Towards the generation of glioblastoma atlases with deep learning methods : Tumor segmentation and metamorphic image registration

Matthis Maillard

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2023IPPAT020

Thèse de doctorat

Towards the generation of glioblastoma atlases with deep learning methods:
Tumor segmentation and metamorphic image registration

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

École doctorale n°626 Ecole Doctorale de l'Institut Polytechnique de Paris (ED IP Paris)
Spécialité de doctorat : Signal, Images, Automatique et robotique

Thèse soutenue et présentée à Palaiseau, le 1er Juin 2023, par

MATTHIS MAILLARD

Composition du Jury :

Diana Mateus Professeure, Centrale Nantes (LS2N)	Présidente/Examinatrice
François Rousseau Professeur, IMT Atlantique (LaTIM)	Rapporteur
Michaël Sdika Ingénieur de recherche, CNRS (CREATIS)	Rapporteur
Maria Vakalopoulou Maîtresse de conférences, CentraleSupélec (MICS)	Examinatrice
Christos Davatzikos Professeur, Université de Pennsylvanie (CBICA)	Examinateur
Isabelle Bloch Professeure, Télécom Paris (LTCI) / Sorbonne Université (LIP6)	Directrice de thèse
Pietro Gori Maître de conférences, Télécom Paris (LTCI)	Co-directeur de thèse

Résumé

Cette thèse s'inscrit dans le cadre de la construction d'un atlas de glioblastomes (tumeurs cérébrales). En imagerie médicale, un atlas est une image ou un ensemble d'images représentant la distribution statistique d'une population. Souvent, cette distribution prend la forme d'une image représentant la moyenne de la population et d'un ensemble de cartes de déformations entre cette moyenne et chaque image. Dans le cas de cerveaux avec tumeurs cette représentation ne peut pas être directement utilisée telle quelle. En effet, la notion d'image moyenne pour des individus ayant des tumeurs situées dans différentes parties du cerveau n'est pas claire. Pour résoudre ce problème, une solution est de fixer une image saine en tant que représentation moyenne et de calculer uniquement les cartes de transformations entre cette image saine et les images de tous les patients avec tumeurs. Pour construire un atlas, il est donc important de correctement définir les transformations entre les images. Les méthodes classiques de recalage considèrent que les deux images sont en correspondance bijective. Or, cela n'est pas le cas dans notre contexte où les deux images n'ont pas le même nombre de composants (une des deux images a la tumeur en plus). Un défi de la thèse a donc été de produire des transformations entre deux images avec des topologies différentes. Dans notre cas, nous supposons qu'il est important de connaître, au préalable, la position où cette différence topologique intervient. Avec des images de tumeurs cérébrales, cela consiste à déterminer la segmentation de la pathologie. Un dernier défi de la thèse a été de développer ces méthodes de telle façon à ce qu'elles aient un temps d'exécution très faible sur des volumes 3D. En effet, pour construire un atlas avec un nombre important de patients, il est primordial que l'exécution de chaque méthode prenne le moins de temps possible. Pour faire cela, nous basons nos deux méthodes (segmentation et recalage) sur des techniques d'apprentissage profond qui ont un temps d'entraînement très élevé mais un temps d'inférence très faible.

La première partie de la thèse a porté sur la segmentation de tumeurs cérébrales sur des IRM, permettant ainsi de déterminer précisément l'endroit avec la différence topologique. Alors que la plupart des algorithmes utilisent plusieurs modalités d'acquisition, dans la pratique clinique souvent une seule est disponible (les images pondérées en T1 par exemple). Notre problématique a donc été de proposer un algorithme qui soit performant sur une seule modalité tout en utilisant les informations des bases de données multi-modales pendant

l'apprentissage. Pour cela, nous avons utilisé une technique de distillation de connaissances (Hinton et al., 2015). Nous utilisons un réseau maître prenant quatre modalités en entrée pour entraîner un réseau étudiant qui lui ne prend qu'une seule modalité. Les sorties du réseau maître sont lissées avant d'être comparées à celle du réseau étudiant afin de mieux montrer les relations entre les classes. Une analyse de différentes stratégies de distillation nous a permis de montrer dans quels cas ces méthodes sont utiles. Notamment, il semble qu'avec un nombre important de données, les méthodes de distillation de la connaissance ne permettent pas une amélioration de la performance. Cependant, avec un faible nombre d'images, notre stratégie montre une forte amélioration de la performance du réseau étudiant par rapport à une baseline.

La seconde partie de la thèse porte sur le recalage d'une image d'un patient ayant une tumeur vers une image de sujet sain. Nous avons développé une méthode qui prend en compte à la fois les différences géométriques et les différences topologiques entre deux images. Nous nous sommes inspirés des Métamorphoses (Trouvé and Younès, 2005) qui ont été développées pour transformer la géométrie et les niveaux d'intensité d'une image. Nous avons utilisé un réseau de neurones résiduel pour résoudre les équations aux dérivées partielles qui constituent les métamorphoses. Cela nous permet d'utiliser la méthode en apprentissage, réduisant considérablement le temps d'inférence une fois que le réseau a été entraîné. En outre, nous encourageons une séparation entre les transformations de forme et d'apparence en exploitant un masque de segmentation de la tumeur. De cette façon, nous autorisons les changements d'apparence uniquement dans les régions où des différences topologiques apparaissent entre les images source et cible (par exemple, la tumeur). Cet algorithme a été appliqué sur des IRM 3D de cerveaux mais, en réalité, il est applicable pour n'importe quel type de données, 2D ou 3D, IRM ou non. Nous démontrons que cette méthode permet un meilleur recalage des tissus sains que les méthodes de recalage classique. La méthode de recalage développée constitue ainsi un outil important dans le but de construire un atlas de glioblastomes.

Abstract

The aim of this thesis was to build an atlas of glioblastoma (brain tumors). In medical imaging, an atlas is an image or a set of images that are meant to represent the statistical distribution of a population. Often, this distribution takes the form of an image representing the population average and a set of deformation maps between this mean and each image. To construct an atlas, it is therefore important to correctly define the transformations between the images. Conventional registration methods assume that the two images have only a geometric difference - that is, the first image is the bijective deformation of the other. However, this is not the case in our context, where the two images do not have the same number of components (one of the two images has the tumor in addition). A challenge of this thesis was therefore to produce transformations between two images with different topologies.
The first part of the thesis focused on the segmentation of brain tumors on MRI. Indeed, it is important to segment the tumors in order to precisely detect the location with the topological differences. Since our goal is to build an atlas from clinical images, we need a segmentation algorithm that performs well on patients with only one acquisition modality available (such as T1-weighted images). However, most of the state-of-the-art (SOTA) tumor segmentation algorithms need four modalities to perform well. The first goal of this thesis was thus to produce a segmentation algorithm that performs well on test images from a single modality, while leveraging information from multi-modal databases during training. To this end, we proposed a new method based on knowledge distillation (Hinton et al., 2015). We use a teacher network that takes four modalities as input and helps training a student network that takes as input only one of the teacher modalities. We compare the proposed method with several knowledge distillation strategies and show that this kind of methods performs well in a low-data regime and becomes less useful in a high-data regime.
The second part of the thesis deals with the registration of a cancerous image onto a healthy image. We developed a method that, in addition to taking into account the geometric differences, it also considers the topological differences between two images. Inspired by Metamorphosis (Trouvé and Younès, 2005), a method developed to transform the geometry and intensity levels of an image, we used a residual neural network to solve the partial differential equations that encode the Metamorphosis framework. This allowed us to re-

formulate the method in a learning context, which greatly reduced the inference time once the network has been trained. Additionally, we encouraged an anatomically meaningful disentanglement between shape and appearance transformations by leveraging the (previously estimated) segmentation mask of the tumor. In this way, we allow appearance changes only in the regions where topological differences occur between source and target images (e.g., tumor). The developed registration method is thus an important tool in the construction of the glioblastoma atlas.

List of Figures

List of Tables

Contents

1 Introduction

Contents

1.1 Context

Glioblastoma (GBM) is a type of aggressive brain cancer that is still considered incurable (Tykocki and Eltayeb, 2018). The symptoms include headaches, focal neurologic deficits, confusion, memory loss, personality changes, or seizures (Alifieris and Trafalis, 2015). Furthermore, the median overall survival ranges from 12 to 20 months depending on the study (Lacroix et al., 2001; Stummer et al., 2006; Pallud et al., 2015). In the United States, the 5 and 10-year survival rate is estimated to be respectively 5% and 2.6% (Ostrom et al., 2014). A common treatment against GBM is the resection of the tumor followed by radiotherapy and adjuvant therapy (Alifieris and Trafalis, 2015). Despite this aggressive treatment, recurrence almost always occurs in proximity to the original lesion (75 to 90 percent of patients according to Tykocki and Eltayeb (2018)). The low survival rate and negative prognosis have fostered a lot of research for a better understanding of the behavior of this kind of tumor. Clinical evidence suggests that tumor size, location, and shape could be important factors related to recurrence and seizures.

The standard protocol to detect brain tumors is Magnetic Resonance Imaging (MRI) (Thust et al., 2018) as it constitutes a non-invasive method to produce detailed images of the brain internal structures. MRI uses powerful magnets and radio waves, therefore it does not involve the use of ionizing radiation or X-rays, making it a safe, non-invasive, and painless way to obtain images of the body. Furthermore, the MRI technique can generate several modalities where each imaging modality highlights specific tissues, or provides different

types of contrast between tissues. In the case of brain tumors, the commonly acquired modalities are T1, T1 contrast-enhanced (T1ce), T2, and Flair (Menze et al., 2015). As seen in Figure 1.1, the contrast in each modality is different and each one exhibits different tumor regions. For instance, the T1ce is necessary to show the necrotic region and the enhancing tissue, while the Flair and T2 better reveal the edema.

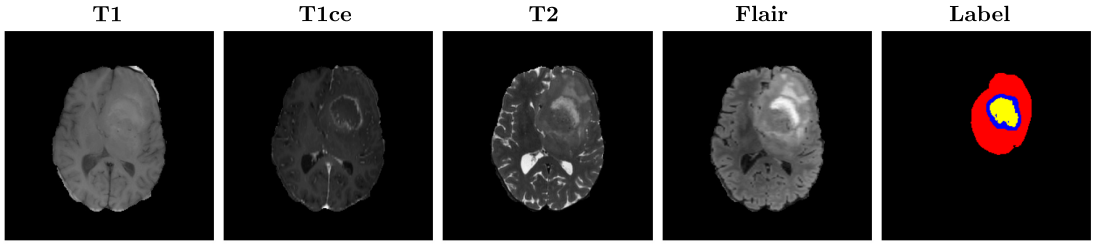


Figure 1.1: The four MRI modalities and the associated brain tumor tissue segmentation for a patient with a brain tumor. The yellow zone is the necrotic tumor, blue is the tumor core, and red is the edema.

Due to the goal of this thesis, which we detail in section 1.2, the scientific context of this thesis falls within the topics of medical image segmentation and registration with deep learning methods which have been widely studied in the previous years. For the segmentation of brain tumors, however, state-of-the-art methods focus on multi-modal data and do not retrieve as good results with one input modality. As for image registration, aligning two images with different topologies (for instance a healthy brain and a cancerous brain) is still an open problem. This thesis is therefore an attempt to solve those two problems.

1.2 Goal

In this thesis, we are interested in building an atlas of brain glioblastoma. In medical imaging, an atlas provides a framework for understanding and interpreting other images of the same body part. Atlases often include detailed information about the normal appearance and variations of the tissue or organ in question. In other words, building an atlas consists in constructing the statistical representation of a population (Joshi et al., 2004; Gori et al., 2017). The average representation cannot be estimated by simply computing the voxel-wise average of every image because, as shown in Figure 1.2, brain images can have substantial shape differences between patients. Each image needs to be mapped into a reference coordinate system. Therefore, an atlas results in an average scan (or template) of the population under analysis and in a set of subject-specific deformations which align the estimated template onto the subject scan. The variability in the population is expressed through the transformations. The estimation is commonly done by minimizing a function of the

form:

$$(\hat{I}, \hat{T}) = \arg\min_{I,T} \sum_{i=1}^{N} S(I, I_i \circ T_i) + R(T_i) \tag{1.1}$$

where $\hat{I}$ is the estimated template,$\hat{T} = (T_i)_{i\in[|1,N|]}$ is the set of estimated transformations, $(I_i)_{i\in[|1,N|]}$ is the set of images, S is a function to measure the similarity between two images and R is a regularization function on the transformations. The choice of the type of transformation is therefore of the utmost importance to correctly estimate the average representation.

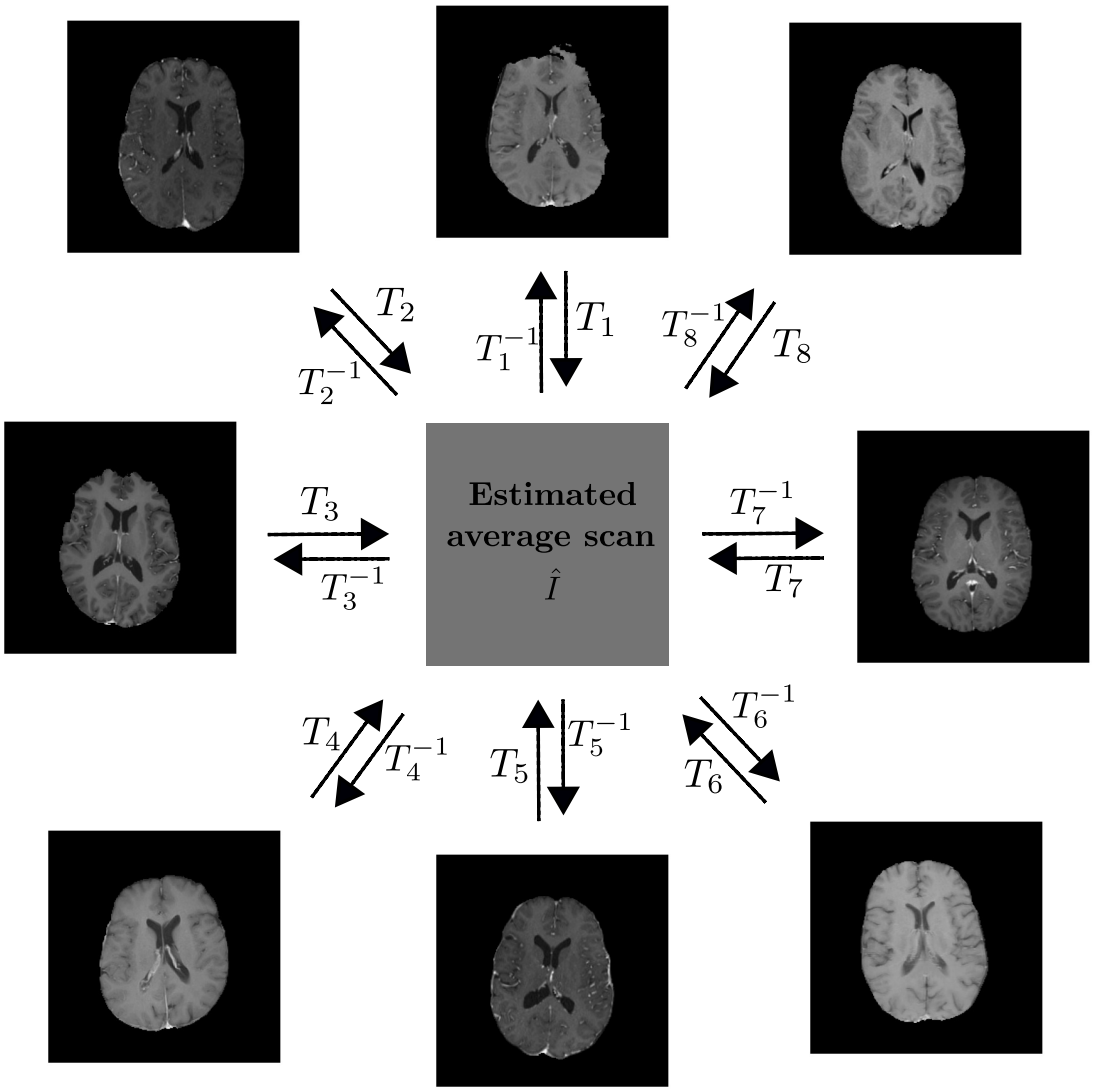


Figure 1.2: Framework for the construction of a template $\hat{I}$ and the associated deformations.

Classical registration algorithms implicitly assume that the anatomical structure of both images is the same and that one can therefore build a one-to-one correspondence between patients' images. Although this assumption is satisfactory with healthy patients, the assumption is too strong in the presence of pathologies such as glioblastoma. A glioblastoma may entail three kinds of

variations between two images: 1- *Topology*, 2- *Appearance* and 3- *Shape*. The first one is due to the presence of the tumor, which means that images might have a different number of components. The second difference is related to a variation in intensity (appearance) which can be caused, for instance, by the tumor infiltration. The third variation is instead caused by the tumor growth which deforms the surrounding healthy tissues (mass effect).

Because of these variations, building an atlas of glioblastoma is still not a well-defined problem. The literature contains a few papers that build a probabilistic atlas which consists in overlaying the spatial distribution of the tumors on a healthy template (Gooya et al., 2012; Roux et al., 2019). However, this type of atlas does not allow performing statistical analysis on the variations caused by the tumor such as the mass effect or the tumor infiltration, for instance.

Furthermore, building an average representation of glioblastoma is not straightforward since the locations of the tumors differ across patients. A potential approach could be to divide the brain into relevant anatomical regions (e.g. temporal lobe) and then estimate one atlas per region using, for each atlas, only the images with tumors present in the respective region. Subsequently, one could estimate, for instance, the average shape and variability of the tumor, the morphological variations due to the mass-effect or the appearance changes due to an infiltration.

As a result, to build an atlas of brain glioblastoma, one needs to first build a transformation method than can deal with shape, appearance *and* topology differences. The main goal of this thesis is therefore to determine a method that can correctly align images with varying topology.

1.3 Challenges

An image can be transformed in two manners: changing the intensities of the image, *i.e.* modifying the values of a voxel, or changing its shape, *i.e.* modifying the position of a voxel. The healthy tissues in the cancerous image need to be dealt with a shape deformation since the corresponding tissue is present in the healthy image. By contrast, there is no correspondence for the tumor in the healthy image so it needs to be dealt with the intensity transformation. To prevent the intensity transformation from modifying healthy tissues, we propose to limit its reach to the tumor region. Therefore, we require to first detect the position where the topology difference occurs, *i.e.* to segment the brain tumor. State-of-the-art brain segmentation methods are all based on convolutional neural networks and use four imaging modalities (T1, T1ce, T2, and Flair). When applying those methods with only one modality, their performance significantly decreases.

Thus, the **first challenge** of this thesis is to design a strategy to improve the segmentation when using only one modality. Since we have access to a large multi-modal database, we want to use it to help the model during training. The proposed approach, inspired by generalized knowledge distillation (Lopez-

Paz et al., 2016), consists in first training a multi-modal teacher network with the four modalities. Then, the teacher guides a student model during its training with only one input modality.

The **second challenge** is to register two images with different topologies when the localization of topological change is known. The proposed method should be able to disentangle the differences in shape and appearance. This means a method that can simultaneously modify the appearance and the geometry of the source image in an anatomically plausible and interpretable manner. The designed method changes the gray-levels of the source image to "erase" the tumor and simultaneously aligns the source image with the target. It requires the segmentation mask to only modify the intensities in the tumor region.

Furthermore, the **third challenge** is to design a solution that offers fast registration time. Indeed, conventional methods can require a lot of memory and computational time, especially in 3D. The method should process 3D images quickly in order to be able to construct an atlas. The method is incorporated into a learning-based framework that offers significantly faster registration times than classical methods.

1.4 Outline

This dissertation is organized into 6 chapters. In Chapter 2, we introduce important background notions on medical image segmentation with a focus on brain tumor segmentation. Namely, we describe the dataset that we used throughout this thesis and present the recent advancement made in learning-based segmentation.

Chapter 3 describes our knowledge distillation approach to transfer the knowledge of a multi-modal teacher network to a mono-modal student model. We compare its performance with other existing strategies such as attention transfer, contrastive distillation, and adversarial approaches. We show that the strategies are able to distill additional knowledge into the student model only when the number of training data is limited.

In Chapter 4, we explain the classical registration algorithms and, in particular, the LDDMM framework (Dupuis et al., 1998; Beg et al., 2005) which is at the core of our proposed registration method. Then, we introduce the learning-based strategies which offer faster registration times than previous methods.

In Chapter 5, we detail our transformation model. It is based on the Metamorphosis framework developed by Trouvé and Younès (2005), which simultaneously alters both the shape and intensities of the source image. We propose two methods: the first one relies on the geodesic shooting of Metamorphosis and the second one uses a residual neural network (ResNet) to solve the system of differential equations that control the framework. We show that our models outperform existing strategies while being very fast at inference time.

In the concluding chapter, we present how the methods developed in this thesis could be used for atlas construction, namely for brain glioblastoma.

1.5 Contributions

This thesis has two main contributions:

- We propose a learning framework to improve the accuracy of single-modality segmentation neural networks. The method is generic in the sense that it can be used for any neural network architecture and with any type of multi-modal images. To the best of our knowledge, this is the first time that the generalized knowledge distillation strategy is adapted to guide the learning of a mono-modal student network using a multi-modal teacher network.
- We propose a learning-based implementation of Metamorphosis (Trouvé and Younès, 2005), which considerably decreases the computational time at inference. We introduce a regularization mask to prevent the intensity changes from actually modifying the shape of the image. We use this mask as prior for topological changes. Indeed, the topological and appearance changes only occur in the spatial location of the mask, therefore we force any transformation outside of the mask to be a shape deformation. The model has been developed for brain glioma but it is generic and can be used with any imaging modality and any tumor type.

1.6 Publications

International Conferences:

M. Hu*, M. Maillard*, Y. Zhang, T. Ciceri, G. La Barbera, I. Bloch, and P. Gori. Knowledge distillation from multi-modal to mono-modal segmentation networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 772–781. Springer, 2020

M. Maillard, A. François, J. Glaunès, I. Bloch, and P. Gori. A deep residual learning implementation of metamorphosis. In *IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–4, 2022b

Workshop:

A. François, M. Maillard, C. Oppenheim, J. Pallud, I. Bloch, P. Gori, and J. Glaunès. Weighted metamorphosis for registration of images with different topologies. In *Biomedical Image Registration: 10th International Workshop, WBIR 2022, Munich, Germany, July 10–12*, pages 8–17. Springer, 2022

This work received the best paper award during the workshop.

National Conference:

* equal contribution

M. Maillard, I. Bloch, and P. Gori. Recalage métamorphique d'images par réseau de neurones résiduels. In *Groupe de Recherche et d'Etudes de Traitement du Signal et des Images (GRETSI)*, 2022a

Submitted to IEEE Transactions in Medical Imaging:

M. Maillard, I. Bloch, and P. Gori. Metamorph: Learning-based metamorphic registration of pathological images. *Submitted*, 2023

2

Medical Image Segmentation

Contents

Medical image segmentation is a crucial technique in medical imaging analysis that involves separating the image into different regions or segments based on the anatomical or pathological features of interest. This technique is useful in many medical applications including:

- Improving the accuracy in diagnosis: segmentation allows for a more accurate and detailed analysis of medical images, which can help clinicians to identify and diagnose various diseases, such as tumors, infections, and other anomalies (Jiji et al., 2013; Heinonen et al., 1998).
- Treatment planning: Accurately segmented medical images can be used to plan and guide medical interventions, such as surgeries, radiation therapy, and other procedures. By understanding the location, size, and shape of the target area, clinicians can plan treatment strategies that maximize effectiveness and minimize side effects (Kikinis et al., 1996; Virzì et al., 2020).
- Disease monitoring: segmentation is also useful in monitoring the progression of diseases, such as cancer, over time. By comparing segmented images taken at different intervals, clinicians can assess the effectiveness

of treatments and make necessary adjustments to treatment plans. (Behbehani et al., 2018; Zhang et al., 2021; Amoruso et al., 2021)

There is very extensive literature on medical image segmentation. Classical methods include several strategies such as thresholding (Batenburg and Sijbers, 2009), level-sets (Mumford and Shah, 1989; Chan and Vese, 2001), active contours models (Kass et al., 1988; Cohen, 1991; Xu and Prince, 1998), clustering (Coleman and Andrews, 1979; Pappas and Jayant, 1989) or atlas-based methods (Lorenzi et al., 2013; Wang et al., 2012). The main drawback of these methods is that their results are not easily generalizable. Either they are semi-automatic and tuning the hyper-parameters can be long, and requires experience or, in the case of automatic methods, the parameters that correctly segment one image do not necessarily generate a satisfactory segmentation for other images. With the advent of large databases and better computational capabilities, deep learning methods have become state-of-the-art algorithms for the segmentation of medical images.

In this chapter, we first present the Brain Tumor Segmentation (BraTS) dataset, which is the main dataset used in this thesis. Secondly, we explain the learning methods for segmentation and especially the best-performing ones on BraTS. Then, we describe the data processing strategies. Finally, we discuss the different training loss functions and the measures to evaluate the quality of a segmentation result.

2.1 BraTS Dataset

The BraTS dataset is a widely-used publicly available dataset of brain tumor MRI scans (Menze et al., 2015; Bakas et al., 2017; Baid et al., 2021). It was created in 2012 to support the development and evaluation of algorithms for the automatic segmentation of brain tumors into three different subregions: enhancing tumor, necrotic tumor core, and edema. The BraTS dataset includes four modalities of MRI scans (T1, T1ce, T2, Flair), along with corresponding ground truth labels that indicate the location and extent of different tumor subregions. The dataset has been updated and expanded over time, and several versions of the BraTS dataset are available. The most recent version, BraTS 2021, includes images from 2,000 patients with corresponding annotations. It is important to note that every patient has a tumor, and there is no healthy element in the database.

All volumes, coming from different institutions, have been co-registered to the anatomical template SRI24 (Rohlfing et al., 2010) and resampled to a 1mm^3 voxel size. Afterward, the skulls have been removed from the MRI scans. The size of the final images is $240 \times 240 \times 155$ for every patient. The reference segmentations were collected by manual annotations from expert radiologists. The segmentations were reviewed by two senior neuroradiologists. If rejected, the scans were returned to the expert until the satisfaction of both reviewers. The reference segmentations contain four labels: background or non-

cancerous tissue (label 0), the necrotic and non-enhancing tumor core (label 1), the peritumoral edematous (label 2), and the enhancing tumor (label 4). However, the evaluation of each segmentation is done on three hierarchical regions. First, the enhancing tumor (ET) which corresponds to label 4. Second, the tumor core (TC) which is the union of the enhancing tumor and the necrotic tumor core (label 4 and 1). Finally, the whole tumor (WT) which is the union of the tumor core and the edema (all labels).

Figure 2.1 displays the four modalities for several patients. It illustrates the variability in terms of tumor location, size, and shape. Additionally, it shows that some patients do not have an enhancing tumor (ET). It is interesting to notice that each modality highlights a different label. The contrast between the edema and the other tissue types is much higher on the T2 and Flair modalities than on T1 and T1ce. By contrast, the necrotic tumor core stands out much more in the T1ce than in the other modalities.

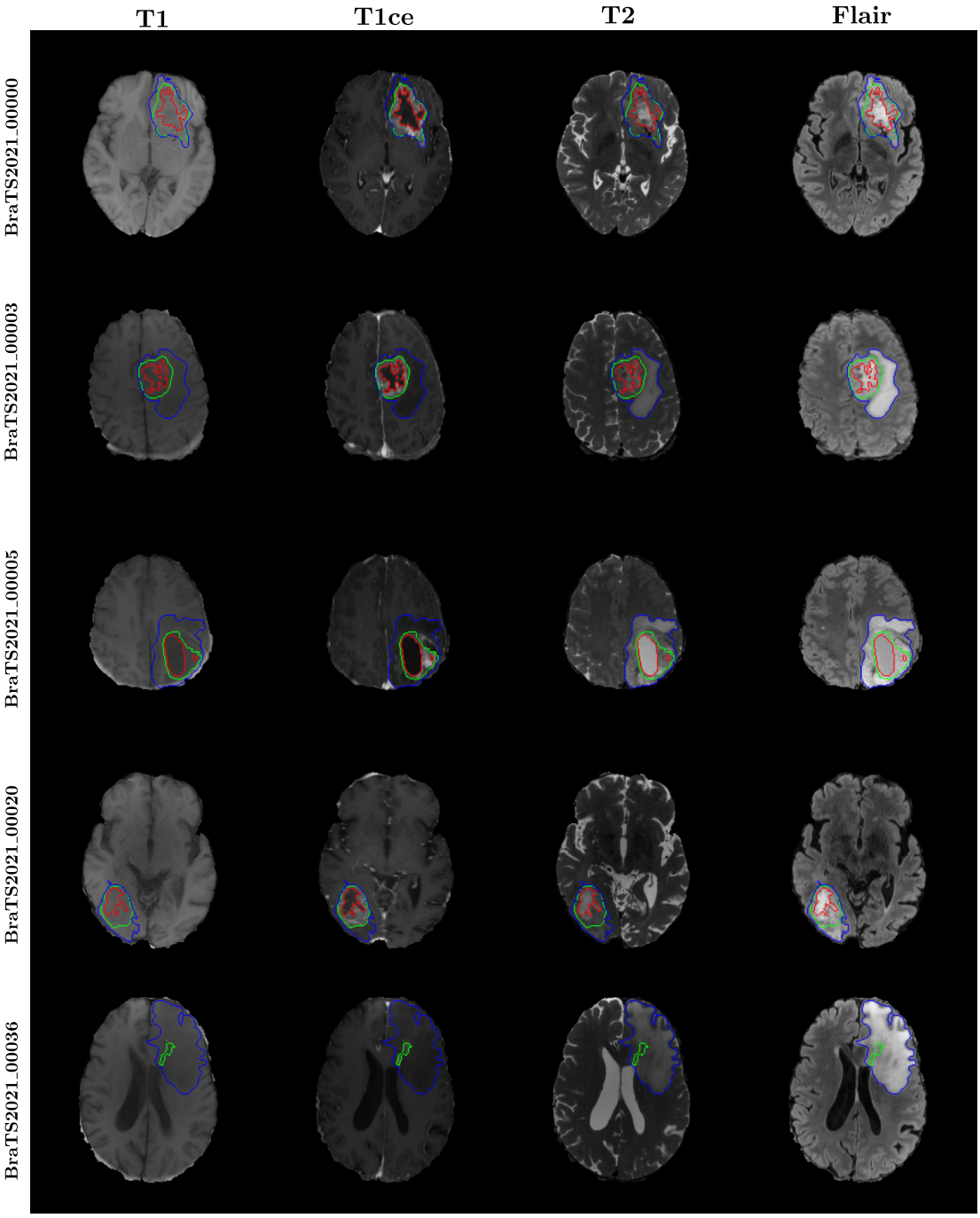


Figure 2.1: Example of images from the BraTS 2021 dataset for 5 patients. The red line is the outline of the enhancing tumor, the green line is the outline of the tumor core and the blue line is the outline of the whole tumor.

2.2 Learning-based segmentation

Early learning methods used classical supervised algorithms, such as support vector machines (García and Moreno, 2004; Zhang et al., 2004; Lee et al., 2005) or decision trees (Zikic et al., 2012; Mitra et al., 2014; Goetz et al., 2014) , which required to pre-specify a set of features. Recently, convolutional neural

networks (CNN) have become the standard method for image segmentation, in particular because they are powerful feature extractors. CNNs automatically extract features and classify each voxel based on them. For most methods, the training of the model takes the form of the one presented in Figure 2.2. What differentiates the methods essentially depends on the following criteria:

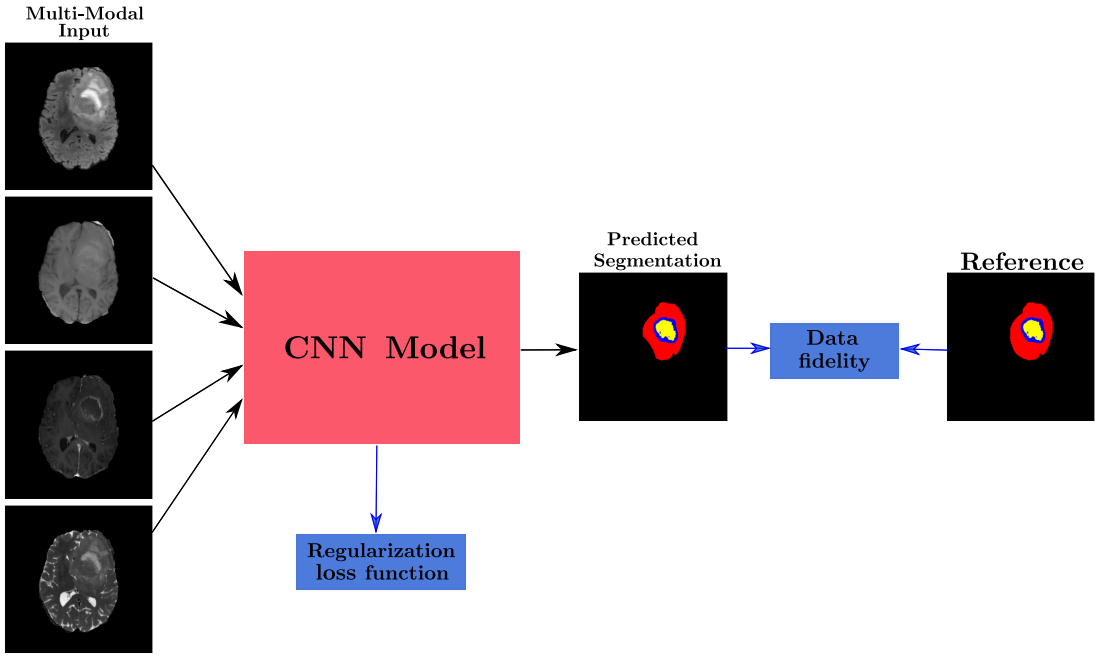


Figure 2.2: General architecture of CNN-based segmentation models.

- how the model runs through the 3D volume;
- the architecture of the CNN;
- the data pre-processing and augmentation;
- the training loss function;

2.2.1 Reading the 3D volume

Due to the heavy load of CNNs, early methods did the segmentation on 2D slices (Zikic et al., 2014; Havaei et al., 2017). To retrieve the whole segmentation of the brain, the model is applied to the 3D volume using a sliding window. Another strategy, called 2.5D, consists in selecting a limited amount of consecutive slices (for instance 3, so that ImageNet initialization can be leveraged) and using 3D convolutions on this limited volume (Ziabari et al., 2018; da Cruz et al., 2022). A similar perspective, between 2D and 3D, is to segment the axial, coronal, and sagittal views with three parallel CNNs and fuse the predictions (Hu et al., 2019; McKinley et al., 2016). Currently the most commonly used technique is to work on 3D patches of the volume (Myronenko, 2019; Isensee et al., 2019a). The benefit of these last three strategies is

that it gives 3D information to the network while limiting the memory load. Nowadays, the GPU capacities allow to feed the full 3D volume into the neural networkLuu and Park (2022). Although the computation is more cumbersome, having the full 3D volume greatly helps the segmentation.

2.2.2 Architectures

The most common segmentation architecture is the U-Net model (Ronneberger et al., 2015) and its extension to 3D, like the V-Net (Milletari et al., 2016; Çiçek et al., 2016). There have been several variations of it (Ibtehaz and Rahman, 2020; Zhou et al., 2020b; Huang et al., 2020; Isensee et al., 2021; Luu and Park, 2022) but the general architecture stays the same: an encoder model progressively downsamples the feature maps and a decoder upsamples them back to the original image size. To keep the spatial information, the encoder's feature maps at each scale are fed to the decoder with skip connections (see Figure 2.3). The main characteristics that differentiate U-Net models from one another are:

- the number of down/upsampling layers and their type (strided convolutions, max pooling, average pooling, etc.);
- the number of convolution kernels, their size, and the number of channels.
- the non-linear activation functions (sigmoid, ReLU (Nair and Hinton, 2010), LeakyReLU (He et al., 2015), etc.);
- the type of normalization layer (Batch normalization(Ioffe and Szegedy, 2015), Instance normalization (Ulyanov et al., 2016), etc.);

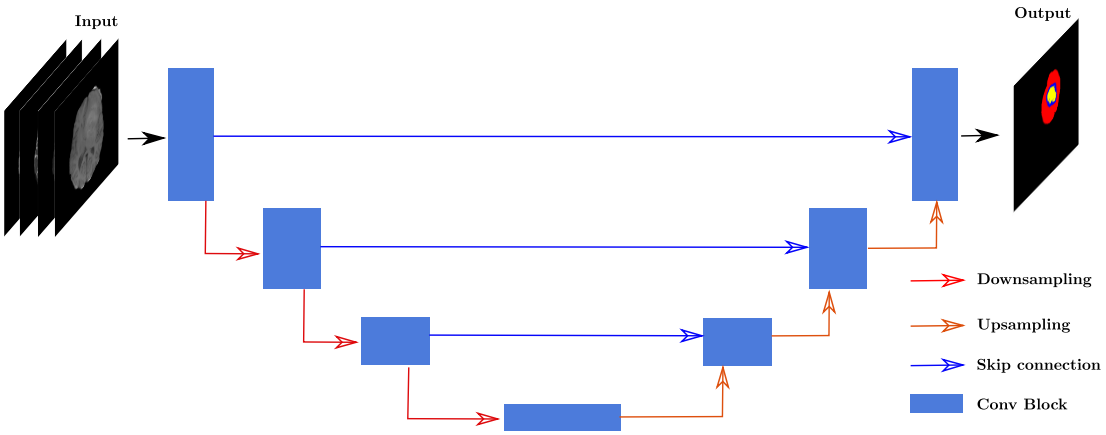


Figure 2.3: Architecture of a U-Net model (Isensee et al., 2019a). Conv block indicates a sequence of convolutions and non-linear operations on the features.

Other architectures such as cascaded networks (Wang et al., 2018; Jiang et al., 2020; Li et al., 2020) have been proposed. They consist in predicting the labels in a sequence of neural networks where each network predicts one label.

Other methods add a decoder model to a U-Net to reconstruct the input image (Myronenko, 2019; Jiang et al., 2020). The reconstruction branch is used as regularization. Furthermore, Ibtehaz and Rahman (2020) speculate that fusing the low-level features from the encoder and the corresponding high-level ones after the skip-connection is a flaw. They propose to incorporate convolutional layers in the skip-connection to alleviate the level disparity.

The fusion of modality information can be done in three different ways: *1-early fusion*, the modalities are processed simultaneously at the input level by concatenating them and treating them as channels; *2-middle fusion*, the information of the modalities is fused in the middle layers of a network; *3-late fusion*, the modal information is merged at the decision level of the model. In the second category, we can cite Dolz et al. (2018) who use one neural network per modality, and each branch is connected by taking as input the intermediate representations of the other branches. Similarly, Zhou et al. (2020a) use an attention mechanism to fuse the representations of each modality at different scales. Among the late fusion methods, Nie et al. (2016) process each modality independently with a fully convolutional network and fuse their representations in the last layer. In this category, we could also add ensembling methods which consist in predicting the segmentation map with several models and merging them by computing an average (Kamnitsas et al., 2018). It's not strictly a late fusion of the modalities because each model takes all the modalities as input but they merge the decision in the last layer.

Although these methods reach good results, Isensee et al. (2019a) and Luu and Park (2022) have shown that a well-trained U-Net-like model can achieve better scores on BraTS dataset. U-Net based models have also reached the first place in other segmentation challenges such as Kidney Tumor segmentation (KiTS) (Heller et al., 2021). The three best-performing methods of the 2021 occurrence of the challenge used 3D U-Net models (Zhao et al., 2022; Golts et al., 2022; George, 2022), proving that this architecture is currently the state-of-the-art method for 3D medical image segmentation. A central idea in (Isensee et al., 2019a) is that, as important as the architecture, the data preprocessing and the training strategy are crucial when it comes to achieving the best results.

2.3 Data preparation

2.3.1 Pre-processing

MRI can suffer from several quality limitations that are important to address before segmentation. Ideally, an MRI would be a piecewise constant function where voxels from the same tissue type have the same value. This is not the case in practice and the acquired images suffer from intensity nonuniformity (Figure 2.4). A popular strategy to deal with this is the N3 Bias correction algorithm (Sled et al., 1998) and its improved version, the N4 Bias field correction (Tustison et al., 2010). They consider the formed image as a product of the

ideal image and a bias field. The latter is estimated to produce the corrected image.

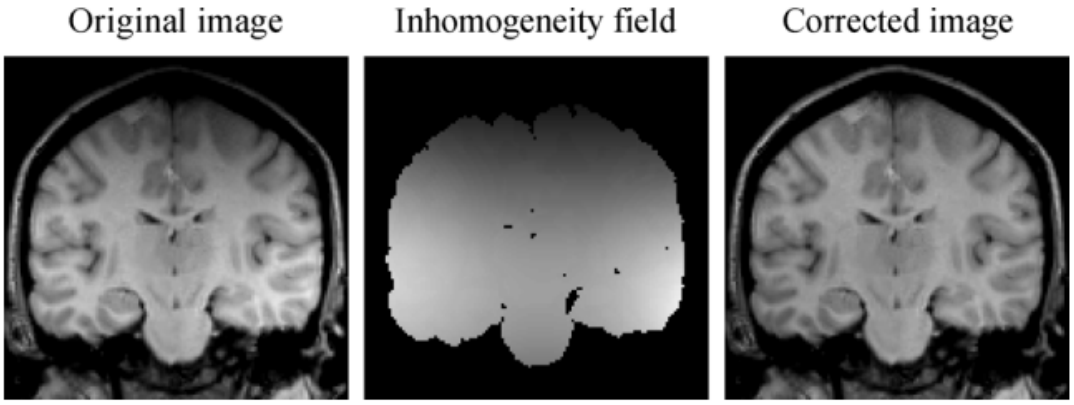


Figure 2.4: Image with an intensity inhomogeneity and the associated correction (Vovk et al., 2007).

Another limitation of MRI is the non-normalization of intensity values. Images are taken from different scanners, with different parameters. Thus, there is a high variability in image intensities between patients (Carré et al., 2020). This can seriously hinder the generalization of segmentation models. First, the min-max normalization is rarely applied as it is very sensitive to outliers. A fast and easy-to-implement standardization is the z-score. It consists in subtracting the average values in the brain region (μ_{brain}) from the image and dividing by the standard deviation in the same region (σ_{brain}):

$$I_z = \frac{I - \mu_{brain}}{\sigma_{brain}}.$$

Figure 2.5 shows the effect of the z-score standardization on the histograms of the four modalities of a patient from the BraTS dataset. Following standardization, the values are centered around 0 and are spread on a similar scale.

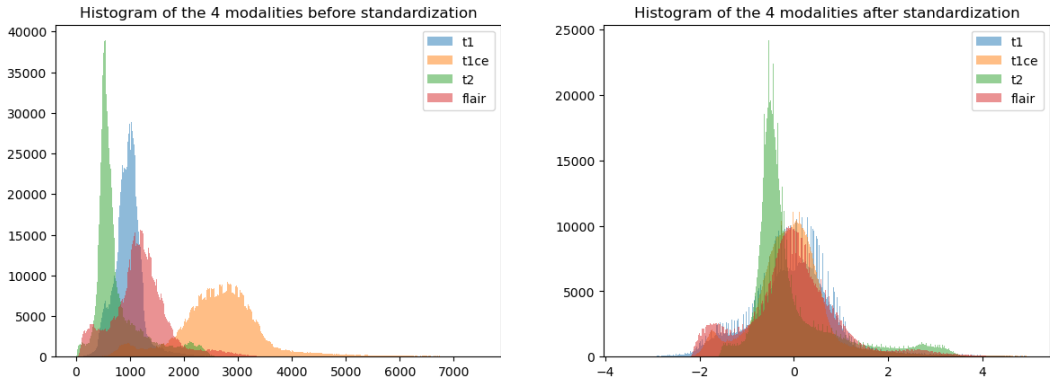


Figure 2.5: Histograms for the four modalities of a single patient from the BraTS dataset before and after standardization with the z-score.

Other methods include WhiteStripe (Shinohara et al., 2014), piecewise linear histogram matching (Nyúl et al., 2000), Ravel (Fortin et al., 2016). However, Carré et al. (2020) and Reinhold et al. (2019) showed that for a variety of tasks, there is no significant difference in terms of results. The only difference occurs between no-standardization and standardization. As a result, the z-score method is used in most state-of-the-art brain tumor segmentation methods for its simplicity (Myronenko, 2019; Jiang et al., 2020; Isensee et al., 2021; Luu and Park, 2022).

2.3.2 Augmentation

Due to the limited amount of patients in medical datasets, data augmentation is a crucial step to avoid the overfitting of the model. It consists in synthesizing new data from the existing one. Most common strategies use simple transformations such as random scaling, rotations, intensity shifts, and flipping of the image along an axis. The best method for BraTS 2021, is the one by Luu and Park (2022) and uses random rotation and scaling, elastic deformation, additive brightness augmentation, and gamma scaling. Isensee et al. (2021) add the random mirroring of the volumes. Extracting random crops from the image can also be considered as data augmentation since it allows getting several views from a single scan. The benefit is double since it also reduces the computational load of the training. Furthermore, it introduces samples without tumors, making the model more robust to potential healthy cases.

2.3.3 Post-processing

For the task of brain tumor segmentation, a common post-processing of the predicted segmentation map is to label enhancing tumor as necrosis if the total number of voxels labeled as enhancing tumor is smaller than a pre-specified threshold (Isensee et al., 2019b). This stems from the fact that some patients do not have an ET label and a single mislabeling in such cases leads to a large error. Another post-processing methods consits in removing all but the largest connected component (Heller et al., 2021). Nevertheless, this is not adapted to brain tumors since multi-focal tumors are possible

2.4 Loss function and evaluation measures

2.4.1 Training Loss functions

The problem of segmentation can be framed as a voxel-by-voxel classification problem. For every voxel $x \in \Omega$, the model returns the probability $p_c(x)$ of belonging to each class c. To train the model, a common strategy is to minimize the cross entropy between the predicted probabilities p_c of each class c and the

reference segmentation q_c:

$$CE(p,q) = -\frac{1}{|\Omega|}\sum_{x\in\Omega}\sum_{c=1}^{C} q_c(x)\log p_c(x).$$

where C is the number of classes. However, in 3D segmentation, there is often a class imbalance. Namely, in brain tumor segmentation, the number of non-cancerous voxels is much bigger than the one of cancerous voxels. Handling the voxels independently with a cross-entropy loss favors the majority label. To prevent this, the focal loss (Chang et al., 2018) balances the contribution of each class in the cross-entropy loss function with their probabilities:

$$FL(p,q) = -\frac{1}{|\Omega|}\sum_{x\in\Omega}\sum_{c=1}^{C} \alpha_c q_c(x)(1-p_c)^{\gamma_c}\log p_c(x)$$

where $\alpha_c > 0$ and $\gamma_c \geq 1$ are fixed hyper-parameters for each class. The idea is to emphasize the badly segmented voxels in the training loss. Furthermore, the weights α_c are selected depending on the number of occurrences: with fewer occurrences, a class should have a higher weight relative to the other classes.

Milletari et al. (2016) proposed to minimize a global functional based on the Dice score to deal with class imbalance. The Dice score measures how two ensembles A and B overlap:

$$Dice(A,B) = \frac{2|A\cap B|}{|A|+|B|}.$$

The Dice score is equal to 1 if both shapes perfectly match and 0 if there is no common element between them. In the context of training a segmentation neural network, we need to minimize a function therefore, the Dice score becomes

$$Dice(p,q) = 1 - \frac{1}{C}\sum_{c=1}^{C}\frac{2\sum_{x\in\Omega} p_c(x)q_c(x)}{\sum_{x\in\Omega} q_c(x) + \sum_{x\in\Omega} p_c(x)}.$$

To further address the class imbalance issue, a focal version of the Dice score has been proposed by Wang and Chung (2018). Similarly to the focal cross-entropy, it uses hyper-parameters $\gamma_c \geq 1$ and $\alpha_c > 0$ to weigh each class contribution:

$$FDice(p,q) = \sum_{c}^{C}\alpha_c(1 - Dice(p_c,q_c)^{\frac{1}{\gamma_c}}).$$

Both focal methods require setting numerous hyper-parameters which can be complicated in presence of several classes. Namely, on BraTS, it would require choosing eight hyper-parameters for this loss function only. As a result, most papers use the combination of the Dice and the cross-entropy since they are simple to implement, introduce both global and local constraints, and have shown to be sufficient to deal with the problem of class imbalance.

Another type of training loss function is inspired by generative adversarial networks (Goodfellow et al., 2020). In contrast to the two other loss functions, it consists of an implicit strategy. Indeed, in such a context, a neural network (called discriminator D) is trained to predict whether a segmentation map is the reference segmentation or comes from the segmentation model (called generator G). In parallel, the goal of the generator is to "trick" the discriminator into thinking that the prediction is the ground truth. The framework is shown in Figure 2.6. For a batch of size N constituted with N pairs of image I^n and the associated reference segmentation q^n, a common training loss function is:

$$\min_G \max_D \sum_{n=1}^{N} \log D(q^n) + \log(1 - D(G(I^n))).$$

However, since the discriminator returns only a single value for the whole image, the above loss function might not offer the information to localize the difference with the reference segmentation. Xue et al. (2018) used the feature maps at every scale of the discriminator to introduce a spatial component in the adversarial loss function. Let $f^l(I^n, q^n)$ be the output of the l^{th} layer of the discriminator, the adversarial loss function is then:

$$\min_G \max_D \sum_{n=1}^{N} \sum_{l=1}^{L} \|f^l(I^n, q^n) - f^l(I^n, G(I^n)\|_1.$$

The main drawbacks of adversarial methods are that the training can be unstable and the discriminator model needs more memory and time compared with a non-adversarial strategy.

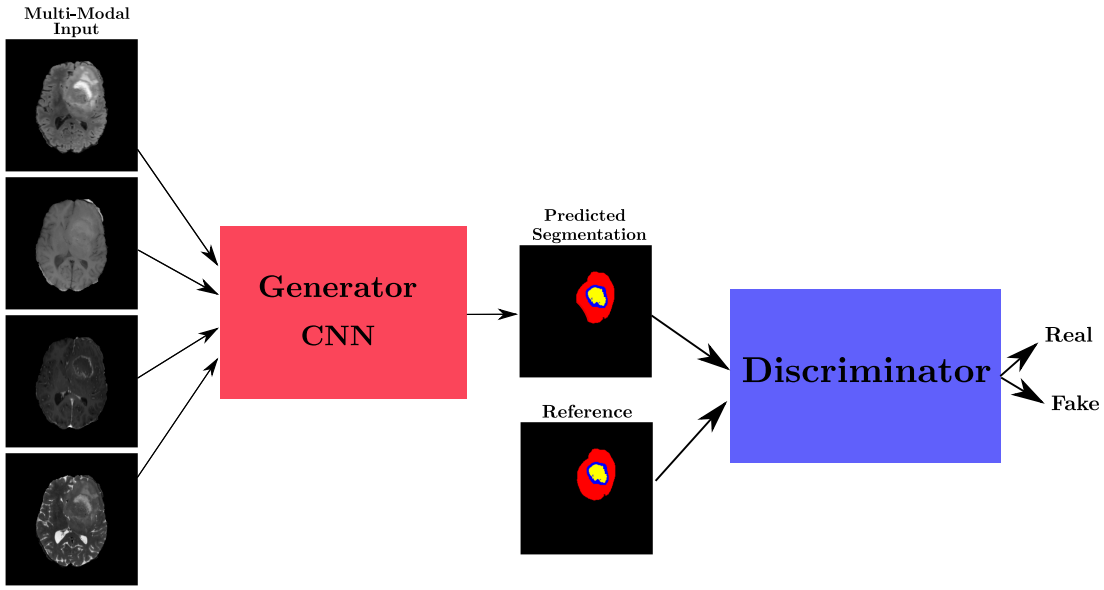


Figure 2.6: Adversarial framework for segmentation.

2.4.2 Evaluation

Evaluation functions are essential to assess the quality of the model. Classical measures use the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FP) notions, also used for classification. Their representation is visible in Figure 2.7. First, the most widely used measure is the Dice score, which, as we have seen before, measures the overlap between two shapes. In a similar fashion, the Jaccard index can also be used:

$$Jac(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{TP}{FN + FP + TP} = \frac{Dice}{2 - Dice}$$

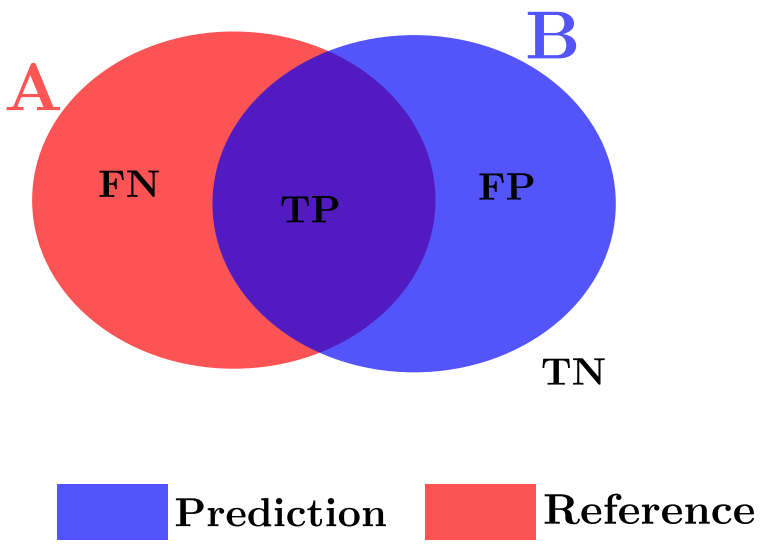


Figure 2.7: Overlap between a predicted segmentation and the reference. TP stands for True Positive, TN for True Negative, FN for False Negative, and FP for False positive. Naturally, we want to maximize the number of TP and TN for a better prediction.

Additionally, the precision, sensitivity, and specificity scores are also commonly used measures.

$$Precision(A,B) = \frac{TP}{TP + FP}$$

$$Sensitivity(A,B) = \frac{TP}{TP + FN}$$

$$Specificity(A,B) = \frac{TN}{TN + FP}$$

Finally, the Hausdorff distance measures the distance between the contours of A and B. The idea is that for any point on the contour of A, its distance from the closest point in B should be small, and vice-versa. This is done with the following distance:

$$HD(A,B) = \max\left\{\max_{a\in A}\min_{b\in B}||a-b||\,,\quad \max_{b\in B}\min_{a\in A}||a-b||\right\}$$

It is worth noting that the Hausdorff distance is sensitive to outliers. For instance, in Figure 2.1, the ET label for patient BraTS_00005 is constituted of a large component and a much smaller element next to it. If the predicted segmentation accurately segments the large part but does not detect the smaller one, the Hausdorff distance will be relatively high. By contrast, due to the disparity of size between both components, the Dice score or the Jaccard index would only be slightly affected by the misclassification. Thus, the user needs to choose one measure over the other depending on their task.

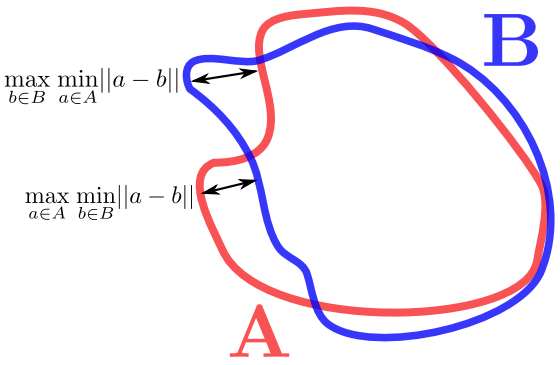


Figure 2.8: Schematic view of the Hausdorff distance. It corresponds to the maximum value between the two distances shown above.

2.5 Conclusion

In this chapter, we reviewed the learning strategies for 3D segmentation, with a focus on multi-modal MRIs of brains containing a tumor. With appropriate data pre-processing and multiple augmentation strategies, the U-Net framework is the state-of-the-art strategy for this task as it obtained first place in several challenges. Hence, we will use this framework as a backbone neural network in the next chapter.

The presented methods have been developed for multi-modal data, however, if only a single modality is available at inference time, the performance of those algorithms significantly decreases. This is expected since each modality only highlights a specific type of tissue. However, in a clinical setting, it is common to have only one available modality for a patient, thus, seriously reducing the applicability of automatic segmentation models in the clinical context. The problem evaluated in the next chapter is how can we use large multi-modal datasets, developped for research, such as BraTS, to improve the segmentation of a neural network that takes ony one modality as input.

3

Transferring knowledge from a multi-modal network to a mono-modal network

Contents

3.1 Introduction

Using multiple modalities to automatically segment medical images has become a common practice in several applications, such as brain tumor segmentation (Menze et al., 2015) or ischemic stroke lesion segmentation (Maier et al., 2017). Although multi-modal models usually give the best results, it is often difficult to obtain multiple modalities in a clinical setting due to a limited

number of physicians and scanners, and to limit costs and scan time. In many cases, especially for patients with pathologies or in case of emergency, only one modality is acquired. Therefore, in this chapter, the context is the segmentation of brain tumors with only one available modality at test time, while multi-modal data are available during training.

Three main strategies have been proposed in the literature to deal with problems where multiple modalities are available at training time but some or most of them are missing at inference time. The first one is to train a generative model to synthesize the missing modalities and then perform multi-modal segmentation. The principal issue with this strategy is that it is quite tedious to train generative models and even more so when more than one modality is missing. The second strategy consists in learning a modality-invariant feature space that encodes the multi-modal information during training, and that allows for all possible combinations of modalities during inference. This strategy is well-adapted when the modalities are randomly missing but when the missing modality is fixed, training a U-Net on the fixed subset of available modalities is more efficient. Finally, the last strategy consists in distilling the knowledge of a trained multi-modal teacher network into a student model. The student model is trained only on a fixed subset of modalities. This last method fits our context since only one modality is available at inference time, and we build on this idea.

This chapter is organized as follows. First, we present all the existing work on the topic of segmentation with missing modalities with a focus on the teacher-student strategy. Then, we describe the method that we published during the MICCAI 2020 conference. Furthermore, we evaluate and compare it with other teacher-student strategies for the segmentation of brain tumors. Finally, we show that this approach is mainly beneficial in the presence of a limited amount of training data.

3.2 Related Works

3.2.1 Modality synthesis

The earliest strategy to deal with missing modality was to simply generate it from the available modalities. It was first executed with two basic generation models: a three-layer neural network and a Restricted Boltzmann Machine (RBM). The models were trained by minimizing the $\mathcal{L}_1$ distance between the predicted modality and the ground truth. Subsequently, a linear SVM or Random Forest were used for the classification of brain tumors (van Tulder and de Bruijne, 2015). The method showed an improvement compared to replacing missing sequences with zeros. However, the authors speculated that this is mainly due to the simplistic nature of the classifier. With more complex models such as deep CNNs, they argued that the accuracy would not improve.

In the previous method, the synthesis and segmentation models were trained independently. Orbes-Arteaga et al. (2018) proposed to simultaneously train a generative model and a segmentation network. In this case, both the generation and the segmentation were done with U-Net models (Ronneberger et al., 2015). The generative neural network was trained with the same reconstruction error as the previous method but it also benefited from the back-propagation of the segmentation error. This strategy achieved a better score than the previous proposition for the segmentation of white matter hypointensities segmentation.

The most popular framework for image generation in recent years is the Generative Adversarial Network (GAN). Naturally, it has also been experimented for the purpose of missing modality synthesis. Ben-Cohen et al. (2018) generated PET images from CT scans using a conditional GAN. The synthesis process operated in two stages: a fully convolutional network predicted a coarse virtual PET image from a CT scan following which the coarse image and the CT scan were fed to the cGAN to refine the first prediction. The authors demonstrated that this generative process helped reduce the number of false positives for the detection of the lesion in livers. Nevertheless, due to the blurry nature of PET images, they claimed that the cGAN is not adapted for segmentation.

Similarly to Orbes-Arteaga et al. (2018), a generative adversarial network and a segmentation model can be trained in an end-to-end fashion (Huang et al., 2022). In this framework, the authors showed an improvement in accuracy but only with one missing modality.

Yu et al. (2018) took an interesting perspective where they synthesized the missing modality by linearly combining the real target modalities from the training set. They used a cGAN to estimate the combination weights. Like the former strategy, a cGAN predicted the missing modality (in this case Flair), and separately a neural network was trained to segment a brain tumor from concatenated T1 and Flair volumes. At testing time, for each image I_{test} the following convex problem was solved:

$$\min_{w} \| \sum_{i=1}^{N} w_i GAN(I^i_{train}) - GAN(I_{test}) \|_2^2$$
$$\text{s.t.} \sum_{i=1}^{N} w_i = 1,\ w_i \geq 0$$

where N is the number of images in the training set. Afterward, the estimated Flair image $\hat{F}_{test}$ was the linear combination of the real Flair images from the training set F^i_{train}:

$$\hat{F}_{test} = \sum_{i=1} w_i F^i_{train}.$$

The segmentation model took $\hat{F}_{test}$ and I_{test} as input. The results on brain tumor segmentation indicated that it offered better predictions than just directly using the output of cGAN for the segmentation.

The main drawback of generating the missing modality is that it is computationally cumbersome, especially when many modalities are missing. In fact, one needs to train one generative network per missing modality in addition to a multi-modal segmentation network. Additionally, except in the work of Yu et al. (2018), this strategy has been mainly applied to 2D segmentation. Finally, the training of GANs is very unstable and long.

3.2.2 Learning a shared representation

A second strategy for the segmentation with missing modalities is to learn a common representation space. The idea is to train several encoders to map each modality into a modality-invariant latent space and sample the segmentation map from that space with a decoder, see Figure 3.1.

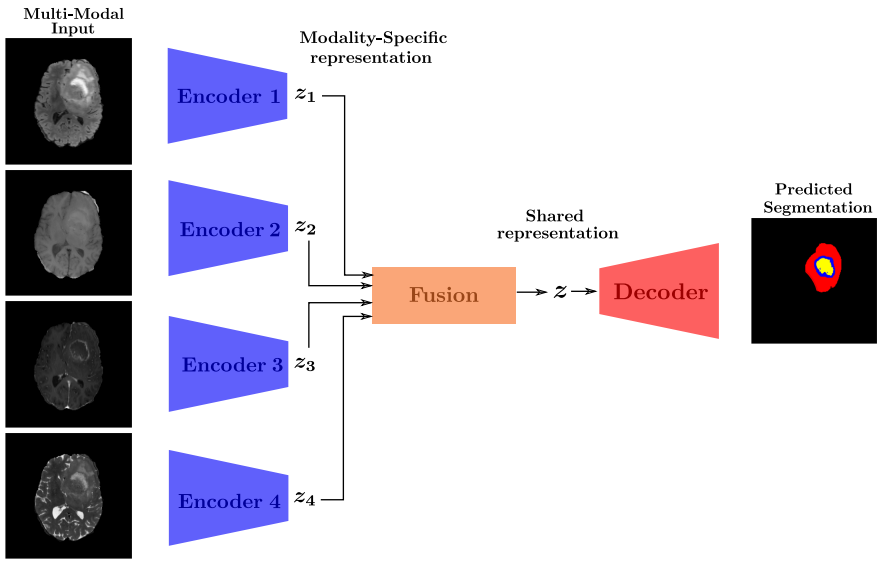


Figure 3.1: General framework of learning a shared latent space with missing modalities.

Havaei et al. (2016) were the first to introduce the method, called HeMIS. Each modality was mapped into a modality-specific feature space with a 2-layer fully convolutional encoder. The dimension of the representation map was the same as the one of the input image, simply it has 48 channels. Later, the representation vectors were fused by computing their mean and variance across the available modalities. If only one modality was available, the variance map was defined to be zero. Therefore the shared latent space did not depend on the number of available modalities. Subsequently, a 2-layer decoder predicted the segmentation. During training, modalities were randomly dropped to simulate missing modalities. This model being very shallow, it did not fully exploit the power of CNNs and therefore, the segmentation results were not satisfying.

Dorent et al. (2019) adapted the Multi-modal Variational Auto-Encoders (MVAE Wu and Goodman (2018)) to the missing modality case. The modality-specific latent spaces were trained to follow a Gaussian distribution. They were therefore represented by two vectors: one for the mean, and the other for the variance. In contrast to the former method, the volumes were downsampled through the encoder which allowed for deeper models. The embeddings were fused using a closed-form formula for the fusion of Gaussian processes. A sample was randomly drawn from the latent space and was decoded into the segmentation map. On top of this decoder, the model also learned a reconstruction decoder for each modality. The reconstruction of each modality served as regularization to enforce the representation space to encode all the information. To avoid the loss of spatial information when downsampling in the encoder, Dorent et al. (2019) considered the features before every downsampling stage as a representation map. With four downsampling stages, the model, therefore, had four embeddings. This strategy introduced skip connections in variational auto-encoders. The authors show that this strategy retrieved better segmentation maps than HeMIS and MVAE.

Using averaging or the product of Gaussians to merge the representation makes each modality contribute equally to the space which is inconsistent with the fact that each modality highlights different tissues. Chen et al. (2019) proposed to learn the fusion method rather than to hard code it. They used a gated fusion method where the modality-specific embeddings were concatenated and fed to a convolution layer that predicts weighting maps for each modality. If a modality was not available, a zero vector was used. Afterward, each embedding was multiplied by the weighting map, concatenated, and transformed into the shared latent vector. In parallel, the model reconstructed the input modalities for the purpose of regularization.

Similarly, Zhou et al. (2021) designed a vector fusion procedure by extracting spatial and channel attention. The spatial attention module was similar to the gated fusion method, in the sense that they used a convolution layer to compute weighting maps. For channel attention, the feature maps underwent a global average pooling so that each channel had only one value. Then, a 2-layer fully connected neural network predicted a weighting vector. The latter was multiplied by the input feature map to retrieve the channel attention representation. The final embedding corresponded to the sum of the channel attention representation and the spatial attention representation.

Learning a shared latent space was designed to deal with randomly missing modalities. The interest was that it could deal with any combination of modalities. Naturally, the best results are reached when only one or two modalities are missing. When only one modality is available, its performance is worse than a U-Net trained to only segment with this specific modality. This kind of method is therefore not suitable for a clinical setting where only one modality is usually acquired, such as pre-operative neurosurgery or radiotherapy.

3.2.3 Teacher-Student learning

A third approach consists in leveraging the knowledge of a trained multi-modal network (teacher) to train a model with missing modalities (Student). This is a case of learning with privileged information (Vapnik and Izmailov, 2015) because the teacher has additional knowledge compared to the student. The general structure of this strategy is displayed in Figure 3.2. We respectively call the teacher and student models f_T and f_S. I^i refers to the i^{th} multi-modal image in the training set and I_k^i denotes that only the modality k of the subject is selected.

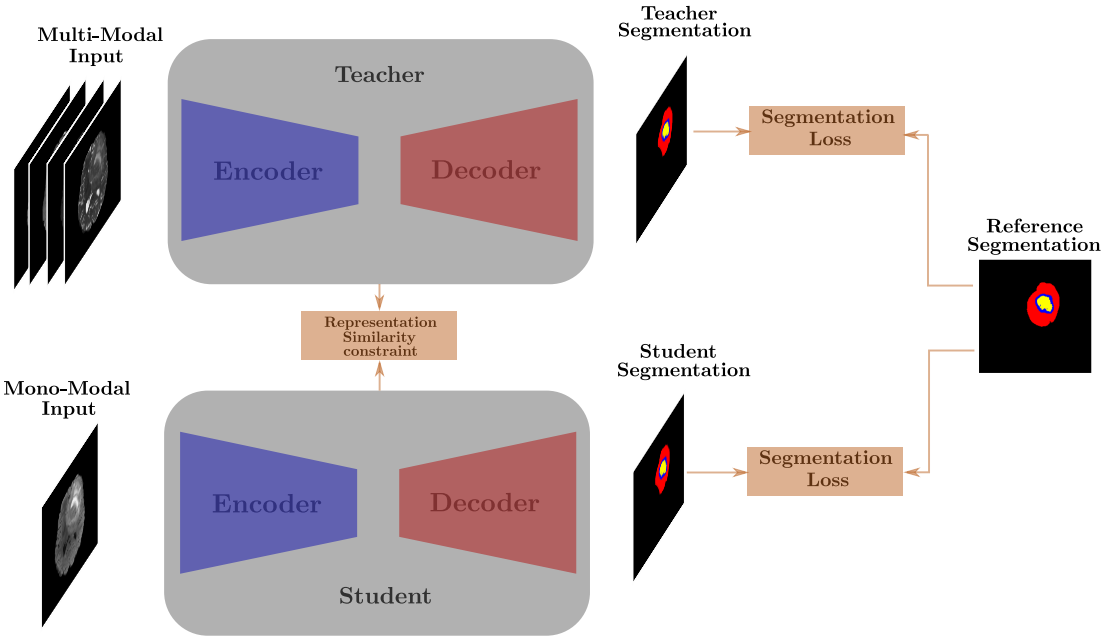


Figure 3.2: Overview of the Teacher-Student framework for transferring knowledge from the teacher model to the student model.

Knowledge distillation

The original framework of Teacher-Student learning has been presented by Hinton et al. (2015) and was called knowledge distillation. Originally, it was designed for model compression: distilling the knowledge of large teacher neural networks to small student models. Later, Lopez-Paz et al. (2016) incorporated it into the framework of learning with privileged information.

The key idea of generalized knowledge distillation is to transfer useful knowledge from the additional information of the teacher to the student using the soft label targets of the teacher. These are computed as follows:

$$s_i = \sigma(f_T(\mathrm{I}^i)/T) \tag{3.1}$$

where σ is the softmax function and T, the temperature parameter, is a strictly positive value. The parameter T controls the softness of the target, and the higher it is, the softer the target. The idea of using soft targets is to uncover

relations between classes that would be harder to detect with hard labels. This idea is illustrated in Figure 3.3 where the softened and non-softened predictions for the classification in three classes have been plotted. Neural networks have the tendency to be overconfident and produce prediction vectors that are very close to a one-hot-encoder, such as the blue one in the figure. Therefore, directly matching the teacher's output with the student's one might not be very efficient as the prediction might be similar to the ground-truth vector and, thus, not provide additional information. However, smoothing with the temperature parameter better reveals the relative importance of each class as shown in Figure 3.3 for the smoothened output.

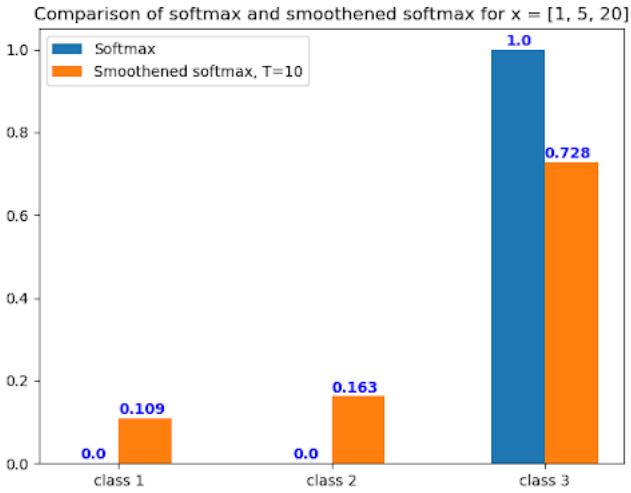


Figure 3.3: Effect of dividing the logits by a temperature parameter before the softmax function. The values have been rounded to the third decimal.

Subsequently, the knowledge distillation loss function consists in minimizing the cross-entropy loss function between the softened outputs of both models:

$$\mathcal{L}_{KD} = \sum_{i=1}^{N} CE[s_i, \sigma(f_S(I_k^i)/T)]. \tag{3.2}$$

This work was originally designed for classification. However, the loss function has been adapted for the compression of semantic segmentation models (Liu et al., 2019; Yang et al., 2022), and medical image segmentation models (Qin et al., 2021; Noothout et al., 2022) by computing the above loss independently for each voxel and summing it. It has also been proven to be efficient in the context of segmentation with privileged information (Chen et al., 2022; Rahimpour et al., 2022).

Despite its effectiveness, knowledge distillation only constrains the output probability distributions of a model. Other approaches built on the Teacher-Student framework to further constrain the intermediate feature maps between both models.

Attention Transfer

A spatial attention map aims at showcasing the location on which the model focuses to "make a decision". A common assumption is that the absolute value of a neuron expresses the importance of that neuron (Zagoruyko and Komodakis, 2017). For a given feature map F with C channels, the activation can be defined as:

$$A = \sum_{i=1}^{C} |F_i|^p$$

or

$$A = \max_{i=1,..,C} |F_i|^p$$

where $p \geq 1$. A visualization of the spatial attention map with $p = 1$ for a network trained with four input modalities and another one trained with one modality is presented in Figure 3.4. We can notice that the high values of the attention map are concentrated around the tumor location, validating the assumption made earlier. Furthermore, the model trained on multiple modalities has an even more concentrated attention map than the other model. This is namely noticeable in the background where the attention map of the unimodal net is more spread out.

As a result, Zagoruyko and Komodakis (2017) introduced an attention transfer loss function by minimizing a distance between the teacher's attention maps and the student's attention maps:

$$\mathcal{L}_{att} = \sum_{j \in \mathcal{I}} \left\| \frac{A_j^T}{||A_j^T||} - \frac{A_j^S}{||A_j^S||} \right\| \tag{3.3}$$

where $\mathcal{I}$ denotes the set of indices of the layers for which one wants to transfer the attention. This spatial attention transfer loss function can also be applied for the task of segmentation (Qin et al., 2021; Cho and Kang, 2022).

Computing the sum over the channel axis makes each channel contribute equally to the spatial attention map. Jang et al. (2019) proposed to learn custom weights to balance the contribution of each channel by feeding the features of the student to a small neural network. Building on this, Ji et al. (2021) used the channel attention, *i.e.* the sum or average along the image dimensions to predict the balancing weights. With a similar perspective, Kim et al. (2018) extract the spatial attention from an auxiliary neural network. During the training of the teacher, a three-layer CNN called paraphraser is trained to extract meaningful features. The attention of the student is pulled with another small CNN that is trained simultaneously. The attention transfer is still done by minimizing a distance between both attention maps.

Zagoruyko and Komodakis (2017) proposed a second attention transfer scheme based on gradients. Indeed, they simply defined the gradient attention of a layer as the gradient of the training loss function with respect to the layer. The transfer was performed by minimizing the $\mathcal{L}_2$ distance between the teacher

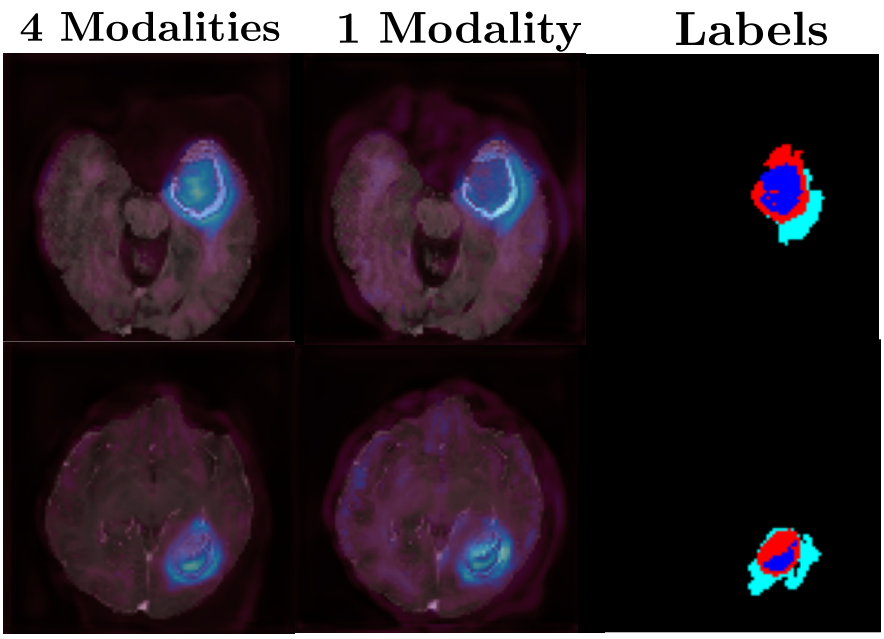


Figure 3.4: Spatial attention maps of multi-modal and mono-modal networks overlayed on the T1ce modality for two subjects. The last column presents the reference segmentation for each scan. The attention maps are computed as the sum of the features over the channel axis with $p = 1$.

and the student's gradients. This required first computing the forward and backward propagation of the training loss function and then, computing the attention error and back-propagate it a second time. The results indicated that gradient transfer performed worse than spatial attention transfer and Srinivas and Fleuret (2018) showed that combining both loss terms was useful only with small datasets. In their experiments, with 500 data points, it was better to use only the spatial attention transfer.

Contrastive Distillation

The idea of contrastive training is to learn a meaningful embedding by forcing representations to be close for similar pairs while pushing apart the representations for different pairs (Hadsell et al., 2006; Bachman et al., 2019; Chen et al., 2020; Tian et al., 2020b). In most contrastive learning frameworks, similar pairs are built by applying custom data augmentation functions such as random cropping, random color distortion, or random Gaussian blur. Any pair constituted with two transformations of the same subject is called a positive pair, otherwise it is called a negative pair. In the context of teacher-student learning, a positive pair is defined as the teacher and student representations of the same image, while a negative pair is composed of the representations from two different subjects (Tian et al., 2020a; Zhu et al., 2021; Chen et al., 2021).

Contrastive training requires choosing the working space in which the loss function is computed. The early methods directly operated in the representation space (Hadsell et al., 2006; Wu et al., 2018; Chen et al., 2021), however, recent strategies have found it beneficial to use a *projection head* g that maps the representation in a smaller space (Chen et al., 2020, 2022). Typically, g is a small feed-forward neural network with one hidden layer.

A widely used contrastive loss function, InfoNCE loss, maximizes a lower bound on the mutual information (Oord et al., 2018). In the context of teacher-student learning, the loss function can be written as:

$$\mathcal{L}_{InfoNCE} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{sim(g(F_i^S), g(F_i^T))}{\sum_{j\neq i} sim(g(F_j^S), g(F_i^T))}$$

where N is the batch size and sim is a similarity function, such as the exponential cosine similarity for instance. Minimizing the above loss function brings positive pairs together with the numerator while pushing all the negative pairs in the batch apart with the denominator.

For the segmentation of brain tumors with privileged information, we found the work of Chen et al. (2022) to be the only one that used contrastive distillation. Their training loss function was:

$$\mathcal{L}_{ct} = \sum_{i=1}^{N}\left[\|g(F_i^S) - g(F_i^T)\|_2^2 + \sum_{j\neq i}\max\{0,\ \xi - \|g(F_j^S) - g(F_i^T)\|_2\}^2\right] \tag{3.4}$$

where $\xi > 0$ is a distance margin. The margin makes the loss fuction ignore pairs for which the distance is already large enough. According to the authors, it helped the model to focus on harder pairs. They evaluated the method on BraTS 2018 and showed an improvement in the Dice score with respect to a baseline mono-modal network.

Adversarial loss functions

Finally, the last strategy consists in incorporating the representations in an adversarial framework to drive them to become indistinguishable (Shen et al., 2019; Chung et al., 2020; Liu et al., 2020; Vadacchino et al., 2021). This learning scheme works by training a discriminator network to determine whether a representation vector is generated by the teacher or the student model. In parallel, the goal of the student is to fool the discriminator by making its representation vectors resemble the feature map of the teacher.

Liu et al. (2020) transfered the knowledge for only one layer and, as a result, uses one discriminator. However, it is also possible to train several discriminators for several layers (Shen et al., 2019). Vadacchino et al. (2021) combined the two strategies by taking several intermediate feature maps from the decoders and feeding them into one discriminator (see Figure 3.5). The last method has been developed for brain tumor segmentation when three modalities (T1, T2,

and Flair) are available to the student. In this context, they show a significant improvement compared with the baseline.

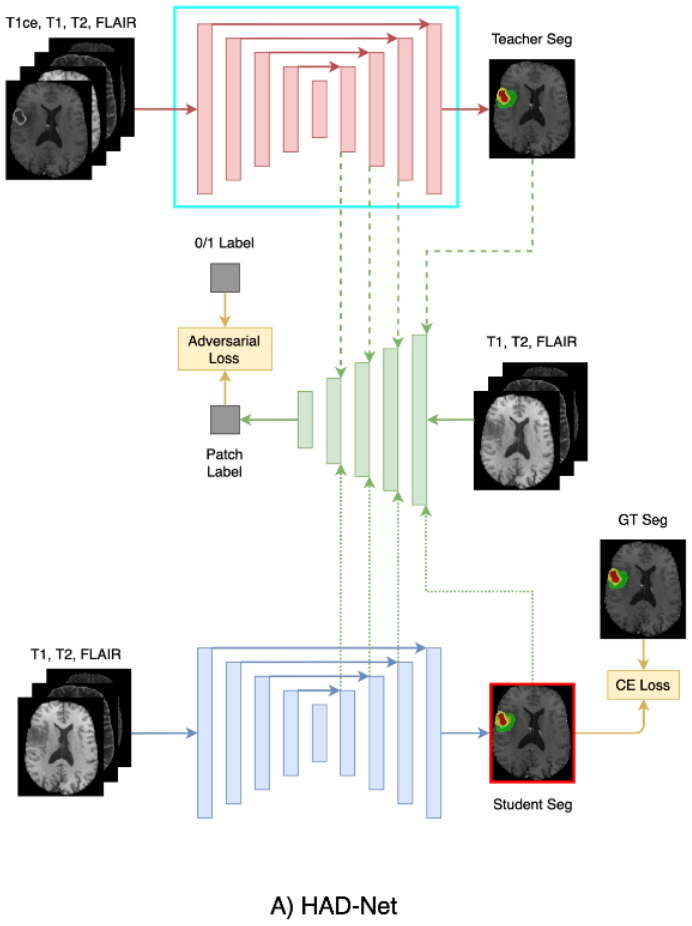


Figure 3.5: Illustration of the HAD-Net framework from Vadacchino et al. (2021).

In this section we described several strategies for transferring the knwoledge of a teacher model to a student one. However, they have been rarely applied in the context of brain tumor segmentation with missing modalities. Thus, in the next section, we propose a strategy mainly based on knowledge distillation adapted for this context. Then, we compare it with several methods presented above.

3.3 Weight sharing models

Our first strategy for transferring the knowledge of a multi-modal teacher network to a mono-modal one was based on the observation that both models have different inputs and that we would like them to return the same output. We speculated that the teacher and the student, having different inputs, should

also encode differently the information in the first layers, the ones related to low-level image properties, such as color, texture, and edges. By contrast, the deepest layers closer to the bottleneck, and related to higher-level properties, should be more similar. Therefore, in our original framework, the student and the teacher had two different encoders but they shared the same decoder.

3.3.1 First model

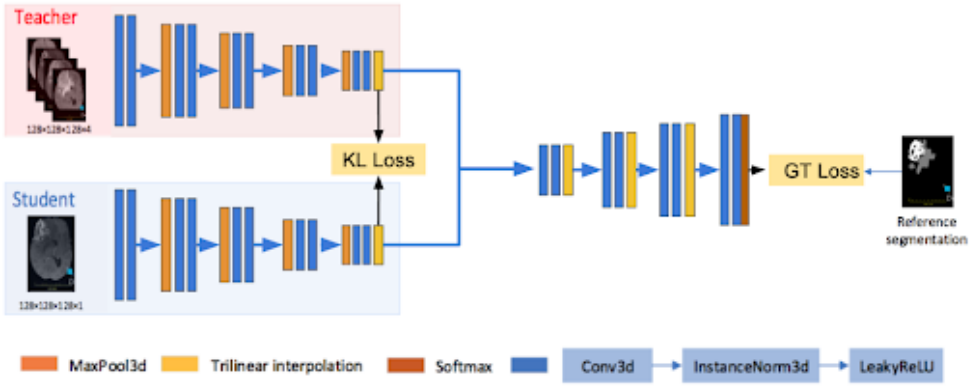


Figure 3.6: Teacher-student framework composed of two independent encoders and one shared decoder. The decoder takes either the representation of the student as input or the one of the teacher.

The architecture of the proposed framework is illustrated in Figure 3.6. The teacher network $f_T(\mathrm{I}^{\mathrm{i}})$ receives as input multiple modalities whereas the student network f_S only one modality I_k^i, k being the index of the chosen modality.

Loss functions

We proposed to force the student to learn from the additional information of the teacher encoded in its bottleneck (and partially in the deepest layers) by making their latent representations as close as possible. To this end, we apply the Kullback-Leibler (KL) divergence as a loss function between the teacher and student's bottleneck representations:

$$\mathcal{L}_{KL}(p,q) = \sum_{i=1}^{N} \sum_{j} q_i(j) \log\left(\frac{q_i(j)}{p_i(j)}\right). \tag{3.5}$$

where p_i (respectively q_i) are the flattened and normalized feature maps of the student (respectively teacher). Note that this function is not symmetric and we put the vectors in that order because we want the distribution of the student's bottleneck to be similar to the one of the teacher.

We add a second term to the objective function to make the predicted segmentation as close as possible to the reference segmentation. It is the sum of the Dice loss function ($Dice$) and the cross-entropy (CE). We call it L_{GT}:

$$L_{GT} = \sum_{i=1}^{N} \Big[(1 - Dice(y_i, \sigma(f_S(I_k^i)))) + CE(y_i, \sigma(f_S(I_k^i))) \Big]. \tag{3.6}$$

where y_i denotes the reference segmentation of the i^{th} sample in the dataset.

Results

We first trained the teacher, using only the reference segmentation as target. Then, we trained the student using the two different loss functions: the dissimilarity between the latent spaces and the reference segmentation loss function. Note that the weights of the teacher were frozen during the training of the student and the error of the student was not back-propagated to the teacher.

We trained the models on BraTS 2018 by doing a 3-fold cross-validation. The teacher received the four modalities (T1, T1ce, T2, Flair) while the student only received the T1ce modality. We compared the student model with a baseline network composed with the same encoder-decoder architecture as the student and only trained with the reference segmentation loss function.

Table 3.1: Average and standard-deviation Dice scores of our first proposed approach.

Model	ET	TC	WT
Baseline	42.02±2.34	67.17±2.09	65.9±1.01
Student	44.71±0.37	71.19±2.14	70.46±1.27

The results, presented in Table 3.1, showed a significant improvement of the Dice score for the three segmentation labels. However, they were also significantly worse than a nnU-Net model trained with the T1ce modality as input (see Table 3.2). We explain this situation with the absence of skip-connections in the student model. Indeed, skip-connections convey highly localized information, which is crucial for image segmentation. This is particularly visible for label ET, which is the smallest of the three labels. The information to properly segment it has therefore a higher chance of getting damaged through the down-sampling stages.

Nevertheless, the Kullback-Leibler loss function between the bottlenecks has proven to be effective at transferring the knowledge of the teacher network. Thus, we incorporated skip-connections in the framework.

3.3.2 Adding skip-connections

To feed the local information to the decoder, we added skip-connections to the previous framework. In this manner, the teacher and the student each had an architecture like nnU-Net (Isensee et al., 2019b), only the weights of their

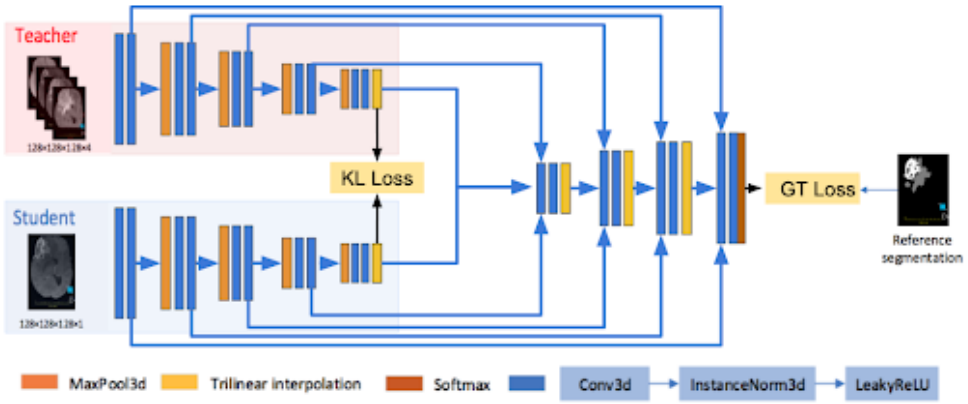


Figure 3.7: Teacher-student framework composed of two independent encoders and one shared decoder with skip-connections. The decoder takes either the representations of the student as input or the ones of the teacher.

decoders were the same. The models were trained in the same way as in the former framework. The results are shown in Table 3.2.

Table 3.2: Average and standard-deviation Dice scores for the teacher-student framework with shared decoder and skip-connections.

Model	ET	TC	WT
Baseline (nn-UNet)	68.1 ± 1.27	80.28 ± 2.44	77.06 ± 1.47
Student	67.56±3.66	80.45±1.35	75.8±0.87

Unfortunately, the scores of the student were still worse than the ones of the baseline, although they were much higher than in Table 3.1. We explain this with the nature of the information contained in the skip-connections. Namely, at the highest scale, the tensor passed through the skip-connection has undergone very few transformations. Thus, the tensor coming from the student and the one from the teacher contained very different pieces of knowledge. We believe that this knowledge gap made it difficult for the decoder to segment the mono-modal input.

In conclusion, we have seen earlier that skip-connections are absolutely necessary to produce results that are competitive with a baseline nnU-Net model. In addition, the skip-connections contain low-level features that are very different depending on the model (teacher of student). Since those features contain dissimilar information, they cannot be dealt with by the same weights. Therefore, in the next section we present our framework with two distinct teacher and student models (i.e., decoders).

3.4 KDNet

In this section, we describe the teacher-student framework that we presented during the MICCAI 2020 conference.

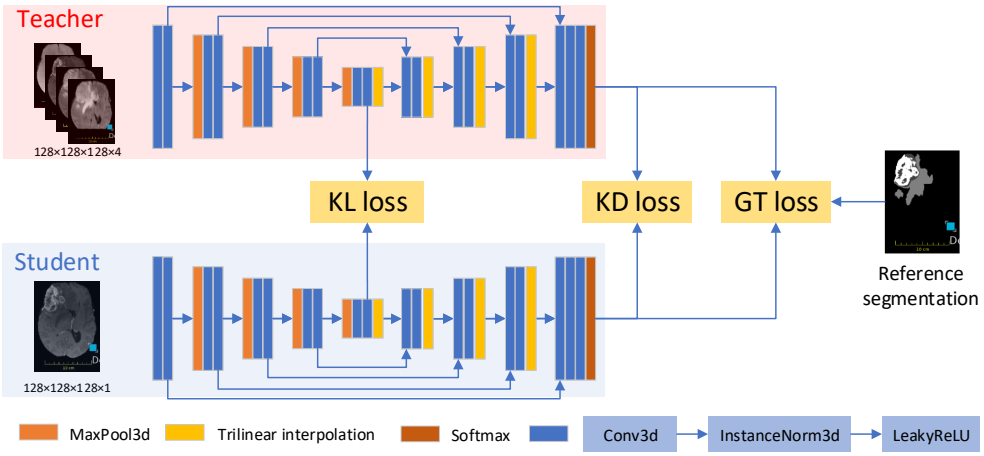


Figure 3.8: Illustration of the proposed framework. Both Teacher and Student have the same architecture adapted from nnUNet (Isensee et al., 2019b). First, the Teacher is trained using only the reference segmentation (GT loss). Then, the student network is trained using all proposed loss functions: KL loss, KD loss and GT loss.

3.4.1 Method

Except for the number of input channels, the teacher and the student networks have the same encoder-decoder architecture with skip connections (see Figure 3.8). We use the same loss functions than the two previous frameworks and we add a third term to further force the student representation vectors to be similar to the ones of the teacher.

Generalized knowledge distillation

Following the strategy of generalized knowledge distillation (Lopez-Paz et al., 2016), we transfer useful knowledge from the additional information of the teacher to the student using the soft label targets of the teacher. We follow the same loss term presented in Section 3.2.3 where the predictions of the teacher is soften with a temperature parameter T as in Equation 3.1. Subsequently, the knowledge distillation loss function in Equation 3.7 is adapted for the context of segmentation. It consists in computing the voxel-wise cross-entropy between the softened outputs of the teacher and the student:

$$\mathcal{L}_{KD} = \sum_{n} \sum_{x \in \Omega} CE[\sigma(f_T(I_k^n)/T)(x),\ \sigma(f_S(I_k^n)/T)(x)]. \tag{3.7}$$

where Ω is the domain of the image.

Objective function

The complete objective function is then the combination of the three loss terms:

$$L = \lambda L_{KD} + (1 - \lambda) L_{GT} + \alpha L_{KL} \tag{3.8}$$

Table 3.3: Comparison of three models using the Dice score on the tumor regions. Results of U-HVED and HeMIS are taken from (Dorent et al., 2019), where the standard deviations were not provided.

Model	ET	TC	WT
Teacher (4 modalities)	69.47 ± 1.86	80.77 ± 1.18	88.48 ± 0.79
Baseline	68.1 ± 1.27	80.28 ± 2.44	$\mathbf{77.06 \pm 1.47}$
U-HVED	65.5	66.7	62.4
HeMIS	60.8	58.5	58.5
Ours	$\mathbf{71.67 \pm 1.22}$	$\mathbf{81.45 \pm 1.25}$	76.98 ± 1.54

with $\lambda \in [0,1]$ and $\alpha \in \mathbb{R}^+$. The imitation parameter λ balances the influence of the reference segmentation with the one of the teacher's soft labels. The greater the λ value, the greater the influence of the teacher's soft labels. The α parameter is instead needed to balance the magnitude of the KL loss term with respect to the other two loss terms.

3.4.2 Experiments

Dataset

We evaluated the performance of the proposed framework on the publicly available dataset from the BraTS 2018 Challenge (Menze et al., 2015). This version of the dataset contains 285 patients. We applied a central crop of size $128 \times 128 \times 128$ and a random flip along each axis for data augmentation. For each modality, the values have been normalized by subtracting the mean and dividing by standard deviation for non-zero voxels.

Results

To demonstrate the effectiveness of the proposed framework, we first compared it to the baseline nnU-Net model. We also compared it to two other models, U-HVED (Dorent et al., 2019) and HeMIS (Havaei et al., 2016), using only T1ce as input. Results were directly taken from Dorent et al. (2019). The results are visible in Table 3.3. Our method outperforms U-HVED and HeMIS in the segmentation of all three tumor components.

Ablation study: To evaluate the contribution of each loss term, we did an ablation study by removing each term from the objective function defined in Equation 3.8. Table 3.4 shows the results. We observe that both the KL and KD loss functions improved the results with respect to the baseline model, especially for the enhanced tumor and tumor core.

Table 3.4: Ablation study of the loss terms. We compare the results of the model trained with three different objective functions: the baseline using only the GT loss term, KD-Net trained with only the KL term and KD-Net with the complete objective function.

Model	Loss	ET	TC	WT
Teacher	GT	69.47 ± 1.86	80.77 ± 1.18	88.48 ± 0.79
Baseline	GT	68.1 ± 1.27	80.28 ± 2.44	77.06 ± 1.47
Student	GT+KL	70.00 ± 1.51	80.85 ± 1.82	$\mathbf{77.08 \pm 1.29}$
Student	GT+KD	69.22 ± 1.19	80.54 ± 1.66	76.83 ± 1.36
Student	GT+KL+KD	$\mathbf{71.67 \pm 1.22}$	$\mathbf{81.45 \pm 1.25}$	76.98 ± 1.54

Networks similarity.

To verify whether the proposed framework made the student more similar to the teacher, we used singular vector canonical correlation analysis (SVCCA) (Raghu et al., 2017). SVCCA compares two representations by computing a singular value decomposition to get a subspace of each and then performing a cross correlation analysis on these subspaces. We applied this method for each layer, before an upsampling/downsampling stage, on the three following combinations: teacher-student, teacher-baseline, baseline-student. The results are plotted in Figure 3.9. They show that, for every layer, the teacher has a higher correlation with the student than with the baseline. This indicates that the framework manages not only to make the segmentation map more similar to the teacher's one but also for every intermediate representation too. Interestingly, one of the highest improvements in correlation is for the representation of the fourth skip-connection. This reinforces the idea that the Kullback-Leibler loss term helps to better train this layer.

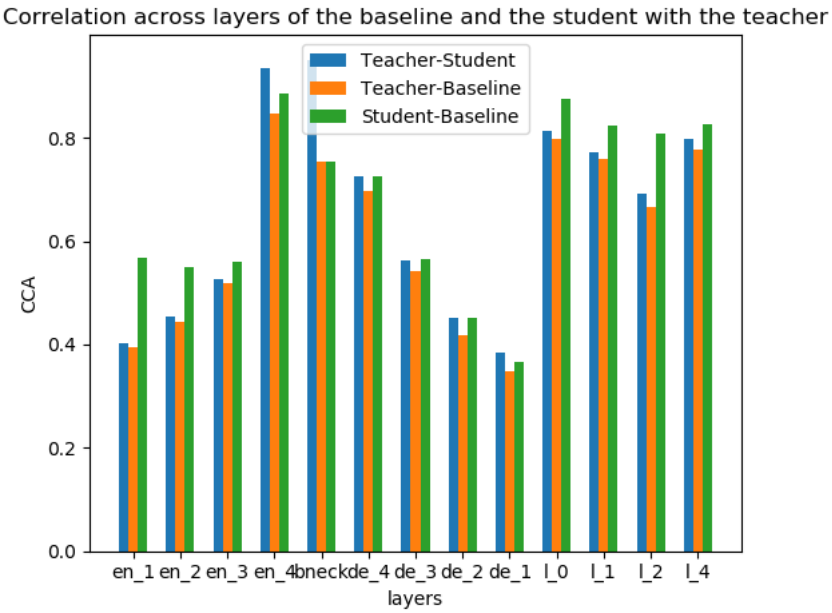


Figure 3.9: SV-CCA values across layers. The last four bars correspond to the representation of the four segmentation labels.

3.5 Effect of the training set size

Since this work has been presented in 2020, the number of subjects in BraTS dataset has significantly increased and it contains now 1251 images. When we applied our method on this dataset, we found that KDNet did not significantly improve the results compared with the baseline. Hence, in this section, we evaluate the results of KDNet with different numbers of training data. Additionally, we compare it with other knowledge-transfer strategies.

3.5.1 Model comparison

We further compare the proposed method with several knowledge transfer loss functions: attention transfer $\mathcal{L}_{Att}$ and contrastive distillation $\mathcal{L}_{CT}$. $\mathcal{L}_{Att}$ refers to the loss function in Equation 3.3, we use $p = 2$ to compute the spatial attention. We found that computing the attention loss function on the features from the bottleneck layer retrieves the best results.

For the contrastive distillation loss function, we use the one from Equation 3.4 defined by Chen et al. (2022) as it has already been applied for brain tumor segmentation. As in the paper, we apply the contrastive loss function in the second to last layer and map the representation vectors in a smaller space with a projection head. The latter consists of a global average pooling that reduces the size of the volume by two, followed by a 2-layer feed-forward network with a ReLU activation between both layers. We also execute it with the combination of the KD loss function and the contrastive loss term as in the work by Chen et al. (2022).

3.5.2 Datasets

We first evaluate the performance of the proposed framework on the 2018 version of the BraTS dataset. It contains MR scans from 285 patients. We apply a central crop of size 192×192×144 as it is the minimal crop that contains every brain and for which the dimensions are a multiple of 2^4. The second condition is necessary to down-sample the input images at least four times in the U-Net. Additionally, we run the models on the 2021 version of BraTS which contains scans for 1251 patients including the 285 ones from BraTS 2018. Moreover, we apply the z-score to normalize the images. For the data augmentation, we randomly mirror the image along an axis, shift the image intensities and apply gamma correction. As post-processing, we change the voxels labeled ET to TC if their total number is less than 20. We chose this threshold because the smallest size of non-zero ET region in BraTS 2021 is 22.

3.5.3 Evaluation protocol

Both datasets are randomly divided into training, validation, and test sets with respective ratios $\frac{2}{3}$, $\frac{1}{6}$, and $\frac{1}{6}$. To measure the robustness of the methods with the number of images, we randomly select one-fourth and one-half of each training set; thus creating four other training sets. Therefore we have 6 training sets with respectively 47, 95, 190, 208, 417, and 834 patients. Finally, to fairly compare all the models, they are tested on the same set which is constituted of 178 patients.

3.5.4 Implementation details

The optimizer is the same for every model, teacher, or student. We used the Adam optimizer for 1000 epochs with a learning rate equal to 0.0001 which is multiplied by 0.2 when the validation loss value has not decreased for 50 epochs. We use a weight decay of 10^{-5}, and a batch size of 4. Early stopping is applied if the value of the learning rate is smaller than 10^{-8}. All models were trained on an NVIDIA Tesla V100 GPU with 32 GB of VRAM.

3.5.5 Model Selection

The goal is to choose the number of down-sampling layers in the network. Down-sampling, in a CNN, has proven to be very efficient, namely, it helps reduce the memory use of the model. However, we want to avoid doing too much down-sampling, to the point where the feature's spatial resolution is so small that it does not offer any localization power. To that end, we train a U-Net model with the same architecture as nnU-Net (Isensee et al., 2019a). In particular, it has four down-sampling layers. For the complete architecture, it corresponds to the one of the student or the teacher in Figure 3.8.

Once the model is trained, to evaluate the usefulness of the low-resolution layers, we set their output feature maps to zero during inference on the validation set. The idea behind this is that if a layer is useful for the final decision,

zeroing it would deteriorate the final result. We select the output of the bottleneck, and the feature maps corresponding to the deepest skip-connection (SC4) and the one before (SC3). In Table 3.5 (respectively Table 3.6) we show the results on the validation set for a teacher model (respectively a student model). Zeroying the bottleneck layer had almost no impact on the Dice score for the three labels. It slightly decreases the results for ET and TC but also slightly increases the results for WT. None of these gaps are significant. Modifying the feature maps from SC4 has more impact on the end result, even if it is also moderate. Unsurprisingly the higher the resolution, the more effect it has on the segmentation map and this is visible with the results of SC3. Consequently, we decide to put 3 down-sampling layers in our U-Net architecture to avoid doing unnecessary computations. Our model has therefore the general teacher-student architecture presented in Figure 3.8 with the backbone model illustrated in Figure 3.10.

Table 3.5: Results of a teacher model on BraTS 2018 dataset. Zeroed layer indicates the name of the layer for which we set to zero all the feature maps at test time.

Zeroed Layer	ET	TC	WT
None	72.05	83.85	90.46
Bottleneck	72.01	83.82	90.51
sc4	71.23	83.22	89,43
sc3	69.14	77.71	84.94

Table 3.6: Results of a mono-modal model on BraTS 2018 dataset, trained with the T1ce modality. Zeroed layer indicates the name of the layer for which we set to zero all the feature maps at test time.

Zeroed Layer	ET	TC	WT
None	67.43	78.23	75.19
Bottleneck	67.25	77.95	75.13
sc4	67.56	76.98	72.91
sc3	60.54	59.24	64.85

Note that the results presented in Table 3.5 and Table 3.6 are not comparable with the other results presented later because the models have been trained and tested on different data splits.

3.5.6 Results

In our all experiments below, the teacher uses all four modalities (T1, T2, T1ce, and Flair concatenated) and the student uses only T1ce. It is interesting to notice that the T1ce modality highlights the enhancing tumor and the necrotic tumor core, therefore the baseline reaches comparable results with the teacher for ET and TC but has a significantly lower WT score. Thus, we are interested in improving the results for the WT label in priority. In this section, our framework KDNet corresponds to the model trained with the KD+KL loss.

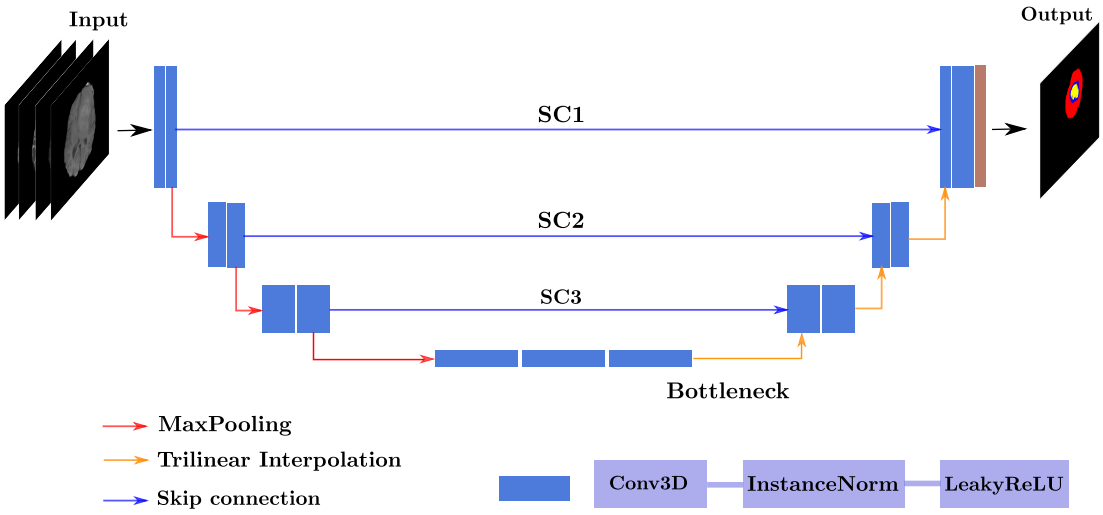


Figure 3.10: Architecture of the selected backbone model.

We present the Dice score and Hausdorff distance on the test set for BraTS 2018 in Table 3.7 and for BraTS 2021 in Table 3.8. The statistical significance of the differences with the baseline are evaluated with a paired t-test in both tables. Furthermore, for every method, we present the improvement of the Dice score for every label with respect to the baseline mono-modal network in Figure 3.11. The improvement is plotted against the number of images in the training set. When the number of training data is small, using a knowledge transfer loss function is very beneficial. In fact, when training on 47 subjects, apart from the student model trained with contrastive distillation, all methods significantly improve the Dice score and Hausdorff distance for the WT label. The Dice score for TC also significantly increases but not for ET which means that it better segments the necrotic tumor core only. For the three training set sizes of BraTS 2018, KDNet is the method that shows the most consistent results.

When the number of training data increases, the benefits of using a teacher-student framework are less clear. Namely, when using the 834 subjects for training, no student significantly increases the Dice score. For WT only KD generates an augmentation but only by a small margin. However, it does significantly improve the Hausdorff distance for that label. Therefore, using a knowledge transfer strategy would only be beneficial for a context where it is more important to minimize the Hausdorff distance.

In Figure 3.12, we present visual results of two patients for which the average Dice score of the student model has increased. Interestingly, CNNs with only T1ce as input always overestimate the size of the edema. It seems that the models "guess" the presence of the edema since it is hardly visible on the T1ce modality. With more training data, the "guess" appears more informed and the shape of the edema resembles more the one of the reference segmentation. For the first patient, the baseline produces very large and round edema,

Table 3.7: Dice score and Hausdorff distance for the models trained on the three training sets from BraTS 2018. Bold indicates the best score. The symbol $*$ (respectively †) indicates that the improvement (respectively deterioration) with respect to the baseline is statistically significant ($p < 0.05$).

Training set	Model	Dice			Hausdorff		
		ET	TC	WT	ET	TC	WT
N=190	Teacher	82.88	85.19	83.69	8.73	7.58	15.4
	Baseline	82.77	85.68	71.05	8.25	7.4	14.08
	Att	82.59	85.43	73.68*	7.84	7.31	13.48
	KL	**83.32**	86.38	**74.86***	7.52	6.69*	**12.28***
	KL + KD	82.48	86.42	74.19*	7.76	6.69*	12.95*
	KD	83.15	**87.2***	73.14*	**7.29***	**6.34***	13.19*
	CT	83.2	85.47	68.51†	9.2†	8.33†	13.52
	CT + KD	82.66	85.27	73.39*	9.05	8.15†	13.87
N=95	Teacher	79.92	83.83	82.14	12.46	9.69	14.06
	Baseline	78.8	82.6	70.0	10.89	9.27	13.8
	Att	79.53	**84.35***	72.15*	9.2*	**7.74***	**12.26***
	KL	79.37	83.75*	70.42	**8.41***	7.94*	12.46*
	KL + KD	**79.75***	82.2	**73.24***	9.57*	9.55	13.28
	KD	79.0	82.76	72.21*	9.89*	9.22	13.73
	CT	**79.75***	82.21	70.09	10.45	9.6	12.96*
	CT + KD	79.19	82.59	70.32	9.52*	9.25	13.66
N=47	Teacher	76.08	77.66	77.25	13.9	14.43	13.88
	Baseline	75.36	75.98	64.35	15.96	15.56	16.73
	Att	74.66	78.6*	66.49*	12.62*	**11.3***	**13.81***
	KL	74.22	76.72	66.88*	12.88*	11.38*	15.51*
	KL + KD	75.69	78.25*	68.09*	12.89*	11.49*	14.16*
	KD	**76.44**	**78.76***	**70.04***	13.06*	11.65*	14.65*
	CT	71.74†	74.83	64.15	15.36	14.52	15.95
	CT + KD	73.1†	76.35	67.52*	**12.24***	11.68*	14.24*

and the student models can produce a smaller region. For the second patient, the baseline trained with little data is not able to detect the presence of the tumor while the students model do. With more training data, all the models better segment the tumor. In Figure 3.13, the student models show worse segmentations than the baseline. Overall, it is still striking that most models overestimate the size of the real tumor. This is especially visible on the second patient where the tumor core is very large for most student models.

3.5.7 Discussion

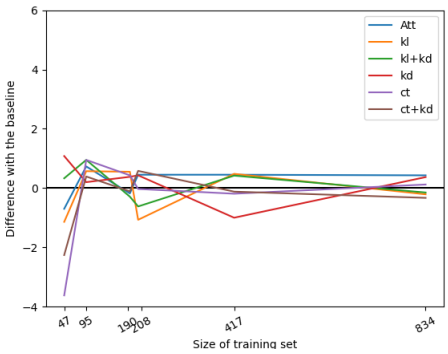
The results indicate that, in the context of missing modalities, the teacher-student framework is beneficial only when little data is available. Figure 3.14 shows the distribution in the test set of the improvements for the KD + KL model trained with 47 and 834 subjects. Namely, Figure 3.14a is very asym-

Table 3.8: Dice score and Hausdorff distance for the models trained on the three training sets from BraTS 2021. Bold indicates the best score. The symbol $*$ (respectively †) indicates that the improvement (respectively deterioration) with respect to the baseline is statistically significant ($p < 0.05$).

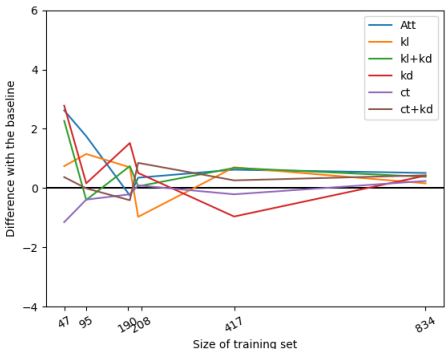
Training set	Model	Dice			Hausdorff		
		ET	TC	WT	ET	TC	WT
N=834	Teacher	87.61	89.95	91.44	6.05	5.63	9.27
	Baseline	86.96	89.68	77.86	6.9	6.1	13.51
	Att	**87.39**	**90.19**	77.77	**6.18**	5.49	**11.48***
	KL	86.75	89.84	77.43	7.33	5.75	11.51*
	KL + KD	86.81	90.06	77.58	6.48	**5.37***	12.89
	KD	87.33	90.1	**77.93**	6.26	5.76	12.28*
	CT	87.08	89.92	76.3†	6.63	5.57	12.59*
	CT + KD	86.63	90.11	76.37†	6.86	5.89	13.42
N=417	Teacher	86.9	89.31	90.5	7.87	7.16	11.54
	Baseline	86.02	88.8	76.77	7.13	6.42	12.79
	Att	86.47	89.42	**77.68***	7.11	6.39	**12.66**
	KL	**86.5**	**89.5**	77.29	7.03	**6.31**	13.9†
	KL + KD	86.44	89.48	75.85†	**6.72**	6.74	12.75
	KD	85.02†	87.84†	75.74†	8.21†	8.31†	13.87†
	CT	85.83	88.59	75.44†	6.8	6.37	12.98
	CT + KD	85.9	89.06	75.23†	7.6	7.3†	13.47†
N=208	Teacher	84.85	86.88	89.23	9.06	7.9	11.1
	Baseline	84.67	87.45	74.1	8.25	7.64	13.3
	Att	85.12	87.8	**77.09***	8.14	6.6*	12.76
	KL	83.6†	86.48†	74.77	8.8	7.17	**12.57***
	KL + KD	84.05	87.51	76.41*	8.78	6.74*	13.6
	KD	85.1	87.96	75.88*	7.67	7.1	13.25
	CT	84.64	87.54	76.17*	8.15	6.94*	12.96
	CT + KD	**85.25**	**88.3**	76.78*	**7.59**	**6.29***	12.78

metrical, the Dice score is improved for more than 75% and the amplitude of the increase is also higher. A similar distribution occurs for the Hausdorff distance in Figure 3.14c but it is less striking. For the model trained on 834, for both the Dice score and the Hausdorff distance, the distribution is very symmetrical showing that there is no benefit of using our method in this case. These results are very similar to the one of Srinivas and Fleuret (2018) where they found that using gradient-based attention transfer was only beneficial on small training sets.

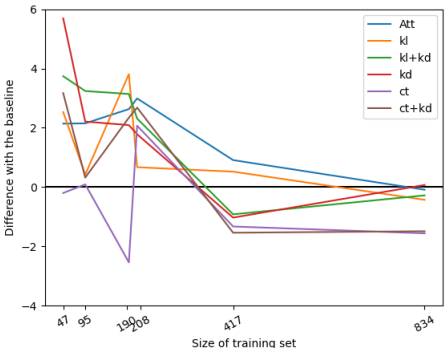
Additionally, we present the improvement/deterioration in Dice score sorted by the size of each tumor label in Figure 3.15. Depending on the tumor part and the model, the size of the tumor seems to have an effect on the improvement. Namely, for the models trained with 47 subjects, KDNet better segments the whole tumor when its size is smaller. Similarly, this occurs also for the en-



(a) Enhancing Tumor



(b) Tumor Core



(c) Whole Tumor

Figure 3.11: Evolution of the improvement of the average Dice score on the test set for every method with respect to the baseline depending on the size of the training set.

hancing tumor. With the exception of the first subject, which appears as an outlier, most of the improvement in the training set occurs for the smallest tumors. For the tumor core, there do not seem to be any effect linked with the tumor size, positively or negatively. By contrast, for the model trained on 834 data points, the tumor size has very little effect on the improvement for the three labels. However, it is interesting to notice that the variability in the results is larger for subjects with small tumors. This can be explained with the fact that a few mislabeled voxels change the Dice score by a larger value for a small tumor than for a bigger one. From this analysis, it seems that in a low-data setting, the knowledge transfer loss functions help the student to segment smaller structures. This is in line with the visual results presented in Figure 3.12 where the student models were able to detect the presence of a small tumor when the baseline could not.

Furthermore, the quality of the segmentation is improved for the structures that are less visible on the available modality. For instance, in this context, the edema is better segmented while it is less visible on the T1ce modality. This indicates that the teacher model does distill additional knowledge into the student. However, this additional knowledge is not necessary when more

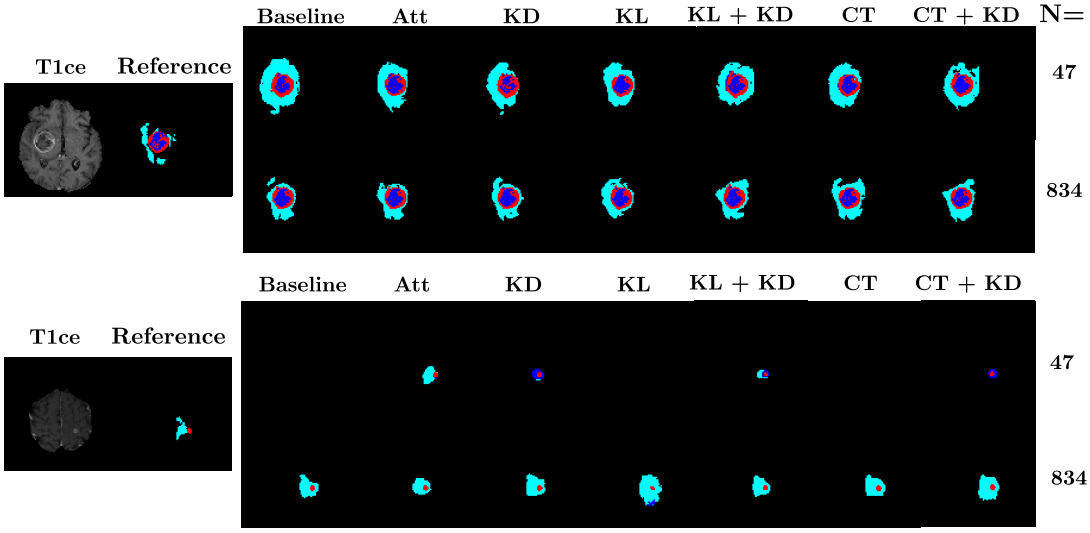


Figure 3.12: Qualitative results for the baseline and the models trained with Att, KD, KL, KL+KD, CT and CT + KD loss functions on the smallest and largest training sets. The red region corresponds to the enhancing tumor, the blue region to the necrotic tumor core, and the sky blue region to the edema. For both subjects, the Dice score of KDNet trained on 47 images has improved w.r.t. the baseline.

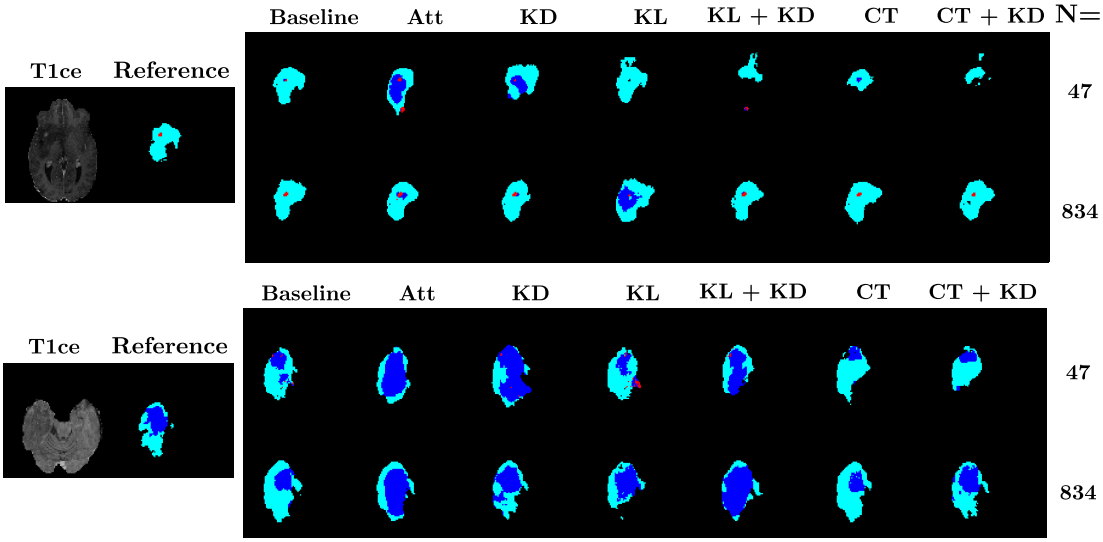
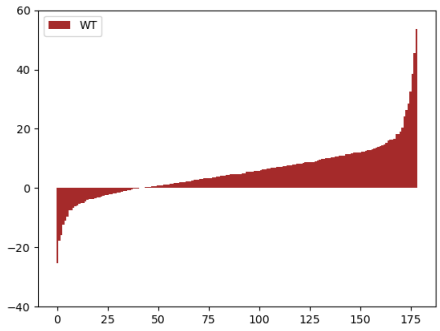
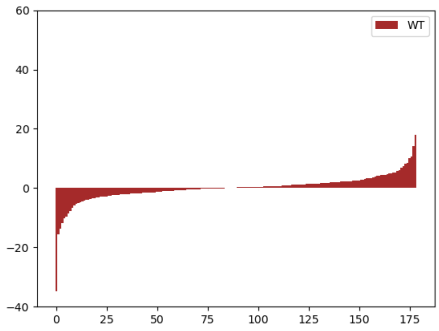


Figure 3.13: Qualitative results for the baseline and the models trained with Att, KD, KL, KL+KD, CT and CT + KD loss functions on the smallest and largest training sets. The red region corresponds to the enhancing tumor, the blue region to the necrotic tumor core, and the sky blue region to the edema. For both subjects, the Dice score of KDNet trained on 47 images has deteriorated w.r.t. the baseline.

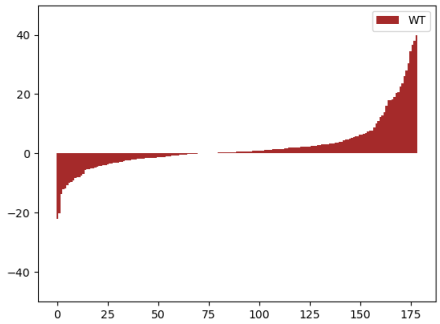
training images are available. Potentially, this could come from the fact that with more data, the student encounters more subjects for which the edema is



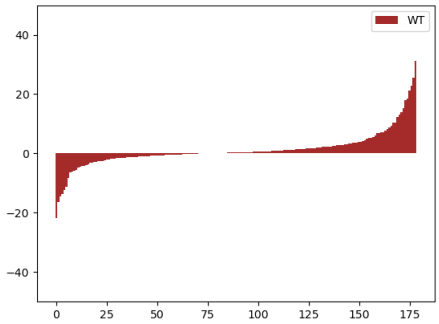
(a) Dice improvement: model trained on 47 subjects



(b) Dice improvement: model trained on 834 subjects



(c) Hausdorff improvement: model trained on 47 subjects



(d) Hausdorff improvement: model trained on 834 subjects

Figure 3.14: Distribution of the improvement with respect to the baseline of the Dice score and the Hausdorff distance for every subject in the test set for the model trained with the KD+KL loss function. The results on the WT label are presented for the model trained with the least data and the one trained with the most.

better visible on the T1ce modality. The model could be more able to infer knowledge about the edema from these images and therefore, not require the teacher's supervision.

The only loss term that systematically generates worse segmentations or at best equivalent to the baseline is the CT. Our hypothesis is that the contrastive loss function, as it is used here, is not adapted for brain tumors. Indeed, the loss function pushes representation vectors to be dissimilar in the same manner for every pair of patients. This is not adequate in this context because the representations for two subjects with the tumor in the same brain region should be more similar than for two subjects where the tumor is in different locations. A potential solution could be to compute weights depending on the relative positions of the tumors in the brain and incorporate them in the contrastive loss function. A similar strategy has been incorporated by Dufumier et al. (2021)

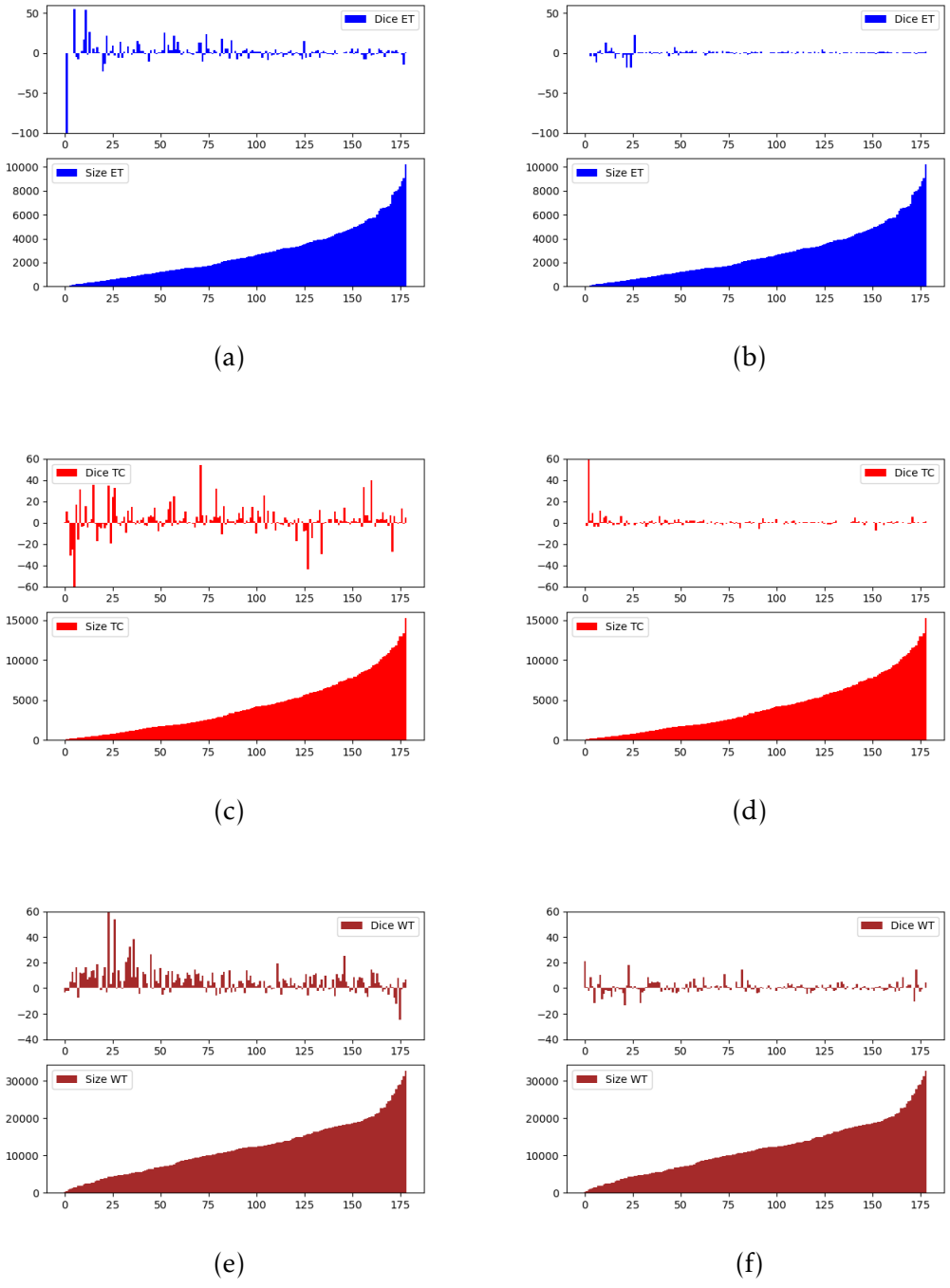


Figure 3.15: Improvement (or deterioration) of the Dice score of the three tumors labels (ET first row, TC second row and WT third row) with respect to the baseline when using KDNet. The results are sorted by the size of the label (i.e., size of the tumor part). For each tumor label, the improvement/deterioration is in the top row while the tumor size is in the bottom row. The x-axis represents the index of the test subject. Figures (a), (c) and (e) show the model trained with 47 subjects. Figures (b), (d) and (f) the one trained with 834 subjects.

where weak labels are used to weigh the contrastive loss function.

Finally, we would like to point out that we also used the adversarial loss from Vadacchino et al. (2021). We first experimented with their code by only feeding the T1ce modality to the student - so with a different backbone neural network from ours. For several hyperparameters, we found that the results have systematically deteriorated when using the discriminator. We raise two potential explanations for this. First, the training of adversarial models is often very unstable and we speculate that our attempts were subject to this issue. The other explanation is that the loss function was originally meant with three input modalities for the student. In this case, the inputs of the student and the teacher differ by only one modality. In our case, the inputs are very different which makes it easier for the discriminator to distinguish between the feature maps. A too strong discriminator leads to a weaker student. For this reason, we did not incorporate the adversarial loss function in our framework.

3.6 Conclusion

In this chapter, we proposed a method to distill the knowledge of a multi-modal teacher network into a mono-modal student network for the segmentation of brain tumors. We compared it with several existing strategies.We showed that using those strategies is only beneficial when the number of training images is small. In the different low data settings that we experimented, our method was consistently among the best ones. However, when using more images (around 250 subjects), all the tested loss functions did not improve the quality of the segmentation.

In this thesis, we require the brain tumor segmentation map as it is a key element in the registration strategy that we present in the following chapters. Since we are working with only one modality, we were interested in training a model that retrieves the best segmentation possible. Therefore, in our case, it is better to train a mono-modal U-Net on BraTS 2021 without any teacher supervision. Nevertheless, using our method can be beneficial for other applications where there is little annotated images such as myocardial pathology segmentation (Li et al., 2022).

4
Image Registration

Contents

In this chapter, we present usefull knowledge about image registration for the understanding of chapter 5. Image registration is the process of aligning two images so that they can be compared. It consists in changing the position of the pixels or voxels, but not their intensity values. This is done with a deformation function which operates as a change of coordinate in the image defined as a function $I : \Omega \to \mathbb{R}$, where $\Omega \subset \mathbb{R}^d$ and d is the dimension of the image. Let Ω_t be the coordinate system in which one wants to map the source image and J be the target image defined on Ω_t. The goal of image registration is to find the mapping $\phi : \Omega \to \Omega_t$ so that for all $x \in \Omega_t$:

$$J(x) \approx I(\phi^{-1}(x))$$

where $\approx$ denotes a similarity function between two images, discussed next. The reason for using the inverse of the transformation as the change of coordinates is that, computationally, image transformation typically does a "pull-back" *i.e.* each voxel in the deformed image is filled by transforming its position with the inverse function and interpolating the neighboring values of the source image. This is opposed to a "push-forward" method where voxels are moved from the source image to a new position.

Numerically, the images stored in the computer are not continuous, they are discrete. The value of a voxel can only be prompted only if its coordinates are integers. Since the values taken by the transformation function are unlikely

to be integers, it is required to interpolate the neighboring values to get the new value (see Figure 5.3). The interpolation function, therefore, has a significant impact on the result of the registration. A simple approach is to get the value of the nearest voxel. Although very simple and fast to implement, this interpolation can “pixelize" the output. Another approach, called bi-linear in 2D, consists in linearly combining the 4 neighboring pixels depending on their distance with $\phi^{-1}(x)$. This produces less “pixelized" outputs but if repetitively applied to the image, it can blur the image. More interpolation functions are presented by Ashburner and Friston (2007).

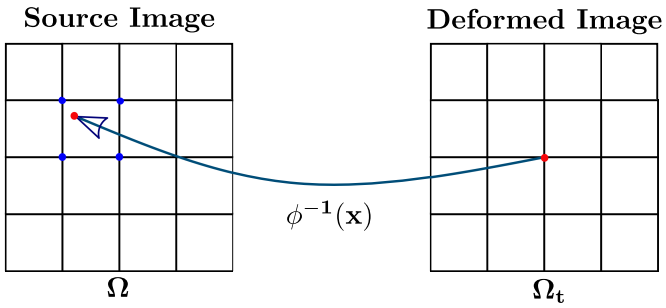


Figure 4.1: Image deformation process. The deformed image is filled by iterating through each pixel value x, computing the corresponding position in the source image $\phi^{-1}(x)$, and interpolating the neighboring values to get the new value in x.

Measuring the similarity between two images is very important to evaluate the quality of the registration. A common technique is to use additional information about the image like segmentation maps or key points (landmarks) and measure how they match. However, the process of collecting those annotations is very time-consuming and is still subject to human errors. In this thesis, we are interested in intensity-based image registration which directly evaluates the alignment with the intensity values of both images.

4.1 Intensity-based Registration

This section presents common similarity functions used during intensity-based registration, for more detail see (Hill et al., 2001). A straightforward and easy-to-implement similarity function is the sum-of-squared distance (SSD):

$$SSD(I,J) = \sum_{x \in \Omega_t} |I(x) - J(x)|^2$$

This also corresponds to the squared $\mathcal{L}_2$-distance. Naturally, the measure is equal to zero if and only if both images are the same. Since the values of both images are directly compared, it is necessary to scale them to be in the same range. Furthermore, this makes the method sensitive to outliers which is common in MRI.

Similarly, we can compute the ratios rather than compute the difference between voxel values. In this case, the goal is to minimize the variance of the ratio:

$$VIR(I,J) = \frac{\sqrt{\frac{1}{|\Omega_t|}\sum_{x\in\Omega_t}|R(x)-\overline{R}|^2}}{\overline{R}}$$

where $R = \frac{I}{J}$ and $\overline{R}$ is the average ratio.

Another common approach is to evaluate if there is a linear relation between the images with the cross-correlation:

$$CC(I,J) = \frac{\sum_{x\in\Omega_t}(I(x)-\overline{I})(J(x)-\overline{J})}{\sqrt{(\sum_{x\in\Omega_t}(I(x)-\overline{I}))^2(\sum_{x\in\Omega_t}(J(x)-\overline{J}))^2}}$$

This can also be done with the local cross-correlation which is the sum of the correlation in a window around every voxel. Let W_x be a window centered around the voxel $x \in \Omega_t$, I_{W_x} and J_{W_x} be the images I and J in the window W_x, the local cross-correlation is:

$$LCC(I,J) = \sum_{x\in\Omega_t} CC(I_{W_x}, J_{W_x})$$

The previous approaches are well-suited for the registration of images from the same modality. Another approach, based on joint histograms, allows for the registration of different modalities. A joint histogram measures the number of occurrences of every possible pair of voxel values (one in each image). For instance, the joint histogram of I and J at position (u,v) measures the number of voxels for which the value in I is equal to u and the value in J is equal to v. By dividing by the number of voxels, we get a joint probability function $p_{I,J}$. The mutual information between two random variables measures how much information is present in one random variable about another. In image registration, if I and J are considered random variables, ideally, we would like the mutual information to be the highest possible. Indeed, given the value at every voxel of image I, we would be able to predict the value of image J which is only possible if the two images are aligned. The mutual information is measured by:

$$M(I,J) = \sum_{u,v} p_{I,J}(u,v)\log\frac{p_{I,J}(u,v)}{p_I(u)p_J(v)}$$

where p_I and p_J are the probability densities computed by normalizing the histograms of I and J respectively. In the context of image registration, we want to maximize the mutual information between the target and the warped image.

Once the similarity function S is chosen, the optimal transformation is usually computed by minimizing a functional of the form:

$$\phi = \arg\min_{\phi} S(I\circ\phi^{-1}, J) + R(\phi)$$

where R is a regularization function.

4.2 Affine Registration

Registration can be categorized into two main categories: linear and deformable (non linear) registration. The first one encodes global transformations of the image while the second one can encapsulate global and localized deformations. For that reason, we are more interested in deformable registration since they are better suited for dealing with the natural shape difference between brains. However, affine registration is often used as a preprocessing step to ease the computations of the non-linear methods. Hence, we first briefly present affine registration.

We present all the functions in 2D for coordinates $(x, y) \in \mathbb{R}^2$ but it is easily extendable to 3D. With affine registration, each coordinate is deformed by the same affine function:

$$\phi(x, y) = \mathbf{A}\mathbf{x} + \mathbf{T} = \begin{bmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_0 \\ t_1 \end{bmatrix}.$$

This transformation is composed of a combination of translation, rotation, scaling, reflection, and shear. The matrix $\mathbf{T}$ determines the translation of the image. Conveniently, the matrix $\mathbf{A}$ can be decomposed into several matrices where each one controls an aspect of the deformation.

With these deformations, parallel lines remain parallel and ratios between components are preserved.

4.3 Classical Non-Linear Methods

4.3.1 Displacement Field

With an affine warping, the transformation is global because the parameters are the same for every voxel. For that reason, non-linear strategies determine voxel-specific parameters. A common method is to compute a dense displacement field u that, to every voxel x, associates the target position:

$$\phi(x) = x + u(x)$$

Computing the optimal deformation is a very cumbersome task since the dimension of the optimization space is much bigger than for affine registration - there are d parameters for every voxel of the image. Additionally, such transformations need to be heavily constrained to generate realistic deformations *i.e.* the distorted image must have the same topology. The goal is therefore to produce smooth displacement fields. The intuition behind it is that a smooth displacement field keeps neighboring voxels as neighbors, hence neighboring structures stay connected. The smoothness property of a deformation

$\phi = (\phi_1, ..., \phi_d)$ is measured with its Jacobian:

$$J_\phi(x) = \begin{bmatrix} \frac{\partial \phi_1}{\partial x_1}(x) & ... & \frac{\partial \phi_1}{\partial x_d}(x) \\ \vdots & & \vdots \\ \frac{\partial \phi_d}{\partial x_1}(x) & ... & \frac{\partial \phi_d}{\partial x_d}(x) \end{bmatrix}$$

The deformation is considered smooth if the determinant of the Jacobian is strictly positive for all $x \in \Omega_t$.

Early approaches constrained the smoothness by generating the displacement fields from laws of the kinematics of elastic solids (Bajcsy and Kovačič, 1989; Miller MI, 1993) and viscous-fluids (Christensen et al., 1994, 1996; Christos, 1997). However, in these methods the displacement field is computed for every voxel which is very time-consuming. Other approaches find the optimal transformation on control points and determine the warping on the rest of the image by interpolation. For N control points $(\psi_n)_{n\in[|1,N|]}$, the deformation at $x \in \Omega_t$ is retrieved with:

$$u(x) = \sum_{i=1}^{N} f(x,i)\psi_n$$

where f is the interpolation function. A wide variety of interpolation functions have been used, including linear interpolation (Kjems et al., 1999), radial basis functions (Fornefett et al., 1999, 2001; Shusharina and Sharp, 2012) and B-splines (Rueckert et al., 1999; Schnabel et al., 2001; Kybic and Unser, 2003). The control points usually are chosen in a uniform grid of points but Schnabel et al. (2001) extended it to non-uniform grids. The control points can also be a set of manually pre-selected landmarks (Fornefett et al., 1999, 2001; Shusharina and Sharp, 2012). The number of control points is important to limit the computational complexity and enforce the smoothness of the displacement field. Users have to find the optimal number of points to correctly align the images and have a smooth warping. This method can also be used to reduce the computation time by starting with a low-resolution grid and progressively increasing the resolution of the grid. The low-resolution displacement field is therefore used as an initializer for the higher resolutions.

Another set of methods, called Demons, uses optical flow equations to generate the displacement (Thirion, 1998; Pennec et al., 1999). A vector field is recursively computed, until convergence, as:

$$v(x) = \frac{[I(x+u(x)) - J(x)]\nabla J(x)}{||J(x)||^2 + ||J - I(x+u(x))||^2} \tag{4.1}$$

and it is smoothed using a Gaussian kernel K to remove the high frequencies and get the displacement field:

$$u = K(v)$$

With the displacement field strategy, computing the inverse of the deformation is complex. When the deformation is small, the approximation $u(x - u(x)) \simeq u(x)$ is always true. Consequently, the inverse of ϕ can be estimated with

$$\phi^{-1}(x) \simeq x - u(x).$$

However, as shown in Figure 4.2 if the deformation is not small, this estimation produces a very inaccurate inverse. To solve this issue, several methods optimize the deformation directly in the space of diffeomorphisms.

4.3.2 Diffeomorphic Registration

A function is a diffeomorphism if it is bijective, differentiable and its inverse is also differentiable. Building diffeomorphic warpings is very convenient for image registration since it ensures to get a one-to-one correspondence between images.

The deformation is usually of the form:

$$\phi(x) = (x + v_1(x)) \circ (x + v_2(x)) \circ ... \circ (x + v_T(x)).$$

The intuition behind it is that for all $t \in [|1, T|]$, v_t is a small displacement field, hence, $Id + v_t$ is considered as a diffeomorphism. Yet, the composition of diffeomorphisms is a diffeomorphism. Thus, ϕ is considered a diffeomorphism. Additionally, the inverse can be approximated with:

$$\phi^{-1}(x) \simeq (x - v_T(x)) \circ (x - v_{T-1}(x)) \circ ... \circ (x - v_1(x)).$$

Figure 4.2 shows that this framework allows performing large deformations and to retrieve their inverse. Rather than computing only one displacement field, diffeomorphic registration requires computing a set of displacement fields which are usually called velocity fields.

Large Deformation Diffeomorphic Metric Mapping (LDDMM) (Dupuis et al., 1998; Beg et al., 2005; Miller et al., 2006; Ashburner and Friston, 2011; Vialard et al., 2011) is a very elegant framework that produces diffeomorphic transformations. The deformations are considered time-dependent where $\phi(x, t = 0)$ is the initial position of x and $\phi(x, t = 1)$ its final position. For simplification, $\phi(x, t)$ is written as $\phi_t(x)$. Let $v(\cdot, t) = v_t : \Omega \to \mathbb{R}^d$ be a time-dependent vector field. At each time t, the velocity of the position $\phi(\cdot, t)$, which is the time derivative, is constrained to be equal to the vector field v_t (see Figure 4.3). Hence, we get the flow equation that parametrizes the whole transformation:

$$\forall x \in \Omega, \ \frac{\partial \phi_t(x)}{\partial t} = v_t(\phi_t(x)) \quad \phi_0(x) = Id(x) = x \tag{4.2}$$

The vector fields $(v(\cdot, t))_{t\in[0,1]}$ (also called velocity fields) are the parameters of the transformation since the endpoint of the transformation can be retrieved by integration of the flow equation:

$$\phi_1(x) = \phi_0(x) + \int_0^1 v_t(\phi_t(x))dt.$$

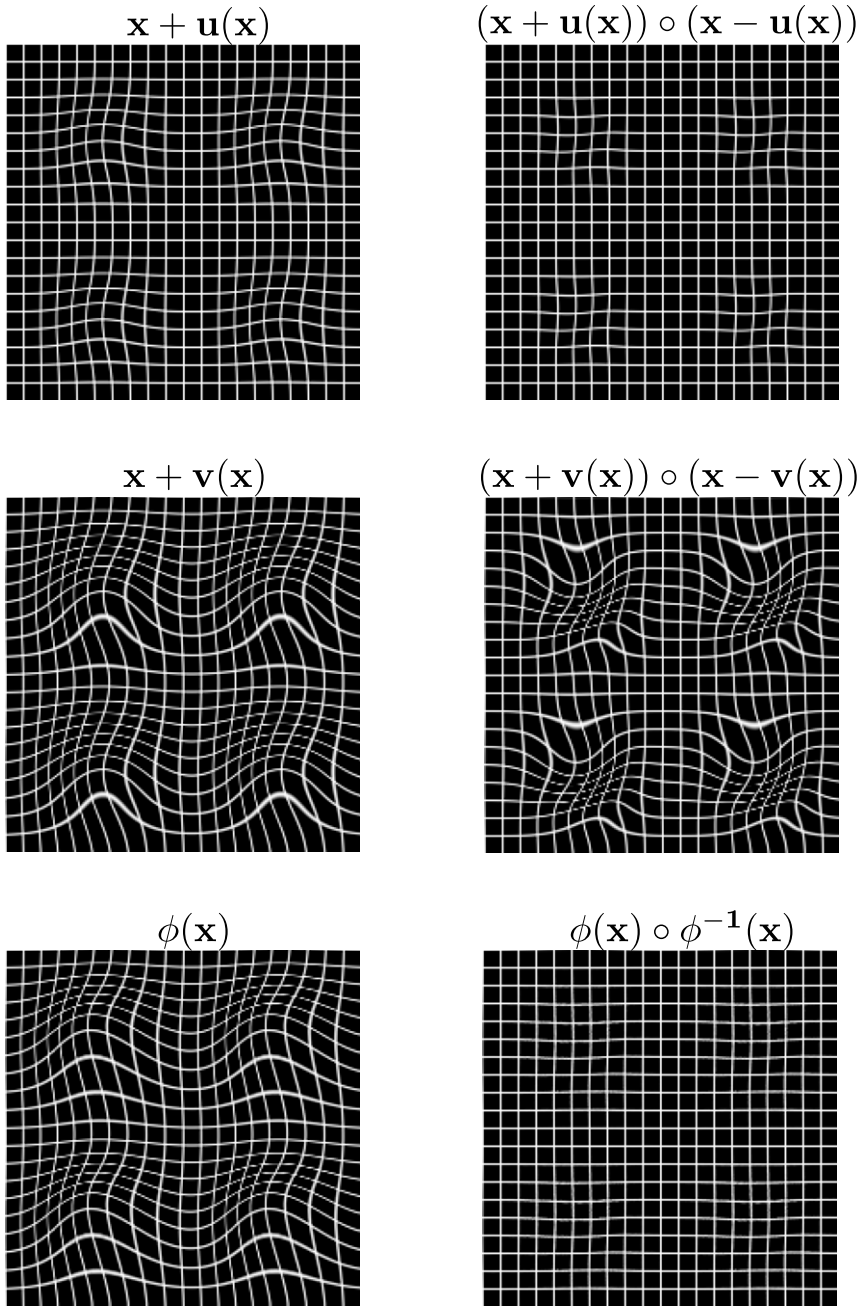


Figure 4.2: Illustration of the problem of estimating the inverse in the displacement field framework. The first row shows the warping of a grid by a small displacement field u and its composition with the estimated inverse $Id - u$. The return to identity is correct thus validating the choice of $x - u(x)$ as the estimated inverse. In the second row, with a bigger displacement field v, the return to identity is not satisfied. In this case, there is no easy method to estimate the inverse deformation. The third row illustrates the deformation and estimated inverse with a diffeomorphic transformation ϕ. ϕ is computed by deforming 15 times the grid with $x + \frac{v(x)}{15}$ and the inverse by warping it 15 times with $x - \frac{v(x)}{15}$.

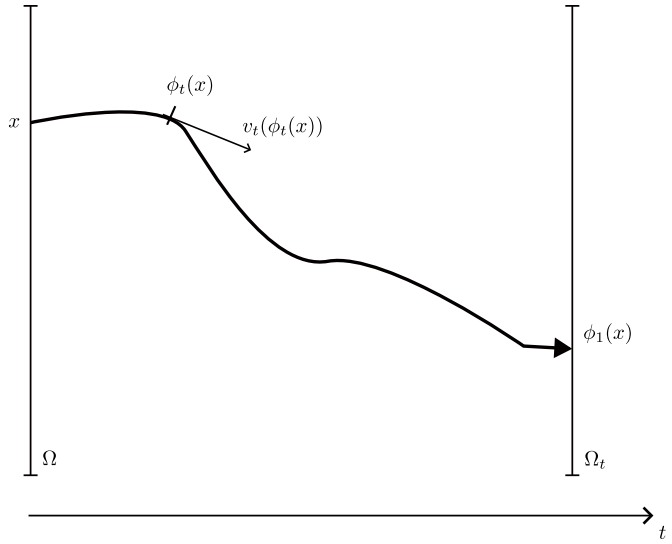


Figure 4.3: Transport of the voxel x in Ω to the position $\phi_1(x)$ in Ω_t. At each time-step t, the movement of the voxel is parametrized by the velocity-field $v_t(\phi_t(x))$.

Therefore, the choice of the velocity field needs to be done carefully to reach the correct alignment and get a smooth deformation. We call V the space of velocity fields from which we sample $v = (v(\cdot, t))_{t\in[0,1]}$. Dupuis et al. (1998) have shown that if all elements $v_t \in V$ are sufficiently smooth, then the solution of Equation 4.2 is in the space of diffeomorphisms. In the literature, V is usually modeled as a Reproducing Kernel Hilbert Space (RKHS) with kernel K. The space V is equipped with the norm $\|v_t\|_V = \sqrt{\langle v_t, Lv_t\rangle}$ with $L = K^{-1}$ and $\langle\cdot,\cdot\rangle$ being the scalar product in $\mathbb{R}^d$. Another common choice for L is a differential operator of the form $(-\alpha\nabla^2 + \gamma)Id$. In this case, the kernel K is computed with Green's function of the operator. The general optimization problem can then be summarized as:

$$\underset{(v_t)_{t\in[0,1]}}{\arg\min}\left(\|I\circ\phi_1^{-1} - J\|_{L^2}^2 + \lambda\int_0^1 \|v_t\|_V^2\,dt\right). \tag{4.3}$$

Early approaches of LDDMM directly optimize every velocity field by performing a gradient descent (Beg et al., 2005). However, computing the gradient for every velocity field is cumbersome. Later methods only optimize the initial momentum Lv_0 (Miller et al., 2006; Ashburner and Friston, 2011; Vialard et al., 2011) and use geodesic shooting to compute the deformation with respect to the initial momentum. Indeed, it was proven by Miller et al. (2006) that

$$v_t = K(|D\phi_t^{-1}|(D\phi_t^{-1})^T(Lv_0\circ\phi_t^{-1}))$$

where $D\phi_t^{-1}$ denotes the jacobian of ϕ_t^{-1} and $|\cdot|$ denotes the determinant operator. From the above equation and equation 4.2, we can iteratively compute v_t and ϕ_t^{-1} to retrieve the final deformation. This offers faster convergence

since the deformation is entirely parametrized by the initial momentum, thus making it the only parameter updated through the gradient descent.

Symmetric Normalization (Avants et al., 2008) further constrains the invertibility of the deformation by computing the time-dependent velocity fields of both the forward transformation ϕ and the backward transformation ψ. The authors constrain the inverse of ψ to be ϕ and the inverse of ϕ to be equal to ψ. Let v_t be the velocity fields of ϕ and w_t be the velocity fields of ψ. Both the source and target images are transformed by their respective warpings and the matching is compared at the middle time-point $t = 0.5$. The optimization problem is of the form:

$$\begin{aligned}
&\underset{v_t, w_t}{\arg\min} \left(\|I \circ \phi(\cdot, 0.5) - J \circ \psi(\cdot, 0.5)\|_{L^2}^2 + \lambda \int_0^{0.5} \|v_t\|_V^2 + \|w_t\|_V^2 dt \right) \\
&\text{s.t.} \quad \frac{\partial \phi_t}{\partial t} = v_t(\phi_t) \quad \phi_0 = Id \\
&\text{and} \quad \frac{\partial \psi_t}{\partial t} = w_t(\psi_t) \quad \psi_0 = Id.
\end{aligned}$$

This has also been implemented with the cross-correlation rather than the SSD (Avants et al., 2009).

Other approaches ease the computations required by LDDMM by relaxing its formulation. Rather than enforcing time-dependent velocity fields, they use a stationary vector field (SVF) $v : \Omega \to \mathbb{R}^d$ (Arsigny et al., 2006; Ashburner, 2007). The integration of the flow equation is therefore considerably easier since no geodesic shooting is required beforehand. The flow equation is then:

$$\forall x \in \Omega, \ \frac{\partial \phi_t(x)}{\partial t} = v(\phi_t(x)) \quad \phi_0(x) = x. \tag{4.4}$$

To numerically solve this equation, we set a number of time-step T. Using Euler's method to solve ordinary differential equations, we get:

$$\phi_{t+1/T} = \phi_t + \frac{1}{T} v(\phi_t) = (Id + \frac{1}{T} v) \circ \phi_t = \phi_{1/T} \circ \phi_t.$$

Thus, computing the final deformation consists in iteratively warping the identity grid with the field $\phi_{1/T} = (Id + \frac{1}{T} v)$ T times. However, by choosing T as a power of two so that $T = 2^N$ we can compute an integration with T time steps in only N iterations. Indeed, based on the previous equation, with $t = 1/T$, we have

$$\phi_{1/2^{N-1}} = \phi_{1/2^N} \circ \phi_{1/2^N}.$$

By recursively applying this step N times, we obtain the deformation ϕ_1. This method is called the *scaling and squaring* algorithm. Although it is easier to work with and enforces small velocity fields, this method also has its drawbacks. Namely, since the velocity fields are not time-dependent, the computed

transformation is not the geodesic *i.e.* the path taken by the transformation is not necessarily the shortest path.

A diffeomorphic Demons approach based on the *scaling and squaring* method has been proposed by Lorenzi et al. (2013). The idea is that rather than using the vector field computed with Equation 4.1 as a displacement field, it is used as a stationary vector field and the deformation is computed with the *scaling and squaring* algorithm.

The B-splines approach has also been adapted into a diffeomorphic framework (Rueckert et al., 2006). Several transformations ϕ_i are computed for several uniform grids of various resolutions. They are then assembled together into the final transformation by composition:

$$\phi = \phi_1 \circ ... \circ \phi_N$$

4.4 Learning Approaches for Registration

Classical registration methods are very computationally demanding and when one wants to register several image pairs, it needs to be done independently. Learning methods offer much faster registration time once the model is trained and during the training, the information learned from one iteration is re-used to better align another image pair. Given an ensemble of pair images (source and target), instead of optimizing each pair-specific deformation field, one performs an optimization at the population level of a global function (often a neural network). At test time, a deformation field can be obtained by simply evaluating the function on a given pair of images.

4.4.1 Supervised Learning

Early methods tackle the registration problem in a supervised manner, relying on segmentation maps or previously computed reference deformation fields (Rohé et al., 2017; Yang et al., 2017; Krebs et al., 2017; Cao et al., 2018). Most methods take the image pair as an input of a convolutional neural network (CNN) which predicts the deformation. The reference deformation fields are either obtained by running one of the classical (and computationally intensive) registration methods on a pair of images (Rohé et al., 2017; Yang et al., 2017; Cao et al., 2018) or by computing a deformation between the segmentation of regions of interest and extend it to the whole image using spline interpolations (Rohé et al., 2017; Krebs et al., 2017). In the first case, acquiring the reference deformation fields can be a long and cumbersome process and the performance of the learning method highly depends on the accuracy of the conventional method. In the second case, it is necessary to have a very precise segmentation to get a correct deformation, and the spline interpolation can introduce errors in non-segmented areas.

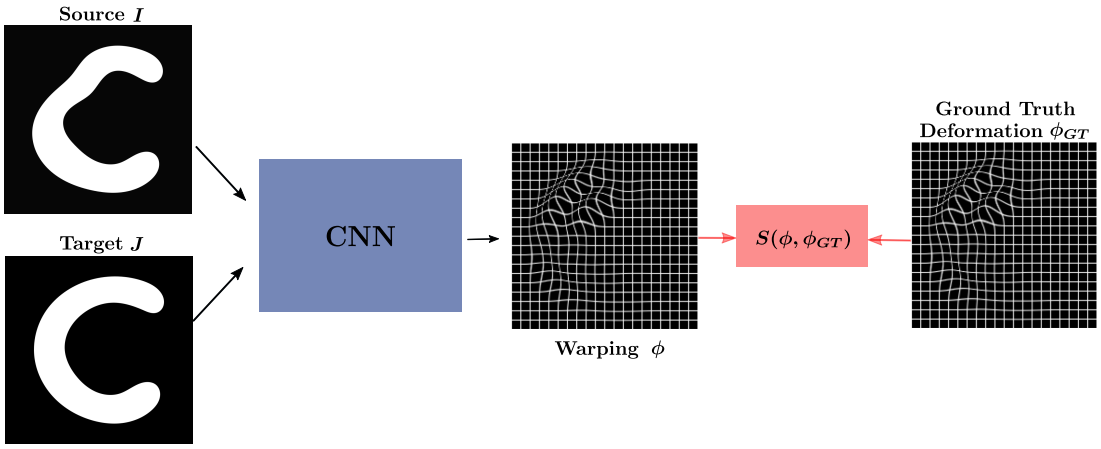


Figure 4.4: Supervised learning framework for image registration. The weights of the model are directly optimized by maximizing a similarity loss function between the output warping and the ground truth deformation.

4.4.2 Unsupervised Learning

To avoid such limitations, unsupervised methods have been developed. The weights of the neural network are optimized by minimizing a dissimilarity measure between the warped image and the target one. Thus, reference deformation fields are not required to train the model. The back-propagation of the gradient from the cost function through the CNN is possible thanks to the spatial transformer layer (Jaderberg et al., 2015) which is a differentiable warping layer. The layer samples the image with a differentiable kernel k. Let $\phi : \Omega_t \rightarrow \Omega$, the sampling of image I is written in the form:

$$\hat{I}(x) = \sum_{n \in \Omega} I(n) k(\phi(x) - n)$$

In the case of bilinear interpolation, the kernel is equal to

$$k(\phi(x) - n) = \prod_{i=1}^{d} \max(0; 1 - |\phi_i(x) - n_i|)$$

which linearly combines the values of the voxels in the neighborhood of $\phi(x)$.

The first approaches directly predict the displacement field (Vos et al., 2017; Li and Fan; Balakrishnan et al., 2019; Sun and Simon, 2021). However, they suffer from the same drawback as classical methods that compute a displacement field: the deformation is not necessarily invertible. Hence, diffeomorphic deep learning-based methods have emerged (Detlefsen et al., 2018; Dalca et al., 2019; Krebs et al., 2019; Mok and Chung, 2020a). The network outputs a vector field and the *scaling-and-squaring* algorithm is used (Arsigny et al., 2006) to generate a diffeomorphic deformation. Most methods integrate a stationary vector field, however, Yang et al. (2017) predict a momentum and compute the deformation using LDDMM geodesic shooting. Although this method has been included in a supervised framework, it could easily be transferred into

an unsupervised context. This has probably not been done due to the longer integration time with the geodesic shooting.

Overall, the objective function is similar to classical registration with the exception that the gradient is computed for several images at the same time:

$$\theta = \arg\min_{\theta} \sum_{i=1}^{N} S(I_i \circ \phi_i^\theta, J_i) + R(\phi_i^\theta) \tag{4.5}$$

where N is the number of image pairs (I_i, J_i) in the batch and ϕ_i^θ is the predicted deformation for the pair i. To further enforce the diffeomorphic propriety, similar strategies as for classical methods have been proposed, namely computing the warping at different scales (Krebs et al., 2019; Mok and Chung, 2020b) or predicting both the forward and backward deformations and imposing the invertibility between them (Mok and Chung, 2020a).

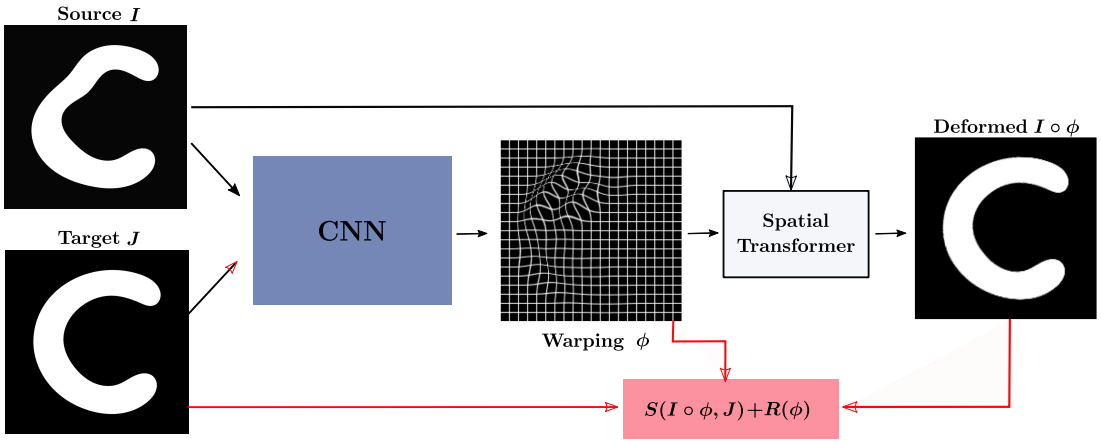


Figure 4.5: Unsupervised learning framework for image registration. The weights are optimized by maximizing the similarity loss between the distorted image and the target.

4.5 Conclusion

In this chapter, we first presented the classical approaches for deformable registration. Among them, the diffeomorphic methods offer nice theoretical guarantees about the smoothness and invertibility of the deformation function. However, they remain time-consuming methods that can require up to several hours to register 3D volumes. Recent deep learning strategies, combined with large datasets, have brought the processing time down to less than a second with a minimal decrease in the performance.

Nevertheless, these methods cannot be applied as such to our problem. Namely, they only tackle the shape differences between images but were not designed to deal with topology variations such as the one that occurs in the presence of a tumor. In the next chapter, we develop a method that combines a diffeomorphic deformation with an intensity change to deal with the topological

variations. Furthermore, we incorporate it in a learning framework to leverage their fast computations.

5

MetaMorph: Learning-Based Metamorphic Registration of Pathological Images

Contents

5.1 Introduction

Image registration is an essential step in medical imaging to perform, for instance, statistical analysis, modality fusion, surgical modeling/planning, or longitudinal studies. Classical techniques assume that source (i.e., moving, the one to be deformed) and target (i.e., fixed) images share the same number of components, and therefore that one can build a one-to-one correspondence between them using diffeomorphic transformations (Ashburner, 2007; Beg et al., 2005; Rueckert et al., 2006). However, some clinical applications

require the alignment of images characterized by a different number of anatomical components, such as a healthy template and a pathological image with a tumor or a lesion (Roux et al., 2019). In particular, when dealing with brain tumors, registration algorithms may be used to correlate tumor location and shape with patient characteristics, surgical treatment, or outcome (Roux et al., 2019). Additionally, the alignment of pre-operative and post-operative images in pathological cases may help the surgeon to assess the quality of the surgical resection (De Witt Hamer et al., 2013). In the case of brain glioblastoma, which is a highly locally recurrent tumor, registration is used to develop tools to predict the location of tumor recurrence (Akbari et al., 2016). Hence, to better understand brain tumors and improve their treatment, providing the most accurate registration algorithm in presence of pathologies is very crucial.

Classical registration approaches based on diffeomorphisms (Ashburner and Friston, 2011; Avants et al., 2008) optimize a functional to maximize the similarity between the deformed source and target image while controlling the bijective property of the deformation and its regularity. This is a slow process since the optimization procedure needs to be repeated for every pair of images. Recent learning strategies (Balakrishnan et al., 2019) offer fast computation at inference. However, like classical registration algorithms, they do not take into account topological differences.

Most of the existing approaches dealing with pathological images minimize a functional, like in classical registration methods, producing very slow registration (Brett et al., 2001; Trouvé and Younès, 2005; Gooya et al., 2012; Liu et al., 2015). Furthermore, they do not properly cope with large lesions and can be specific to a given clinical context (for instance several modalities are required, only for one type of pathology). Recent deep learning methods (Bône et al., 2020; Han et al., 2020b) offer a fast computation time at inference but they do not guarantee a proper disentanglement between shape (*i.e.* geometric) and appearance (*i.e.* intensity) transformations. This means that, when dealing for instance with brain tumors, the appearance transformation should only account for the tumor core (topological difference) and infiltration (i.e. edema) but it should not deal with shape changes due to anatomical differences or tumor mass effect. Morphological variations should be taken into account only with a non-linear geometric deformation. This is critical for correctly interpreting the estimated alignment and for using the estimated transformations in further statistical analysis, such as atlas construction (Gori et al., 2017).

In this chapter, we propose a deep learning approach inspired by Metamorphosis (Trouvé and Younès, 2005) that modifies both the shape and the appearance of the image at the same time. Our method works on single modality images and is generic. It makes no assumption on the imaging modality, and it can be used in any clinical context that requires the alignment of images with different topologies. We use a residual deep network (He et al., 2016) to solve the system of differential equations of Metamorphosis. Furthermore, we spatially limit the appearance changes with a pre-specified mask to better disentangle shape and appearance modifications.

5.1.1 Related Works

Masking Methods

An early approach for the registration of images with a different topology is the cost function masking (CFM) (Brett et al., 2001; Stefanescu et al., 2004), where the tumor/lesion region is ignored when evaluating the cost function. This strategy has also been tied with the creation of an intermediate, cohort-specific template in (Pappas et al., 2021). Unfortunately, the CFM method falls short with large tumors/lesions (Kim et al., 2007). To cope with that, geometric metamorphosis (Niethammer et al., 2011) adds a specific deformation to the masked area, but it works only when the lesion/tumor is present in both source and target images. Additionally, segmentation masks are required for both images.

Tumor Growth Models

In the context of aligning a healthy image with one showing a tumor or a lesion, it has been proposed to first make both images topologically identical and then perform the registration. A first approach has been to simulate the growth of the tumor in the healthy image with a biophysical model (Zacharaki et al., 2009; Gooya et al., 2012; Gholami et al., 2017; Scheufele et al., 2019), and then register it onto the pathological scan. This strategy requires user initialization, and extensive computations to estimate the model parameters, which are specific to a particular kind of tumor. Although a recent fully-automatic method was introduced in Scheufele et al. (2021), it is based on a rather simplistic biophysical growth model. In Nielsen et al. (2019), authors use a similar perspective with a non-biophysical growth model computed simultaneously with the diffeomorphic warping. Despite being more generic than the previous methods, it still requires user initialization and extensive computations.

Healthy Image Synthesis

An opposite strategy consists in removing the tumor to generate a healthy image. In (Liu et al., 2015; Yang et al., 2016; Han et al., 2017; Tang et al., 2019), the pathological region is removed by synthesizing a quasi-normal image via low-rank approaches. This approach can effectively recover tumor regions, but at the same time distort or blur the healthy regions. Furthermore, it is a statistical technique that needs lesions to be homogeneously (and randomly) distributed across the population (Liu et al., 2015), which is not the case for all kinds of lesions or tumors (*e.g.*, brain glioblastoma). With a similar perspective, inpainting techniques on brain MRI have also been proposed (Almansour et al., 2021; Liu et al., 2021). However, in the presence of a strong mass effect (deformation of healthy tissues surrounding the tumor), the inpainting of a tumor might not produce realistic results (Almansour et al., 2021).

Metamorphosis

A mathematical elegant method, called Metamorphosis (Trouvé and Younès, 2005; Garcin and Younes, 2005; Fletcher et al., 2009), has been developed to align images with different shapes and appearances. It does not assume a one-to-one correspondence between the source and target images. It basically consists in alternately deforming the input image (using diffeomorphisms) and adding (small) intensity variations several times. In the context of brain tumors, it can be used in both directions: synthesizing a healthy image and aligning it on the non-cancerous target or generating a tumor in the healthy patient and registering it on the diseased brain. The main drawbacks of this method are that it's computationally cumbersome (François et al., 2021) and finding the parameters that perfectly disentangle shape and appearance is rather difficult. Indeed, morphological and topological differences should be taken into account by the diffeomorphic deformation and intensity addition respectively. Appearance modifications should not account for shape differences.

Learning Approaches

Deep learning methods dealing with images showing different topologies have also been developed. Czolbe et al. (Czolbe et al., 2021), used registration networks for the detection of topological differences. This could be used as a pre-processing step when no segmentation labels are available. In Bône et al. (2020), the authors proposed a Metamorphic Variational Auto-Encoder (MVAE) with two branches to modify both the geometry and the appearance of an image at the same time. However, as shown in the experiments, finding the correct hyper-parameters to balance shape and appearance modifications is complicated. Similarly, in Han et al. (2020a), a network with three branches performs the synthesis of a healthy counterpart of a pathological image, the deformation, and the segmentation of the tumor. Yet, since the method does not explicitly restrict the synthesis to the tumor region, it may result in poor disentanglement, where the synthesis may actually modify the shape of the source image.

5.1.2 Metamorphosis

The aim of Metamorphosis is to modify the source image I so that it perfectly aligns with a target image J. The model joins diffeomorphic deformations with additive intensity changes. Metamorphosis is cast in a similar framework as LDDMM. In fact, LDDMM is a specific case of Metamorphosis where the appearance change is set to 0. As a reminder, in LDDMM, the deformation verifies the flow equation

$$\frac{\partial \phi_t}{\partial t} = v_t(\phi_t).$$

If we define $I_t = I \circ \phi_t^{-1}$ as before, it is possible to prove that

$$\frac{\partial I_t}{\partial t} = -\langle \nabla I_t, v_t \rangle$$

First, we show that for all $x \in \Omega$:

$$\frac{\partial \phi_t^{-1}(x)}{\partial t} = -(D\phi_t)^{-1} v_t(x).$$

where $D\phi_t$ is the Jacobian matrix of ϕ_t. Using the fact that for all $x \in \Omega$, $\phi_t^{-1}(\phi_t(x)) = x$ and writing $y_t = (y_t^1, ..., y_t^d)^T = \phi_t(x)$, by applying the chain rule, we get:

$$\begin{aligned}
\frac{\partial \phi_t^{-1}(\phi_t(x))}{\partial t} = 0 &\iff \frac{\partial \phi_t^{-1}}{\partial t}(y_t) + \sum_{i=1}^{d} \frac{\partial \phi_t^{-1}}{\partial y_t^i}^T \frac{\partial y_t^i}{\partial t} = 0 \\
&\iff \frac{\partial \phi_t^{-1}}{\partial t}(y_t) = -D\phi_t^{-1} \frac{\partial \phi_t(x)}{\partial t} \\
&\iff \frac{\partial \phi_t^{-1}}{\partial t}(y_t) = -(D\phi_t)^{-1} v_t(y_t)
\end{aligned}$$

because $\frac{\partial \phi_t(x)}{\partial t} = v_t(\phi_t(x))$ and the Jacobian of the inverse of a function is the inverse of the Jacobian of that function. Since ϕ_t is a bijection in Ω, we have that for all $x \in \Omega$

$$\frac{\partial \phi_t^{-1}(x)}{\partial t} = -(D\phi_t)^{-1} v_t(x).$$

Now, when differentiating I_t and writing $\gamma_t = (\gamma_t^1, ..., \gamma_t^d)^T = \phi_t^{-1}(x)$ we have for all $x \in \Omega$:

$$\begin{aligned}
\frac{\partial I_t}{\partial t}(x) &= \frac{\partial I_0 \circ \phi_t^{-1}(x)}{\partial t} \\
&= \sum_{i=1}^{d} \frac{\partial I_0}{\partial \gamma_t^i}(\gamma_t) \frac{\partial \gamma_t^i}{\partial t} \\
&= \nabla I_0^T(\gamma_t) \frac{\partial \phi_t^{-1}}{\partial t}(x) \\
&= -\nabla I_0^T(\gamma_t)(D\phi_t)^{-1}(x) v_t(x)
\end{aligned}$$

By applying the chain rule, we also get that

$$\nabla I_0^T(\phi_t^{-1}(x))(D\phi_t)^{-1}(x) = \nabla I_t^T(x)$$

Thus, we obtain for all $x \in \Omega$:

$$\frac{\partial I_t}{\partial t}(x) = -\nabla I_t^T v_t(x) = -\langle \nabla I_t(x), v_t(x) \rangle.$$

Geodesic Shooting

In Metamorphosis (Trouvé and Younès, 2005; François et al., 2021), the intensities of image I are modified by adding a residual image $z_t : \Omega \to \mathbb{R}$ corresponding to the infinitesimal intensity variation (called the residual image or momentum):

$$\frac{\partial I_t}{\partial t} = -\langle \nabla I_t, v_t \rangle + \mu^2 z_t \quad \text{s.t. } I_0 = I \text{ and } I_1 = J \tag{5.1}$$

The parameter $\mu^2 \in \mathbb{R}^+$ balances the intensity and geometric changes and if it is equal to 0, we get the LDDMM framework. The goal of Metamorphosis is to compute the minimal geodesic path by optimizing the energy of the transformation, $\int_0^1 \|v_t\|_V^2 + \|\mu z_t\|_2^2 dt$, under the condition in Equation 5.1. As shown in Trouvé and Younès (2005); Holm et al. (2008), by computing the Euler-Lagrange equations, one obtains the following geodesic equations for Metamorphosis:

$$\begin{cases} v_t = -K(z_t \nabla I_t) & \text{(5.2a)} \\ \dfrac{\partial z_t}{\partial t} = -\nabla \cdot (z_t v_t) & \text{(5.2b)} \\ \dfrac{\partial I_t}{\partial t} = -\langle \nabla I_t, v_t \rangle + \mu^2 z_t & \text{(5.2c)} \end{cases}$$

with $\|v_t\|_V^2 = \langle K(z_t \nabla I_t), z_t \nabla I_t \rangle = \langle v_t, L v_t \rangle$, where, as for LDDMM, K is a kernel.

As in François et al. (2021), we cast the metamorphic registration as an inexact matching problem minimizing the cost function:

$$E = \frac{1}{2}\|I_1 - J\|_2^2 + \lambda [\int_0^1 \|v_t\|_V^2 + \|\mu z_t\|_2^2 dt] \tag{5.3}$$

From this system of equations, we can notice that v_t is completely defined by z_t and I_t, thus making z_0 the only unknown. The momentum z_t has therefore a double role. It represents the additive intensity variation *and* it is also the parameter of the deformation. This eases the computation but at the same time it makes the disentanglement between geometry and intensity variations more difficult.

From Equation 5.2c, we define the infinitesimal action of v_t on I_t as:

$$v_t \cdot I_t = -\langle \nabla I_t, v_t \rangle.$$

Therefore, we obtain

$$\frac{\partial I_t}{\partial t} = v_t \cdot I_t + \mu^2 z_t.$$

This is useful when discretizing Equation 5.1 since we get:

$$I_{t+1} = I_t + \delta v_t \cdot I_t + \delta \mu^2 z_t = (Id + \delta v_t) \cdot I_t + \delta \mu^2 z_t$$

where Id is the identity function on Ω, and $\delta > 0$ is the discretization step. Computationally, to retrieve I_{t+1}, it is the inverse of $(Id + \delta v_t)$ that is applied to I_t. Since v_t belongs to V, the inverse can be approximated with $Id - \delta v_t$. Thus, to compute I_{t+1} from I_t, one can warp the latter with $(Id - \delta v_t)$:

$$I_{t+1} = I_t \circ (Id - \delta v_t) + \delta \mu^2 z_t.$$

5.2 Methods

As previously explained, solving the optimization problem of Metamorphosis can be computationally expensive and slow. We propose here two strategies to redefine it into a learning-based method, that is fast at inference. By leveraging CNNs or ResNets, we can estimate either only z_0 or the entire flow z_t. We found that computing the entire flow with a ResNet offers the best results.

5.2.1 MetaMorph - Geodesic shooting

Our first approach only predicts the initial momentum z_0. This strategy is similar to the diffeomorphic learning-based registration methods (Krebs et al., 2019; Dalca et al., 2019). We determine the initial residuals z_0 with a U-Net that takes both source and target images as input. The total transformation is computed using the shooting equations of Metamorphosis (Equation 5.2). We use the semi-Lagrangian scheme introduced in François et al. (2021) to solve the differential equations. The scheme is presented in Algorithm 5.1. The network is trained by minimizing the same data and regularization terms as Metamorphosis for each pair in the database. Therefore, one needs to minimize the following energy:

$$E_G(\theta) = \sum_{n=1}^{N} \left[\frac{1}{2} \|I_T^n - J\|_2^2 + \frac{\lambda}{T} [\sum_{t=0}^{T-1} (\|v_t^n\|_V^2 + \mu^2 \|z_t^n\|_2^2)] \right] \tag{5.4}$$

where n is the index of image I^n in the dataset and θ the parameters of the neural network. We call this method MetaMorph-G.

Algorithm 5.1 Geodesic shooting with a semi-Lagrangian scheme

Data: Initial residual z_0, Source image I_0, Number of time steps T, $\mu > 0$
Result: Transformed image I_T

Initialize identity grid Id s.t. $Id(x) = x$
for $t \leftarrow 0$ *to* $T-1$ **do**
  $v_t = -K * (z_t \nabla I_t)$
  $I_{t+1} = (Id + \frac{1}{T} v_t) \cdot I_t + \frac{1}{T} \mu^2 z_t$
  $z_{t+1} = (Id + \frac{1}{T} v_t) \cdot z_t - \frac{1}{T} \nabla \cdot (v_t) z_t$
Return I_T

Optionally, we can add an inverse-consistency term to further reinforce the diffeomorphic property of the warping field. The inverse of each small deformation $Id - \delta v_t$ can be approximated with $Id + \delta v_t$. Thus, the inverse of ϕ_T^{-1} can be estimated with

$$\hat{\phi}_T = (Id + \delta v_{T-1}) \circ ... \circ (Id + \delta v_0).$$

The inverse consistency term is then:

$$R_{inv}(\phi_T^{-1}) = \|Id - \phi_T^{-1} \circ \hat{\phi}_T\|_2^2$$

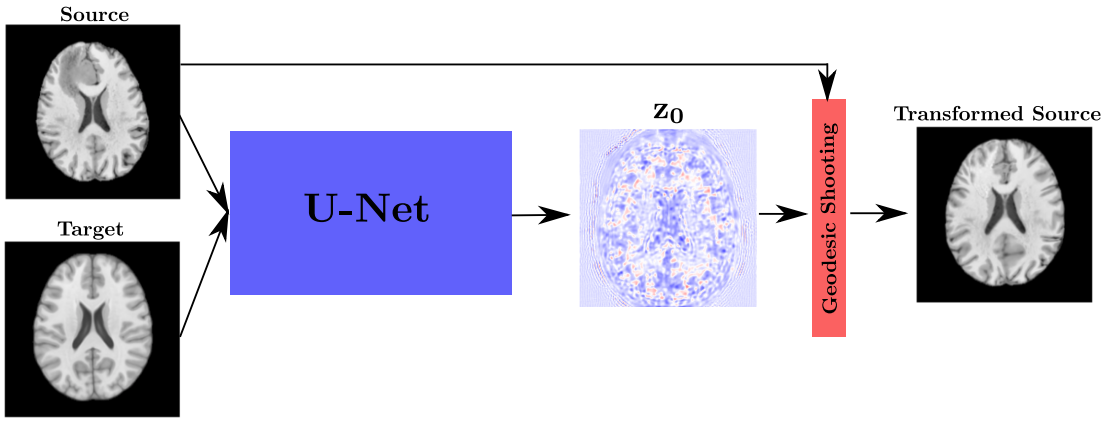


Figure 5.1: MetaMorph-G framework.

Interestingly, although the model is primarily meant to be used in a learning context, it is possible to use it on a single pair of images and optimize its parameters by repetitively minimizing the above functional on that same pair. In this case, the algorithm is similar to the original metamorphosis.

5.2.2 MetaMorph - Resnet integration

As second strategy, we propose to directly estimate all z_t. Inspired by Amor et al. (2021); Rousseau et al. (2020), we propose to use a residual neural network (ResNet) to find the solution of the system of differential equations 5.2. We take advantage of the similarity between ResNets and the numerical solutions of ODEs using Euler's method, given an initial value. Indeed, the numerical integration of Equation 5.2b, using discrete time steps t, is:

$$z_{t+1} = z_t - \delta\nabla\cdot(z_t v_t) \text{ for } t \in 0,..,T-1 \tag{5.5}$$

where T is the number of steps and δ is the integration step equal to $\frac{1}{T}$. By replacing the divergence with a neural network, we obtain a ResNet:

$$z_{t+1} = z_t - \delta f_{\theta_t}(z_t, I_t, J) \tag{5.6}$$

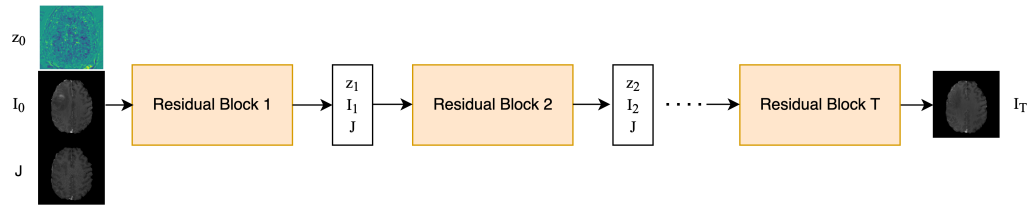
where f_{θ_t} is a convolutional neural network with parameters θ_t. The benefit of using a neural network is that metamorphoses can be applied in a learning context rather than just in an optimization scheme. For that reason, the source and target images are also given as input to f_{θ_t}. The network is built as a sequence of T convolutional blocks f_{θ_t}. At each time step t, z_{t+1} is computed using Equation 5.6. Subsequently, v_{t+1} is calculated directly with Equation 5.2a and one determines I_{t+1} by applying the geometric transformation induced by v_t and adding the residuals z_t as in Equation 5.2c. The architecture of the model is detailed in Figure 5.2.

The parameters of this model are optimized by minimizing the same data and regularization terms as for the previous method. However, in this case, the initial residual z_0 is the same for every image pair and is learnt during training.

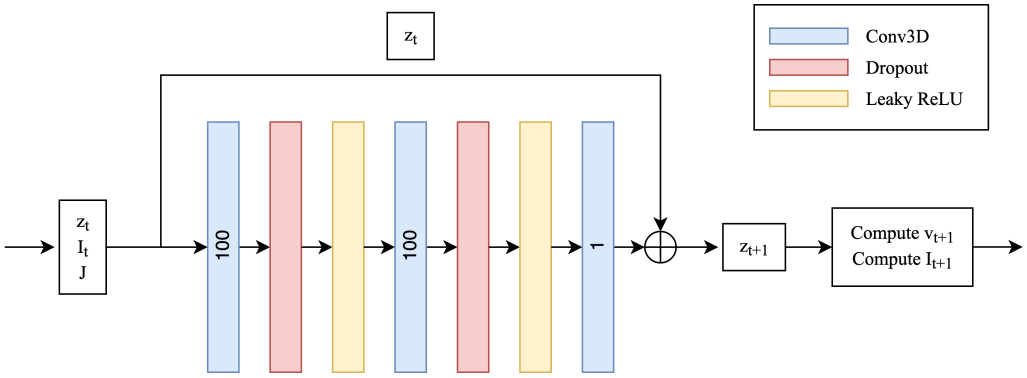
The training loss is thus:

$$E_R(\theta, z_0) = \sum_{n=1}^{N} \left[\frac{1}{2} \|I_T^n - J\|_2^2 + \frac{\lambda}{T} \left[\sum_{t=0}^{T-1} \|v_t^n\|_V^2 + \mu^2 (\|z_0\|_2^2 + \sum_{t=1}^{T-1} \|z_t^n\|_2^2) \right] \right] \tag{5.7}$$

Furthermore, as before, we can also add an inverse-consistency term. We call this implementation MetaMorph.



(a) Overall design of the neural network.



(b) Composition of a residual block. z_t, I_t and J are concatenated before the convolution. The output of the block is added to z_t to form z_{t+1}. The numbers on the blue layers are the number of channels of the output tensor.

Figure 5.2: The residual network is composed of T residual blocks. All residual blocks have the exact same architecture.

5.2.3 Integration Scheme

The numerical solution of Equation 5.2c using Euler's method is

$$\begin{aligned} I_{t+1} &= (Id + \delta v_t) \cdot I_t + \delta \mu^2 z_t \\ &= I_t \circ (Id - \delta v_t) + \delta \mu^2 z_t \end{aligned} \tag{5.8}$$

where the integration step is $\delta = \frac{1}{T}$. This indicates that at each time-step t, I_t is deformed by the vector field v_t. Computationally, the resulting image is an interpolation of I_t. Therefore, the final image I_T is obtained after T interpolations of I_0. Using a high number of successive nearest-neighbor interpolations creates a pixelation effect and a bilinear (or trilinear) interpolation blurs the image. To avoid both of these downsides, we rewrite Equation 5.8 so that I_{t+1} is a direct deformation of I_0 and not I_t. We recursively replace I_t by its expression

in function of I_{t-1} in Equation 5.8 until we reach I_0. We get:

$$I_{t+1} = I_0 \circ \phi_{0,t+1}^{-1} + \delta\mu^2 \sum_{i=0}^{t} z_i \circ \phi_{i+1,t+1}^{-1} \tag{5.9}$$

where $\phi_{i,t+1}(x)$ is the position at time $t+1$ of the element at position x at time i:

$$\begin{cases} \phi_{i,t}^{-1} = Id & \text{if } i = t \quad (5.10a) \\ \phi_{i,t}^{-1} = Id - \delta v_{t-1} & \text{if } i = t-1 \quad (5.10b) \\ \phi_{i,t}^{-1} = (Id - \delta v_i) \circ \ldots \circ (Id - \delta v_{t-1}) & \text{otherwise.} \quad (5.10c) \end{cases}$$

Here, Id denotes the identity function: $Id(x) = x$.

With this expression, each image I_t is resulting from only one resampling of I_0. Note that even if I_{t+1} is not directly computed from I_t, it is still necessary to calculate the latter as it is required to get v_t and z_t.

As seen in Figure 5.3, this integration scheme produces more accurate and less blurry deformations. Moreover, this integration scheme is not specific to our ResNet model, and it can be applied to both MetaMorph-G and standard Metamorphosis.

5.2.4 Local regularization.

The main inconvenience with Metamorphosis is that it is hard to control the disentanglement between shape and appearance. For instance, a trivial solution would be to set the overall geometrical deformation function to the identity (no geometrical change) and the overall appearance deformation map to $J - I_0$. In that case, the L_2 distance between the deformed image and J would be 0 but it would not be a satisfactory result since homologous structures should be matched using only geometric deformations whereas appearance and topological changes (*i.e.*, new components) should be taken into account by the intensity modifications. The disentanglement can be controlled by tuning the hyper-parameters μ and λ. However, finding the right ones is a difficult task and they are different for each setting. If they are not correctly chosen, the appearance map could, for instance, modify the shape of the image, thus distorting the results and their interpretations.

To this end, we propose to restrict the intensity modifications (i.e. z) only to the regions showing a topological or appearance difference between the source and target images. Here, we do that by multiplying z by a mask m of the region where the topological/appearance changes occur (a tumor for instance). Equation 5.2c then becomes

$$\frac{\partial I_t}{\partial t} = v_t \cdot I_t + \mu^2 m_t z_t \tag{5.11}$$

with $m_0(x) = 1$ if x is a voxel in the selected region and 0 otherwise. Since the region varies along t with the source image, the mask must follow the deformation generated by the velocity fields. Consequently, the mask is not

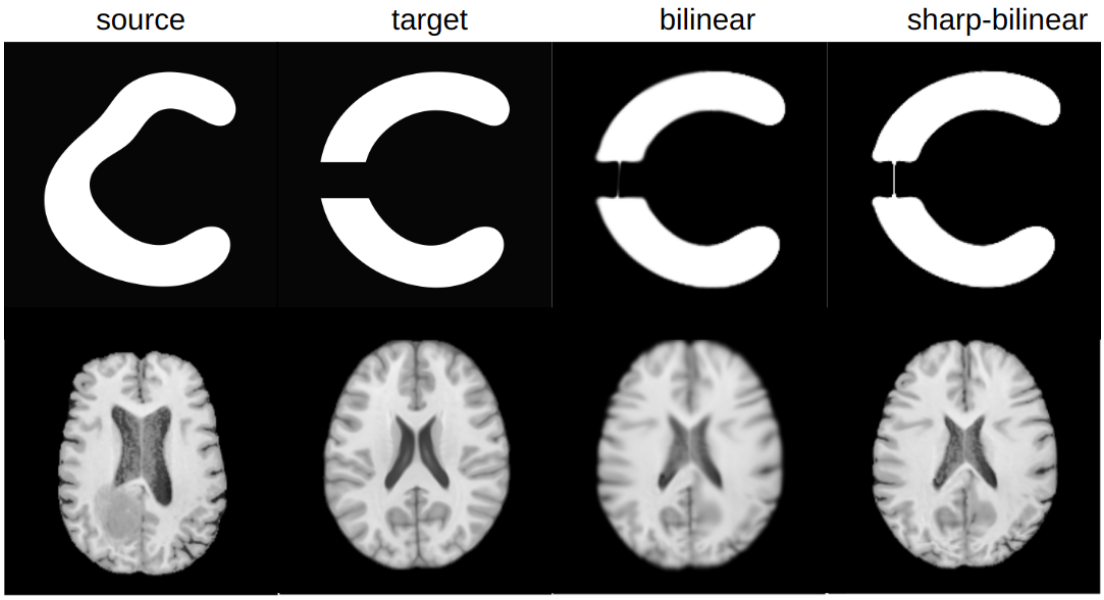


Figure 5.3: Results of MetaMorph-R with different integration schemes: *bilinear* corresponds to the Eulerian integration, applying T times a bilinear interpolation; *sharp-bilinear* corresponds to the new integration scheme using the bilinear interpolation. In the first row, μ is set to 0 (i.e., LDDMM) and the resulting warping field should therefore be diffeomorphic. This means that the topology of the source image should be preserved. However, only the proposed sharp interpolation preserves it (small white strip) wheres the successive bilinear interpolations blur the images and create two different components. In the second row, μ is equal to 0.1 (i.e., Metamorphosis) to show both geometric and intensity changes. Results are less blurry with the proposed sharp interpolation.

fixed but it follows the equation:

$$\frac{\partial m_t}{\partial t} = v_t \cdot m_t.$$

5.3 Evaluations

5.3.1 Data

Synthetic Data

We first evaluate our method on a database of 2D-generated “C-like" images. It is constituted of a template image and 2000 transformations of that image. The original image is a white “C" on a black background of size 200×200 pixels. Each transformation is first computed by changing the topology of the template and then applying a random deformation. The topological change is done by cutting the “C" with a rectangle of random width and random angle. The deformation is generated by randomly sampling a momentum and integrating it with the shooting equations of LDDMM (Equation 5.2). The genera-

tion process is detailed in Algorithm 5.2. In this way, due to Equation 5.2a and the fact that images are binary, we will have a smooth deformation that will be strong around the contour of the C (high values of ∇I_t), close to the identity in the homogeneous black areas of the image ($\nabla I_t \sim 0$) and will ignore the topological differences. The resulting deformations can thus be used as "ground truth" for comparing different registration methods. The rectangle to cut the "C" is used as mask for the proposed local regularization.

Algorithm 5.2 Generation of the "C-like" dataset

Data: $N > 0$, $T > 0$, Image I_0, Gaussian kernel K
Result: Database of N transformed images and the associated transformation.
$h, w \leftarrow \text{shape}(I_0)$
for $n \leftarrow 1$ *to* N **do**
  $I_0^n \leftarrow \text{change\_topology}(I_0)$
  Initialize identity grid ϕ_0^n s.t. $\phi_0^n(x) = x$
  $z_0^n \leftarrow uniform(-1000, 1000)^{h \times w}$
  for $t \leftarrow 0$ *to* $T-1$ **do**
    $v_t^n = -K * (z_t^n \nabla I_t^n)$
    $\phi_{t+1}^n = \phi_t^n \circ (Id - \frac{1}{T} v_t^n)$
    $I_{t+1}^n = I_0^n \circ \phi_{t+1}^n$
    $z_{t+1}^n = z_t - \frac{1}{T} \nabla \cdot (z_t^n v_t^n)$
  Save I_T^n, ϕ_T^n

BraTS 2021

The second dataset is from the Brain Tumor Segmentation Challenge (BraTS) 2021 (Menze et al., 2015; Bakas et al., 2017; Baid et al., 2021) comprising four MR modalities and the associated tumor segmentation image for 1251 brains with tumor. The experiments are only conducted using the T1 modality. We register the scans on the healthy sri24 template (Rohlfing et al., 2010). As a preprocessing, we perform histogram matching on every scan with the template as the target image and crop it to the size of $192 \times 192 \times 144$ voxels. We randomly pick 34 images from the dataset to form an evaluation set. We use the segmentation of the tumor as a mask for local regularization.

BraTS-Reg

The method is evaluated on the BraTS-Reg dataset (Baheti et al., 2021), which includes 140 training and 20 validation subjects. For each subject, the pre-operative and follow-up T1, T1 contrast-enhanced (T1ce), T2, and Flair modalities are available. The scans are manually annotated with 6 to 50 landmarks. Evaluation is based on the alignment quality between the landmarks of the follow-up scan to the landmarks of the pre-operative scan. For the validation set, the landmarks are only available for the follow-up scan and the evaluation is done by uploading their deformation on an online evaluation plat-

form[1]. All scans have been skull-stripped and rigidly registered to the same template. The images have a 1 mm^3 isotropic voxel size, and dimensions equal to 240×240×155. As pre-processing, we first used ANTs (Avants et al., 2008) to rigidly register the follow-up scans to the pre-operative scans. Then, we crop the images to the size $192 \times 192 \times 144$. Finally, we register the pre-operative T1ce scan to the follow-up T1ce and use the estimated inverse deformation $\phi^{-1}(l_n)$ to retrieve the position of each landmark l_n from the follow-up space to the pre-operative space. The segmentation of the whole tumor is used as local regularization. Segmentation masks are obtained using a U-Net (Ronneberger et al., 2015) model trained on BraTS 2021 (average dice score of 0.90 in the validation set). Please note that, differently from the best-performing methods in the online platform, here we use a very simple pre-processing, no post-processing and only one modality. The goal here is to compare different registration methods with topologically different source and target images, and not to find the best algorithm for this specific challenge.

5.3.2 Scoring functions

C-shape dataset. For the synthetic dataset, we have access to the ground-truth backward deformation which is the deformation from the target image to the source image. The composition of the latter with the predicted forward warping should therefore be close to the identity if the prediction is correct. We call ψ_1 the ground truth warping, and we measure the registration score with:

$$s(\psi_1, \phi_T^{-1}) = \|Id - \psi_1 \circ \phi_T^{-1}\|_2$$

For the overall transformation quality measure, we use the $\mathcal{L}_2$ distance between the transformed and target images. A perfect alignment would give a score equal to 0.

BraTS 2021. Measuring the quality of the registration is not a straightforward task since there is no well-defined ground truth. We use the $\mathcal{L}_2$ distance between the target and the deformed source intensities and the $\mathcal{L}_2$ distance outside the mask to evaluate the registration. Additionally, we manually segment the ventricles of 34 images and warp them with the computed deformation to measure the overlap with the ventricles of the target image (i.e., sri24 template) with the Dice score.

BraTS-Reg. For this dataset, we evaluate the mean and median absolute error between the deformed and target landmarks. Furthermore, we compute the robustness which is the proportion of key points for which the absolute distance has decreased after registration.

Finally, for all three datasets, we calculate the number of negative elements in the determinant of the Jacobian matrix to measure the diffeomorphic properties of the warping fields.

[1] https://www.cbica.upenn.edu/BraTSReg2022/lboardValidation.html

5.3.3 Baselines

We first compare our methods and Metamorphosis (Meta) with rigid registration, symmetric normalization (SyN) using the ANTs package in python (Avants et al., 2009), and voxelmorph (VM) (Balakrishnan et al., 2019) to show that one-to-one methods are not adapted in this context. Additionally, we compare it with their cost masking versions (rigid-CFM, SyN-CFM, VM-CFM). Since Voxelmorph is a learning-based method, the cost function is masked during the training of the network and the model takes the source image masked by the tumor segmentation during both training and test. Furthermore, we use Metamorphic Auto-Encoders (MAE) (Bône et al., 2020) only with 2D data since the code is not available in 3D.

5.3.4 Numerical aspects

We set the number of integration step T to 15. We found that for both BraTS datasets, the best values for λ, σ, and μ are respectively 1e-7, 8, and 0.1. For the methods computed with ANTs we use the default parameters and for voxelmorph, we set the hyper-parameter λ to 0.1.

All the deep learning models and Metamorphosis are computed on an Nvidia A100 GPU with 40GB of memory. All methods from the ANTs package are computed on an Intel Xeon Silver 4110 CPU, 2.10GHz with a memory of 32 GB.

The 3D network trained on Brats has about 1.5 million parameters, it requires about 30 GB of GPU memory during training and takes 24 hours. This is mainly because, unlike most learning methods, there is no intermediate downscaling of the input image. With 3D images of size $192 \times 192 \times 144$, computing gradients is therefore more memory expensive. Nevertheless, during inference, since no gradients are computed, the model only requires 10 GB of VRAM memory, which allows the use of standard and accessible GPU cards. Furthermore, running the model on a single image at inference time takes less than one second. In comparison, Metamorphosis takes around 6 hours to converge for a single pair of the same volumes. Voxelmorph requires 13 GB of VRAM memory and takes 16 hours.

Table 5.1: Comparison of computation time and memory use between MetaMorph-R, Metamorphosis, and VoxelMorph on one image from Brats 2021.

Method	Time	Memory (GB)
Meta	6h	21
MetaMorph-R (Inference)	<1s	10
MetaMorph-R (Training)	24h	30
VoxelMorph (Training)	16h	13
Voxelmorph (Inference)	0.1s	4

5.3.5 Results

Results on the “C-shape" dataset are provided in Table 5.2 and in Figure 5.4. As expected, classical methods cannot properly align images due to the “cut" in the source image. To compensate, they need to merge the two white components together to fill the “cut" which creates unrealistic deformations. This can be illustrated by Voxelmorph which gets scores relatively close to ours regarding $\mathcal{L}_2$ distance but the dissimilarity between the warping field and the ground truth is much higher. Naturally, the cost-function masking methods have a high $\mathcal{L}_2$ distance since the “cut" is ignored during optimization. However, the similarity with the ground truth warping field is much higher than the classical versions. Our methods can do both: having a low $\mathcal{L}_2$ distance and producing deformation fields similar to the ground truth. The Metamorphic Auto-Encoder manages to get a low $\mathcal{L}_2$ distance but the dissimilarity with the ground truth is very high. This is due to the fact that finding the optimal hyper-parameters to generate smooth warping fields is difficult. Figure 5.4 shows that a small change in λ_s (the parameter for the shape regularization) completely changes the smoothness of the deformation. Either the shape deformation contains more than a thousand folds and both images are correctly aligned, or it is too smooth and the images are not aligned. In both cases, we found that the appearance branch is unable to locate the region with topological change. With our methods, by constraining the intensity modification to a pre-specified zone, finding satisfactory hyper-parameters is much easier. Hence, our methods obtain better results for all values of the hyper-parameters. It is interesting to notice that MetaMorph-R with the least regularization (*i.e.* $\lambda = 3 \times 10^{-7}$) produces non-diffeomorphic deformations and as result has a higher distance with the ground-truth. Furthermore, for equal hyper-parameters, the Resnet approach reaches better results than the shooting method.

On the BraTS dataset, MetaMorph-R outperforms all others regarding $\mathcal{L}_2$ distances and only Metamorphosis has a comparable dice score (Table 5.3). The Dice score of the ventricles is significantly higher than the one for every other learning method and the $\mathcal{L}_2$ distance computed outside the tumor location is also lower. This shows a better alignment of the healthy tissues of the brain. MetaMorph-R produces few folds on average (about 0.01% of voxels) which is less than the other learning methods. Results show that predicting z_0 to compute the geodesic shooting is not efficient since results are even worse than Voxelmorph. Figure 5.5 shows the visual comparison for three different subjects of MetaMorph, Voxelmorph, and SyN-CFM. On all three patients, our model better aligns the ventricles. Namely, on the first row, it seems that the tumor prevents VM and SyN-CFM from correctly aligning the left ventricle, which is not the case for our method. It seems that because our model changes both the appearance and the shape of the image simultaneously, it can better align healthy tissues. Furthermore, the method generates rather realistic healthy images, although some edges of the tumor mask can be spotted due to a sudden intensity change between healthy and masked regions.

Table 5.2: Evaluation of the methods on the "C-shape" dataset. Results are shown as mean(standard deviation). For Meta, MetaMorph-R and Metamorph-G if hyper parameter values are not specified then $\sigma = 6$, $\lambda = 3 \times 10^{-6}$, and $\mu = 0.05$

Method	$\mathcal{L}_2$	$s(\psi_1, \phi_T^{-1})$	$\lvert J_\phi \rvert < 0$
Rigid	1616(236)	4575(2110)	0(0)
Rigid-CFM	1605(250)	18271(17141)	0(0)
SyN	282(113)	3904(1989)	0.23(2.02)
SyN-CFM	538(177)	623(110)	0(0)
VM ($\lambda = 0.1$)	97.18(33.6)	2195(580)	13.9(38.9)
VM ($\lambda = 0.5$)	331(115)	1932(742)	0(0)
VM-CFM ($\lambda = 0.1$)	576(184)	443(86)	0(0)
Meta ($\sigma = 4$)	17.1(13.9)	390(38)	0(0)
Meta ($\sigma = 6$)	45.3(21)	386(35)	0(0)
Meta ($\sigma = 10$)	53(28)	277(30)	0(0)
MetaMorph-G ($\sigma = 4$)	30.9(12.5)	370(29)	0(0)
MetaMorph-G ($\sigma = 6$)	45.4(16.7)	320(32)	0(0)
MetaMorph-G ($\sigma = 10$)	70.5(32.3)	282(28)	0(0)
MetaMorph-R ($\sigma = 4$)	9.65(7.64)	381(41.8)	0(0)
MetaMorph-R ($\sigma = 6$)	20.95(11.9)	352(35)	0(0)
MetaMorph-R ($\sigma = 10$)	69.7(24.5)	**265(37)**	0(0)
MetaMorph-R ($\lambda = 3 \times 10^{-5}$)	37.7(9.5)	275(26)	0(0)
MetaMorph-R ($\lambda = 3 \times 10^{-7}$)	**8.16(3.2)**	601(31)	4.9(18)
MetaMorph-R ($\mu = 0.03$)	60(128)	297(40)	0(0)
MetaMorph-R ($\mu = 0.1$)	16.4(4.9)	442(45)	0(0)
MAE ($\lambda_s = 269$)	77(32)	4492(1501)	15543(12005)
MAE ($\lambda_s = 269.5$)	2003(511)	3051(985)	0(0)

Table 5.4 shows the result for the BraTS-Reg dataset. We did not evaluate MetaMorph-G on this dataset since it performed significantly worse than MetaMorph-R on the two previous dataset. Our method produces the best results among all baselines. It reduces respectively by 33 % and 24 % the median absolute error and the mean absolute error with respect to the second best-performing method SyN. Furthermore, it is as robust as the latter and considerably more than all others. It is also interesting to notice that, compared to the other two learning methods, it is the one that has the least folds: only 2.7 on average per image during validation.

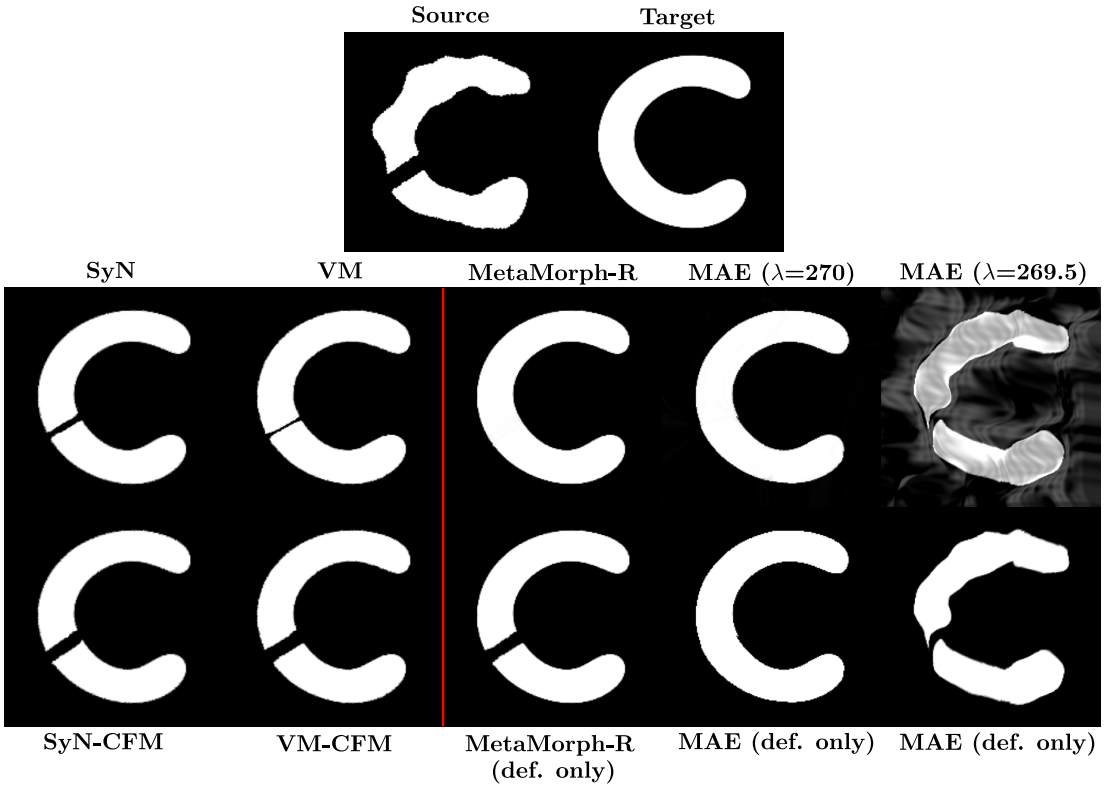


Figure 5.4: Results on the synthetic dataset. The left part shows methods that cannot modify image intensities in contrast to the ones on the right part. For the latter, the bottom row shows the source image deformed by the predicted warping field. For MAE, λ_s is the parameter that controls the smoothness of the deformation.

Table 5.3: Results on BraTS 2021 dataset. The $\mathcal{L}_2$ scores were divided by 10^4 for better readability. Classical registration methods are compared with their cost function masking version and our method. The "Dice" column provides the Dice scores between the segmentation of the ventricles in both images after registration. $\mathcal{L}_2$ is the $\mathcal{L}_2$ distance between the deformed and target image. $\mathcal{L}_2$ w/o tumor excludes the tumor region when computing the latter distance. $|J_\phi| < 0$ measures the number of folds in the image.

Method	Dice	$\mathcal{L}_2$	$\mathcal{L}_2$ w/o tumor	$\lvert J_\phi\rvert < 0$
Rigid	43.9(12.3)	8.1(1.0)	7.4(0.9)	**0(0)**
SyN	55.7(12.9)	5.5(1.0)	4.9(0.8)	**0(0)**
VM	65.5(7.2)	4.4(0.8)	3.3(0.6)	4427(1821)
Rigid-CFM	43.9(12.3)	8.1(1.1)	7.5(0.9)	**0(0)**
SyN-CFM	56.4(12.8)	5.5(1.0)	4.9(0.9)	**0(0)**
VM-CFM	61.5(8.1)	4.9(1.0)	3.4(0.6)	3423 1733)
Meta	**70.3(5.3)**	4.2(0.6)	3.9(0.6)	**0 (0)**
MetaMorph-G	65.32(6.25)	5.1(0.7)	4.9(0.7)	283(666)
MetaMorph-R	69.2(6.7)	**3.4(0.57)**	**3.1(0.55)**	366(594)

Table 5.4: Evaluation of the methods on BraTS-Reg. AE stands for absolute error. Bold numbers indicate the best scores.

Method	Median AE	Mean AE	Robustness	$\lvert J_\phi\rvert < 0$
Initial	7.8(5.6)	8.41(5.5)	0(0)	-
Rigid	5.2(3.3)	6.03(3.3)	0.65(0.32)	**0 (0)**
SyN	4.6(2.7)	5.4(2.9)	0.71(0.29)	**0 (0)**
VM	4.9(3.3)	5.8(3.3)	0.71(0.27)	50.9(227)
Rigid-CFM	4.7(2.7)	5.6(2.8)	0.66(0.32)	**0 (0)**
SyN-CFM	4.5(2.7)	5.4(2.9)	0.70(0.3)	**0 (0)**
VM-CFM	4.8(3.2)	5.8(3.3)	0.71(0.28)	36.3(162)
MetaMorph-R	**2.8(0.81)**	**3.8(1.5)**	**0.78(0.24)**	2.7(9.0)

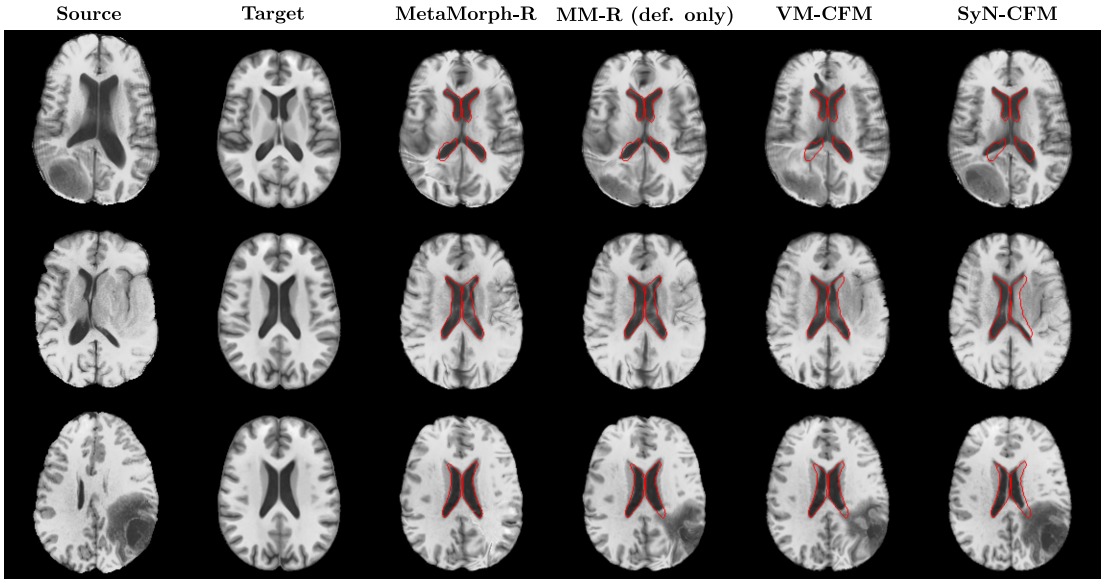


Figure 5.5: Results of registration on BraTS 2021 dataset for our method (deformation only and total transformation), Voxelmorph (VM), and symmetric normalization with cost function masking (SyN-CFM). The blue line on the source images delineates the tumor zone (tumor core + edema) which is used as a mask. The red line delineates the ventricles of the target image. The Dice score of each method is written in the top right corner of each image.

5.4 Discussion

5.4.1 Learning curve

We found that MetaMorph-R is surprisingly fast at learning the deformations in terms of the number of epochs. As can be seen in Figure 5.6, after only one epoch the Dice score is 0.65. For comparison, the initial Dice score is 0.30 and the Dice score obtained by Voxelmorph and MetaMorph-G after one epoch is only 0.45. We suspect that this fast learning curve is due to the computation of the velocity fields (Equation 5.2a). Indeed, in contrast to the other methods, v_t is directly computed by the neural network given I_t which makes it easier for the model to adjust the direction of the deformation. For the two other methods, the only variable of adjustment is z_0 which might make them less stable. As a result, our model is trained on fewer epochs than Voxelmorph and MetaMorph-G.

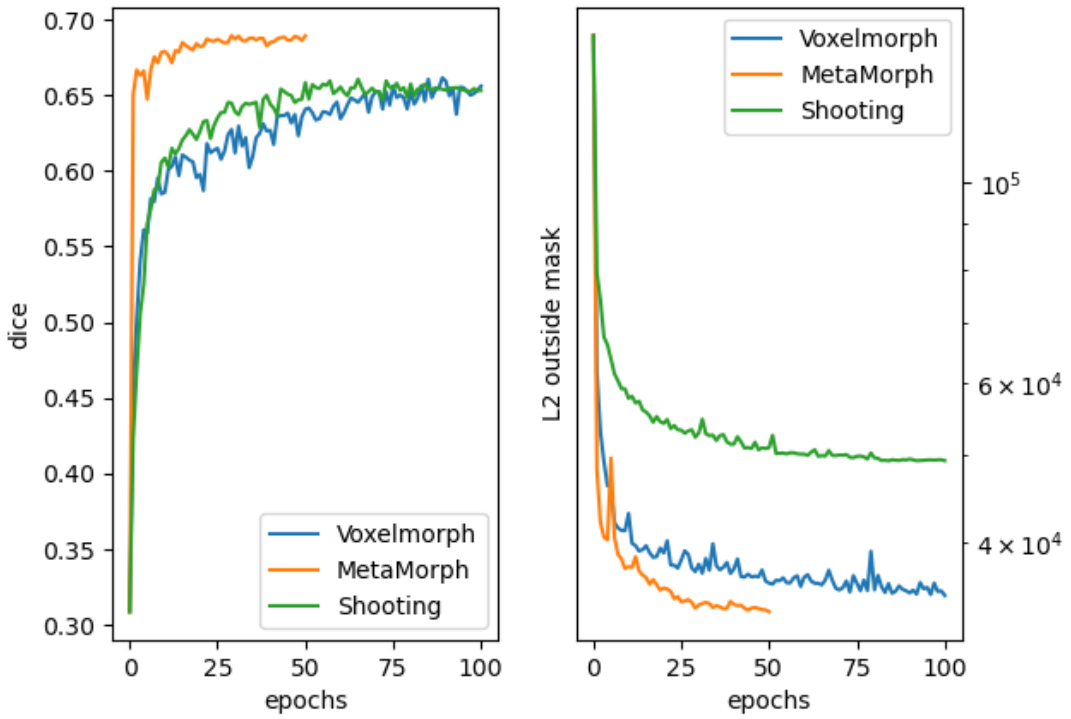


Figure 5.6: Evolution of the Dice score and $\mathcal{L}_2$ distance outside of the mask over the epochs on BraTS 2021 dataset.

5.4.2 Shape and Appearance disentanglement

As visible in Figure 5.7, the introduction of the local regularization makes the disentanglement between shape and appearance transformations easier. Indeed, for the 3 values of λ and μ, the results without masking are very different. For a high value of λ, the intensity transformation is nonexistent whereas

the shape deformation is too small to properly align the images. For lower values, the appearance transformation modifies the shape of the images, to the point where the total transformation is almost perfectly the target image. This means that the model essentially learned the intensity differences between the source and target images. On the other hand, the masked method obtains better results for the 3 hyper-parameters - only with λ =3e-5, the result is less satisfying since the geometric deformation is too constrained. Similar results are obtained with the variations of μ. When the appearance transformation is boosted (higher values of μ), it overtakes the shape deformation.

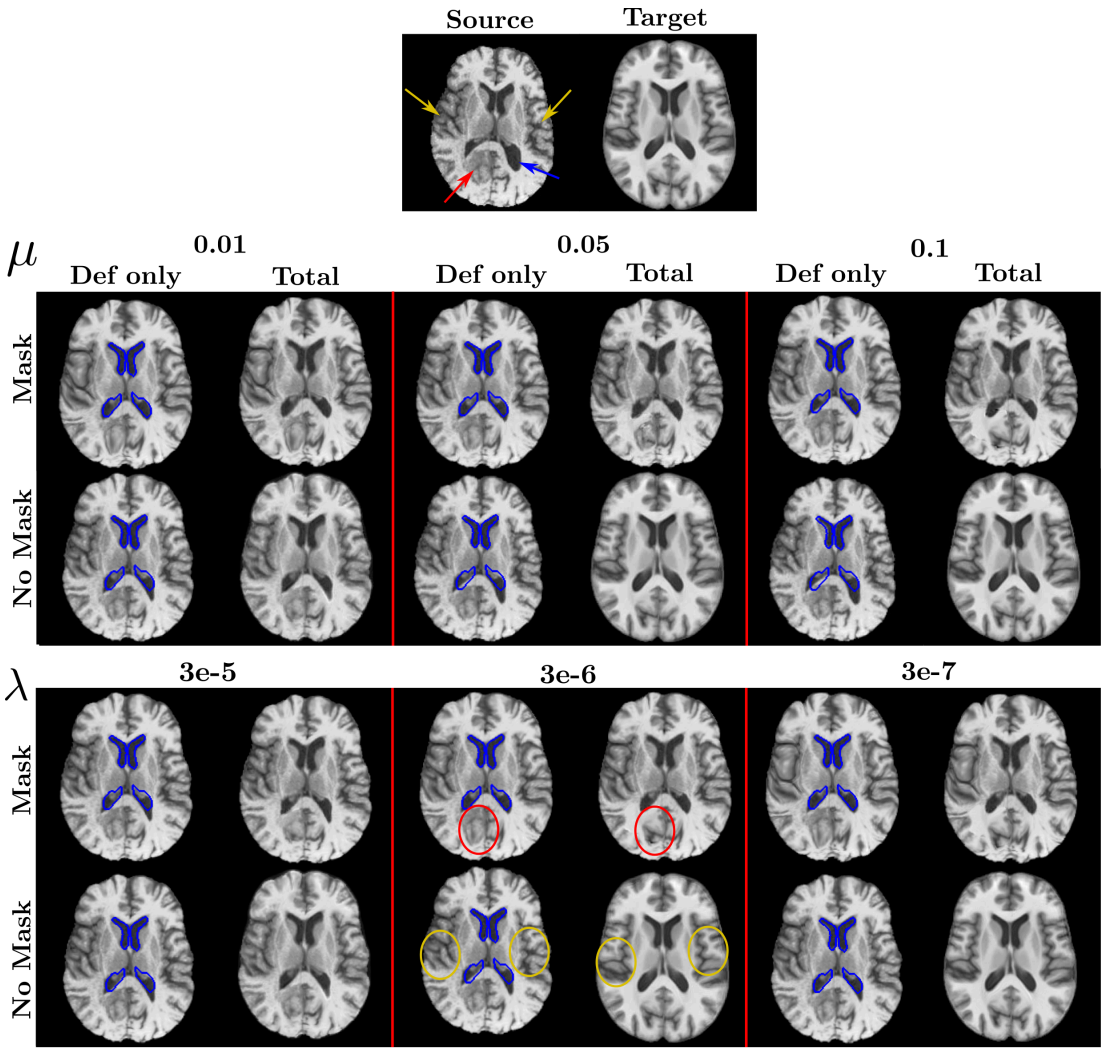


Figure 5.7: Deformation and total transformation for the masked and non-masked versions of MetaMorph-R with 3 values of μ (first two rows) and λ (last two rows). When μ is varying, λ = 3e-6 and when λ is varying μ = 0.04. Red arrow indicates the tumor, blue arrow shows the ventricle that is incorrectly aligned without masking (manual segmentation in blue), yellow arrows show healthy tissues that are incorrectly modified by the appearance transformation without masking.

The problem of disentangling shape and appearance changes is not specific to out method, it is also the case for MAE (Bône et al., 2020). In Figure 5.8, we show the results when varying the parameter λ_s that controls the regularization of the transformation. In this case, a slight change of the hyper-parameter value modifies the deformation from a smooth one to a very irregular one. Furthermore, with this method, the localization of the appearance transformation isn't precise at all.

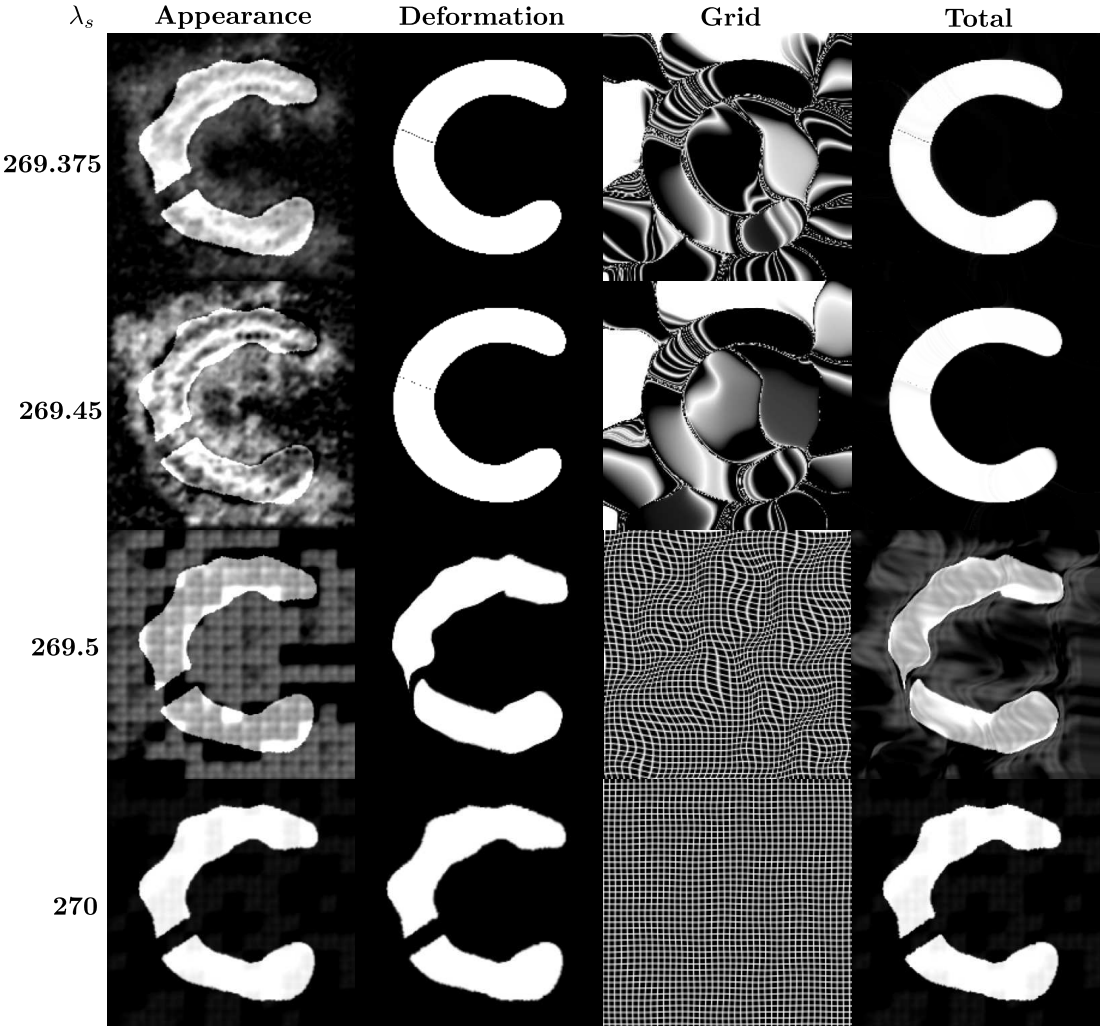


Figure 5.8: Results of Metamorphic Auto-Encoders for several values of λ_s, the hyper-parameter controlling the disentanglement between shape and appearance modifications.

5.4.3 Invertibility

The proposed method is invertible *i.e.* by computing the forward transformation from the source to the target image, one can also estimate the inverse transformation from the target to the source space. Figure 5.9 shows the forward transformation of an image and then its composition with the inverse. As expected for a diffeomorphic transformation, the estimated inverse image

is almost the same as the source (with the difference mainly caused by the bilinear interpolation).

An interesting phenomenon during the forward transformation is the intensity modification in the tumor zone. Intuitively, at each time step, one would expect the model to add the difference between source and target voxels divided by the number of time steps. However, this is not the case as can be seen in Figure 5.9 where the model modifies the intensity a lot from $t = 0$ to $t = 8$ and then reduces it until the final time step $t = 15$. This suggests that our model does not fully minimize the geodesic path between both images. Additionally, this highlights that MetaMorph-R does not constitute a tumor growth model (or in our case, a tumor shrinking model), and actually, it was not intended to. The movement of the tumor mask during registration is not based on a physical model. Adding a physics-based growth model could be a potential improvement to our method, however, it requires to have access to the tissue segmentation of the whole brain.

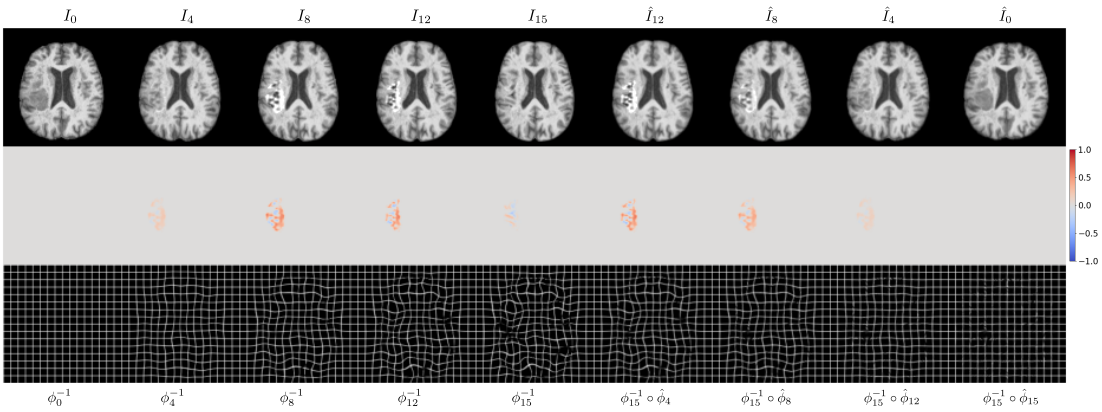


Figure 5.9: Evolution of the forward transformation at different time points followed by its composition with the estimated backward transformation. The bottom row shows the geometric deformation of a grid over time, demonstrating a return to identity when the forward and backward deformations are composed. Similarly, the second row illustrates the backward and forward cumulative intensity addition.

5.5 Conclusion

MetaMorph-R outperforms state-of-the-art methods on the three datasets. Qualitative and quantitative results indicate that it better aligns healthy tissues of the brain than the classical cost function masking strategy. The method applies to 2D and 3D gray-scale images and is invertible.

Our method has been conceived for a clinical context, where a single modality is usually available. The generalization to multiple modalities is not so trivial. This arises from the fact that the residuals z constitute the intensity modification but also generate the velocity fields v. In a multi-modal context, one

would like the appearance change to be specific to each modality (*i.e.* different z for each modality) but the shape deformation to be the same (*i.e.* same v for every modality). To solve that, one would need to generate one velocity field from the multiple residuals, changing therefore the cost function and the geodesic equations. In the appendix A.1, we show a first lead into building multi-modal metamorphosis.

Finally, the method is not specific to a certain imaging modality or anatomical location. In this paper, we showed that it works with synthetic data, T1, and T1ce MRI scans on the brain. In addition, it could also be used with other modalities, such as CT or PET, and with pathological images of other anatomical areas, such as the abdomen (La Barbera et al., 2021).

6

Conclusions and Perspectives

Contents

The goal of this thesis was to build an atlas of glioblastoma. As we have seen in the introduction the concept of glioblastoma atlas is still not well-defined. Nevertheless, we determined that a key step in doing so is to build a transformation function between images with different topologies. We proposed to compute the segmentation mask of the tumor and use it as a topological prior in the transformation function. We constrain the frameworks to be usable on a single modality to make it applicable in a clinical context. The main challenges of the thesis were therefore to: 1) improve the segmentation algorithms working on a unique modality, and 2) design the transformation method using the segmentation masks. Finally, the last challenge was to develop methods with little computation time in order to use them for the construction of the atlas. In the next section, we discuss our contributions to solve these challenges.

6.1 Conclusions

Knowledge transfer:

Prior to this PhD, existing work on the topic of segmentation with missing modality focused on two strategies: synthesizing the missing modality or learning a shared representation space. However, the first strategy requires extensive computations while the second does not perform well when the missing modality is fixed. Therefore, we proposed to distill the knowledge of a multi-modal teacher network into a mono-modal student model. The method is largely inspired by knowledge distillation (Hinton et al., 2015) which was originally designed for model compression. We found promising results on BraTS 2018 and presented them at the MICCAI 2020 conference (Hu* et al., 2020). Since then, other methods based on the teacher-student framework have been proposed (Chen et al., 2022; Vadacchino et al., 2021; Rahimpour et al., 2022). Later during the PhD, BraTS 2021 dataset has been released with more than 1200 annotated subjects. Therefore, in Chapter 3, we evaluted our method on BraTS 2021 and compared it with other techniques which include attention

transfer and contrastive distillation. We found that these strategies were able to improve the results of the student network only when the training data was limited. With a larger training set, the benefit of using knowledge transfer strategies was less clear. For instance, in terms of Dice score, the results did not significantly improve. The methods proposed by Chen et al. (2022); Vadacchino et al. (2021) and Rahimpour et al. (2022) were only tested on BraTS 2018 but we believe a similar phenomenon would occur with more training data since they report improvements in a similar range as ours on BraTS 2018.

For the task of brain tumor segmentation, using current teacher-student strategies is therefore probably not beneficial since BraTS dataset contains enough subjects to properly train a mono-modal network. However, BraTS is a bit of an exception that required the collaboration of multiple international institutions and the manual annotation of more than 50 experts. Not every anatomical region has benefited from such attention. For instance, in myocardial pathology segmentation, the only publicly available multi-modal dataset contains only 45 annotated subjects (Li et al., 2022). For this type of applications, where only a small database of annotated images is available, the current teacher-student knowledge distillation approach should be beneficial.

Registration with varying morphology/appearance/topology:

The second part of this thesis consisted in designing a registration method that deals with images with morphological, appearance and topological differences. Currently, the most commonly used method consists in masking the tumor region when computing the cost function (Brett et al., 2001; Stefanescu et al., 2004). However, it has been proven that it does not work as intended with large tumors (Kim et al., 2007). Other strategies such as tumor growth models (Zacharaki et al., 2009; Gooya et al., 2012; Gholami et al., 2017; Scheufele et al., 2019) and healthy image synthesis (Liu et al., 2015; Yang et al., 2016) have been proposed. Nevertheless, they either require extensive computations or the disentanglement between the shape and appearance change might not be guaranteed.

We proposed two approaches based on the Metamorphosis framework (Trouvé and Younès, 2005) that allow for the modification of both the geometry and the intensity levels of an image. The first one, called MetaMorph-G, consists in using a backbone CNN (such as U-Net) to predict an initial momentum and from it, using the geodesic equations of Metamorphosis, to compute the shape and appearance transformations. The second method, MetaMorph-R, inspired by Amor et al. (2021), uses a ResNet-like architecture to retrieve the residuals of the transformation and subsequently the deformation. For both methods, we employ the segmentation mask of the tumor to limit the appearance change only to the tumor region. We found that MetaMorph-R reaches better results than state-of-the-art methods for both 2D and 3D data. Furthermore, since it is incorporated in a learning framework, at inference, the execution time is less than a second. In this way, this registration strategy is well adapted for atlas construction. The main drawback of our method is the GPU memory requirements. Namely, during training it requires 30 GB of VRAM which only

a few GPUs verify. Nevertheless, once it is trained, the model requires 10 GB, allowing for the use on more accessible GPU cards.

6.2 Perspectives

In this section, we discuss the perspectives of the work presented in this thesis and the issues that they raise.

Knowledge transfer:

In this thesis, the knowledge transfer methods have only been tested in the context of brain tumor segmentation with missing modalities. It would be interesting to experiment on different data types and different applications. It has been applied to myocardial pathology segmentation by Chen et al. (2022). However, as stated before, it is a small database. Therefore, we cannot test if the increase in the segmentation Dice score vanishes with more training data. Unfortunately, we did not find another large multi-modal segmentation database to test the knowledge transfer strategies on another data type.

In this work we only evaluated segmentation networks, but knowledge transfer methods can also be used on other tasks such as classification. Experimenting if the same decrease occurs with more training data for classification would be interesting. This could be done on the BraTS 2021 dataset. Indeed, a second task of the challenge is the prediction of the MGMT promoter methylation status which is an important biomarker for the prediction of the outcome (Weller et al., 2010).

Registration with varying topology:

One of the drawbacks of our registration method is that it is only applicable on a single modality. On the one hand, this is useful for a clinical context where only one modality is acquired. But on the other hand, when a public multi-modal database such as BraTS-Reg is available, being able to process the multiple modalities simultaneously would produce a more accurate registration. Furthermore, using the multi-modal images allows to predict more precise tumor segmentation masks.

The metamorphosis framework could be extended to multimodal data by considering the image I_t as a function from $\Omega \subset \mathbb{R}^d$ to $\mathbb{R}^C$ with C being the number of modalities. Then, we set $z_t : \Omega \subset \mathbb{R}^d \to \mathbb{R}^C$, and the velocity-fields v_t are the same as for Metamorphosis on single data. In this case, the images are transformed with the equation:

$$\frac{\partial I_t}{\partial t} = -\langle DI_t^T, v_t \rangle + \mu^2 z_t. \tag{6.1}$$

Indeed, with this formulation, we have that, for each modality c, the deformation of image I_t^c is induced by the common velocity field v_t and the intensity transformation is specific to that modality with z_t^c.

Furthermore, the norm of the residual becomes:

$$\|z_t\|_2 = \sqrt{\int_\Omega \langle z_t(x), z_t(x)\rangle dx} = \sqrt{\int_\Omega z_t(x)^T z_t(x) dx}.$$

We define the energy of the transformation as $E(I, v) = \int_0^1 \|v_t\|_V^2 + \|\mu z_t\|_2^2 dt$. It is important to note that if we set $C = 1$, we get the same equations as in Chapter 5.

Using Equation 6.1, and the energy of the transformation, we can compute the Euler-Lagrange equations of $E(\cdot, v)$ and $E(I, \cdot)$ to retrieve the geodesic equations of multi-modal Metamorphosis (see Appendix A.1 for the proof):

$$\begin{cases} v_t = -K(DI_t^T z_t) & \text{(6.2a)} \\ \dfrac{\partial z_t}{\partial t} = -div(z_t v_t^T) & \text{(6.2b)} \\ \dfrac{\partial I_t}{\partial t} = -\langle DI_t^T, v_t\rangle + \mu^2 z_t. & \text{(6.2c)} \end{cases}$$

where K refers to the kernel operator and DI_t the jacobian of I_t. These equations could then be used similarly as we did in Chapter 5, by using a U-Net or ResNet architecture. Unfortunately, we did not have the time to code and test this system of equations.

Atlas construction:

The original aim of this thesis was to build an atlas of glioblastoma. The proposed registration method could be used to compute a frequentist atlas of glioblastoma. This would require to first register a set of pathological brain images onto a common healthy template. Subsequently, we would fetch the tumor masks in the template space and compute the spatial probabilities in that space. This type of atlas has already been proposed with different registration methods such as cost function masking (Roux et al., 2019) or tumor growth model (Gooya et al., 2012). From this frequentist atlas, it is also possible to derive several statistical maps based on the patients characteristics such as age, sex, or outcome. Although computing such maps are informative, they ignore several crucial characteristics such as the size, the shape and the tissue subdivision (necrotic tumor, enhancing tumor, etc.) which could give important information for understanding the behavior of glioblastoma.

Computing an atlas that takes the previous characteristics into account is still not well-defined. Indeed, a common atlas representation consists in an average scan and a set of image-specific transformations. However, for tumors with very different positions int the brain, the concept of average representation is not clear. A tumor in the frontal lobe and a tumor in the occipital lobe cannot stem from the same average tumor representation. A potential lead could be to compute an atlas per brain region. In this way, only tumors with similar positions would be used to compute the every average representation. This type of atlas would allow computing more thorough statistics than the frequentist

atlas. Namely, one could perform statistical analyses on the shape *and* appearance transformations, such as, for instance, the average tumor growth given the position.

In order to do so, we would probably require to adapt our registration method. Indeed, the appearance change determined by our method is not constrained by a biophysical model. It does not contain any prior knowledge about brain tumors or the surrounding tissue. Using our methods, would therefore probably not produce meaningful statistics about the appearance transformation. For this reason, future work on MetaMorph should focus on incorporating a biophysical model into the geodesisc equations. This could be first experimented using the simple tumor growth model presented by Scheufele et al. (2021). In this case, the evolution of the tumor mask would be set by the tumor growth model and the residual inside the mask as well. Further work could focus on more complex tumor growth models such as the ones introduced by Scheufele et al. (2019) or Subramanian et al. (2023).

Tumor growth models require the full segmentation of healthy tissues into white matter, grey matter and cerebrospinal fluid regions. This process is tedious in the presence of a tumor. A straightforward technique would be to train a U-Net model on fully segmented brain images with tumors. However, we did not find any brain dataset with such annotations. A method on 2D data (Gholami et al., 2019) proposes to compute the full segmentation of the image of a cancerous brain. However, it is done by overlaying the segmentation of a healthy atlas on the cancerous image with a classical registration method. This might produce an inaccurate segmentation of the healthy tissues. Therefore, a future work would be to develop a method able to fully segment a brain with a tumor, which would be used as a preliminary step for the registration method with a biophysical model.

A
Appendix

Contents

A.1 Multi-modal Metamorphosis

Let $I_t : \mathbb{R}^d \to \mathbb{R}^C$, $z_t : \mathbb{R}^d \to \mathbb{R}^C$ and $v \in V$ with C being the number of modalities and d the dimension of the image. The evolution of the image follows the equation:

$$\frac{\partial I_t}{\partial t} = -\langle DI_t^T, v_t \rangle + \mu^2 z_t. \tag{A.1}$$

The energy of the transformation is:

$$E(I, v) = \int_0^1 \|v_t\|_V^2 + \|\mu z_t\|_2^2 dt. \tag{A.2}$$

Then, the geodesics associated with energy A.2 are:

$$\begin{cases} v_t = -K(DI_t^T z_t) & \text{(A.3a)} \\ \dfrac{\partial z_t}{\partial t} = -div(z_t v_t^T) & \text{(A.3b)} \\ \dfrac{\partial I_t}{\partial t} = -\langle DI_t^T, v_t \rangle + \mu^2 z_t. & \text{(A.3c)} \end{cases}$$

Proof. This proof is similar to the one by François et al. (2022) but is adapted for the multi-modal case. For simplification, we replace $\frac{\partial I}{\partial t}$ with $\dot{I}$ and $\frac{\partial v}{\partial t}$ with $\dot{v}$. We first consider the Lagrangian $L_v(t, v, \dot{v}) = E(I, v)$. Using $z_t = \frac{1}{\mu^2}(\dot{I}_t + \langle DI_t^T, v_t \rangle)$, we compute the variations h with respect to v:

$$\begin{aligned} L_v(t, v+h, \dot{v}) - L_v(t, v, \dot{v}) &= 2 \int_0^1 \langle v_t, h_t \rangle_V + \mu^2 \langle z_t, \frac{1}{\mu^2} DI_t h_t \rangle dt + o(\|h\|_2) \\ &= 2 \int_0^1 \langle K^{-1} v_t + DI_t^T z_t, h_t \rangle dt + o(\|h\|_2) \end{aligned}$$

Therefore, we get $\nabla_v L_v = 2(Kv_t + DI_t^T z_t)$. Since $\nabla_{\dot{v}} L_v = 0$, by computing the Euler-Lagrange equation of L_v: $\nabla_v L_v - \frac{d}{dt}\nabla_{\dot{v}} L_v = 0$, we obtain Equation A.3a:

$$v_t = -K(DI_t^T z_t)$$

To retrieve Equation A.3b, we compute the variations of the Lagrangian $L_I(t, I, \dot{I}) = E(I, v)$. First, we get the variations for I:

$$\begin{aligned}
L_I(t, I+h, \dot{I}) - L_v(t, I, \dot{I}) &= 2\int_0^1 \mu^2 \langle z_t, \frac{1}{\mu^2} Dh_t v_t \rangle dt + o(\|h\|_2) \\
&= 2\int_0^1 \int_\Omega \sum_{c=1}^{C} z_t^c(x) \sum_{i=1}^{d} \frac{\partial h_t^c}{\partial x_i}(x) v_t^i(x) dx dt + o(\|h\|_2) \\
\text{with integration by parts,} \quad &= -2\int_0^1 \int_\Omega \sum_{c=1}^{C} h_t^c(x) \sum_{i=1}^{d} \frac{\partial z_t^c v_t^i}{\partial x_i}(x) dx dt + o(\|h\|_2) \\
\text{and since } v_t \text{ vanishes on } \partial\Omega \quad &= -2\int_0^1 \int_\Omega \sum_{i=1}^{d} \left(\frac{\partial z_t v_t^i}{\partial x_i}(x) \right)^T h_t(x) dx dt + o(\|h\|_2) \\
&= -2\int_0^1 \langle div(z_t v_t^T), h_t \rangle dt + o(\|h\|_2)
\end{aligned}$$

Thus, we get $\nabla_I L_I = -2div(z_t v_t^T)$. Similarly, we obtain $\nabla_{\dot{I}} L_I = 2z_t$. Thus, when computing the Euler-Lagrange equation, we get:

$$\frac{\partial z_t}{\partial t} = -div(z_t v_t^T)$$

and therefore, we retrieve the geodesic equations A.3.

Bibliography

H. Akbari, L. Macyszyn, X. Da, M. Bilello, et al. Imaging surrogates of infiltration obtained via multiparametric imaging pattern analysis predict subsequent location of recurrence of glioblastoma. *Neurosurgery*, 78(4):572–580, 2016. page 78

C. Alifieris and D. T. Trafalis. Glioblastoma multiforme: Pathogenesis and treatment. *Pharmacology Therapeutics*, 152:63–82, 2015. page 13

M. Almansour, N. M. Ghanem, and S. Bassiouny. High-resolution mri brain inpainting. In *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–6, 2021. page 79

B. B. Amor, S. Arguillère, and L. Shao. ResNet-LDDMM: Advancing the LDDMM Framework Using Deep Residual Networks. *CoRR*, 2021. pages 84, 102

L. Amoruso, S. Geng, N. Molinaro, P. Timofeeva, S. Gisbert-Muñoz, S. Gil-Robles, I. Pomposo, I. Quiñones, and M. Carreiras. Oscillatory and structural signatures of language plasticity in brain tumor patients: A longitudinal study. *Human Brain Mapping*, 42(6):1777–1793, 2021. page 22

V. Arsigny, O. Commowick, X. Pennec, and N. Ayache. A log-euclidean framework for statistics on diffeomorphisms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 924–931. Springer, 2006. pages 71, 73

J. Ashburner. A fast diffeomorphic image registration algorithm. *NeuroImage*, 38(1):95–113, Oct. 2007. pages 71, 77

J. Ashburner and K. J. Friston. Rigid body registration. *Statistical parametric mapping: The analysis of functional brain images*, pages 49–62, 2007. page 64

J. Ashburner and K. J. Friston. Diffeomorphic registration using geodesic shooting and Gauss–Newton optimisation. *NeuroImage*, 55(3):954–967, Apr. 2011. pages 68, 70, 78

B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1): 26–41, 2008. pages 71, 78, 89

B. B. Avants, N. Tustison, G. Song, et al. Advanced normalization tools (ants). *Insight Journal*, 2(365):1–35, 2009. pages 71, 90

P. Bachman, R. D. Hjelm, and W. Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019. page 43

B. Baheti, D. Waldmannstetter, S. Chakrabarty, et al. The brain tumor sequence registration challenge: Establishing correspondence between pre-operative and follow-up mri scans of diffuse glioma patients, 2021. page 88

U. Baid, S. Ghodasara, et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification, 2021. pages 22, 88

R. Bajcsy and S. Kovačič. Multiresolution elastic matching. *Computer Vision, Graphics, and Image Processing*, 46(1):1–21, 1989. page 67

S. Bakas, H. Akbari, et al. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci Data*, 2017. pages 22, 88

G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, , and A. V. Dalca. VoxelMorph: A Learning Framework for Deformable Medical Image Registration. *IEEE TMI*, 38(8):1788–1800, Aug. 2019. pages 73, 78, 90

K. Batenburg and J. Sijbers. Adaptive thresholding of tomograms by projection distance minimization. *Pattern Recognition*, 42(10):2297–2305, 2009. Selected papers from the 14th IAPR International Conference on Discrete Geometry for Computer Imagery 2008. page 22

M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes. Computing Large Deformation Metric Mappings via Geodesic Flows of Diffeomorphisms. *IJCV*, 61(2): 139–157, 2005. pages 17, 68, 70, 77

R. Behbehani, H. Adnan, A. A. Al-Hassan, A. Al-Salahat, and R. Alroughani. Predictors of retinal atrophy in multiple sclerosis: a longitudinal study using spectral domain optical coherence tomography with segmentation analysis. *Multiple Sclerosis and Related Disorders*, 21:56–62, 2018. page 22

A. Ben-Cohen, E. Klang, S. Raskin, S. Soffer, S. Ben-Haim, E. Konen, M. Amitai, and H. Greenspan. Cross-modality synthesis from CT to PET using FCN and GAN networks for improved automated lesion detection. *Engineering Applications of Artificial Intelligence*, 78:186–194, 2018. page 37

A. Bône, P. Vernhet, O. Colliot, and S. Durrleman. Learning joint shape and appearance representations with metamorphic auto-encoders. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 202–211. Springer, 2020. pages 78, 80, 90, 98

M. Brett, A. P. Leff, C. Rorden, and J. Ashburner. Spatial Normalization of Brain Images with Focal Lesions Using Cost Function Masking. *NeuroImage*, 14(2):486–500, Aug. 2001. pages 78, 79, 102

X. Cao, J. Yang, J. Zhang, Q. Wang, P.-T. Yap, and D. Shen. Deformable image registration using a cue-aware deep regression network. *IEEE Transactions on Biomedical Engineering*, 65(9):1900–1911, 2018. page 72

A. Carré, G. Klausner, M. Edjlali, M. Lerousseau, J. Briend-Diop, R. Sun, S. Ammari, S. Reuzé, E. Alvarez Andres, T. Estienne, et al. Standardization of brain mr images across machines and protocols: bridging the gap for mri-based radiomics. *Scientific reports*, 10(1):1–15, 2020. pages 28, 29

T. F. Chan and L. A. Vese. Active contours without edges. *IEEE Transactions on image processing*, 10(2):266–277, 2001. page 22

J. Chang, X. Zhang, J. Chang, M. Ye, D. Huang, P. Wang, and C. Yao. Brain tumor segmentation based on 3d unet with multi-class focal loss. In *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–5, 2018. page 30

C. Chen, Q. Dou, Y. Jin, H. Chen, J. Qin, and P.-A. Heng. Robust Multimodal Brain Tumor Segmentation via Feature Disentanglement and Gated Fusion. In *MICCAI*, volume LNCS 11766, pages 447–456, Cham, 2019. Springer. page 39

C. Chen, Q. Dou, Y. Jin, Q. Liu, and P. A. Heng. Learning with privileged multimodal knowledge for unimodal segmentation. *IEEE Transactions on Medical Imaging*, 41(3):621–632, 2022. pages 41, 44, 52, 101, 102, 103

L. Chen, D. Wang, Z. Gan, J. Liu, R. Henao, and L. Carin. Wasserstein contrastive representation distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16296–16305, 2021. pages 43, 44

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. pages 43, 44

Y. Cho and S. Kang. Class attention transfer for semantic segmentation. In *2022 IEEE 4th International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, pages 41–45, 2022. page 42

G. Christensen, R. Rabbitt, and M. Miller. Deformable templates using large deformation kinematics. *IEEE Transactions on Image Processing*, 5(10):1435–1447, 1996. page 67

G. E. Christensen, R. D. Rabbitt, and M. I. Miller. 3d brain mapping using a deformable neuroanatomy. *Physics in Medicine & Biology*, 39(3):609, 1994. page 67

D. Christos. Spatial transformation and registration of brain images using elastically deformable models. *Comput Vis Image Underst.*, 66(2):207–22, 1997. page 67

I. Chung, S. Park, J. Kim, and N. Kwak. Feature-map-level online adversarial knowledge distillation. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2006–2015. PMLR, 13–18 Jul 2020. page 44

Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016. page 26

L. D. Cohen. On active contour models and balloons. *CVGIP: Image understanding*, 53(2):211–218, 1991. page 22

G. B. Coleman and H. C. Andrews. Image segmentation by clustering. *Proceedings of the IEEE*, 67(5):773–785, 1979. page 22

S. Czolbe, A. Feragen, and O. Krause. Spot the Difference: Detection of Topological Changes via Geometric Alignment. In *NeurIPS*, 2021. page 80

L. B. da Cruz, D. A. D. Júnior, J. O. B. Diniz, A. C. Silva, J. D. S. de Almeida, A. C. de Paiva, and M. Gattass. Kidney tumor segmentation from computed tomography images using deeplabv3+ 2.5 d model. *Expert Systems with Applications*, 192:116270, 2022. page 25

A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical image analysis*, 57:226–236, 2019. pages 73, 83

P. C. De Witt Hamer, E. J. Hendriks, E. Mandonnet, F. Barkhof, A. H. Zwinderman, and H. Duffau. Resection probability maps for quality assessment of glioma surgery without brain location bias. *PloS one*, 8(9):e73353, 2013. page 78

N. S. Detlefsen, O. Freifeld, and S. Hauberg. Deep Diffeomorphic Transformer Networks. In *CVPR*, pages 4403–4412. IEEE, June 2018. page 73

J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. B. Ayed. Hyperdense-net: a hyper-densely connected cnn for multi-modal image segmentation. *IEEE transactions on medical imaging*, 38(5):1116–1126, 2018. page 27

R. Dorent, S. Joutard, M. Modat, S. Ourselin, and T. Vercauteren. Hetero-Modal Variational Encoder-Decoder for Joint Modality Completion and Segmentation. In *MICCAI*, volume LNCS 11765, pages 74–82. Springer, 2019. pages 38, 39, 50

B. Dufumier, P. Gori, J. Victor, A. Grigis, M. Wessa, P. Brambilla, P. Favre, M. Polosan, C. Mcdonald, C. M. Piguet, et al. Contrastive learning with continuous proxy meta-data for 3d mri classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 58–68. Springer, 2021. page 60

P. Dupuis, U. Grenander, and M. I. Miller. Variational problems on flows of diffeomorphisms for image matching. *Quarterly of applied mathematics*, pages 587–600, 1998. pages 17, 68, 70

P. T. Fletcher, S. Venkatasubramanian, and S. Joshi. The geometric median on riemannian manifolds with application to robust atlas estimation. *NeuroImage*, 45(1):S143–S152, 2009. page 80

M. Fornefett, K. Rohr, and H. Stiehl. Elastic registration of medical images using radial basis functions with compact support. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, volume 1, pages 402–407 Vol. 1, 1999. page 67

M. Fornefett, K. Rohr, and H. Stiehl. Radial basis functions with compact support for elastic registration of medical images. *Image and Vision Computing*, 19(1):87–96, 2001. page 67

J.-P. Fortin, E. M. Sweeney, J. Muschelli, C. M. Crainiceanu, R. T. Shinohara, A. D. N. Initiative, et al. Removing inter-subject technical variability in magnetic resonance imaging studies. *NeuroImage*, 132:198–212, 2016. page 29

A. François, M. Maillard, C. Oppenheim, J. Pallud, I. Bloch, P. Gori, and J. Glaunès. Weighted metamorphosis for registration of images with different topologies. In *Biomedical Image Registration: 10th International Workshop, WBIR 2022, Munich, Germany, July 10–12*, pages 8–17. Springer, 2022. page 107

A. François, P. Gori, and J. Glaunès. Metamorphic image registration using a semi-Lagrangian scheme. In *GSI*, 2021. pages 80, 81, 82, 83

C. García and J. A. Moreno. Kernel based method for segmentation and modeling of magnetic resonance images. In *Advances in Artificial Intelligence–IBERAMIA 2004: 9th Ibero-American Conference on AI, Puebla, Mexico, November 22-26, 2004. Proceedings 9*, pages 636–645. Springer, 2004. page 24

L. Garcin and L. Younes. Geodesic image matching: A wavelet based energy minimization scheme. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 349–364. Springer, 2005. page 80

Y. George. A coarse-to-fine 3d u-net network for semantic segmentation of kidney ct scans. In *Kidney and Kidney Tumor Segmentation: MICCAI 2021 Challenge, KiTS 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings*, pages 137–142. Springer, 2022. page 27

A. Gholami, A. Mang, K. Scheufele, C. Davatzikos, M. Mehl, and G. Biros. A Framework for Scalable Biophysics-based Image Analysis. In *ACM International Conference for High Performance Computing, Networking, Storage and Analysis, SC '17*, pages 19:1–19:13, New York, USA, 2017. ISBN 978-1-4503-5114-0. pages 79, 102

A. Gholami, S. Subramanian, V. Shenoy, N. Himthani, X. Yue, S. Zhao, P. Jin, G. Biros, and K. Keutzer. A novel domain adaptation framework for medical image segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4*, pages 289–298. Springer, 2019. page 105

M. Goetz, C. Weber, J. Bloecher, B. Stieltjes, H.-P. Meinzer, and K. Maier-Hein. Extremely randomized trees based brain tumor segmentation. *Proceeding of BRATS challenge-MICCAI*, 14:6–11, 2014. page 24

A. Golts, D. Khapun, D. Shats, Y. Shoshan, and F. Gilboa-Solomon. An ensemble of 3d u-net based models for segmentation of kidney and masses in ct scans. In *Kidney and Kidney Tumor Segmentation: MICCAI 2021 Challenge, KiTS 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings*, pages 103–115. Springer, 2022. page 27

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. page 31

A. Gooya, K. M. Pohl, M. Bilello, L. Cirillo, G. Biros, E. R. Melhem, and C. Davatzikos. GLISTR: Glioma Image Segmentation and Registration. *IEEE TMI*, 31(10):1941–1954, 2012. pages 16, 78, 79, 102, 104

P. Gori, O. Colliot, L. Marrakchi-Kacem, Y. Worbe, et al. A Bayesian framework for joint morphometry of surface and curve meshes in multi-object complexes. *Medical Image Analysis*, 35:458–474, 2017. pages 14, 78

R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. pages 43, 44

X. Han, X. Yang, S. Aylward, R. Kwitt, and M. Niethammer. Efficient registration of pathological images: a joint pca/image-reconstruction approach. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 10–14. IEEE, 2017. page 79

X. Han, Z. Shen, Z. Xu, S. Bakas, H. Akbari, M. Bilello, C. Davatzikos, and M. Niethammer. A deep network for joint registration and reconstruction of images with pathologies. In *International Workshop on Machine Learning in Medical Imaging*, pages 342–352. Springer, 2020a. page 80

X. Han et al. A Deep Network for Joint Registration and Reconstruction of Images with Pathologies. In *MLMI - MICCAI*, pages 342–352, 2020b. ISBN 978-3-030-59861-7. page 78

M. Havaei, N. Guizard, N. Chapados, and Y. Bengio. HeMIS: Hetero-Modal Image Segmentation. In *MICCAI*, volume LNCS 9901, pages 469–477. Springer, 2016. pages 38, 50

M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017. page 25

K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. page 26

K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *CVPR*, pages 770–778. IEEE, 2016. page 78

T. Heinonen, P. Dastidar, H. Eskola, H. Frey, P. Ryymin, and E. Laasonen. Applicability of semi-automatic segmentation for volumetric analysis of brain lesions. *Journal of medical engineering & technology*, 22(4):173–178, 1998. page 21

N. Heller, F. Isensee, K. H. Maier-Hein, X. Hou, C. Xie, F. Li, Y. Nan, G. Mu, Z. Lin, M. Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical image analysis*, 67:101821, 2021. pages 27, 29

D. L. G. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes. Medical image registration. *Physics in Medicine Biology*, 46(3):R1, mar 2001. page 64

G. Hinton, O. Vinyals, and J. Dean. Distilling the Knowledge in a Neural Network. *Deep Learning and Representation Learning Workshop: NIPS 2015*, 2015. pages 3, 4, 40, 101

D. D. Holm, A. Trouve, and L. Younes. The Euler-Poincare theory of Metamorphosis. *arXiv:0806.0870 [cs, nlin]*, June 2008. page 82

M. Hu*, M. Maillard*, Y. Zhang, T. Ciceri, G. La Barbera, I. Bloch, and P. Gori. Knowledge distillation from multi-modal to mono-modal segmentation networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 772–781. Springer, 2020. page 101

Y. Hu, X. Liu, X. Wen, C. Niu, and Y. Xia. Brain tumor segmentation on multimodal mr imaging using multi-level upsampling in decoder. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4*, pages 168–177. Springer, 2019. page 25

H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1055–1059. IEEE, 2020. page 26

Z. Huang, L. Lin, P. Cheng, L. Peng, and X. Tang. Multi-modal brain tumor segmentation via missing modality synthesis and modality-level attention fusion. *arXiv preprint arXiv:2203.04586*, 2022. page 37

N. Ibtehaz and M. S. Rahman. MultiResUNet : Rethinking the U-Net Architecture for Multimodal Biomedical Image Segmentation. *Neural Networks*, 121: 74–87, 2020. pages 26, 27

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR. page 26

F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein. No new-net. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4*, pages 234–244. Springer, 2019a. pages 25, 26, 27, 53

F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein. No New-Net. In *BrainLes - MICCAI Workshop*, volume LNCS 11384, pages 234–244. Springer, 2019b. pages 29, 47, 49

F. Isensee, P. F. Jäger, P. M. Full, P. Vollmuth, and K. H. Maier-Hein. nnu-net for brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part II 6*, pages 118–132. Springer, 2021. pages 26, 29

M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. page 73

Y. Jang, H. Lee, S. J. Hwang, and J. Shin. Learning what and where to transfer. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3030–3039. PMLR, 09–15 Jun 2019. page 42

M. Ji, B. Heo, and S. Park. Show, attend and distill: Knowledge distillation via attention-based feature matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7945–7952, 2021. page 42

Z. Jiang, C. Ding, M. Liu, and D. Tao. Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part I 5*, pages 231–241. Springer, 2020. pages 26, 27, 29

S. Jiji, K. A. Smitha, A. K. Gupta, V. P. M. Pillai, and R. S. Jayasree. Segmentation and volumetric analysis of the caudate nucleus in alzheimer's disease. *European journal of radiology*, 82(9):1525–1530, 2013. page 21

S. Joshi, B. Davis, M. Jomier, and G. Gerig. Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage*, 23:S151–S160, 2004. Mathematics in Brain Imaging. page 14

K. Kamnitsas, W. Bai, E. Ferrante, S. McDonagh, M. Sinclair, N. Pawlowski, M. Rajchl, M. Lee, B. Kainz, D. Rueckert, et al. Ensembles of multiple models and architectures for robust brain tumour segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3*, pages 450–462. Springer, 2018. page 27

M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331, 1988. page 22

R. Kikinis, M. E. Shenton, D. V. Iosifescu, R. W. McCarley, P. Saiviroonporn, H. H. Hokama, A. Robatino, D. Metcalf, C. G. Wible, C. M. Portas, et al. A digital brain atlas for surgical planning, model-driven segmentation, and teaching. *IEEE Transactions on visualization and computer graphics*, 2(3):232–241, 1996. page 21

J. Kim, B. Avants, S. Patel, and J. Whyte. Spatial normalization of injured brains for neuroimaging research: An illustrative introduction of available options. *NCRRN Methodology Papers*, 2007. pages 79, 102

J. Kim, S. Park, and N. Kwak. Paraphrasing complex network: Network compression via factor transfer. *Advances in neural information processing systems*, 31, 2018. page 42

U. Kjems, S. C. Strother, J. Anderson, I. Law, and L. K. Hansen. Enhancing the multivariate signal of [/sup 15/o] water pet studies with a new nonlinear neuroanatomical registration algorithm [mri application]. *IEEE Transactions on Medical Imaging*, 18(4):306–319, 1999. page 67

J. Krebs, T. Mansi, H. Delingette, L. Zhang, F. C. Ghesu, S. Miao, A. K. Maier, N. Ayache, R. Liao, and A. Kamen. Robust non-rigid registration through agent-based action learning. In *Medical Image Computing and Computer Assisted Intervention MICCAI 2017*, pages 344–352, Cham, 2017. Springer International Publishing. page 72

J. Krebs, H. Delingette, B. Mailhé, N. Ayache, and T. Mansi. Learning a probabilistic model for diffeomorphic registration. *IEEE transactions on medical imaging*, 38(9):2165–2176, 2019. pages 73, 74, 83

J. Kybic and M. Unser. Fast parametric elastic image registration. *IEEE Transactions on Image Processing*, 12(11):1427–1442, 2003. page 67

G. La Barbera et al. Automatic size and pose homogenization with Spatial Transformer Network to improve and accelerate pediatric segmentation. In *IEEE ISBI*, pages 1773–1776, 2021. page 100

M. Lacroix, D. Abi-Said, D. R. Fourney, Z. L. Gokaslan, W. Shi, F. DeMonte, F. F. Lang, I. E. McCutcheon, S. J. Hassenbusch, E. Holland, et al. A multivariate analysis of 416 patients with glioblastoma multiforme: prognosis, extent of resection, and survival. *Journal of neurosurgery*, 95(2):190–198, 2001. page 13

C.-H. Lee, M. Schmidt, A. Murtha, A. Bistritz, J. Sander, and R. Greiner. Segmenting brain tumors with conditional random fields and support vector machines. In *Computer Vision for Biomedical Image Applications: First International Workshop, CVBIA 2005, Beijing, China, October 21, 2005. Proceedings 1*, pages 469–478. Springer, 2005. page 24

H. Li and Y. Fan. Non-rigid image registration using self-supervised fully convolutional networks without training data. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1075–1078. page 73

L. Li, F. Wu, S. Wang, X. Luo, C. Martin-Isla, S. Zhai, J. Zhang, Y. Liu, Z. Zhang, M. J. Ankenbrand, et al. Myops: A benchmark of myocardial pathology segmentation combining three-sequence cardiac magnetic resonance images. *arXiv preprint arXiv:2201.03186*, 2022. pages 62, 102

X. Li, G. Luo, and K. Wang. Multi-step cascaded networks for brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part I 5*, pages 163–173. Springer, 2020. page 26

P. Liu, W. Liu, H. Ma, Z. Jiang, and M. Seok. Ktan: knowledge transfer adversarial network. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020. page 44

X. Liu, M. Niethammer, R. Kwitt, N. Singh, M. McCormick, and S. Aylward. Low-Rank Atlas Image Analyses in the Presence of Pathologies. *IEEE TMI*, 34:2583–2591, 2015. pages 78, 79, 102

X. Liu, F. Xing, C. Yang, C.-C. J. Kuo, G. El Fakhri, and J. Woo. Symmetric-constrained irregular structure inpainting for brain mri registration with tumor pathology. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 80–91, 2021. page 79

Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2604–2613, 2019. page 41

D. Lopez-Paz, L. Bottou, B. Schölkopf, and V. Vapnik. Unifying distillation and privileged information. In *ICLR*, 2016. pages 16, 40, 49

M. Lorenzi, N. Ayache, G. Frisoni, and X. Pennec. LCC-Demons: A robust and accurate symmetric diffeomorphic registration algorithm. *NeuroImage*, 81: 470–483, Nov. 2013. pages 22, 72

H. M. Luu and S.-H. Park. Extending nn-unet for brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 173–186, Cham, 2022. Springer International Publishing. pages 26, 27, 29

O. Maier et al. ISLES 2015 - A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. *Medical Image Analysis*, 35: 250–269, 2017. page 35

M. Maillard, I. Bloch, and P. Gori. Recalage métamorphique d'images par réseau de neurones résiduels. In *Groupe de Recherche et d'Etudes de Traitement du Signal et des Images (GRETSI)*, 2022a.

M. Maillard, A. François, J. Glaunès, I. Bloch, and P. Gori. A deep residual learning implementation of metamorphosis. In *IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–4, 2022b.

M. Maillard, I. Bloch, and P. Gori. Metamorph: Learning-based metamorphic registration of pathological images. *Submitted*, 2023.

R. McKinley, R. Wepfer, T. Gundersen, F. Wagner, A. Chan, R. Wiest, and M. Reyes. Nabla-net: A deep dag-like convolutional architecture for biomedical image segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Second International Workshop, BrainLes*

2016, with the Challenges on BRATS, ISLES and mTOP 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Revised Selected Papers 2, pages 119–128. Springer, 2016. page 25

B. H. Menze et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015. pages 14, 22, 35, 50, 88

M. I. Miller, A. Trouvé, and L. Younes. Geodesic shooting for computational anatomy. *Journal of mathematical imaging and vision*, 24(2):209–228, 2006. pages 68, 70

A. Y. G. U. Miller MI, Christensen GE. Mathematical textbook of deformable neuroanatomies. *Proc Natl Acad Sci U S A.*, (24), 1993. page 67

F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. pages 26, 30

J. Mitra, P. Bourgeat, J. Fripp, S. Ghose, S. Rose, O. Salvado, A. Connelly, B. Campbell, S. Palmer, G. Sharma, et al. Lesion segmentation from multimodal mri using random forest following ischemic stroke. *NeuroImage*, 98: 324–335, 2014. page 24

T. C. Mok and A. Chung. Fast symmetric diffeomorphic image registration with convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4644–4653, 2020a. pages 73, 74

T. C. W. Mok and A. C. S. Chung. Large Deformation Diffeomorphic Image Registration with Laplacian Pyramid Networks. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 211–221. Springer International Publishing, 2020b. ISBN 978-3-030-59716-0. page 74

D. B. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on pure and applied mathematics*, 1989. page 22

A. Myronenko. 3D MRI Brain Tumor Segmentation Using Autoencoder Regularization. In A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, volume 11384 of *Lecture Notes in Computer Science*, pages 311–320, Cham, Jan. 2019. Springer International Publishing. ISBN 978-3-030-11726-9. pages 25, 27, 29

V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. page 26

D. Nie, L. Wang, Y. Gao, and D. Shen. Fully convolutional networks for multi-modality isointense infant brain image segmentation. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 1342–1345, 2016. page 27

R. K. Nielsen, S. Darkner, and A. Feragen. TopAwaRe: Topology-Aware Registration. In *MICCAI*, volume 11765, pages 364–372. Cham, 2019. page 79

M. Niethammer, G. L. Hart, D. F. Pace, P. M. Vespa, A. Irimia, J. D. V. Horn, , and S. R. Aylward. Geometric metamorphosis. *MICCAI*, 14(2):639–646, 2011. page 79

J. M. Noothout, N. Lessmann, M. C. Van Eede, L. D. van Harten, E. Sogancioglu, F. G. Heslinga, M. Veta, B. van Ginneken, and I. Išgum. Knowledge distillation with ensembles of convolutional neural networks for medical image segmentation. *Journal of Medical Imaging*, 9(5):052407–052407, 2022. page 41

L. G. Nyúl, J. K. Udupa, and X. Zhang. New variants of a method of mri scale standardization. *IEEE transactions on medical imaging*, 19(2):143–150, 2000. page 29

A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. page 44

M. Orbes-Arteaga, M. J. Cardoso, L. Sørensen, M. Modat, S. Ourselin, M. Nielsen, and A. Pai. Simultaneous synthesis of FLAIR and segmentation of white matter hypointensities from T1 MRIs. In *MIDL*, 2018. page 37

Q. T. Ostrom, H. Gittleman, P. Liao, C. Rouse, Y. Chen, J. Dowling, Y. Wolinsky, C. Kruchko, and J. Barnholtz-Sloan. Cbtrus statistical report: primary brain and central nervous system tumors diagnosed in the united states in 2007–2011. *Neuro-oncology*, 16(suppl_4):iv1–iv63, 2014. page 13

J. Pallud, E. Audureau, G. Noel, R. Corns, E. Lechapt-Zalcman, J. Duntze, V. Pavlov, J. Guyotat, P. D. Hieu, P.-J. Le Reste, T. Faillot, C.-F. Litre, N. Desse, A. Petit, E. Emery, J. Voirin, J. Peltier, F. Caire, J.-R. Vignes, J.-L. Barat, O. Langlois, E. Dezamis, E. Parraga, M. Zanello, E. Nader, M. Lefranc, L. Bauchet, B. Devaux, P. Menei, P. Metellus, C. de Neuro-Oncologie of the Société Française de Neurochirurgie, G. A. Lahoud, F. Andreiuolo, A. Borha, A. Busson, L. Capelle, F. Chapon, F. Chassoux, I. Catry-Thomas, F. Chrétien, P. Colin, A. Czorny, J.-M. Derlon, M.-D. Diebold, H. Duffau, M. Edjlali-Goujon, J. Eskandari, A. Fustier, C. Gantois, R. Gadan, J. Geffrelot, E. Gimbert, J. Godard, S. Godon-Hardy, M. Gueye, J.-S. Guillamo, N. Heil, D. Hoffmann, N. Jovenin, M. Kalamarides, H. Katranji, S. Khouri, M. Koziak, E. Landré, V. Leon, D. Liguoro, E. Mandonnet, M. Mann, E. Méary, J.-F. Meder, C. Mellerio, S. Michalak, C. Miquel, K. Mokhtari, P. Monteil, O. Naggara, F. Nataf, C. Oppenheim, I. Quintin-Roue, P. Page, P. Paquis, D. Pedenon, P. Peruzzi, T. Riem, V. Rigau, O. Rigaux-Viodé, A. Rougier, F.-X. Roux,

C. Salon, E. Théret, B. Turak, D. Trystram, F. Vandenbos, P. Varlet, G. Viennet, and C. d. N.-O. o. t. S. F. d. N. Vital, Anne. Long-term results of carmustine wafer implantation for newly diagnosed glioblastomas: a controlled propensity-matched analysis of a French multicenter cohort. *Neuro-Oncology*, 17(12):1609–1619, 07 2015. page 13

I. Pappas, H. Hector, K. Haws, B. Curran, A. S. Kayser, and M. D'Esposito. Improved normalization of lesioned brains via cohort-specific templates. *Hum Brain Mapp*, 42(13):4187–4204, 2021. page 79

T. N. Pappas and N. S. Jayant. An adaptive clustering algorithm for image segmentation. In *International Conference on Acoustics, Speech, and Signal Processing,*, pages 1667–1670. IEEE, 1989. page 22

X. Pennec, P. Cathier, and N. Ayache. Understanding the "demon's algorithm": 3d non-rigid registration by gradient descent. In *MICCAI*, 1999. page 67

D. Qin, J.-J. Bu, Z. Liu, X. Shen, S. Zhou, J.-J. Gu, Z.-H. Wang, L. Wu, and H.-F. Dai. Efficient medical image segmentation based on knowledge distillation. *IEEE Transactions on Medical Imaging*, 40(12):3820–3831, 2021. pages 41, 42

M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems 30*, pages 6076–6085. Curran Associates, Inc., 2017. page 51

M. Rahimpour, J. Bertels, A. Radwan, H. Vandermeulen, S. Sunaert, D. Vandermeulen, F. Maes, K. Goffin, and M. Koole. Cross-modal distillation to improve mri-based brain tumor segmentation with missing mri sequences. *IEEE Transactions on Biomedical Engineering*, 69(7):2153–2164, 2022. pages 41, 101, 102

J. C. Reinhold, B. E. Dewey, A. Carass, and J. L. Prince. Evaluating the impact of intensity normalization on mr image synthesis. In *Medical Imaging 2019: Image Processing*, volume 10949, pages 890–898. SPIE, 2019. page 29

M.-M. Rohé, M. Datar, T. Heimann, M. Sermesant, and X. Pennec. SVF-Net: Learning Deformable Image Registration Using Shape Matching. In *MICCAI*, volume 10433, pages 266–274. 2017. page 72

T. Rohlfing, N. M. Zahr, E. V. Sullivan, and A. Pfefferbaum. The sri24 multichannel atlas of normal adult human brain structure. *Human brain mapping*, 31(5):798–819, 2010. pages 22, 88

O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. pages 26, 37, 89

F. Rousseau, L. Drumetz, and R. Fablet. Residual Networks as Flows of Diffeomorphisms. *JMIV*, 62(3):365–375, 2020. page 84

A. Roux et al. MRI atlas of IDH wild-type supratentorial glioblastoma: Probabilistic maps of phenotype, management, and outcomes. *Radiology*, 293(3): 633–643, 2019. pages 16, 78, 104

D. Rueckert, L. Sonoda, C. Hayes, D. Hill, M. Leach, and D. Hawkes. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE Transactions on Medical Imaging*, 18(8):712–721, 1999. page 67

D. Rueckert, P. Aljabar, R. A. Heckemann, J. V. Hajnal, and A. Hammers. Diffeomorphic registration using b-splines. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006*, pages 702–709, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-44728-3. pages 72, 77

K. Scheufele, A. Mang, A. Gholami, C. Davatzikos, G. Biros, and M. Mehl. Coupling brain-tumor biophysical models and diffeomorphic image registration. *Comput Methods Appl Mech Eng*, 347:533–567, 2019. pages 79, 102, 105

K. Scheufele, S. Subramanian, and G. Biros. Fully automatic calibration of tumor-growth models using a single mpmri scan. *IEEE Transactions on Medical Imaging*, 40(1):193–204, 2021. pages 79, 105

J. A. Schnabel, D. Rueckert, M. Quist, J. M. Blackall, A. D. Castellano-Smith, T. Hartkens, G. P. Penney, W. A. Hall, H. Liu, C. L. Truwit, et al. A generic framework for non-rigid registration based on non-uniform multi-level free-form deformations. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2001: 4th International Conference Utrecht, The Netherlands, October 14–17, 2001 Proceedings 4*, pages 573–581. Springer, 2001. page 67

Z. Shen, Z. He, W. Cui, J. Yu, Y. Zheng, C. Zhu, and M. Savvides. Adversarial-based knowledge distillation for multi-model ensemble and noisy data refinement. *arXiv preprint arXiv:1908.08520*, 2019. page 44

R. T. Shinohara, E. M. Sweeney, J. Goldsmith, N. Shiee, F. J. Mateen, P. A. Calabresi, S. Jarso, D. L. Pham, D. S. Reich, C. M. Crainiceanu, et al. Statistical normalization techniques for magnetic resonance imaging. *NeuroImage: Clinical*, 6:9–19, 2014. page 29

N. Shusharina and G. C. Sharp. Image registration using radial basis functions with adaptive radius. *Medical physics*, 39 11:6542–9, 2012. page 67

J. G. Sled, A. P. Zijdenbos, and A. C. Evans. A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE transactions on medical imaging*, 17(1):87–97, 1998. page 27

S. Srinivas and F. Fleuret. Knowledge transfer with jacobian matching. In *International Conference on Machine Learning*, pages 4723–4731. PMLR, 2018. pages 43, 57

R. Stefanescu, O. Commowick, G. Malandain, P.-Y. Bondiau, N. Ayache, and X. Pennec. Non-rigid Atlas to Subject Registration with Pathologies for Conformal Brain Radiotherapy. In *MICCAI*, volume LNCS 3216, pages 704–711, Sept. 2004. ISBN 978-3-540-22976-6 978-3-540-30135-6. pages 79, 102

W. Stummer, U. Pichlmeier, T. Meinel, O. D. Wiestler, F. Zanella, H.-J. Reulen, A.-G. S. Group, et al. Fluorescence-guided surgery with 5-aminolevulinic acid for resection of malignant glioma: a randomised controlled multicentre phase iii trial. *The lancet oncology*, 7(5):392–401, 2006. page 13

S. Subramanian, A. Ghafouri, K. M. Scheufele, N. Himthani, C. Davatzikos, and G. Biros. Ensemble inversion for brain tumor growth models with mass effect. *IEEE Transactions on Medical Imaging*, 42(4):982–995, 2023. page 105

K. Sun and S. Simon. Fdrn: a fast deformable registration network for medical images. *Medical Physics*, 48(10):6453–6463, 2021. page 73

Z. Tang, P. Yap, and D. Shen. A New Multi-Atlas Registration Framework for Multimodal Pathological Images Using Conventional Monomodal Normal Atlases. *IEEE Transactions on Image Processing*, 28(5):2293–2304, 2019. page 79

J.-P. Thirion. Image matching as a diffusion process: an analogy with maxwell's demons. *Medical Image Analysis*, 2(3):243–260, 1998. page 67

S. Thust, S. Heiland, A. Falini, H. R. Jäger, A. Waldman, P. Sundgren, C. Godi, V. Katsaros, A. Ramos, N. Bargallo, et al. Glioma imaging in europe: a survey of 220 centres and recommendations for best clinical practice. *European radiology*, 28(8):3306–3317, 2018. page 13

Y. Tian, D. Krishnan, and P. Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020a. page 43

Y. Tian, D. Krishnan, and P. Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020b. page 43

A. Trouvé and L. Younès. Metamorphoses Through Lie Group Action. *Foundations of Computational Mathematics*, 5(2):173–198, 2005. pages 3, 4, 17, 18, 78, 80, 81, 82, 102

N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee. N4itk: improved n3 bias correction. *IEEE transactions on medical imaging*, 29(6):1310–1320, 2010. page 27

T. Tykocki and M. Eltayeb. Ten-year survival in glioblastoma. a systematic review. *Journal of Clinical Neuroscience*, 54:7–13, 2018. page 13

D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. page 26

S. Vadacchino, R. Mehta, N. M. Sepahvand, B. Nichyporuk, J. J. Clark, and T. Arbel. Had-net: A hierarchical adversarial knowledge distillation network for improved enhanced tumour segmentation without post-contrast images. In *Medical Imaging with Deep Learning*, pages 787–801. PMLR, 2021. pages 44, 45, 62, 101, 102

G. van Tulder and M. de Bruijne. Why Does Synthesized Data Improve Multi-sequence Classification? In *MICCAI*, volume LNCS 9349, pages 531–538, Cham, 2015. Springer. page 36

V. Vapnik and R. Izmailov. Learning using privileged information: Similarity control and knowledge transfer. *Journal of Machine Learning Research*, 16 (61):2023–2049, 2015. page 40

F.-X. Vialard, L. Risser, D. Rueckert, and C. J. Cotter. Diffeomorphic 3d image registration via geodesic shooting using an efficient adjoint calculation. *International Journal of Computer Vision*, 97:229–241, 2011. pages 68, 70

A. Virzì, C. O. Muller, J.-B. Marret, E. Mille, L. Berteloot, D. Grévent, N. Boddaert, P. Gori, S. Sarnacki, and I. Bloch. Comprehensive review of 3d segmentation software tools for mri usable for pelvic surgery planning. *Journal of digital imaging*, 33(1):99–110, 2020. page 21

B. D. d. Vos, F. F. Berendsen, M. A. Viergever, M. Staring, and I. Išgum. End-to-end unsupervised deformable image registration with a convolutional neural network. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 204–212. Springer, 2017. page 73

U. Vovk, F. Pernus, and B. Likar. A review of methods for correction of intensity inhomogeneity in mri. *IEEE transactions on medical imaging*, 26(3):405–421, 2007. pages 6, 28

G. Wang, W. Li, S. Ourselin, and T. Vercauteren. Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3*, pages 178–190. Springer, 2018. page 26

H. Wang, J. W. Suh, S. R. Das, J. B. Pluta, C. Craige, and P. A. Yushkevich. Multi-atlas segmentation with joint label fusion. *IEEE transactions on pattern analysis and machine intelligence*, 35(3):611–623, 2012. page 22

P. Wang and A. C. Chung. Focal dice loss and image dilation for brain tumor segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA*

2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4, pages 119–127. Springer, 2018. page 30

M. Weller, R. Stupp, G. Reifenberger, A. A. Brandes, M. J. Van Den Bent, W. Wick, and M. E. Hegi. Mgmt promoter methylation in malignant gliomas: ready for personalized medicine? *Nature Reviews Neurology*, 6(1):39–51, 2010. page 103

M. Wu and N. Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in neural information processing systems*, 31, 2018. page 39

Z. Wu, Y. Xiong, S. X. Yu, and D. Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. page 44

C. Xu and J. L. Prince. Generalized gradient vector flow external forces for active contours. *Signal processing*, 71(2):131–139, 1998. page 22

Y. Xue, T. Xu, H. Zhang, L. R. Long, and X. Huang. Segan: Adversarial network with multi-scale l 1 loss for medical image segmentation. *Neuroinformatics*, 16:383–392, 2018. page 31

C. Yang, H. Zhou, Z. An, X. Jiang, Y. Xu, and Q. Zhang. Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12319–12328, 2022. page 41

X. Yang, X. Han, E. Park, S. Aylward, R. Kwitt, and M. Niethammer. Registration of pathological images. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 97–107. Springer, 2016. pages 79, 102

X. Yang, R. Kwitt, M. Styner, and M. Niethammer. Quicksilver: Fast Predictive Image Registration - a Deep Learning Approach. *NeuroImage*, 158:378–396, 2017. pages 72, 73

B. Yu, L. Zhou, L. Wang, J. Fripp, and P. Bourgeat. 3d cgan based cross-modality mr image synthesis for brain tumor segmentation. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 626–630. IEEE, 2018. pages 37, 38

E. I. Zacharaki, C. S. Hogea, D. Shen, G. Biros, and C. Davatzikos. Non-diffeomorphic registration of brain tumor images by simulating tissue loss and tumor growth. *NeuroImage*, 46(3):762–774, July 2009. pages 79, 102

S. Zagoruyko and N. Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017. page 42

J. Zhang, K.-K. Ma, M.-H. Er, and V. Chong. Tumor segmentation from magnetic resonance imaging by learning via one-class support vector machine. In *International Workshop on Advanced Image Technology (IWAIT'04)*, pages 207–211, 2004. page 24

P. Zhang, B. Yu, R. Zhang, X. Chen, S. Shao, Y. Zeng, J. Cui, and J. Zhao. Longitudinal study of the morphological and t2* changes of knee cartilages of marathon runners using prototype software for automatic cartilage segmentation. *The British Journal of Radiology*, 94(1119):20200833, 2021. page 22

Z. Zhao, H. Chen, and L. Wang. A coarse-to-fine framework for the 2021 kidney and kidney tumor segmentation challenge. In *Kidney and Kidney Tumor Segmentation: MICCAI 2021 Challenge, KiTS 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings*, pages 53–58. Springer, 2022. page 27

T. Zhou, S. Ruan, Y. Guo, and S. Canu. A multi-modality fusion network based on attention mechanism for brain tumor segmentation. In *2020 IEEE 17th international symposium on biomedical imaging (ISBI)*, pages 377–380. IEEE, 2020a. page 27

T. Zhou, S. Canu, P. Vera, and S. Ruan. Latent correlation representation learning for brain tumor segmentation with missing mri modalities. *IEEE Transactions on Image Processing*, 30:4263–4274, 2021. page 39

Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 39(6):1856–1867, 2020b. page 26

J. Zhu, S. Tang, D. Chen, S. Yu, Y. Liu, M. Rong, A. Yang, and X. Wang. Complementary relation contrastive distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9260–9269, 2021. page 43

A. Ziabari, D. H. Ye, S. Srivastava, K. D. Sauer, J.-B. Thibault, and C. A. Bouman. 2.5d deep learning for ct image reconstruction using a multi-gpu implementation. In *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, pages 2044–2049, 2018. page 25

D. Zikic, B. Glocker, E. Konukoglu, A. Criminisi, C. Demiralp, J. Shotton, O. M. Thomas, T. Das, R. Jena, and S. J. Price. Decision forests for tissue-specific segmentation of high-grade gliomas in multi-channel mr. In *MICCAI (3)*, pages 369–376, 2012. page 24

D. Zikic, Y. Ioannou, M. Brown, and A. Criminisi. Segmentation of brain tumor tissues with convolutional neural networks. *Proceedings MICCAI-BRATS*, 36 (2014):36–39, 2014. page 25

ECOLE
DOCTORALE

Titre : Vers la génération d'altlas de glioblastome avec des méthodes d'apprentissage profond : segmentation de tumeurs et recalage métamorphique d'images.

Mots clés : apprentissage profond, imagerie medicale, construction d'atlas, segmentation

Résumé : Cette thèse s'inscrit dans le cadre de la construction d'un atlas de glioblastomes. En imagerie médicale, un atlas est une image ou un ensemble d'images représentant la distribution statistique d'une population. Souvent, cette distribution prend la forme d'une image représentant la moyenne de la population et d'un ensemble de cartes de déformations entre cette moyenne et chaque image. Pour construire un atlas, il est donc important de correctement définir les transformations entre les images. Les méthodes classiques de recalage considèrent que les deux images sont en correspondance bijective. Or, cela n'est pas le cas dans notre contexte où les deux images n'ont pas le même nombre de composants. Un défi de la thèse a donc été de produire des transformations entre deux images avec des topologies différentes.
La première partie de la thèse a porté sur la segmentation de tumeurs cérébrales sur des IRM, permettant ainsi de déterminer précisément l'endroit avec la différence topologique. Alors que la plupart des algorithmes utilisent plusieurs modalités d'acquisition, dans la pratique clinique souvent une seule est disponible. Notre problématique a donc été de proposer un algorithme qui soit performant sur une seule modalité tout en utilisant les informations des bases de données multi-modales pendant l'apprentissage. Pour cela, nous avons utilisé une technique de distillation de connaissances. Une analyse de différentes stratégies de distillation nous a permis de montrer dans quels cas ces méthodes sont utiles.
La seconde partie de la thèse porte sur le recalage d'une image d'un patient ayant une tumeur vers une image de sujet sain. Nous avons développé une méthode qui prend en compte à la fois les différences géométriques et les différences topologiques entre deux images. Nous nous sommes inspirés des Métamorphoses qui ont été développées pour transformer la géométrie et les niveaux d'intensité d'une image. Nous avons utilisé un réseau de neurones résiduel pour résoudre les équations aux dérivées partielles qui constituent les métamorphoses. Cela nous permet d'utiliser la méthode en apprentissage, réduisant considérablement le temps d'inférence une fois que le réseau a été entraîné. En outre, nous encourageons une séparation entre les transformations de forme et d'apparence en exploitant un masque de segmentation de la tumeur. La méthode de recalage développée constitue ainsi un outil important dans le but de construire un atlas de glioblastomes.

Title : Towards the generation of glioblastoma atlases with deep learning methods: Tumor segmentation and Metamorphic image registration.

Keywords : deep learning, medical imaging, atlas construction, segmentation

Abstract : The aim of this thesis was to build an atlas of glioblastoma. In medical imaging, an atlas is an image or a set of images that are meant to represent the statistical distribution of a population. Often, this distribution takes the form of an image representing the population average and a set of deformation maps between this mean and each image. To construct an atlas, it is therefore important to correctly define the transformations between the images. Conventional registration methods assume that the two images have only a geometric difference - that is, the first image is the bijective deformation of the other. However, this is not the case in our context, where the two images do not have the same number of components. A challenge of this thesis was therefore to produce transformations between two images with different topologies.
The first part of the thesis focused on the segmentation of brain tumors on MRI. Indeed, it is important to segment the tumors in order to precisely detect the location with the topological differences. Since our goal is to build an atlas from clinical images, we need a segmentation algorithm that performs well on patients with only one acquisition modality available (such as T1-weighted images). However, most of the state-of-the-art tumor segmentation algorithms need four modalities to perform well. The first goal of this thesis was thus to produce a segmentation algorithm that performs well on test images from a single modality, while leveraging information from multi-modal databases during training. To this end, we proposed a new method based on knowledge distillation. We compare the proposed method with several knowledge distillation strategies and show that this kind of methods performs well in a low-data regime and becomes less useful in a high-data regime.
The second part of the thesis deals with the registration of a cancerous image onto a healthy image. We developed a method that, in addition to taking into account the geometric differences, it also considers the topological differences between two images. Inspired by Metamorphosis, a method developed to transform the geometry and intensity levels of an image, we used a residual neural network to solve the partial differential equations that encode the Metamorphosis framework. This allowed us to reformulate the method in a learning context, which greatly reduced the inference time once the network has been trained. Additionally, we encouraged an anatomically meaningful disentanglement between shape and appearance transformations by leveraging the (previously estimated) segmentation mask of the tumor. The developed registration method is thus an important tool in the construction of the glioblastoma atlas.

Institut Polytechnique de Paris
91120 Palaiseau, France