A modality-adaptive method for segmenting brain tumors and organs-at-risk in radiation therapy planning

Agn, Mikael; Munck af Rosenschöld, Per; Puonti, Oula; Lundemann, Michael J.; Mancini, Laura; Papadaki, Anastasia; Thust, Steffi; Ashburner, John; Law, Ian; Van Leemput, Koen

Published in:
Medical Image Analysis

Link to article, DOI:
10.1016/j.media.2019.03.005

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

Medical Image Analysis 54 (2019) 220–237

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

A modality-adaptive method for segmenting brain tumors and organs-at-risk in radiation therapy planning

Mikael Agn[a,*], Per Munck af Rosenschöld[b], Oula Puonti[c], Michael J. Lundemann[d], Laura Mancini[e,f], Anastasia Papadaki[e,f], Steffi Thust[e,f], John Ashburner[g], Ian Law[h], Koen Van Leemput[a,i]

[a] Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark
[b] Radiation Physics, Department of Hematology, Oncology and Radiation Physics, Skåne University Hospital, Lund, Sweden
[c] Danish Research Centre for Magnetic Resonance, Copenhagen University Hospital Hvidovre, Denmark
[d] Department of Oncology, Copenhagen University Hospital Rigshospitalet, Denmark
[e] Neuroradiological Academic Unit, Department of Brain Repair and Rehabilitation, UCL Institute of Neurology, University College London, UK
[f] Lysholm Department of Neuroradiology, National Hospital for Neurology and Neurosurgery, UCLH NHS Foundation Trust, UK
[g] Wellcome Centre for Human Neuroimaging, UCL Institute of Neurology, University College London, UK
[h] Department of Clinical Physiology, Nuclear Medicine and PET, Copenhagen University Hospital Rigshospitalet, Denmark
[i] Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, USA

ARTICLE INFO

Article history:
Received 18 July 2018
Revised 14 March 2019
Accepted 21 March 2019
Available online 22 March 2019

Keywords:
Glioma
Whole-brain segmentation
Generative probabilistic model
Restricted Boltzmann machine

ABSTRACT

In this paper we present a method for simultaneously segmenting brain tumors and an extensive set of organs-at-risk for radiation therapy planning of glioblastomas. The method combines a contrast-adaptive generative model for whole-brain segmentation with a new spatial regularization model of tumor shape using convolutional restricted Boltzmann machines. We demonstrate experimentally that the method is able to adapt to image acquisitions that differ substantially from any available training data, ensuring its applicability across treatment sites; that its tumor segmentation accuracy is comparable to that of the current state of the art; and that it captures most organs-at-risk sufficiently well for radiation therapy planning purposes. The proposed method may be a valuable step towards automating the delineation of brain tumors and organs-at-risk in glioblastoma patients undergoing radiation therapy.

1. Introduction

Glioblastomas, which are the most common type of malignant tumors originating within the brain (Preusser et al., 2011), are commonly treated with a combination of surgical resection, chemo-therapy and radiation therapy. During radiation therapy, the patient is subjected to radiation beams, typically from different directions and with different intensity profiles, with the aim of maximizing the delivered radiation dose to the targeted tumor while minimizing the dose to sensitive healthy structures, so-called organs-at-risk (OARs) (Shaffer et al., 2010). For the purpose of planning a radiation therapy session, these structures need to be delineated on computed tomography (CT) or magnetic resonance (MR) scans of the patient's head (Munck af Rosenschöld et al., 2011).

In current clinical practice, delineation is performed manually with limited assistance from automatic procedures, which is time consuming for the human expert and typically suffers from high inter-rater variability (Deeley et al., 2011; Dolz et al., 2015b; Menze et al., 2015). These limitations are amplified in emerging techniques for image-guided radiation therapy, which introduce a demand for continuous delineation during treatment (Lagendijk et al., 2014). Consequently, there is an increasing need for fast automated segmentation methods that can robustly segment both brain tumors and OARs from clinically acquired head scans.

Recent years have seen an influx of *discriminative* methods for brain tumor segmentation, with good – although not very robust – performance reported in the annual MICCAI Brain Tumor Segmentation (BRATS) challenges (Menze et al., 2015). Discriminative methods directly exploit the intensity information of annotated training data to discern between tumorous and other tissue in new images. Traditionally, they rely on user-engineered image features that are then fed into classifiers, such as random forests (Zikic et al., 2012; Islam et al., 2013; Tustison et al., 2015; Maier et al., 2016) or support vector machines (Bauer et al., 2011). Lately, however, convolutional neural networks (CNNs), which learn suitable image features simultaneously with their classifiers, have become

* Corresponding author.
E-mail address: miag@dtu.dk (M. Agn).

https://doi.org/10.1016/j.media.2019.03.005
1361-8415/

more prominent (Pereira et al., 2016; Kamnitsas et al., 2017; Havaei et al., 2017).

Although discriminative methods have demonstrated state-of-the-art tumor segmentation performance, they suffer from several drawbacks that limit their practical applicability in radiation therapy planning settings. In particular, what is needed in radiation therapy is an accurate segmentation not just of the tumor, but also of a multitude of OARs. Although CNNs segmenting dozens of brain substructures have recently been demonstrated (Roy et al., 2017; Rajchl et al., 2018), using such methods in the context of radiation therapy planning is complicated by their need for large annotated training datasets, as scans with high-quality segmentations of both tumors and OARs in hundreds of patients are not easily available. Further exacerbating this issue is that both the type and the number of acquired images often differ substantially among treatment centers, not only as a result of differences in imaging protocols and scanner platforms, but also because of the continuous development of novel MR pulse sequences for brain tumor imaging (Mabray et al., 2015; Sauwen et al., 2016). Although an active research area in the field (Havaei et al., 2016; Ghafoorian et al., 2017; Valindria et al., 2018), effectively dealing with the ensuing explosion of possible contrasts and contrast combinations remains an open problem for discriminative segmentation methods.

In order to sidestep these difficulties with discriminative approaches, we present a method in this paper for simultaneously segmenting brain tumors and OARs using a *generative* approach, in which prior knowledge of anatomy and the imaging process are incorporated using Bayesian statistics. Specifically, our method combines an existing contrast-adaptive method for whole-brain segmentation (Puonti et al., 2016) with a new spatial regularization model of tumor shape using generative neural networks. The OARs we consider in this paper are eyes, optic chiasm, optic nerves, brainstem, and hippocampi, but more structures can easily be added. Compared to existing work, the proposed method presents several novel contributions:

1. To the best of our knowledge, this is the first method that addresses the segmentation of both brain tumors and OARs within the same modeling framework. While existing generative methods for tumor segmentation typically also perform classification into white matter, gray matter and cerebrospinal fluid (Moon et al., 2002; Prastawa et al., 2003; Menze et al., 2010; Gooya et al., 2012; Kwon et al., 2014; Bakas et al., 2016), they do not further subdivide these tissue types into OARs, nor do they segment OARs outside the brain. Conversely, with the exception of the optic system (Bekes et al., 2008; Noble and Dawant, 2011; Dolz et al., 2015a), most automated segmentation methods for OARs in radiation therapy applications have been concentrated on label transfer using non-linear registration of manually annotated template data (Dawant et al., 1999; Cuadra et al., 2004; Isambert et al., 2008; Bauer et al., 2013; Bondiau et al., 2005), which does not address the problem of tumor segmentation itself.
2. By adopting a generative approach, the proposed method makes judicious use of readily available training data. In particular, the approach allows merging of disparate models of normal head structures, learned from manually annotated scans of normal subjects, with models of tumor shape derived from brain tumor patients, without requiring that segmentations of these two set of structures are available within the same set of subjects. Importantly, once trained the same method can be readily applied to data from different sites without retraining. As we will demonstrate, this is the case even when data acquisitions are fundamentally different from the data used to train the method, such as CT scans or experimental MR contrasts.
3. In contrast to discriminative methods for brain lesion segmentation, in which large spatial contexts are exploited to achieve state-of-the-art segmentation accuracy (Corso et al., 2008; Geremia et al., 2011; Karimaghaloo et al., 2016; Brosch et al., 2016; Kamnitsas et al., 2017), spatial regularization of lesions in generative methods has so far been limited to local properties, such as local lesion probability in lesion-seeded probabilistic atlases (Moon et al., 2002; Prastawa et al., 2003; Gooya et al., 2012; Kwon et al., 2014; Bakas et al., 2016) or first-order Markov random fields (MRFs) in which only pairwise interactions between neighboring voxels are taken into account (Van Leemput et al., 2001; Menze et al., 2010). In this paper, we explore the potential of convolutional restricted Boltzmann machines (cRBMs) (Lee et al., 2011) to provide long-range spatial regularization through MRFs with high-order clique potentials that are automatically learned from manual segmentations of brain tumors. We empirically demonstrate that these higher-order shape models yield an advantage in segmentation accuracy compared to first-order MRFs.

Table 1
Labels associated with normal head structures, with brain structures in B.

$l \in B$	{white matter (WM), grey matter (GM), cerebrospinal fluid (CSF), brainstem, unspecified brain tissue, and left and right hippocampus}
$l \notin B$	{background, eye socket fat, eye socket muscles, optic chiasm; and left and right optic nerve, eye tissue and eye fluid}

Preliminary versions of this work appeared in two conference contributions (Agn et al., 2016a; 2016b). Here we extend the method to handle more OARs, in particular optic nerves, optic chiasm, and eyes; describe the model and the statistical inference in more detail; and provide an in-depth validation on a large number of patients, evaluating the method's adaptability to varying input data and suitability for radiation therapy planning.

2. Modeling framework

Let $\mathbf{D} = (\mathbf{d}_1, \ldots, \mathbf{d}_I)$ denote the data of N co-registered medical images of a patient's head, where I is the number of image voxels and $\mathbf{d}_i$ contains the log-transformed[1] intensities at voxel i. Each voxel i has a normal label $l_i \in \{1, \ldots, K\}$ that is associated with one of $K = 17$ normal head structures, detailed in Table 1, where B denotes a set of structures located inside the brain. A voxel i can be tumor-affected, indicated by $z_i = 1$, where $z_i \in \{0, 1\}$. Within tumor-affected tissue, a voxel i can be either *edema* or *core*, indicated by $y_i = 0$ and $y_i = 1$, respectively, where $y_i \in \{0, 1\}$. Edema corresponds to the visible peritumoral edema surrounding the core, which corresponds to the gross tumor volume (GTV) used in radiation therapy. To model the labels l_i, z_i and y_i across all voxels, we build a generative model that describes the image formation process, seen in Fig. 1. The model consists of two parts. The first part is a likelihood function $p(\mathbf{D}|\mathbf{l}, \mathbf{z}, \mathbf{y}, \boldsymbol{\theta})$ that links the labels to image intensities, where $\mathbf{l} = (l_1, \ldots, l_I)$, $\mathbf{z} = (z_1, \ldots, z_I)$, and $\mathbf{y} = (y_1, \ldots, y_I)$. This likelihood function depends on a set of parameters $\boldsymbol{\theta}$, governed by a prior distribution $p(\boldsymbol{\theta})$, that allows the model to adapt to images with different contrast properties. The second part is a segmentation prior $p(\mathbf{l}, \mathbf{z}, \mathbf{y}|\boldsymbol{\eta}) = \sum_{\mathbf{H}^z} \sum_{\mathbf{H}^y} p(\mathbf{l}, \mathbf{z}, \mathbf{y}, \mathbf{H}^z, \mathbf{H}^y|\boldsymbol{\eta})$, where $\boldsymbol{\eta}$, with prior $p(\boldsymbol{\eta})$,

[1] We work with log-transformed intensities to model the MR bias field effect as an additive (rather than multiplicative) process, see Section 2.2.

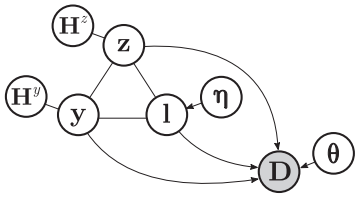

Fig. 1. Graphical representation of the model. The atlas-based prior on **l** is defined by parameters $\boldsymbol{\eta}$ governing the deformation of the atlas. The tumor-affected map **z** and the tumor core map **y** are connected to auxiliary variables $\mathbf{H}^z$ and $\mathbf{H}^y$, respectively. The variables **l, z** and **y** jointly predict the data **D** according to the likelihood parameters $\boldsymbol{\theta}$. Shading indicates observed variables.

are parameters governing the deformation of a probabilistic atlas, and $\mathbf{H}^z$ and $\mathbf{H}^y$ are auxiliary variables that help encode high-order shape models of **z** and **y**.

We use this model to obtain a fully automated segmentation algorithm by evaluating the posterior of the labels given the data:

$$p(\mathbf{l}, \mathbf{z}, \mathbf{y}|\mathbf{D}) \propto p(\mathbf{D}|\mathbf{l}, \mathbf{z}, \mathbf{y})p(\mathbf{l}, \mathbf{z}, \mathbf{y}), \tag{1}$$

where $p(\mathbf{l}, \mathbf{z}, \mathbf{y}) = \int_{\boldsymbol{\eta}} p(\mathbf{l}, \mathbf{z}, \mathbf{y}|\boldsymbol{\eta})p(\boldsymbol{\eta})\mathrm{d}\boldsymbol{\eta}$ and $p(\mathbf{D}|\mathbf{l}, \mathbf{z}, \mathbf{y}) = \int_{\theta} p(\mathbf{D}|\mathbf{l}, \mathbf{z}, \mathbf{y}, \theta)p(\theta)\mathrm{d}\theta$ will be detailed in Sections 2.1 and 2.2, respectively, and computationally evaluating Eq. (1) will be addressed in Section 2.3.

2.1. Segmentation prior

We obtain the segmentation prior $p(\mathbf{l}, \mathbf{z}, \mathbf{y}|\boldsymbol{\eta})$ by defining

$$p(\mathbf{l}, \mathbf{z}, \mathbf{y}, \mathbf{H}^y, \mathbf{H}^z|\boldsymbol{\eta}) \propto \exp\left[-E(\mathbf{l}, \mathbf{z}, \mathbf{y}, \mathbf{H}^y, \mathbf{H}^z|\boldsymbol{\eta})\right]$$

with an energy

$$\begin{aligned} E(\mathbf{l}, \mathbf{z}, \mathbf{y}, \mathbf{H}^y, \mathbf{H}^z|\boldsymbol{\eta}) = {} & E^z(\mathbf{z}, \mathbf{H}^z) + E^y(\mathbf{y}, \mathbf{H}^y) \\ & - \log q(\mathbf{l}|\boldsymbol{\eta}) + \sum_i f(l_i, z_i, y_i), \end{aligned} \tag{2}$$

where $E^z(\mathbf{z}, \mathbf{H}^z)$ and $E^y(\mathbf{y}, \mathbf{H}^y)$ are the energy terms of two cRBMs that model tumor shape in **z** and **y**, respectively, and $q(\mathbf{l}|\boldsymbol{\eta})$ is a deformable atlas that models the spatial configuration of the normal labels in **l**. Additionally, we use a restriction function defined as

$$f(l, z, y) = \begin{cases} \infty & \text{if } z = 0 \text{ and } y = 1 \\ \infty & \text{if } z = 1 \text{ and } l \notin B \\ 0 & \text{otherwise} \end{cases}. \tag{3}$$

This function encodes that a core voxel can never appear outside the tumor-affected region **z**, and that a tumor-affected voxel can never appear outside the brain. Note that it is only this restriction function that ties the labels **l, z**, and **y** to each other. Without it, the segmentation prior would simply devolve into $p(\mathbf{l}, \mathbf{z}, \mathbf{y}|\boldsymbol{\eta}) = p(\mathbf{l}|\boldsymbol{\eta})p(\mathbf{z})p(\mathbf{y})$.

We will now present the two types of models that are included in this prior: the cRBMs on tumor shape in Section 2.1.1, and the atlas on the spatial configuration of normal head structures in Section 2.1.2.

2.1.1. Prior on tumor shape using cRBMs

In order to model the spatial configuration of tumor tissue, we use cRBMs – neural networks that can be interpreted as MRFs encoding high-order interactions among voxels ("visible units") through local connections to latent variables ("hidden units") (Fischer and Igel, 2014). In contrast to a standard restricted Boltzmann machine (Smolensky, 1986; Freund and Haussler, 1992; Hinton, 2002), where arbitrary weights can be assigned between the visible and the hidden units, the weights of the connections in a cRBM are in the form of filters that are much smaller than the image size and that are shared among all locations in the image (Lee et al., 2011). This allows us to infer over large images without a predefined size. We now present the model in only 1D for the sole purpose of avoiding cluttered equations, but it directly generalizes to 3D images.

The distribution over visible units **v** in a cRBM is defined as

$$p(\mathbf{v}) = \sum_{\mathbf{H}^v} \exp\left[-E^v\left(\mathbf{v}, \mathbf{H}^v\right)\right] \tag{4}$$

with the energy term (Lee et al., 2011)

$$E^v\left(\mathbf{v}, \mathbf{H}^v\right) = -\sum_{m=1}^{M} \mathbf{h}_m^v \bullet \left(\mathbf{w}_m^v * \mathbf{v}\right) - \sum_{m=1}^{M} b_m^v \sum_{j=1}^{J} h_{mj}^v - a^v \sum_{i=1}^{I} v_i,$$

where $\mathbf{H}^v = \{\mathbf{h}_m^v\}_{m=1}^M$ contains M hidden groups, $\bullet$ denotes element-wise product followed by summation, and $*$ denotes spatial convolution. Each hidden group $\mathbf{h}_m^v$ is connected to the visible units in **v** with a convolutional filter $\mathbf{w}_m^v$ of size r, and contains $J = I - r + 1$ hidden units. The filter $\mathbf{w}_m^v$ models interactions between the hidden and visible units, effectively detecting specific features in **v**. Furthermore, each hidden group has a bias b_m^v and visible units have a bias a^v. These bias terms encourage units to be enabled or disabled when set to non-zero values. A small example of a cRBM can be seen in Fig. 2.

The computational appeal of this model is that no direct connections exist between two visible units or two hidden units, so that the visible units are independent of each other given the state of the hidden ones, and vice versa:

$$p\left(\mathbf{v}|\mathbf{H}^v\right) = \prod_i p\left(v_i|\mathbf{H}^v\right) \quad \text{and} \quad p\left(\mathbf{H}^v|\mathbf{v}\right) = \prod_m \prod_j p\left(h_{mj}^v|\mathbf{v}\right) \tag{5}$$

$$\text{with} \quad p\left(v_i|\mathbf{H}^v\right) \propto \exp\left[v_i\left(\sum_m \left(\tilde{\mathbf{w}}_m^v * \mathbf{h}_m^v\right)_i + a^v\right)\right]$$

$$\text{and} \quad p\left(h_{mj}^v|\mathbf{v}\right) \propto \exp\left[h_{mj}^v\left(\left(\mathbf{w}_m^v * \mathbf{v}\right)_j + b_m^v\right)\right],$$

where $\tilde{\mathbf{w}}$ denotes a mirror-reversed version of the filter **w**. Although no direct connections exist among visible units, high-order connections are still obtained among them through the connections to the hidden units. This can be seen clearly by summing out the hidden units in Eq. (4) analytically (Fischer and Igel, 2014), which gives us $p(\mathbf{v}) \propto \exp[-E^v(\mathbf{v})]$ with

$$E^v(\mathbf{v}) = \sum_{i=1}^{I-r+1} g(\mathbf{v}_{i:i+r-1}) - a^v \sum_{i=1}^{I} v_i, \tag{6}$$

where $i{:}i'$ denotes elements from i to i', and $g(\mathbf{v}_{i:i+r-1}) = -\sum_m \log\left[1 + \exp(\mathbf{w}_m^v \cdot \mathbf{v}_{i:i+r-1} + b_m^v)\right]$ is a high-order MRF clique potential defined over groups of visible units as

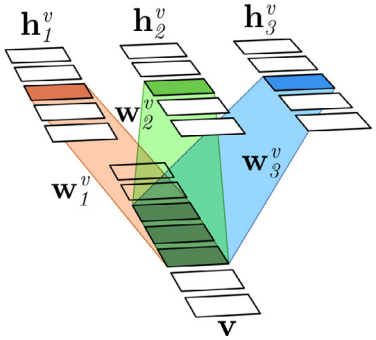

Fig. 2. A small 1D example of a cRBM with $\mathbf{v} = (v_1, \ldots, v_7)$ and $\mathbf{H}^v = \{\mathbf{h}_m^v\}_{m=1}^3$. Visible units (image voxels) are connected to hidden units in a hidden group $\mathbf{h}_m^v$ through a convolutional filter $\mathbf{w}_m^v$ of size 3. All locations in **v** share the same filter weights. The connections are exemplified by the three central visible units which are connected to the central hidden unit in each group.

large as the filter size r. This can be contrasted to traditionally used MRF models for brain lesion shape, e.g., (Van Leemput et al., 1999a; Menze et al., 2010), where a^v is set to zero and the clique potentials are only between pairs of voxels in $\mathbf{v}$, i.e., $r = 2$, defined as $g(\mathbf{v}_{i:i+1}) = \beta^v |v_i - v_{i+1}|$, where β^v is a user-tunable hyperparameter.

In this paper, we use two separate binary cRBMs: one that models shape in the tumor-affected map $\mathbf{z}$ and one that models shape in the core map $\mathbf{y}$, with energies $E^z(\mathbf{z}, \mathbf{H}^z)$ and $E^y(\mathbf{y}, \mathbf{H}^y)$, defined exactly as for $\mathbf{v}$. We learn suitable values for the filters and biases of these cRBMs by stochastic gradient descent on the log-likelihood using expert segmentations obtained from training data, as detailed in Section 3.2.

2.1.2. Atlas-based prior on normal head structures

To model the spatial configuration of normal head structures $q(\mathbf{l}|\boldsymbol{\eta})$, we use the type of probabilistic atlas introduced in Van Leemput (2009) and further validated in Puonti et al. (2016). It is based on a deformable tetrahedral mesh, where the parameters $\boldsymbol{\eta}$ are the spatial positions of the mesh nodes and $p(\boldsymbol{\eta})$ is a topology-preserving deformation prior (Ashburner et al., 2000). Each mesh node in the atlas is associated with a probability vector containing the probabilities of the K normal head structures to occur at that node; for a given mesh deformation, these vectors are interpolated using barycentric interpolation to yield probabilities $\pi_i(k|\boldsymbol{\eta})$ for each structure k in all voxels i. Assuming that structure labels at different voxels are conditionally independent given the node positions, this finally yields

$$q(\mathbf{l}|\boldsymbol{\eta}) = \prod_{i=1}^{I} \pi_i(l_i|\boldsymbol{\eta}).$$

As described in Van Leemput (2009), the atlas can be trained by a non-linear, group-wise registration of expert segmentations obtained from training data. The node positions in atlas space with associated label probabilities are optimized during this training process, as well as the topology of the mesh, where the mesh resolution adapts to be sparse in large uniform regions and dense at label borders. Fig. 3 shows the atlas that we built for the current paper; more details will be given in Section 3.1.

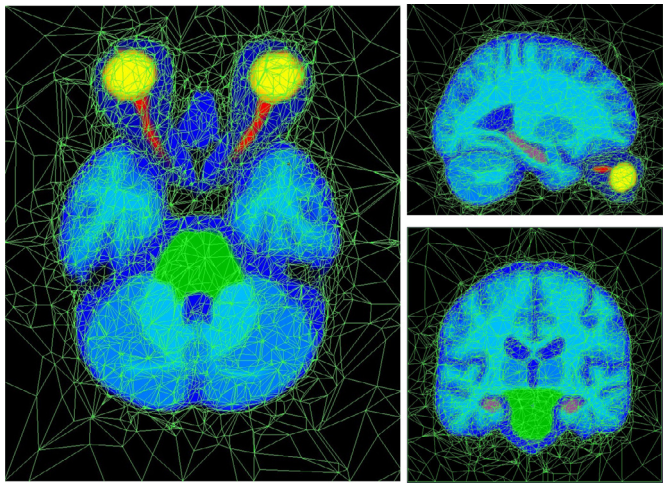

Fig. 3. The built atlas in axial, sagittal, and coronal view; shown in atlas space. Nodes and connections between nodes are shown in light green and probabilities of normal labels, interpolated between the nodes, are shown in varying colors (yellow = eye fluid, orange = eye tissue, red = optic nerves, green = brainstem, lilac = hippocampi, shades of blue = other normal labels). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.2. Likelihood

To link the labels $\mathbf{l}$, $\mathbf{z}$ and $\mathbf{y}$ to image intensities, we use $X = 12$ Gaussian mixture models (GMMs) in the likelihood function $p(\mathbf{D}|\mathbf{l}, \mathbf{z}, \mathbf{y}, \boldsymbol{\theta})$, where each GMM models the intensity distribution of certain label combinations. Some GMMs are connected to several label combinations, e.g., left and right hippocampus are modeled by the same GMM as both hippocampi have the same intensity properties, and any voxel i that belongs to edema (i.e., $z_i = 1$, $y_i = 0$ and $l_i \in B$) is modeled by a single GMM. In order to map a voxel i with l_i, z_i and y_i to a specific GMM, we therefore introduce a mapping function $x(l_i, z_i, y_i)$, which is detailed in Table 2. Additionally, we model so-called bias fields that typically corrupt MR scans as additive effects by linear combinations of spatially smooth basis functions. A bias field is a multiplicative low-frequency imaging artifact, so to model it as an additive effect we work with log-transformed intensities throughout this paper, as in Wells et al. (1996); Van Leemput et al. (1999b).

Specifically, we define the likelihood function as

$$p\big(\mathbf{D}|\mathbf{l}, \mathbf{z}, \mathbf{y}, \boldsymbol{\theta}\big) = \prod_i p_i\big(\mathbf{d}_i | x(l_i, z_i, y_i), \boldsymbol{\theta}\big)$$

$$\text{with} \quad p_i\big(\mathbf{d}_i|x, \boldsymbol{\theta}\big) = \sum_{g=1}^{G_x} \gamma_{xg} \mathcal{N}\big(\mathbf{d}_i | \boldsymbol{\mu}_{xg} + \mathbf{C}\boldsymbol{\phi}_i, \boldsymbol{\Sigma}_{xg}\big),$$

where $\mathcal{N}(\mathbf{d}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$; G_x is the number of components in the xth GMM; and γ_{xg}, $\boldsymbol{\mu}_{xg}$ and $\boldsymbol{\Sigma}_{xg}$ are the weight, mean and covariance matrix of component g. The weights satisfy the constraints $\gamma_{xg} \geq 0$ and $\sum_{g=1}^{G_x} \gamma_{xg} = 1$. Furthermore, the bias fields corrupting MR scans are modeled by $\boldsymbol{\phi}_i$ and $\mathbf{C}$. The column vector $\boldsymbol{\phi}_i \in \mathbb{R}^P$ evaluates P spatially smooth basis functions at voxel i and $\mathbf{C} = (\mathbf{c}_1, \ldots, \mathbf{c}_N)^T$ denotes the parameters of the bias field model, where $\mathbf{c}_n \in \mathbb{R}^P$ are the parameters for image contrast n. Finally, all likelihood parameters are jointly collected in $\boldsymbol{\theta} = \{\{\gamma_{xg}, \boldsymbol{\mu}_{xg}, \boldsymbol{\Sigma}_{xg}\} \forall xg, \mathbf{C}\}$.

We use a restricted conjugate prior $p(\boldsymbol{\theta})$ on the likelihood parameters:

$$p(\boldsymbol{\theta}) \propto \begin{cases} \prod_x \Big[\text{Dir}(\boldsymbol{\gamma}_x|\alpha_0) \prod_{g=1}^{G_x} \text{IW}\big(\boldsymbol{\Sigma}_{xg}|\upsilon_x^0, \mathbf{S}_x^0\big)\Big] \\ \quad \text{if constraints on } \{\boldsymbol{\mu}_{xg}\} \text{ are satisfied} \\ 0 \quad \text{otherwise,} \end{cases} \tag{7}$$

where we have used uniform priors on the bias field parameters $\mathbf{C}$ and the mean vectors $\{\boldsymbol{\mu}_{xg}\}$, and conjugate priors on the covariance matrix of each component and mixture weights of each GMM following the definitions in Murphy (2012). To avoid extreme numerical values in Gaussian components representing only

Table 2
Mapping function $x(l, z, y)$ that maps combinations of l, z and y to 12 distinct GMMs in the model. Note that combinations $\{z = 1, \forall y, l \notin B\}$ and $\{z = 0, y = 1, \forall l\}$ will never occur due to the restriction function in Eq. (3). The right column shows the number of components G_x in each GMM – these values are based on pilot experiments detailed in Section 3.3.

Combinations of l, z, and y	$x(l, z, y)$	G_x
$z = 1, y = 1,$ and $l \in B$	core	3
$z = 1, y = 0,$ and $l \in B$	edema	1
$z = 0, y = 0,$ and $l \in$		
{GM, L/R hippocampus}	global gray matter (GGM)	1
{WM, brainstem}	global white matter (GWM)	1
{L/R optic nerve, L/R eye tissue}	global nerves/eye tissue (GNE)	2
{L/R eye fluid}	global eye fluid	1
CSF	CSF	2
background	background	3
unspecified brain tissue	unspecified brain tissue	1
optic chiasm	optic chiasm	1
eye socket fat	eye socket fat	2
eye socket muscles	eye socket muscles	3

a handful of voxels, we regularize the covariance matrices using inverse-Wishart distributions $\mathrm{IW}(\boldsymbol{\Sigma}|\upsilon_x^0, \mathbf{S}_x^0)$, where $\mathbf{S}_x^0$ is a prior scatter matrix with strength υ_x^0. Furthermore, to discourage numerical removal of components, we use symmetric Dirichlet distributions $\mathrm{Dir}(\boldsymbol{\gamma}|\alpha_0)$ where $\alpha_0 > 1$, since these have their mode at $\gamma_{xg} = 1/G_x, \forall x, g$. Finally, we add certain linear constraints on $\{\boldsymbol{\mu}_{xg}\}$ to encode prior knowledge about overall tumor appearance relative to normal brain tissue in typical MR sequences for brain tumor imaging. These constraints allow for a wide variability of tumor appearance across subjects, while imposing plausible limits on how similar to normal tissue tumors can look. Tuning of the likelihood function and its parameter prior is detailed in Section 3.3.

2.3. Inference

Exact inference of the posterior $p(\mathbf{l}, \mathbf{y}, \mathbf{z}|\mathbf{D})$ in Eq. (1) is computationally intractable because it marginalizes over all of the uncertainty in the model parameters and the hidden units of the cRBM models. We therefore resort to Markov chain Monte Carlo (MCMC) techniques to sample from all unknown variables (except the atlas node positions $\boldsymbol{\eta}$, as detailed below), followed by voxel-wise majority voting on the segmentation samples to obtain the final segmentation. This procedure is detailed in Section 2.3.1.

Although it is possible to also sample from $\boldsymbol{\eta}$, as shown in Iglesias et al. (2013), this is considerably more computationally expensive and was not implemented in this paper. Instead, we ignore the uncertainty on deformations and use a suitable point estimate of the atlas node positions $\hat{\boldsymbol{\eta}}$ obtained with a simplified model, which we describe in Section 2.3.2. We also obtain an initial state of the sampler from this simplified model.

2.3.1. MCMC sampler

Given a point estimate of the atlas node positions $\hat{\boldsymbol{\eta}}$, we generate samples of the labels $\mathbf{l}$, $\mathbf{z}$ and $\mathbf{y}$ from $p(\mathbf{l}, \mathbf{z}, \mathbf{y}|\mathbf{D}, \hat{\boldsymbol{\eta}})$ by sampling from $p(\mathbf{l}, \mathbf{z}, \mathbf{y}, \mathbf{H}^z, \mathbf{H}^y, \boldsymbol{\theta}|\mathbf{D}, \hat{\boldsymbol{\eta}})$ using a blocked Gibbs sampler, and discarding the samples of $\mathbf{H}^y$, $\mathbf{H}^z$ and $\boldsymbol{\theta}$. The sampler, which is illustrated in Algorithm 1, iteratively draws each set of variables

Algorithm 1 MCMC sampler to obtain $\hat{\mathbf{l}}, \hat{\mathbf{z}}, \hat{\mathbf{y}}$.

Input: $\mathbf{l}^{(0)}, \mathbf{z}^{(0)}, \mathbf{y}^{(0)}, \hat{\boldsymbol{\eta}}$
Output final estimates of labels $\hat{\mathbf{l}}, \hat{\mathbf{z}}, \hat{\mathbf{y}}$
for $s = 1$ to $(S_{\text{burn-in}} + S)$
  Sample $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|\mathbf{D}, \mathbf{l}^{(s-1)}, \mathbf{z}^{(s-1)}, \mathbf{y}^{(s-1)})$, detailed in Appendix A
  Sample $\mathbf{H}^z$ from $p(\mathbf{H}^z|\mathbf{z}^{(s-1)})$, see Eq. 5
  Sample $\mathbf{H}^y$ from $p(\mathbf{H}^y|\mathbf{y}^{(s-1)})$, see Eq. 5
  Sample $\mathbf{l}^{(s)}, \mathbf{z}^{(s)}, \mathbf{y}^{(s)}$ from $p(\mathbf{l}, \mathbf{z}, \mathbf{y}|\mathbf{D}, \mathbf{H}^z, \mathbf{H}^y, \boldsymbol{\theta}, \hat{\boldsymbol{\eta}})$ in Eq. 8
end for
Final $\hat{\mathbf{l}}, \hat{\mathbf{z}}, \hat{\mathbf{y}}$ obtained by voxel-wise majority voting of samples in $\{\mathbf{l}^{(s)}, \mathbf{z}^{(s)}, \mathbf{y}^{(s)}\}_{s=S_{\text{burn-in}}+1}^{S_{\text{burn-in}}+S}$

from its conditional distribution given the other variables; with the exception of $\boldsymbol{\theta}$ this is straightforward to implement as each conditional distribution factorizes over its components. The hidden units $\mathbf{H}^z$ and $\mathbf{H}^y$ are sampled as in Eq. (5), and the labels are sampled from

$$p\left(\mathbf{l}, \mathbf{z}, \mathbf{y}|\mathbf{D}, \mathbf{H}^z, \mathbf{H}^y, \boldsymbol{\theta}, \hat{\boldsymbol{\eta}}\right) = \prod_i p_i\left(l_i, z_i, y_i|\mathbf{d}_i, \mathbf{H}^z, \mathbf{H}^y, \boldsymbol{\theta}, \hat{\boldsymbol{\eta}}\right) \tag{8}$$

with

$$p_i\left(l_i, z_i, y_i|\mathbf{d}_i, \mathbf{H}^z, \mathbf{H}^y, \boldsymbol{\theta}, \hat{\boldsymbol{\eta}}\right) \propto p_i\left(\mathbf{d}_i|l_i, z_i, y_i, \boldsymbol{\theta}\right)\pi_i(l_i) \exp\left[z_i\left(\textstyle\sum_m (\tilde{\mathbf{w}}_m^z * \mathbf{h}_m^z)_i + a^z\right)\right] \exp\left[y_i\left(\textstyle\sum_m \left(\tilde{\mathbf{w}}_m^y * \mathbf{h}_m^y\right)_i + a^y\right)\right] \exp\left[-f(l_i, z_i, y_i)\right].$$

Sampling from the conditional distribution $p(\boldsymbol{\theta}|\mathbf{D}, \mathbf{l}, \mathbf{z}, \mathbf{y})$ is more difficult due to interdependencies among the various components of $\boldsymbol{\theta}$ (including those imposed by the linear constraints on the Gaussian means $\{\boldsymbol{\mu}_{xg}\}$), and is detailed in Appendix A.

We obtain the final estimate of the labels $\hat{\mathbf{l}}$, $\hat{\mathbf{z}}$, and $\hat{\mathbf{y}}$ by voxel-wise majority voting, separately on each variable, over S collected samples after an initial burn-in period of $S_{\text{burn-in}}$ samples.

2.3.2. Simplified model to obtain atlas node position estimates and initial state of sampler

For the purpose of estimating appropriate atlas node positions $\hat{\boldsymbol{\eta}}$ and to obtain an initial state $\{\mathbf{l}^{(0)}, \mathbf{z}^{(0)}, \mathbf{y}^{(0)}\}$ for the MCMC sampler, we use a simplified model in which the non-local dependencies among the voxels introduced by the cRMB shape models are removed. In particular, we set the filter weights $\{\mathbf{w}_m^z\}_{m=1}^M$ and $\{\mathbf{w}_m^y\}_{m=1}^M$ to zero values, effectively removing the hidden units from the model, and set the visual bias values so that a fraction $w = 0.1$ of normal voxels is expected to be tumorous, and within these voxels a fraction $u = 0.5$ is expected to be tumor core. We achieve this by setting the visual biases $a_y = \log(\frac{u}{1-u})$ and $a_z = \log(\frac{w-wu}{1-w})$. This reduces the model to the same form as in Puonti et al. (2016), and we can therefore use the same approach for optimization, i.e., by alternating between optimizing the likelihood parameters $\boldsymbol{\theta}$ with a generalized expectation-maximization (GEM) algorithm (Dempster et al., 1977) and optimizing the atlas node positions $\boldsymbol{\eta}$ with a general-purpose gradient-based optimizer.

Algorithm 2 Initial algorithm to obtain $\mathbf{l}^{(0)}, \mathbf{z}^{(0)}, \mathbf{y}^{(0)}, \hat{\boldsymbol{\eta}}$.

Input: $\mathbf{D}$, initial affine transformation of atlas $\hat{\boldsymbol{\eta}}$
Output: $\mathbf{l}^{(0)}, \mathbf{z}^{(0)}, \mathbf{y}^{(0)}, \hat{\boldsymbol{\eta}}$
Change tumor prior to a simplified version
Initialize $\hat{\boldsymbol{\theta}}$
until convergence
  Optimize $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{D}, \hat{\boldsymbol{\eta}})$
  Optimize $\hat{\boldsymbol{\eta}} = \arg\max_{\boldsymbol{\eta}} p(\boldsymbol{\eta}|\mathbf{D}, \hat{\boldsymbol{\theta}})$
end until
Record $\hat{\boldsymbol{\eta}}$
Compute *maximum a posteriori* segmentation
$\{\mathbf{l}^{(0)}, \mathbf{z}^{(0)}, \mathbf{y}^{(0)}\} = \arg\max_{\mathbf{l},\mathbf{z},\mathbf{y}} p(\mathbf{l}, \mathbf{z}, \mathbf{y}|\mathbf{D}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}})$

Algorithm 2 illustrates this approach, which is implemented as in Puonti et al. (2016) with a few exceptions. In particular, for the atlas node positions a more efficient optimizer is used (limited-memory BFGS (Liu and Nocedal, 1989)). Furthermore, the linear constraints in the prior $p(\boldsymbol{\theta})$ (Eq. (7)) alter the relevant update equations in the GEM algorithm to involve a so-called quadratic programming problem, as detailed in Appendix B. Finally, as in Puonti et al. (2016), all Gaussian component parameters in $\boldsymbol{\theta}$ are initialized based on the atlas prior after affine registration, except the mean values for the tumor GMMs. These are instead initialized based on prior knowledge about overall tumor appearance in typical MR sequences for brain tumor imaging, as detailed in Section 3.3.

After convergence of the parameter optimization with this simplified model, we record $\hat{\boldsymbol{\eta}}$ and compute the *maximum a posteriori* segmentation

$$\left\{\mathbf{l}^{(0)}, \mathbf{z}^{(0)}, \mathbf{y}^{(0)}\right\} = \arg\max_{\mathbf{l},\mathbf{z},\mathbf{y}} p\left(\mathbf{l}, \mathbf{y}, \mathbf{z}|\mathbf{D}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}}\right)$$
$$= \arg\max_{\{l_i,z_i,y_i\}} \prod_i p\left(l_i, z_i, y_i|\mathbf{d}_i, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}}\right),$$

which is used as the initial state for the MCMC sampler.

3. Training and tuning of the model

In this section, we describe how we trained the deformable atlas $q(\mathbf{l}|\boldsymbol{\eta})$ (in Section 3.1) and the two cRBMs modeling $\mathbf{z}$ and $\mathbf{y}$ (in Section 3.2), which together make up the segmentation prior in our model. Furthermore, we describe overall tuning of the method in Section 3.3.

To train the deformable atlas, we used the same training dataset as in Puonti et al. (2016), which is also the training data of the publicly available software package FreeSurfer (Fischl, 2012). This dataset consists of 39 subjects (without any tumors) with dozens of neuroanatomical structures within the brain segmented by experts, following a validated semi-automated protocol developed at the Center for Morphometric Analysis (CMA), MGH, Boston (Caviness et al., 1989; 1996; Kennedy et al., 1989). We call this dataset *the atlas training dataset*.

For all other parts of the model, we used the training dataset of the brain tumor segmentation (BRATS) challenge that was held in conjunction with the BrainLes workshop at the 2015 MICCAI conference. This dataset consists of 220 high-grade gliomas and 54 low-grade gliomas of varying types, with publicly available ground truth segmentations of tumor, which include annotations of four tumor regions: edema and three regions inside tumor core. 30 subjects were manually segmented (20 high-grade, 10 low-grade), while the rest have fused segmentations from highly ranked algorithms from previous editions of the BRATS challenge. The included MR sequences are T2-weighted FLAIR (2D acquisition), T2-weighted (2D acquisition), T1-weighted (2D acquisition), and T1-weighted with contrast enhancement (T1c, 3D acquisition). The scans have been acquired at different centers, with varying magnetic field strength and resolution. All data were resampled to 1 mm isotropic resolution by the challenge organizers. We call this dataset *the BRATS 2015 training dataset*.

3.1. Training the deformable atlas

We automatically trained the tetrahedral mesh atlas, shown in Fig. 3 and described in Section 2.1.2, from expert segmentations from the atlas training dataset. We emphasize that only the manual segmentations are needed for this purpose, and that the intensity information of the original MR scans from which these were derived was not used.

As we are specifically interested in structures applicable to radiation therapy, we merged some of the manually segmented structures into larger labels before building the atlas. Specifically, we kept the segmentations for the OARs *brainstem, optic chiasm* and left and right *hippocampus*; as well as the background label. We merged all other structures into the following catch-all labels: cerebrospinal fluid (CSF), and remaining white matter (WM) and gray matter (GM). Two important OARs were not included in the available expert segmentations, as they are located outside of the brain – namely *optic nerves* and *eyes*. We therefore performed additional manual delineations for the left and right structures of these two extra OARs. To provide some context around these structures, we also delineated the muscles and fat in the eye sockets into two separate labels. We further separated the left and right eye into two labels each: *eye fluid* describing the fluid and gel inside an eye and *eye tissue* describing the lens and the solid outer layer of an eye.

To build the atlas, we chose the resulting segmentations of a representative subset of 10 subjects. We selected 10 subjects as manual delineations are time consuming and we have previously shown that adding more subjects does not substantially increase the average segmentation performance (Puonti et al., 2016). After building the atlas, we added an unspecified brain tissue label designed to capture normal structures that are not specified in the atlas, such as blood vessels. Towards this end, we added a constant prior probability of 0.01 for this label in each mesh node's probability vector and re-normalized the probability vector to ensure that the values sum to one. Overall, we use $K = 17$ normal head structure labels, listed in Table 1.

3.2. Training the cRBMs

To learn suitable values for the filters and biases of the cRBMs modeling $\mathbf{z}$ and $\mathbf{y}$, described in Section 2.1.1, we used the 30 *manual* tumor segmentations from the BRATS 2015 training dataset, again without using any associated intensity information. As the number of segmentations is small, we augmented the dataset by flipping the segmentations in eight different directions, yielding a dataset of 240 tumor segmentations. To form binary segmentations corresponding to $\mathbf{z}$ and $\mathbf{y}$, we merged tumor regions in the manual segmentations: all four regions for $\mathbf{z}$ and the three tumor core regions for $\mathbf{y}$. We learned the filters and bias terms through stochastic gradient ascent on the log-probability of the tumor segmentations under the cRBM model. To efficiently approximate the gradients, we used the contrastive divergence (CD) approximation with one block-Gibbs sampling step (Hinton, 2002) together with the so-called enhanced gradient which has been shown to improve learning (Cho et al., 2013; Melchior et al., 2013). Each cRBM was trained with 9600 gradient steps of size 0.1. A subset of 10 randomly selected segmentations (a so-called mini-batch) was used to approximate the gradient at each step.

We used the same settings for both cRBMs. The filter size and number of filters were set by pilot experiments on a separate subset of the BRATS 2015 training dataset. Choosing a larger filter size would increase the number of parameters which may result in overfitting, while a smaller filter size might not capture long-range features. Empirically, we found that by tying neighboring parameters in a filter we can reduce the number of parameters while still capturing long-range features. Specifically, we tied filter parameters in $(2 \times 2 \times 2)$ blocks of voxels, effectively treating each block as one parameter. We used $M = 40$ filters of size $(14 \times 14 \times 14)$ (i.e., $7 \times 7 \times 7$ blocks) corresponding to 40 hidden groups. In our pilot experiments, this configuration performed better than other combinations of 20, 30 and 40 filters of sizes between 10 and 18.

3.3. Tuning

The tuning of the model described in this section is based on initial experiments on the full BRATS 2015 training dataset. We use $S = 50$ samples from the MCMC sampler, after an initial burn-in period of $S_{\text{burn-in}} = 200$ (cf. Algorithm 1). In the likelihood function $p(\mathbf{D}|\mathbf{l}, \mathbf{z}, \mathbf{y}, \boldsymbol{\theta})$, described in Section 2.2, we associate three Gaussian components (i.e., $G_x = 3$) with the GMMs of *core, eye socket muscle*, and *background*; two components with the GMMs of *eye socket fat, CSF,* and *GNE* (global optic nerves/eye tissue); and one component with all other GMMs (cf. Table 2). Additionally, we use the 64 lowest frequencies of the 3D DCT as bias field basis functions, i.e., $P = 64$.

In the likelihood parameter prior $p(\boldsymbol{\theta})$ defined in Eq. (7), the linear constraints on the Gaussian means $\{\boldsymbol{\mu}_{xg}\}$ were set by building statistics of their values in the BRATS 2015 training data. Specifically, we estimated the average Gaussian mean values using automatic segmentations produced by our method, but with the tumor

labels fixed to the ground truth. Based on the resulting statistics, we set constraints for the Gaussian mean values relating to edema and enhanced core in the MR sequences FLAIR and T1c. Enhanced core, which is the core region that is enhanced in T1c, is specifically targeted by setting constraints on only one of the Gaussian components associated with core. Additionally, we set constraints on the mean values relating to the unspecified brain tissue and optic chiasm as to ascertain that these labels will not interfere with the tumor segmentation. All constraints are in relation to the mean values of global WM (GWM) and global GM (GGM), and are shown in Table 3. Note that the image intensities are log-transformed, so an added logarithm of a value is equivalent to that value being multiplied by the original intensities.

For the inverse Wishart distribution in Eq. (7), we set the scatter matrix $\mathbf{S}_x^0 = \upsilon_x^0 X^{-2}\mathrm{diag}\left[\sum_i (\mathbf{d}_i - \bar{\mathbf{d}})(\mathbf{d}_i - \bar{\mathbf{d}})^T / I\right]$, with $\bar{\mathbf{d}} = \sum_i \mathbf{d}_i / I$ and strength $\upsilon_x^0 = N + 10^{-1} I_x / G_x$, where I_x is the expected number of voxels for each GMM, obtained from the atlas for normal structures and from the BRATS 2015 training data for tumor. Because the unspecified brain tissue label should catch any unspecified brain tissue, we use a wider scatter matrix for the GMM of this label, with X replaced by 1. Finally, we set $\alpha_0 = 1 + 10^{-4} I$ in the Dirichlet prior of Eq. (7) for each GMM.

Initialization of the simplified model of Algorithm 2. As described in Section 2.3.2, all Gaussian component parameters are initialized based on the atlas prior, except the mean values associated with tumor. If the flat tumor prior in the simplified model of Algorithm 2 would be used, these mean values would be initialized as the average intensities within the brain, which are far away from typical tumor intensities. Therefore, we instead initialize each of these mean values a certain distance (measured in standard deviations) away from the average data intensity in the corresponding image. Based on initial pilot experiments on the BRATS 2015 training data, we set the distances as in Table 4, e.g., the T2 mean value for edema is initialized 0.7 standard deviations above the average T2 data intensity.

Specific settings for tumor core. The GMM connected to tumor core needs special care due to the flat tumor prior used in the simplified model of Algorithm 2. Tumor core regions can vary widely in their intensity distribution and can also have a similar intensity distribution to edema and normal tissue. This fact creates challenges when estimating the parameters of the core GMM during inference in Algorithm 2, as the flat tumor prior has no notion of tumor shape. The easiest region to recognize only by intensity is the region that is enhanced in T1c. Thus, we temporarily restrict all three Gaussian components associated with core to have identical mixture parameters while using the simplified model, and specifically target the enhanced region. We then release the restriction before starting the sampler (Algorithm 1). Additionally, to help the full cRBM-based model to capture other core regions in the vicinity of the enhanced region, we randomly change a fifth of the edema voxels ($z_i^{(0)} = 1$ and $y_i^{(0)} = 0$) in the initial state to core voxels ($z_i^{(0)} = 1$ and $y_i^{(0)} = 1$).

Table 3
Constraints on mean values of Gaussian components.

Edema (TE)		
$\mu_{\mathrm{TE}}^{\mathrm{FLAIR}}$	$\geq$	$\max\left(\mu_{\mathrm{GWM}}^{\mathrm{FLAIR}}, \mu_{\mathrm{GGM}}^{\mathrm{FLAIR}}\right) + \log 1.15$
Core, Gaussian component relating to enhanced core (denoted TC1)		
$\mu_{\mathrm{TC1}}^{\mathrm{FLAIR}}$	$\geq$	$\max\left(\mu_{\mathrm{GWM}}^{\mathrm{FLAIR}}, \mu_{\mathrm{GGM}}^{\mathrm{FLAIR}}\right)$
$\mu_{\mathrm{TC1}}^{\mathrm{T1c}}$	$\geq$	$\max\left(\mu_{\mathrm{GWM}}^{\mathrm{T1c}}, \mu_{\mathrm{GGM}}^{\mathrm{T1c}}\right) + \log 1.10$
Unspecified brain tissue (US)		
$\mu_{\mathrm{US}}^{\mathrm{FLAIR}}$	$\leq$	$\min\left(\mu_{\mathrm{GWM}}^{\mathrm{FLAIR}}, \mu_{\mathrm{GGM}}^{\mathrm{FLAIR}}\right) - \log 1.05$
$\mu_{\mathrm{US}}^{\mathrm{T1c}}$	$\leq$	$\min\left(\mu_{\mathrm{GWM}}^{\mathrm{T1c}}, \mu_{\mathrm{GGM}}^{\mathrm{T1c}}\right) - \log 1.05$
Chiasm (CH)		
$\mu_{\mathrm{CH}}^{\mathrm{FLAIR}}$	$\leq$	$\min\left(\mu_{\mathrm{GWM}}^{\mathrm{FLAIR}}, \mu_{\mathrm{GGM}}^{\mathrm{FLAIR}}\right)$

Table 4
Distances used to initialize tumor GMMs, in standard deviations away from the average image intensity.

x	FLAIR	T2	T1	T1c
Core	1	0.7	0.2	1.5
Edema	1	0.7	0.2	0.2

4. Experiments and results

To evaluate our method, we conduct experiments on three different datasets from different imaging centers with varying input data, including CT images and several MR sequences. The varying input data enables us to assess our method's ability to handle images from different modalities, MR sequences and scanner platforms. In Section 4.1, we test our method on a dataset of 70 glioblastoma patients that have undergone radiation therapy treatment at Rigshospitalet in Copenhagen, Denmark. We call this dataset *the Copenhagen dataset*. It includes all data needed for a radiation therapy session, which enables us to test our method's performance on both tumor and OAR segmentation, as well as to conduct a dosimetric evaluation. In this dataset, we will also vary the input data to the method from the available images to test the effect this has on the segmentation performance. Furthermore, we will compare our cRBM-based method to that of the same method but instead using first-order MRFs. In Section 4.2, we compare our method's performance on segmenting tumors to that of top-performing methods in the 2015 BRATS challenge, using the challenge's test dataset of 53 patients from varying centers, which we call *the BRATS 2015 test dataset*. Lastly, in Section 4.3, we further test our method's ability to adapt to varying input data by using a dataset of seven patients with a different set of acquired images, including an MR sequence not present in the other datasets, scanned at the National Hospital for Neurology and Neurosurgery, UCLH NHS Foundation Trust, London, UK. We call this dataset *the London dataset*.

Throughout this section, we employ two widely used metrics – Dice score and Hausdorff distance – to compare our method's segmentations to the manual segmentations in the datasets. A Dice score measures overlap between two segmentations, where a score of zero means no overlap and a score of one means a perfect overlap. In contrast, a Hausdorff distance evaluates the distance between the surfaces of two segmentations. As in the BRATS challenges, we use a robust version of this metric. A further description of these two metrics can be found in the BRATS reference paper (Menze et al., 2015).

The entire algorithm was implemented in MATLAB 2015b, except for the atlas mesh deformation which was implemented in C++. Segmenting one subject takes around 40 minutes on a Core i7-5930K CPU with 32 GB of memory, with roughly equal time spent on Algorithms 1 and 2 described in Section 2.3.

4.1. Results for Copenhagen dataset

To evaluate our method's performance on segmenting both OARs and tumors, we use the Copenhagen dataset, which consists of 70 glioblastoma patients that have undergone radiation therapy treatment at Rigshospitalet in Copenhagen, Denmark, in 2016 (GTV size range: 5–205 cm^3). As part of their radiation therapy workup, these patients have been scanned with a CT scanner and a Siemens

Magnetom Espree 1.5T MRI scanner. The dataset includes three MR sequences: T2-weighted FLAIR (transversal 2D-acquisition), T2-weighted (T2, transversal 2D-acquisition) and T1-weighted with contrast enhancement (T1c, 3D-acquisition); with a voxel size of $(1 \times 1 \times 3)$, $(1 \times 1 \times 3)$ and $(0.5 \times 0.5 \times 1)$ mm^3 respectively. The CT scans have a voxel size of $(0.5 \times 0.5 \times 1)$ mm^3. As part of the treatment planning, the GTV (corresponding to tumor core) and several OARs (including hippocampi, brainstem, eyes, optic nerves and chiasm) have been manually delineated in CT-space, with the MR sequences transformed to this space. As the only pre-processing step for our method, we co-register the MR and CT scans and resample them to 1 mm isotropic resolution.

4.1.1. Evaluation of results on three data combinations

To test the ability of our method to adapt to varying input data, we evaluate the segmentation results obtained with three different data combinations. In the first combination, we use all available data, i.e., {T1c, FLAIR, T2, CT}. We include CT scans as they are used in manual delineation of the optic system. CT scans do not exhibit bias field artefacts, so we clamp the bias field parameters in our model to zero for this image type. Additionally, as CT scans have a low contrast within the brain, we can initialize the tumor-associated mean values in the same way as for normal labels. In the second combination, we only use the MR sequences, i.e., {T1c, FLAIR, T2}. In the last combination, we use T1c and a new combinatory sequence named $FLAIR^2$ that is designed to improve lesion detection (Wiggermann et al., 2016). This image is computed by multiplying FLAIR with T2. For this image, we use the same settings in our model as for FLAIR. We emphasize that some of the modalities under consideration – in particular CT and $FLAIR^2$ – were not included in any training data available to the proposed segmentation algorithm. We start with an overall visual inspection of the segmentations and then analyze the performance scores, followed by a more in-depth visual inspection of some of the segmentations.

Fig. 4 shows slices of the segmentations using the three data combinations for four representative subjects. We can see that the method in general seems to work well and consistently across all three data combinations. The atlas deforms well to fit subjects with varying shapes, and the method is capable of segmenting tumor cores of varying size, shape and intensity profile; although it underestimates the tumor size in some cases. Eyes, hippocampi and brainstem seem to be consistently well-captured, while optic nerves and chiasm are less well-captured, but better for the data combination including CT, which is because the difference in intensity between the optic nerves and surrounding tissue is larger in CT than MR. Finally, as can be noticed in the last subject, many subjects show some ambiguity in the intensity profile of the optic nerves.

Fig. 5 shows box plots of the Dice scores and Hausdorff distances for the three data combinations, with the following structures: tumor core (TC), brainstem (BS), hippocampi (HC), eyes (EB), optic nerves (ON), and chiasm (CH). The left and right structures are included as separate scores in the plots for hippocampi, eyes, and optic nerves. As can be seen, the method readily adapts to the various included and excluded images in the three data combinations without the need for adjustment. The scores are consistent across the three data combinations for all regions except optic nerve and chiasm. The average Dice scores for tumor core are fair, but the range of scores is large. However, this is consistent with the state of the art in brain tumor segmentation, as will be shown in Section 4.2. Furthermore, this dataset includes a number of difficult subjects with large resections, small and thin contrast-enhanced tumor regions in T1c and small bright tumor regions in FLAIR.

The Dice scores for brainstem are high and consistent across the subjects and comparable to the ones obtained with the healthy whole-brain segmentation method that our method is based on Puonti et al. (2016). Furthermore, the Hausdorff distances are low and consistent as well. For eyes, the Dice scores are generally high, except for a few outliers that were affected by a very thin outer eye wall, and the Hausdorff distances are generally low, indicating a good performance. Hippocampi, on the other hand, have a range of generally lower Dice scores than in Puonti et al. (2016). Their Hausdorff distances are also fairly large. In the majority of the outliers, the method has segmented the hippocampus near to the tumor border while the manual segmentations either lack that hippocampus or have undersegmented it. Finally, the Dice scores for optic nerves and chiasm are generally low and with a large range. These structures are very small and thin, which significantly affects this metric. The Hausdorff distances for these structures are reasonably low however, which indicates that the manual and automatic segmentations are in fact fairly close. The Dice scores for the data combination including CT are higher, due to the better contrast in CT between the optic nerve and surrounding structures.

Fig. 6 shows sagittal slices of two representative segmentations of hippocampi, together with surface plots of the manual and automatic segmentation (for {T1c, CT, FLAIR, T2}). In both cases, the automatic segmentations are larger and seem to capture the hippocampi somewhat better than the manual segmentations. As can be seen in the surface plots, the manual segmentations are not very consistent with each other. The head and subiculum of the hippocampi are also excluded, due to a difference in segmentation protocol compared to the healthy segmentations used to build the method's atlas. To a large extent, this explains the fairly low and inconsistent Dice scores. Another reason for the lower Dice scores compared to Puonti et al. (2016) could be the large slice thickness in FLAIR and T2, which introduces large partial volume effects.

Fig. 7 shows slices of two representative segmentations of the optic system (including eyes, optic nerves and chiasm), together with surface plots of the manual and automatic segmentation (for {T1c, CT, FLAIR, T2}). The method captures the eyes well, although in some cases the wall of the eye is slightly oversegmented. By visual inspection, we found that the method has some difficulties when a subject has the eye lids open, as the solid wall between eye and air becomes very thin. Furthermore, when guided by CT, the method captures the optic nerve (the thin nerve going from an eye in one end to the chiasm in the other end) reasonably well. However, the method has problems in the region where the nerve goes through the skull, as the nerve is especially thin in this region. Because the nerve is thin, the method is also sensitive to intensity ambiguities in the data, such as artifacts or movement of the optic nerve between image acquisitions. In general, the method finds the location of chiasm, but because this structure is so small, the segmentation is to an even larger extent affected by partial volume effects and intensity ambiguities. Finally, the manual segmentations are quite variable in where the borders are placed between the optic nerves and chiasm, as well as between chiasm and the optic tracts (the continuation of the optic system into the brain).

Fig. 8 shows slices of two problematic tumor core segmentations (for data combination {T1c, FLAIR, T2}) that are representative of cases when the method struggles. The first case includes a very large resection at the border of the brain, which the method has difficulty to adapt to for three main reasons: (1) resectioned tumor regions close to the border of the brain can be interpreted as CSF by the method; (2) the method relies on the contrast-enhanced tumor region, which in this case is thin and with weak contrast-enhancement; (3) the method also relies on a bright tumor region in FLAIR, which in this case is small and only slightly brighter than surrounding tissue. In the second case, the method struggles to fill in the inner part of the tumor core. This is an is-

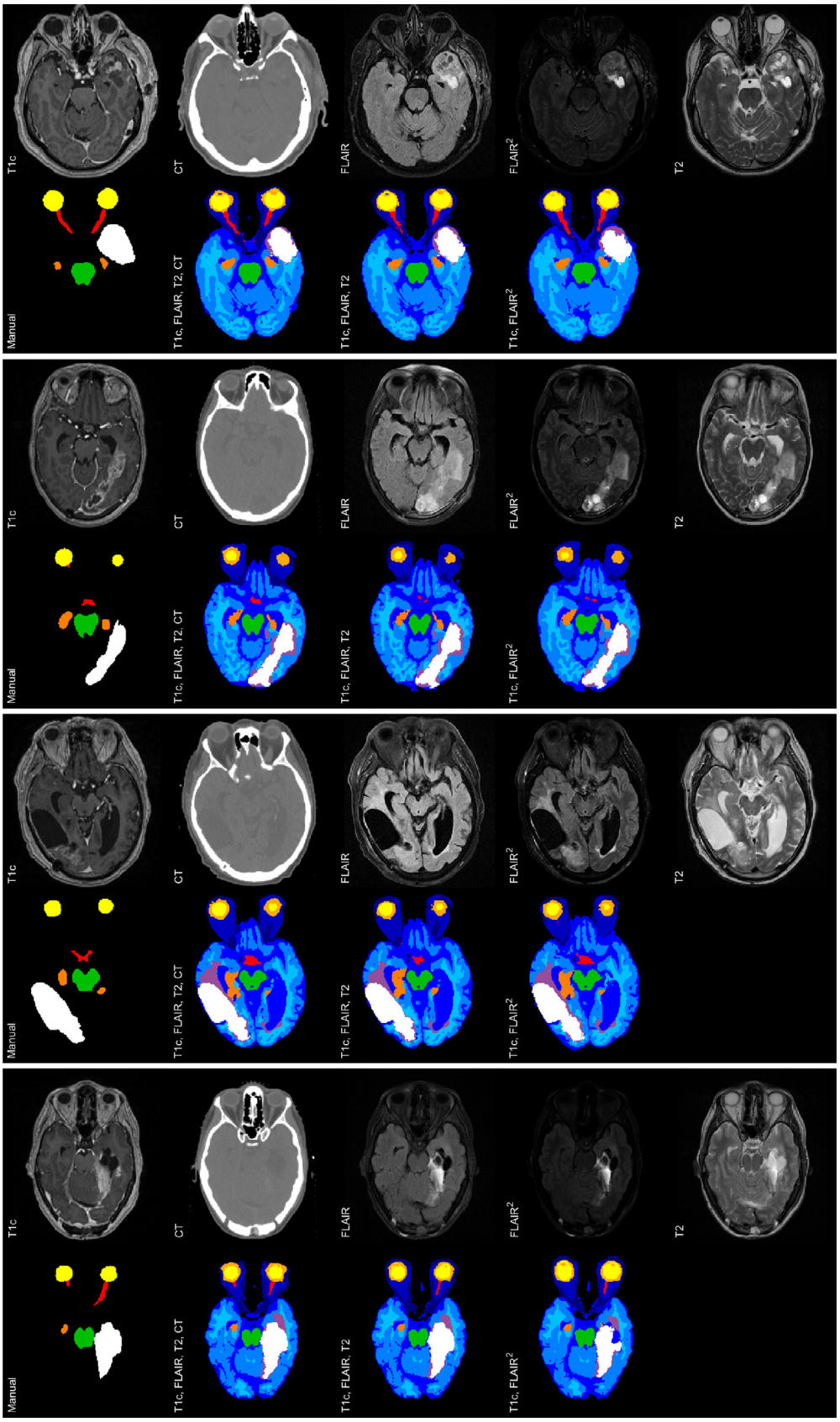

Fig. 4. Segmentations of four representative subjects in the Copenhagen dataset. For each subject, the top row shows slices of the data (from left to right: T1c, CT, FLAIR, FLAIR^2 and T2), whereas the bottom row shows, from left to right, the manual segmentation and automatic segmentations for data combinations {T1c, FLAIR, T2, CT}, {T1c, FLAIR, T2} and {T1c, FLAIR^2}. Label colors: white = TC, lilac = edema, green = BS, dark orange = HC, yellow/light orange = EB, red = ON/CH, shades of blue = other normal labels. For TC in order of appearance: Dice score: {0.68, 0.67, 0.62}, {0.93, 0.93, 0.91}, {0.86, 0.85, 0.85}, {0.61, 0.72, 0.73}, Hausdorff distance: {10, 10, 10}, {2, 3, 5}, {7, 7, 6}, {42, 25, 8}. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

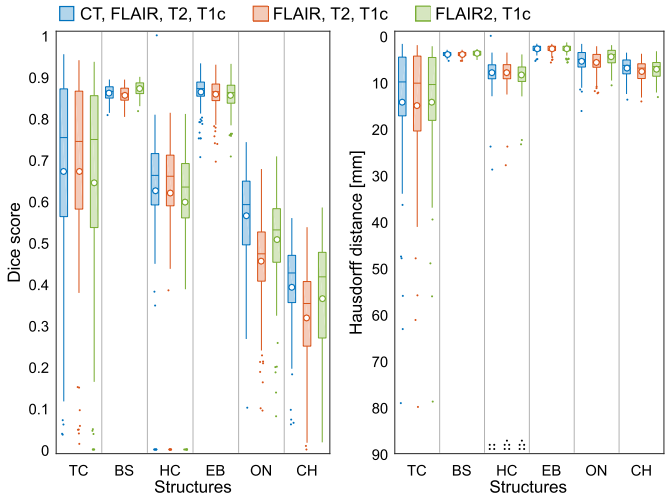

Fig. 5. Boxplots of Dice scores (left) and Hausdorff distances (right) for structures in the Copenhagen dataset, for three data combinations in blue, red and green, respectively. 70 subjects in total. On each box, the central line is the median, the circle is the mean and the edges of the box are the 25th and 75th percentiles. Outliers are shown as dots. Black dots at the bottom of the Hausdorff distance boxplot indicate structures for which scores could not be calculated due to missing ground truth. Note that scores for the left and right structures are included separately in the box plots for HC, EB and ON. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

sue in a few cases where the intensity profile of the inner part of the core is similar to that of edema or healthy tissues.

4.1.2. Dosimetric evaluation

To estimate whether the use of automatic, rather than manual, segmentations introduces any differences in metrics typically reviewed when planning a radiation therapy session, we conduct an additional dosimetric evaluation of our results.

During radiation therapy planning, the segmentations of tumor core (clinically defined as GTV) and OARs are used to optimize a radiation dose distribution that will be used during treatment. Fig. 9 shows an example of such a radiation dose plan. Note that, to form the target to be irradiated, a margin is added around the tumor core to cover likely subclinical spread of tumor cells, which is defined as the clinical target volume (CTV). Finally, a margin stemming from any geometrical uncertainties adhering to the treatment planning and radiation delivery is added, and this volume is defined as the planning target volume (PTV). During the treatment planning process, each OAR and target structure (usually only the PTV) is given a dose-volume objective and a priority that varies with the clinical relevance. A more detailed explanation of the dose plan optimization is given in Munck af Rosenschöld et al. (2011).

To assess the delivered dose to different structures, cumulative dose-volume histograms (DVHs) are often used. Each bin in a DVH represents a certain dose and shows the volume percentage of a structure that receives at least that dose. Fig. 10 shows the DVHs of all relevant structures for the example in Fig. 9, i.e., tumor core (GTV), brainstem (BS), hippocampi (HC), eyes (EB), optic nerves (ON), and chiasm (CH). We show DVHs for both the manual and the automatic segmentations for the data combination {T1c, CT, FLAIR, T2}. Although ideally the DVHs for the automatic segmentations would be obtained by recalculating the dose distributions based on these automatic segmentations and then superimposing the manual segmentations on the resulting distributions (Kieselmann et al., 2018), for the current study we simply superimposed the automatic segmentations on the original dose plan instead. The wide margin added around the tumor core means that the hippocampus in the same hemisphere is frequently located almost completely inside the tumor target. This is the case for the example we show, which is why almost half of the hippocampi volume is irradiated as much as the tumor core, as can be seen in Fig. 10. The maximum accepted dose to the optic chiasm and optic nerves during the treatment planning phase is generally 54 Gy, though small volumes may exceed that dose occasionally. Using the automatic segmentation of the optic chiasm, the radiation dose maximum is somewhat above 54 Gy, suggesting some clinically relevant disagreement between the manual and automatic chiasm segmentations.

To ease the comparison of the DVH results of the automatic and manual segmentations for all subjects, we summarize them as in Conson et al. (2014) by using three points in the histograms. To cover a large part of the cumulative histograms, we use the dose at 5% of volume (D5), 50% of volume (D50), and 95% of volume (D95). Fig. 11 shows the summarized results for all structures, with values for the manual segmentations plotted against values for the automatic segmentations. In the plots, the closer a point is to the diagonal line, the closer the results of the manual and au-

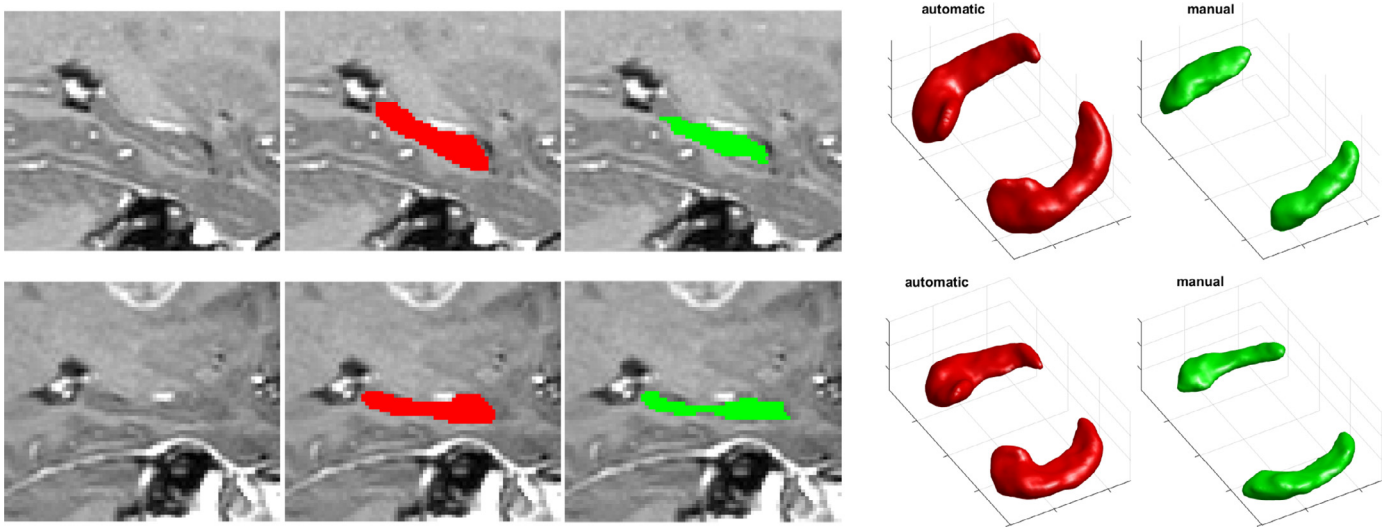

Fig. 6. Hippocampi on two representative subjects in the Copenhagen dataset. Automatic segmentations (for {T1c, CT, FLAIR, T2}) in red and manual segmentations in green. Slice of segmentation overlaid on the T1-weighted scan and 3D surface plot of full structure. For left and right hippocampus: Dice score: {0.54, 0.58}, {0.63, 0.67}; Hausdorff distance: {13, 10}, {8, 7}. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

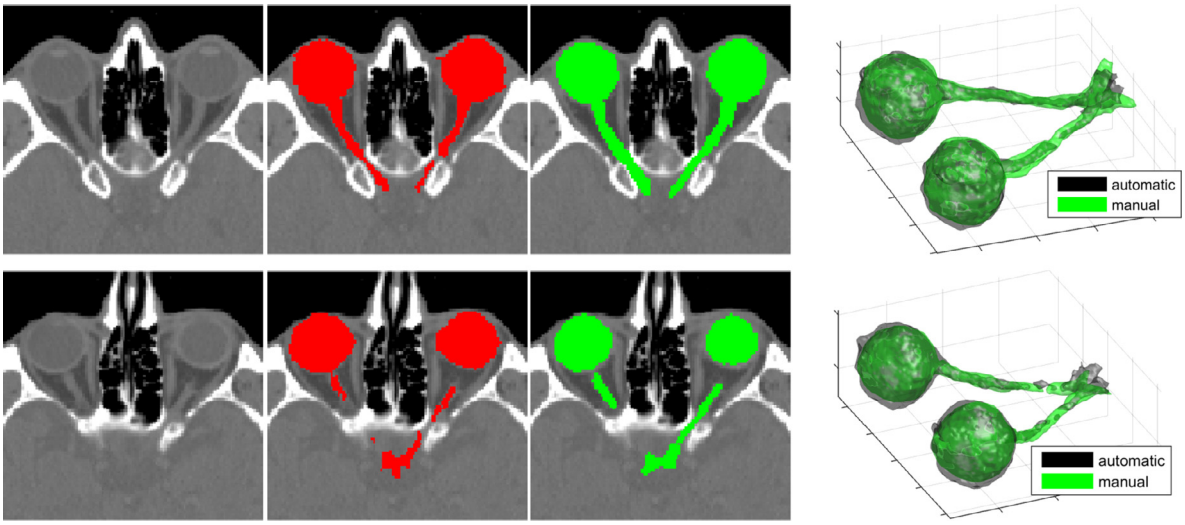

Fig. 7. Optic system on two representative subjects in the Copenhagen dataset. Automatic segmentations (for {T1c, CT, FLAIR, T2}) in red and manual segmentations in green. Slice of segmentation overlaid on the CT scan and 3D surface plot of full structure. For right and left eye; right and left optic nerve; and chiasm: Dice score: {0.91, 0.89}, {0.91, 0.87}, {0.67, 0.67}, {0.48, 0.55} and {0.49, 0.44}; Hausdorff distance: {2, 2}, {2, 2}, {4, 4}, {4, 6} and {4, 6} (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

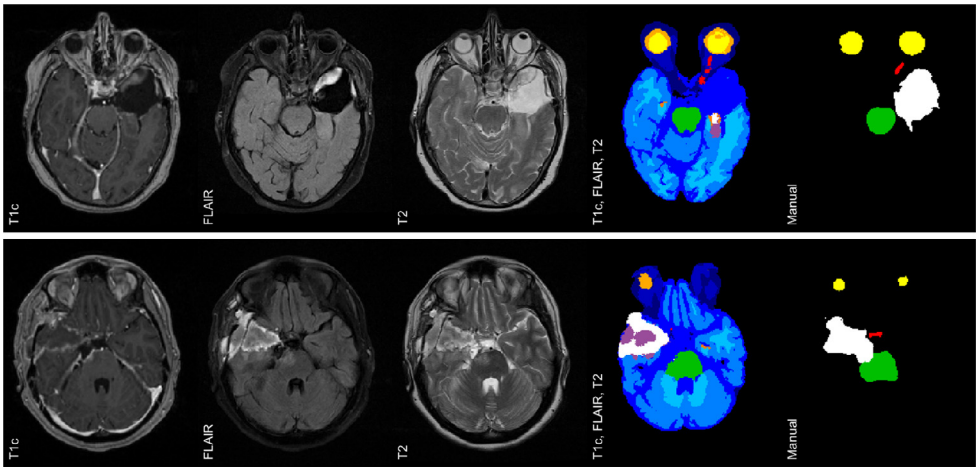

Fig. 8. Two problematic tumor core segmentations in the Copenhagen dataset. Data slices shown together with automatic segmentation (for {T1c, FLAIR, T2}) and manual segmentation. For tumor core: Dice score: {0.04, 0.45}, Hausdorff distance: {41, 28}.

tomatic segmentations are. For tumor core, most points are very close to the line, which is unsurprising considering the wide margin added around tumor. The four D95 outliers belong to subjects where small regions in the brain were erroneously segmented as tumor core by our method, for some cases because of co-occurring pathologies. The results for the organs-at-risk largely confirm the findings using Dice scores and Hausdorff distances. Brainstem and eyes are delineated in close agreement, and the issue with over-segmentation when the outer eye wall is very thin does not affect the dosimetric measure, because that region will always be far away from tumor. The results for hippocampi are varying for subjects where a hippocampus is on the border of the tumor target, mainly due to the difference in protocol between the manual and automatic segmentations. Furthermore, the results for optic nerves vary widely for a few subjects. However, at the maximum dose target of 54 Gy the results of the manual and automatic segmentations match fairly well. For the optic chiasm, on the other hand, some results for the automatic segmentations are significantly beyond its dose objective of maximum 54 Gy. This suggests that significant differences to treatments could be expected if the automatic segmentation of this structure would be used instead of the manual segmentation when optimizing the radiation dose plan.

4.1.3. Comparing our tumor prior to first-order MRFs

To demonstrate the benefits of modeling high-order interactions with the cRBM-based tumor prior, we will contrast it to a tumor prior based on more traditional first-order MRFs. As mentioned before, first-order MRFs only have pairwise clique potentials, compared to the potentials in cRBMs that are defined over groups of voxels as large as the size of the convolutional filters. The inference of the model is kept exactly the same except for the tumor prior in Algorithm 1: there are no hidden units to sample and therefore the labels in a voxel i are sampled from

$$p_i\left(l_i, z_i, y_i | \mathbf{d}_i, \boldsymbol{\theta}, \hat{\boldsymbol{\eta}}\right) \propto p_i\left(\mathbf{d}_i | l_i, z_i, y_i, \boldsymbol{\theta}\right) \pi_i(l_i) \exp\left[-\beta^z \sum_{j \in \mathfrak{N}_i} |z_i - z_j|\right] \exp\left[-\beta^y \sum_{j \in \mathfrak{N}_i} |y_i - y_j|\right] \exp\left[-f(l_i, z_i, y_i)\right],$$

where $\mathfrak{N}_i$ is the set of 26 voxels that form neighboring pairs with voxel i.

To find suitable values for the user-tunable hyperparameters β_z and β_y, we performed a grid search with steps of 0.5 using the same 30 manually segmented BRATS training subjects that we used for training the cRBMs (see Section 2.1.1). For each hyperparameter combination, we segmented the subjects using the

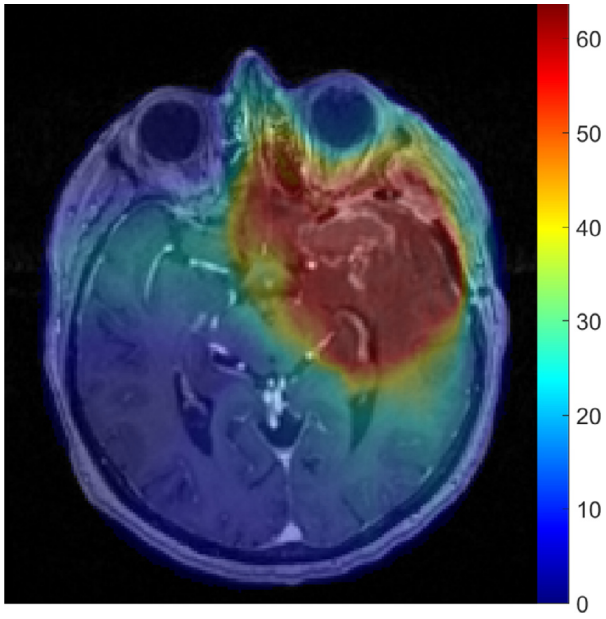

Fig. 9. A radiation dose plan overlaid on a T1c image slice for a representative subject. The dose is measured in Gy.

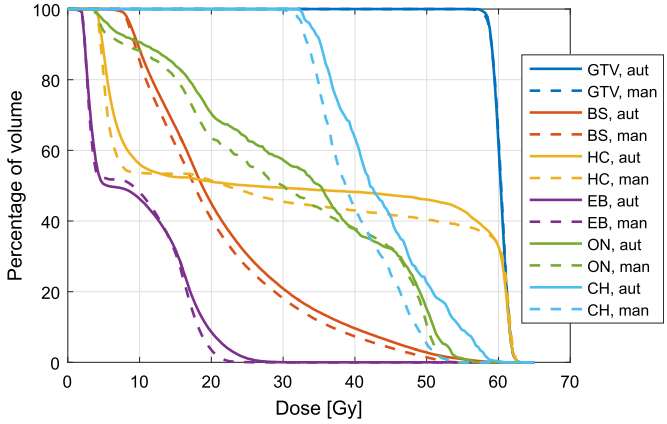

Fig. 10. Dose volume histogram (DVH) of several structures for the representative subject in Fig. 9, i.e., tumor core (GTV), brainstem (BS), hippocampi (HC), eyes (EB), optic nerves (ON), and chiasm (CH). Solid lines and broken lines correspond to automatic and manual segmentations, respectively. Note that all DVHs were computed using the original treatment dose plan, which was based on the manual segmentations. .

method with first-order MRFs. By comparing the average performance (using Dice scores and Hausdorff distances) we found the combination $\{\beta_z = 4, \beta_y = 1\}$ to have the best overall performance. With these optimized hyperparameter values, we compare the tumor core segmentation performance on the data combination {T1c, FLAIR, T2} when using the two different priors. The average and median Dice score for the cRBM-based method is 0.67 and 0.74 respectively, compared to 0.58 and 0.57 when using the first-order MRFs described here. Furthermore, the average and median Hausdorff distance for the cRBM-based method is 14 mm and 10 mm respectively, compared to 23 mm and 17 mm when using first-order MRFs. This demonstrates the benefit of modeling high-order interactions among voxels.

4.2. Results for 2015 BRATS test dataset

To further evaluate our method's performance on segmenting tumors and compare it to that of other methods, we use the test dataset of the 2015 BRATS challenge – at the time of writing the

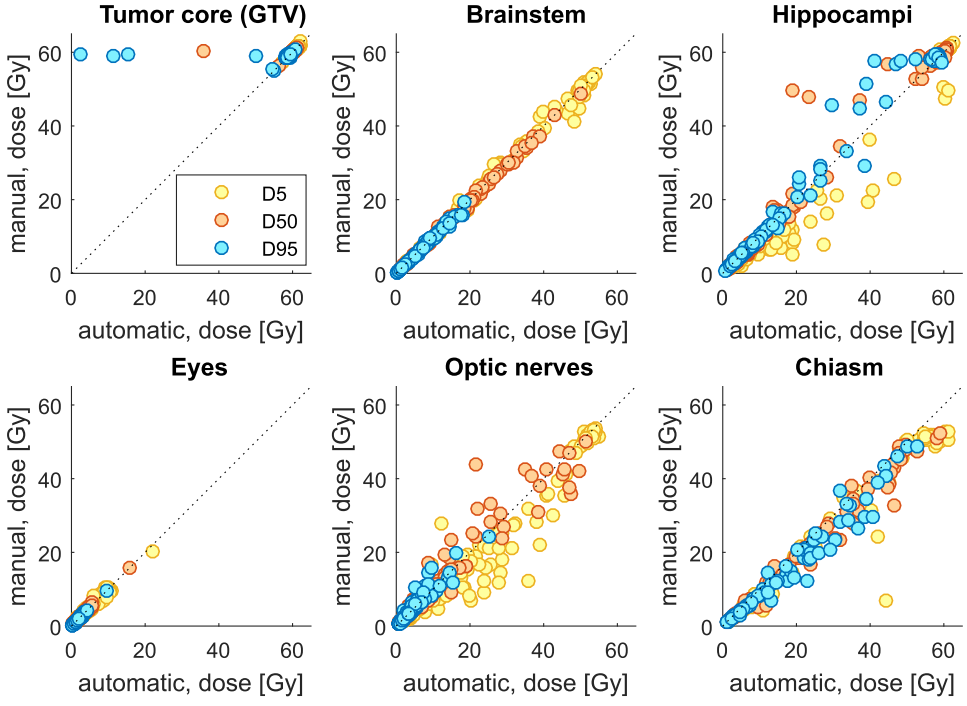

Fig. 11. Summary statistics of DVH results for all subjects and structures, showing 5% volume (D5), 50% volume (D50), and 95% volume (D95), for manual versus automatic segmentations. Note that left and right hippocampus, eye and optic nerve are included as separate points in their respective plots.

latest edition with datasets available to us.[2] We participated in this challenge and were among the top-performing methods out of a total of 12 methods. This dataset includes non-enhanced T1 scans, which the dataset in Section 4.1 lacks, and data with varying magnetic field strength and resolution from several imaging centers. The dataset is skull-stripped, so we merge all non-brain labels used in our method into the background label. We stress that we did not need to change anything else in our method.

The dataset is publicly available at the virtual skeleton online platform (Kistler et al., 2013). It consists of 53 patients with varying high- and low-grade gliomas, and a mix of pre-operative and post-operative scans. The included MR contrasts are T2-weighted FLAIR (2D acquisition), T2-weighted (2D acquisition), T1-weighted (2D acquisition) and T1-weighted with contrast enhancement (T1c, 3D-acquisition). All data were resampled to 1 mm isotropic resolution, aligned to the same anatomical template and skull-stripped by the challenge organizers. The dataset includes manual annotations of four tumor regions, which are not publicly available. Instead, the performance of a method can be evaluated by uploading segmentations to the online platform. On the online platform and during the challenge, scores are reported on enhanced core, core (which includes enhanced core and other core regions), and whole tumor (which includes core and edema).

Fig. 12 shows slices of three representative segmentations with: T1c, FLAIR, T2 and T1, and the segmentation by our method as presented in this paper. Note that the manual segmentations compared against are not publicly available. We can see that the atlas deforms well to the subjects, and brainstem and hippocampi are well-captured. Furthermore, our method can segment brain tumors with large variations in size, location and appearance. Also note the low resolution and image quality in some of the images.

For the purpose of comparing against the manual segmentations, we focus on the core region, as this corresponds to the GTV used in radiation therapy. We compare the performance of our method to that of three other top-performing tumor segmentation methods that also participated in the 2015 BRATS challenge.

(1) GLISTRboost (Bakas et al., 2016): This semi-automated method is based on a modified version of the generative atlas-based method GLISTR (Kwon et al., 2014; Gooya et al., 2012), which uses a tumor growth model. The method requires manual input of a seed-point for each tumor center and a radius of the extent of the tumor. To increase the segmentation performance, the method is extended with a discriminative post-processing step using a gradient boosting multi-label classification scheme followed by a patient-wise refinement step.

(2) Grade-specific CNNs (Pereira et al., 2016): This semi-automated method uses a discriminative 2D Convolutional Neural Network (CNN) approach. The method takes advantage of the fact that high- and low-grade tumors exhibit differences in intensity and spatial distribution. To do this, it uses two CNNs: one trained on high-grade tumors and one trained on low-grade tumors. The CNN to use for a specific subject is then chosen manually based on visual assessment, which is the only manual step in the method.

(3) Two-way CNN (Havaei et al., 2017): This fully automated method uses a similar discriminative 2D CNN approach to the previous method. The method forms a cascaded architecture with two parts, where the voxel-wise label predictions from the first part are added as additional input to the second part. Each part has two pathways, where intensity features are automatically learned: one learning local details of tumor appearance and one learning larger contexts.

[2] We note that the more recent BRATS 2017 and 2018 editions have since released new training and benchmark datasets; in Section 5 we will briefly discuss the results we report here in the context of these more recent challenges.

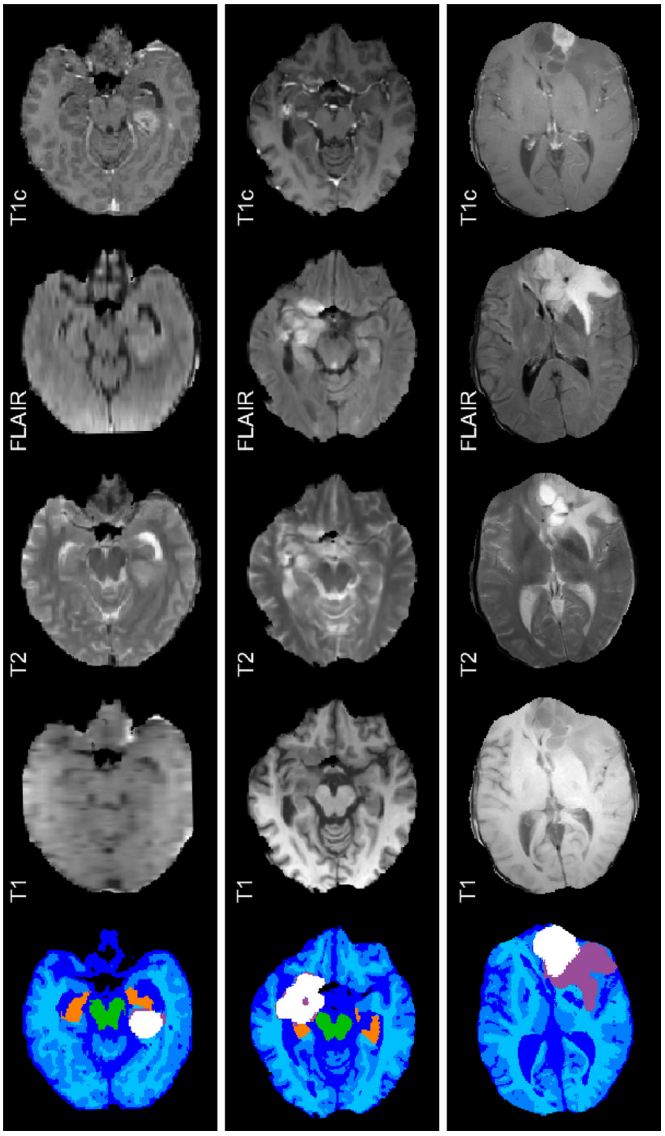

Fig. 12. Three representative segmentations in the BRATS test dataset. Slices of T1c, FLAIR, T2, T1, and automatic segmentation. Label colors: white = TC, lilac = edema, green = BS, dark orange = HC, shades of blue = other brain tissues. Note that the images are skull-stripped by the BRATS challenge organizers. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 13 shows box plots of the Dice scores and Hausdorff distances for tumor core. We show scores for our method and the three benchmark methods as reported at the challenge. The scores for our method are for the version we participated with in the challenge, as presented in Agn et al. (2016b). The main difference, compared to the current version, is the use of an affinely registered atlas, instead of the mesh-based deformable atlas presented in this paper to enable a detailed segmentation of normal head structures. This, however, does not significantly affect the tumor segmentation; we also segmented the dataset with our current version and obtained similar Dice scores from the online platform, with just a 4% increase in the average Dice score. As seen in the figure, the range of Dice scores is similar to our results in Section 4.1 (Fig. 5), which shows that our method readily adapts to the included non-enhanced T1 scans and data from different imaging centers. Comparing to the other benchmark methods, our method performs significantly better on tumor core when considering Dice scores. The range of values are large for all methods, illustrating the

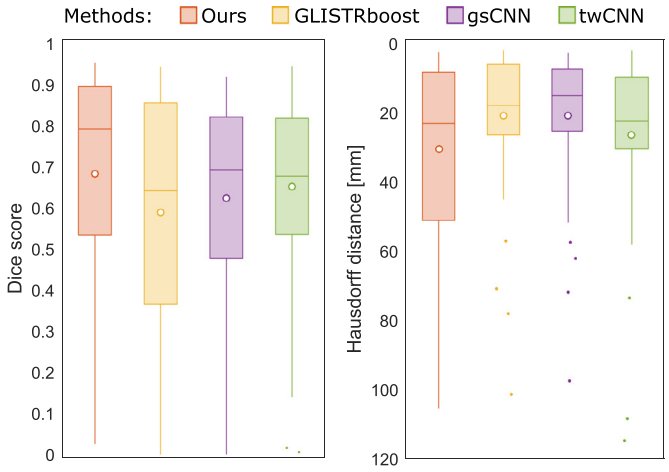

Fig. 13. Box plots of Dice scores and Hausdorff distances for tumor core on the BRATS 2015 test dataset. 53 subjects in total. Scores are as reported in the challenge. On each box, the central line is the median, the circle is the mean and the edges of the box are the 25th and 75th percentiles. Outliers are shown as dots.

difficulty of segmenting tumors. This dataset includes a number of subjects with large resections and a wide variety of tumors, e.g., low-grade tumors that have been shown to be difficult to segment in Menze et al. (2015). The Hausdorff distances for our method are somewhat worse than for the other methods, which could be explained by a better capability of their methods to remove small erroneous tumor clusters, e.g., because of the deep architecture in a CNN. The Hausdorff distances for our method are also worse for this dataset than for the dataset in Section 4.1 (cf. Fig. 5), which is explained by the generally lower resolution and image quality.

4.3. Results for London dataset

As a final experiment, we investigate the ability of our method to adapt to yet a different set of acquired images using the London dataset. In contrast to the other datasets, this one completely lacks T1-weighted images and includes a new MR sequence: double inversion recovery (DIR). The data set consists of seven patients with varying low- and high-grade gliomas, which were scanned with a Siemens Trio 3T scanner at the National Hospital for Neurology and Neurosurgery, UCLH NHS Foundation Trust, London, as part of a registered clinical audit. The following MR images were acquired with 1 mm isotropic resolution: T2-weighted (3D acquisition) and T2-weighted DIR (3D-acquisition). We use exactly the same settings in our method for the DIR images as we would for FLAIR, without any changes. As no manual segmentation has been performed on this dataset, we only perform a qualitative analysis of the results.

Fig. 14 shows slices for three representative subjects with DIR, T2 and the method's segmentation. As seen in the Figure, our method can easily segment datasets that lack T1-weighted images and include a DIR image instead of FLAIR without any changes to the method. Visual inspection of all seven segmentations revealed no significant deviations from other results presented in this paper.

5. Discussion and conclusion

In this paper, we have presented a generative method for simultaneous segmentation of brain tumors and an extensive set of organs-at-risk (OARs) applicable to radiation therapy planning for glioblastomas. To the best of our knowledge, this is the first time a segmentation method has been presented that encompasses both brain tumors and OARs within the same modeling framework. The

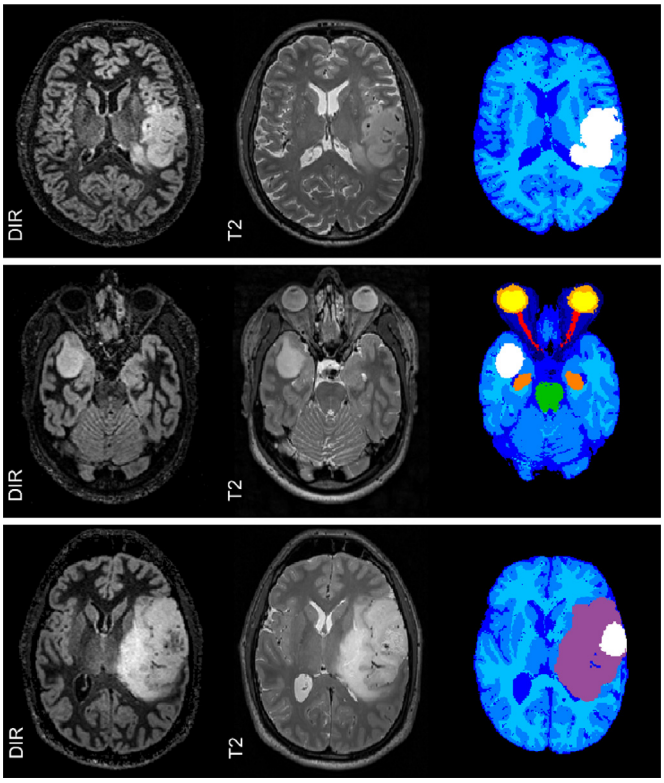

Fig. 14. Three representative segmentations in the London dataset. Slices of DIR, T2 and automatic segmentation.

method combines a previously validated atlas-based model for detailed segmentation of normal brain structures with a model for brain tumor segmentation based on convolutional restricted Boltzmann machines (cRBMs). In contrast to generative lesion shape models proposed in the past, cRBMs are capable of modeling long-range spatial interactions among tumor voxels. Furthermore, by completely separating the modeling of anatomy from the modeling of image intensities, the method is able to adapt to heterogeneous data that differs substantially from any available training data, including unseen (e.g., CT or FLAIR^2) or missing (e.g., T1) contrasts.

Although the method we propose is demonstrated to be applicable across data with various image contrast properties without retraining, it does rely on contrast-specific settings to constrain and to initialize tumor-specific appearance parameters, especially in the MR-sequences FLAIR and T1c (see Tables 3 and 4, respectively). We found that this was necessary to guide the model to the correct intensities for tumor in these sequences, which are typically acquired for brain tumor imaging. Ideally, such hand-crafted rules would be replaced by a prior on model parameters that can be learned automaticaly from example cases; however, because tumor appearance can vary widely across subjects, robustly establishing such a prior may be challenging. With the current setup, our results demonstrate that the same settings work robustly across FLAIR and T1c images acquired with a variety of scanners and imaging protocols, and even when FLAIR is replaced with FLAIR^2 or DIR. In data where FLAIR and/or T1c is entirely missing, however, the method may need to be adjusted by modifying the corresponding lines in Tables 3 and 4.

Our experiments show that the method's performance in segmenting tumors is comparable to that of some of the best methods benchmarked by the BRATS 2015 challenge. We note that, since the time of writing, the more recent BRATS 2017 and 2018 editions have released new training and benchmark datasets, and that top-performing methods in these challenges obtain significantly

better Dice and Hausdorff scores than the ones reported here. However, some care is needed when comparing the results of the various BRATS challenges. Unlike the 2015 edition, which contained a mix of pre- and post-operative scans including several cases with large resections, the more recent editions only involve pristine, pre-operative cases which are arguably more uniform and somewhat easier to segment. This difference is especially important in the given context of radiation therapy planning of glioblastoma patients, where the vast majority of patients has undergone resective surgery (Davis, 2016; Munck af Rosenschöld et al., 2014) so that segmentation performance on pre-operative scans only (as benchmarked by BRATS 2017 and 2018) is less relevant. A second difference between older and newer BRATS challenges is that the number of manually annotated subjects available for training models differs by almost an order of magnitude (285 in 2017–2018, vs. the 30 from 2015 we used for the current paper), making the obtained numerical scores difficult to compare directly. While the cRBM tumor shape model proposed in the current paper is still fairly local, a ten-fold increase in manually annotated training data should allow one to use generative shape models with a deeper structure, such as variational autoencoders (Kingma and Welling, 2013), which could potentially eliminate the occasional false-positive tumor detections that remain for the current method. Nevertheless, within the given application area of radiation therapy planning, it is worth remembering that further increases in segmentation overlap scores may not necessarily translate into meaningful improvements in radiation therapy delivery, given the wide margins that are added around the tumor to obtain final radiation target volumes. Indeed, the results shown in Figs. 10 and 11 (top left) indicate that, with the exception of a few outliers, tumor segmentation performance of the current method may already be quite adequate for this specific purpose.

In addition to delineating tumors, the proposed method is also capable of segmenting the OARs hippocampi, brainstem, eyes, optic nerves and optic chiasm. We quantitatively evaluated out method's OAR segmentation performance in 70 patients with manual segmentations used when planning a radiation therapy session. The evaluation showed a generally good performance in segmenting hippocampi (HC), brainstem (BS) and eyes (EB); but lower performance in segmenting the very small structures optic nerves (ON) and chiasm (CH). The overall performance of our method (average Dice scores for BS: 0.86, EB: 0.86, ON: 0.56, CH: 0.39 when using the image combination {CT, T1c, FLAIR,T2}) is comparable to the human inter-rater variability reported in Deeley et al. (2011), where eight experts segmented OARs in 20 high-grade glioma patients, with average Dice scores BS: 0.83, EB: 0.84, ON: 0.50, CH: 0.39. It is clear that the Dice scores for optic nerves and chiasm can be low even for experts. Nevertheless, the dosimetric evaluation and visual inspection of our automated segmentation of these structures point to the need for further research to obtain better results. An improvement could possibly be achieved by incorporating dedicated geometrical information in the prior, e.g., about the tubular structure of the optic system which was successfully used in Noble and Dawant (2011).

Using manual segmentations from radiation therapy planning as ground truth complicates our findings, as these segmentations themselves might be suboptimal with large inter-rater variability. Different clinics might also use differing delineation protocols. In our experiments, the Dice score for hippocampi was significantly affected by differing delineation protocols between the experts at the clinic and the expert segmentations used to train the atlas in our method. The manual segmentations at the clinic were also found to be of variable quality in regions where the segmented structures have a similar intensity profile to neighboring structures – such as the chiasm and brainstem compared to neighboring white matter structures. Additionally, structures far away from a tumor are sometimes not carefully delineated because they will not significantly affect the radiation therapy plan anyway.

The segmentation method we proposed in this paper can be further extended in a number of ways. First, the original segmentations we used to train our atlas for normal brain structures include dozens of segmented structures. The method could directly handle any of these structures by simply retraining the atlas on segmentations in which these structures have not been merged into global catch-all labels as we did in the current paper. This may be helpful if additional OARs need to be segmented or for automating CTV decisions based on anatomical context (Unkelbach et al., 2014). A detailed whole-brain segmentation can also be useful for training outcome prediction models, e.g., to study the effect of the radiation received by various structures on cognition (Conson et al., 2014). A second aspect that we did not explore in the current work is the method's innate ability to quantify uncertainty in the produced segmentations, by analyzing the variation across the MCMC segmentation samples instead of simply retaining the mode in each voxel. As shown in Lê et al. (2016, 2017), uncertainty in segmentation boundaries of tumors and OARs can be propagated onto uncertainty in radiation dose distributions, which has interesting potential applications in the optimization and the personalization of radiation therapy planning. In such applications, however, it will likely be imperative to also take into account the uncertainty on atlas deformations instead of using a point estimate for the atlas node positions $\boldsymbol{\eta}$, as we did in the current work, for instance by using the Hamiltonian Monte Carlo (Duane et al., 1987) technique we used for this purpose in Iglesias et al. (2013).

Segmenting one subject with the proposed method currently takes around 40 min. Although a manual delineation procedure is typically faster, the method can still be a useful aid in the clinical work flow, as no manual input is needed before or during the segmentation procedure. A further speed-up would be necessary to use the method for continuous segmentation during an image-guided radiation therapy session (Lagendijk et al., 2014). Since the implementation used in this paper has mainly been focused on demonstrating the feasibility of the method rather than optimizing speed, a further speed-up would be expected with a more efficient implementation, especially with one that utilizes GPUs.

Declarations of interest

None.

Acknowledgments

This research was supported by the NIH NCRR (P41RR14075, 1S10RR023043), NIBIB (R01EB013565) and the Lundbeck foundation (R141-2013-13117). This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 765148. The Wellcome Centre for Human Neuroimaging is supported by core funding from the Wellcome Trust (203147/Z/16/Z).

Appendix A. Sampling from $p(\boldsymbol{\theta}|\mathbf{l}, \mathbf{z}, \mathbf{y}, \mathbf{D})$

Here we describe how we sample from $p(\boldsymbol{\theta}|\mathbf{l}, \mathbf{z}, \mathbf{y}, \mathbf{D})$ in the blocked Gibbs sampler used in Section 2.3.

Table 3 specifies a number of linear constraints on the Gaussian means $\{\boldsymbol{\mu}_{xg}\}$ in the prior $p(\boldsymbol{\theta})$, encoding prior knowledge about tumor appearance relative to normal brain tissue. Stacking all Gaussian means into a single vector $\boldsymbol{\mu} = (\ldots, \boldsymbol{\mu}_{xg}^T, \ldots)^T$ allows us to express these constraints in the form

$$\mathbf{A}\boldsymbol{\mu} \leq \mathbf{b},$$

where the values in each row of $\mathbf{A}$ and $\mathbf{b}$ are chosen to match the corresponding line in Table 3.

Introducing the "one-hot" auxiliary variable $\mathbf{t}_i = \{t_i^{xg}\}$ to indicate which individual Gaussian component the ith voxel is associated with (t_i^{xg} has value one when the voxel belongs to the gth component of the xth GMM, and zero otherwise) the target distribution is obtained as a marginal distribution of $p(\boldsymbol{\theta}, \{\mathbf{t}_i\}|\mathbf{l}, \mathbf{z}, \mathbf{y}, \mathbf{D})$: $p(\boldsymbol{\theta}|\mathbf{l}, \mathbf{z}, \mathbf{y}, \mathbf{D}) = \sum_{\{\mathbf{t}_i\}} p(\boldsymbol{\theta}, \{\mathbf{t}_i\}|\mathbf{l}, \mathbf{z}, \mathbf{y}, \mathbf{D})$. Therefore, samples of $p(\boldsymbol{\theta}|\boldsymbol{l}, \boldsymbol{zy}, \boldsymbol{D})$ can be obtained with a blocked Gibbs sampler of $p(\boldsymbol{\theta}, \{\boldsymbol{t}_i\}|\boldsymbol{l}, \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{D})$ cyclically sampling from the following conditional distributions and subsequently discarding the samples of $\{\mathbf{t}_i\}$:

$$p\big(\{\mathbf{t}_i\}|\boldsymbol{\theta}, \mathbf{l}, \mathbf{z}, \mathbf{y}, \mathbf{D}\big) = \prod_i p\big(\mathbf{t}_i|\boldsymbol{\theta}, x(l_i, z_i, y_i), \mathbf{d}_i\big) \tag{A.1}$$

$$\text{with} \quad p\big(\mathbf{t}_i|\boldsymbol{\theta}, x, \mathbf{d}_i\big) = \frac{\sum_{g=1}^{G_x} t_i^{xg} \gamma_{xg} \mathcal{N}\big(\mathbf{d}_i|\boldsymbol{\mu}_{xg} + \mathbf{C}\boldsymbol{\phi}_i, \boldsymbol{\Sigma}_{xg}\big)}{\sum_{g=1}^{G_x} \gamma_{xg} \mathcal{N}\big(\mathbf{d}_i|\boldsymbol{\mu}_{xg} + \mathbf{C}\boldsymbol{\phi}_i, \boldsymbol{\Sigma}_{xg}\big)},$$

$$p\big(\{\boldsymbol{\gamma}_x\}|\boldsymbol{\theta}_{\setminus\{\boldsymbol{\gamma}_x\}}, \mathbf{t}, \mathbf{l}, \mathbf{z}, \mathbf{y}, \mathbf{D}\big) = \prod_x \mathrm{Dir}\big(\boldsymbol{\gamma}_x|\{\alpha_{xg}\}_{g=1}^{G_x}\big), \tag{A.2}$$

$$p\big(\{\boldsymbol{\mu}_{xg}\}|\boldsymbol{\theta}_{\setminus\{\boldsymbol{\mu}_{xg}\}}, \mathbf{t}, \mathbf{l}, \mathbf{z}, \mathbf{y}, \mathbf{D}\big) \propto \begin{cases} \mathcal{N}\big(\boldsymbol{\mu}|\mathbf{m}_\mu, \mathbf{S}_\mu\big) & \text{if } \mathbf{A}\boldsymbol{\mu} \le \mathbf{b} \\ 0 & \text{otherwise,} \end{cases} \tag{A.3}$$

$$p\big(\{\boldsymbol{\Sigma}_{xg}\}|\boldsymbol{\theta}_{\setminus\{\boldsymbol{\Sigma}_{xg}\}}, \mathbf{t}, \mathbf{l}, \mathbf{z}, \mathbf{y}, \mathbf{D}\big) = \prod_x \prod_g \mathrm{IW}(\boldsymbol{\Sigma}_{xg}|\mathbf{S}_{xg}, \upsilon_{xg}), \tag{A.4}$$

and finally

$$p\big(\mathbf{C}|\boldsymbol{\theta}_{\setminus\mathbf{C}}, \mathbf{t}, \mathbf{l}, \mathbf{z}, \mathbf{y}, \mathbf{D}\big) = \mathcal{N}(\mathbf{c}|\mathbf{m}_c, \mathbf{S}_c) \quad \text{with} \quad \mathbf{c} = \begin{pmatrix} \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_N \end{pmatrix}. \tag{A.5}$$

Here we have defined the following variables:

$$\alpha_{xg} = \alpha_0 + N_{xg} \quad \text{with} \quad N_{xg} = \sum_i t_i^{xg}$$

$$\mathbf{S}_\mu = \begin{pmatrix} \ddots & & \\ & N_{xg}^{-1}\boldsymbol{\Sigma}_{xg} & \\ & & \ddots \end{pmatrix}$$

$$\mathbf{m}_\mu = \begin{pmatrix} \vdots \\ \mathbf{m}_{xg} \\ \vdots \end{pmatrix} \quad \text{with} \quad \mathbf{m}_{xg} = \frac{\sum_i t_i^{xg}\big(\mathbf{d}_i - \mathbf{C}\boldsymbol{\phi}_i\big)}{N_{xg}}$$

$$\mathbf{S}_{xg} = \mathbf{S}_x^0 + \sum_i t_i^{xg}\big(\mathbf{d}_i - \mathbf{C}\boldsymbol{\phi}_i - \boldsymbol{\mu}_{xg}\big)\big(\mathbf{d}_i - \mathbf{C}\boldsymbol{\phi}_i - \boldsymbol{\mu}_{xg}\big)^T$$

$$\upsilon_{xg} = \upsilon_x^0 + N_{xg}$$

$$\mathbf{S}_c = \begin{pmatrix} \boldsymbol{\Phi}^T\mathbf{W}^{11}\boldsymbol{\Phi} & \cdots & \boldsymbol{\Phi}^T\mathbf{W}^{1N}\boldsymbol{\Phi} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Phi}^T\mathbf{W}^{N1}\boldsymbol{\Phi} & \cdots & \boldsymbol{\Phi}^T\mathbf{W}^{NN}\boldsymbol{\Phi} \end{pmatrix}^{-1}$$

$$\text{and} \quad \mathbf{m}_c = \mathbf{S}_c \begin{pmatrix} \boldsymbol{\Phi}^T\big(\sum_{n=1}^N \mathbf{W}^{1n}\mathbf{r}^{1n}\big) \\ \vdots \\ \boldsymbol{\Phi}^T\big(\sum_{n=1}^N \mathbf{W}^{Nn}\mathbf{r}^{Nn}\big) \end{pmatrix},$$

$$\text{where} \quad \boldsymbol{\Phi} = \begin{pmatrix} \phi_1^1 & \cdots & \phi_P^1 \\ \vdots & \ddots & \vdots \\ \phi_1^I & \cdots & \phi_P^I \end{pmatrix} \quad \text{and} \quad \mathbf{W}^{mn} = \mathrm{diag}\big(w_i^{mn}\big)$$

$$\text{with} \quad w_i^{mn} = \sum_x \sum_{g=1}^{G_x} w_{ixg}^{mn}, \quad w_{ixg}^{mn} = t_i^{xg}\big(\boldsymbol{\Sigma}_{xg}^{-1}\big)_{mn},$$

$$\mathbf{r}^{mn} = (r_1^{mn}, \ldots, r_I^{mn})^T, \quad r_i^{mn} = d_i^n - \frac{\sum_x \sum_{g=1}^{G_x} w_{ixg}^{mn}\big(\boldsymbol{\mu}_{xg}\big)_n}{w_i^{mn}}.$$

In order to sample from the truncated multivariate Gaussian distribution in Eq. (A.3), we use the Gibbs sampling approach proposed in Kotecha and Djuric (1999) and Rodriguez-Yam et al. (2004), which cycles through the conditional distributions of each component of $\boldsymbol{\mu}$ and samples from the corresponding truncated univariate normal distributions using inverse transform sampling.

In our implementation, rather than repeating the Gibbs sampler steps described in Eqs. (A.1)–(A.5) until the Markov chain reaches equilibrium and an independent sample of $\boldsymbol{\theta}$ is obtained, we only make a single sweep before obtaining new samples of $\mathbf{H}^y$, $\mathbf{H}^z$, and $\{\mathbf{l}, \mathbf{z}, \mathbf{y}\}$ in the main loop described in Algorithm 1, effectively implementing a so-called partially collapsed Gibbs sampler (Van Dyk and Park, 2008).

Appendix B. Optimizing likelihood parameters in GEM algorithm

Here we describe how we optimize the likelihood parameters $\boldsymbol{\theta}$ for a given value of the atlas node positions $\boldsymbol{\eta}$ in the simplified model of the label prior described in Section 2.3.2.

We use a generalized expectation-maximization (GEM) algorithm (Dempster et al., 1977) that is very similar to the ones proposed in Van Leemput et al. (1999b) and Puonti et al. (2016). In short, the algorithm iteratively updates the various components of $\boldsymbol{\theta}$ to the mode of the conditional distributions given by Eqs. (A.1)–(A.5):

$$\gamma_{xg} \leftarrow \frac{\alpha_{xg} - 1}{\sum_{g'=1}^{G_x}\big(\alpha_{xg'} - 1\big)}, \quad \forall x, g$$

$$\boldsymbol{\mu} \leftarrow \arg\max_{\boldsymbol{\mu}} \Big[\big(\boldsymbol{\mu} - \mathbf{m}_\mu\big)^T \mathbf{S}_\mu^{-1}\big(\boldsymbol{\mu} - \mathbf{m}_\mu\big)\Big] \quad \text{s.t.} \quad \mathbf{A}\boldsymbol{\mu} \le \mathbf{b} \tag{B.1}$$

$$\boldsymbol{\Sigma}_{xg} \leftarrow \frac{\mathbf{S}_{xg}}{\upsilon_{xg} + N + 1}, \quad \forall x, g$$

$$\mathbf{c} \leftarrow \mathbf{m}_c$$

where the "one-hot" auxiliary variables $\{\mathbf{t}_i\}$ are replaced by their expected values:

$$t_i^{xg} = \frac{\gamma_{xg}\mathcal{N}\big(\mathbf{d}_i|\boldsymbol{\mu}_{xg} - \mathbf{C}\boldsymbol{\phi}_i, \boldsymbol{\Sigma}_{xg}\big)p_i(x|\boldsymbol{\eta})}{\sum_{x'=1}^{X} p_i\big(\mathbf{d}_i|x', \boldsymbol{\theta}\big)p_i(x'|\boldsymbol{\eta})}, \quad \forall x, g, i. \tag{B.2}$$

Solving Eq. (B.1) is a so-called quadratic programming problem, for which an implementation is directly available in MATLAB.

References

Agn, M., Law, I., af Rosenschöld, P.M., Van Leemput, K., 2016a. A generative model for segmentation of tumor and organs-at-risk for radiation therapy planning of glioblastoma patients. In: Proceedings of the SPIE Medical Imaging. International Society for Optics and Photonics. 97841D–97841D.

Agn, M., Puonti, O., Munck af Rosenschöld, P., Law, I., Van Leemput, K., 2016b. Brain tumor segmentation using a generative model with an rbm prior on tumor shape. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: First International Workshop, Brainles 2015, Held in Conjunction with MICCAI, Munich, Germany Revised Selected Papers. Springer, pp. 168–180.

Ashburner, J., Andersson, J.L., Friston, K.J., 2000. Image registration using a symmetric prior – in three dimensions. Hum. Brain. Mapp 9 (4), 212–225.

Bakas, S., Zeng, K., Sotiras, A., Rathore, S., Akbari, H., Gaonkar, B., Rozycki, M., Pati, S., Davatzikos, C., 2016. Glistrboost: combining multimodal mri segmentation, registration, and biophysical tumor growth modeling with gradient boosting machines for glioma segmentation. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: First International Workshop, Brainles 2015, Held in Conjunction with MICCAI Munich, Germany, Revised Selected Papers. Springer, pp. 144–155.

Bauer, S., Lu, H., May, C.P., Nolte, L.-P., Büchler, P., Reyes, M., 2013. Integrated segmentation of brain tumor images for radiotherapy and neurosurgery. Int. J. Imaging Syst. Technol. 23 (1), 59–63.
Bauer, S., Nolte, L.-P., Reyes, M., 2011. Fully automatic segmentation of brain tumor images using support vector machine classification in combination with hierarchical conditional random field regularization. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 354–361.
Bekes, G., Máté, E., Nyúl, L.G., Kuba, A., Fidrich, M., 2008. Geometrical model-based segmentation of the organs of sight on ct images. Med. Phys. 35 (2), 735–743.
Bondiau, P.-Y., Malandain, G., Chanalet, S., Marcy, P.-Y., Habrand, J.-L., Fauchon, F., Paquis, P., Courdi, A., Commowick, O., Rutten, I., et al., 2005. Atlas-based automatic segmentation of mr images: validation study on the brainstem in radiotherapy context. Int. J. Radiat. Oncol. Biol. Phys. 61 (1), 289–298.
Brosch, T., Tang, L.Y., Yoo, Y., Li, D.K., Traboulsee, A., Tam, R., 2016. Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. IEEE Trans. Med. Imaging 35 (5), 1229–1239.
Caviness, V.S., Filipek, P.A., Kennedy, D.N., 1989. Magnetic resonance technology in human brain science: blueprint for a program based upon morphometry. Brain Devel. 11 (1), 1–13.
Caviness, V.S., Meyer, J., Makris, N., Kennedy, D.N., 1996. Mri-based topographic parcellation of human neocortex: an anatomically specified method with estimate of reliability. J. Cogn. Neurosci. 8 (6), 566–587.
Cho, K., Raiko, T., Ilin, A., 2013. Enhanced gradient for training restricted Boltzmann machines. Neural Comput. 25 (3), 805–831.
Conson, M., Cella, L., Pacelli, R., Comerci, M., Liuzzi, R., Salvatore, M., Quarantelli, M., 2014. Automated delineation of brain structures in patients undergoing radiotherapy for primary brain tumors: from atlas to dose–volume histograms. Radioth. Oncol. 112 (3), 326–331.
Corso, J.J., Sharon, E., Dube, S., El-Saden, S., Sinha, U., Yuille, A., 2008. Efficient multilevel brain tumor segmentation with integrated Bayesian model classification. IEEE Trans. Med. Imaging 27 (5), 629–640.
Cuadra, M.B., Pollo, C., Bardera, A., Cuisenaire, O., Villemure, J.-G., Thiran, J.-P., 2004. Atlas-based segmentation of pathological mr brain images using a model of lesion growth. IEEE Trans. Med. Imaging 23 (10), 1301–1314.
Davis, M.E., 2016. Glioblastoma: overview of disease and treatment. Clin. J. Oncol. Nurs. 20 (5), S2.
Dawant, B.M., Hartmann, S., Gadamsetty, S., 1999. Brain atlas deformation in the presence of large space-occupying tumors. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 589–596.
Deeley, M., Chen, A., Datteri, R., Noble, J., Cmelak, A., Donnelly, E., Malcolm, A., Moretti, L., Jaboin, J., Niermann, K., et al., 2011. Comparison of manual and automatic segmentation methods for brain structures in the presence of space-occupying lesions: a multi-expert study. Phys. Med. Biol. 56 (14), 4557.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the em algorithm. J. R. Stat. Soc. Ser. B (Methodol.) 1–38.
Dolz, J., Leroy, H.-A., Reyns, N., Massoptier, L., Vermandel, M., 2015a. A fast and fully automated approach to segment optic nerves on mri and its application to radiosurgery. In: Proceedings of the IEEE 12th International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 1102–1105.
Dolz, J., Massoptier, L., Vermandel, M., 2015b. Segmentation algorithms of subcortical brain structures on MRI for radiotherapy and radiosurgery: a survey. IRBM 36 (4), 200–212.
Duane, S., Kennedy, A., Pendleton, B., Roweth, D., 1987. Hybrid monte carlo. Phys. Lett. B 195 (2), 216–222.
Fischer, A., Igel, C., 2014. Training restricted boltzmann machines: an introduction. Pattern Recognit. 47 (1), 25–39.
Fischl, B., 2012. Freesurfer. Neuroimage 62 (2), 774–781.
Freund, Y., Haussler, D., 1992. Unsupervised learning of distributions on binary vectors using two layer networks. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 912–919.
Geremia, E., Clatz, O., Menze, B.H., Konukoglu, E., Criminisi, A., Ayache, N., 2011. Spatial decision forests for ms lesion segmentation in multi-channel magnetic resonance images. Neuroimage 57 (2), 378–390.
Ghafoorian, M., Mehrtash, A., Kapur, T., Karssemeijer, N., Marchiori, E., Pesteie, M., Guttmann, C.R., de Leeuw, F.-E., Tempany, C.M., van Ginneken, B., et al., 2017. Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 516–524.
Gooya, A., Pohl, K.M., Bilello, M., Cirillo, L., Biros, G., Melhem, E.R., Davatzikos, C., 2012. Glistr: glioma image segmentation and registration. IEEE Trans. Med. Imaging 31 (10), 1941–1954.
Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., Larochelle, H., 2017. Brain tumor segmentation with deep neural networks. Med. Image Anal. 35, 18–31.
Havaei, M., Guizard, N., Chapados, N., Bengio, Y., 2016. Hemis: Hetero-modal image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 469–477.
Hinton, G.E., 2002. Training products of experts by minimizing contrastive divergence. Neural Comput. 14 (8), 1771–1800.
Iglesias, J.E., Sabuncu, M.R., Van Leemput, K., Initiative, A.D.N., et al., 2013. Improved inference in Bayesian segmentation using monte carlo sampling: application to hippocampal subfield volumetry. Med. Image Anal. 17 (7), 766–778.
Isambert, A., Dhermain, F., Bidault, F., Commowick, O., Bondiau, P.-Y., Malandain, G., Lefkopoulos, D., 2008. Evaluation of an atlas-based automatic segmentation software for the delineation of brain organs at risk in a radiation therapy clinical context. Radiother. Oncol. 87 (1), 93–99.
Islam, A., Reza, S.M., Iftekharuddin, K.M., 2013. Multifractal texture estimation for detection and segmentation of brain tumors. IEEE Trans. Biomed. Eng. 60 (11), 3204–3215.
Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. 36, 61–78.
Karimaghaloo, Z., Arnold, D.L., Arbel, T., 2016. Adaptive multi-level conditional random fields for detection and segmentation of small enhanced pathology in medical images. Med. Image Anal. 27, 17–30.
Kennedy, D.N., Filipek, P.A., Caviness, V.S., 1989. Anatomic segmentation and volumetric calculations in nuclear magnetic resonance imaging. IEEE Trans. Med. Imaging 8 (1), 1–7.
Kieselmann, J.P., Kamerling, C.P., Burgos, N., Menten, M.J., Fuller, C.D., Nill, S., Cardoso, M.J., Oelfke, U., 2018. Geometric and dosimetric evaluations of atlas-based segmentation methods of mr images in the head and neck region. Phys. Med. Biol. 63 (14), 145007.
Kingma, D. P., Welling, M., Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114, 2013.
Kistler, M., Bonaretti, S., Pfahrer, M., Niklaus, R., Büchler, P., 2013. The virtual skeleton database: an open access repository for biomedical research and collaboration. J. Med. Internet Res. 15 (11).
Kotecha, J.H., Djuric, P.M., 1999. Gibbs sampling approach for generation of truncated multivariate gaussian random variables. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 3. IEEE, pp. 1757–1760.
Kwon, D., Shinohara, R.T., Akbari, H., Davatzikos, C., 2014. Combining generative models for multifocal glioma segmentation and registration. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 763–770.
Lagendijk, J.J., Raaymakers, B.W., van Vulpen, M., 2014. The magnetic resonance imaging–linac system. Sem. Radiat. Oncol. 24, 207–209.
Lê, M., Delingette, H., Kalpathy-Cramer, J., Gerstner, E.R., Batchelor, T., Unkelbach, J., Ayache, N., 2017. Personalized radiotherapy planning based on a computational tumor growth model. IEEE Trans. Med. Imaging 36 (3), 815–825.
Lê, M., Unkelbach, J., Ayache, N., Delingette, H., 2016. Sampling image segmentations for uncertainty quantification. Med. Image Anal. 34, 42–51.
Lee, H., Grosse, R., Ranganath, R., Ng, A.Y., 2011. Unsupervised learning of hierarchical representations with convolutional deep belief networks. Commun. ACM 54 (10), 95–103.
Liu, D.C., Nocedal, J., 1989. On the limited memory BFGS method for large scale optimization. Math. Program 45 (1–3), 503–528.
Mabray, M.C., Barajas, R.F., Cha, S., 2015. Modern brain tumor imaging. Brain Tumor. Res. Treat. 3 (1), 8–23.
Maier, O., Wilms, M., Handels, H., 2016. Image features for brain lesion segmentation using random forests. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: First International Workshop, Brainles Held in Conjunction with MICCAI Munich, Germany Revised Selected Papers. Springer, pp. 119–130.
Melchior, J., Fischer, A., Wang, N., Wiskott, L., How to center binary restricted Boltzmann machines. arXiv preprint arXiv:1311.1354, 2013.
Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al., 2015. The multimodal brain tumor image segmentation benchmark (brats). IEEE Trans. Med. Imaging 34 (10), 1993–2024.
Menze, B.H., Van Leemput, K., Lashkari, D., Weber, M.-A., Ayache, N., Golland, P., 2010. A generative model for brain tumor segmentation in multi-modal images. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 151–159.
Moon, N., Bullitt, E., Van Leemput, K., Gerig, G., 2002. Automatic brain and tumor segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 372–379.
Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. MIT Press.
Munck af Rosenschöld, P., Costa, J., Engelholm, S.A., Lundemann, M.J., Law, I., Ohlhues, L., Engelholm, S., 2014. Impact of [18f]-fluoro-ethyl-tyrosine pet imaging on target definition for radiation therapy of high-grade glioma. Neuro-Oncol. 17 (5), 757–763.
Munck af Rosenschöld, P., Engelholm, S., Ohlhues, L., Law, I., Vogelius, I., Engelholm, S.A., 2011. Photon and proton therapy planning comparison for malignant glioma based on ct, fdg-pet, dti-mri and fiber tracking. Acta Oncol. (Madr.) 50 (6), 777–783.
Noble, J.H., Dawant, B.M., 2011. An atlas-navigated optimal medial axis and deformable model algorithm (nomad) for the segmentation of the optic nerves and chiasm in mr and ct images. Med. Image Anal. 15 (6), 877–884.
Pereira, S., Pinto, A., Alves, V., Silva, C.A., 2016. Brain tumor segmentation using convolutional neural networks in MRI images. IEEE Trans. Med. Imaging 35 (5), 1240–1251.
Prastawa, M., Bullitt, E., Moon, N., Van Leemput, K., Gerig, G., 2003. Automatic brain tumor segmentation by subject specific modification of atlas priors1. Acad. Radiol. 10 (12), 1341–1348.
Preusser, M., de Ribaupierre, S., Wöhrer, A., Erridge, S.C., Hegi, M., Weller, M., Stupp, R., 2011. Current concepts and management of glioblastoma. Ann. Neurol. 70 (1), 9–21.

Puonti, O., Iglesias, J.E., Van Leemput, K., 2016. Fast and sequence-adaptive whole-brain segmentation using parametric Bayesian modeling. NeuroImage 143, 235–249.

Rajchl, M., Pawlowski, N., Rueckert, D., Matthews, P. M., Glocker, B., 2018. Neuronet: fast and robust reproduction of multiple brain image segmentation pipelines. arXiv:1806.04224.

Rodriguez-Yam, G., Davis, R. A., Scharf, L. L., Efficient Gibbs sampling of truncated multivariate normal with application to constrained linear regression. *Unpublished Manuscript*, 2004.

Roy, A.G., Conjeti, S., Sheet, D., Katouzian, A., Navab, N., Wachinger, C., 2017. Error corrective boosting for learning fully convolutional networks with limited data. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 231–239.

Sauwen, N., Acou, M., Van Cauter, S., Sima, D., Veraart, J., Maes, F., Himmelreich, U., Achten, E., Van Huffel, S., 2016. Comparison of unsupervised classification methods for brain tumor segmentation using multi-parametric MRI. NeuroImage: Clinical 12, 753–764.

Shaffer, R., Nichol, A.M., Vollans, E., Fong, M., Nakano, S., Moiseenko, V., Schmuland, M., Ma, R., McKenzie, M., Otto, K., 2010. A comparison of volumetric modulated arc therapy and conventional intensity-modulated radiotherapy for frontal and temporal high-grade gliomas. Int. J. Radiat. Oncol. Biol. Phys. 76 (4), 1177–1184.

Smolensky, P., 1986. Information processing in dynamical systems: Foundations of harmony theory. In: Rumelhart, D.E., McClelland, J.L. (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition, 1. MIT Press, pp. 194–281.

Tustison, N.J., Shrinidhi, K., Wintermark, M., Durst, C.R., Kandel, B.M., Gee, J.C., Grossman, M.C., Avants, B.B., 2015. Optimal symmetric multimodal templates and concatenated random forests for supervised brain tumor segmentation (simplified) with ANTSR. Neuroinformatics 13 (2), 209–225.

Unkelbach, J., Menze, B.H., Konukoglu, E., Dittmann, F., Le, M., Ayache, N., Shih, H.A., 2014. Radiotherapy planning for glioblastoma based on a tumor growth model: improving target volume delineation. Phys. Med. Biol. 59 (3), 747.

Valindria, V. V., Lavdas, I., Bai, W., Kamnitsas, K., Aboagye, E. O., Rockall, A. G., Rueckert, D., Glocker, B., Domain adaptation for MRI organ segmentation using reverse classification accuracy. arXiv preprint arXiv:1806.00363, 2018.

Van Dyk, D.A., Park, T., 2008. Partially collapsed Gibbs samplers: theory and methods. J. Am. Stat. Assoc. 103 (482), 790–796.

Van Leemput, K., 2009. Encoding probabilistic brain atlases using Bayesian inference. IEEE Trans. Med. Imaging 28 (6), 822–837.

Van Leemput, K., Maes, F., Bello, F., Vandermeulen, D., Colchester, A., Suetens, P., 1999. Automated segmentation of ms lesions from multi-channel mr images. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 11–21.

Van Leemput, K., Maes, F., Vandermeulen, D., Colchester, A., Suetens, P., 2001. Automated segmentation of multiple sclerosis lesions by model outlier detection. IEEE Trans. Med. Imaging 20 (8), 677–688.

Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 1999. Automated model-based bias field correction of MR images of the brain. IEEE Trans. Med. Imaging 18 (10), 885–896.

Wells, W.M., Grimson, W.E.L., Kikinis, R., Jolesz, F.A., 1996. Adaptive segmentation of MRI data. IEEE Trans. Med. Imaging 15 (4), 429–442.

Wiggermann, V., Hernandez-Torres, E., Traboulsee, A., Li, D., Rauscher, A., 2016. Flair2: a combination of flair and t2 for improved ms lesion detection. Am. J. Neuroradiol. 37 (2), 259–265.

Zikic, D., Glocker, B., Konukoglu, E., Criminisi, A., Demiralp, C., Shotton, J., Thomas, O., Das, T., Jena, R., Price, S., 2012. Decision forests for tissue-specific segmentation of high-grade gliomas in multi-channel mr. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 369–376.