Deep learning for segmentation of brain tumors and organs at risk in radiotherapy planning

Pawel Mlynarski

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

THÈSE DE DOCTORAT

Apprentissage profond pour la segmentation des tumeurs cérébrales et des organes à risque en radiothérapie

Pawel MLYNARSKI

Inria Sophia Antipolis - Méditerranée, équipe Epione

Présentée en vue de l'obtention du grade de docteur en Automatique, Traitement du Signal et des Images **d'**Université Côte d'Azur

Dirigée par : Nicholas Ayache
Co-encadrée par : Hervé Delingette

Soutenue le : 15/11/2019

Devant le jury, composé de :
Nicholas Ayache, Inria Sophia Antipolis
Isabelle Bloch, Télécom ParisTech
Pierre-Yves Bondiau, Centre Antoine Lacassagne
Hervé Delingette, Inria Sophia Antipolis
Bjoern Menze, Technical University of Munich
Nikos Paragios, CentraleSupélec

Inria

Microsoft Research - Inria JOINT CENTRE

Apprentissage profond pour la segmentation des tumeurs cérébrales et des organes à risque en radiothérapie

Deep Learning for Segmentation of Brain Tumors and Organs at Risk in Radiotherapy Planning

Jury :

Rapporteurs
Isabelle Bloch, Télécom ParisTech
Bjoern Menze, Technical University of Munich

Examinateurs
Nikos Paragios, CentraleSupélec
Pierre-Yves Bondiau, Centre Antoine Lacassagne

Directeur de thèse
Nicholas Ayache, Inria Sophia Antipolis

Co-directeur de thèse
Hervé Delingette, Inria Sophia Antipolis

Deep Learning for Segmentation of Brain Tumors and Organs at Risk in Radiotherapy Planning

Abstract: Medical images play an important role in cancer diagnosis and treatment. Oncologists analyze images to determine the different characteristics of the cancer, to plan the therapy and to observe the evolution of the disease. The objective of this thesis is to propose efficient methods for automatic segmentation of brain tumors and organs at risk in the context of radiotherapy planning, using Magnetic Resonance (MR) images.

First, we focus on segmentation of brain tumors using Convolutional Neural Networks (CNN) trained on MRIs manually segmented by experts. We propose a segmentation model having a large 3D receptive field while being efficient in terms of computational complexity, based on combination of 2D and 3D CNNs. We also address problems related to the joint use of several MRI sequences (T1, T2, FLAIR).

Second, we introduce a segmentation model which is trained using weakly-annotated images in addition to fully-annotated images (with voxelwise labels), which are usually available in very limited quantities due to their cost. We show that this mixed level of supervision considerably improves the segmentation accuracy when the number of fully-annotated images is limited.

Finally, we propose a methodology for an anatomically consistent segmentation of organs at risk in the context of radiotherapy of brain tumors. The segmentations produced by our system on a set of MRIs acquired in the Centre Antoine Lacassagne (Nice, France) are evaluated by an experienced radiotherapist.

Keywords: Convolutional Neural Networks, semi-supervised learning, MRI, radiotherapy, brain tumor, organs at risk

Apprentissage profond pour la segmentation des tumeurs cérébrales et des organes à risque en radiothérapie

Résumé: Les images médicales jouent un rôle important dans le diagnostic et la prise en charge des cancers. Les oncologues analysent des images pour déterminer les différentes caractéristiques de la tumeur, pour proposer un traitement adapté et suivre l'évolution de la maladie. L'objectif de cette thèse est de proposer des méthodes efficaces de segmentation automatique des tumeurs cérébrales et des organes à risque dans le contexte de la radiothérapie, à partir des images de résonance magnétique (IRM).
Premièrement, nous nous intéressons à la segmentation des tumeurs cérébrales en utilisant des réseaux neuronaux convolutifs entrainés sur des IRM segmentés par des experts. Nous proposons un modèle de segmentation ayant un grand champ récepteur 3D tout en étant efficace en termes de complexité de calcul, en combinant des réseaux neuronaux convolutifs 2D et 3D. Nous abordons aussi les problèmes liés à l'utilisation conjointe des différentes séquences IRM (T1, T2, FLAIR).
Nous introduisons ensuite un modèle de segmentation qui est entrainé avec des images faiblement annotées en complément des images segmentées, souvent disponibles en quantités très limitées du fait de leur coût. Nous montrons que ce niveau mixte de supervision améliore considérablement la performance de segmentation quand le nombre d'images entièrement annotées est limité.
Finalement, nous proposons une méthodologie pour segmenter, de manière cohérente anatomiquement, les organes à risque dans le contexte de la radiothérapie des tumeurs cérébrales. Les segmentations produites par notre système sur un ensemble d'IRM acquis dans le Centre Antoine Lacassagne (Nice) sont évaluées par un radiothérapeute expérimenté.

Mots clés: Réseau neuronal convolutif, apprentissage semi-supervisé, IRM, radiothéraphie, tumeur cérébrale, organes à risque

Acknowledgements

I thank my supervisors, Nicholas Ayache and Hervé Delingette, for these years of working together and for giving me the opportunity to be part of Epione, an excellent research team. Thank you for supervising my PhD during which I learned a lot, both from methodological and technical point of view.

I would like to thank the reviewers of my thesis, Isabelle Bloch and Bjoern Menze, for their time and their constructive remarks on the manuscript. I thank Nikos Paragios and Pierre-Yves Bondiau for accepting to be members of my jury.

I would like to thank Antonio Criminisi for our scientific discussions and for giving me the opportunity to visit Microsoft Research in Cambridge, which was a great experience.

Thanks to Pierre-Yves Bondiau and Hamza Alghamdi, from the Centre Antoine Lacassagne, for launching the interesting project of segmentation of organs at risk. I was very happy to work with you and to visit the Centre Antoine Lacassagne. I thank Hamza for the clinical evaluation of segmentation results, which considerably enriched this research work.

I am grateful to Microsoft for funding my thesis, for our cooperation and for inviting me to an interesting summer school in Cambridge.

I would like to thank all Inria staff for the professional work, in particular the IT service which often helped me to solve tricky technical problems related to installation of libraries and all mysteries of Linux.

I thank all members of our team Epione and all nice people I met in Inria. Thank you for all the good moments and for making Inria Sophia such a nice place. For our coffee breaks, beach volley afterworks, tournaments of table football (also called *futbolín*), gym sessions and deep philosophical discussions in the famous bus 230 Nice-Sophia. Special thanks to our futbolín team including such stars as Clément, Luigi, Jaume, Benoît, Marco M (also known as 'The Legend'), Marco L, Nicolas G, Bastien and Santiago. To Raphaël, for sharing our office for 3 years and half, answering my numerous questions and being so nice. To all interns, PhDs, postdocs and engineers of Epione: Marc-Michel, Roch, Tania, Nicolas C, Loïc C, Fanny, Wilhelm, Gaëtan, Qiao, Julian, Sofia, Shuman, Yann, Rocio, Nina, Mehdi, Wen, Zihao, Bishesh and many others.

Special thanks to all my friends and my family for the support during my PhD and all the great moments we shared.

In particular, I would like to thank my mother, Jadwiga, who has always supported me. Thank you for everything.

Pawel

Contents

Chapter 1

Introduction

Contents

The objective of this chapter is to situate the thesis in its clinical and technical context. First, we discuss aspects that are important to understand the clinical motivations of the proposed methods. We start by defining brain tumors, which are the main focus of this thesis. We then present the commonly used types of medical images and we discuss the role of segmentation tasks in neuro-oncology.
Second, we discuss important aspects related to deep learning, which forms the basis of most of the current state-of-the-art methods for image segmentation. In particular, we discuss its main advantages and inconvenients in the context of segmentation of brain tumors and organs at risk.
Finally, we introduce our main contributions and we present the organization of the thesis.

1.1 Clinical context

1.1.1 Brain tumors

Cancer is a life-threatening disease involving an abnormal proliferation of cells. It originates from one cell which developped several characteristics, often called *hallmarks of cancer* [Hanahan 2000, Hanahan 2011]. In particular, cancer cells are able to replicate infinitely and autonomously, they invade other tissues and ignore natural regulatory mechanisms such as programmed cell death. The capacity of invading neighboring tissues and spreading to distant sites of the body (metastasis) distinguishes malignant tumors (cancers) from the benign ones. Every year, approximately 9 millions people in the world die from different forms of cancer.

In this thesis, we focus on neuro-oncology. Brain tumors include several types of primary and secondary tumors which develop within the brain region.

Most brain cancers are secondary [Board 2018], i.e. corresponding to metastases of primary cancers which developped outside the brain, in particular the lung cancer. The most common types of primary brain tumors [Mehta 2011] are gliomas [Schwartzbaum 2006] and meningiomas [Wiemels 2010]. Meningiomas originate from meninges, which are membranes protecting the brain and the spinal cord. Most of them are benign and they can often be cured by surgery (if needed). Gliomas originate from glial cells, which are one of the two main components of the nervous tissue (with neurons) and who have several functions related to support and protection of neurons. Gliomas represent approximately 80 % of primary malignant brain tumors and their malignants forms, such as glioblastoma, are among the most aggressive cancers. The first two chapters of this thesis are related to segmentation of gliomas, using a publicly available database of the Multimodal Brain Tumor Segmentation Challenge (BRATS) [Menze 2015, Bakas 2017].

1.1.2 Medical imaging in neuro-oncology

Medical images are extensively used in oncology for diagnosis, therapy planning and monitoring of tumors. Oncologists analyze images to locate tumors and assess their different characteristics.
Different types of medical images are used, depending on the task (search of metastases, radiotherapy planning) and the region of interest (brain, lungs, digestive system). The commonly used types of imaging include computed tomography (CT), magnetic resonance imaging (MRI) and positron-emission tomography (PET).

Positron-emission tomography [Gambhir 2002] is based on injection of a radioactive tracer in the blood of the patient in order to observe the metabolism of different tissues. A commonly used tracer is fludeoxyglucose which is a structural analog of glucose. As cancer cells need an important glucose supply due to their divisions, the tumoral tissues may be detected by their abundant absorption of the radioactive tracer. PET scan is particularly useful for diagnosis and staging of tumors, for detecting cancer metastases and monitoring effects of a therapy. However, due to physical limitations, PET scans have usually a considerably lower spatial resolution than MRI and CT scans.

Computed tomography [Hsieh 2009] measures the absorption of X-rays of different tissues in the body. The radiation is emmited from different angles in order to acquire a series of 2D radiographic images from which a 3D scan is then reconstructed. Even if CT scans have generally a better spatial resolution than MRI, they offer a significantly weaker contrast between soft tissues such as the ones present in the brain. Morever, the exposure to X-rays may induce cancers by damaging DNA of body cells.

Acquisition of MRI [Atlas 2009] is based on the detection of signals emitted by the nuclear magnetic resonance of atoms in the body. The detected signal

is usually produced by protons of hydrogen, present in abundance in the human body (water, fat). The atoms are set in a strong magnetic field and are then perturbed by a radio wave, called *pulse sequence*. By modifying the parameters of the pulse sequence and pulsed field gradients, different contrasts are obtained, corresponding to specific *MRI sequences*. The MRI sequences commonly used for brain tumor imaging are T1, T2 and FLAIR (T2 with suppression of fluids). T1 is often acquired after injection of a gadolinium-based contrast agent [Zhou 2013] in the blood of the patient, in particular to highlight the tumor angiogenesis, i.e. creation of new vascular networks by the tumor.
Magnetic resonance images are particularly suitable for imaging of brain tumors and organs. In particular, they offer a high contrast between soft tissues in the brain (compared to other types of imaging) and the use of different MRI sequences offers the possibility to highlight different tumoral compartments (edema, tumor vascularisation, necrosis).

1.1.3 Radiotherapy planning and organs at risk

Treatment of brain tumors often includes radiotherapy [Khan 2014], which uses a ionizing radiation to kill cancer cells or to stop their division by damaging their DNA. The most common type of radiotherapy is external beam radiotherapy, in which the radiation is emitted from the exterior of the patient.

Radiotherapy planning is a particularly important application of automatic segmentation. The objective of radiotherapy planing is to compute optimal irradiation doses, i.e. to deliver a radiation which destroys tumoral cells while sparing healthy structures. Computation of the irradiation doses (with a dedicated software) requires an accurate segmentation of target volumes, containing cancer cells, and a large number of healthy structures which may be damaged by the therapy [Scoccianti 2015]. The target volumes include the gross tumor volume (GTV) and the clinical target volume (CTV). The GTV corresponds to the visible part of the tumor. The CTV includes the GTV and a region which is likely to contain cancer cells which could not be imaged with the current technologies such as MRI. The planning target volume (PTV) is a margin around the CTV which ensures the delivery of the necessary irradiation dose to the CTV. A similar margin is usually considered around the organs at risk. Figure 1.1 displays an example of the different volumes segmented during the radiotherapy planning, on a real clinical case from the Centre Antoine Lacassagne (Nice, France).
The segmentation process requires medical expertise and takes typically several hours per patient for an experienced clinican. It represents therefore a considerable cost and eventually delays the therapy.

The objective of this thesis is to propose efficient methods for segmentation tasks in neuro-oncology. Two chapters of this thesis are related to segmentation of

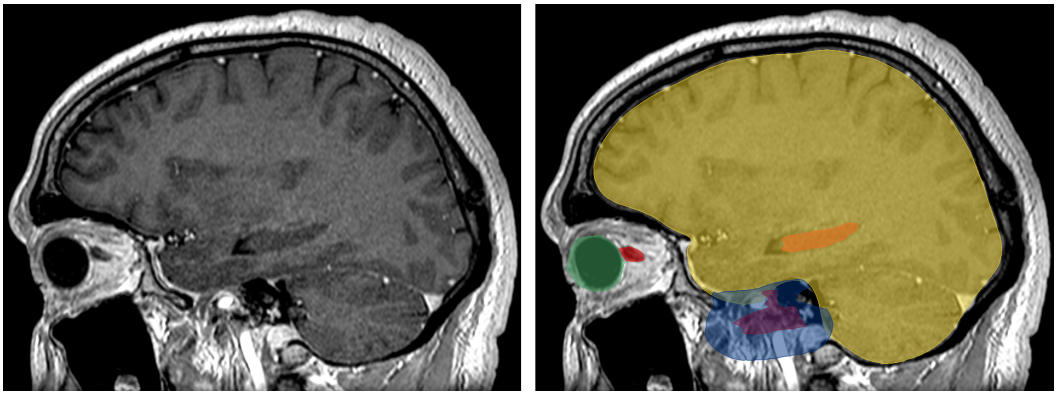

Figure 1.1: Segmentation of the target volumes and organs at risk in radiotherapy planning. Left: T1-weighted MRI acquired after injection of a gadolinium-based contrast agent. Right: manual delineation of the therapy targets and anatomical structures which should not be irradiated excessively. The displayed organs at risk are respectively the eye (green), the optic nerve (red), the hippocampus (orange) and the brain (yellow). The blue and magenta regions correspond respectively to the clinical target volume (CTV) and the gross tumor volume (GTV).

brain tumors, using the database of the Multimodal Brain Tumor Segmentation Challenge (BRATS). Automatic segmentation of organs at risk is addressed in the third main chapter of this thesis.

1.2 Deep learning in medical imaging

The methods presented in this thesis are mainly based on deep learning, which is a branch of machine learning. In this section, we briefly present the general principles of deep learning, we motivate its use for segmentation tasks in neuro-oncology and we discuss its limitations, some of which are addressed in this thesis.

Given an input space X and a label space Y, the objective of supervised machine learning is to find a predictive function $f : X \to Y$, using a database of training examples (x_i, y_i), where $x_i \in X$ and $y_i \in Y$. To achieve this goal, three main elements have to be defined:

- Family of candidate functions f_θ, parametrized by a vector of parameters $\theta \in \Theta$
- Loss function $L : \Theta \to \mathbb{R}$, which quantifies the mismatch between the outputs predicted by a candidate function f_θ and the ground truth.
- Training algorithm, which minimizes the loss function (with respect to the parameters θ) over the training data

The main particularity of deep learning is the nature of the considered candidate functions. The term *deep* is related to multiple compositions of functions. The considered composed functions are differentiable and organized in layers, with the idea to progressively transform the input vector, extracting more and more complex information. The term *neural network* is related to the considered family of functions, represented typically by a graph. Training of the model (minimization of the loss function) is typically based on iterative optimization with variants of the stochastic gradient descent.

Convolutional Neural Networks (CNN) [LeCun 1995] are a commonly used type of neural networks for image processing and analysis (classification, segmentation). They exploit spatial relations between pixels (or voxels, in 3D) and are based on application of local operations such convolution, pooling (maximum, average) and upsampling. The objectives of such design are to limit the number of parameters of the network and to limit computational costs, as images correspond generally to very large inputs. In fact, an important property of the operations used in CNNs is that they can be parallelized and efficiently computed on a Graphical Processing Unit (GPU).
CNNs for image segmentation are usually trained in an end-to-end manner, i.e. their input is the image and the output is the segmentation. With an end-to-end training, the model automatically learns to extract relevant information from images, using the training database. This property is particularly important for very challenging tasks in medical imaging, such as tumor segmentation. Most of the current state-of-the-art methods for image segmentation are based on CNNs, in particular the methods for brain tumor segmentation [Kamnitsas 2017a, Myronenko 2018].

However, even if CNNs have recently obtained state-of-the-art results in many recognition tasks, they still have important limitations in the context of segmentation in medical imaging. The objective of methodological contributions of this thesis is to address these limitations.
Despite the progress of GPU capacities, computational costs still severely limit the potential of CNNs for segmentation tasks in medical imaging. A typical segmentation network, such as U-net [Ronneberger 2015] performs thousands of convolutions, max-poolings and upsamplings. Outputs of these operations have to be stored in the memory of the GPU during each iteration of the the training, in order to compute gradients of the loss function by the Backpropagation algorithm. A typical MRI is composed of several millions of voxels. Training of neural networks for an end-to-end segmentation on entire MRIs requires therefore a huge amount of GPU memory and is often impossible using the currently available GPUs. For this reason, current segmentation models are usually trained on subvolumes of limited size and have limited receptive fields.

Another important problem is the cost of the ground truth annotations necessary to train neural networks, and machine learning models in general. Manual

segmentation of tumors is particularly costly as it is not only time-consuming but also requires medical exprtise and therefore has to be performed by experienced clinicians.

Other difficulties are related to the use of multimodal data. Usually, different types of images (e.g. different MRI sequences) are used in oncology. Most of the current CNN-based models consider MRIs as 4D tensors and assume presence of all modalities for all patients in the training database, which is rarely the case in practice.

Finally, commonly used segmentation CNNs may produce spatially inconsistent results, as they are based on individual classification of voxels given their receptive fields. It means that, in general, the model does not explicitely analyze aspects related, for instance, to the connectivity of the output segmentation and the spatial relations between the different segmentation classes.

1.3 Thesis overview

The objective of this thesis is to propose efficient methods for segmentation of brain tumors and organs at risk in radiotherapy planning. The three main chapters correspond to journal articles which have been published or submitted during the preparation of the thesis. The manuscript is organized as follows.

In chapter 2, published as a journal article [Mlynarski 2019b], we introduce a CNN-based system for brain tumor segmentation which addresses two important problems of current deep learning models. First, we propose a methodology to obtain a large 3D receptive field of a segmentation model without requiring an excessive computational load. The main idea of our approach is to use features learned by segmentation networks (representing rich information and capturing a large spatial context) as an additional input of another segmentation network. Second, we address the problem of missing image modalities (in particular, the different MRI sequences) in training databases, by proposing a model architecture with modality-specific subnetworks. Our method was tested on a publicly available database of the BRATS 2017 challenge and obtained one of the best performances of the challenge.

In chapter 3, corresponding to the second journal article [Mlynarski 2019c], we exploit the use of less costly forms of annotations to train segmentation networks for brain tumor segmentation. We assume that the training database contains a small number of segmented images and a large number of images with global labels, simply indicating presence or absence of a tumor tissue within the

image (without any information on the location of the tumor, if present). This setting represents therefore a mixed level of supervision and differs from the standard semi-supervised learning as it uses weakly-annotated data rather than totally unlabelled data. The main idea of our approach is to extend segmentation networks with a branch performing image-level classification of tumors and to train the model for the two tasks jointly, using the two types of training images. To assess the effects of using this mixed level of supervision, we perform a series of cross-validated experiments on the database of the BRATS challenge. We show that the mixed supervision significantly improves segmentation accuracy compared to the standard supervised learning. The improvement is proportional to the ratio between the numbers of weakly-annotated and fully-annotated images.

In chapter 4, corresponding to the third journal article [Mlynarski 2019a], we propose a CNN-based system for an anatomically consistent segmentation of organs at risk in the context of radiotherapy planning. This work is done in cooperation with Centre Antoine Lacassagne in Nice, France. First, we propose a methodology to train neural networks for segmenting multiple and non-exclusive classes, where one voxel may belong to zero or several classes (in contrast to standard segmentation problems, where each voxel is assigned one, unique class). In particular we address problems related to computational costs and missing annotations of different classes, resulting from the fact that the ground truth segmentation of one anatomical structure is usually available only for a subset of patients, as the different structures are segmented according to clinical needs. Then, we propose procedures to enforce the anatomical consistency of the segmentation in a postprocessing stage. In particular, we propose a graph-based algorithm for segmentation of the optic nerves, which are among the most challenging organs for automatic segmentation. Our method is tested on clinical data acquired in the Centre Antoine Lacassagne. In particular, the segmentations produced by our system are evaluated by an experienced radiotherapist on a set of 50 non-annotated MRIs, for several anatomical structures in the brain region. A large majority of output segmentations were found acceptable for radiotherapy planning.

Finally, in chapter 5, we summarize the contributions of the thesis and we propose directions for future research works.

Chapter 2

3D Convolutional Neural Networks for Tumor Segmentation using Long-range 2D Context

Contents

In this first main chapter, published as a journal article [Mlynarski 2019b], we focus on segmentation of brain tumors using Convolutional Neutal Networks trained on manually segmented images. We focus mainly on the notion of receptive field of segmentation networks but we also address the problem of training of neural networks on databases containing cases with missing image modalities. Morever, a voting strategy is proposed to combine multiclass segmentations produced by several models, in order to improve the robustness of the system. Aspects related to mathematical optimization (the training algorithm) are also discussed. Our method is evaluated on a publicly available database from the BRATS challenge.

2.1 Introduction

Gliomas are the most frequent primary brain tumors and represent approximatively 80% of primary malignant brain tumors [Goodenberger 2012]. They originate from

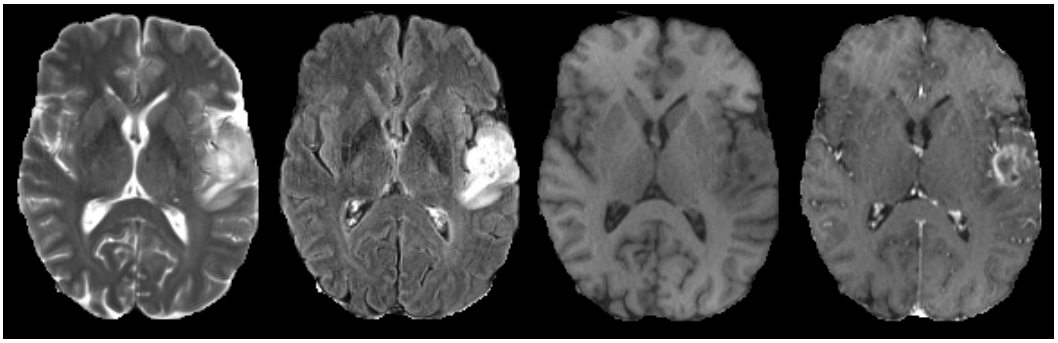

Figure 2.1: Multisequence MR scan of a patient suffering from a glioblastoma. From left to right: T2-weighted, FLAIR, T1-weighted, post-contrast T1-weighted.

glial cells of the brain or the spine and can be classified according to the cell type, the grade and the location. High grade gliomas (grades III and IV) are associated with a particularly poor prognosis: patients diagnosed with glioblastoma multiforme survive on average 12-14 months under therapy. Medical images such as MRI [Bauer 2013] are used for diagnosis, therapy planning and monitoring of gliomas.

Different tumor tissues (necrotic core, active rim, edema) can be imaged using multiple MR sequences. For instance, T2-FLAIR sequence is suitable for detecting edema while T1-weighted MR images acquired after the injection of a gadolinium-based contrast agent are suitable to detect active parts of the tumor core (Fig. 2.1). These tumor tissues may be treated with different therapies [Gillies 2015] and their analysis is important to assess the tumor characteristics, in particular its malignity.

Manual segmentation of tumors is a challenging and time-consuming task. Moreover, there is a significant variability between segmentations produced by human experts. An accurate automatic segmentation method could help in therapy planning and in monitoring of the tumor progression by providing the exact localization of tumor subregions and by precisely quantifying their volume.

Tumor variability in location, size and shape makes it difficult to use probabilistic priors. Image intensities of voxels representing tumor tissues in MR images highly overlap with intensities of other pathologies or healthy structures. Furthermore, ranges of MR image intensities highly vary from one imaging center to another depending on the acquisition system and the clinical protocol. Due to these aspects, in order to determine the presence of a tumor at a given position, high-level contextual information has to be analyzed.

A large variety of methods have been proposed for multiclass tumor segmentation. In 2012, the Multimodal Brain Tumor Segmentation Challenge (BRATS) [Menze 2015, Bakas 2017] was launched. The first group of methods corresponds to generative models based on the registration of the patient scan to a brain atlas providing a spatially varying probabilistic prior of different tissues. In the method of Prastawa et al [Prastawa 2004], tumor segmentation is guided by differences between

the patient scan and the atlas of healthy brain. One limitation of this approach is the fact that it ignores the mass effect (deformation of neighboring healthy structures) caused by the tumor, which can lead to incorrect registration. In methods such as GLISTR [Gooya 2012] or [Kwon 2014], the authors propose to modify a healthy atlas by using tumor growth models and to perform a joint segmentation and registration to a modified brain atlas. These methods have the advantage of taking into account the characterics of tumors, however the use of tumor growth models comes with an additional complexity and the estimation of the number of tumor seeds is non trivial. A multi-atlas method, based on the search of similar image patches, was also proposed by Cordier et al [Cordier 2016].

Promising results were obtained by discriminative models corresponding to voxelwise classifiers such as SVM [Bauer 2011, Lee 2005] or Random Forests [Ho 1995, Zikic 2012, Geremia 2012, Le Folgoc 2016, Bauer 2012, Tustison 2015]. For instance, Geremia et al [Geremia 2012] propose to classify each voxel of a multimodal MR brain image by a random forest using features capturing information from neighbooring voxels and from distant regions such as the symmetric part of the brain. More recently, Le Folgoc et al proposed Lifted Auto-Context Forests [Le Folgoc 2016], an efficient method based on cascaded Random Forests progressively segmenting tumor subclasses exploiting the semantics of labels.

In recent years, Convolutional Neural Networks [LeCun 1995] achieved state-of-the-art results in many tasks of image classification [He 2016, Krizhevsky 2012, Simonyan 2014], detection [Sermanet 2013] and segmentation [Long 2015, Chen 2014]. In particular, the representation learning ability of CNNs is a considerable advantage for the task of tumor segmentation, where the design of discriminant image features is non trivial. The CNN-based methods of Pereira et al [Pereira 2015] and Kamnitsas et al [Kamnitsas 2016] obtained respectively the best performance in BRATS 2015 and BRATS 2016 challenges. Fully-convolutional neural networks [Long 2015, Ronneberger 2015, Havaei 2017, Zheng 2018] were used in most state-of-the-art segmentation methods, in particular, recently we observe a particular interest for 3D fully-convolutional neural networks [Dou 2017, Çiçek 2016, Kamnitsas 2017a, Wang 2017, Isensee 2017]. Many methods include postprocessing steps, often based on Conditional Random Fields [Lafferty 2001] or mathematical morphology [Serra 2012].

Despite promising results obtained by these methods, segmentation of tumors in large medical images is still a very challenging task. One of the main drawbacks of CNNs is their computational cost resulting from application of thousands of costly operations (convolutions, poolings, upsamplings) on input images. This aspect is particularly problematic for segmentation problems in large medical images such as MRI or CT scans. Despite the variety of proposed neural network architectures, current CNN-based systems struggle to capture a large 3D context from input images. Moreover, most methods implicitly assume the presence of all MR sequences for all patients and the correct registration between sequences whereas these conditions do not necessarily hold in practice.

In this work, we propose an efficient system based on a 2D-3D model in which features extracted by 2D CNNs (capturing a rich information from a long-range 2D context in three orthogonal directions) are used as an additional input to a 3D CNN.

We propose a 2D model (processing axial, coronal or sagittal slices of the input image) in which we introduce an alternative approach for treating different MR sequences. In many CNNs, including the state-of-the-art deep learning models mentioned before, all channels of the input MR image are directly combined by the first convolutional layers of the network. We propose an architecture composed of modality-specific subnetworks (which can be trained independently) and of a joint part combining all input modalities. Such design allows to train one part of the network on images with missing MR sequences while also extracting a rich information resulting from the combination of all MR sequences.

We propose to use features learned by 2D CNNs as an additional input to a 3D CNN in order to capture rich information extracted from a very large spatial context while bypassing computational constraints. Such design considerably increases the size of the receptive field compared to standard 3D models taking as input only the raw intensities of voxels of a subvolume.

In order to combine the strengths of different network architectures, we introduce a voxelwise voting strategy to merge multiclass segmentations produced by several models. Finally, we designed a simple and stable training algorithm which is particularly well adapted for training large models.

We have evaluated our method on the challenging task of multiclass tumor segmentation of malignant brain tumors in multisequence MR images from the Validation set of BRATS 2017 challenge, using a public benchmark. In the performed experiments, our 2D-3D approach has outperformed the standard 3D model (where a CNN takes as input only the raw intensities of voxels of a subvolume) and our system has obtained promising results with median Dice scores of 0.918, 0.883 and 0.854 respectively for the three tumor subregions considered in the challenge (whole tumor, tumor core and contrast-enhancing core). Our method can be adapted to a large variety of multiclass segmentation tasks in medical imaging.

2.2 Methods

Our generic 2D-3D approach is illustrated on Fig. 2.2.

The main components of our method are described in the following. First, we introduce an efficient 2D-3D model with a long-range 3D receptive field. Second, we present our neural network architecture with modality-specific subnetworks. Loss functions and the optimization algorithm are presented in the third subsection. In order to be more robust to limitations of specific choices of neural network architectures, we propose a simple hierarchical decision process to merge multiclass segmentations produced by several models.

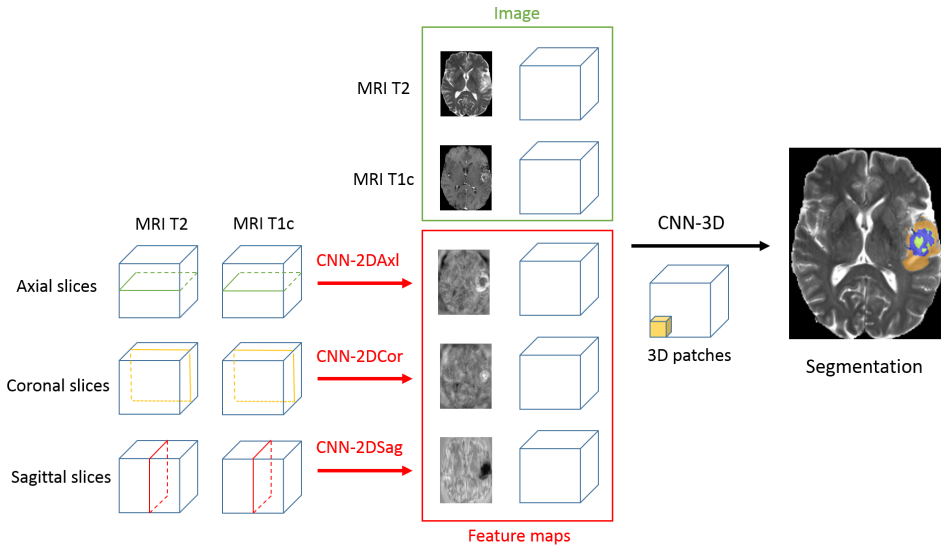

Figure 2.2: Illustration of our 2D-3D model. Features extracted by 2D CNNs (processing the image by axial, coronal and sagittal slices) are used as additional channels of the patch processed by a 3D CNN. As these features encode a rich information extracted from a large spatial context, their use significantly increases the size of the receptive field of the 3D model.

2.2.1 Spatial context and 3D models

A typical multisequence MR scan is composed of several millions of voxels. Convolutional neural networks transform input images by applying hundreds of convolutions and other operations whose outputs have to be stored in memory during iterations of the training in order to compute gradients of the loss by Backpropagation algorithm [Dreyfus 1990]. Training of CNNs requires typically dozens of thousands of iterations. Because of high computational costs of CNNs, large medical images are generally processed by subvolumes of limited size.

The obvious limitation of standard 2D approaches is to ignore one spatial dimension. However networks processing images by planes (axial, coronal or sagittal) have the ability to compare a studied voxel with distant voxels within the same plane and to capture a relevant information while keeping the input size reasonable. In the single-scale setting, the choice between the 2D and 3D option can therefore be seen as the choice between comparing distant voxels within the same plane (long-range 2D context) or comparing close voxels in three dimensions (short-range 3D context). Fig. 2.3 depicts the comparison of the information represented by a 2D patch of dimensions 125x125 and a 3D patch of dimensions 25x25x25 (both having the same number of voxels).

Another option is to process three orthogonal planes and classify the voxel at the intersection of three planes. This approach was successfully aplied by Ciompi et al. [Ciompi 2017] for the problem of classification of lung nodules. The system proposed by the authors is composed of 9 streams processing 2D patches in three orthogonal

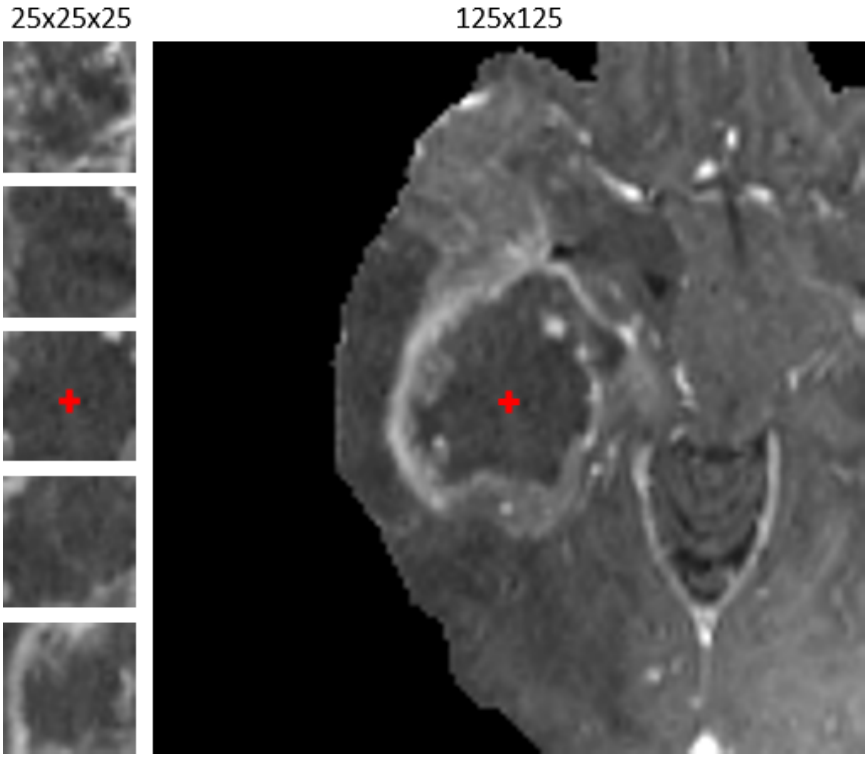

Figure 2.3: Comparison of information represented by a 25x25x25 patch (left: 5 slices shown) and a 125x125 axial 2D patch centered at the same point. While both patches have the same number of voxels, the spatial context is considerably different. While the first patch captures local 3D shapes, the second patch captures information from distant points within the same plane.

planes centered at a givel voxel and at three different scales. The streams are then combined by fully-connected layers with the last layer performing classification. However, CNN-based systems with fully-connected layers are computationally less efficient for the segmentation task compared to fully-convolutional networks such as U-Net, that classify simultaneously several neighboring voxels and take advantage of shared computations. On modern GPUs, fully-convolutional networks are able to classify hundreds of thousands of voxels in each iteration of the training.

A larger 3D context can be analyzed by extracting multiscale 3D patches as in Deep Medic [Kamnitsas 2016], a state-of-the-art CNN-based system which processes two-scale 3D patches by two streams of convolutional layers. The main characteristic of this design is the separate processing at two scales. A more global information is captured by the stream processing the patch from the image downsampled by a factor 3. However, this global information is not of the same nature as the one extracted by U-net [Ronneberger 2015] in which it results from a long sequence of convolutions and max-poolings starting from the original scale of the image (from local and low-level information to global and high-level information). A possible limitation of the model is its sequential aspect: the only concatenation is before the

two last hidden layers of the network whereas skip-connections seem to improve the performance of neural networks [He 2016].

The idea of our 2D-3D approach is to take into account a very large 3D context by using features learned by 2D networks rather than simply processing downsampled versions of the input image. In fact, features learned by 2D CNNs encode a rich information extracted from a large spatial context and the use of these features allows to considerably increase the size of the receptive field of the model.

In our method we use fully-convolutional neural networks [Long 2015]. A network processes the input image by a sequence of spatially-invariant transformations in order to output voxelwise classification scores for all classes. The outputs of transformations at the same level of processing form a *layer* which can be seen as a multi-channel image when arranged in a grid as in commonly used deep learning libraries such as Theano [Bergstra 2010] or TensorFlow [Abadi 2016]. In 3D CNNs, each layer of the network corresponds to a multi-channel image with three spatial coordinates. A convolutional layer whose number of feature maps is equal to the number of classes and whose ouput is penalized during the training is called *classification layer*. The channels of a layer are called *feature maps* whose points represent *neurons*. The set of voxels in the input layer which are taken into account in the computation of the output of a given neuron is called the *receptive field* of the neuron.

Our 2D-3D model (Fig. 2.4) is similar to 3D U-Net [Çiçek 2016] whose input is a 3D patch of a multimodal image along with a set of feature maps produced by networks trained on axial, coronal and sagittal slices (three versions of one 2D network). The extracted feature maps are concatenated to the input patch as additional channels. The network processes 3D patches of size 70x70x70 and has the receptive field of size 41x41x41. However, given that the network takes as input not only the raw intensities of voxels but also the values of features extracted by 2D

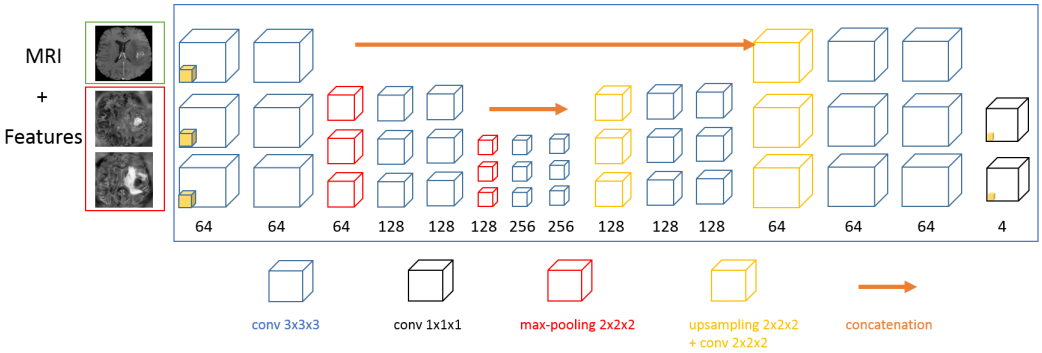

Figure 2.4: Architecture of the main 2D-3D model used in our experiments (named '2D-3D model A' in the remainder). The channels of the input 3D patch are all MR sequences and feature maps extracted by 2D CNNs. The number of feature maps in the last convolutional layer is equal to the number of classes (4 in our case).

neural networks analyzing a large spatial context, the effective receptive field of the 2D-3D model is strikingly larger. Each feature represents a semantic information extracted from a large patch in axial, coronal or sagittal plane. The model uses the values of these features computed for all voxels. Therefore, classification of one voxel is performed using not only the raw intensities of voxels within the surrounding 41x41x41 patch but also from all axial, coronal and sagittal planes passing by the voxels of this patch (Fig. 2.5). To the best of our knowledge, this is a novel way to capture a large 3D context with CNNs. The idea of using outputs of a CNN as additional input to another CNN was recently used for tumor segmentation in the work of Havaei et al [Havaei 2017], however the system proposed in [Havaei 2017] is significantly different from our 2D-3D approach, in particular as it processes the image by axial slices, considered independently from each other.

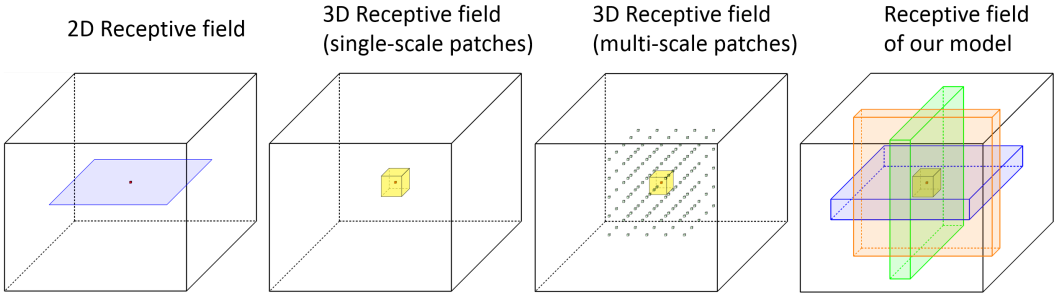

Figure 2.5: Illustration of the receptive field of our 2D-3D model and the comparison with other approaches. The use of features extracted by 2D CNNs significantly increases the size of the receptive field compared to standard 3D approches which only use raw intensities of voxels of a subvolume.

The steps of the training of our model are the following:

1. Train three versions of the 2D network respectively on axial, coronal and sagittal slices. We refer to these three versions respectively as CNN-2DAxl, CNN-2DCor and CNN-2DSag, according to the nature of the captured 2D context.

2. For all images of the training database, extract the learned features from final convolutional layers (without softmax normalization) of the 2D neural networks (CNN-2DAxl, CNN-2DCor and CNN-2DSag) and save their outputs in files.

3. Train the 3D model using the extracted 2D features as additional channels to the input image patches.

The choice of extracting features from the last convolutional layer is motivated by the fact that this layer has the largest receptive field and represents a semantic information while being composed of a small number of feature maps.

The two-step training (2D, then 3D) significantly reduces computational costs compared to an end-to-end training of the 2D-3D architecture. In each iteration of the training of the 3D CNN, the 2D features are already computed and there is therefore no need to store three 2D CNNs in the memory of a GPU. Moreover, the 2D networks can be trained in parallel on different GPUs.

2.2.2 2D model and modality-specific processing

Our generic 2D deep learning model performs segmentation of tumors in axial, coronal or sagittal slices of a multisequence MRI. Our model is similar to U-net [Ronneberger 2015] in which we introduce a system of co-trained subnetworks processing different input MR sequences (Fig. 4.2). This design can be seen as a hybrid approach in which one part of the network processes independently different MR sequences and another part extracts features resulting from the combination of all sequences. Independent processing of input channels has the considerable advantage of being more robust to missing data. On the other hand, models using data from all input channels can extract important information resulting from relations between channels and therefore are likely to obtain better segmentation performance. Our goal is to combine these two aspects.

Given an input image with K channels, we consider K+1 subnetworks: one subnetwork per input channel and one subnetwork directly combining all channels. The subnetworks learn therefore features specific to each MR sequence (except the last subnetwork which learns features related to the direct combination of sequences) and can be trained on images with missing MR sequences.

During the training phase we attach a classification layer to each subnetwork: more precisely, if a subnetwork has n layers, then during the training phase we add one convolutional layer whose number of feature maps is equal to the number of classes and whose input is the n^{th} layer of the subnetwork. The outputs of these additional layers, that we call *auxiliary classification layers*, are penalized during the training in order to force the subnetworks to extract the most pertinent information from each MR sequence. If the training database contains images with missing MR sequences, each modality-specific subnetwork can be pretrained independently of the others, on images for which the given MR sequence is provided. During the test phase, the auxiliary classification layers are ignored. The idea of using of intermediate losses to perform *deep supervision* was succesfully used in the method of Dou et al [Dou 2017] for the problems of liver segmentation and vessel segmentation in 3D medical images.

Final convolutional layers of the subnetworks are concatenated and fed to the main part of the network similar to U-net [Ronneberger 2015]. The main network is composed of two sections connected by concatenations of feature maps between layers at the same scale. The downsampling section is composed of convolutions and max-poolings. The upsampling section is composed of bilinear upsamplings, convolutions and concatenations with feature maps from the downsampling part.

If the training database contains cases with missing modalities, the steps of the

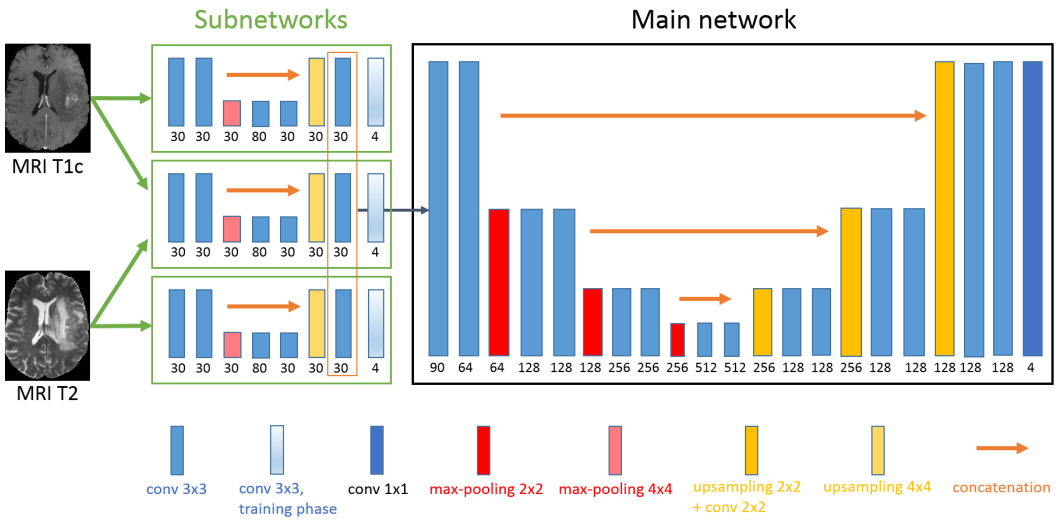

Figure 2.6: Architecture of the main 2D model used in our experiments (named '2D model 1' in the remainder). The numbers of feature maps are specified below rectangles representing layers. In each subnetwork the first layer is concatenated to an upsampling layer in order to combine local and global information. Each subnetwork learns features specific to one image modality, except one subnetwork which directly combines all modalities. The classification layers of subnetworks are ignored during the test phase. For clarity purposes, we display the case with two MR sequences. The modality-specific subnetworks (top-left and bottom-left rectangles) can be pretrained independently as they are separated and have a different input.

training are the following:

1. Train each modality-specific subnetwork on images for which its input modality (e.g. MRI T2-FLAIR) is provided.

2. Train the entire network on images for which all modalities provided.

During the test phase, we assume that all modalities are provided. The segmentation is produced by the main part of the network.

2.2.3 Training of the model

2.2.3.1 Loss functions and dealing with class imbalance

To train our models, we use a weighted cross-entropy loss. In the 3D case, given a training batch b and the estimated model parameters θ, the loss function penalizes the output of the classification layer:

$$Loss_b^{3D}(\theta) = -\sum_{i=1}^{|b|}\sum_{(x,y,z)}\sum_{c=0}^{C-1}\delta(G^{i,b}_{(x,y,z)},c)W_{c,b}\log(p^c_{i,(x,y,z)}(\theta)) \tag{2.1}$$

where δ denotes the Kronecker delta, $W_{c,b}$ is a voxelwise weight of the class c for the batch b, $p^c_{i,(x,y,z)}(\theta)$ is the classification softmax score given by the network to the class c for the voxel at the position (x,y,z) in the i^{th} image of the batch and $G^{i,b}_{(x,y,z)}$ is the ground truth class of this voxel. The purpose of using weights is to counter the problem of severe class imbalance, tumor subclasses being considerably under-represented. In contrast to common approaches, the voxelwise weights are set automatically depending on the composition of the batch (number of examples of each class greatly varies accross batches). We suppose that in each training batch there is at least one voxel of each class. Let's note C the number of classes and N^c_b the number of voxels of the class c in the batch b. For each class c we set a target weight t_c with $0 \leq t_c \leq 1$ and $\sum_{c=0}^{C-1} t_c = 1$. Then all voxels of the class c are assigned the weight $W_{c,b} = t_c/N^c_b$ so that the total sum of their weights accounts for the proportion t_c of the loss function. To better understand the effect of this parameter, note that in the standard non-weighted cross-entropy each voxel has a weight of 1 and the total weight of the class c is proportional to the number of voxels labeled c. It implies that setting a target weight t_c larger than the proportion of voxels labeled c increases the total weight of the class c (favoring its sensitivity) and conversely.

The same strategy is applied in the 2D case, for each classification layer of the model. The final loss of the 2D model is a convex combination of all intermediate losses, associated respectively with the main network and all subnetworks:

$$Loss^{2D}_b(\theta) = c^{main} Loss^{main}_b(\theta) + \sum_{k=1}^{K+1} c^k Loss^k_b(\theta) \tag{2.2}$$

where K is the number of input channels, $0 \leq c^{main} \leq 1$, $0 \leq c^k \leq 1 \; \forall \; k \in [1..K+1]$ and $c^{main} + \sum_{k=1}^{K+1} c^k = 1$.

2.2.3.2 Training algorithm

Our training algorithm is a modified version of Stochastic Gradient Descent (SGD) with momentum [Rumelhart 1988]. In each iteration of the standard SGD with momentum, the loss is computed on one batch b of training examples and the vector v of updates is computed as a linear combination of the previous update and the gradient of the current loss with respect to the parameters of the network: $v^{t+1} = \mu v^t - \alpha_t \nabla Loss_b(\theta^t)$ where θ^t are the current parameters of the network, μ is the momentum and α_t is the current learning rate. The parameters of the network are then updated: $\theta^{t+1} = \theta^t + v^{t+1}$. We apply two main modifications to this scheme.

First, in each iteration of the training, we minimize the loss over several training batches in order to take into account a large number of training examples while bypassing hardware constraints. In fact, due to GPU memory limits, backpropagation can only be performed on a training batch of limited size. For large models, training batches may be too small to correctly represent the training database, which would result in large oscillations of the loss and a difficult convergence. If we note

N the number of training batches per iteration, the loss at one iteration is given by $Loss^N(\theta) = \sum_{b=1}^{N} Loss_b(\theta)$ where $Loss_b(\theta)$ is the loss over one training batch. Given the linearity of derivatives, the gradient of this loss with respect to the parameters of the network is simply the sum of gradients of losses over the N training batches: $\nabla Loss^N(\theta) = \sum_{b=1}^{N} \nabla Loss_b(\theta)$. Each of the N gradients is computed by backpropagation.

The second modification is to divide the gradient by its norm. With the update rule of the standard SGD, strong gradients would cause too high updates of the parameters which can even result in the divergence of the training and numerical problems. Conversely, weak gradients would result in too small updates and then a very slow training. We want therefore to be independent of the magnitude of the gradient in order to guarantee a stable training. To summarize, our update vector v is computed as following:

$$v^{t+1} = \mu v^t - \alpha_t \frac{\nabla Loss^N(\theta^t)}{\|\nabla Loss^N(\theta^t)\|} \tag{2.3}$$

In order to converge to a local minimum, we decrease the learning rate automatically according to the observed convergence speed. We fix the initial value α_{init} and the minimal value α_{min} of the learning rate. After each F iterations we compute the mean loss accross the last $F/2$ iterations ($Loss_{current}$) and we compare it with the mean loss accross the previous $F/2$ iterations ($Loss_{previous}$) . We fix a threshold $0 < d_{loss} < 1$ on the relative decrease of the loss: if we observe $Loss_{current} > d_{loss} \times Loss_{previous}$ then the learning rate is updated as follows: $\alpha_{t+1} = max(\frac{\alpha_t}{2}, \alpha_{min})$. Given that the loss is expected to decrease slower with the progress of the training, the value of F is doubled when we observe an insufficient decrease of the loss two times in a row. For the training of our models we fixed $\alpha_{init} = 0.25$, $\alpha_{min} = 0.001$, $F = 200$ and $d_{loss} = 0.98$, i.e. initially we expect a 2% decrease of the loss every 200 iterations. The high values of the learning rate are due to the fact that we divide gradients by their norm. The values of these hyperparameters were chosen by observing the convergence of performed trainings for different values of α_{init} and choosing a high value for which the convergence is still observed. Subsequently, the value of the learning rate is automatically adapted by the algorithm following the observed relative decrease of the loss (if the loss stops to decrease, the learning rate is halved). The parameter α_{min} (minimal value of the learning rate) was introduced in order to prevent the learning rate to decrease infinitely after convergence.

2.2.4 Fusion of multiclass segmentations

In order to be robust to limitations of particular choices of neural network architectures (kernels, strides, connectivity between layers, numbers of features maps, activation functions) we propose to combine multiclass segmentations produced by several models. The final segmentation is obtained by a voxelwise voting strategy exploiting the following relations between tumor subclasses:

- Whole tumor region includes tumor-induced edema (class 2) and tumor core
- Tumor core region includes contrast-enhancing core (class 3) and non-enhancing core (class 1)

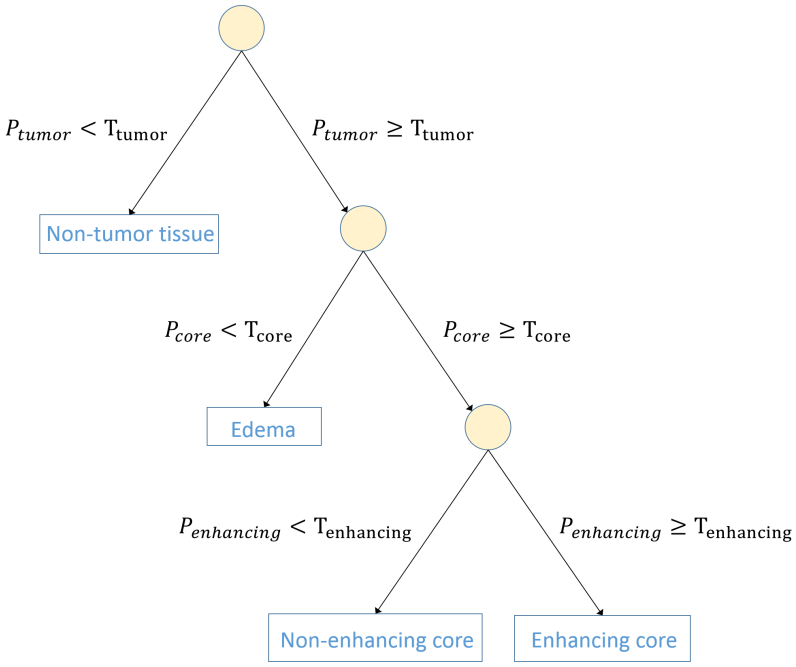

Figure 2.7: Tree representing our decision process: leaves represent classes and nodes represent decisions according to aggregated votes for tumor subregions. The class of a voxel is progressively determined by thresholding on proportions of models which voted for given subregions.

Suppose we have n multiclass segmentations produced by different models and let's note v_c the number of models which classified voxel (x, y, z) as belonging to the class c, with $c \in \{0, 1, 2, 3\}$. The main idea is to aggregate the votes for classes according to their common regions and to take the decision in the hierarchical order, progressively determining the tumor subregions. The number of votes for one region is the sum of votes for all classes belonging to the region (for example the votes for 'tumor core' are either votes for 'enhancing core' or 'non-enhancing core'). We define the following quantities:

- $P_{tumor} = (v_1 + v_2 + v_3)/(v_0 + v_1 + v_2 + v_3)$ (proportion of votes for the whole tumor region in the total number of votes)
- $P_{core} = (v_1 + v_3)/(v_1 + v_2 + v_3)$ (proportion of votes for the 'tumor core' region among all votes for tumor subclasses)
- $P_{enhancing} = v_3/(v_1 + v_3)$ (proportion of votes for the contrast-enhancing core among all votes for the tumor core)

The decision process can be represented by a tree (Fig. 2.7) whose internal nodes represent the application of thresholding on the quantities defined above and whose leaves represent classes (final decision). The first decision is therefore to determine if a given voxel represents a tumor tissue, given the proportion of networks which voted for one of the tumor subclasses. If this proportion is above a chosen threshold, we consider that the voxel represents a tumor tissue and we apply the same strategy to progressively determine the tumor subclass.

For each internal node R (corresponding to a tumor subregion) of the decision tree, we therefore have to choose a threshold T_R with $0 < T_R \leq 1$. A high T_R implies that a large proportion of models have to vote for this tumor subregion in order to consider its presence. The choice of this threshold therefore allows the user to control the trade-off between sensitivity and specificity of the corresponding tumor subregion. A low threshold gives priority to the sensitivity while a high threshold gives priority to the specificity.

A voting strategy was also used by the organizers of the BRATS 2015 challenge [Menze 2015] to combine multiclass segmentations provided by few experts. In the merging scheme of BRATS 2015, the tumor subregions are ordered and the votes for different subregions are successively thresholded by the number of total votes divided by 2. In contrast to this approach, in each step of our decision process we only consider the votes for the 'parent' region in the decision tree and we consider varying thresholds.

2.3 Experiments

We perform a series of experiments in order to analyze the effects of the main components of our method and to compare our results with the state of the art. Our method is evaluated on a publicly available database of the BRATS 2017 challenge.

2.3.1 Data and evaluation

The datasets of BRATS 2017 contain multisequence MR preoperative scans of patients diagnosed with malignant brain tumors. For each patient, four MR sequences were acquired: T1-weighted, post-contrast (gadolinium) T1-weighted, T2-weighted and FLAIR (Fluid Attenuated Inversion Recovery). The images come from 19 imaging centers and were acquired with different MR systems and with different clinical protocols. The images are provided after the pre-processing performed by the organizers: skull-stripped, registered to the same anatomical template and interpolated to $1mm^3$ resolution.

The *Training* dataset contains 285 scans (210 high grade gliomas and 75 low grade gliomas) with provided ground truth segmentation. The *Validation* dataset consists of 46 patients without provided segmentation and without provided information on the tumor grade. The evaluation on this dataset is performed via a public benchmark.

The first test dataset used in our experiments is composed of 50 randomly chosen patients from the *Training* dataset and the networks are trained on the remaining 235 patients. We refer to this dataset as 'test dataset' in the remainder (locally generated split training/test). We then evaluate our method on the *Validation* dataset of BRATS 2017 (networks are trained on all 285 patients of the *Training* dataset).

The ground truth corresponds to voxelwise annotations with 4 possible classes: non-tumor (class 0), contrast-enhancing tumor (class 3), necrotic and non-enhancing tumor (class 1), tumor-induced edema (class 2). The performance is measured by the Dice score between the segmentation $\tilde{Y}$ produced by the algorithm and the ground truth segmentation Y:

$$DSC(\tilde{Y}, Y) = \frac{2|\tilde{Y} \cap Y|}{|\tilde{Y}| + |Y|} \tag{2.4}$$

We perform t-tests (paired, one-tailed) to measure statistical significance of the observed improvements provided by the main components of our method (2D-3D model, modality-specific subnetworks, merging strategy). We consider the significance level of 5%.

2.3.2 Technical details

The ranges of image intensities highly vary between the scans due to image acquisition differences. We perform therefore a simple intensity normalization: for each patient and each MR sequence separately, we compute the median value of non-zero voxels, we divide the sequence by this median and we multiply it by a fixed constant. In fact, median is likely to be more stable than the mean, which can be easily impacted by the tumor zone. Experimentation with other normalization approaches such as histogram-matching methods [Nyúl 2000] will be a part of the future work. Another potentially useful pre-processing could be bias field correction [Sled 1998].

Models are trained with our optimization algorithm described previously. In each iteration of the training, gradients are computed on 10 batches (parameter N introduced in section 2.2.3.2) in the 2D case and on 5 batches in the 2D-3D case. Batch normalization [Ioffe 2015] was used in the 2D model but was not required to train the 2D-3D model. In the latter case, we normalized the input images to approximatively match the ranges of values of extracted 2D features.

To train the 2D model, the following target weights (defined in section 2.2.3.1) were fixed: $t_0 = 0.7$, $t_1 = 0.1$, $t_2 = 0.1$, $t_3 = 0.1$, corresponding respectively to 'non-tumor', 'non-enhancing core', 'edema' and 'enhancing core' classes. The choice of these values has an influence on the sensitivity to different tumor subclasses, however, the final segmentation performance in terms of Dice score was not found to be very sensitive to these hyperparameters. We fixed the same target weight for all tumor subclasses and we fixed a relatively high target weight for the non-tumor class to limit the risk of oversegmentation. However, given that non-tumor voxels represent approximately 98% of voxels of the batch, we significantly decreased the

weight of the non-tumor class compared to a standard cross-entropy loss (0.98 vs 0.7). In the 3D case, the following weights were fixed: $t_0 = 0.4$, $t_1 = 0.2$, $t_2 = 0.2$, $t_3 = 0.2$. We observe a satisfying convergence of the training both for the 2D and the 2D-3D model. Fig. 2.8 shows the evolution of the training loss of the 2D model along with Dice scores of tumor subclasses.

The weights of the classification layers of the 2D model (section 2.2.3.1) were the following: $c^{main} = 0.75$, $c^k = 0.05\ \forall k \in [1..5]$ (4 modality-specific subnetworks, one subnetwork combining all modalities and the main part of the network having a weight of 0.75 in the loss function). A high weight was given for the main classification layer as it corresponds to the final output of the 2D model. The classification layers of subnetworks were all given the same weight.

2.3.3 Training with missing modalities

We test our 2D model with modality-specific subnetworks in the context of missing MR sequences in the training database. In this setting, we suppose that the four MR sequences are available only for 20% of patients and that for the remaining patients, one MR sequence out of the four is missing. More precisely, we randomly split the training set of 235 patients in five equal subsets (47 patients in each) and we consider that only the first subset contains all the four MR sequences whereas the four other subsets exclusively miss one MR sequence (T1, T1c, T2 or T2-FLAIR). We previously noted that modality-specific subnetworks can be trained independently: in this case, a subnetwork specific to a given MR sequence can be trained on 80% of the training database (on all training images except the ones for which the MR sequence is missing). The goal of the experiment is to test if the training of these subnetworks improves the segmentation performance in practice. We first evaluate the performance obtained by 2D model 1 (version CNN-2DAxl) trained only on the training subset containing all MR sequences (47 patients). Then we evaluate the performance obtained when the subnetworks are pretrained, each of them using 80% of the training database.

The results are reported in Table 2.1. Pretraining of the modality-specific subnetworks improved the segmentation performance on the test set for all tumor sub-

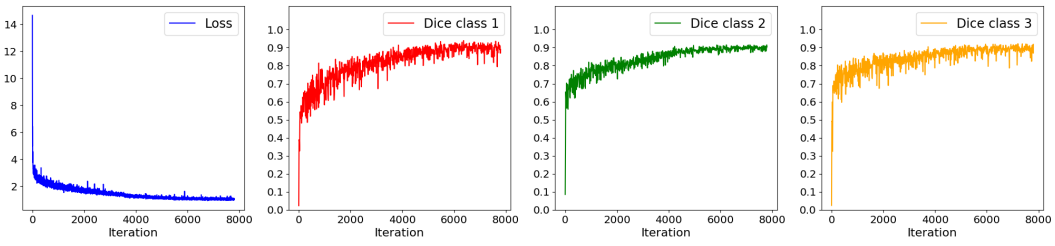

Figure 2.8: Evolution of the loss and of Dice scores of tumor subclasses during the training of the 2D model.

regions. Even if the multiclass segmentation problem is very difficult for a small network using only one MR sequence, this pretraining forces the subnetwork to learn the most relevant features, which will then be used by the main part of the network, trained on the subset of training cases for which all MR sequences are available. The improvement was found statistically significant (p-value < 0.05) for all the three tumor subregions (Table 2.5).

2.3.4 Using long-range 2D context

We perform a series of experiments to analyze the effects of using features learned by 2D networks as an additional input to 3D networks. In the first step, 2D model 1 is trained separately on axial, coronal and sagittal slices and the standard 3D model is trained on 70x70x70 patches. Then we extract the features produced by the 2D model for all images of the training database and we train the same 3D model on 70x70x70 patches using these extracted features (Fig. 2.9) as an additional input (2D-3D model A specified on Fig. 2.4). The experiment is performed on two datasets: the test dataset of 50 patients (networks trained on the remaining 235 patients) and the *Validation* dataset of BRATS 2017 (networks trained on 285 patients). The results on the two datasets are reported respectively in Table 2.2 and Table 2.3. Further experiments, involving varying 2D and 3D architectures are presented in section 2.3.5. Qualitative analysis is performed on the first dataset, for which the ground truth segmentation is provided. For comparison, we also display the scores obtained by U-net processing axial slices, using our implementation (with batch-normalization).

On the two datasets and for all tumor subregions, our 2D-3D model obtained a

Table 2.1: Mean Dice scores on the test dataset (50 patients) in the context of misssing MR sequences in the training database. EC, TC and WT refer respectively to 'Enhancing Core', 'Tumor Core' and 'Whole Tumor' regions. The numbers in brackets denote standard deviations.

	EC	TC	WT
2D model 1, missing data	70.2 (22.3)	68.6 (27.9)	83.0 (14.6)
2D model 1 missing data + pretrained subnetworks	**71.9** (20.9)	**73.7** (23.7)	**84.1** (13.6)
2D model 1 full data	73.6 (19.8)	79.4 (15.7)	86.6 (11.1)

Table 2.2: Mean Dice scores on the test dataset (50 patients). The numbers in brackets denote standard deviations.

	EC	TC	WT
Unet axial slices	73.9 (19.7)	78.1 (17.9)	86.5 (11.6)
2D model 1 axial slices	73.6 (19.8)	79.4 (15.7)	86.6 (11.1)
Standard 3D model (without 2D features)	73.7 (19.9)	77.0 (18.5)	85.7 (8.3)
2D-3D model A, features from 2D model 1	**77.4** (16.6)	**80.9** (16.9)	**87.3** (11.7)

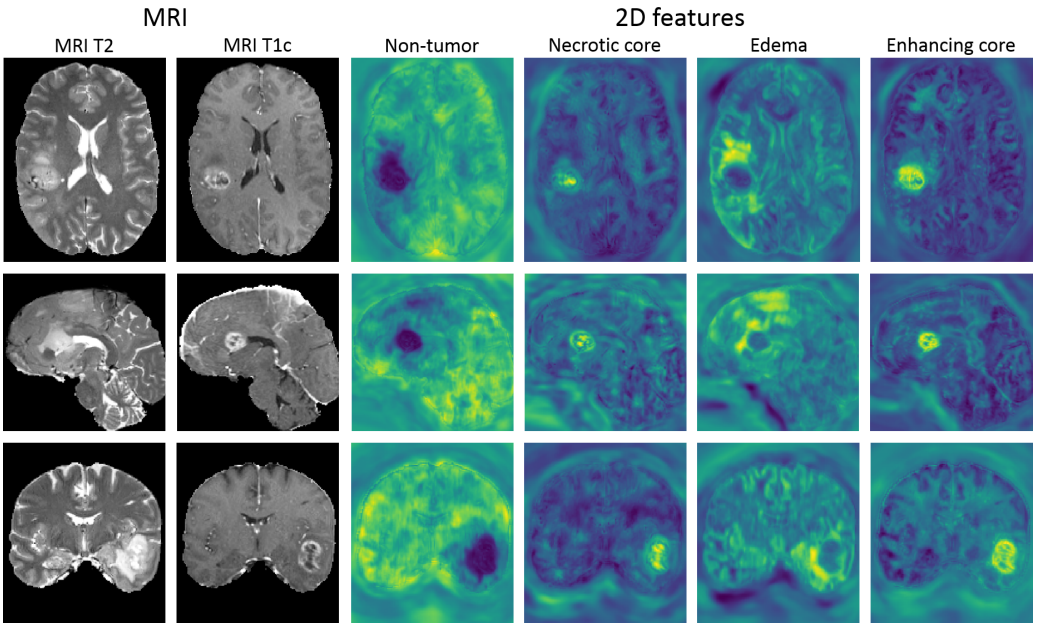

Figure 2.9: 2D features computed for three different patients from the test set. These features correspond to unnormalized outputs of the final convolutional layers of three versions of a 2D model (CNN-2DAxl, CNN-2DSag, CNN-2DCor). The values of these features are used as an additional input to a 3D CNN. Each feature highlights one of the tumor classes (columns 3-6) and encodes a rich information extracted from a long-range 2D context within an axial, sagittal or coronal plane (rows 1-3). Each row displays a different case from the test set (unseen by the network during the training).

better performance than the standard 3D CNN (without the use of 2D features) and than 2D model 1 from which the features were extracted (Table 2.2 and Table 2.3). The qualitative analysis (Fig. 2.10) of outputs of 2D networks highlights two main problems of 2D approaches. First, as expected, the produced segmentations show discontinuities which appear as patterns parallel to the planes of processing. The second problem are false positives in the slices at the borders of the brain and containing artefacts of skull-stripping. Segmentations produced by the standard 3D model are more spatially consistent but the network suffers from a limited input information from distant voxels. The use of learned features as an additional input to

Table 2.3: Mean Dice scores on the *Validation* dataset of BRATS 2017 (46 patients).

	EC	TC	WT
Unet axial slices	71.4 (27.4)	76.6 (22.4)	87.7 (10.6)
2D model 1 axial slices	71.1 (28.8)	78.4 (21.3)	88.6 (8.7)
Standard 3D model (without 2D features)	68.7 (30.0)	74.2 (23.7)	85.4 (10.9)
2D-3D model A, features from 2D model 1	**76.7** (27.6)	**79.5** (21.3)	**89.3** (8.5)

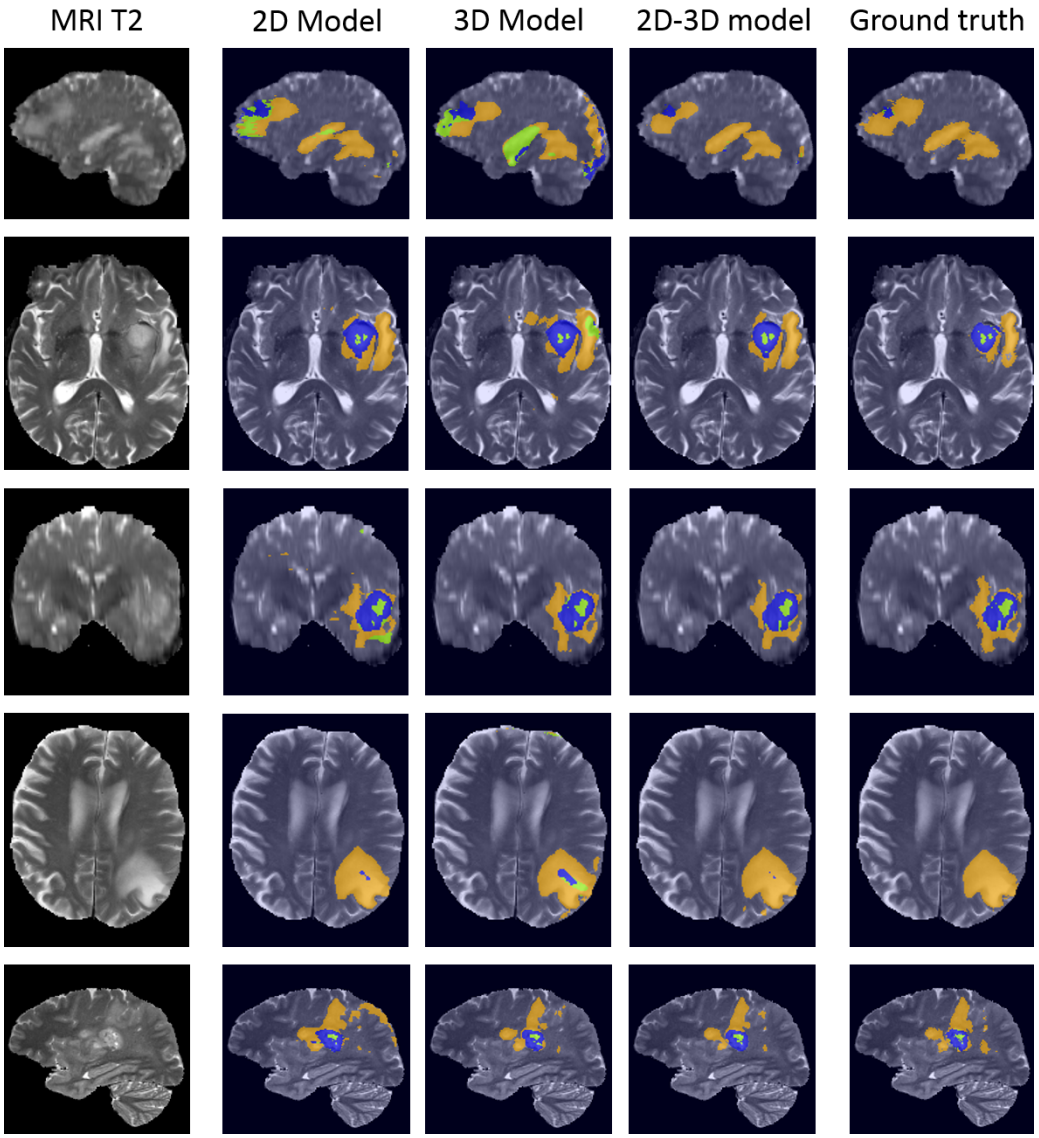

Figure 2.10: Examples of segmentations obtained with models using a different spatial context. Each row represents a different patient from the local test dataset (images unseen during the training). From left to right: MRI T2, '2D model 1' processing the image by axial slices, standard 3D model (without 2D features), '2D-3D model A' using the features produced by '2D model 1', ground truth segmentation. Orange, blue and green zones represent respectively edema, contrast-enhancing core and non-enhancing core.

the network gives a considerable advantage by providing rich information extracted from distant points. The difference of performance is particulary visible for 'tumor core' and 'enhancing core' subregions. The improvements of our 2D-3D approach compared to the standard 3D CNN (without the use of 2D features) were found statistically significant (p-value < 0.05) in all cases except the 'whole tumor' region

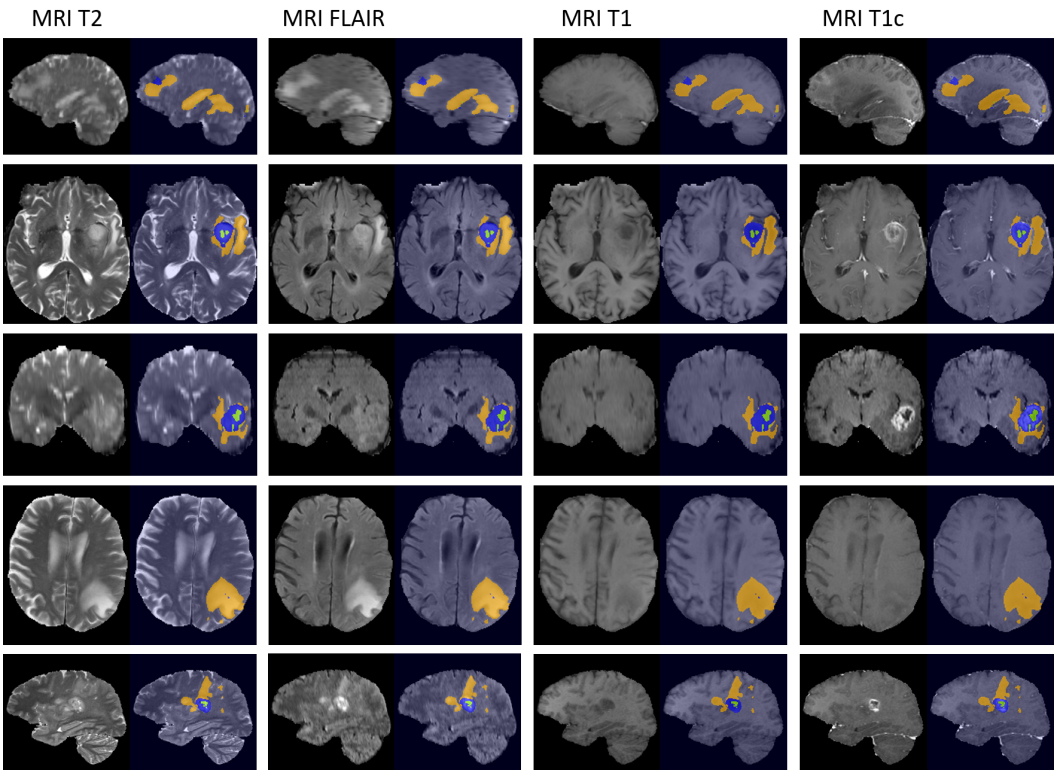

Figure 2.11: Results obtained by the 2D-3D model, displayed for each available MR sequence. While both T2 and T2-FLAIR higlight the edema, T2-FLAIR allows for distinguishing it from the cerebrospinal fluid. T1 with injection of a gadolinium-based contrast agent highlights the degradation of the blood-brain barrier induced by the tumor.

in the first dataset (Table 2.5).

2.3.5 Varying network architectures and combining segmentations

We perform experiments with varying architectures of 2D and 2D-3D models. The first objective is to test if the use of 2D features provides an improvement when different 2D and 2D-3D architectures are used. The second objective is to test our decision process combining different multiclass segmentations. The third goal is to compare performances obtained by different models. The experiments are performed on the *Validation* set of BRATS 2017, the performance is evaluated by the public benchmark of the challenge.

In our experiments we use two architectures of our 2D model and three architectures of the 2D-3D model. The main difference between the two 2D networks used in experiments is the architecture of subnetworks processing the input MR sequences. In the first 2D model, the subnetworks correspond to reduced versions

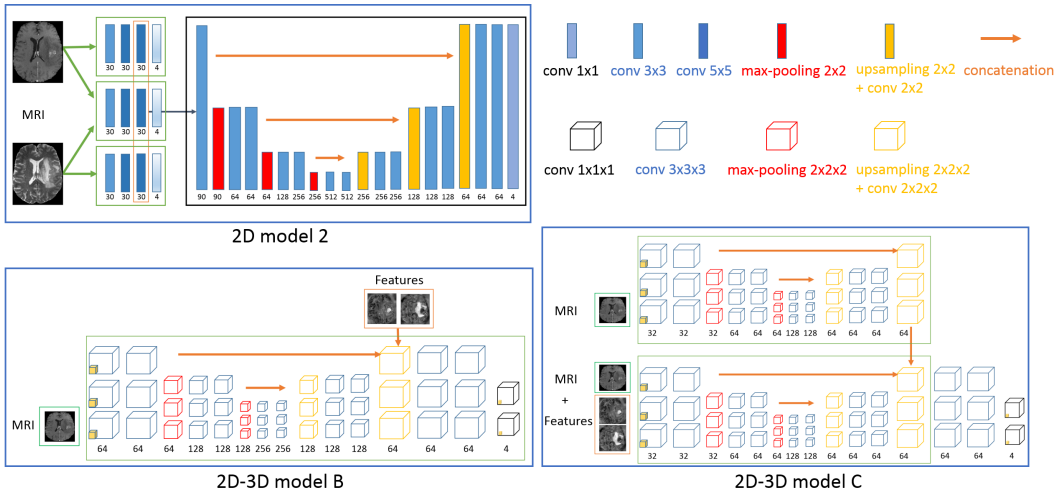

Figure 2.12: Architectures of complementary networks used in our experiments.

of U-Net (Fig. 4.2) whereas in the second model, the subnetworks are composed of three convolutional layers (Fig. 2.12, top). In the remainder, we refer to these models as '2D model 1' and '2D model 2'. The difference between the two first 2D-3D models is the choice of the layer in which the 2D features are imported: in the first layer of the network (Fig. 2.4) or before the final sequence of convolutional layers (Fig. 2.12, bottom left). The third 2D-3D model (Fig. 2.12, bottom right) is composed of two streams, one processing only the 3D image patch and the other stream taking also the 2D features as input. We refer to these models as 2D-3D model A, 2D-3D model B and 2D-3D model C. Please note that the two first models correspond to a standard 3D model with the only difference of taking an additional input.

Each of the 2D-3D models is trained twice using respectively features learned by 2D model 1 or features learned by 2D model 2. We combine the trained 2D-3D models with the voting strategy described in section 2.2.4. As we observe that 2D model 1 performs better than 2D model 2, we consider two ensembles: combination of all trained 2D-3D models and combination of three models using features from 2D model 1. We use the following thresholds for merging (defined in section 2.2.4): $T_{tumor} = 0.4$, $T_{core} = 0.3$, $T_{enhancing} = 0.4$.

The results are reported in Table 2.4. In all experiments, the 2D-3D models obtain better performances than their standard 3D counterparts and than 2D networks from which the features were extracted. The merging of segmentations with our decision rule further improves the performance. For all tumor subregions, the ensemble of 6 models (the last row of Table 2.4) outperforms each of the individual models. The improvement over the main 2D-3D model (2D-3D model A with features from 2D model 1) was found statistically significant (p-value < 0.05) for 'whole tumor' and 'tumor core' subregions, as reported in the last row of Table 2.5.

Table 2.4: Mean Dice scores on the *Validation dataset* of BRATS 2017 (46 patients).

	EC	TC	WT
2D model 1 axial slices	71.1	78.4	88.6
2D model 2 axial slices	68.0	78.3	88.1
Standard 3D model (without 2D features)	68.7	74.2	85.4
* 2D-3D model A, features from 2D model 1	76.7	79.5	89.3
* 2D-3D model B, features from 2D model 1	76.6	79.1	89.1
* 2D-3D model C, features from 2D model 1	76.9	78.3	89.4
* 2D-3D model A, features from 2D model 2	73.4	79.5	89.7
* 2D-3D model B, features from 2D model 2	74.1	79.4	89.5
* 2D-3D model C, features from 2D model 2	74.3	79.4	89.6
Combination of models A-C features from model 1	76.7	79.6	89.4
Combination of all models * (final segmentation)	**77.2**	**80.8**	**90.0**

Table 2.5: p-values of the t-tests (in bold: statistically significant results, with p < 0.05) of the improvement provided by the different components of our method. To lighten the notations, '2D' refers to '2D model 1 axial slices' and '2D-3D' refers to '2D-3D model A, features from 2D model 1'. 'Combination of 2D-3D' refers to the result obtained by merging 6 models with our hierarchical decision process.

	EC	TC	WT
2D vs 2D with pretrained subnetworks, missing data	**0.0054**	**0.0003**	**0.0074**
Standard 3D vs 2D-3D, dataset 1	**0.0082**	**0.0016**	0.0729
Standard 3D vs 2D-3D, dataset 2	**0.0077**	**0.0005**	**<0.0001**
2D-3D vs combination of 2D-3D	0.1058	**0.0138**	**0.0496**

While the three 2D-3D architectures yield similar performances, 2D model 1 (subnetworks similar to U-net) performs better than 2D model 2 for all three tumor regions. However, the 2D-3D models trained with the features from 2D model 2 are useful for the merging of segmentations: the ensemble of all models yields better performances than the ensemble of three models (two last rows of Table 2.4).

2.3.6 Comparison to the state of the art

We have evaluated our segmentation performance on the public benchmark of the challenge to compare our results with few dozens of teams from renowned research institutions worldwide. Our method compares favorably with competing methods of BRATS 2017 (Table 2.6): among 55 teams which evaluated their methods on all test patients of the *Validation* set, we obtain top-3 performance for 'core' and 'enhancing core' tumor subregions. We obtain mean Dice score of 0.9 for the 'whole tumor' region, which is almost equal to the one obtained by the best scoring team (0.905).

The winning method of UCL-TIG [Wang 2017] proposes to sequentially use three

Table 2.6: Mean Dice scores of the 10 best scoring teams on the validation leaderboard of BRATS 2017 (state of January 22, 2018)

	EC	TC	WT	Rank EC	Rank TC	Rank WT	Average rank
UCL-TIG	78.6	83.8	90.5	1 / 55	1	1	1.0
MIC_DKFZ	77.6	81.9	90.3	2 / 55	2	2	2.0
inpm (our method)	77.2	80.8	90.0	3 / 55	3	7	4.3
UCLM_UBERN	74.9	79.1	90.1	9 / 55	6	3	6.0
biomedia1	73.8	79.7	90.1	12 / 55	5	5	7.3
stryker	75.5	78.3	90.1	6 / 55	10	6	7.3
xfeng	75.1	79.9	89.2	8 / 55	4	11	7.7
Zhouch	75.4	77.8	90.1	7 / 55	12	4	7.7
tkuan	76.5	78.2	88.9	4 / 55	11	13	9.3
Zhao	75.9	78.9	87.2	5 / 55	7	16	9.3

Table 2.7: Distribution of Dice scores (final result). The numbers in brackets denote standard deviations.

	EC	TC	WT
Mean	77.2 (24.4)	80.8 (18.9)	90.0 (8.1)
Median	85.4	88.3	91.8
Quantile 25 %	76.9	75.0	89.6
Quantile 75 %	90.0	93.5	94.5

3D CNNs in order to progressively determine the tumor subclass. Each of the networks performs a binary segmentation (tumor/not tumor, core/edema, enhancing core/non-enhancing core) and was designed for one tumor subregion of BRATS. A common point with our method is the hierarchical process, however in our method all models perform multiclass segmentation. The method of the team MIC_DKFZ, according to [Isensee 2017], is based on an optimized version of 3D U-net and an extensive use of data augmentation.

The leaderboard of BRATS 2017 only shows mean performances obtained by participating teams. However, the benchmark individually provides detailed scores and complementary statistics, in particular quartiles and standard deviations reported in Table 2.7. Our method yields promising results with median Dice score of 0.918 for the *whole tumor*, 0.883 for the *tumor core* and 0.854 for the *enhancing core*. While the Dice scores for the *whole tumor* region are rather stable (generally between 0.89 and 0.95), we observe a high variability of the scores obtained for the tumor subregions. In particular the obtained median Dices are much higher than the means, due to the sensitivity of Dice score to outliers.

2.4 Discussion and conclusion

In this work, we presented a deep learning system for multiclass segmentation of tumors in multisequence MR scans. The goal of our work was to propose elements to improve performance, robustness and applicability of commonly used CNN-based systems. In particular, we proposed a new methodology to capture a long-range 3D context with CNNs, we introduced a network architecture with modality-specific subnetworks and we proposed a voting strategy to merge multiclass segmentations produced by different models.

First, we proposed to use features learned by 2D CNNs (capturing a long-range 2D context in three orthogonal directions) as an additional input to a 3D CNN. Our approach combines the strengths of 2D and 3D CNNs and was designed to capture a very large spatial context while being efficient in terms of computations and memory load. Our experiments showed that this hybrid 2D-3D model obtains better performances than both the standard 3D approach (considering only the intensities of voxels of a subvolume) and than the 2D models which produced the features. Even if the use of the additional input implies supplementary reading operations, the simple importation of few features to a CNN does not considerably increase the number of computations and the memory load. In fact, in typical CNNs performing hundreds of convolutions, max-poolings and upsamplings, the data layer represents typically a very small part of the memory load of the network. One solution to limit the reading operations could be to read downsampled versions of features or to design a 2D-3D architecture in which the features are imported in a part of the network where the feature maps are relatively small.

The improvement provided by the 2D-3D approach has the cost of increasing the complexity of the method compared to a pure 3D approach as it requires a two-step processing (first 2D, then 3D). However, its implementation is rather simple as the only supplementary element to implement is the extraction of 2D features, i.e. computation of outputs of trained 2D networks (with a deep learning software such as TensorFlow) and saving the obtained tensors in files. In the 3D part, the extracted features are then simply read as additional channels of the input image.

Despite the important recent progress of GPUs, pure 3D approaches may be easily limited by their computational requirements when the segmentation problem involves an analysis of a very large spatial 3D context. In fact, Convolutional Neural Networks require an important amount of GPU memory and a high computational power as they perform thousands of costly operations on images (convolutions, max-poolings, upsamplings). The main advantage of our 2D-3D approach is to considerably increase the size of the receptive field of the model while being efficient in terms of the computational load. The use of our 2D-3D model may therefore be particularly relevant in the case of very large 3D scans.

Second, we proposed a novel approach to process different MR sequences, using an architecture with modality-specific subnetworks. Such design has the considerable advantage of offering a possibility to train one part of the network on databases containing images with missing MR sequences. In our experiments, training of

modality-specific subnetworks improved the segmentation performance in the setting with missing MR sequences in the training database. Moreover, the fact that our 2D model obtained promising segmentation performance is particularly encouraging given that 2D networks are easier to apply for the clinical use where images have a variable number of acquired slices. Our approach can be easily used with any deep learning software (e.g. Keras). In the case of databases with missing MR sequences, the user only has to perform a training of a subnetwork (on images for which the given MR sequence is provided) and then read the learned parameters for the training of the main part of the network (on images for which all MR sequences are available).

In order to be less prone to limitations of particular choices of neural network architectures, we proposed to merge outputs of several models by a voxelwise voting strategy taking into account the semantics of labels.

In constrast to most methods, we do not apply any postprocessing on the produced segmentations.

Our methodological contributions can be easily included separately or jointly into a CNN-based system to solve specific segmentation problems. The implementation of our method will be made publicly available on https://github.com/PawelMlynarski.

Chapter 3

Deep Learning with Mixed Supervision for Brain Tumor Segmentation

Contents

In this chapter, published as a journal article [Mlynarski 2019c], we focus on the cost of manual segmentation of brain tumors and we propose a method to exploit a less costly form of annotations (image-level labels). We introduce a CNN-based model which is trained with a mixed level of supervision, using both fully-annotated images and weakly-annotated images. A large number of cross-validated experiments in performed to analyze the effect of the mixed supervision compared to the standard supervised learning for segmentation. The obtained improvement is proportional to the radio between the numbers of weakly-annotated and fully-annotated images.

3.1 Introduction

Cancer is today the third cause of mortality worldwide. In this work, we focus on segmentation of gliomas, which are the most frequent primary brain cancers [Goodenberger 2012]. Gliomas are particularly malignant tumors and can be

broadly classified according to their grade into low grade gliomas (grades I and II defined by World Health Organization) and high grades gliomas (grades III-IV). Glioblastoma multiforme is the most malignant form of glioma and is associated with a very poor prognosis: the average survival time under therapy is between 12 and 14 months.

Medical images play a key role in diagnosis, therapy planning and monitoring of cancers. Treatment protocols often include evaluation of tumor volumes and locations. In particular, for radiotherapy planning, clinicians have to manually delineate target volumes, which is a difficult and time-consuming task. Magnetic Resonance (MR) images [Bauer 2013] are particularly suitable for brain cancer imaging. Different MR sequences (T2, T2-FLAIR, T1, T1+gadolinium) highlight different tumor subcomponents such as edema, necrosis or contrast-enhancing core.

In recent years, machine learning methods have achieved impressive performance in a large variety of image recognition tasks. Most of the recent state-of-the-art segmentation methods are based on Convolutional Neural Networks (CNN) [LeCun 1995, Long 2015]. CNNs have the considerable advantage of automatically learning relevant image features. This ability is particularly important for the tumor segmentation task. CNN-based methods [Pereira 2015, Kamnitsas 2016, Kamnitsas 2017a, Wang 2017] have obtained the best performances on the four last editions of Multimodal Brain Tumor Segmentation Challenge (BRATS) [Menze 2015, Bakas 2017].

Most of the segmentation methods based on machine learning rely uniquely on manually segmented images. The cost of this annotation is particularly high in medical imaging where manual segmentation is not only time-consuming but also requires high medical competences. Image intensity of cancerous tissues in MRI or CT scans is often similar to the one of surrounding healthy or pathological tissues, making the exact tumor delineation difficult and subjective. In the case of brain tumors, according to [Menze 2015], the inter-rater overlap of expert segmentations is between 0.74 and 0.85 in terms of Dice coefficient. For these reasons, high-quality manual tumor segmentations are generally available in very limited numbers. Segmentation approaches able to exploit images with weaker forms of annotations are therefore of particular interest.

In this work, we assume that the training dataset contains two types of images: fully-annotated (with provided ground truth segmentation) and weakly-annotated, with an image-level label indicating presence or absence of a tumor tissue within the image (Fig. 4.1). We refer to this setting as 'mixed supervision'. The latter type of annotations can be obtained at a substantially lower cost as it is less time-consuming, potentially requires less medical expertise and can be obtained without the use of a dedicated software.

We introduce a novel CNN-based segmentation model which can be trained using weakly-annotated images in addition to fully-annotated images. We propose to extend segmentation networks, such as U-Net [Ronneberger 2015], with an additional branch, performing image-level classification. The model is trained jointly for both tasks, on fully-annotated and weakly-annotated images. The goal is to

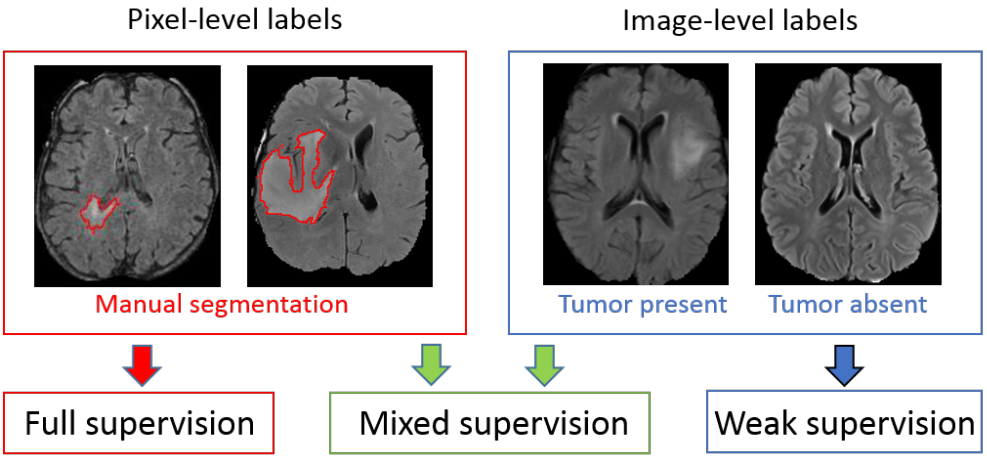

Figure 3.1: Different levels of supervision for training of segmentation models. Standard models are trained on fully-annotated images only, with pixel-level labels. Weakly-supervised approaches aim to train models using only weakly-annotated images, e.g. with image-level labels. Our model is trained with a mixed supervision, exploiting both types of training images.

exploit the representation learning ability of CNNs to learn from weakly-annotated images while supervising the training using fully-annotated images in order to learn features relevant for the segmentation task. Our approach differs from the standard semi-supervised learning as we consider weakly-annotated data instead of totally unlabelled data. To the best of our knowledge, we are the first to combine pixel-level and image-level labels for training of models for tumor segmentation.

We perform a series of cross-validated tests on the challenging task of segmentation of gliomas in MR images from BRATS 2018 challenge. We evaluate our model both for binary and multiclass segmentation using a variable number of ground truth segmentations available for training. Since all 3D images from the BRATS 2018 contain brain tumors, we focus on the 2D problem of tumor segmentation in axial slices of a MRI and we assume slice-level labels for weakly-annotated images. Using approximately 220 MRI with slice-level labels and a varying number of fully-annotated MRI, we show that our approach significantly improves the segmentation accuracy when the number of fully-annotated cases is limited.

3.2 Related work

In the literature, there are several works related to weakly-supervised and semi-supervised learning for object segmentation or detection. Most of the related works were applied to natural images.

The first group of weakly-supervised methods aims to localize objects using

only weakly-annotated images for training. When only image-level labels are available, one approach is to design a neural network which outputs two feature maps per class (interpreted as 'heat maps' of the class) which are then pooled to obtain an image-level classification score penalized during the training [Pathak 2014, Pinheiro 2015, Saleh 2016, Bearman 2016, Wang 2018b]. At test time, these 'heat maps' are used for detection (determining a bounding box of the object) or segmentation. To guide the training process, some works use self-generated spatial priors [Pinheiro 2015, Saleh 2016, Bearman 2016] or inconsistency measures [Wang 2018b] in the loss function. To obtain an image-level score, in [Pathak 2014, Bearman 2016], global maximum pooling is used. Application of the maximum function on large feature maps may cause optimization problems as training of neural networks is based on computation of gradients [Dreyfus 1990]. Log-SumExp approximation of the maximum [Boyd 2004] is therefore used in the works [Pinheiro 2015, Saleh 2016] in order to partially limit this problem. Average pooling on small feature maps was used by Wang et al [Wang 2018b] for the problem of detection of lung nodules.

Dubost et al [Dubost 2017] propose to extend a network similar to 3D U-net with a subnetwork performing image-level regression of the number of present lesions. The model is trained using only image-level labels (lesion counts) and the weights learned by the regression subnetwork are used during the test phase to construct heat maps of lesions. Detection of lesions is obtained by thresholding of the heat maps. The common point with our model is the extension of U-net with a subnetwork performing an image-level task. One of the key differences is that our model is trained using both image-level and pixel-level labels and has a dedicated segmentation layer (trained with pixelwise labels and producing the final segmentation).

Another type of weakly-supervised methods aims to detect objects in natural images based on classification of image subregions [Girshick 2014, Oquab 2015] using pre-trained classification networks such as VGG-Net [Simonyan 2014] or AlexNet [Krizhevsky 2012]. In fact, one particularity of natural images is their recursive aspect: one image can correspond to a subpart of another image (e.g. two images of the same object taken from different distances). A classification network trained on a large dataset may therefore be used on a subregion of a new image in order to determine if it contains an object of interest.

Pre-trained classification networks were also used to detect objects by determining image subregions whose modification influences the global classification score of a class. In [Simonyan 2013], Simonyan et al. propose to compute the gradient of the classification score with respect to the intensities of pixels and to threshold it in order to localize the object of interest. However, these partial derivatives represent a very weak information for tumor segmentation, which requires a complex analysis of the spatial context. The method proposed in [Bergamo 2014] is based on replacing image subregions by the mean value in order to measure the drop of the classification score.

Overall, the reported segmentation performances of weakly-supervised methods

are considerably lower than the ones obtained by semi-supervised and supervised approaches. In absence of pixel-level labels, a model may learn irrelevant features, due for example to co-occurrences of objects or image acquisition differences in the case of multicenter medical data. Despite the cost of manual segmentation, at least few fully-annotated images can still be obtained in many cases.

In standard semi-supervised learning [Cheplygina 2018] for classification, the training data is composed both of labelled samples and unlabelled samples. Unlabelled samples can be used to encourage the model to satisfy some properties on relations between labels and the feature space. Common properties include smoothness (points close in the feature space should be close in the target space), clustering (labels form clusters in the feature space) and low density separation (decision boundaries should be in low density regions of the feature space). Semi-supervised learning based on these properties can be performed by graph-based methods such as the recent work of Kamnitsas et al. [Kamnitsas 2018]. The main idea of such methods is to propagate labels in a fully-connected graph whose nodes are samples (labelled and unlabelled) and whose edges are weighted by similarities between samples. The use of graph-based semi-supervised methods is difficult for segmentation, in particular because it implies computation of similarity metrics between samples, whereas each single image is generally composed of millions of samples (pixels or voxels).

Relatively few works were proposed for semi-supervised learning for image segmentation. Some semi-supervised approaches are based on self-training, i.e. training of a machine learning model on self-generated labels. Iterative algorithms similar to EM [Zhang 2001] were proposed for natural images [Papandreou 2015] and medical images [Rajchl 2016]. Recently, Hung et al [Hung 2018] proposed a method based on Generative Adversarial Networks [Goodfellow 2014] where the generator network performs image segmentation and the discriminator network tries to determine if a segmentation corresponds to the ground truth or the segmentation produced by the generator. The discriminator network is used to produce confidence maps for self-training. The approaches based on self-training have the drawback of learning on uncertain labels (produced by the model itself) and training of such models is difficult.

Other approaches assume mixed levels of supervision similarly to our approach. Hong et al [Hong 2015, Hong 2016] proposed decoupled classification and segmentation, an approach for segmentation of objects in natural images based on a two-step training with a varying level of supervision. This architecture is composed of two separate networks trained sequentially, one performing image-level classification and used as encoder, and the another one taking as input small feature maps extracted from the encoder and performing segmentation. An important drawback of such design, in the case of tumor segmentation, is that the segmentation network does not take as input the original image and can therefore miss important details of the image (e.g. small tumors).

Our approach is related to multi-task learning [Evgeniou 2004]. In our case, the goal of training for two tasks (segmentation and classification) is to exploit all the

available labels and to guide the training process to learn relevant features. The approach closest to ours is the one of Shah et al. [Shah 2018]. In this work, the authors consider three types of annotations: segmentations, bounding boxes and seed points at the borders of objects. A neural network is trained using these three types of training data. In our work, we exploit the use of a significantly weaker form of annotations, image-level labels.

3.3 Joint classification and segmentation with Convolutional Neural Networks

3.3.1 Deep learning model for binary segmentation

We designed a novel deep learning model, which aims to take advantage of all available voxelwise and image-level annotations. We propose to extend a segmentation CNN with an additional subnetwork performing image-level classification and to train the model for the two tasks jointly. Most of the layers are shared between the classification and segmentation subnetworks in order to transfer the information between the two subnetworks. In this work we present the 2D version of our model, which can be used on different types of medical images such as slices of a CT scan or a multisequence MRI.

The proposed network takes as input a multimodal image of dimensions 300x300 and extends U-Net [Ronneberger 2015] which is currently one of the most used architectures for segmentation tasks in medical imaging. The different image modalities (e.g. sequences of a MRI) correspond to channels of the data layer and are the input of the first convolutional layer of the network (as in most of the currently used CNNs for image segmentation). U-Net is composed of an encoder part and a decoder part which are connected by concatenations between layers at the same scale, in order to combine low-level and local features with high-level and global features. This design is well suited for the tumor segmentation task since the classification of a voxel as tumor requires to compare its value with its close neighborhood but also taking into account a large spatial context. The last convolutional layer of U-net produces pixelwise classification scores, which are normalized by softmax function during the training phase. We apply batch normalization [Ioffe 2015] in all convolutional layers except the final layer.

We propose to add an additional branch to the network, performing image-level classification (Fig. 3.2), in order to exploit the information contained in weakly-annotated images during the training. This classification branch takes as input the second to last convolutional layer of U-net (representing a rich information extracted from a local and a long-range spatial context) and is composed of one mean-pooling, one convolutional layer and 7 fully-connected layers.

The goals of taking a layer from the final part of U-Net as input of the classification branch are both to guide the image-level classification task and to force the major part of the segmentation network to take into account weakly-annotated

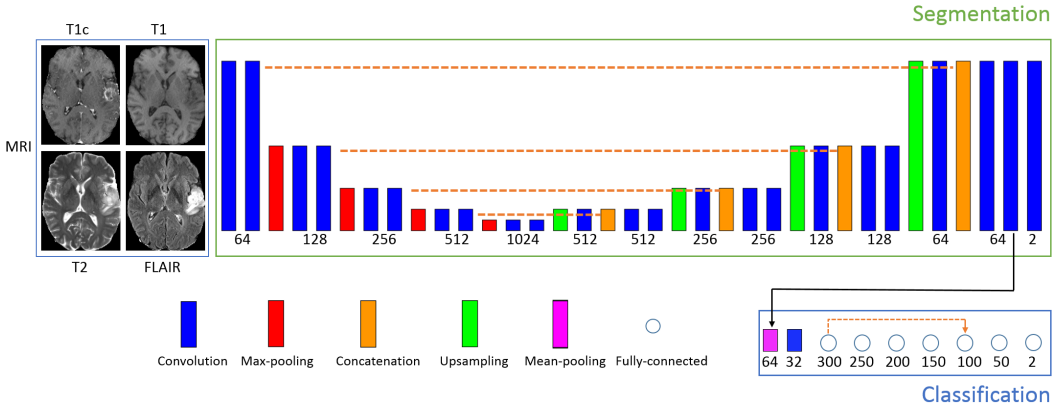

Figure 3.2: Architecture of our model for binary segmentation. The numbers of outputs are specified below boxes representing layers. The height of rectangles represents the scale (increasing with pooling operations). The dashed lines represent concatenation operations. The proposed architecture is an extended version of U-net, with a subnetwork performing image-level classification. Training of the model corresponds to a joint minimization of two loss functions related respectively to segmentation and image-level classifcation tasks.

images. This also helps the optimization process by taking advantage of the connectivity of layers in U-Net, helping the flow of gradients of the loss function during the training (in particular, note the connection between the first part and the last part of U-Net).

The second to last layer of the segmentation network outputs 64 feature maps of size 101x101 from which the classification branch has to output two global (image-level) classification scores (tumor absent/tumor present). We first reduce the size of these feature maps by applying a mean-pooling with kernels of size 8x8 and the stride of 8x8. We use the mean pooling rather than max-pooling in order to avoid information loss and optimization problems. One convolutional layer, with ReLU activation and kernels of size 3x3, is then added to reduce the number of feature maps from 64 to 32. The resulting 32 feature maps of size 11x11 are the input of the first fully-connected layer of the classification branch.

According to our experiments, a relatively deep architecture of the classification branch with a limited number of parameters and a skip-connection between layers yields the best performance. This observation is in agreement with current common designs of neural networks. Deep networks have the capacity to learn more complex features, due to applied non-linearities. The connectivity between layers at different depths helps the optimization process (e.g. Res-Net [He 2016]). In our case, we use 7 fully-connected layers with ReLU activations (except the final layer) and we concatenate the outputs of the first and the fifth fully-connected layer. The role of this concatenation is similar to the one connecting the first and the last sequence of convolutional layers in U-Net. The concatenation is used before the second to last

layer in order to have one layer to process the mixed information (concatenation of two layers) before the final decision in the seventh fully-connected layer. We use only one concatenation as the subnetwork is composed of only few layers while concatenations increase the number of parameters in the network. The last fully-connected layer outputs image-level classification scores (tumor tissue absent or present).

The model is trained both on fully-annotated and weakly-annotated images for the two tasks jointly (segmentation and classification). We can distinguish between three types of training images. First, images containing a tumor and with provided ground truth segmentation are the most costly ones. The second type are images that do not contain tumor, which implies that none of their pixels corresponds to a tumor. In this case, the ground truth segmentation is simply the zero matrix. The only problematic case is the third one, when the image is labelled as containing a tumor but without provided segmentation.

To train our model, we propose to form training batches containing the three mentioned types of images: k positive cases (containing a tumor) with provided segmentation, m negative cases and n positive cases without provided segmentation.

Given a training batch b and the network parameters θ, we use a weighted pixelwise cross-entropy loss on images of types 1 and 2: $Loss_s^b(\theta) = -\sum_{i=1}^{k+m}\sum_{(x,y)} w_{(x,y)}^i \log(p_{i,(x,y)}^l(\theta))$ where $p_{i,(x,y)}^l$ is the classification score given by the network to the ground truth label for pixel (x,y) of the i^{th} image of the batch and $w_{(x,y)}^i$ is the weight given to this pixel. The weights are used to limit the effect of class imbalance, since tumor pixels represent a small portion of the image. Weights of pixels are set automatically according to the composition of the training batch (number of pixels of each class) so that pixels associated with healthy tissues have a total weight of t_0 in the loss function and the pixels of the tumor class have a total weight of t_1, where t_0 and t_1 are target weights fixed manually. It means that if the training batch contains N_t pixels labelled as tumor, then each tumor pixel has a weight of t_1/N_1 (the pixelwise weight is high when the number of tumor pixels is low). This type of loss function was used in our previous work [Mlynarski 2019b].

The classification loss is a standard cross-entropy loss on all images of the training batch: $Loss_c = -\frac{1}{k+m+n}\sum_{i=1}^{k+m+n} \log(p_i^l(\theta))$ where p_i^l is the global classification score given by the network to the ground truth global label for the i^{th} image of the batch. In particular, fully-annotated images are also used for training of the classification branch in order to transfer the knowledge from the segmentation task to the image-level classification. We do not apply weights on the classification loss as image-level labels are balanced through the sampling of training batches (having a fixed number of non-tumor images).

Since both segmentation and classification losses are normalized, we define the total loss as a convex combination of the classification and segmentation losses: $Loss = a * Loss_s + (1-a) * Loss_c$.

We train our model with a variant of Stochastic Gradient Descent (SGD) with momentum [Rumelhart 1988], used also in our previous work [Mlynarski 2019b].

The main differences with the standard SGD are to divide the gradient by its norm and to compute gradients on several training batches in each iteration, in order to take into account many training examples while bypassing GPU memory constraints.

3.3.2 Extension to the multiclass problem

We extend our model to the multiclass case where each pixel has to be labelled with one of K classes, such as the four ones considered in BRATS challenge (non-tumor, contrast-enhancing core, edema, non-enhancing core). We now assume that image-level labels are provided for each class (absent/present in the image).

Extension of the segmentation subnetwork to the multiclass problem is straightforward, by changing the number of final feature maps to match the number of classes. However, image-level labels are not exclusive, i.e. an image may contain several tumor subclasses. For this reason, we propose to consider one image-level classification output per tumor subclass, indicating absence or presence of the given subclass.

According to our experiments, better performances are obtained when each subclass has its dedicated entire classification branch (Fig. 3.3). A possible reason is that the image-level classification of tumor subclasses is a challenging task requiring a sufficient number of dedicated parameters.

Training batches are sampled similarly to the binary case, however each tumor subclass has to be present at least once in each training batch. In our implementa-

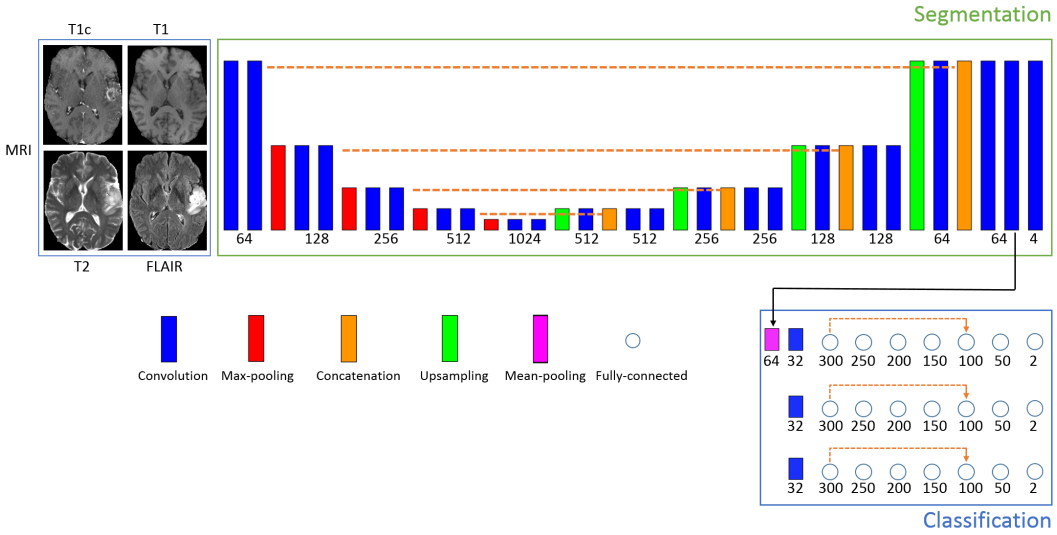

Figure 3.3: Extension of our model to the multiclass problem. The number of final feature maps of the segmentation subnetwork is equal to the number of classes (4 in our case). As image-level labels (class present/absent) are not exclusive, we consider one classification branch per tumor subclass.

tion, we store lists of paths of images containing tumor subclasses in order to sample from these lists during the training of the model.

In the segmentation loss we empirically fix the following target weights for the four classes (non-tumor, non-enhancing tumor core, edema, enhancing-core): $t_0 = 0.7$, $t_1 = 0.1$, $t_2 = 0.1$, $t_3 = 0.1$ (all tumor subclasses have an equal weight in the loss function). The loss associated with each classification branch is the same as in the binary case and the total classification loss is the average across all classification branches.

3.4 Experiments

3.4.1 Data

We evaluate our method on the challenging task of brain tumor segmentation in multisequence MR scans, using the *Training* dataset of BRATS 2018 challenge. It contains 285 multisequence MRI of patients diagnosed with low-grade gliomas or high-grade gliomas. For each patient, manual ground truth segmentation is provided. In each case, four MR sequences are available: T1, T1+gadolinium, T2 and FLAIR (Fluid Attenuated Inversion Recovery). Preprocessing performed by the organizers includes skull-stripping, resampling to 1 mm^3 resolution and registration of images to a common brain atlas. The resulting volumes are of size 240x240x155. The images were acquired in 19 different imaging centers. In order to normalize image intensites, each image is divided by the median of non-zero voxels (which is supposed to be less affected by the tumor zone than the mean) and multiplied the image by a fixed constant.

Each voxel is labelled with one of the following classes: non-tumor (class 0), contrast-enhancing core (class 3), non-enhancing core (class 1), edema (class 2). The benchmark of the challenge groups classes in three regions: *whole tumor* (formed by all tumor subclasses), *tumor core* (classes 1 and 3, corresponding to the visible tumor mass) and *enhancing core* (class 3).

Given that all 3D images of the database contain tumors (no negative cases to train a 3D classification network), we consider the 2D problem of tumor segmentation in axial slices of the brain.

3.4.2 Test setting

The goal of our experiments is to compare our approach with the standard supervised learning. In each of the performed tests, our model is trained on fully-annotated and weakly-annotated images and is compared with the standard U-Net trained on fully-annotated images only. The goal is to compare our model with a commonly used segmentation model on a publicly available database.

We consider three diffferent training scenarios, with a varying number of patients for which we assume a provided manual tumor segmentation. In each scenario we perform a 5-fold cross-validation. In each fold, 57 patients are used for test and

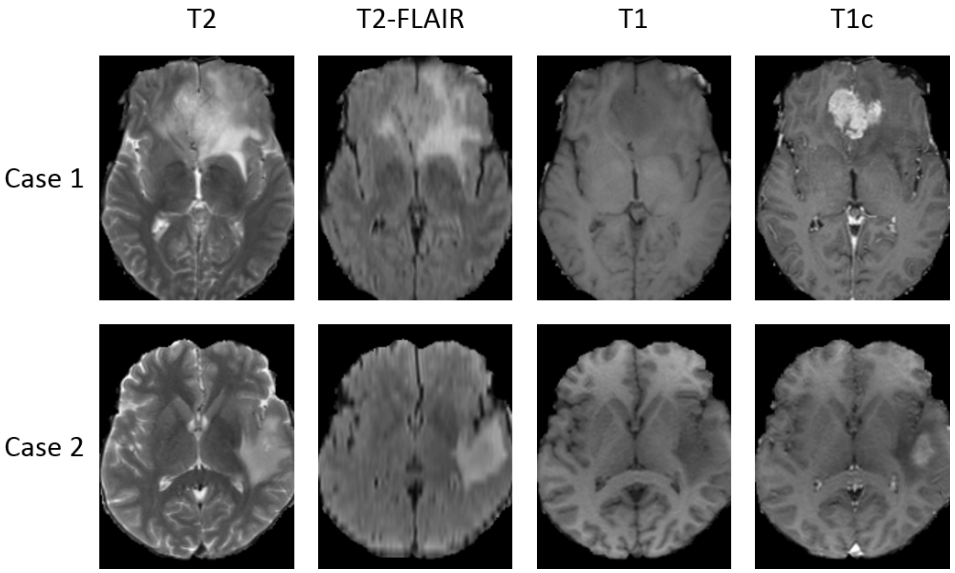

Figure 3.4: Examples of multisequence MRI from the BRATS 2018 database. While T2 and T2-FLAIR highlight the edema induced by the tumor, T1 is suitable for determining the tumor core. In particular, T1 acquired after injection of a contrast agent (T1c) highlights the tumor angiogenesis, indicating presence of highly proliferative cancer cells.

228 patients are used for training. Among the 228 training images, few cases are assumed to be fully-annotated and the remaining ones are considered to be weakly-annotated, with slice-level labels. The fully-annotated images are different in each fold. If the 3D volumes are numbered from 0 to 284, then in k^{th} fold, the test images correspond to the interval [(k-1)*57, k*57 -1], the next few images correspond to fully-annotated images and the remaining ones are considered as weakly-annotated (the folds are generated in a circular way). In the following, *FA* denotes the number of fully-annotated cases and *WA* denotes the number of weakly-annotated cases (with slice-level labels). In particular, note that the split training/test is on 3D MRIs, i.e. the different slices of the same patient are always in the same set (training or test).

In the first training scenario, 5 patients are assumed to be provided with a manual segmentation and 223 patients have slice-level labels. In the second and the third scenario, the numbers of fully-annotated cases are respectively 15 and 30 and the numbers of weakly-annotated images are therefore respectively 213 and 198. The three training scenarios are independent, i.e. folds are re-generated randomly (the list of all images is permuted randomly and the folds are generated). In fact, results are likely to depend not only on the number of fully-annotated images but also on qualitative factors (for example the few fully-annotated images may correspond to atypical cases), and the goal is to test the method in various settings. Overall, our approach is compared to the standard supervised learning on 60 tests (5-fold cross-validation, three independent training scenarios, three binary problems and

one multiclass problem).

We evaluate our method both on binary segmentations problems (separately for each of three tumor regions considered in the challenge) and on the end-to-end multiclass segmentation problem. In each binary case, the model is trained for segmentation and classification of one tumor region (whole tumor, tumor core or enhancing core).

Segmentation performance is expressed in terms of Dice score quantifying the overlap between the ground truth (Y) and the output of a model ($\tilde{Y}$):

$$DSC(\tilde{Y}, Y) = \frac{2|\tilde{Y} \cap Y|}{|\tilde{Y}| + |Y|} \tag{3.1}$$

In order to measure the statistical significance of obtained results, we perform two-tailed and paired t-tests. Pairs of observations correspond to segmentation scores obtained with the standard supervised learning (U-Net trained on fully-annotated images) and with our approach. Dice scores for all patients from 5 folds are concatenated to form a set of 285 pairs of observations. The statistical test is performed for each training scenario and for each segmentation task (three binary problems and one multiclass problem). We consider a significance level of 5%.

3.4.3 Model hyperparameters

3.4.3.1 Loss function and training of the model

The main introduced training hyperparameter is the parameter a, corresponding to the trade-off between classification and segmentation losses. We report mean Dice scores obtained with a varying value of the parameter a, on a validation set of 57 patients (20% of the database used for testing and 80% used for training) in the case with 5 fully-annotated cases and 223 weakly-annotated cases. Segmentation accuracy obtained for the *whole tumor* in the binary case is reported on Fig. 3.5. The peak of performance is observed for $a = 0.7$ (improvement of approximately 12 points of Dice over the standard supervised learning on this validation set), i.e. for the configuration where the segmentation loss accounts for 70% of the total loss. With high values of a, the improvement over the standard supervised learning is limited: around 2.5 points of Dice for $a = 0.9$. In fact, setting a high value of a corresponds to giving less importance to the image-level classification task and therefore ignoring weakly-annotated images. For too low values of a, segmentation accuracy decreases too, probably because the model focuses on the secondary task, of image-level classification. In the end-to-end multiclass case (Fig. 3.6), lower values of a seem more suitable, possibly because of an increased complexity of the image-level classification task. In all subsequent tests, we fix $a = 0.7$ for binary segmentations problems and $a = 0.3$ for the end-to-end multiclass segmentation.

Training batches in our experiments contain 10 images, including 8 images with tumors (4 images with provided tumor segmentation and 4 without provided segmentation) and 2 images without tumors. The number of images was fixed to fit in

the memory of the used GPUs (Nvidia GeForce GTX 1080 Ti), i.e. to form training batches for which Backpropation can be performed using the memory of the GPU. In each training batch there are only 2 images without tumors because most of the pixels of tumor images correspond to non-tumor zones.

The parameters t_c, corresponding to target weights of classes in the segmentation loss, were fixed manually. Both in binary and multiclass cases, we chose $t_0 = 0.7$, which corresponds to giving a target weight of 70% to non-tumor voxels. In fact, tumor pixels represent approximately 1% of pixels of the training batch and therefore non-tumor pixels account approximately for 99% of non-weighted cross-entropy segmentation loss. With $t_0 = 0.7$, relative weight of non-tumor pixels is therefore decreased compared to the standard, non-weighted cross entropy, while still giving the non-tumor class a high weight in order to avoid oversegmentation. In the multiclass setting, we fixed the same target weight to all three tumor subclasses, i.e. $t_1 = 0.1$, $t_2 = 0.1$, $t_3 = 0.1$. As a good convergence of the training was obtained in terms of Dice scores of tumor subclasses, we did not further need to optimize these hyperparameters. Morever, U-Net trained with these weights and using 228 fully-annotated images obtained a mean Dice score of almost 0.87 for *whole tumor* (last row of Table 3.1), which is a satisfactory performance for a model independently processing axial slices without any postprocessing.

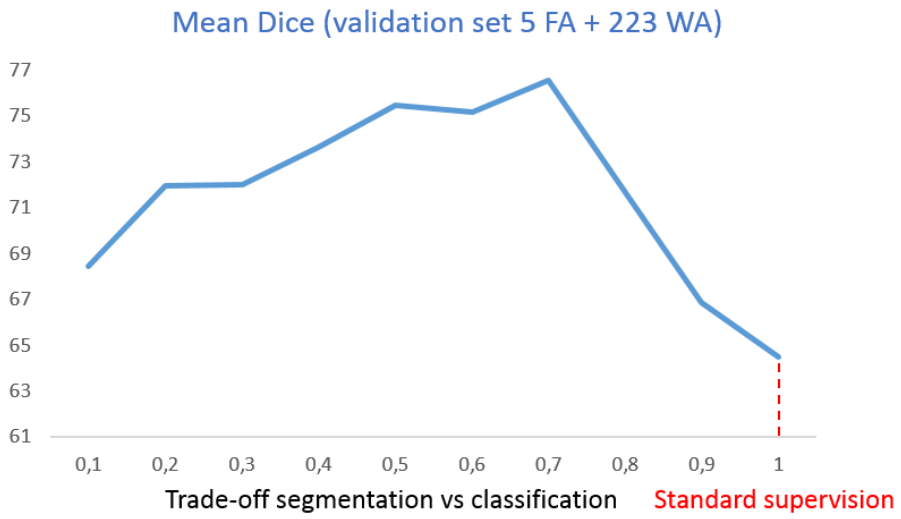

Figure 3.5: Mean Dice scores for the 'whole tumor' region obtained with a varying value of the parameter 'a', corresponding to the trade-off between segmentation and image-level classification losses. Segmentation scores are evaluated on a validation set of 57 MRI in the training scenario where 5 fully-annotated MRI and 223 weakly-annotated MRI are available for training. The case a=1.0 corresponds to ignoring the classification loss and therefore ignoring weakly-annotated images.

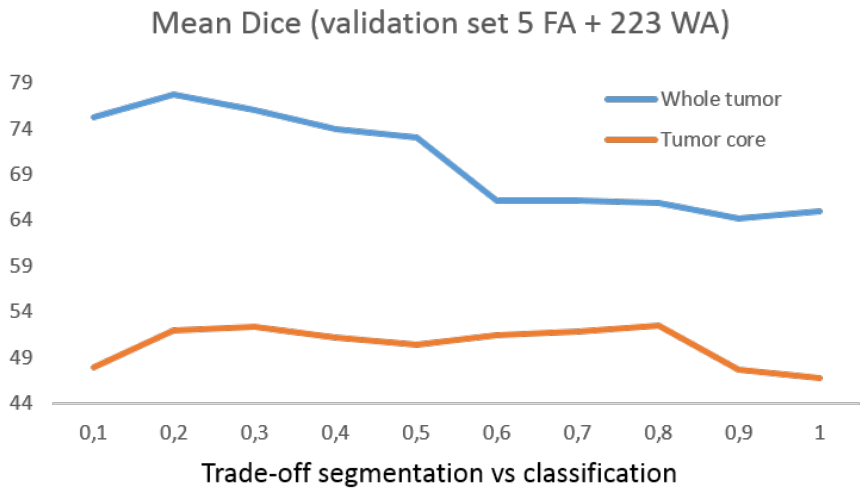

Figure 3.6: Mean Dice scores for 'whole tumor' and 'tumor core' regions obtained with a varying value of the parameter 'a' in the multiclass case. Segmentation scores are evaluated on a validation set of 57 MRI in the training scenario where 5 fully-annotated MRI and 223 weakly-annotated MRI are available for training. The case a=1.0 corresponds to ignoring the classification loss and weakly-annotated images.

3.4.3.2 Model architecture

One of the most important attributes of our method is the architecture of classification branches extending segmentation networks. We perform experiments to compare our model with alternative types of architectures of classification subnetworks. We report the segmentation accuracy obtained on the previously defined validation set of 57 patients.
In the binary case, we consider two alternative architectures of classification subnetworks. The first one is composed of four fully-connected layers having respectively 2000, 500, 100 and 2 neurons. It corresponds therefore to a shallow variant of the classification subnetwork with a relatively high number of parameters. We name this architecture *Shallow* model. The second variant has the same architecture as our model (7 fully-connected layers) but with removed concatenation between the first and the fifth fully-connected layer. We name this architecture *Deep-sequential*. The comparison of segmentation accuracy for *whole tumor* obtained by these two variants and by our model is reported on Fig. 3.7. All three models using mixed level of supervision obtain a better segmentation accuracy than the standard U-Net using 5 fully-annotated images (64.48). Among the three architectures, the shallow variant yields the lowest accuracy (72.29). Our model obtains the highest accuracy (76.56) and performs slightly better than its counterpart with removed concatenation, *Deep-sequential* model (75.78). The improvements over the standard model and the *Shallow* model were found statistically significant (two-tailed and paired t-test).

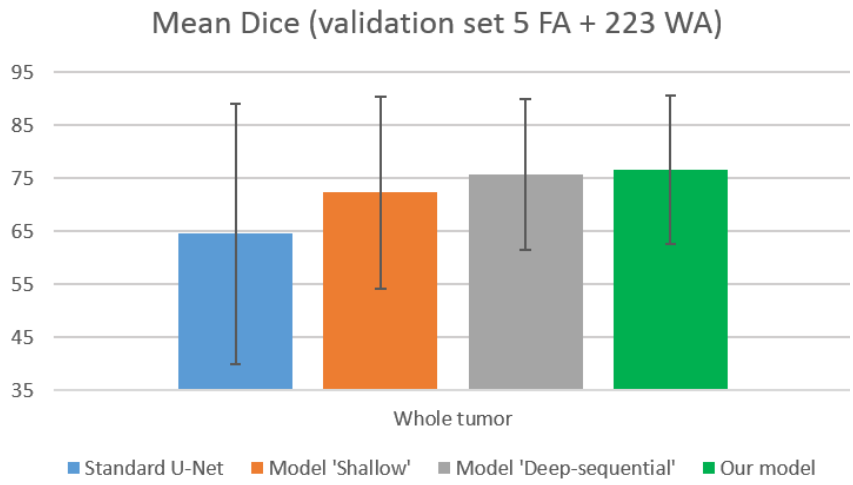

Figure 3.7: Mean Dice scores for the 'whole tumor' region obtained by the standard U-Net and by different models using mixed level of supervision. Standard deviations are represented by error bars. The segmentation scores are evaluated on a validation set of 57 MRI in the training scenario where 5 fully-annotated MRI and 223 weakly-annotated MRI are available for training. Our model corresponds to U-Net extended with a classification branch composed of 7 fully-connected layers and containing one skip-connection.

We also report results obtained with an alternative architecture of the multi-class model. In our model, we considered separate classification branches for all tumor subclasses. We consider an alternative architecture, having only one classification branch (with the same architecture as our model for binary segmentation and classification) shared between the three final fully-connected layers performing image-level classification. In this configuration, the classification layer of each tumor subclass takes as input the 6th fully-connected layer of the shared classification branch. We name this architecture *Shared classification*. The comparison with our multiclass model (separate classification branches for all tumor subclasses) on the same validation set as previously is reported on Fig. 3.8. Our model obtains the highest accuracy for the three tumor subregions while the alternative model (*Shared classification*) obtains higher accuracy than the standard multiclass U-Net for *whole tumor* and *tumor core*. The improvements of our model over the standard model were found statistically significant for *whole tumor* and *tumor core* regions. The improvements over the alternative model with mixed supervision (*Shared classification*) were not found statistically significant (p-values > 0.05).

3.4.4 Results

The main observation is that our model with mixed supervision provides a significant improvement over the standard supervised approach (U-Net trained on fully-annotated images) when the number of fully-annotated images is limited. In the two

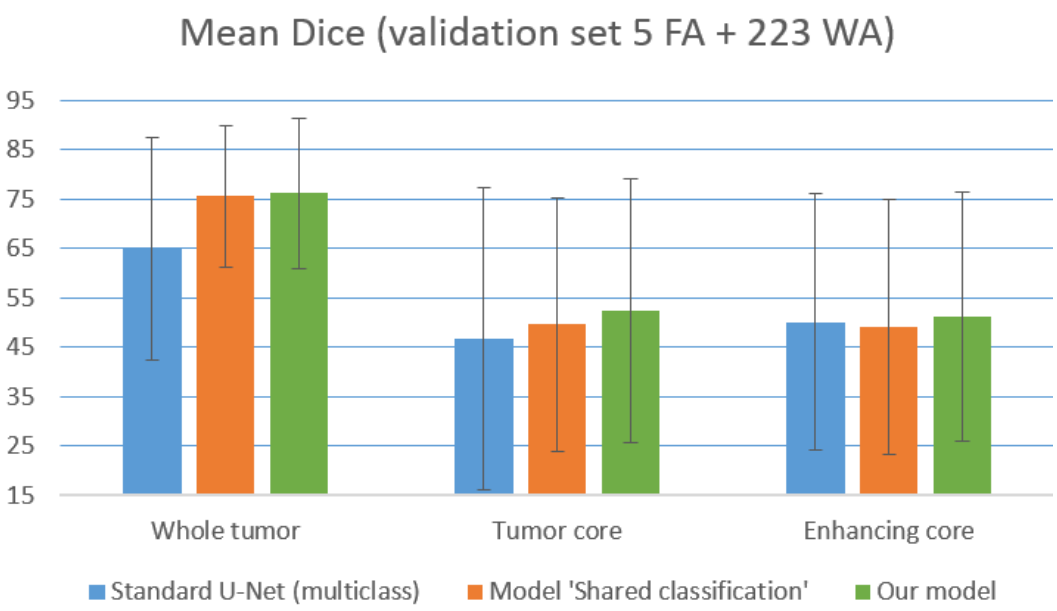

Figure 3.8: Mean Dice scores for the 'whole tumor' region obtained by the standard multiclass U-Net and by different multiclass models using mixed level of supervision. The error bars represent standard deviations. The segmentation scores are evaluated on a validation set of 57 MRI in the training scenario where 5 fully-annotated MRI and 223 weakly-annotated MRI are available for training. Our model is multiclass U-Net extended with three separate classification branch (for each tumor subclass), each branch having the same architecture as in the binary segmentation/classification problem.

first training scenarios (5 FA and 15 FA), our model outperformed the supervised approach on the three binary segmentation problems (Table 3.1) and in the multiclass setting (Table 3.3). The largest improvements are in the first scenario (5 FA) for the *whole tumor* region where the improvement is of 8 points of the mean Dice score in the binary setting and of 9 points of Dice in the multiclass seting. Results on different folds of the second scenario (intermediate case, 15 FA) are displayed in Table 3.2 for the binary problems and in Table 3.4 for the multiclass problem. Our approach provided an improvement in all folds of the second scenario and for all tumor regions, except one fold for *enhancing core* in the binary setting. In the third scenario (30 FA + 198 WA), our approach and the standard supervised approach obtained similar performances. Furthermore, we observe that standard deviations are consistently lower with our approach, in all training scenarios and for all tumor subregions. The results obtained with mixed supervision are therefore more stable than the ones obtained with the standard supervised learning.

All improvements were found statistically significant for binary segmentations problems. In the multiclass case, all improvements were found statistically significant except for *enhancing core* in the first training scenario and for *whole tumor* in

the third training scenario.

Qualitative results are displayed on Figures 3.9, 3.10 and 3.11. Each figure shows segmentations of one tumor region (*whole tumor*, *tumor core*, *enhancing core*) produced by models trained with a varying number of fully-annotated and weakly-annotated images available for training.

Segmentation performance increases quickly with the first fully-annotated cases, both for the standard supervised learning and the learning with mixed supervision. For instance, mean Dice score obtained by the supervised approach for *whole tumor* increases from 70.39, in the case with 5 fully-annotated images, to 77.9 in the case with 15 fully-annotated images. Our approach using 5 fully-annotated images and 223 weakly-annotated images obtained a slightly better performance (78.3) than the supervised approach using 15 fully-annotated cases (77.9). This result is represented on Fig. 3.12.

On Fig. 3.13, we report cross-validated results obtained with a varying number of weakly-annotated while images keeping a fixed number of fully-annotated images. This complementary experiment is performed for segmentation of *whole tumor* in the first training scenario (5 fully-annotated images). We observe that the improvement slows down with the number of added weakly-annotated scans. Inclusion of the first 100 weakly-annotated MRIs yields an improvement of approximately 5 points of the cross-validated mean Dice score (from 70.39 to 75.28), while addition of the remaining 123 weakly-annotated images improves this score by 3 points (from 75.28 to 78.34).

Note that each fully-annotated case corresponds to a large 3D volume with voxelwise annotations. Each manually segmented axial slice of size 240x240 corresponds to 57 600 labels, which represents indeed a huge amount of information compared to one global label simply indicating presence of absence of a tumor tissue within the slice.

Table 3.1: Mean Dice scores (5-fold cross-validation, 57 test cases in each fold) in the three binary segmentation problems obtained by the standard supervised approach and by our model trained with mixed supervision. The numbers in brackets denote standard deviations computed on the distribution of Dice scores for all patients of the 5 folds. The asterisks represent statistically significant improvements (p-value < 0.05) provided by our method compared to the standard supervised learning.

	Whole Tumor	Tumor Core	Enhancing core
Standard supervision 5 FA	70.39 (21.78)	48.14 (28.31)	55.74 (26.73)
Mixed supervision 5 FA + 223 WA	**78.34*** (13.01)	**50.11*** (25.95)	**60.06*** (22.72)
Standard supervision 15 FA	77.91 (16.77)	58.33 (29.00)	62.88 (25.80)
Mixed supervision 15 FA + 213 WA	**80.92*** (11.17)	**63.23*** (26.40)	**66.61*** (23.12)
Standard supervision 30 FA	**83.95** (11.84)	66.17 (25.61)	**69.15** (23.51)
Mixed supervision 30 FA + 198 WA	83.84 (9.68)	**68.30*** (23.73)	67.18 (21.69)
Standard supervision 228 FA	86.80 (8.47)	77.09 (18.58)	72.20 (19.11)

Table 3.2: Results obtained for the three binary problems (whole tumor, tumor core, enhancing core) on differents folds in the case with 15 fully-annotated images and 213 weakly-annotated images. The numbers in brackets denote standard deviations computed on the distribution of Dice scores for all patients.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Total
Standard, WT	76.23 (14.68)	78.15 (19.24)	78.13 (16.88)	77.67 (18.46)	79.35 (13.76)	77.91 (16.77)
Mixed, WT	**82.36** (9.28)	**81.03** (10.21)	**78.96** (12.47)	**79.88** (13.60)	**82.35** (9.16)	**80.92** (11.17)
Standard, TC	61.46 (28.94)	61.17 (26.55)	56.68 (27.90)	56.42 (28.63)	55.94 (32.17)	58.33 (29.00)
Mixed, TC	**63.15** (25.92)	**66.82** (21.74)	**63.45** (26.73)	**60.83** (27.22)	**61.91** (29.40)	**63.23** (26.40)
Standard, EC	66.33 (24.51)	61.08 (26.49)	57.86 (25.85)	**68.09** (22.40)	61.02 (27.82)	62.88 (25.80)
Mixed, EC	**68.72** (23.66)	**70.65** (17.91)	**60.34** (25.84)	67.55 (20.49)	**65.80** (25.46)	**66.61** (23.12)

In terms of the annotation cost, manual delineation of tumor tissues in one MRI may take about 45 minutes for an experienced oncologist using a dedicated segmentation tool. Determing the range of axial slices containing tumor tissues may take 1-2 minutes but can be done without a specialized software. More importanty, determining global labels may require less medical expertise than performing an exact tumor delineation and can therefore be performed by a larger community.

Table 3.3: Mean Dice scores (5-fold cross-validation, 57 test cases in each fold) obtained by the standard supervised approach and by our model in the multiclass setting. The numbers in brackets denote standard deviations computed on the distribution of Dice scores for all patients of the 5 folds. The asterisks represent statistically significant improvements (p-value < 0.05) provided by our method compared to the standard supervised learning.

	Whole Tumor	Tumor Core	Enhancing core
Standard supervision 5 FA	67.61 (22.24)	51.12 (26.98)	58.15 (24.65)
Mixed supervision 5 FA + 223 WA	**76.64*** (14.14)	**56.30*** (22.65)	**58.19** (23.05)
Standard supervision 15 FA	74.46 (18.04)	59.87 (25.97)	61.85 (24.86)
Mixed supervision 15 FA + 213 WA	**79.39*** (12.99)	**63.91*** (24.72)	**65.71*** (23.07)
Standard supervision 30 FA	81.10 (14.29)	**67.48** (24.78)	**68.67** (22.79)
Mixed supervision 30 FA + 198 WA	**81.23** (10.90)	66.33 (24.12)	67.69 (21.87)
Standard supervision 228 FA	85.67 (9.66)	78.78 (18.31)	74.14 (19.62)

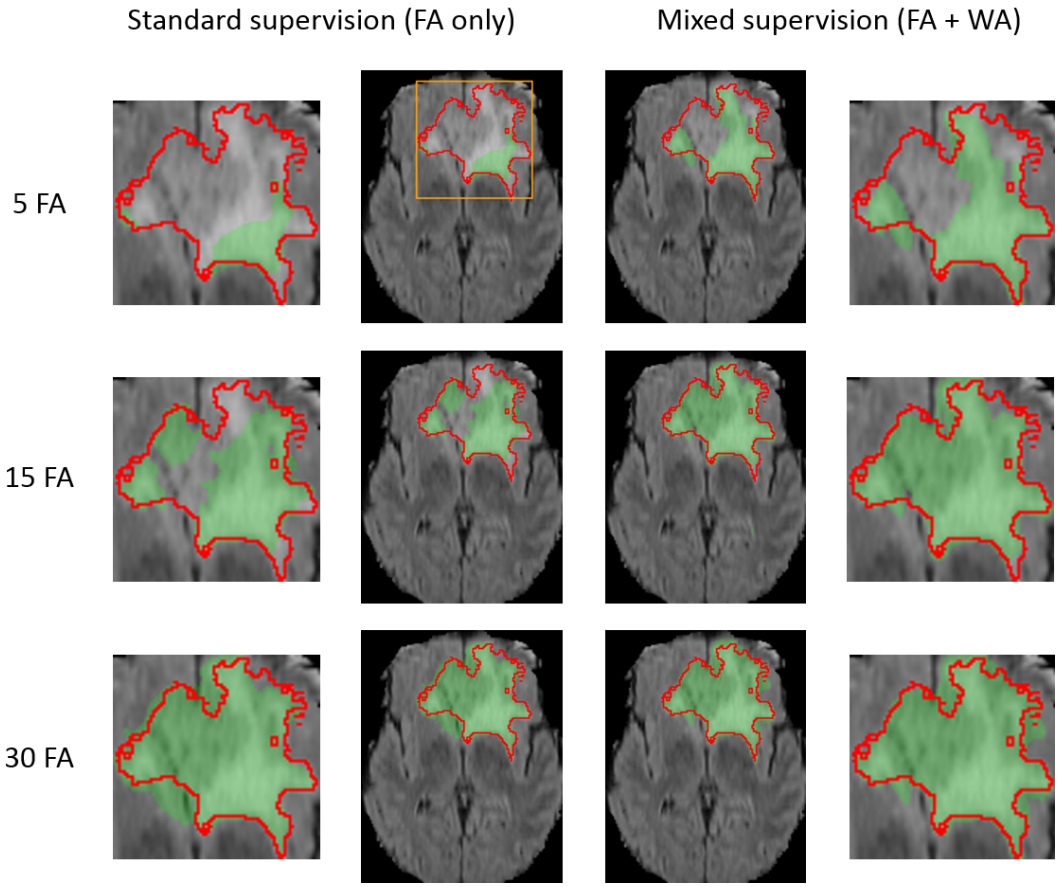

Figure 3.9: Comparison of our approach with the standard supervised learning for binary segmentation of the 'whole tumor' region. Each row represents the same test example (first image of Fig. 4) from a different training scenario (5, 15 or 30 fully-annotated scans available for training). FA and WA refer respectively to the number of fully-annotated MRI and weakly-annotated MRI (with slice-level labels). The results are displayed on MRI T2-FLAIR sequence. The performance of both models improves with the number of manual segmentations available for training.

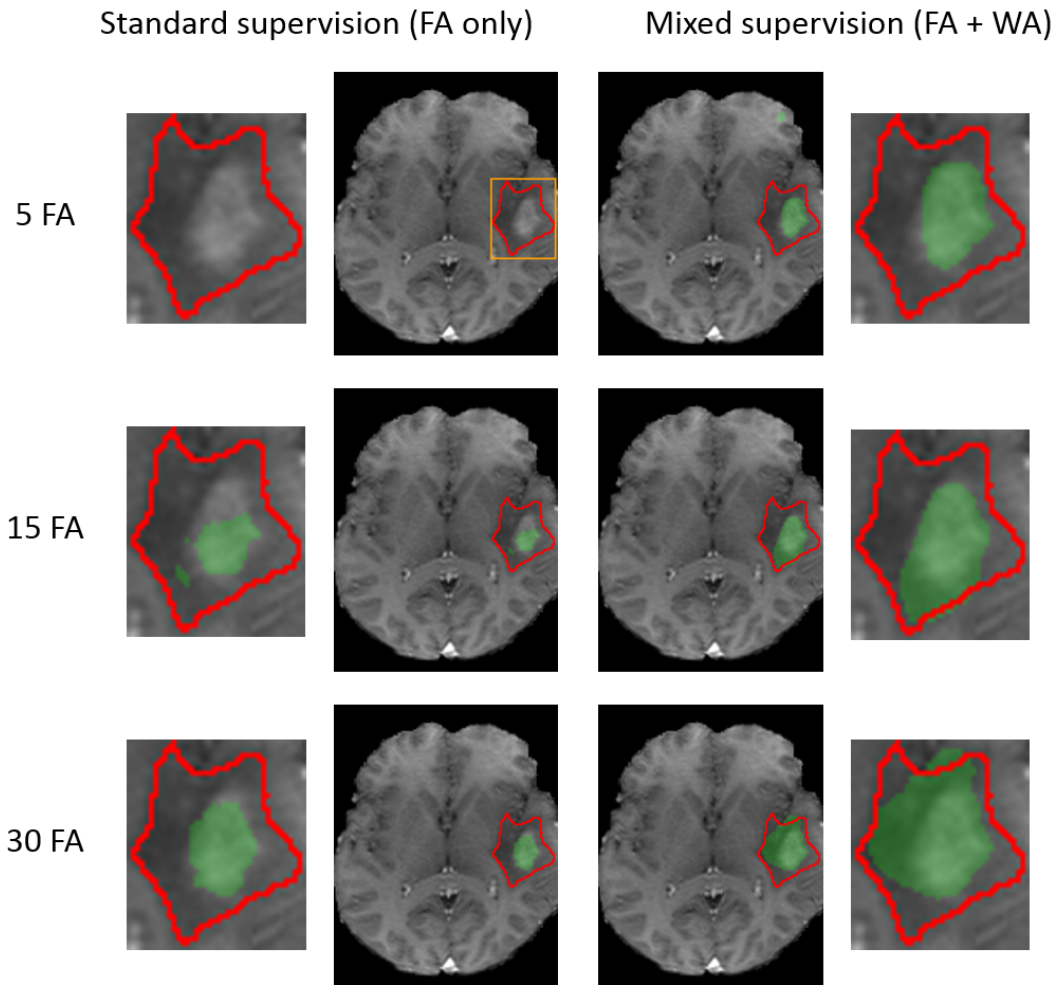

Figure 3.10: Comparison of our approach with the standard supervised learning for binary segmentation of the 'tumor core' region (test example corresponding to the bottom image of Fig. 4). Each row corresponds to a different training scenario (5, 15 or 30 fully-annotated scans available for training). FA and WA refer to the numbers of fully-annotated and weakly-annotated scans. The results are displayed on MRI T1+gadolinium. The observations are similar to the problem of binary segmentation of the 'whole tumor' region. In particular, in the first training scenario, the standard supervised approach does not detect the tumor core zone, in contrast to our method.

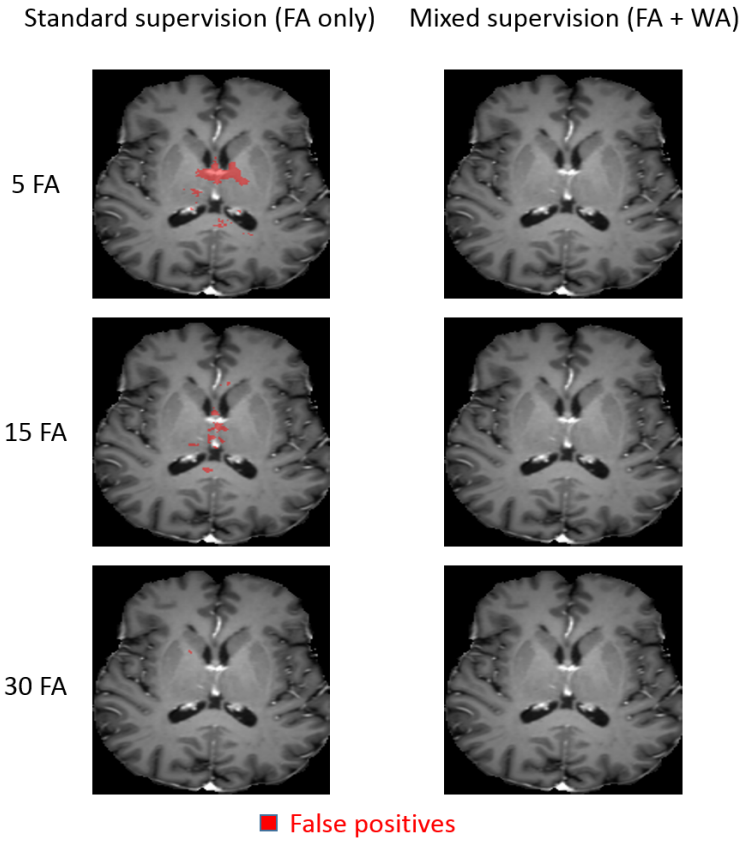

Figure 3.11: Comparison of our approach with the standard supervised learning for binary segmentation of the 'enhancing core' region. Each row corresponds to a different training scenario (5, 15 or 30 fully-annotated scans available for training). FA and WA refer to the numbers of fully-annotated and weakly-annotated scans. The results are displayed on MRI T1+gadolinium. The example shows false positives obtained by the model trained with standard supervision. The number of false positives decreases with the number of fully-annotated images available for training. No false positives are observed for our model trained with mixed supervision, in any of the three training scenarios.

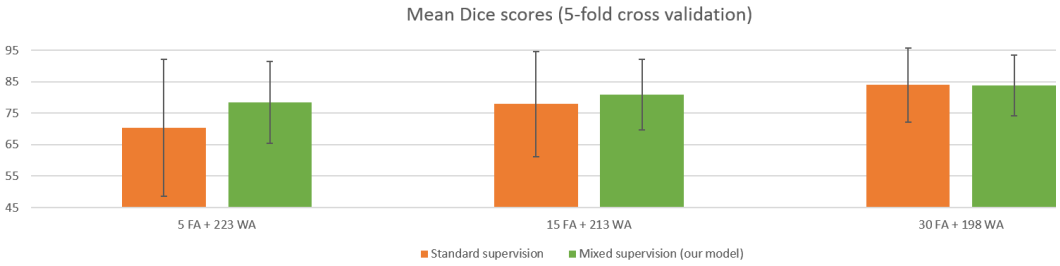

Figure 3.12: Illustration of the improvement provided by the mixed supervision for binary segmentation of the 'whole tumor' region (mean Dice scores and their standard deviations). Mixed supervision using 5 fully-annotated MRI and 223 weakly-annotated MRI obtains a slightly better performance than the standard supervised approach using 15 fully-annotated MRI. The improvement provided by the weakly-annotated images decreases with the number of available ground truth segmentations.

Table 3.4: Results obtained in the multiclass setting on differents folds in the case with 15 fully-annotated images and 213 weakly-annotated images. The numbers in brackets denote standard deviations computed on the distribution of Dice scores for all patients.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Total
Standard, WT	74.31 (13.78)	78.91 (15.41)	67.57 (23.14)	75.55 (17.59)	75.96 (16.85)	74.46 (18.04)
Mixed, WT	**77.53** (12.81)	**82.20** (9.39)	**73.72** (16.37)	**80.96** (13.40)	**82.55** (9.38)	**79.39** (12.99)
Standard, TC	61.17 (23.64)	63.89 (22.79)	55.72 (26.34)	55.36 (28.33)	63.18 (27.06)	59.87 (25.97)
Mixed, TC	**62.83** (24.65)	**65.26** (22.63)	**62.23** (25.82)	**61.99** (27.87)	**67.23** (21.74)	**63.91** (24.72)
Standard, EC	66.15 (24.58)	64.83 (23.14)	53.83 (25.52)	61.68 (24.38)	62.77 (24.77)	61.85 (24.86)
Mixed, EC	**68.33** (21.70)	**68.39** (18.55)	**59.51** (26.07)	**68.63** (21.76)	**63.70** (25.14)	**65.71** (23.07)

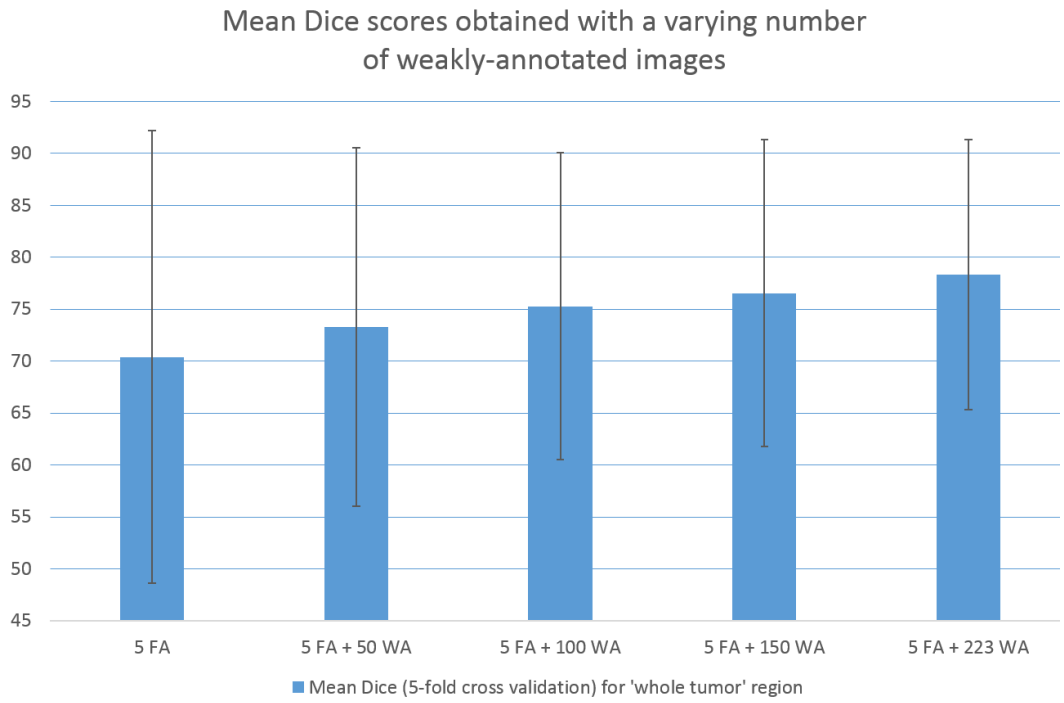

Figure 3.13: Mean Dice scores (5-fold cross-validation, 57 test cases in each fold) obtained for binary segmentation of *whole tumor* with training on 5 fully-annotated scans and a varying number of weakly-annotated scans. The error bars represent standard deviations.

3.5 Conclusion and future work

In this work we proposed a new deep learning approach for tumor segmentation which takes advantage of weakly-annotated medical images during the training of neural networks, in addition to a small number of manually segmented images. In our approach, we propose to use neural networks producing both voxelwise and image-level outputs. The classification and segmentation subnetworks share most of their layers and are trained jointly using both fully-annotated and weakly-annotated data. We performed a large number of cross-validated experiments to test our method both in binary and multiclass settings. Our experiments showed that the use of weakly-annotated data improves the segmentation performance significantly when the number of manually segmented images is limited. Our model is end-to-end and straightforward to implement with common deep learning libraries such as Theano [Bergstra 2010] or TensorFlow [Abadi 2016]. In order to encourage other researchers to continue the research in the field, the code of our method will be made publicly available on `https://github.com/PawelMlynarski/segmentation_mixed_supervision`.

In our work we focused on the 2D segmentation problem, in particular because all 3D images from the BRATS 2018 database contain tumors whereas we also need non-tumor images to train the classification part of our model. A practical difficulty of collecting databases containing both tumor and non-tumor 3D scans is the heterogeneity of available imaging modalities. For example, MRI + gadolinium, commonly used for tumor imaging, is generally available for patients with suspected tumors or vascular problems (requiring imaging of blood vessels using a contrast agent). In this work, we chose to focus only on the problem of available ground truth annotations, assuming availability of the same imaging modalities for all patients, both for supervised learning and learning with mixed supervision. Dealing with the variability of available modalities is a very important problem of medical imaging and is beyond the scope of this work.

Extension of our model to an end-to-end segmentation of entire 3D scans could be difficult with the current GPUs because of computational costs of CNNs. One advantage of a 3D model would be to take into account a richer spatial context in the case of MRI or CT scans. Furthermore, volume-level labels require less effort than slice-level labels and would therefore be easier to obtain, even if these labels are also less informative. However, 2D CNNs still perform reasonably well on 3D scans. As reported in the last row of Table 3.1, U-Net processing independently axial slices obtains a mean Dice of almost 0.87 for the *whole tumor* region and of 0.77 for the *tumor core* region, using 228 fully-annotated images (80% of the database of BRATS), without any postprocessing.

In our tests, we used approximately 220 weakly-annotated MRI, which is relatively a limited number. An important future step would be to test our method on a database containing a considerably larger number of weakly-annotated images (thousands, millions).

Chapter 4

Anatomically Consistent Segmentation of Organs at Risk in MRI with Convolutional Neural Networks

Contents

In this chapter, published as a journal article [Mlynarski 2019a], we focus on segmentation of organs at risk in the context of radiotherapy of brain tumors. Accurate segmentation of healthy organs is a necessary and time-consuming step of radiotherapy planning. We propose a CNN-based system for segmentation of multiple and non-exclusive anatomical structures (overlaps between classes) and we propose methods to enforce the anatomical consistency of the result. We constructed a database of MRIs acquired in the Centre Antoine Lacassagne (Nice, France). Our method is evaluated on real clinical data, both quantitatively (with cross-validation and using several metrics) and qualitatively, by an experienced radiotherapist. Our system is able to produce accurate segmentations of several anatomical structures in the brain region despite several challenging aspects such as the presence of tumors (deforming healthy structures) and the natural anatomical variablity between patients.

4.1 Introduction and related work

Malignant tumors of the central nervous system cause more than 200 000 deaths per year worldwide [Vos 2016]. Many brain cancers are treated with radiotherapy, often combined with other types of treatment, in particular surgery and chemotherapy. Radiotherapy planning requires segmentation of target volumes (visible tumor mass and areas likely to contain tumor cells) and anatomical structures surrounding lesions. The segmented volumes are used for computation of optimal irradiation doses, with the objective of maximizng irradiation of cancer cells while minimizing damage of neighboring healthy structures, called *organs at risk* (OAR). Magnetic Resonance (MR) images [Bauer 2013] are commonly used for imaging of tumors and organs in the brain region. In this work, we address the challenging problem of multiclass segmentation of organs in MRI of the brain.

Delineation of organs at risk is today manually performed by experienced clinicians. Due to a large number of structures to be accurately segmented, the segmentation process takes usually several hours per patient. Manual segmentation represents therefore a very high cost and eventually delays the beginning of the therapy. Moreover, a high intra-observer and inter-observer variability is observed [Brouwer 2012]. Automatic methods for segmentation of organs at risk are therefore of particular interest. We can distinguish two main types of approaches proposed

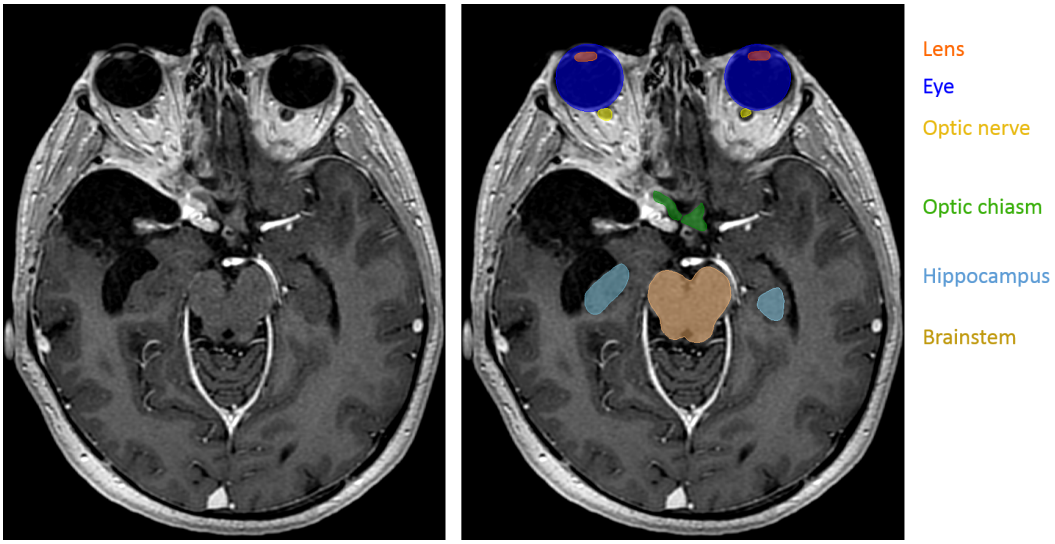

Figure 4.1: Segmentation of organs at risk in radiotherapy planning. Left: T1-weighted MRI acquired after injection of a gadolinium-based contrast agent. Right: manual annotations of several organs at risk. In contrast to standard segmentations problems, one voxel may belong to zero or several classes (for instance, the eye and the lens).

in the literature.

The first type of methods corresponds to atlas-based approaches [Ciardo 2017, Bondiau 2005, Alchatzidis 2015]. The input image is typically registered to one [Commowick 2008, Commowick 2009] or several [Ramus 2010a, Ramus 2010b] annotated images, from which the segmentation is extrapolated. When multiple atlases are used, the candidate segmentations may be combined, for instance, by voting strategies [Ramus 2010b] or by the STAPLE algorithm [Warfield 2004]. An important advantage of atlas-based methods is to produce anatomically consistent results. However, their main drawback is their limited generalization capacity. The important variability between cases results not only from the natural anatomical differences between patients but also from pathological factors. In particular, healthy organs are deformed by growing tumors, which may appear at different locations and which are typically not present in atlases. Some organs may even be missing because of surgeries undergone previously by the patient.
The second group of approaches is based on a discriminative classification of voxels with machine learning models such as Random Forests [Criminisi 2010, Criminisi 2013, Gauriau 2015] or Convolutional Neural Networks (CNN) [LeCun 1995]. These discriminative methods are less constrained than atlas-based approaches and may therefore better adapt to the diversity of cases. However, in general, voxelwise classifiers may produce results which are inconsistent in terms of shapes and locations of organs.

Organs at risk in the brain region have complex shapes and are surrounded by other structures sharing similar voxel intensities in MRI. Moreover, there are large differences related to acquisition of MRI, especially when images come from different medical centers. In order to segment organs from MRI, a complex and abstract information has therefore to be extracted. Convolutional Neural Networks are suitable for this task, as they have the ability to automatically learn complex and relevant image features. In this work, we propose a system based on CNNs for multiclass segmentation of organs at risk in brain MRI. Anatomical consistency of the result is enforced in a postprocessing step.

In this work we assume non-exclusive classes, i.e. that one voxel may belong to zero or several classes (Fig. 4.1). This is in contrast with the large majority of segmentation models, which assign one unique label to each voxel [Long 2015, Kamnitsas 2016]. For OAR segmentation, the previously proposed methods assume either exclusive classes [Zhu 2019, Roth 2017] or non-exclusive classes [Nikolov 2018, Wang 2018a, Larsson 2018, Ibragimov 2017] similarly to our work. An important difficulty to train machine learning models for multiclass OAR segmentation is the varying availability of ground truth segmentations of different classes among patients, depending on clinical needs. While some organs, such as the optic nerve, are systematically segmented during radiotherapy planning, annotation

of other structures may be available only for a subset of patients. One solution to this problem is to independently train one model per class, as it was proposed in some recent deep learning works [Larsson 2018, Ibragimov 2017, Men 2019]. A limitation of this approach is, however, the need to perform time-consuming trainings for every class, while the number of classes of interest may be large. In this work, we propose a loss function and an algorithm to train neural networks for an end-to-end multiclass segmentation, taking into account the problem of missing annotations. To the best of our knowledge, the only deep learning method for end-to-end multiclass OAR segmentation which addresses this issue is the one proposed in [Zhu 2019] for the segmentation of head and neck organs at risk in CT scans.

The network architecture used in our work is a modified version of 2D U-net [Ronneberger 2015]. The choice of a 2D architecture rather than variants of 3D U-Net [Çiçek 2016] is motivated by the ability of 2D CNNs to capture a long-range spatial context without downsampling the image. This property is important in our problem as we segment several anatomical structures in large images, including very small structures such as the lens, the pituitary gland or the optic nerve. 2D CNNs were recently applied in [Kodym 2018] for segmentation of head and neck organs in CT scans.

Even if most of the proposed deep learning methods for OAR segmentation do not apply anatomical constraints on the output of neural networks, some approaches include shape priors in models. For instance, [Tong 2018] propose to learn latent representations of shapes of organs by a stacked autoencoder and to use these learned representations in the loss function of a segmentation network, in order to compare the shape of the output with the shape of the ground truth. The works [Brosch 2018, Orasanu 2018] propose to adapt triangulated meshes representing organ boundaries to medical images and to use neural networks for regression of distances between centers of triangles and organ boundaries. This type of approach may therefore be seen as atlas-based with the use of deep learning for boundary detection.
However, inclusion of constraints related to connectivity and relative positions of organs in loss functions of CNNs is non trivial due to considerable computational costs. In order to apply such constraints, a neural network would have to segment large regions of the input images during the training phase, which requires a considerable amount of GPU memory. To the best of our knowledge, none of the proposed deep learning methods explicitely enforces consistency of OAR segmentation in terms of relative positions of organs. However, some methods define regions of interest of organs, for instance by registering the image to a set of atlases [Larsson 2018].

In our work, we enforce some anatomical constraints in a postprocessing stage, starting from the segmentation produced by majority voting of 2D CNNs processing the image by axial, coronal and sagittal slices. In particular, we propose an

anatomically consistent segmentation of the optic nerves, with an approach based on the search of the shortest path in a graph, using outputs of neural networks to define weights of edges in the graph.

We consider eight classes of interest, corresponding to anatomical structures systematically segmented during radiotherapy planning for brain cancers: eye, lens, optic nerve, optic chiasm, pituitary gland, hippocampus, brainstem and brain (including cerebrum, cerebellum and brainstem). The anatomical structures composed of left and right components (eye, lens, optic nerve, hippocampus) are seen as one entity by the neural network but are separated in the postprocessing step.

Most of the proposed deep learning methods for segmentation of organs at risk were applied on CT scans in the context of head and neck cancers [Argiris 2008], i.e. cancers of the upper parts of respiratory and digestive systems (mouth, larynx, throat). To the best of our knowledge, the only deep learning method for segmentation of organs at risk in MRIs of the brain is the one proposed in [Orasanu 2018] (MRI T1 and T2).

Our method is tested on a set of contrast-enhanced T1-weighted MRIs acquired in the Centre Antoine Lacassagne in Nice (France). First, our method is quantitatively evaluated on a set of 44 MRIs with provided segmentation of different anatomical structures. Segmentation performances are measured by three different metrics: Dice score, Hausdorff distance and the mean distance between the output and the ground truth. Then, the segmentations produced by our method on a different set of 50 MRIs are qualitatively evaluated by an experienced radiotherapist. Our system was able to produce segmentations with an accuracy level which was found acceptable for radiotherapy planning in a large majority of cases (96%). The mean distances between the output segmentation and the ground truth for different organs were between 0.1 mm and 0.7 mm.

4.2 Methods

4.2.1 Deep learning model

4.2.1.1 Network architecture

The architecture used in our work is a modified version of 2D U-Net [Ronneberger 2015], which is composed of an encoding part and a decoding part. The encoding part is a sequence of convolutional and max-pooling layers. The number of feature maps is doubled after each pooling, taking advantage of their reduced dimensions. The decoding part is composed of convolutional and upsampling layers. Feature maps of the encoding part as concatenated in the decoding part in order to combine low-level and high-level features and to ease the flow of gradients during the optimization process. The final convolutional layer (the segmentation layer) of the standard U-Net has two feature maps, representing pixelwise classification scores of the class 0 ('background') and the class 1. During training, these two final feature maps are normalized by the softmax function.

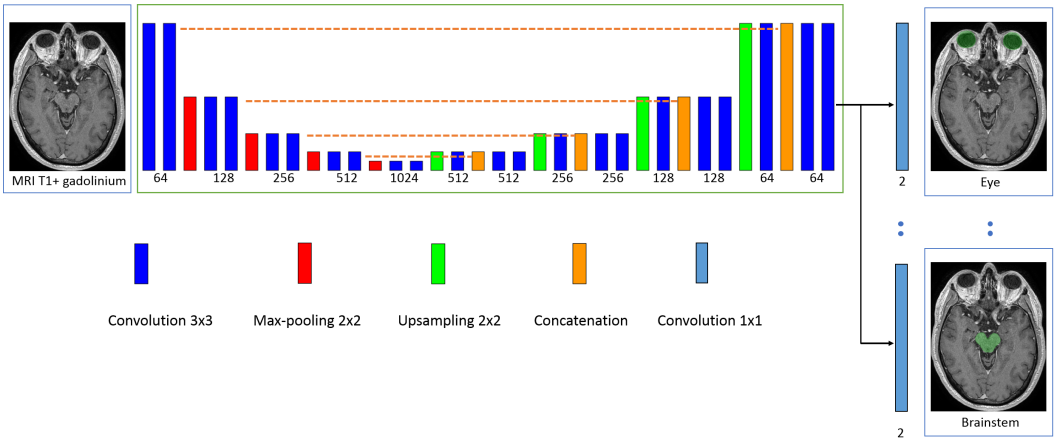

Figure 4.2: Architecture of our model. The rectangles represent layers and their height represents the sampling factor (increasing with max-poolings, decreasing with upsamplings). The numbers of features maps are specified below layers. The proposed model is a modified version of U-Net, having one segmentation layer per class in order to perform an end-to-end multiclass segmentation with non-exclusive classes.

We adapt this architecture to our problem of multiclass segmentation with non-exclusive classes, where each pixel may belong to zero or several classes. In the following, C denotes the number of classes (in our experiments, $C = 8$) and the classes are numbered from 1 to C. In our model, each class c has its dedicated binary segmentation layer (Fig. 4.2), composed of two feature maps corresponding to pixelwise scores of the class and of the background. Each segmentation layer takes as input the second to last convolutional layer of U-Net. We use batch normalization [Ioffe 2015] in all convolutional layers of the network, except segmentation layers.

2D CNNs have the advantage of being able to capture information from distant pixels without the need to downsample the input image. In general, CNNs cannot be trained on whole 3D images such as MRI because of their considerable computional costs and GPU memory limitations. In some recent works [Roth 2017, Wang 2018a], 3D CNNs have been applied sequentially in two steps, on a downsampled version of the image and on the original version. As in this work we simultaneously segment several classes which are generally of small size (some may be hardly visible after downsampling) and distant from each other, a 2D architecture is suitable.

4.2.1.2 Training of the model

Our loss function and training scheme were designed to deal with class imbalance and the problem of missing annotations (for a given image, the ground truth is

available only for a subset of classes).
In a given training image i, each pixel (x,y) has 3 possible labels for the class c: 0 (negative), 1 (positive) or -1 (unknown). If the ground truth segmentation of the class c is unavailable for the image i, all pixels are labelled as unknown for the class c by default. However, missing annotations may be partially reconstructed from segmentations of other classes. For example, if the segmentation of the 'lens' class is not available but the 'eye' class is segmented, all pixels outside the eye may be labelled as negative for the lens.

Given a training batch of M images and the estimated parameters θ of the network, the segmentation layer of the class c is penalized by the following loss function, which can be seen as pixelwise cross-entropy with adaptative weights. Let's note N_0^c, N_1^c and N_{-1}^c the numbers of pixels labelled respectively 0, 1 and -1 for the class c in the training batch. The weight $w_{(x,y)}^i$ of the pixel (x,y) of the image i has three possible values, according to the label of the pixel. If the label is unknown, then $w_{(x,y)}^i = 0$. If its label is 1, then $w_{(x,y)}^i = t_c/N_1^c$ where $0 < t_c < 1$ is a fixed hyperparameter, which we call the *target weight*. If the pixel is labelled 0, then $w_{(x,y)}^i = (1-t_c)/N_0^c$. The introduced hyperparameter t_c controls therefore the relative weight of positive and negative pixels of the class c (positive pixels have the total weight of t_c and negative pixels have the total weight of $1-t_c$). This type of weighting strategy has been used in our previous work [Mlynarski 2019b] to counter the problem of class imbalance. The loss function of the segmentation layer of the class c is defined by $Loss_c(\theta) = -\sum_{i=1}^{M}\sum_{(x,y)} w_{(x,y)}^i \log(p_{i,(x,y)}^l(\theta))$ where $p_{i,(x,y)}^l$ is the softmax score given by the network for the ground truth label l of the pixel.
The loss function of the model is a convex combination of losses of all segmentation layers: $Loss(\theta) = (1/C)\sum_{c=1}^{C} Loss_c(\theta)$.

We propose a sampling strategy to construct training batches so that there are positive and negative pixels for each of the C classes in each training batch.
For each image of the training database, we precompute bounding boxes of all classes with provided segmentations. For bilateral classes such as the eyes, there are generally two bounding boxes per image corresponding to left and right components, unless one of the components is missing (e.g. an organ removed by surgery). The precomputed bounding boxes are used during the training in order to sample patches containing positive pixels of different classes.
At the beginning of the training, for each class c, we construct a list I_c of training images with provided ground truth segmentation of the class c. To sample a 2D patch which is likely to contain positive pixels of the class c, we randomly choose an image i from I_c and a random point (x, y, z) from the bounding box (or two bounding boxes if the class has left and right components) of the class c in the chosen image. Once the point is chosen, a 2D patch (axial, coronal or sagittal) centered on this point is extracted from the image i and segmentations of all available classes are read. In the following, we refer to this procedure as extracting

a patch centered on the class c.
We assume that the number of images in each training batch (M) is larger than the number of classes C, in order to be able to sample at least one image/patch centered on each of the classes. Each training batch is constructed as follows. The first C images of the batch are centered respectively on each of the C classes. At this stage, the batch is likely to contain positive and negative pixels of each class. The remaining $M - C$ images may be chosen randomly or be centered on larger classes. In our case, $C = 8$, $M = 10$ and the last images are centered on the largest class we segment, the brain, whose bounding box occupies almost an entire volume of the head.

As the model is trained for multiclass segmentation with non-exclusive classes, several binary segmentation maps have to be read in each iteration of the training. If the ground truth segmentations are not optimally stored in the memory, these reading operations may considerably slow down the training. The ground truth label of a given pixel can be represented by one bit (0 or 1). However, to store binary segmentation masks in commonly used formats such as HDF5 [Folk 2011], each label would have to be represented by at least one byte. We propose therefore to store multiclass segmentations in a specifically encoded format, where every bit represents a label of a given class c. A binary segmentation mask of the class c is retrieved by the bitwise 'and' operation between the encoded multiclass segmentation and the code of the class, corresponding to a power of 2.
The size of extracted 2D patches should be chosen according to the capacities of the GPU. In our experiments, the training batches were composed of 10 patches of size 230x230. Given that in our network we use unpadded operations (convolutions, max-poolings, etc.), the dimensions of the outputs of segmentation layers are considerably smaller.
The model is trained with a variant of Stochastic Gradient Descent with momentum presented in our previous work [Mlynarski 2019b]. The main characteristics of this algorithm is that gradients are computed over several batches in each iteration of the training, in order to use many training examples despite GPU memory limitations.

4.2.2 Postprocessing and enforcing anatomical consistency

Fully-convolutional neural networks such as our model produce segmentations by individually classifying every voxel based on intensities of voxels within the corresponding receptive field. Such classification is performed by extracting powerful and automatically learned image features. However, as this classification is performed on a voxel by voxel basis, there is no guarantee of obtaining an anatomically consistent result, especially when the number of training images is limited. In particular, CNNs do not explicitely take into account aspects such as relative positions of different structures or adjacency of voxels belonging to the same structure. Including constraints related to these aspects in loss functions of neural

networks or conceiving architectures which produce anatomically consistent results is difficult, in particular because of computational costs (need to simulatenously segment large 3D regions of input images). We propose therefore to improve consistency of segmentations in a postprocessing step. We also separate left and right components of classes such as the eye, as these components are considered separately for radiotherapy planning.
We combine, by majority voting, segmentations produced by three networks trained respectively on axial, coronal and sagittal slices. The goal of this combination is to take into account the three dimensions and to improve the robustness of the method. We subsequently apply a few rules described in the following, in order to correct some observed inconsistencies.

4.2.2.1 Segmentation of the brain

Brain (including the cerebrum, the cerebellum and the brainstem) is the largest class to be segmented. For various reasons, some voxels within this structure may be inconsistently classified as negative by networks, which appear as 'holes' in the segmentation or unrealistically sharp borders. We propose therefore a procedure that we call triplar hole-filling (Fig. 4.3). For each axial, coronal and sagittal plan of the 3D image, we compute connected components of the background (negative voxels) and we remove components (changing their label from 0 to 1) which are not connected to the border of the plan. The reason of applying this procedure in 2D is that some holes may easily be connected to the outside of the class in 3D.

The bounding box of the segmentation of the brain is subsequently used to separate left and right components of bilateral classes. Please note that the head of the patient may appear at different locations of the image, depending on acquisition conditions and performed preprocessings. For a given class expected to have left and right components (eye, lens, optic nerve, hippocampus), barycenter of each connected component is computed. In order to decide to which side corresponds a connected component, the coordinate x (right-left) of its barycenter is compared to min and max coordinates x of the bounding box of the brain.

4.2.2.2 Segmentation of the visual system

We propose an anatomically consistent segmentation of the visual system (eyes, lenses, optic nerves and chiasm), starting from the segmentations predicted by neural networks.
The eye is probably the less challenging organ for automatic segmentation as it has a simple spherical shape. However, some false positives are possible, especially in cases where an eye has been removed by surgery, resulting in false positives within the orbit. We propose therefore to remove connected components of eye

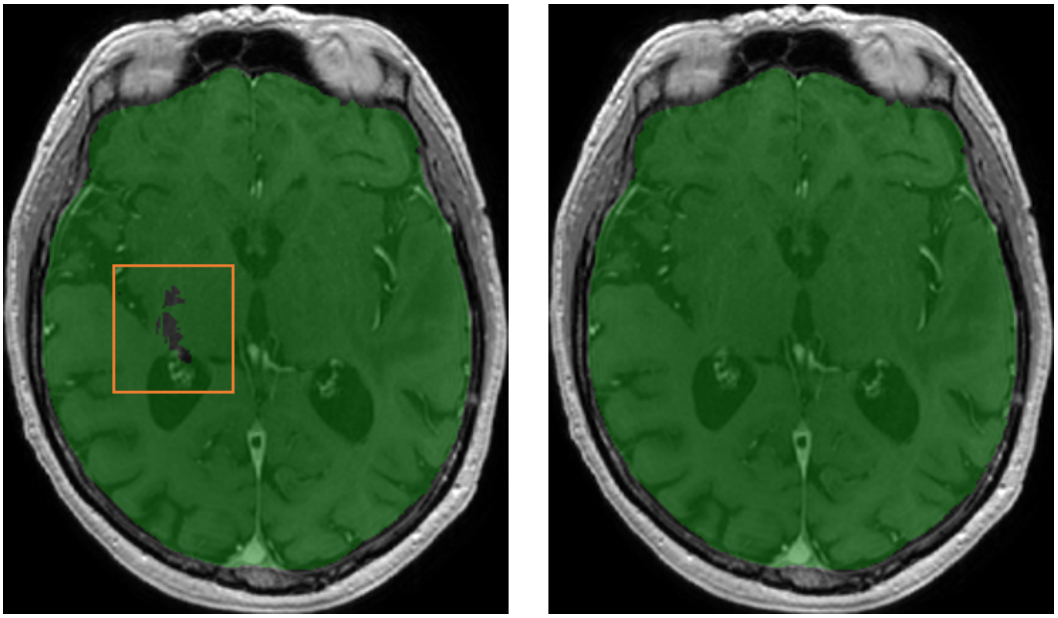

Figure 4.3: Example of 'holes' in the original output segmentation (left image) on a test example. Right image: segmentation obtained after our postprocessing (triplanar hole-filling).

segmentation whose volume is below an expected minimum value, which is set to 4 cm^3.

We constraint segmentation of the lenses to be inside the eyes, i.e. we assign the 0 label to all voxels outside the predicted masks of the eyes. Segmentation of the optic chiasm is obtained by taking the largest connected component of the segmentation predicted by the networks. We distinguish left and right sides of the chiasm in order to compute landmarks for segmentation of the two optic nerves as described in the following.

Segmentation of the optic nerve in MR images is particularly challenging as the nerve is thin and may have an appearance similar to neighboring structures at some locations. However it has a rather regular shape which can be seen as a tube connecting an eye and the optic chiasm. The nerve is generally well visible at some locations, in particular close to the eye. A human expert is able to track the trajectory of the nerve to distinguish it from neighboring structures at more difficult locations. Based on this observation, we propose a graph-based algorihm for segmentation of the optic nerves in order to guarantee connectivity between the eyes and the optic chiasm and to decrease the number of false positives. The algorithm is based on the search of the shortest path between two nodes in a graph. Outputs of neural networks are used to define weights of the edges in the graph. The different steps of the algorithm (applied separately for left and right nerves) are described below.

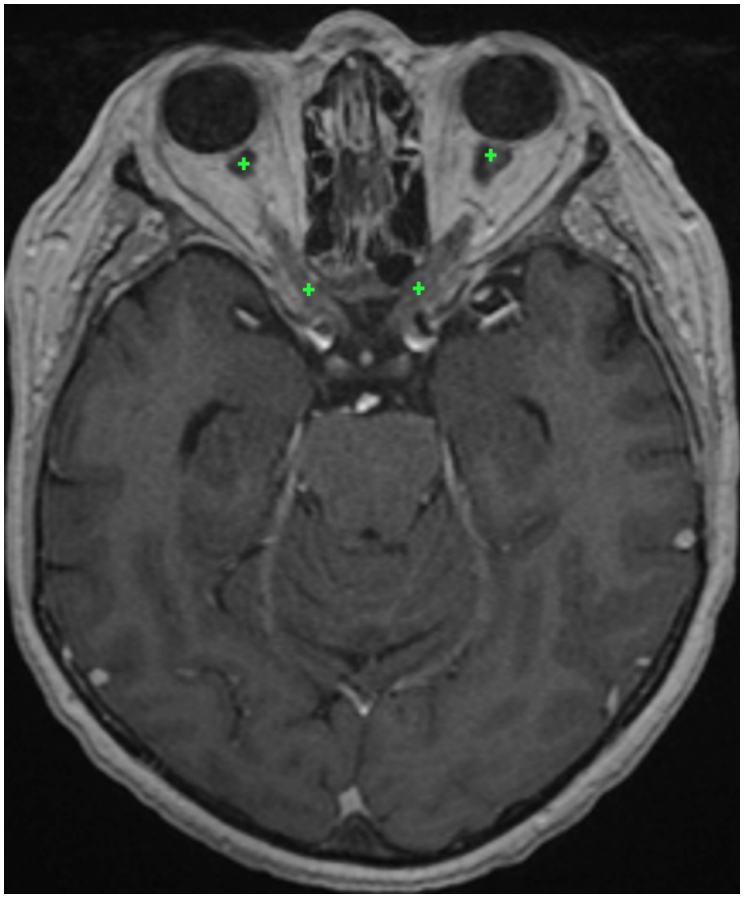

Figure 4.4: Approximate position of the optic nerve landmarks (displayed on the same axial slice) found by the system on a test example. For each of the optic nerves, the graph-based algorithm ensures the connectivity between the two landmarks.

First, we detect landmarks corresponding to the two endpoints of an optic nerve based on the initial segmentation of the visual system produced by neural networks (Fig. 4.4). The first landmark of the left optic nerve is the barycenter of P points initially predicted as the left optic nerve and which are closest to the left eye. The second landmark is computed similarly but searching P points of the left side of the optic chiasm which are the closest to the initial prediction of the left optic nerve. We take the barycenter of several points (in our experiments $P = 30$) in order to obtain a point which is more likely to be close to the centerline of the nerve. If the detected chiasm landmarks for the two optic nerves are anormally close, the procedure is applied only for one nerve, connecting the landmark with the closest eye.

Before applying the graph-based algorithm, we refine the initial segmentation of the optic nerves based on voxel intensities (specific to each image). In fact, the

optic nerves are surrounded by fat, which appears hyperintense on MR T1-weighted images and can be rather easily distinguished from the optic nerve. We compute an approximate range of intensities of voxels of the fat by computing the 98% quantile of a small volume surrounding the eye-nerve landmark. Voxels whose intensites are above 80% of this value are classified negative for the optic nerve in order to eliminate common false positives.

Given the two computed landmarks and the refined inital segmentation, we estimate the centerline of the optic nerve (Fig. 4.5) by computing the shortest path in an oriented graph. The nodes of the graph correspond to voxels within a region of interest (cuboid containing the two landmarks) and which are reachable from the starting point. The connectivty of nodes is defined by adjacence of voxels with increasing y coordinate, i.e. the childs of the node (x, y, z) are nodes $(x + d_x, y + 1, z + d_z)$ with $d_x \in \{-1, 0, 1\}$ and $d_z \in \{-1, 0, 1\}$. We therefore assume strictly increasing y of the centerline towards the second landmark (from anterior to posterior).
Each node (x, y, z) of the graph has its associated cost based on three criteria (listed by decreasing importance):

- Label $l_{(x,y,z)}$ initially assigned to the voxel (x, y, z). A strong penalty is applied to voxels predicted as negative, in order to force the centerline to pass by points initially predicted as positive. The associated cost is $c^{label}_{(x,y,z)} = 0$ if $l_{(x,y,z)} = 1$ and $c^{label}_{(x,y,z)} = C^l$ otherwise, where C^l is a fixed number controlling the importance of this cost (we set $C^l = 100$).

- If the predicted label $l_{(x,y,z)}$ is positive: distance d_{border} to the closest point classified as negative. The penalty is inversely proportional to this distance, to give priority to points which are far from predicted borders of the optic nerve (preference to central points). This cost is expressed by $c^{border}_{(x,y,z)} = 0$ if $l_{(x,y,z)} = 0$ and $c^{border}_{(x,y,z)} = R - d_{border}$ otherwise, where R is the radius of a search zone around the voxel (x, y, z). As the visible nerve is larger close to the eye, R varies with the coordinate y (interpolatation between $R = 7$ and $R = 3$, expressed in number of voxels).

- Distance d_{target} to the target point (i.e. the nerve-chiasm landmark). The penalty is proportional to this distance in order to force the centerline to immediately go towards the target point if other criteria do not give priority to some points. In particular when one part of the optic nerve has not been initially detected (negative voxels), the line should go in the direction of the target point. The associated cost is $c^{distance}_{(x,y,z)} = C^t d_{target}$ where C^t controls the importance of this cost. We fixed $C^l = 0.001$, to make it negligible compared to the previous criteria.

The cost of the node (x, y, z) is the sum of the three components: $c_{(x,y,z)} = c^{label}_{(x,y,z)} + c^{border}_{(x,y,z)} + c^{distance}_{(x,y,z)}$. The introduced cost determines the weights of edges in

the graph. A directed edge between the point (x_1, y_1, z_1) and (x_2, y_2, z_2) has the weight of $c_{(x_2,y_2,z_2)}$. The shortest path between nodes corresponding to the two endpoints of the optic nerve is computed by Dijkstra's algorithm [Cormen 2009, Zhan 1998]. The start point is the eye-nerve landmark as the optic nerve is generally well visible close to the eye. To the best of our knowledge, our approach is the first to combine deep learning with the search of the shortest path in a graph for segmentation of tubular anatomical structures. However, the idea of computing optimal distances for segmentation of tubular structures appears in interactive level-set methods presented in [Deschamps 2001, Cohen 1997, Benmansour 2011]. The objective of these methods is to find a geodesic between two points in the image chosen by the user. The Eikonal equation is constructed based on voxel intensites and contrasts, and the problem is solved by Fast Marching [Sethian 1999], similar to Dijkstra's algorithm. Application of methods based only on image intensities may be difficult for segmentation of the optic nerves in MRI due, for instance, to the noise in images and local inhomogeneity of intensities within the optic nerve.

The final segmentation of the optic nerve is constructed from the centerline. As the optic nerve has a variable thickness, around each point (x, y, z) of the centerline we consider two spherical volumes $S^1_{(x,y,z)}$ and $S^2_{(x,y,z)}$ with associated radii $R_1 \leq R_2$. All voxels within $S^1_{(x,y,z)}$ are classified positive (optic nerve). Voxels of $S^2_{(x,y,z)}$ which are not within $S^1_{(x,y,z)}$ are classified positive only if they were positive in the original segmentation. We fixed $R_1 = 2.5$ and R_2 corresponds to the radius R defined previously (large close to the eye, smaller close to the optic chiasm).
Finally, we apply mathematical morphology [Zana 2001] to reduce false positives corresponding to structures which 'attach' to the optic nerve and have a similar appearance. As these false positives are often connected to the correct segmentation by thin segments (Fig. 4.6), we apply the morphological opening with three 1D structing elements of size 2 in the three directions and we take the largest connected component.

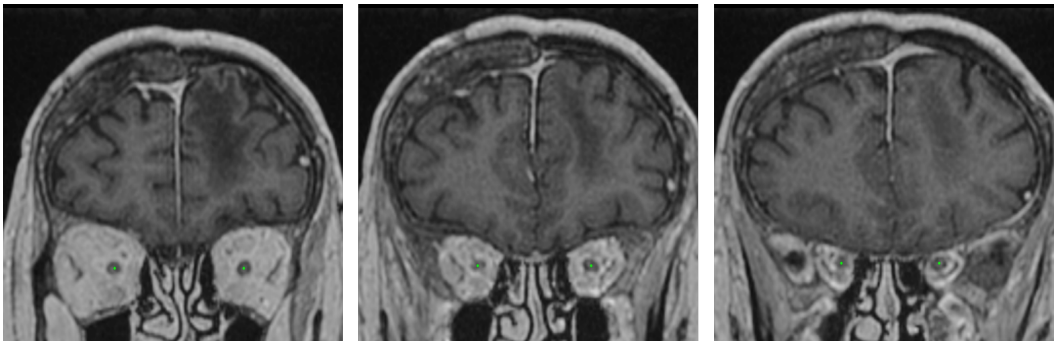

Figure 4.5: The centerlines of the optic nerves computed by our system on a test example (displayed on three different coronal slices). We assume one point of the centerline for each coronal slice between the two landmarks of the optic nerve.

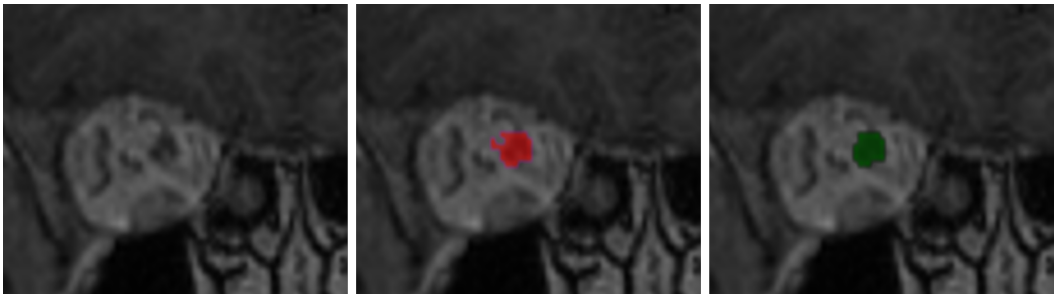

Figure 4.6: Use of mathematical morphology for reduction of false positives. Left: a coronal patch centered on an optic nerve. Middle: result obtained by the system on a test example without using mathematical morphology. Right: result obtained after application of morphological opening followed by taking the largest connected component.

4.3 Experiments

4.3.1 Data and preprocessing

We constructed a database of contrast-enhanced T1 MRIs acquired in the Centre Antoine Lacassagne (Nice, France), which is one of the three cancer centers in France equipped with proton therapy systems. Proton therapy [Levin 2005] is an external beam radiotherapy which irradiates cancer cells with beams of protons. Ionizing radiation by protons focuses over a narrow range of depth: the energy loss of the radiation follows the *Bragg curve* achieving a pronounced peak just before the particles stop. An important advantage of proton therapy is therefore to offer the possibility to deliver high doses to target volumes while sparing healthy structures. Proton therapy is particularly suitable for treatment of cancers which are very close to critical organs or which are located deep in the body. Currently, one of the main applications of the protontherapy is treatement of brain cancers.
The database contains 44 MRIs with provided segmentations of organs at risk and 50 non-annotated MRIs. The annotated images are used for training and cross-validated quantitative evaluation. For each scan, the ground truth segmentation was provided only for a subset of classes. The numbers of available segmentations for each class are reported in Table 4.1. The images without annotations are used for qualitative evaluation by a radiotherapist, as described in section 4.3.4.

The images were originally provided in Dicom format [Mildenberger 2002] and were heterogenous in terms of image intensities and geometrical properties such as size, spatial resolution and the visible part of the head. The ground truth segmentations were the ones used for routine radiotherapy planning and were provided in Dicom RT-Struct files [Law 2009], representing coordinates of polygons corresponding to contours of anatomical structures.

In order to use these images, we performed the following preprocessings. We used 3D Slicer [Fedorov 2012] and its extension SlicerRT [Pinter 2012] to generate 3D volumes (in *nrrd* format) from Dicom slices and to generate binary label masks from RT-Struct files. All images were resampled to the same spatial resolution, 0.7x0.7x0.9 (isotropic in axial slices, 0.9 spacing between slices) and then resized to dimensions 320x365x200. The 200 axial slices start from the top of the head, i.e. if an image originally has more than 200 axial slices, the bottom slices (close to the neck) are ignored. However, the input images had generally around 150 axial slices and the bottom slices were filled with zeros. To approximately normalize the image intensities, first we compute the maximum of an image, which is likely to be reached by a point on a fat or contrast-enhanced blood vessels. Then, all voxel values are divided by the value of the maximum and multiplied by a fixed constant.

4.3.2 Metrics for quantitative evaluation

To quantitatively evaluate our system, we perform 5-fold cross-validation on the set of 44 annotated MRIs. In each fold, 80% of the database is used for training and 20 % is used for test. For each class of interest, two results are reported. First, we report results obtained with our model trained on axial slices (denoted 'U-Net multiclass, axial' in the following), i.e. the raw output of the neural network, without postprocessing. Then we report results obtained after majority voting and postprocessing (denoted 'Final result' in the following).

The first metric we use is the Dice score, which measures the voxelwise overlap between the output and the ground truth segmentation. An important limitation of this metric is that it gives the same importance to very close and very distant mismatches. As the ground truth is often uncertain and noisy close to the boundaries of structures, the Dice scores are generally considerably lower for small structures. This is why, in addition to raw Dice scores, we also report results (Dice, sensitivity, specificity) obtained when a margin of one voxel is allowed, i.e. ignoring

Table 4.1: Numbers of provided ground truth segmentations for different classes (in the database of 44 MRIs).

	Number of segmentations
Hippocampus	39
Brainstem	39
Eye	41
Lens	34
Optic nerves	40
Optic chiasm	41
Pituitary gland	29
Brain	37

mismatches on the borders of the ground truth. This assumption means that a false positive on a voxel (x, y, z) which is directly neighboring with the ground truth segmentation is ignored, i.e. it is neither counted as false positive nor true positive. Similarly, a false negative (non-detection) on the border of the ground truth is ignored.
The second used metric is the undirected Hausdorff distance expressed in millimeters (the coordinates of points are expressed in real values). The Hausdorff distance measures the length of the farthest mismatch between the output and and ground truth (false positive or false negative). It is therefore useful to assess the consistency of the result, i.e. presence of very distant mistmatches. However, its limitation is that it only measures the value of the maximal distance and therefore one misclassified voxel is sufficient to considerably increase the Hausdorff distance.

Therefore, we also measure the mean distance between the output segmentation A and the ground truth B, defined as follows:

$$M(A,B) = \frac{1}{|A| + |B|} \left(\sum_{a \in A} \inf_{b \in B} d(a,b) + \sum_{b \in B} \inf_{a \in A} d(b,a) \right) \tag{4.1}$$

where d is the Euclidean distance.

4.3.3 Quantitative results

The mean distances between produced segmentations and the ground truth segmentation ranged from 0.08 mm (for the brain) to 0.69 mm (for the pituitary gland), as reported in Table 4.5. The results are variable across the different organs, according to their size, the number of ground truth segmentations available for training and the overall complexity of the segmentation task.
The Dice scores are usually higher for large anatomical structures such as the brain and the brainstem. In particular, the borders of the ground truth are usually very uncertain, which represents a problem for quantitative evaluation for smaller classes. In large classes, the border region is small compared to the entire volume of the class and therefore the mismatches on borders do not cause large drops of the metric. The highest Dice score was obtained for the brain (Dice score of 96.8). The lowest performances were obtained for the pituitary gland (mean Dice of 58, mean distance of 0.69 mm between the output and the ground truth). Segmentation of the pituitary gland is particularly challenging as it is small and difficult to be differentiated from surrounding structures. Moreover, the pituitary gland was the class with the lowest number of training examples (29 annotated cases, i.e. around 23 training cases in each of the 5 folds).
To take into account the uncertain borders of the ground truth, we also reported Dice scores, sensitivity and specificity ignoring mismatches on the border of the ground truth, as described previously. As most mismatches between the outputs

and the ground truth are on noisy borders of organs, there is a considerable difference between the raw Dice score (Table 4.2) and the Dice score with tolerance to one voxel (Table 4.3).
However, the measured Hausdorff distances (Table 4.4) are higher for large classes. The highest mean Hausdorff distance is observed for the brain, for which it is of almost equal to 1 cm.

The combination of neural networks (trained respectively on axial, coronal, sagittal slices) by majority voting improved almost all metrics. The improvements were particularly large for the Hausdorff distance (Table 4.4) and the mean distance

Table 4.2: Mean Dice scores (5-fold cross-validation) obtained on a set of 44 MRIs.

	U-Net multiclass, axial	Final result
Hippocampus	69.2	71.4
Brainstem	88.1	88.6
Eye	88.3	89.6
Lens	55.8	58.8
Optic nerves and chiasm	63.9	67.4
Pituitary gland	53.6	58.0
Brain	96.5	96.8

Table 4.3: Mean Dice score (5-fold cross-validation), sensitivity and specificity with tolerance to one voxel (ignoring mismatches on the borders due to the uncertainty of the ground truth).

	Dice score	Sensitivity	Specificity
Hippocampus	88.2	92.7	85.0
Brainstem	95.1	95.5	95.6
Eye	97.5	98.3	96.8
Lens	82.1	88.2	78.4
Optic nerves and chiasm	91.1	96.2	87.1
Pituitary gland	79.7	83.3	77.5
Brain	98.6	98.0	99.4

Table 4.4: Hausdorff distances in millimeters (5-fold cross-validation).

	U-Net multiclass, axial	Final result
Hippocampus	42.1	6.9
Brainstem	45.5	7.8
Eye	75.9	3.0
Lens	31.0	3.7
Optic nerves and chiasm	76.7	6.3
Pituitary gland	52.5	4.6
Brain	30.4	9.8

(Table 4.5). We observe that the majority voting removes almost all distant false positives and yields more robust results than a raw output of one neural network. The results were subsequently improved by additional postprocessings.
The postprocessing of the eyes consisted in setting a lower bound on the physical volume of the output segmentation. This simple procedure allowed to remove false positives and decreased the mean Hausdorff distance from 12.2 mm (result of the majority voting) to 3 mm.
The postprocessing of the optic nerve decreased the number of false positives and enforced connectivity between the eyes and the chiasm, as described in section 4.2.2.2. False positives are removed when they are either too far from the centerline, hyperintense in T1-weighted MRI (fat surrounding eyes) or are disconnected from the main connected component after application of morphological opening removing thin segments. The Dice score with one-voxel tolerance increased from 89.6 (result of the majority voting) to 91.1 (after postprocessing) for the optic nerves and chiasm. The raw Dice score increased from 66.3 to 67.4.
The postprocessing of the brain consisted in taking the largest connected component and filling the 'holes' of the segmentation in axial, coronal and sagittal planes. As these 'holes' are usually small compared to the whole volume of the class (occupying a large part of the image), the variation of the metrics is limited. The Dice score increased from 96.7 to 96.8 and the Hausdorff distance decreased from 10.2 to 9.8.

To the best of our knowledge, the only deep learning work for segmentation of organs at risk in MRI is the one proposed in [Orasanu 2018] which reported cross-validated results (mean distances in mm) on a set of 16 MRIs . The authors used a model-based segmentation [Ecabert 2008] combined with a neural network for detection of boundaries of anatomical structures. The results reported by the authors for the anatomical structures we also segment are: 0.608 mm for the brainstem, 0.563 mm for the eyes, 0.268 mm for the lenses and 0.41 mm for the optic nerves and chiasm. Overall, the ranges of mean distances are therefore comparable to the ours.

Examples of the output segmentations (comparison to the ground truth) are

Table 4.5: Mean distances in millimeters (5-fold cross-validation).

	U-Net multiclass, axial	Final result
Hippocampus	0.97	0.66
Brainstem	0.26	0.26
Eye	0.35	0.11
Lens	1.29	0.63
Optic nerves and chiasm	1.09	0.48
Pituitary gland	2.45	0.69
Brain	0.07	0.08

displayed on Fig. 4.7-4.14.

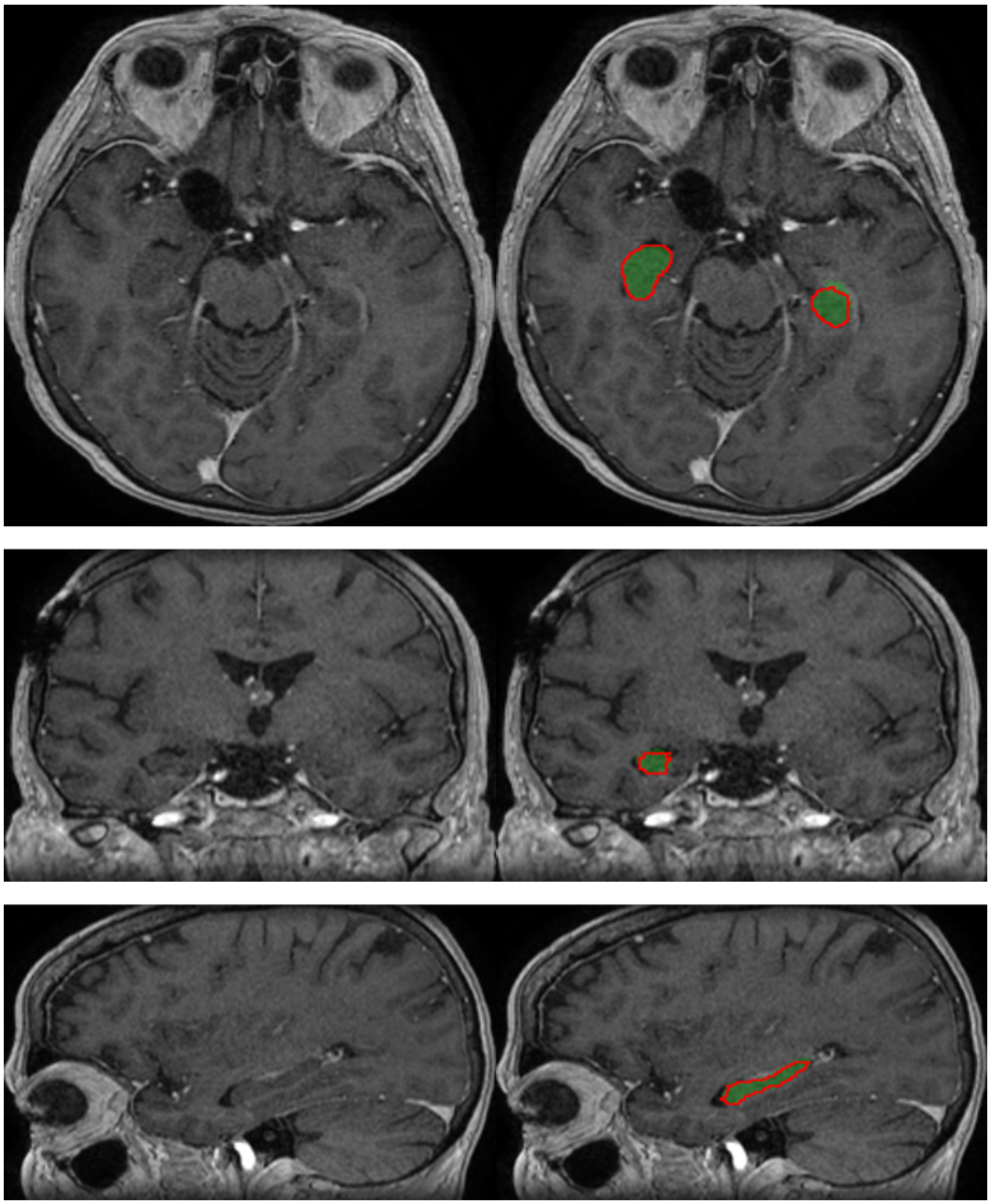

Figure 4.7: Segmentation of the hippocampus produced by our system on a test example (three orthogonal slices passing by the same point). The output segmentation is represented by the green region, the ground truth annotation is represented by the red contour.

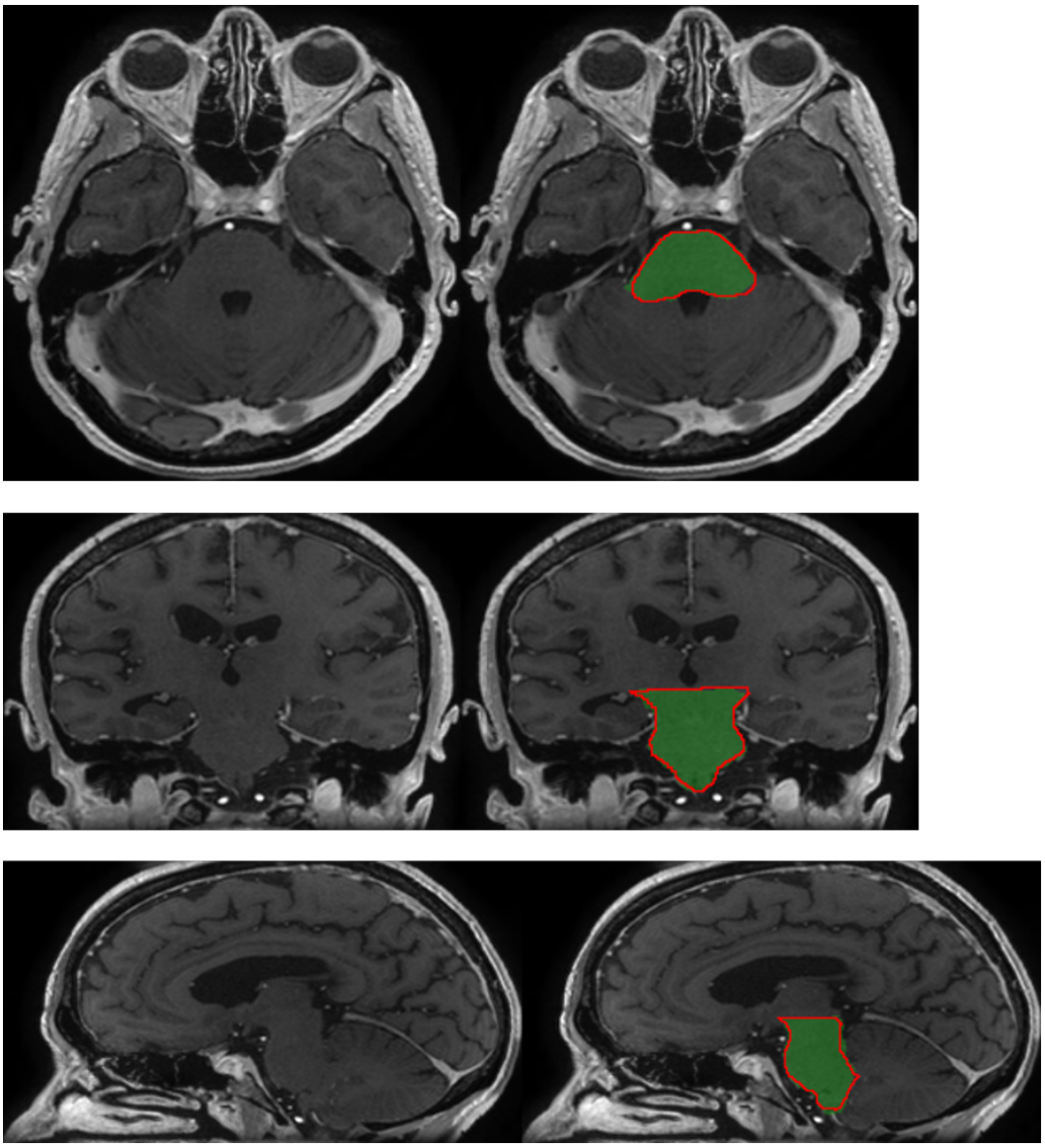

Figure 4.8: Segmentation of the brainstem produced by our system on a test example (three orthogonal slices passing by the same point). The output segmentation is represented by the green region, the ground truth annotation is represented by the red contour.

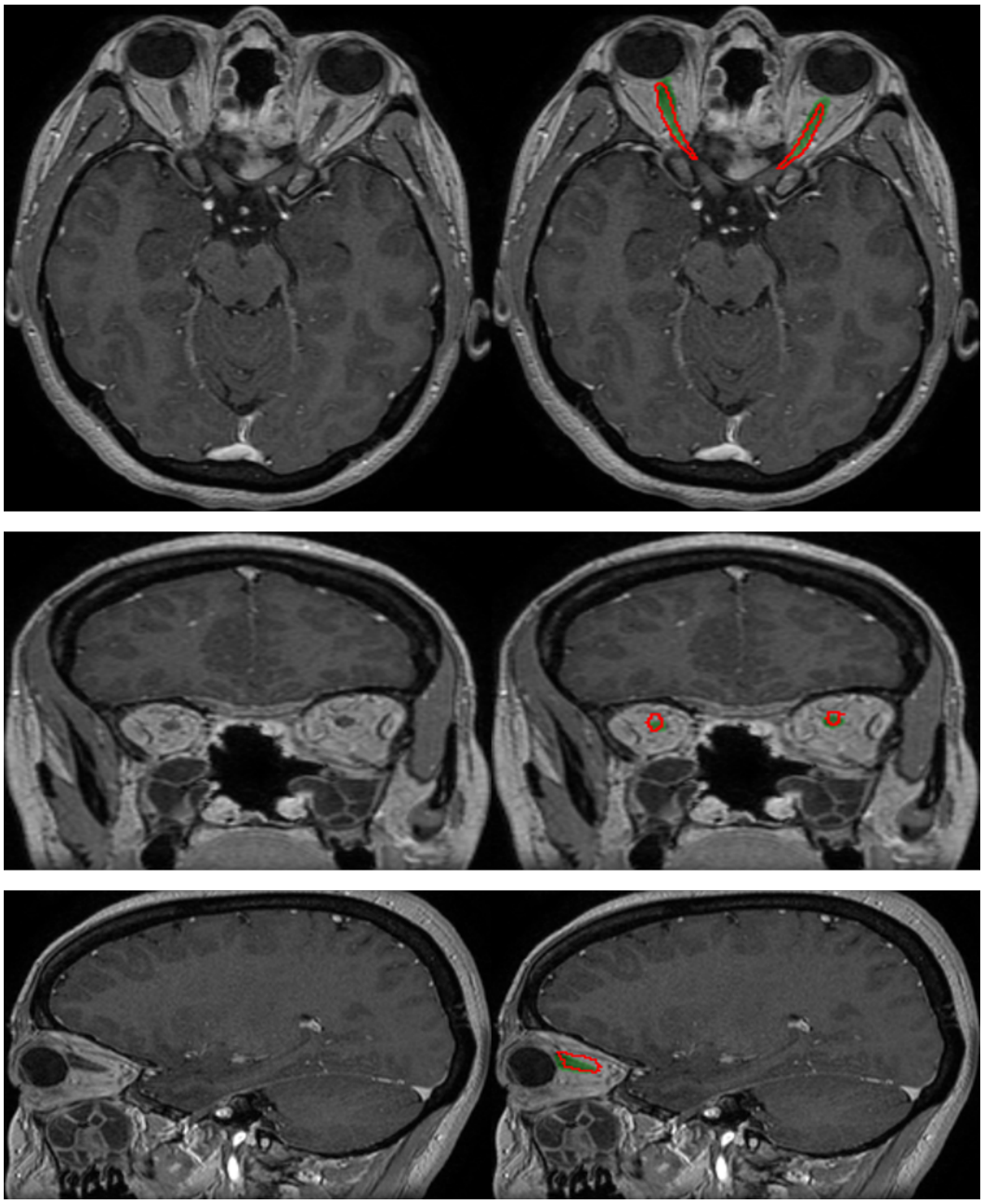

Figure 4.9: Segmentation of the optic nerves produced by our system on a test example (three orthogonal slices passing by the same point). The output segmentation is represented by the green region, the ground truth annotation is represented by the red contour.

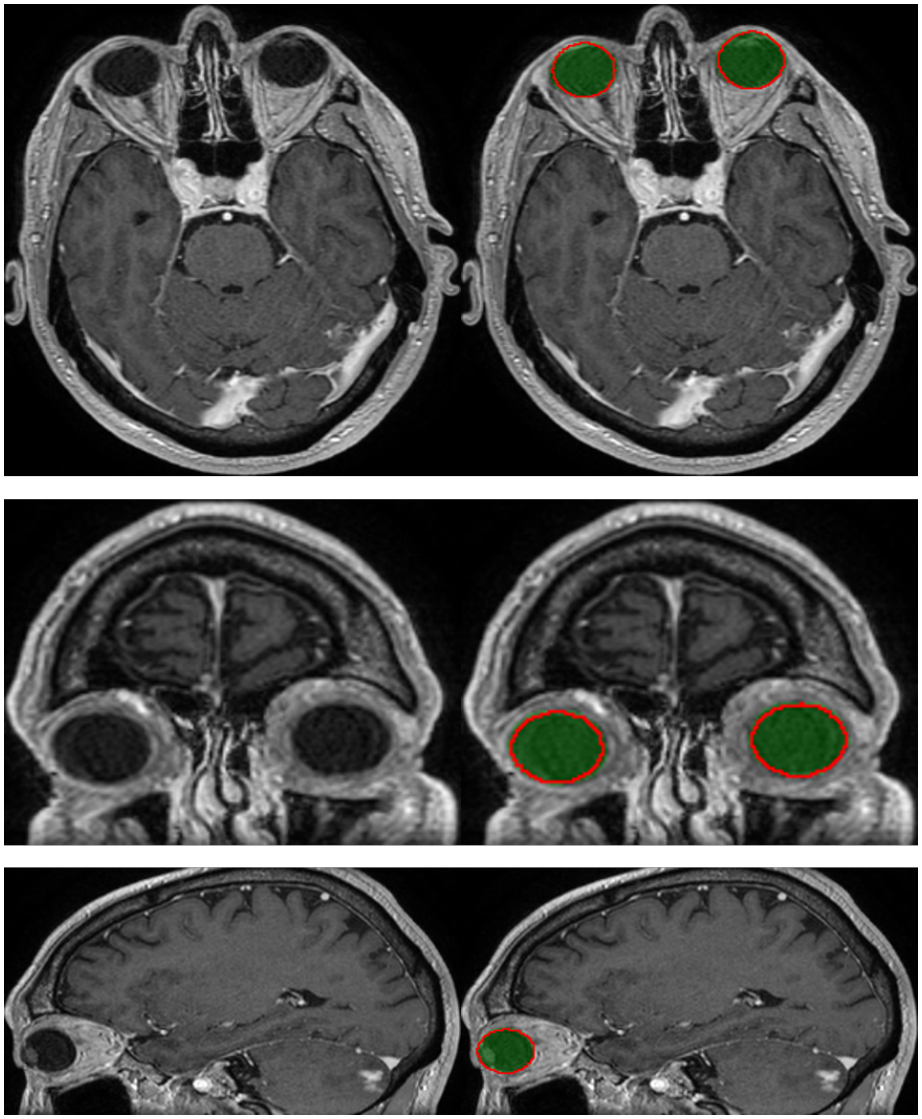

Figure 4.10: Segmentation of the eyes produced by our system on a test example (three orthogonal slices passing by the same point). The output segmentation is represented by the green region, the ground truth annotation is represented by the red contour.

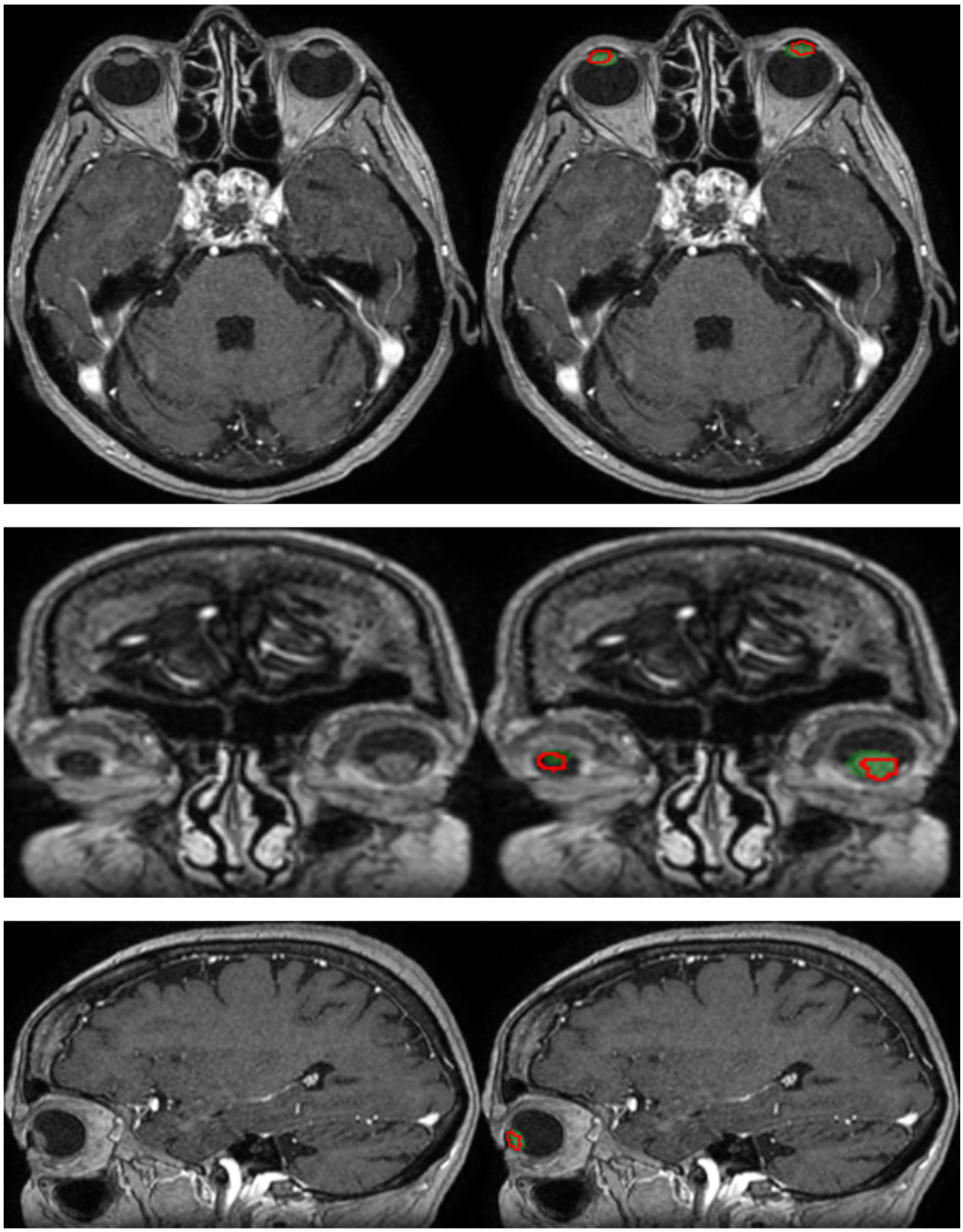

Figure 4.11: Segmentation of the lenses produced by our system on a test example (three orthogonal slices passing by the same point). The output segmentation is represented by the green region, the ground truth annotation is represented by the red contour.

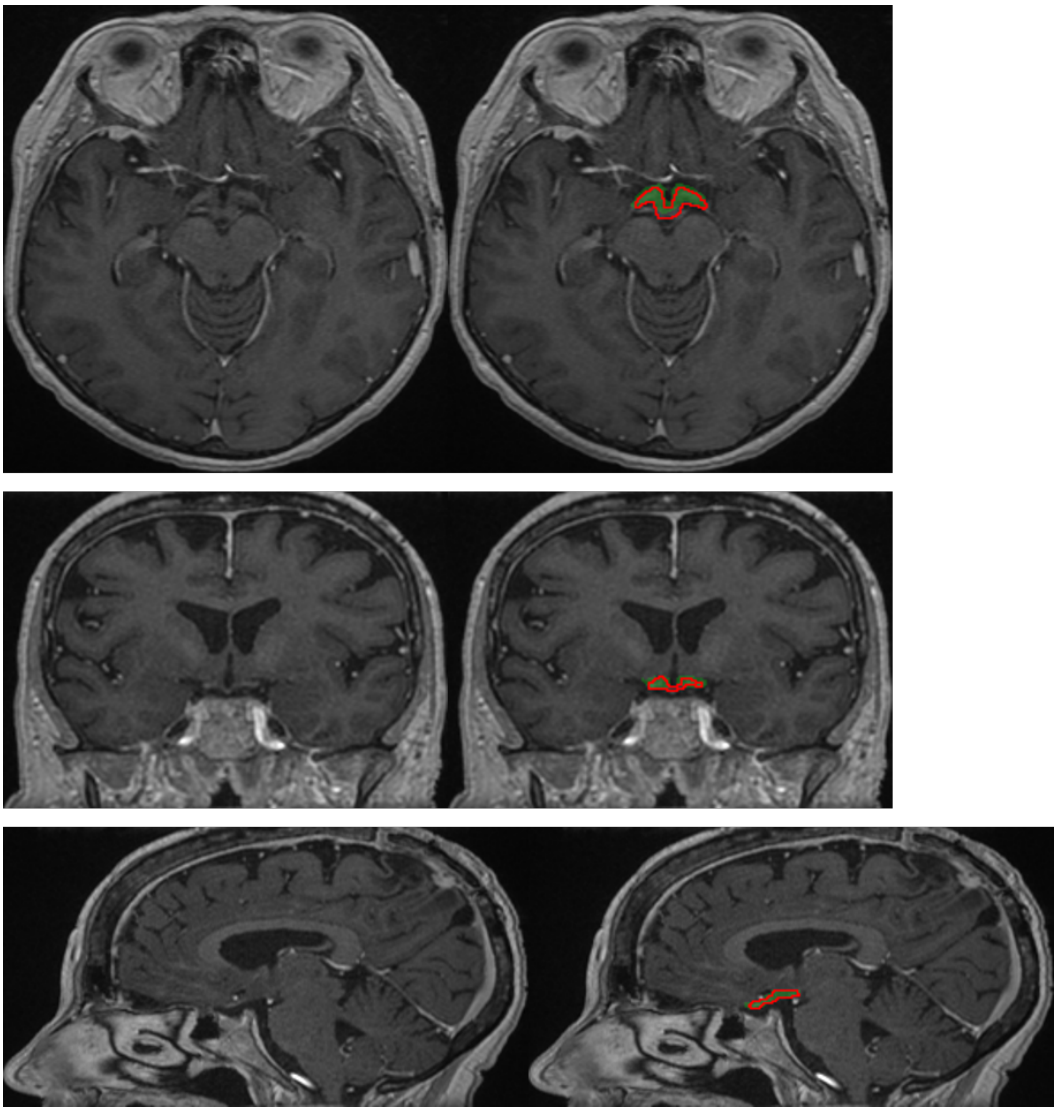

Figure 4.12: Segmentation of the optic chiasm produced by our system on a test example (three orthogonal slices passing by the same point). The output segmentation is represented by the green region, the ground truth annotation is represented by the red contour.

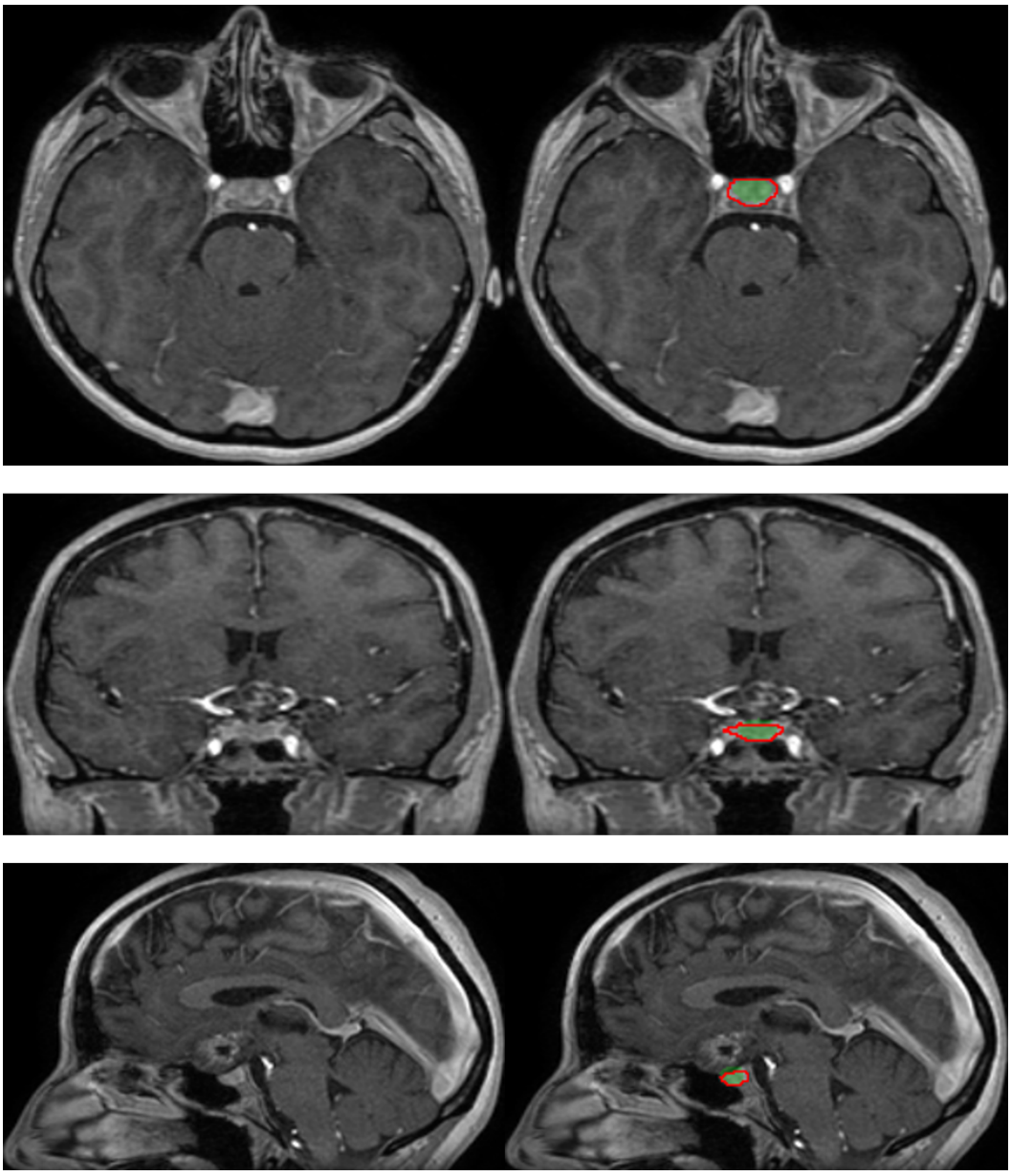

Figure 4.13: Segmentation of the pituitary gland produced by our system on a test example (three orthogonal slices passing by the same point). The output segmentation is represented by the green region, the ground truth annotation is represented by the red contour.

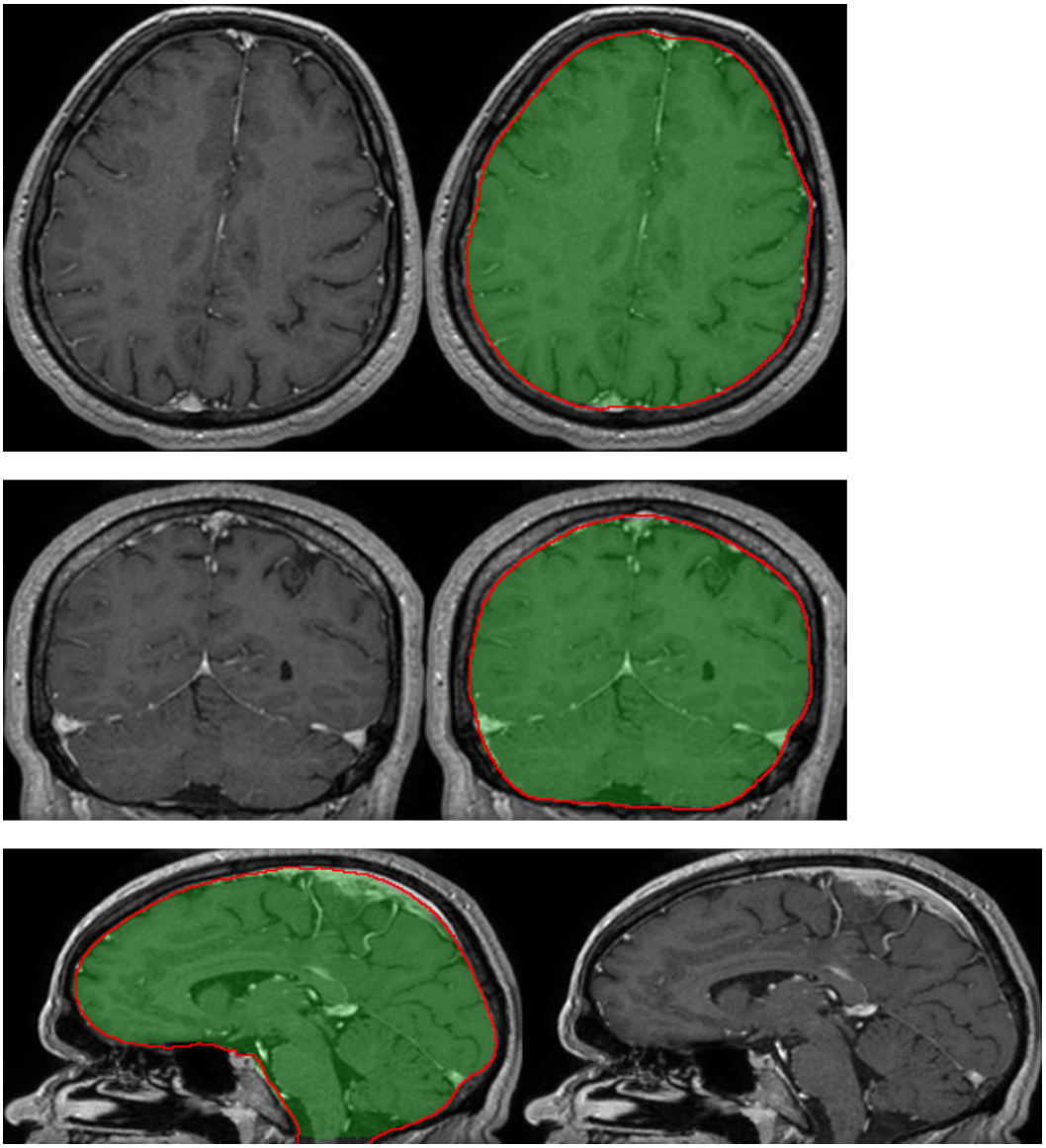

Figure 4.14: Segmentation of the brain produced by our system on a test example (three orthogonal slices passing by the same point). The output segmentation is represented by the green region, the ground truth annotation is represented by the red contour.

4.3.4 Qualitative evaluation by a radiotherapist

The segmentations produced by our system on a set of 50 non-annotated MRIs are qualitatively evaluated by an experienced radiotherapist in order to assess their accuracy and utility for radiotherapy planning. For each of the 50 patients, the radiotherapist qualitatively evaluates the segmentations produced by our system for 12 anatomical structures (counting separately left and right components of bilateral classes), i.e. 600 segmentations are evaluated in total. The segmentations are displayed with 3D Slicer [Fedorov 2012]. Each of the 600 segmentations is assigned to one of the following categories:

- Accept: the radiotherapist would keep the segmentation for radiotherapy planning without any changes
- Accept with minor modifications: the segmentation is still acceptable for radiotherapy planning, i.e. some minor errors are observed but keeping the current segmentation without changes should not impact the irradiation doses
- Accept with major modifications: the segmentation has necessarily to be corrected, i.e. keeping the current segmentation would have an important impact on the irradiation doses (even if only few voxels are misclassified). The segmentation is however still good enough to keep it, i.e. it is less time-consuming to perform the necessary modifications than segmenting the structure from the beginning
- Reject: the segmentation has failed and keeping it would not save time compared to manually segmenting the structure from the beginning
- Not assigned: the structure is absent (e.g. organ removed by surgery) or invisible in the image because of a tumor

The results are summarized in Table 4.6. 73 % of the segmentations were assigned to the category *accept*, i.e. would be kept for radiotherapy planning without any modifications. Approximately 23 % of the segmentations were assigned to the second category, i.e. acceptable for radiotherapy planning but with recommendation to perform some minor corrections, usually on extremities of organs. The system produced therefore satisfactory segmentations in a large majority of cases. It was able to correctly delineate organs despite the important difficulties such as presence of tumors and the resulting mass effects, motion artifacts in MRI, different orientations of heads of patients and anatomical modifications resulting from previous surgeries undergone by the patient (removed tissues).

Segmentations of the eyes had the highest rate of immediate acceptation: 93 out of 100 segmentations were assigned to the *accept* category. The only segmentation which required a major modification was a case with a lesion inside the eye, possibly the polypoidal choroidal vasculopathy. The lesion was not classified by

Table 4.6: Clinical evaluation by a radiotherapist on 50 test cases.

	Accept	Accept, minor corrections	Accept, major corrections	Reject	N/A
Hippocampus left	39/50	8	1	2	0
Hippocampus right	45/50	5	0	0	0
Brainstem	22/50	26	1	1	0
Eye left	48/50	2	0	0	0
Eye right	45/50	4	1	0	0
Lens left	39/50	7	4	0	0
Lens right	42/50	6	2	0	0
Optic nerve left	44/50	6	0	0	0
Optic nerve right	40/50	10	0	0	0
Optic chiasm	19/50	26	4	1	0
Pituitary gland	19/50	25	3	0	3
Brain	36/50	14	0	0	0
Total	438/600	139	16	4	3

the system as part of the eye and therefore one part of the eye was not detected. The minor modifications recommended for other cases were generally to correct few non-detected voxels on the border of the eye (top or bottom axial slices) or few false positives on the anterior part of the orbit.

All segmentations of the optic nerves were found acceptable for radiotherapy planning: 84 out of 100 segmentations were assigned to the *accept* category and the remaining 16 cases required only minor corrections. Most of the minor errors were non-detections for few voxels on the extremity of the optic nerve close to the eye (e.g. on the top axial slice). There was also at least one case of false positives on the neighboring arteries, close to the optic chiasm.

Even if in the previous, quantitative evaluation, the metrics for the lenses were significantly lower than for other structures, most of their segmentations on the set of 50 MRIs were found satisfactory by the radiotherapist. Minor corrections were required in 13 out of 100 cases and major corrections were required in 6 cases. Most of the problems were non-detections, for instance observed in cases where the patient looks to the side and the system does not detect one side of the lens. The lenses are very small structures and their visibility is highly impacted by motion artifacts in MRI.

For the optic chiasm, the corrections were more frequently required but were usually minor: 19 out of 50 cases were assigned to the *accept* category and 26 cases required minor corrections. The minor errors were often false positives on the hypothalamus (the same issue was observed in the ground truth used to train the model) and sometimes on arteries neighboring the chiasm. The major corrections (4 cases) were mainly non-detections of a small subpart of the beginning of an optic nerve. In fact, even if only a small number of voxels is not detected (false

negatives), the corrections are necessary as an excessive irradiation of one part of the optic nerve could make the entire nerve dysfonctional [Källman 1992]. One segmentation was rejected due to non-detection of one part of the chiasm. This error appeared in a challenging case where the anatomy of the patient was modified by an important mass effect caused by a tumor.
Similar performances were obtained for the pituitary gland, located below the optic chiasm. Most of the minor (26 cases) and major (3 cases) required corrections correspond to non detections, typically on the 1-2 lowermost slices. In at least 2 cases, few false negatives were observed on the pituitary stalk (also observed in some ground truth segmentations used for training), which is the connection between the pituitary gland and the hypothalamus.
Even if segmentation of the hippocampus is difficult (low contrast with neighboring structures), in our evaluation it had one of the highest acceptation rates, with 84 segmentations in the *accept* category. However, it is also the only structure for which more than one segmentation was rejected. The two rejected segmentations correspond to cases where a large tumoral mass has grown near to the hippocampus, causing an edema having a similar intensity in T1-weighted MRI. Morever, the tumors had a large necrotic core which may be confused with a ventricle by the system. In other cases, the required corrections (mostly minor) correspond usually to false positives (in particular on the amygdales, neighboring hippocampi and having a similar intensity in MRI T1) or some non-detections on the extremities of the hippocampus.
For the brainstem, 48 out of 50 segmentations were found acceptable for radiotherapy planning but required minor modifications in approximately half cases. The required corrections (false positives or non-detections) were almost exclusively on the uppermost axial slices (typically on 2 slices) which correspond to the top extremity of the brainstem. The only rejected segmentation corresponds to a case with a tumor adjacent to the brainstem and which was mistakenly included in the segmentation (false positives).
Finally, all segmentations of the brain (occupying a large part of the head) were found acceptable for radiotherapy planning even if they required minor corrections in almost one third of cases. The recommended corrections include, for instance, non-detections close to the cribriform plate (between the eyes) and false positives on bones.

In particular, we observe that the only two structures for which all segmentations were found acceptable for radiotherapy planning (without any major correction) are the ones for which a specific postprocessing was performed, i.e. the optic nerves and the brain.

4.4 Conclusion and future work

In this work we proposed a CNN-based method for segmentation of organs at risk from MR images in the context of neuro-oncology. The method was evaluated on clinical data.
First, we proposed a deep learning model and a training algorithm for segmentation of multiple and non-exclusive anatomical structures. The proposed methodology addresses problems related to computational costs and the variable availability of ground truth segmentations of the different anatomical structures (unsegmented classes). The neural network used in our method is a modified version of U-Net. The network is trained separately for segmentation in axial, coronal and sagittal slices. The three versions of the network are combined by majority voting.
Second, we proposed procedures to enforce anatomical consistency of the result in a postprocessing stage. In particular, we proposed a graph-based algorithm for segmentation of the optic nerves, which are among the most difficult anatomical structures for automatic segmentation. The proposed postprocessings have shown their efficiency particularly in the qualitative evaluation by a radiotherapist. In particular, all segmentations of the optic nerves were found acceptable for radiotherapy planning.
The method was evaluated quantitatively on a set of 44 annotated MRIs, with 5 fold cross-validation and using several metrics. The segmentations produced by our system on a set of 50 non-annotated MRIs were qualitatively evaluated by an experienced radiotherapist. Despite the limited size of the training database (44 annotated MRIs) and the different challenges of the segmentation tasks (in particular, presence of tumors), a large majority of the output segmentations were found sufficiently accurate to be used for computation of irradiation doses in radiotherapy.

An important step of the future work is to adapt the method to multimodal data. Often, several types of images are acquired during radiotherapy planning for one patient, including the different MR sequences (T1, T2, FLAIR) and CT scans. Inclusion of different imaging modalities could improve segmentation of several structures but it comes also with new challenges related, for instance, to inter-modality registration and training of models on cases with missing modalities.
As for other segmentation tasks in medical imaging, availability of annotated training data is an important problem. Methods able to exploit weaker forms of annotations (bounding boxes, slice-level labels) for training of segmentation models are therefore of interest. In particular, methods combining weakly-annotated and fully-annotated training images were recently proposed in [Mlynarski 2019c, Shah 2018].
As our system was able to produce accurate segmentations in a large majority of cases and the rare observed errors were mainly on boundaries of organs, the system could be used for generation of bounding boxes (subsequently verified by a human) which could be used to train segmentation models which are able to exploit this type of annotations.

Another important direction of the future work is to combine segmentation of organs at risk and segmentation of radiotherapy target volumes. In particular, a large variability of methods for tumor segmentation [Myronenko 2018, Kamnitsas 2017b, Wang 2017, Mlynarski 2019b, Parisot 2014] were proposed in recent years. Deep learning could also be used for computation of irradiation doses [Andres 2019] in radiotherapy planning.

Chapter 5

Conclusion and perspectives

Contents

5.1 Contributions of the thesis

In this thesis, we proposed methods for segmentation of brain tumors and organs at risk in the context of radiotherapy planning. Most of the proposed methods are based on deep learning and address important limitations of the current deep learning segmentation models.

In Chapter 2, we focused on multiclass segmentation of brain tumors using Convolutional Neural Networks trained in a supervised manner, on manually segmented images. The main methodological contributions of the chapter address problems related to computational costs of CNNs and the joint use of several imaging modalities.
First, we discussed aspects related to the notion of receptive field, which is one of the key attributes of segmentation networks. We proposed to use a cascaded model trained in two steps, where one segmentation network takes, as additional input, features learned by other segmentation networks. Such design offers the possibility to have a large 3D receptive field (resulting from the use of the extracted features) bypassing GPU memory constraints. Our system was tested on a publicly available database of the BRATS challenge and yielded promising results, with high median Dice scores of the three tumor subregions considered in the challenge: 0.918 (whole tumor), 0.883 (tumor core) and 0.854 (enhancing core).
Moreover, we proposed a new approach to deal with several input modalities, such as the different MRI sequences. In most segmentation CNNs, the first convolutional layer of the network takes as input all channels of the input image, assuming therefore the availability of all modalities. In practice, a given MRI sequence is available only for a subset of patients in the training database, especially when images were acquired in different imaging centers. We proposed a hybrid model composed of modality-specific subnetworks and the main part of the network which

extracts features resulting from the combination of the different channels (MRI sequences in our case). Even if our model assumes the presence of all modalities at the test time, it can be trained on databases containing images with missing modalities, as it contains modality-specific subnetworks which can be trained independently.

In Chapter 3 we studied the use of weaker forms of annotations for training of neural networks for tumor segmentation. Most of the current state-of-the-art segmentation models are based on CNNs trained on manually segmented images. Such annotations are very costly in the case of tumor segmentation which is time-consuming and, more importantly, requires medical expertise and has therefore to be performed by experts. Weaker forms of annotations, for instance image-level labels (tumor present or not within the image), can be obtained at a considerably lower cost but are also less informative. Machine learning models trained using only weakly-annotated images, representing a limited information, yield generally significantly lower segmentation accuracies than fully-supervised methods. We proposed therefore a new approach, with a mixed level of supervision. We assumed that the training database contains two types of images: fully-annotated (with provided tumor segmentation) and weakly-annotated (with image-level labels). The main principle of our method is to extend a segmentation network, such as U-Net, with an additional branch performing image-level classification. The segmentation and classification subnetworks share most of their layers and are jointly trained, using fully-annotated and weakly-annotated images. A large number of cross-validated tests was performed, using the data of the BRATS challenge, to study the effects of the mixed level of supervision. We showed that it significantly improves segmentation accuracy compared to the standard supervised learning when the number of fully-annotated images is limited, and yields similar segmentation accuracy when the ratio between the number of weakly-annotated and fully-annotated images decreases.

In Chapter 4, we focused on the challenging task of segmentation of organs at risk, which is a necessary step of radiotherapy planning. The objective is to perform an anatomically consistent segmentation of several structures in the brain such as the brainstem, the hippocampus, the hypophysis and the organs of the visual system. First, we proposed an efficient approach to train CNNs for segmentation of multiple and non-exclusive classes, which differs from standard segmentation problems where one voxel is assigned to exactly one class. Then, we proposed procedures to enforce anatomical consistency of the result in a postprocessing stage. In particular, we segmented the optic nerves (one of the most challenging organs) with an algorithm based on the search of the shortest path in a graph, using outputs of neural networks to define the weights of the edges in the graph.
Our system was extensively evaluated on real clinical data. We constructed a database of MRIs acquired in the Centre Antoine Lacassagne (Nice, France), per-

forming all the necessary processing (cleaning, conversion of formats, resampling, resizing, intensity normalization). In addition to a quantitative cross-validated evaluation, the segmentations produced by our system on a set of non-annotated MRIs were qualitatively evaluated by an radiotherapist from the Centre Antoine Lacassagne. A large majority of the 600 output segmentations (50 patients, 12 structures) evaluated by the radioterapist were found accurate enough to be used for radiotherapy planning.

5.2 Perspectives

Efficiency of machine learning segmentation models depends on several factors including, among others, the quantity and the quality of the training data, the model architecture and the algorithm used to train the model. Even if we addressed several important points, efficiency of segmentations systems for neuro-oncology could still be improved with the future research work and using significantly larger databases than the ones used during this thesis. In this final section, we propose a few directions for the future research work.

First, we believe that the use of weakly-annotated data could have a significant impact in medical imaging. An interesting direction of the future work would be to extend our model with mixed supervison (Chapter 3) to 3D CNNs and for segmentation of organs at risk. In this case, for each anatomical structure, weak annotations would correspond to ground truth bounding boxes of the structure. The model would be a variant of 3D U-Net with an additional branch for patch-level classification, i.e. predicting if the 3D patch contains the organ of interest. To train such model, a strategy should be designed to sample training examples, i.e. to form training batches. During the training phase, three types of 3D patches could be extracted: the ones with provided segmentation of the anatomical structure, patches outside the bounding box of the structure (assumed to be provided for all training images) and patches corresponding to a bounding box of the structure (positive examples) without provided segmentation. With the currently available GPUs, segmentation of large 3D patches would be difficult but such model could be used for segmentation of smaller structures such as the pituitary gland or the optic chiasm.
Moreover, weak annotations could be generated by fully-supervised segmentation models, after verification by a human observer. In particular, our system for segmentation of organs at risk trained on a small database (30-40 ground truth segmentations per class) was able to produce very accurate segmentations in almost all cases. As most observed errors were only at the borders of organs, bounding boxes of the different anatomical structures could be generated with a very high confidence. Such annotations could be subsequently used for training of models using a mixed level of supervision.

Another important aspect is the quality of the annotations in the training database. Segmentations of organs and lesions often contain errors and are subjective, due to several factors including human errors and the different technical issues related, for instance, to the inter-modality registration (when several types of images are used). This 'noise' in the ground truth not only impacts the accuracy of segmentation systems but also represents a difficulty for quantitative evaluation of segmentation methods. For these reasons, methods able to estimate the uncertainty of the segmentation, such as the one proposed in [Lê 2016b], are of particular interest.

In this thesis, we addressed the segmentation of tumors following the format of the BRATS challenge, in which three types of tumor tissues are considered: the tumor-induced edema, the necrotic core and the contrast-enhancing core. An interesting direction would be to directly estimate the target volumes for radiotherapy planning: the gross tumor volume (GTV) and the clinical target volume (CTV). Determination of the CTV is particularly challenging as individual cancer cells could not be imaged using MRI and the delimitation of this region is based on medical expertise. Automatic determination of the CTV could potentially benefit from the use of tumor growth models [Angelini 2007]. In particular, a tumor growth model was recently used in [Lê 2016a] for optimization of irradiation doses in radiotherapy planning.

We separately addressed segmentation of tumor lesions and organs at risk, due particularly to the constraints of the available databases. In particular, the data of the BRATS challenge contains already preprocessed images (in particular, skull-stripped) and the challenge focuses only on segmentation of gliomas. An important direction of the future work would be to combine segmentation of tumors and organs at risk.

In general, current segmentation systems usually address very specific tasks and are often applied in a very limited context. We believe that future systems may benefit from unification of different tasks and from the use of different types of data.

5.3 List of publications

Journal articles:

- **P. Mlynarski**, H. Delingette, A. Criminisi, N. Ayache, *3D Convolutional Neural Networks for Tumor Segmentation using Long Range 2D Context*, Computerized Medical Imaging and Graphics, vol. 73, pages 60-72, 2019

- **P. Mlynarski**, H. Delingette, A. Criminisi, N. Ayache, *Deep Learning with Mixed Supervision for Brain Tumor Segmentation*, SPIE Journal of Medical Imaging, vol. 6, no. 3, 2019
- **P. Mlynarski**, H. Delingette, H. Alghamdi, P. Bondiau, N. Ayache, *Anatomically Consistent Segmentation of Organs at Risk in MRI with Convolutional Neural Networks*, submitted to SPIE Journal of Medical Imaging

Bibliography

[Abadi 2016] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin*et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems.* arXiv preprint arXiv:1603.04467, 2016. (Cited on pages 15 and 58.)

[Alchatzidis 2015] Stavros Alchatzidis, Aristeidis Sotiras and Nikos Paragios. *Local atlas selection for discrete multi-atlas segmentation.* In 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), pages 363–367. IEEE, 2015. (Cited on page 61.)

[Andres 2019] E Alvarez Andres, L Fidon, M Vakalopoulou, G Noël, S Niyoteka, N Benzazon, E Deutsch, N Paragios and C Robert. *PO-1002 Pseudo Computed Tomography generation using 3D deep learning–Application to brain radiotherapy.* Radiotherapy and Oncology, vol. 133, page S553, 2019. (Cited on page 89.)

[Angelini 2007] Elsa D Angelini, Olivier Clatz, Emmanuel Mandonnet, Ender Konukoglu, Laurent Capelle and Hugues Duffau. *Glioma dynamics and computational models: a review of segmentation, registration, and in silico growth algorithms and their clinical applications.* Current Medical Imaging Reviews, vol. 3, no. 4, pages 262–276, 2007. (Cited on page 94.)

[Argiris 2008] Athanassios Argiris, Michalis V Karamouzis, David Raben and Robert L Ferris. *Head and neck cancer.* The Lancet, vol. 371, no. 9625, pages 1695–1709, 2008. (Cited on page 63.)

[Atlas 2009] Scott W Atlas. Magnetic resonance imaging of the brain and spine, volume 1. Lippincott Williams & Wilkins, 2009. (Cited on page 2.)

[Bakas 2017] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani and Christos Davatzikos. *Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features.* Scientific data, vol. 4, page 170117, 2017. (Cited on pages 2, 10 and 36.)

[Bauer 2011] Stefan Bauer, Lutz-P Nolte and Mauricio Reyes. *Fully automatic segmentation of brain tumor images using support vector machine classification in combination with hierarchical conditional random field regularization.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 354–361. Springer, 2011. (Cited on page 11.)

[Bauer 2012] Stefan Bauer, Thomas Fejes, Johannes Slotboom, Roland Wiest, Lutz-P Nolte and Mauricio Reyes. *Segmentation of brain tumor images based on integrated hierarchical classification and regularization.* In MICCAI BraTS Workshop. Nice: Miccai Society, 2012. (Cited on page 11.)

[Bauer 2013] Stefan Bauer, Roland Wiest, Lutz-P Nolte and Mauricio Reyes. *A survey of MRI-based medical image analysis for brain tumor studies.* Physics in medicine and biology, vol. 58, no. 13, page R97, 2013. (Cited on pages 10, 36 and 60.)

[Bearman 2016] Amy Bearman, Olga Russakovsky, Vittorio Ferrari and Li Fei-Fei. *What's the point: Semantic segmentation with point supervision.* In European Conference on Computer Vision, pages 549–565. Springer, 2016. (Cited on page 38.)

[Benmansour 2011] Fethallah Benmansour and Laurent D Cohen. *Tubular structure segmentation based on minimal path method and anisotropic enhancement.* International Journal of Computer Vision, vol. 92, no. 2, pages 192–210, 2011. (Cited on page 71.)

[Bergamo 2014] Alessandro Bergamo, Loris Bazzani, Dragomir Anguelov and Lorenzo Torresani. *Self-taught object localization with deep networks.* arXiv preprint arXiv:1409.3964, 2014. (Cited on page 38.)

[Bergstra 2010] James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley and Yoshua Bengio. *Theano: A CPU and GPU math compiler in Python.* In Proc. 9th Python in Science Conf, pages 1–7, 2010. (Cited on pages 15 and 58.)

[Board 2018] PDQ Adult Treatment Editorial Board. *Adult Central Nervous System Tumors Treatment (PDQ®).* In PDQ Cancer Information Summaries [Internet]. National Cancer Institute (US), 2018. (Cited on page 2.)

[Bondiau 2005] Pierre-Yves Bondiau, Grégoire Malandain, Stéphane Chanalet, Pierre-Yves Marcy, Jean-Louis Habrand, Francois Fauchon, Philippe Paquis, Adel Courdi, Olivier Commowick, Isabelle Ruttenet *al. Atlas-based automatic segmentation of MR images: validation study on the brainstem in radiotherapy context.* International Journal of Radiation Oncology* Biology* Physics, vol. 61, no. 1, pages 289–298, 2005. (Cited on page 61.)

[Boyd 2004] Stephen Boyd and Lieven Vandenberghe. Convex optimization. Cambridge university press, 2004. (Cited on page 38.)

[Brosch 2018] Tom Brosch, Jochen Peters, Alexandra Groth, Thomas Stehle and Jürgen Weese. *Deep learning-based boundary detection for model-based segmentation with application to MR prostate segmentation.* In International

Conference on Medical Image Computing and Computer-Assisted Intervention, pages 515–522. Springer, 2018. (Cited on page 62.)

[Brouwer 2012] Charlotte L Brouwer, Roel JHM Steenbakkers, Edwin van den Heuvel, Joop C Duppen, Arash Navran, Henk P Bijl, Olga Chouvalova, Fred R Burlage, Harm Meertens, Johannes A Langendijk*et al. 3D variation in delineation of head and neck organs at risk.* Radiation Oncology, vol. 7, no. 1, page 32, 2012. (Cited on page 60.)

[Chen 2014] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy and Alan L Yuille. *Semantic image segmentation with deep convolutional nets and fully connected crfs.* arXiv preprint arXiv:1412.7062, 2014. (Cited on page 11.)

[Cheplygina 2018] Veronika Cheplygina, Marleen de Bruijne and Josien PW Pluim. *Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis.* arXiv preprint arXiv:1804.06353, 2018. (Cited on page 39.)

[Ciardo 2017] Delia Ciardo, Marianna Alessandra Gerardi, Sabrina Vigorito, Anna Morra, Veronica Dell'Acqua, Federico Javier Diaz, Federica Cattani, Paolo Zaffino, Rosalinda Ricotti, Maria Francesca Spadea*et al. Atlas-based segmentation in breast cancer radiotherapy: evaluation of specific and generic-purpose atlases.* The Breast, vol. 32, pages 44–52, 2017. (Cited on page 61.)

[Çiçek 2016] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox and Olaf Ronneberger. *3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation.* arXiv preprint arXiv:1606.06650, 2016. (Cited on pages 11, 15 and 62.)

[Ciompi 2017] Francesco Ciompi, Kaman Chung, Sarah J Van Riel, Arnaud Arindra Adiyoso Setio, Paul K Gerke, Colin Jacobs, Ernst Th Scholten, Cornelia Schaefer-Prokop, Mathilde MW Wille, Alfonso Marchianò*et al. Towards automatic pulmonary nodule management in lung cancer screening with deep learning.* Scientific Reports, vol. 7, 2017. (Cited on page 13.)

[Cohen 1997] Laurent D Cohen and Ron Kimmel. *Global minimum for active contour models: A minimal path approach.* International journal of computer vision, vol. 24, no. 1, pages 57–78, 1997. (Cited on page 71.)

[Commowick 2008] Olivier Commowick, Vincent Grégoire and Grégoire Malandain. *Atlas-based delineation of lymph node levels in head and neck computed tomography images.* Radiotherapy and Oncology, vol. 87, no. 2, pages 281–289, 2008. (Cited on page 61.)

[Commowick 2009] Olivier Commowick, Simon K Warfield and Grégoire Malandain. *Using Frankenstein's creature paradigm to build a patient specific atlas.* In In-

ternational Conference on Medical Image Computing and Computer-Assisted Intervention, pages 993–1000. Springer, 2009. (Cited on page 61.)

[Cordier 2016] Nicolas Cordier, Hervé Delingette and Nicholas Ayache. *A patch-based approach for the segmentation of pathologies: Application to glioma labelling.* IEEE transactions on medical imaging, vol. 35, no. 4, pages 1066–1076, 2016. (Cited on page 11.)

[Cormen 2009] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest and Clifford Stein. Introduction to algorithms. MIT press, 2009. (Cited on page 71.)

[Criminisi 2010] Antonio Criminisi, Jamie Shotton, Duncan Robertson and Ender Konukoglu. *Regression forests for efficient anatomy detection and localization in CT studies.* In International MICCAI Workshop on Medical Computer Vision, pages 106–117. Springer, 2010. (Cited on page 61.)

[Criminisi 2013] Antonio Criminisi, Duncan Robertson, Ender Konukoglu, Jamie Shotton, Sayan Pathak, Steve White and Khan Siddiqui. *Regression forests for efficient anatomy detection and localization in computed tomography scans.* Medical image analysis, vol. 17, no. 8, pages 1293–1303, 2013. (Cited on page 61.)

[Deschamps 2001] Thomas Deschamps and Laurent D Cohen. *Fast extraction of minimal paths in 3D images and applications to virtual endoscopy.* Medical image analysis, vol. 5, no. 4, pages 281–299, 2001. (Cited on page 71.)

[Dou 2017] Qi Dou, Lequan Yu, Hao Chen, Yueming Jin, Xin Yang, Jing Qin and Pheng-Ann Heng. *3D deeply supervised network for automated segmentation of volumetric medical images.* Medical Image Analysis, 2017. (Cited on pages 11 and 17.)

[Dreyfus 1990] Stuart E Dreyfus. *Artificial neural networks, back propagation, and the Kelley-Bryson gradient procedure.* Journal of Guidance, Control, and Dynamics, vol. 13, no. 5, pages 926–928, 1990. (Cited on pages 13 and 38.)

[Dubost 2017] Florian Dubost, Gerda Bortsova, Hieab Adams, Arfan Ikram, Wiro J Niessen, Meike Vernooij and Marleen De Bruijne. *GP-Unet: lesion detection from weak labels with a 3D regression network.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 214–221. Springer, 2017. (Cited on page 38.)

[Ecabert 2008] Olivier Ecabert, Jochen Peters, Hauke Schramm, Cristian Lorenz, Jens von Berg, Matthew J Walker, Mani Vembar, Mark E Olszewski, Krishna Subramanyan, Guy Lavi*et al.* *Automatic model-based segmentation of the heart in CT images.* IEEE transactions on medical imaging, vol. 27, no. 9, pages 1189–1201, 2008. (Cited on page 76.)

[Evgeniou 2004] Theodoros Evgeniou and Massimiliano Pontil. *Regularized multi–task learning*. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 109–117. ACM, 2004. (Cited on page 39.)

[Fedorov 2012] Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean-Christophe Fillion-Robin, Sonia Pujol, Christian Bauer, Dominique Jennings, Fiona Fennessy, Milan Sonka*et al.* *3D Slicer as an image computing platform for the Quantitative Imaging Network*. Magnetic resonance imaging, vol. 30, no. 9, pages 1323–1341, 2012. (Cited on pages 73 and 85.)

[Folk 2011] Mike Folk, Gerd Heber, Quincey Koziol, Elena Pourmal and Dana Robinson. *An overview of the HDF5 technology suite and its applications*. In Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases, pages 36–47. ACM, 2011. (Cited on page 66.)

[Gambhir 2002] Sanjiv Sam Gambhir. *Molecular imaging of cancer with positron emission tomography*. Nature Reviews Cancer, vol. 2, no. 9, page 683, 2002. (Cited on page 2.)

[Gauriau 2015] Romane Gauriau, Rémi Cuingnet, David Lesage and Isabelle Bloch. *Multi-organ localization with cascaded global-to-local regression and shape prior*. Medical image analysis, vol. 23, no. 1, pages 70–83, 2015. (Cited on page 61.)

[Geremia 2012] Ezequiel Geremia, Bjoern H Menze, Nicholas Ayache*et al.* *Spatial decision forests for glioma segmentation in multi-channel MR images*. MICCAI Challenge on Multimodal Brain Tumor Segmentation, vol. 34, 2012. (Cited on page 11.)

[Gillies 2015] Robert J Gillies, Paul E Kinahan and Hedvig Hricak. *Radiomics: images are more than pictures, they are data*. Radiology, vol. 278, no. 2, pages 563–577, 2015. (Cited on page 10.)

[Girshick 2014] Ross Girshick, Jeff Donahue, Trevor Darrell and Jitendra Malik. *Rich feature hierarchies for accurate object detection and semantic segmentation*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 580–587, 2014. (Cited on page 38.)

[Goodenberger 2012] McKinsey L Goodenberger and Robert B Jenkins. *Genetics of adult glioma*. Cancer genetics, vol. 205, no. 12, pages 613–621, 2012. (Cited on pages 9 and 35.)

[Goodfellow 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio. *Generative adversarial nets*. In Advances in neural information processing systems, pages 2672–2680, 2014. (Cited on page 39.)

[Gooya 2012] Ali Gooya, Kilian M Pohl, Michel Bilello, Luigi Cirillo, George Biros, Elias R Melhem and Christos Davatzikos. *GLISTR: glioma image segmentation and registration.* IEEE transactions on medical imaging, vol. 31, no. 10, pages 1941–1954, 2012. (Cited on page 11.)

[Hanahan 2000] Douglas Hanahan and Robert A Weinberg. *The hallmarks of cancer.* cell, vol. 100, no. 1, pages 57–70, 2000. (Cited on page 1.)

[Hanahan 2011] Douglas Hanahan and Robert A Weinberg. *Hallmarks of cancer: the next generation.* cell, vol. 144, no. 5, pages 646–674, 2011. (Cited on page 1.)

[Havaei 2017] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin and Hugo Larochelle. *Brain tumor segmentation with deep neural networks.* Medical image analysis, vol. 35, pages 18–31, 2017. (Cited on pages 11 and 16.)

[He 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. *Deep residual learning for image recognition.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. (Cited on pages 11, 15 and 41.)

[Ho 1995] Tin Kam Ho. *Random decision forests.* In Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on, volume 1, pages 278–282. IEEE, 1995. (Cited on page 11.)

[Hong 2015] Seunghoon Hong, Hyeonwoo Noh and Bohyung Han. *Decoupled deep neural network for semi-supervised semantic segmentation.* In Advances in neural information processing systems, pages 1495–1503, 2015. (Cited on page 39.)

[Hong 2016] Seunghoon Hong, Junhyuk Oh, Honglak Lee and Bohyung Han. *Learning transferrable knowledge for semantic segmentation with deep convolutional neural network.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3204–3212, 2016. (Cited on page 39.)

[Hsieh 2009] Jiang Hsieh*et al. Computed tomography: principles, design, artifacts, and recent advances.* SPIE Bellingham, WA, 2009. (Cited on page 2.)

[Hung 2018] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin and Ming-Hsuan Yang. *Adversarial Learning for Semi-Supervised Semantic Segmentation.* arXiv preprint arXiv:1802.07934, 2018. (Cited on page 39.)

[Ibragimov 2017] Bulat Ibragimov and Lei Xing. *Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks.* Medical physics, vol. 44, no. 2, pages 547–557, 2017. (Cited on pages 61 and 62.)

[Ioffe 2015] Sergey Ioffe and Christian Szegedy. *Batch normalization: Accelerating deep network training by reducing internal covariate shift.* arXiv preprint arXiv:1502.03167, 2015. (Cited on pages 23, 40 and 64.)

[Isensee 2017] Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus and Klaus H Maier-Hein. *Brain Tumor Segmentation and Radiomics Survival Prediction: Contribution to the BRATS 2017 Challenge.* 2017 International MICCAI BraTS Challenge, 2017. (Cited on pages 11 and 31.)

[Källman 1992] P Källman, A Ågren and A Brahme. *Tumour and normal tissue responses to fractionated non-uniform dose delivery.* International journal of radiation biology, vol. 62, no. 2, pages 249–262, 1992. (Cited on page 87.)

[Kamnitsas 2016] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert and Ben Glocker. *Efficient Multi-Scale 3D CNN with Fully Connected CRF for Accurate Brain Lesion Segmentation.* arXiv preprint arXiv:1603.05959, 2016. (Cited on pages 11, 14, 36 and 61.)

[Kamnitsas 2017a] Konstantinos Kamnitsas, Wenjia Bai, Enzo Ferrante, Steven McDonagh, Matthew Sinclair, Nick Pawlowski, Martin Rajchl, Matthew Lee, Bernhard Kainz, Daniel Rueckert*et al.* *Ensembles of Multiple Models and Architectures for Robust Brain Tumour Segmentation.* arXiv preprint arXiv:1711.01468, 2017. (Cited on pages 5, 11 and 36.)

[Kamnitsas 2017b] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert and Ben Glocker. *Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation.* Medical image analysis, vol. 36, pages 61–78, 2017. (Cited on page 89.)

[Kamnitsas 2018] Konstantinos Kamnitsas, Daniel C Castro, Loic Le Folgoc, Ian Walker, Ryutaro Tanno, Daniel Rueckert, Ben Glocker, Antonio Criminisi and Aditya Nori. *Semi-Supervised Learning via Compact Latent Space Clustering.* arXiv preprint arXiv:1806.02679, 2018. (Cited on page 39.)

[Khan 2014] Faiz M Khan and John P Gibbons. Khan's the physics of radiation therapy. Lippincott Williams & Wilkins, 2014. (Cited on page 3.)

[Kodym 2018] Oldřich Kodym, Michal Španěl and Adam Herout. *Segmentation of Head and Neck Organs at Risk Using CNN with Batch Dice Loss.* In German Conference on Pattern Recognition, pages 105–114. Springer, 2018. (Cited on page 62.)

[Krizhevsky 2012] Alex Krizhevsky, Ilya Sutskever and Geoffrey E Hinton. *Imagenet classification with deep convolutional neural networks.* In Advances in neural information processing systems, pages 1097–1105, 2012. (Cited on pages 11 and 38.)

[Kwon 2014] Dongjin Kwon, Russell T Shinohara, Hamed Akbari and Christos Davatzikos. *Combining generative models for multifocal glioma segmentation and registration.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 763–770. Springer, 2014. (Cited on page 11.)

[Lafferty 2001] John Lafferty, Andrew McCallum and Fernando CN Pereira. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data.* 2001. (Cited on page 11.)

[Larsson 2018] Måns Larsson, Yuhang Zhang and Fredrik Kahl. *Robust abdominal organ segmentation using regional convolutional neural networks.* Applied Soft Computing, vol. 70, pages 465–471, 2018. (Cited on pages 61 and 62.)

[Law 2009] Maria YY Law and Brent Liu. *DICOM-RT and its utilization in radiation therapy.* Radiographics, vol. 29, no. 3, pages 655–667, 2009. (Cited on page 72.)

[Le Folgoc 2016] Loic Le Folgoc, Aditya V Nori, Siddharth Ancha and Antonio Criminisi. *Lifted Auto-Context Forests for Brain Tumour Segmentation.* In International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, pages 171–183. Springer, 2016. (Cited on page 11.)

[Lê 2016a] Matthieu Lê, Hervé Delingette, Jayashree Kalpathy-Cramer, Elizabeth R Gerstner, Tracy Batchelor, Jan Unkelbach and Nicholas Ayache. *Personalized radiotherapy planning based on a computational tumor growth model.* IEEE transactions on medical imaging, vol. 36, no. 3, pages 815–825, 2016. (Cited on page 94.)

[Lê 2016b] Matthieu Lê, Jan Unkelbach, Nicholas Ayache and Hervé Delingette. *Sampling image segmentations for uncertainty quantification.* Medical image analysis, vol. 34, pages 42–51, 2016. (Cited on page 94.)

[LeCun 1995] Yann LeCun, Yoshua Bengio*et al.* *Convolutional networks for images, speech, and time series.* The handbook of brain theory and neural networks, vol. 3361, no. 10, page 1995, 1995. (Cited on pages 5, 11, 36 and 61.)

[Lee 2005] Chi-Hoon Lee, Mark Schmidt, Albert Murtha, Aalo Bistritz, Jöerg Sander and Russell Greiner. *Segmenting brain tumors with conditional random fields and support vector machines.* In International Workshop on Computer Vision for Biomedical Image Applications, pages 469–478. Springer, 2005. (Cited on page 11.)

[Levin 2005] WP Levin, H Kooy, JS Loeffler and TF DeLaney. *Proton beam therapy.* British journal of Cancer, vol. 93, no. 8, page 849, 2005. (Cited on page 72.)

[Long 2015] Jonathan Long, Evan Shelhamer and Trevor Darrell. *Fully convolutional networks for semantic segmentation.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3431–3440, 2015. (Cited on pages 11, 15, 36 and 61.)

[Mehta 2011] Minesh Mehta, Michael A Vogelbaum, Susan Chang and Neha Patel. *Neoplasms of the central nervous system.* Cancer: principles and practice of oncology, vol. 9, pages 1700–49, 2011. (Cited on page 2.)

[Men 2019] Kuo Men, Huaizhi Geng, Chingyun Cheng, Haoyu Zhong, Mi Huang, Yong Fan, John P Plastaras, Alexander Lin and Ying Xiao. *More accurate and efficient segmentation of organs-at-risk in radiotherapy with convolutional neural networks cascades.* Medical physics, vol. 46, no. 1, pages 286–292, 2019. (Cited on page 62.)

[Menze 2015] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest*et al.* *The multimodal brain tumor image segmentation benchmark (BRATS).* IEEE transactions on medical imaging, vol. 34, no. 10, pages 1993–2024, 2015. (Cited on pages 2, 10, 22 and 36.)

[Mildenberger 2002] Peter Mildenberger, Marco Eichelberg and Eric Martin. *Introduction to the DICOM standard.* European radiology, vol. 12, no. 4, pages 920–927, 2002. (Cited on page 72.)

[Mlynarski 2019a] Pawel Mlynarski, Hervé Delingette, Hamza Alghamdi, Pierre-Yves Bondiau and Nicholas Ayache. *Anatomically Consistent Segmentation of Organs at Risk in MRI with Convolutional Neural Networks (submitted to SPIE Journal of Medical Imaging).* arXiv preprint arXiv:1907.02003, 2019. (Cited on pages 7 and 59.)

[Mlynarski 2019b] Pawel Mlynarski, Hervé Delingette, Antonio Criminisi and Nicholas Ayache. *3D convolutional neural networks for tumor segmentation using long-range 2D context.* Computerized Medical Imaging and Graphics, vol. 73, pages 60–72, 2019. (Cited on pages 6, 9, 42, 65, 66 and 89.)

[Mlynarski 2019c] Pawel Mlynarski, Hervé Delingette, Antonio Criminisi and Nicholas Ayache. *Deep learning with mixed supervision for brain tumor segmentation.* Journal of Medical Imaging, vol. 6, no. 3, page 034002, 2019. (Cited on pages 6, 35 and 88.)

[Myronenko 2018] Andriy Myronenko. *3D MRI brain tumor segmentation using autoencoder regularization.* In International MICCAI Brainlesion Workshop, pages 311–320. Springer, 2018. (Cited on pages 5 and 89.)

[Nikolov 2018] Stanislav Nikolov, Sam Blackwell, Ruheena Mendes, Jeffrey De Fauw, Clemens Meyer, Cían Hughes, Harry Askham, Bernardino Romera-Paredes, Alan Karthikesalingam, Carlton Chu*et al.* *Deep learning to achieve*

clinically applicable segmentation of head and neck anatomy for radiotherapy. arXiv preprint arXiv:1809.04430, 2018. (Cited on page 61.)

[Nyúl 2000] László G Nyúl, Jayaram K Udupa and Xuan Zhang. *New variants of a method of MRI scale standardization.* IEEE transactions on medical imaging, vol. 19, no. 2, pages 143–150, 2000. (Cited on page 23.)

[Oquab 2015] Maxime Oquab, Léon Bottou, Ivan Laptev and Josef Sivic. *Is object localization for free?-weakly-supervised learning with convolutional neural networks.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 685–694, 2015. (Cited on page 38.)

[Orasanu 2018] Eliza Orasanu, Tom Brosch, Carri Glide-Hurst and Steffen Renisch. *Organ-At-Risk Segmentation in Brain MRI Using Model-Based Segmentation: Benefits of Deep Learning-Based Boundary Detectors.* In International Workshop on Shape in Medical Imaging, pages 291–299. Springer, 2018. (Cited on pages 62, 63 and 76.)

[Papandreou 2015] George Papandreou, Liang-Chieh Chen, Kevin P Murphy and Alan L Yuille. *Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation.* In Proceedings of the IEEE international conference on computer vision, pages 1742–1750, 2015. (Cited on page 39.)

[Parisot 2014] Sarah Parisot, William Wells III, Stéphane Chemouny, Hugues Duffau and Nikos Paragios. *Concurrent tumor segmentation and registration with uncertainty-based sparse non-uniform graphs.* Medical image analysis, vol. 18, no. 4, pages 647–659, 2014. (Cited on page 89.)

[Pathak 2014] Deepak Pathak, Evan Shelhamer, Jonathan Long and Trevor Darrell. *Fully convolutional multi-class multiple instance learning.* arXiv preprint arXiv:1412.7144, 2014. (Cited on page 38.)

[Pereira 2015] Sérgio Pereira, Adriano Pinto, Victor Alves and Carlos A Silva. *Deep convolutional neural networks for the segmentation of gliomas in multi-sequence MRI.* In International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, pages 131–143. Springer, 2015. (Cited on pages 11 and 36.)

[Pinheiro 2015] Pedro O Pinheiro and Ronan Collobert. *From image-level to pixel-level labeling with convolutional networks.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1713–1721, 2015. (Cited on page 38.)

[Pinter 2012] Csaba Pinter, Andras Lasso, An Wang, David Jaffray and Gabor Fichtinger. *SlicerRT: radiation therapy research toolkit for 3D Slicer.* Medical physics, vol. 39, no. 10, pages 6332–6338, 2012. (Cited on page 73.)

[Prastawa 2004] Marcel Prastawa, Elizabeth Bullitt, Sean Ho and Guido Gerig. *A brain tumor segmentation framework based on outlier detection.* Medical image analysis, vol. 8, no. 3, pages 275–283, 2004. (Cited on page 10.)

[Rajchl 2016] Martin Rajchl, Matthew CH Lee, Ozan Oktay, Konstantinos Kamnitsas, Jonathan Passerat-Palmbach, Wenjia Bai, Bernhard Kainz and Daniel Rueckert. *DeepCut: Object Segmentation from Bounding Box Annotations using Convolutional Neural Networks.* arXiv preprint arXiv:1605.07866, 2016. (Cited on page 39.)

[Ramus 2010a] Liliane Ramus and Grégoire Malandain. *Assessing selection methods in the context of multi-atlas based segmentation.* In 2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pages 1321–1324. IEEE, 2010. (Cited on page 61.)

[Ramus 2010b] Liliane Ramus, Grégoire Malandain*et al. Multi-atlas based segmentation: Application to the head and neck region for radiotherapy planning.* In MICCAI Workshop Medical Image Analysis for the Clinic-A Grand Challenge, pages 281–288, 2010. (Cited on page 61.)

[Ronneberger 2015] Olaf Ronneberger, Philipp Fischer and Thomas Brox. *U-net: Convolutional networks for biomedical image segmentation.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 234–241. Springer, 2015. (Cited on pages 5, 11, 14, 17, 36, 40, 62 and 63.)

[Roth 2017] Holger R Roth, Hirohisa Oda, Yuichiro Hayashi, Masahiro Oda, Natsuki Shimizu, Michitaka Fujiwara, Kazunari Misawa and Kensaku Mori. *Hierarchical 3D fully convolutional networks for multi-organ segmentation.* arXiv preprint arXiv:1704.06382, 2017. (Cited on pages 61 and 64.)

[Rumelhart 1988] David E Rumelhart, Geoffrey E Hinton and Ronald J Williams. *Learning representations by back-propagating errors.* Cognitive modeling, vol. 5, no. 3, page 1, 1988. (Cited on pages 19 and 42.)

[Saleh 2016] Fatemehsadat Saleh, Mohammad Sadegh Aliakbarian, Mathieu Salzmann, Lars Petersson, Stephen Gould and Jose M Alvarez. *Built-in foreground/background prior for weakly-supervised semantic segmentation.* In European Conference on Computer Vision, pages 413–432. Springer, 2016. (Cited on page 38.)

[Schwartzbaum 2006] Judith A Schwartzbaum, James L Fisher, Kenneth D Aldape and Margaret Wrensch. *Epidemiology and molecular pathology of glioma.* Nature Reviews Neurology, vol. 2, no. 9, page 494, 2006. (Cited on page 2.)

[Scoccianti 2015] Silvia Scoccianti, Beatrice Detti, Davide Gadda, Daniela Greto, Ilaria Furfaro, Fiammetta Meacci, Gabriele Simontacchi, Lucia Di Brina,

Pierluigi Bonomo, Irene Giacomelli*et al. Organs at risk in the brain and their dose-constraints in adults and in children: a radiation oncologist's guide for delineation in everyday practice*. Radiotherapy and Oncology, vol. 114, no. 2, pages 230–238, 2015. (Cited on page 3.)

[Sermanet 2013] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus and Yann LeCun. *Overfeat: Integrated recognition, localization and detection using convolutional networks*. arXiv preprint arXiv:1312.6229, 2013. (Cited on page 11.)

[Serra 2012] Jean Serra and Pierre Soille. Mathematical morphology and its applications to image processing, volume 2. Springer Science & Business Media, 2012. (Cited on page 11.)

[Sethian 1999] James A Sethian. *Fast marching methods*. SIAM review, vol. 41, no. 2, pages 199–235, 1999. (Cited on page 71.)

[Shah 2018] Meet P Shah, SN Merchant and Suyash P Awate. *MS-Net: Mixed-Supervision Fully-Convolutional Networks for Full-Resolution Segmentation*. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 379–387. Springer, 2018. (Cited on pages 40 and 88.)

[Simonyan 2013] Karen Simonyan, Andrea Vedaldi and Andrew Zisserman. *Deep inside convolutional networks: Visualising image classification models and saliency maps*. arXiv preprint arXiv:1312.6034, 2013. (Cited on page 38.)

[Simonyan 2014] Karen Simonyan and Andrew Zisserman. *Very deep convolutional networks for large-scale image recognition*. arXiv preprint arXiv:1409.1556, 2014. (Cited on pages 11 and 38.)

[Sled 1998] John G Sled, Alex P Zijdenbos and Alan C Evans. *A nonparametric method for automatic correction of intensity nonuniformity in MRI data*. IEEE transactions on medical imaging, vol. 17, no. 1, pages 87–97, 1998. (Cited on page 23.)

[Tong 2018] Nuo Tong, Shuiping Gou, Shuyuan Yang, Dan Ruan and Ke Sheng. *Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks*. Medical physics, vol. 45, no. 10, pages 4558–4567, 2018. (Cited on page 62.)

[Tustison 2015] Nicholas J Tustison, KL Shrinidhi, Max Wintermark, Christopher R Durst, Benjamin M Kandel, James C Gee, Murray C Grossman and Brian B Avants. *Optimal symmetric multimodal templates and concatenated random forests for supervised brain tumor segmentation (simplified) with ANTsR*. Neuroinformatics, vol. 13, no. 2, pages 209–225, 2015. (Cited on page 11.)

[Vos 2016] Theo Vos, Christine Allen, Megha Arora, Ryan M Barber, Zulfiqar A Bhutta, Alexandria Brown, Austin Carter, Daniel C Casey, Fiona J Charlson, Alan Z Chen*et al.* *Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015.* The Lancet, vol. 388, no. 10053, pages 1545–1602, 2016. (Cited on page 60.)

[Wang 2017] Guotai Wang, Wenqi Li, Sebastien Ourselin and Tom Vercauteren. *Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks.* arXiv preprint arXiv:1709.00382, 2017. (Cited on pages 11, 30, 36 and 89.)

[Wang 2018a] Yueyue Wang, Liang Zhao, Zhijian Song and Manning Wang. *Organ at Risk Segmentation in Head and Neck CT Images by Using a Two-Stage Segmentation Framework Based on 3D U-Net.* arXiv preprint arXiv:1809.00960, 2018. (Cited on pages 61 and 64.)

[Wang 2018b] Zhiwei Wang, Chaoyue Liu, Danpeng Cheng, Liang Wang, Xin Yang and Kwang-Ting Cheng. *Automated Detection of Clinically Significant Prostate Cancer in mp-MRI Images Based on an End-to-End Deep Neural Network.* IEEE transactions on medical imaging, vol. 37, no. 5, pages 1127–1139, 2018. (Cited on page 38.)

[Warfield 2004] Simon K Warfield, Kelly H Zou and William M Wells. *Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation.* IEEE transactions on medical imaging, vol. 23, no. 7, page 903, 2004. (Cited on page 61.)

[Wiemels 2010] Joseph Wiemels, Margaret Wrensch and Elizabeth B Claus. *Epidemiology and etiology of meningioma.* Journal of neuro-oncology, vol. 99, no. 3, pages 307–314, 2010. (Cited on page 2.)

[Zana 2001] Frederic Zana and J-C Klein. *Segmentation of vessel-like patterns using mathematical morphology and curvature evaluation.* IEEE transactions on image processing, vol. 10, no. 7, pages 1010–1019, 2001. (Cited on page 71.)

[Zhan 1998] F Benjamin Zhan and Charles E Noon. *Shortest path algorithms: an evaluation using real road networks.* Transportation science, vol. 32, no. 1, pages 65–73, 1998. (Cited on page 71.)

[Zhang 2001] Yongyue Zhang, Michael Brady and Stephen Smith. *Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm.* IEEE transactions on medical imaging, vol. 20, no. 1, pages 45–57, 2001. (Cited on page 39.)

[Zheng 2018] Qiao Zheng, Hervé Delingette, Nicolas Duchateau and Nicholas Ayache. *3D Consistent & Robust Segmentation of Cardiac Images by Deep*

Learning with Spatial Propagation. IEEE Transactions on Medical Imaging, 2018. (Cited on page 11.)

[Zhou 2013] Zhuxian Zhou and Zheng-Rong Lu. *Gadolinium-based contrast agents for magnetic resonance cancer imaging.* Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology, vol. 5, no. 1, pages 1–18, 2013. (Cited on page 3.)

[Zhu 2019] Wentao Zhu, Yufang Huang, Liang Zeng, Xuming Chen, Yong Liu, Zhen Qian, Nan Du, Wei Fan and Xiaohui Xie. *AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy.* Medical physics, vol. 46, no. 2, pages 576–589, 2019. (Cited on pages 61 and 62.)

[Zikic 2012] Darko Zikic, Ben Glocker, Ender Konukoglu, Antonio Criminisi, C Demiralp, Jamie Shotton, OM Thomas, T Das, R Jena and SJ Price. *Decision forests for tissue-specific segmentation of high-grade gliomas in multi-channel MR.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 369–376. Springer, 2012. (Cited on page 11.)