNs within 3D detection contexts. In the same line, we found that even decision level correlations may exploited to improve the performance of CNNs by the aggregating predictions from different input transformations that are spatially correlated. Through our study, we also proposed an activation function to combat the risk of dying neurons associated with rectified linear units (ReLU) based activations by allowing gradient flows for negative inputs while preserving the linear characteristic of such activations. We empirically demonstrated an enhanced performance using the proposed activation in contrast to a number of popular activation functions, including ReLU, Leaky ReLU, and ELU. Additionally, for the false positive reduction task, we presented and utilised a joint analysis scheme to dynamically integrate information of different spatial contextual levels. We also proposed Zoom-in convolutional paths to allow the network dynamically and effectively capture and re-enforce spatial details at different semantic levels and spatial scales. The proposed methods achieve considerable gains in contrast to state-of-the-art pulmonary nodule detection methods.

- **3D detection based on global correlations and self-attention.** We explored the influence of incorporating long-range dependencies between arbitrary positions within 3D detection contexts, in Chapter 7. We presented a hybrid Transformer CNN framework to handle the 3D detection problem of pulmonary nodules in volumetric computed tomography images. The proposed solution utilises CNNs to extract and down-sample 3D feature maps from image inputs into an effective size of which a Transformer network can then efficiently process with respect to the computational cost. As such, the proposed scheme provides complementary information by taking advantage of both, the localisation and

representational characteristics of CNNs, as well as the Transformer's ability in capturing global context. Accordingly, we proposed two hybrid Transformer CNN variants in which we evaluate the impact of incorporating a relatively deep transformer (i.e., using more layers and trainable parameters) when used with high level semantic 3D convolutional descriptors, in contrast to a shallower (i.e., using less layers and trainable parameters), less computation demanding, Transformer CNN design that exploits feature maps of multiple spatial scales and semantic levels. We demonstrate using an ablative study that both Transformer designs can improve the detection when compared to the baseline CNN when no Transformer is used. We observe that using high semantic level convolutional features with Transformers and relatively more trainable parameters induces higher gains in the detection performance (at a higher computational cost) in contrast to Transformer designs that exploit feature maps of multiple spatial scales and semantic levels, but less trainable parameters (at a lower computational cost). The proposed detection approach achieves comparable or higher performance in contrast to state-of-the-art pulmonary nodule detection CNN based methods, as well as popular localised attention schemes (e.g., CBAM spatial and channel wise attention, and Squeeze and Excitation networks). We further studied the possibility of jointly exploiting localised attention schemes with global context based attention and found no significant improvements from such combination. Furthermore, we demonstrate that by ensembling the proposed Transformer CNN based detection network with the false positive reduction module from Chapter 6, the proposed method can obtain state-of-the-art performance in contrast to all baseline pulmonary nodule detection methods. Our study demonstrates that global correlations can be effectively incorporated to infer attention and are indeed useful within 3D detection contexts. To the best of our knowledge, we are the first to explore Transformers within the object detection task in volumetric medical imaging contexts.

Generally, extensive analyses were provided throughout this study and demonstrate the importance of incorporating feature-level correspondence and correlations within deep learning based localisation contexts, in line with our hypothesis in which this work was motivated by.

8.2 Future Work

Feature-level dependency analysis is a fundamental concept that is directly related to the problem of deep learning. It may be formulated in various fashions that are commonly influenced by

the same core philosophy, in which the ultimate motivation is to automatically capture patterns, regularities, and relations present within given data to model some problem or task. The concepts presented in this thesis may be straightforwardly generalised into a wide scope of machine learning problems, particularly, we believe that such analysis could be very opportune within tasks that are, by essence, highly dependant on cross-feature correlations.

For instance, particularly but not exclusively, image registration, video summarising and frame interpolation or recovery, object tracking, multi-modal imaging based tasks, and 3D reconstruction, are all problems that are highly dependant on cross channel and spatial dependencies, and require strong spatial representational and localisation powers. Such tasks may therefore indicate an attractive candidate in which cross-channel and inter-spatial, convolutional based, correlation learning techniques may be exploited. In the same line, long range correlations and global context based approaches (e.g., Transformer networks) also indicate promising potential within such contexts. We believe that hybrid methods (e.g., Transformer CNN hybrid networks) may particularly form a more applicable solution for such scenarios. Indeed, using hybrid approaches takes advantage of the representational power of CNNs, the ability of Transformers in capturing global context, and provide a convenient approach to overcome the high computational cost associated with such tasks.

In this work, we explored joint analysis based approaches within different contexts and tasks. Such analysis provides an effective solution for by inducing deep learning models to capture cross-feature relations by dynamically and gradually cross-integrating information from different input subsets, e.g., multi-modal imaging scenarios. However, the incorporation of parallel convolutional analyses can be computationally limiting (i.e., limited scalability), and from a feature sharing point of view is not ideal. Explicitly modelling such relations in a learnable manner, e.g., using attention mechanisms, may be a good direction for future works. As we demonstrated in our experiment, attention based analysis can provide a feasible solution in context where the computational complexity of the target problem is restraining. Accordingly, we believe that combining the joint analysis principles with explicit correlation modelling schemes, as well as parameter sharing strategies (e.g., sharing of feature extraction backbones), may form a feasible solution for such problems.

During our study, as in many machine learning problems, the limited data availability formed a main challenge in which we attempted to address using different strategies, e.g., data gathering, labelling, data augmentation, transfer learning where applicable, and by using weak learning approaches. While such methods may provide sufficient solutions to the problem, the availability

of extended amounts of data can directly impact the performance of deep learning methods. In this regard, further tests using extended amounts of data, as well as different data types and domains, may provide a closer and a wider insight on the capacity and generalisability of the proposed methods.

In our work, we demonstrated that using cross-sectional information to infer 3D spatial attention is opportune and may be feasibly performed. Our approach focused particularly on axial, coronal and sagittal planes to infer correlations, nonetheless, other cross-sectional planes may also be utilised to conduct such analysis. A potential future direction may involve investigating the possibility of directly inferring cross-sectional importance in 3D contexts, by predicting cross-sectional orientations of which the discriminative ability of the learner model is maximised. Such analysis may be directly exploited, e.g., to perform detection tasks by jointly analysing most relevant cross-sectional planes, or indirectly, e.g., to infer feature correlations and attention. Such understanding may provide a convenient solution for tasks that target volumetric data where computational complexity is typically constraining.

Our study demonstrates a notable potential of global context based self-attention networks, i.e., Transformers, within 3D object detection contexts. We show that utilising convolutional operations and self-attention networks leads to a favorable enhanced performance in contrast to either of these methods when utilised separately. We find that this however, may be directly impacted by the type the convolutional features, e.g., grouped against standard convolutions, used to perform the global analysis. Future works may investigate the impact of different convolutional features, e.g., depth-wise, spatially separable, dilated, and shuffled convolutions, to further explain such observations and potentially improve the quality and usability of these approaches. In the same line, recently emerging self-attention based approaches, e.g., [218–220], investigate and demonstrate the effectiveness of exploiting hierarchical spatial information to infer self-attention in object detection contexts, directly or with the help of convolutional networks. Further studying, evaluating, and extending such methods form an important starting point for future research.

Generally, in line with the unprecedented advancements in the field of deep learning, cross-feature correlation modelling seems to offer a promising scope of which such analysis may be utilised within. We anticipate a continuing trend of positive contributions and improvements towards more robust machine learning, and consequently, benefiting life.

Bibliography

[1] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Proceedings of the European conference on computer vision (European Conference on Computer Vision), 2018, pp. 3–19.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in International Conference on Learning Representations, 2020.

[3] A. Setio, A. Traverso, T. De Bel, M. S. Berens, C. Van Den Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, B. Geurts et al., "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge," Medical image analysis, 2017.

[4] J. Yanase and E. Triantaphyllou, "A systematic survey of computer-aided diagnosis in medicine: Past and present developments," Expert Systems with Applications, vol. 138, p. 112821, 2019.

[5] Q. Ye and D. Doermann, "Text detection and recognition in imagery: A survey," IEEE transactions on pattern analysis and machine intelligence, vol. 37, no. 7, pp. 1480–1500, 2014.

[6] G. Guo and N. Zhang, "A survey on deep learning based face recognition," Computer vision and image understanding, vol. 189, p. 102805, 2019.

[7] J. Cao, Y. Pang, J. Xie, F. S. Khan, and L. Shao, "From handcrafted to deep features for pedestrian detection: A survey," IEEE transactions on pattern analysis and machine intelligence, 2021.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Conference on Computer Vision and Pattern Recognition, 2005.

[9] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in Proceedings of international conference on image processing, vol. 1. IEEE, 2002, pp. I–I.

[10] Z. Guo, L. Zhang, and D. Zhang, "A completed modeling of local binary pattern operator for texture classification," IEEE transactions on image processing, vol. 19, no. 6, pp. 1657–1663, 2010.

[11] D. G. Lowe, "Object recognition from local scale-invariant features," in Proceedings of the seventh IEEE international conference on computer vision, vol. 2. IEEE, 1999, pp. 1150–1157.

[12] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," in European conference on computer vision. Springer, 2006, pp. 404–417.

[13] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," IEEE Intelligent Systems and their applications, vol. 13, no. 4, pp. 18–28, 1998.

[14] Y. Freund, R. E. Schapire et al., "Experiments with a new boosting algorithm," in International Conference on Machine Learning, vol. 96. Citeseer, 1996, pp. 148–156.

[15] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5–32, 2001.

[16] P. Viola and M. J. Jones, "Robust real-time face detection," International Journal of Computer Vision, 2004.

[17] S. Lloyd, "Least squares quantization in pcm," IEEE transactions on information theory, vol. 28, no. 2, pp. 129–137, 1982.

[18] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," Advances in neural information processing systems, vol. 14, 2001.

[19] R. M. Haralick and L. G. Shapiro, "Image segmentation techniques," Computer vision, graphics, and image processing, vol. 29, no. 1, pp. 100–132, 1985.

[20] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in nd images," in IEEE international conference on computer vision, vol. 1. IEEE, 2001, pp. 105–112.

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems, vol. 25, 2012.

[23] X. Wu, D. Sahoo, and S. C. Hoi, "Recent advances in deep learning for object detection," Neurocomputing, vol. 396, pp. 39–64, 2020.

[24] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.

[25] W. Liu et al., "SSD: Single shot multibox detector," in European Conference on Computer Vision, 2016.

[26] J. Redmon et al., "You only look once: Unified, real-time object detection," in Conference on Computer Vision and Pattern Recognition, 2016.

[27] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 4, pp. 640–651, 2016.

[28] 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.

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

[30] S. Dara and P. Tumma, "Feature extraction by using deep learning: A survey," in International Conference on Electronics, Communication and Aerospace Technology. IEEE, 2018, pp. 1795–1801.

[31] F. Shaheen, B. Verma, and M. Asafuddoula, "Impact of automatic feature extraction in deep learning architecture," in International conference on digital image computing: techniques and applications. IEEE, 2016, pp. 1–8.

[32] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, "Dive into deep learning," 2021.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[34] 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.

[35] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[36] S. Zagoruyko and N. Komodakis, "Wide residual networks," in Proceedings of the British Machine Vision Conference. BMVA Press, 2016.

[37] O. Kopuklu, N. Kose, A. Gunduz, and G. Rigoll, "Resource efficient 3d convolutional neural networks," in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019, pp. 0–0.

[38] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in neural information processing systems, 2017, pp. 5998–6008.

[40] Z. Guo et al., "Deep learning-based image segmentation on multimodal medical imaging," IEEE Transactions on Radiation and Plasma Medical Sciences, 2019.

[41] J. Wagner et al., "Multispectral pedestrian detection using deep fusion convolutional neural networks," in European Symposium on Artificial Neural Networks, 2016.

[42] K. Takumi et al., "Multispectral object detection for autonomous vehicles," in Thematic Workshops, 2017.

[43] R. Jarolim et al., "Multi-channel coronal hole detection with a convolutional neural network," in Machine Learning in Heliophysics, 2019.

[44] J. Schmidhuber, "Deep learning in neural networks: An overview," Neural networks, vol. 61, pp. 85–117, 2015.

[45] J. Feng, X. He, Q. Teng, C. Ren, H. Chen, and Y. Li, "Reconstruction of porous media from extremely limited information using conditional generative adversarial networks," Physical Review E, vol. 100, no. 3, p. 033308, 2019.

[46] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," in Neural networks: Tricks of the trade. Springer, 2012, pp. 437–478.

[47] M. Li, T. Zhang, Y. Chen, and A. J. Smola, "Efficient mini-batch training for stochastic optimization," in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014, pp. 661–670.

[48] K. D. P. and B. J., "Adam: A method for stochastic optimization," in International Conference on Learning Representations, 2015.

[49] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," The journal of machine learning research, vol. 15, no. 1, pp. 1929–1958, 2014.

[50] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in International conference on machine learning. PMLR, 2015, pp. 448–456.

[51] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in European conference on computer vision. Springer, 2014, pp. 818–833.

[52] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," Advances in neural information processing systems, vol. 28, pp. 91–99, 2015.

[53] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in European Conference on Computer Vision, 2018.

[54] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in Advances in Neural Information Processing Systems. Curran Associates, Inc., 2016.

[55] X. Wang, S. Zheng, C. Zhang, R. Li, and L. Gui, "R-YOLO: A real-time text detector for natural scenes with arbitrary rotation," Sensors, vol. 21, no. 3, p. 888, 2021.

[56] R.-C. Chen et al., "Automatic license plate recognition via sliding-window darknet-YOLO deep learning," Image and Vision Computing, vol. 87, pp. 47–56, 2019.

[57] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," International journal of computer vision, vol. 104, no. 2, pp. 154–171, 2013.

[58] R. Girshick, "Fast r-cnn," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

[59] P. Soviany and R. Ionescu, "Optimizing the trade-off between single-stage and two-stage deep object detectors using image difficulty prediction," in International Symposium on Symbolic and Numeric Algorithms for Scientific Computing. IEEE, 2018.

[60] J. Huang et al., "Speed/accuracy trade-offs for modern convolutional object detectors," in Conference on Computer Vision and Pattern Recognition, 2017.

[61] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.

[62] A. M. Hafiz and G. M. Bhat, "A survey on instance segmentation: state of the art," International journal of multimedia information retrieval, pp. 1–19, 2020.

[63] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 4, pp. 834–848, 2017.

[64] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in Deep learning and data labeling for medical applications. Springer, 2016, pp. 179–187.

[65] L. Wang, X. Chen, L. Hu, and H. Li, "Overview of image semantic segmentation technology," in IEEE Joint International Information Technology and Artificial Intelligence Conference, vol. 9. IEEE, 2020, pp. 19–26.

[66] X. Liu, L. Song, S. Liu, and Y. Zhang, "A review of deep-learning-based medical image segmentation methods," Sustainability, vol. 13, no. 3, p. 1224, 2021.

[67] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," Medical image analysis, vol. 42, pp. 60–88, 2017.

[68] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3150–3158.

[69] N. Gao, Y. Shan, Y. Wang, X. Zhao, Y. Yu, M. Yang, and K. Huang, "SSAP: Single-shot instance segmentation with affinity pyramid," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 642–651.

[70] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, "SOLO: Segmenting objects by locations," in European Conference on Computer Vision. Springer, 2020, pp. 649–665.

[71] S. Mohajerani, T. A. Krammer, and P. Saeedi, "A cloud detection algorithm for remote sensing images using fully convolutional neural networks," in IEEE Multimedia Signal Processing, 2018.

[72] S. Wang et al., "Weakly supervised deep learning for segmentation of remote sensing imagery," Remote Sensing, 2020.

[73] Q. Li, A. Arnab, and P. H. Torr, "Weakly-and semi-supervised panoptic segmentation," in European Conference on Computer Vision, 2018.

[74] C. Rother, V. Kolmogorov, and A. Blake, ""grabcut" interactive foreground extraction using iterated graph cuts," ACM transactions on graphics, 2004.

[75] J. Pont-Tuset et al., "Multiscale combinatorial grouping," in Conference on Computer Vision and Pattern Recognition. Citeseer, 2014.

[76] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in Conference on Computer Vision and Pattern Recognition, 2016.

[77] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in International Conference on Computer Vision, 2015.

[78] A. Khoreva et al., "Weakly supervised semantic labelling and instance segmentation," in Conference on Computer Vision and Pattern Recognition, 2016.

[79] S. Mohajerani and P. Saeedi, "Cloud-Net: An end-to-end cloud detection algorithm for Landsat 8 imagery," in The International Geoscience and Remote Sensing Symposium, 2019.

[80] O. A. B. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in Conference on Computer Vision and Pattern Recognition, 2015.

[81] T. Ishii et al., "Detection by classification of buildings in multispectral satellite imagery," in International Conference on Pattern Recognition, 2016.

[82] S. Hwang et al., "Multispectral pedestrian detection: Benchmark dataset and baselines," in Conference on Computer Vision and Pattern Recognition, 2015.

[83] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 2014.

[84] M. Weber, C. Rist, and J. M. Zöllner, "Learning temporal features with cnns for monocular visual ego motion estimation," in IEEE International Conference on Intelligent Transportation Systems. IEEE, 2017, pp. 1–6.

[85] L. Liu, B. Zhang, and H. Wang, "Organ localization in pet/ct images using hierarchical conditional Faster R-CNN method," in Proceedings of the Third International Symposium on Image Computing and Digital Medicine, 2019, pp. 249–253.

[86] S. Afshari, A. BenTaieb, and G. Hamarneh, "Automatic localization of normal active organs in 3d pet scans," Computerized Medical Imaging and Graphics, vol. 70, pp. 111–118, 2018.

[87] X. Yang, N. Wang, Y. Wang, X. Wang, R. Nezafat, D. Ni, and P.-A. Heng, "Combating uncertainty with novel losses for automatic left atrium segmentation," in International workshop on statistical atlases and computational models of the heart. Springer, 2018, pp. 246–254.

[88] R. Huang, W. Xie, and J. A. Noble, "Vp-nets: Efficient automatic localization of key brain structures in 3d fetal neurosonography," Medical image analysis, vol. 47, pp. 127–139, 2018.

[89] M. Ebner, G. Wang, W. Li, M. Aertsen, P. A. Patel, R. Aughwane, A. Melbourne, T. Doel, S. Dymarkowski, P. De Coppi et al., "An automated framework for localization, segmentation and super-resolution reconstruction of fetal brain mri," NeuroImage, vol. 206, p. 116324, 2020.

[90] X. Zhou, T. Kojima, S. Wang, X. Zhou, T. Hara, T. Nozaki, M. Matsusako, and H. Fujita, "Automatic anatomy partitioning of the torso region on ct images by using a deep convolutional network with majority voting," in Medical Imaging : Computer-Aided Diagnosis, vol. 10950. International Society for Optics and Photonics, 2019, p. 109500Z.

[91] B. D. de Vos, J. M. Wolterink, P. A. de Jong, T. Leiner, M. A. Viergever, and I. Išgum, "Convnet-based localization of anatomical structures in 3-d medical images," IEEE transactions on medical imaging, vol. 36, no. 7, pp. 1470–1481, 2017.

[92] G. E. Humpire-Mamani, A. A. A. Setio, B. van Ginneken, and C. Jacobs, "Efficient organ localization using multi-label convolutional neural networks in thorax-abdomen ct scans," Physics in Medicine & Biology, vol. 63, no. 8, p. 085003, 2018.

[93] K. C. Kaluva, K. Vaidhya, A. Chunduru, S. Tarai, S. P. P. Nadimpalli, and S. Vaidya, "An automated workflow for lung nodule follow-up recommendation using deep learning," in International Conference on Image Analysis and Recognition. Springer, 2020, pp. 369–377.

[94] X. Xu, F. Zhou, B. Liu, D. Fu, and X. Bai, "Efficient multiple organ localization in ct image using 3d region proposal network," IEEE transactions on medical imaging, vol. 38, no. 8, pp. 1885–1898, 2019.

[95] Z. Qiu, J. Langerman, N. Nair, O. Aristizabal, J. Mamou, D. H. Turnbull, J. Ketterling, and Y. Wang, "Deep bv: A fully automated system for brain ventricle localization and

segmentation in 3d ultrasound images of embryonic mice," in IEEE Signal Processing in Medicine and Biology Symposium. IEEE, 2018, pp. 1–6.

[96] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in International conference on medical image computing and computer-assisted intervention. Springer, 2016, pp. 424–432.

[97] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in International conference on 3D vision. IEEE, 2016, pp. 565–571.

[98] M. Simony, S. Milzy, K. Amendey, and H.-M. Gross, "Complex-YOLO: An euler-region-proposal for real-time 3d object detection on point clouds," in Proceedings of the European Conference on Computer Vision (European Conference on Computer Vision) Workshops, 2018, pp. 0–0.

[99] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7263–7271.

[100] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2019, pp. 770–779.

[101] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2017, pp. 1907–1915.

[102] X. Zhou, K. Yamada, T. Kojima, R. Takayama, S. Wang, X. Zhou, T. Hara, and H. Fujita, "Performance evaluation of 2d and 3d deep learning approaches for automatic segmentation of multiple organs on ct images," in Medical Imaging : Computer-Aided Diagnosis, vol. 10575. International Society for Optics and Photonics, 2018, p. 105752C.

[103] H. Lu, H. Wang, Q. Zhang, S. W. Yoon, and D. Won, "A 3d convolutional neural network for volumetric image semantic segmentation," Procedia Manufacturing, vol. 39, pp. 422–428, 2019.

[104] S. Ji, C. Zhang, A. Xu, Y. Shi, and Y. Duan, "3d convolutional neural networks for crop classification with multi-temporal remote sensing images," Remote Sensing, vol. 10, no. 1, p. 75, 2018.

[105] L. Sun, Z. Wang, H. Pu, G. Yuan, L. Guo, T. Pu, and Z. Peng, "Attention-embedded complementary-stream cnn for false positive reduction in pulmonary nodule detection," Computers in Biology and Medicine, vol. 133, p. 104357, 2021.

[106] X. Lu, E. Y. Chang, C.-n. Hsu, J. Du, and A. Gentili, "Multi-classification study of the tuberculosis with 3d CBAM-ResNet and efficientnet," in Central Europe Workshop Proceedings, 2021.

[107] M. A. Nawshad, U. A. Shami, S. Sajid, and M. M. Fraz, "Attention based residual network for effective detection of covid-19 and viral pneumonia," in International Conference on Digital Futures and Transformative Technologies. IEEE, 2021, pp. 1–7.

[108] B. A. Sangeroki and T. W. Cenggoro, "A fast and accurate model of thoracic disease detection by integrating attention mechanism to a lightweight convolutional neural network," Procedia Computer Science, vol. 179, pp. 112–118, 2021.

[109] J. Park, S. Woo, J.-Y. Lee, and I.-S. Kweon, "Bam: Bottleneck attention module," in British Machine Vision Conference (BMVC). British Machine Vision Association (BMVA), 2018.

[110] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, "Transbts: Multimodal brain tumor segmentation using transformer," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2021, pp. 109–119.

[111] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," arXiv preprint arXiv:2102.04306, 2021.

[112] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," arXiv preprint arXiv:2101.11986, 2021.

[113] B. H. Menze et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," IEEE Transactions on Medical Imaging, 2015.

[114] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in Proceedings of the IEEE International Conference on Computer Vision, 2021, pp. 7262–7272.

[115] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in European Conference on Computer Vision. Springer, 2020, pp. 213–229.

[116] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 6881–6890.

[117] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2022, pp. 574–584.

[118] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation," arXiv preprint arXiv:2103.03024, 2021.

[119] A. Fraknoi, D. Morrison, and S. Wolff, "Astronomy," Houston: Open-Stax, 2016.

[120] J.-P. Delaboudiniere, G. Artzner, J. Brunaud, A. H. Gabriel, J.-F. Hochedez, F. Millier, X. Song, B. Au, K. Dere, R. A. Howard et al., "EIT: extreme-ultraviolet imaging telescope for the soho mission," in The SOHO Mission. Springer, 1995, pp. 291–312.

[121] A. Benkhalil et al., "Active region detection and verification with the solar feature catalogue," Solar Physics, 2006.

[122] K. Revathy, S. Lekshmi, and S. R. P. Nayar, "Fractal-based fuzzy technique for detection of active regions from solar images," Solar Physics, 2005.

[123] J. C. Bezdek, R. Ehrlich, and W. Full, "Fcm: The fuzzy c-means clustering algorithm," Computers & geosciences, vol. 10, no. 2-3, pp. 191–203, 1984.

[124] P. A. Higgins et al., "Solar magnetic feature detection and tracking for space weather monitoring," Advances in Space Research, 2011.

[125] C. Verbeeck et al., "The SPoCA-suite: Software for extraction, characterization, and tracking of active regions and coronal holes on EUV images," Astronomy & Astrophysics, 2013.

[126] R. Krishnapuram and J. Keller, "The possibilistic C-means algorithm: Insights and recommendations," IEEE Transactions on Fuzzy Systems, 1996.

[127] C. Verbeeck et al., "A multi-wavelength analysis of active regions and sunspots by comparison of automatic detection algorithms," Solar Physics, 2013.

[128] D. Riquelme and M. A. Akhloufi, "Deep learning for lung cancer nodules detection and classification in ct scans," Artificial intelligence, vol. 1, no. 1, pp. 28–67, 2020.

[129] M. Berens, R. van der Gugten, M. de Kaste, J. Manders, and G. Zuidhof, "Znet-lung nodule detection," 2016.

[130] J. Shi, "Lung nodule detection using convolutional neural networks," Electrical Engineering and Computer Sciences. Berkeley, California: University of California at Berkeley, 2018.

[131] J. Pearl, Probabilistic reasoning in intelligent systems: networks of plausible inference. Elsevier, 2014.

[132] F. Liao, M. Liang, Z. Li, X. Hu, and S. Song, "Evaluate the malignancy of pulmonary nodules using the 3-d deep leaky noisy-or network," IEEE transactions on neural networks and learning systems, vol. 30, no. 11, pp. 3484–3495, 2019.

[133] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, "Dual path networks," in Advances in Neural Information Processing Systems, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[134] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Annals of statistics, pp. 1189–1232, 2001.

[135] W. Zhu, C. Liu, W. Fan, and X. Xie, "Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification," in IEEE Winter Conference on Applications of Computer Vision, 2018, pp. 673–681.

[136] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1492–1500.

[137] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[138] Y. Li and Y. Fan, "Deepseed: 3d squeeze-and-excitation encoder-decoder convolutional neural networks for pulmonary nodule detection," in IEEE 17th International Symposium on Biomedical Imaging (ISBI). IEEE, 2020, pp. 1866–1869.

[139] H. Tang, X. Liu, and X. Xie, "An end-to-end framework for integrated pulmonary nodule detection and false positive reduction," in IEEE International Symposium on Biomedical Imaging. IEEE, 2019, pp. 859–862.

[140] Y. Chen, P. Cao, L. Dou, and J. Yang, "An end-to-end framework for pulmonary nodule detection and false positive reduction from ct images," in The Fourth International Symposium on Image Computing and Digital Medicine, 2020, pp. 156–162.

[141] H. Tang, D. R. Kim, and X. Xie, "Automated pulmonary nodule detection using 3d deep convolutional neural networks," in IEEE International Symposium on Biomedical Imaging. IEEE, 2018, pp. 523–526.

[142] 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.

[143] G. Polat, U. Halici, and Y. S. Dogrusoz, "False positive reduction in lung computed tomography images using convolutional neural networks," arXiv preprint arXiv:1811.01424, 2018.

[144] M. Gani et al., "Multispectral object detection with deep learning," 2021.

[145] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," arXiv preprint arXiv:1406.2199, 2014.

[146] A. Eitel et al., "Multimodal deep learning for robust rgb-d object recognition," in IEEE International Conference on Intelligent Robots and Systems, 2015.

[147] X. Song et al., "A multispectral feature fusion network for robust pedestrian detection," Alexandria Enginering Journal, 2021.

[148] H. Schunker, D. Braun, A. Birch, R. Burston, and L. Gizon, "Sdo/hmi survey of emerging active regions for helioseismology," Astronomy & Astrophysics, vol. 595, p. A107, 2016.

[149] H. Schunker, A. C. Birch, R. H. Cameron, D. C. Braun, L. Gizon, and R. B. Burston, "Average motion of emerging solar active region polarities-i. two phases of emergence," Astronomy & Astrophysics, vol. 625, p. A53, 2019.

[150] X. Sun, M. G. Bobra, J. T. Hoeksema, Y. Liu, Y. Li, C. Shen, S. Couvidat, A. A. Norton, and G. H. Fisher, "Why is the great solar active region 12192 flare-rich but cme-poor?" The Astrophysical Journal Letters, vol. 804, no. 2, p. L28, 2015.

[151] M. Everingham, S. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," International journal of computer vision, vol. 111, no. 1, pp. 98–136, 2015.

[152] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in European conference on computer vision. Springer, 2014, pp. 740–755.

[153] R. Galvez et al., "A machine-learning data set prepared from the NASA solar dynamics observatory mission," The Astrophysical Journal Supplement Series, 2019.

[154] E. Friis-Christensen and K. Lassen, "Length of the solar cycle: An indicator of solar activity closely associated with climate," Science, 1991.

[155] N. Otsu, "A threshold selection method from gray-level histograms," IEEE Systems, Man, and Cybernetics Society, 1979.

[156] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," International Journal of Computer Vision, 2015.

[157] B. Crabbe et al., "Skeleton-free body pose estimation from depth images for movement analysis," in IEEE International Conference on Computer Vision Workshop, 2015.

[158] C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, no. 3, pp. 273–297, 1995.

[159] A. Bansal et al., "Pixelnet: Representation of the pixels, by the pixels, and for the pixels," arXiv preprint arXiv:1702.06506, 2017.

[160] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the hausdorff distance," IEEE Transactions on pattern analysis and machine intelligence, vol. 15, no. 9, pp. 850–863, 1993.

[161] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: a cancer journal for clinicians, vol. 71, no. 3, pp. 209–249, 2021.

[162] K. Latimer and T. Mott, "Lung cancer: diagnosis, treatment principles, and screening," American family physician, vol. 91, no. 4, pp. 250–256, 2015.

[163] L. G. Collins, C. Haines, R. Perkel, and R. E. Enck, "Lung cancer: diagnosis and management," American family physician, vol. 75, no. 1, pp. 56–63, 2007.

[164] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman et al., "The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans," Medical physics, vol. 38, no. 2, pp. 915–931, 2011.

[165] A. S. S. Tehrani, H. Lee, S. C. Mathews, A. Shore, M. A. Makary, P. J. Pronovost, and D. E. Newman-Toker, "25-year summary of us malpractice claims for diagnostic errors 1986–2010: an analysis from the national practitioner data bank," BMJ quality & safety, vol. 22, no. 8, pp. 672–680, 2013.

[166] C. S. Lee, P. G. Nagy, S. J. Weaver, and D. E. Newman-Toker, "Cognitive and system factors contributing to diagnostic errors in radiology," American Journal of Roentgenology, vol. 201, no. 3, pp. 611–617, 2013.

[167] J. P. Borgstede, R. S. Lewis, M. Bhargavan, and J. H. Sunshine, "Radpeer quality assurance program: a multifacility study of interpretive disagreement rates," Journal of the American College of Radiology, vol. 1, no. 1, pp. 59–65, 2004.

[168] B. Al Mohammad, P. C. Brennan, and C. Mello-Thoms, "A review of lung cancer screening and the role of computer-aided detection," Clinical radiology, vol. 72, no. 6, pp. 433–442, 2017.

[169] R. A. Castellino, "Computer aided detection (cad): an overview," Cancer Imaging, vol. 5, no. 1, p. 17, 2005.

[170] S. Matsumoto, Y. Ohno, T. Aoki, H. Yamagata, M. Nogami, K. Matsumoto, Y. Yamashita, and K. Sugimura, "Computer-aided detection of lung nodules on multidetector ct in concurrent-reader and second-reader modes: a comparative study," European journal of radiology, vol. 82, no. 8, pp. 1332–1337, 2013.

[171] C. Jacobs, E. M. van Rikxoort, K. Murphy, M. Prokop, C. M. Schaefer-Prokop, and B. van Ginneken, "Computer-aided detection of pulmonary nodules: a comparative study using the public lidc/idri database," European radiology, vol. 26, no. 7, pp. 2139–2147, 2016.

[172] S. G. Armato, F. Li, M. L. Giger, H. MacMahon, S. Sone, and K. Doi, "Lung cancer: performance of automated lung nodule detection applied to cancers missed in a ct screening program," Radiology, vol. 225, no. 3, pp. 685–692, 2002.

[173] R. Yuan, P. M. Vos, and P. L. Cooperberg, "Computer-aided detection in screening ct for pulmonary nodules," American Journal of Roentgenology, vol. 186, no. 5, pp. 1280–1287, 2006.

[174] I. J. Lee, G. Gamsu, J. Czum, N. Wu, R. Johnson, and S. Chakrapani, "Lung nodule detection on chest ct: evaluation of a computer-aided detection (cad) system," Korean journal of radiology, vol. 6, no. 2, pp. 89–93, 2005.

[175] B. Xiao, Z. Yang, X. Qiu, J. Xiao, G. Wang, W. Zeng, W. Li, Y. Nian, and W. Chen, "Pamdensenet: A deep convolutional neural network for computer-aided covid-19 diagnosis," IEEE Transactions on Cybernetics, 2021.

[176] X. Fu, L. Bi, A. Kumar, M. Fulham, and J. Kim, "Multimodal spatial attention module for targeting multimodal pet-ct lung tumor segmentation," IEEE Journal of Biomedical and Health Informatics, 2021.

[177] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 761–769.

[178] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in Proceedings of the fourteenth international conference on artificial intelligence and statistics, 2011, pp. 315–323.

[179] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 6546–6555.

[180] G. Zhang, Z. Yang, L. Gong, S. Jiang, and L. Wang, "Classification of benign and malignant lung nodules from ct images based on hybrid features," Physics in Medicine & Biology, vol. 64, no. 12, p. 125011, 2019.

[181] J. GU, F. WANG, Y. QI, Z. SUN, Z. TIAN, and Y. ZHANG, "Retrieval method of pulmonary nodule images based on multi-scale convolution feature fusion," Journal of Computer Applications, vol. 40, no. 2, pp. 561–565.

[182] G. Zhang, Z. Yang, L. Gong, S. Jiang, L. Wang, and H. Zhang, "Classification of lung nodules based on ct images using squeeze-and-excitation network and aggregated residual transformations," La radiologia medica, pp. 1–10, 2020.

[183] A. L. Maas, A. Y. Hannun, A. Y. Ng et al., "Rectifier nonlinearities improve neural network acoustic models," in International Conference on Machine Learning, vol. 30, no. 1. Citeseer, 2013, p. 3.

[184] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 2016.

[185] M. Yu, L. Cai, L. Gao, and J. Gao, "Amplification method of lung nodule data based on dcgan generation algorithm," in International Conference of Pioneering Computer Scientists, Engineers and Educators. Springer, 2020, pp. 563–576.

[186] L. Haibo, T. Shanli, S. Shuang, and L. Haoran, "An improved YOLOv3 algorithm for pulmonary nodule detection," in IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference, vol. 4. IEEE, 2021, pp. 1068–1072.

[187] W. Huang and L. Hu, "Using a noisy U-Net for detecting lung nodule candidates," IEEE Access, vol. 7, pp. 67 905–67 915, 2019.

[188] Q. Wang, F. Shen, L. Shen, J. Huang, and W. Sheng, "Lung nodule detection in ct images using a raw patch-based convolutional neural network," Journal of digital imaging, vol. 32, no. 6, pp. 971–979, 2019.

[189] J. Tan, Y. Huo, Z. Liang, and L. Li, "A fast automatic juxta-pleural lung nodule detection framework using convolutional neural networks and vote algorithm," in International Workshop on Patch-based Techniques in Medical Imaging. Springer, 2018, pp. 85–92.

[190] N. Moshkov, B. Mathe, A. Kertesz-Farkas, R. Hollandi, and P. Horvath, "Test-time augmentation for deep learning-based cell segmentation on microscopy images," Scientific reports, vol. 10, no. 1, pp. 1–7, 2020.

[191] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

[192] D. Shanmugam, D. Blalock, G. Balakrishnan, and J. Guttag, "When and why test-time augmentation works," arXiv preprint arXiv:2011.11156, 2020.

[193] A. R. Larici, A. Farchione, P. Franchi, M. Ciliberto, G. Cicchetti, L. Calandriello, A. Del Ciello, and L. Bonomo, "Lung nodules: size still matters," European Respiratory Review, vol. 26, no. 146, 2017.

[194] L.-H. Shen, X.-H. Wang, M.-X. Gao, and B. Li, "Classification of benign-malignant pulmonary nodules based on multi-view improved dense network," in International Conference on Intelligent Computing. Springer, 2021, pp. 582–593.

[195] B. Zhao, G. Gamsu, M. S. Ginsberg, L. Jiang, and L. H. Schwartz, "Automatic detection of small lung nodules on ct utilizing a local density maximum algorithm," journal of applied clinical medical physics, vol. 4, no. 3, pp. 248–260, 2003.

[196] M. Woźniak, D. Połap, G. Capizzi, G. L. Sciuto, L. Kośmider, and K. Frankiewicz, "Small lung nodules detection based on local variance analysis and probabilistic neural network," Computer methods and programs in biomedicine, vol. 161, pp. 173–180, 2018.

[197] M. Callister, D. Baldwin, A. Akram, S. Barnard, P. Cane, J. Draffan, K. Franks, F. Gleeson, R. Graham, P. Malhotra et al., "British thoracic society guidelines for the investigation and management of pulmonary nodules: accredited by nice," Thorax, vol. 70, no. Suppl 2, pp. ii1–ii54, 2015.

[198] M. K. Gould, J. Fletcher, M. D. Iannettoni, W. R. Lynch, D. E. Midthun, D. P. Naidich, and D. E. Ost, "Evaluation of patients with pulmonary nodules: when is it lung cancer?: Accp evidence-based clinical practice guidelines," Chest, vol. 132, no. 3, pp. 108S–130S, 2007.

[199] M. Das, G. Mühlenbruch, A. H. Mahnken, T. G. Flohr, L. Gündel, S. Stanzel, T. Kraus, R. W. Günther, and J. E. Wildberger, "Small pulmonary nodules: effect of two computer-aided detection systems on radiologist performance," Radiology, vol. 241, no. 2, pp. 564–571, 2006.

[200] H. Robbins and S. Monro, "A stochastic approximation method," The annals of mathematical statistics, pp. 400–407, 1951.

[201] 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.

[202] M. Niemeijer, M. Loog, M. D. Abramoff, M. A. Viergever, M. Prokop, and B. van Ginneken, "On combining computer-aided detection systems," IEEE Transactions on Medical Imaging, vol. 30, no. 2, pp. 215–223, 2010.

[203] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman et al., "The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans," Medical physics, 2011.

[204] E. A. Kazerooni, J. H. Austin, W. C. Black, D. S. Dyer, T. R. Hazelton, A. N. Leung, M. F. McNitt-Gray, R. F. Munden, and S. Pipavath, "Acr–str practice parameter for the performance and reporting of lung cancer screening thoracic computed tomography (ct): 2014 (resolution 4)," Journal of thoracic imaging, vol. 29, no. 5, pp. 310–316, 2014.

[205] D. P. Naidich, A. A. Bankier, H. MacMahon, C. M. Schaefer-Prokop, M. Pistolesi, J. M. Goo, P. Macchiarini, J. D. Crapo, C. J. Herold, J. H. Austin et al., "Recommendations for the management of subsolid pulmonary nodules detected at ct: a statement from the fleischner society," Radiology, vol. 266, no. 1, pp. 304–317, 2013.

[206] D. Manos, J. M. Seely, J. Taylor, J. Borgaonkar, H. C. Roberts, and J. R. Mayo, "The lung reporting and data system (lu-rads): a proposal for computed tomography screening," Canadian Association of Radiologists Journal, vol. 65, no. 2, pp. 121–134, 2014.

[207] D. Aberle, A. Adams, C. Berg, and W. Black, "Reduced lung-cancer mortality with low-dose computed tomographic screening," New England Journal of Medicine, 2011.

[208] K. Murphy, B. van Ginneken, A. M. Schilham, B. De Hoop, H. A. Gietema, and M. Prokop, "A large-scale evaluation of automatic pulmonary nodule detection in chest ct using local image features and k-nearest-neighbour classification," Medical image analysis, 2009.

[209] M. Tan, R. Deklerck, B. Jansen, M. Bister, and J. Cornelis, "A novel computer-aided lung nodule detection system for ct images," Medical physics, 2011.

[210] C. Jacobs, E. M. Van Rikxoort, T. Twellmann, E. T. Scholten, P. A. De Jong, J.-M. Kuhnigk, M. Oudkerk, H. J. De Koning, M. Prokop, C. Schaefer-Prokop et al., "Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images," Medical image analysis, 2014.

[211] A. A. Setio, C. Jacobs, J. Gelderblom, and B. van Ginneken, "Automatic detection of large pulmonary solid nodules in thoracic ct images," Medical physics, 2015.

[212] A. Traverso, E. L. Torres, M. E. Fantacci, and P. Cerello, "Computer-aided detection systems to improve lung cancer early diagnosis: state-of-the-art and challenges," in Journal of Physics: Conference Series. IOP Publishing, 2017.

[213] Q. Dou, H. Chen, L. Yu, J. Qin, and P.-A. Heng, "Multilevel contextual 3-d cnns for false positive reduction in pulmonary nodule detection," IEEE Transactions on Biomedical Engineering, vol. 64, no. 7, pp. 1558–1567, 2016.

[214] S. Zagoruyko and N. Komodakis, "Wide residual networks," in British Machine Vision Conference 2016. British Machine Vision Association, 2016.

[215] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," arXiv preprint arXiv:2009.06732, 2020.

[216] J. Beal, E. Kim, E. Tzeng, D. H. Park, A. Zhai, and D. Kislyuk, "Toward transformer-based object detection," arXiv preprint arXiv:2012.09958, 2020.

[217] M. Dobko, D.-I. Kolinko, O. Viniavskyi, and Y. Yelisieiev, "Combining cnns with transformer for multimodal 3d mri brain tumor segmentation with self-supervised pretraining," arXiv preprint arXiv:2110.07919, 2021.

[218] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE International Conference on Computer Vision, 2021, pp. 10 012–10 022.

[219] P. Zhang, X. Dai, J. Yang, B. Xiao, L. Yuan, L. Zhang, and J. Gao, "Multi-scale vision longformer: A new vision transformer for high-resolution image encoding," in International Conference on Computer Vision, October 2021, pp. 2998–3008.

[220] D.-J. Chen, H.-Y. Hsieh, and T.-L. Liu, "Adaptive image transformer for one-shot object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 12 247–12 256.