Unsupervised quality control of segmentations based on a smoothness and intensity probabilistic model

Benoît Audelan, Hervé Delingette

Unsupervised quality control of segmentations based on a smoothness and intensity probabilistic model

Benoît Audelan[a,*], Hervé Delingette[a]

[a] *Université Côte d'Azur, Inria, Epione project-team, Sophia Antipolis, France*

Abstract

Monitoring the quality of image segmentation is key to many clinical applications. This quality assessment can be carried out by a human expert when the number of cases is limited. However, it becomes onerous when dealing with large image databases, so partial automation of this process is preferable. Previous works have proposed both supervised and unsupervised methods for the automated control of image segmentations. The former assume the availability of a subset of trusted segmented images on which supervised learning is performed, while the latter does not. In this paper, we introduce a novel unsupervised approach for quality assessment of segmented images based on a generic probabilistic model. Quality estimates are produced by comparing each segmentation with the output of a probabilistic segmentation model that relies on intensity and smoothness assumptions. Ranking cases with respect to these two assumptions allows the most challenging cases in a dataset to be detected. Furthermore, unlike prior work, our approach enables possible segmentation errors to be localized within an image. The proposed generic probabilistic segmentation method combines intensity mixture distributions with spatial regularization prior models whose parameters are estimated with variational Bayesian techniques. We introduce a novel smoothness prior based on the penalization of the derivatives of label maps which allows an automatic estimation of its hyperparameter in a fully data-driven way. Extensive evaluation of quality control on medi-

*Corresponding author
Email address: benoit.audelan@inria.fr (Benoît Audelan)

cal and COCO datasets is conducted, showing the ability to isolate atypical segmentations automatically and to predict, in some cases, the performance of segmentation algorithms.

Keywords: Unsupervised quality control, image segmentation, Bayesian learning, spatial regularization

---

1. Introduction

Semantic segmentation of an image is the process of associating a label to every pixel in an image. This task is particularly important in a medical context since it impacts downstream algorithms using image segmentations as input, but also the decisions that clinicians may make about the patient. For instance, in radiotherapy planning, the delineations of tumor lesions directly influence the extent of the dose delivered around the tumor. Also, obtaining reliable image segmentations is mandatory to use image derived biomarkers in a clinical setting (Keshavan et al., 2018). Finally, the development of supervised learning for image segmentation requires the accumulation of potentially large sets of manually or semi-manually segmented image databases that need to be quality controlled. Such segmentations are prone to inter-rater variability (Visser et al., 2019) in addition to plain errors. It is therefore of great importance to automatically detect possible failed segmentation cases, whether those segmentations are generated by an algorithm or a human rater. The challenge is to perform this monitoring in the absence of ground truth segmentations.

In prior work, evaluation methods can be categorized either as supervised or as unsupervised, depending on whether a reference segmentation is required or not (Zhang et al., 2008). A first set of supervised methods is based on a classifier which accepts or rejects the proposed segmentation based on combined features. For instance, in Hui Zhang et al. (2006), decision trees based on handcrafted features depending on the image (texture, color space...) and on the geometry of the segmented region (perimeter, compactness...) are combined in a single classifier. In Shamir and Bomzon (2019), a decision tree predicts the Dice score

of head segmentations with an application to the treatment of brain tumors. In Xu et al. (2009), a framework to detect failures in cardiac segmentation based on a shape parameter and an intensity feature has been proposed. The number of features taken into account is increased in Kohlberger et al. (2012), where the model decision relies on 42 shape and appearance features. They are combined in an SVM classifier regressing the Dice coefficient between the given segmentation and the unknown ground truth. While in Xu et al. (2009) the features were specific to cardiac segmentation, the approach taken in Kohlberger et al. (2012) is more generic and was trained on segmentations of 8 different organs.

Reverse Classification Accuracy (RCA) has also been proposed for quality control assessment in Valindria et al. (2017). Assuming the availability of a set of trusted images with ground truth, the proposed segmentation on a new image is compared to the predicted one based on those reference images, which can result in rejection if discrepancies are too large. This approach was tested on larger databases in Robinson et al. (2019) where the authors showed the ability of the method to highlight poor quality segmentations but pointed out the relatively long computation times as a bottleneck.

Another family of supervised approaches uses deep learning to estimate the quality of a segmentation. For instance, in Robinson et al. (2018), a neural network is trained to predict the Dice coefficient of cardiac segmentations. The Jaccard index (intersection over union) is predicted by neural networks in (Arbelle et al., 2019; Huang et al., 2016; Shi et al., 2017) where the original image and the proposed segmentation mask are provided as input. Some authors have proposed exploiting the uncertainty of segmentations in order to assess their quality, within a deep learning framework. Uncertainty quantification also adds some interpretability to the quality assessment as it provides information about the location of possible errors. Bayesian QuickNat proposed by Roy et al. (2019) uses Monte Carlo dropout at test time to generate several segmentation samples. The average over the samples gives the final segmentation map while variability across the different samples gives an estimate of the uncertainty of the segmentation. The authors show a good correlation between the measured uncertainty

and the Dice coefficient between the segmentation and the unknown ground truth. Other methods to evaluate the uncertainty were explored in Jungo and Reyes (2019) and the results suggest that none is superior to the others. Finally in DeVries and Taylor (2018), a first network outputs a segmentation map and an uncertainty map at the pixel level, which are then taken as inputs by a second network which regresses a quality score at the image level.

A limitation shared by these methods is their supervised design, meaning that they require the extraction of a subset of segmented data that is considered to be "ground truth". This trusted subset is used by the models to learn how a "good" segmentation looks. The resulting decision rules making a new segmentation acceptable or not may thus be biased by the composition of the trusted set, which must be large enough for training a deep-learning-based framework. Further, access to large annotated datasets remains an issue in many domains including medical imaging. Finally, supervised methods often lack generality as their performance depends on the type of images and segmented structures in the training set.

In contrast, unsupervised approaches do not rely on a subset of trusted images but rather on assumptions about the appearance and shape of the foreground and background regions (Rosenberger et al., 2006; Zhang et al., 2008). These assumptions are then translated into a set of segmentation metrics. For instance, common hypothesis is that a "good" segmentation exhibits high levels of intra-region homogeneity and inter-region heterogeneity (Johnson and Xie, 2011), and several handcrafted features have been proposed to measure them (Chabrier et al., 2006; Gao et al., 2017; Johnson and Xie, 2011; Zhang et al., 2008). The main limitation of these approaches is that it is difficult to design discriminative indices and to find a proper way to combine them. Moreover, as mentioned by Zhang et al. (2008), most of those metrics assume a single underlying intensity distribution, typically Gaussian, in both foreground and background regions which is overly simplistic and sensitive to outliers.

Last but not least, interpretability is a desirable property, as knowing the problematic regions could facilitate the segmentation curation. However, it is

often an issue since many of the previous methods, supervised or not, are black boxes outputting a simple score, which does not help to understand why a segmentation has failed.

In this paper, we propose a novel unsupervised approach for automated quality assessment of image segmentations. It is based on the comparison between a proposed segmentation S produced by an algorithm or a human rater and the segmentation M given by a generic probabilistic segmentation model. The generic model is based on two simple intensity and smoothness assumptions, the underlying hypothesis being that explainable segmentations correspond to clearly visible boundaries in the image well captured by M. On the contrary, segmentations far from M are categorized as difficult or challenging as they would require priors other than intensity and smoothness to be explained. The quality assessment of a set of segmented images is then performed by studying how the distance between the proposed segmentation S and the modelled segmentation M varies within the dataset. Segmentations that are lying on the tails of this distance distribution are considered to be atypical and are candidates for manual verification. We show the effectiveness of this approach to extract suspicious segmentations on various public datasets ranging from photographic images for object detection and segmentation (COCO dataset) to lung and brain medical images (LIDC and BRATS datasets). We also show that this approach can be used in some cases to predict the performance of segmentation algorithms.

Our main contributions are twofold:

- Instead of relying on an arbitrary subset of selected segmentations as a training set, we propose an unsupervised approach based on intensity and smoothness hypotheses without any prior knowledge of the structure to be segmented. It removes the bias related to the selection of the reference images and allows the quality of segmentations to be assessed when few or even no other segmentations are available from a database. Our method differs from previous unsupervised segmentation quality indices

with a more complex and robust approach to modeling the intensity of the different regions in the image. In addition, it allows a combination of the key factors defining a "good" segmentation (i.e., the intra-region homogeneity and the inter-region heterogeneity) in a data-driven way. Last but not least, our method is visually interpretable. For instance, when dealing with 3D medical images, it allows automatic retrieval of the slices with suspicious segmentations. Finally, the result can be useful to guide the manual correction of poorly segmented cases.

- We provide different spatial regularization strategies to enforce the spatial continuity. In particular, we introduce a novel prior, denoted by FDSP (Finite Difference Spatial Prior), based on the penalization of the squared norm of the derivatives of the prior label map, which allows an adaptative learning of the hyperparameter. It is compared to the classical Markov random field (MRF) and another spatial prior based on a weighted combination of spatially smooth kernels introduced in an earlier work of the authors (Audelan and Delingette, 2019), which will be denoted by GLSP (Generalized Linear Spatial Prior) throughout the paper.

This paper expands Audelan and Delingette (2019) by proposing a different spatial regularization strategy for which the hyperparameter can be estimated. In addition, the novel regularization prior can be entirely inferred with a variational Bayes method (no Laplace approximation needed) and leads to much faster computations. We also provide more extensive experiments on different datasets and added a qualitative and quantitative comparison with unsupervised segmentation quality control indices proposed in prior works. The code with the different regularization strategies is available in this repository: `https://gitlab.inria.fr/epione/unsegqc`.

The rest of the paper is organized as follows. Section 2 presents the general framework of our unsupervised quality control assessment. In section 3, we present our appearance model and the spatial priors. Section 4 describes the model inference depending on the regularization. Finally, we show in section

5 the relevance of our approach for segmentation quality control on several datasets.

2. Unsupervised quality control workflow

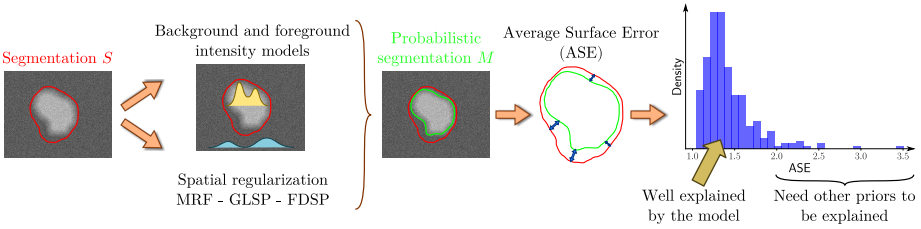


Figure 1: Unsupervised segmentation quality control workflow.

Supervised segmentation quality control methods require the existence of a trusted subset of data from which quality assessment is learned. Instead, we follow an unsupervised approach (see Fig. 1) based on a probabilistic segmentation model relying only on two simple smoothness and intensity assumptions. Its great advantage is that it is agnostic with respect to the structure to be segmented and therefore can be run automatically even in the absence of ground truth.

2.1. Input segmentation

The input of the proposed method is a binary segmentation S on an image I into foreground and background regions for which we would like a quality estimate. There are no restrictions regarding the origin of S as it can have been created by an algorithm or a human rater. Note that this is in contrast to several other methods that require the input segmentation to have been generated by a specific algorithm, like the uncertainty-based methods in deep learning (DeVries and Taylor, 2018; Jungo and Reyes, 2019; Roy et al., 2019).

2.2. Probabilistic model

Given the segmentation S, we produce a smooth contour or surface M close to S which is mostly aligned with visible contours in the image. We stress that

the objective is not to build a surrogate ground truth, but instead to use M only as a comparison tool.

Intensity assumption. The first hypothesis of our approach is that intensity distribution variations in the image can help to understand segmentations. Given the segmentation S, two intensity models are built for the foreground and background regions.

Spatial smoothness assumption. The second hypothesis relies on the generally accepted assumption that two neighbouring voxels share a higher probability of belonging to the same label region. This is classically enforced by the use of discrete priors such as MRF. In Audelan and Delingette (2019), we proposed a regularization strategy based on a combination of spatially smooth kernels (GLSP). In addition to these two possibilities, we introduce in this paper a novel way to take into account the spatial organization of the voxels, which we call the Finite Difference Spatial Prior (FDSP). This approach allows full tractability of the hyperparameter in an efficient manner which is not possible for the MRF and GLSP formulations.

The two assumptions are combined into a probabilistic model that outputs a new segmentation map M. By construction, M is typically a smooth contour which is mostly aligned with the visible intensity boundaries in the image. Again, M should only be seen as a representation used to benchmark the input segmentation S.

To measure the adequacy of S with respect to M, we employ the average asymmetric surface error (ASE) defined as $E_S = d(S, M) = \frac{1}{\partial S} \sum_{x \in \partial S} \min_{y \in \partial M} d(x, y)$ where ∂ denotes the segmentation surface. We discard the metric $d(M, S)$ as being uninformative since M is not a surrogate ground truth. An alternative measure to the ASE used in this paper is the Dice score computed between the segmentations M and S.

2.3. Detection of challenging cases

Segmentations S close to M are identified as being explained by the model. In that case, the two intensity and spatial smoothness assumptions upon which the probabilistic model is based are sufficient to understand the contours. However, segmentations S far from M are classified as unexplained or challenging. Typically, contours crossing large regions of uniform intensity distribution would be identified as unexplained by our model. It is important to note that having an unexplained segmentation does not imply that this segmentation is wrong. It simply means that other priors besides those of smoothness and intensity are required to understand its boundaries.

2.4. Use case

We believe our approach is particularly interesting when dealing with a whole set of segmentations. For instance, say we are given a set of images with corresponding annotations. The comparison of adequacies between S and M for all images allows the detection of atypical cases which behave differently from the majority of the distribution, and for which a visual inspection might be worthy. On the contrary, applying the method on a single image is not the ideal use case as the analysis of the result is difficult without any comparison with similar images.

Our approach is unsupervised, generic, and based on few simple assumptions. However this comes with intrinsic limitations. For instance, any irrelevant contour following visible intensity boundaries will not be considered as suspicious. This limitation is common to all previously proposed unsupervised methods. More generally, the proposed method is not intended to return all erroneous segmentations inside a dataset (which is expected from a supervised approach) but instead to extract some suspicious cases when limited information is available.

3. Method

In this section, we review the details of our probabilistic model. We consider a binary image segmentation problem for isolating a single structure from an image I made of N voxels in a grid of dimension D ($D = 2, 3$) having intensity $I_n \in \mathbb{R}^v$, $n = 1, \ldots, N$, where $v \geq 1$ ($v = 1, 3$ and 4 in practice). We introduce for each voxel a binary hidden random variable $Z_n \in \{0, 1\}$ with $Z_n = 1$ if voxel n belongs to the structure of interest.

3.1. Mixtures of multivariate Student's t-distributions

Appearance models of the foreground and background regions of S are defined respectively by the two image likelihoods $p(I_n|Z_n = 1, \theta_I^1)$ and $p(I_n|Z_n = 0, \theta_I^0)$ where θ_I^0, θ_I^1 are parameters governing those models. In this paper, we consider generic parametric appearance models as variational mixtures of multivariate Student's t-distributions (Archambeau and Verleysen, 2007). The Student's t generalizes the Gaussian distribution with heavy tails and leads to robust mean and covariance estimates. The number of components in the mixture is automatically estimated by using a sparsity-inducing Dirichlet prior over the mixture proportions which automatically prunes the components with a small number of samples. Finally, we introduce the appearance probability ratio r_n defined as:

$$r_n(I, \theta_I^0, \theta_I^1) \triangleq \frac{p(I_n|Z_n = 1, \theta_I^1)}{p(I_n|Z_n = 0, \theta_I^0) + p(I_n|Z_n = 1, \theta_I^1)}\,, \tag{1}$$

which is the posterior label probability with a non-informative prior ($p(Z_n = 1) = 0.5$).

3.2. Spatial smoothness prior

The spatial smoothness prior allows the spatial organization between voxels to be taken into account and a certain degree of continuity to be enforced. To this end, different strategies can be employed. In this paper, we propose to compare one discrete prior (MRF) with two continuous priors (GLSP and FDSP), the third one being novel to the best of our knowledge.

3.2.1. MRF prior

The classical MRF formulation relies on labels of neighbouring voxels. In a binary segmentation problem, a natural way to enforce spatial smoothness is the Ising model. Assuming β to be the hyperparameter of the MRF, the label prior probability is given by:

$$p(Z|\beta) = \frac{1}{T(\beta)} \exp\left\{ \frac{\beta}{2} \sum_{i=1}^{N} \sum_{j \in \delta_i} Z_i Z_j \right\}, \tag{2}$$

where δ_i are the neighbouring voxels of i and $T(\beta)$ is the partition function. In practice, we consider 4- and 6-connectivity neighborhoods for 2D and 3D images, respectively. The value of β represents the strength of association between neighbouring voxels: $\beta = 0$ corresponds to a model with no spatial prior, while large positive values encourage neighbouring voxels to have the same label. The Ising model may be replaced by an image contrast sensitive prior as performed for instance in the GrabCut algorithm (Rother et al., 2004).

The computation of the partition function $T(\beta)$, needed for an automatic estimation of the model's hyperparameter β, requires considering all possible configurations of the MRF which is not computationally tractable for large lattices. Therefore, β has to be fixed by the user.

3.2.2. Generalized Linear Spatial Prior

In Audelan and Delingette (2019), we proposed a continuous label prior denoted by Generalized Linear Spatial Prior (GLSP) to enforce the spatial continuity. The prior is defined through a generalized linear model of spatially smooth functions. More precisely, the prior probability $p(Z_n = 1)$ is defined as a Bernoulli distribution whose parameter is a *spatially random* function specified as a generalized linear model:

$$p(Z_n = 1|W) = \sigma\left(\sum_{l=1}^{L} \Phi_l(\mathbf{x}_n) w_l \right), \tag{3}$$

where $\mathbf{x}_n \in \mathbb{R}^D$ is the voxel position in an image of dimension D and the link function $\sigma(u)$ is the sigmoid function $\sigma(f) = 1/(1+\exp(-f))$. The basis $\{\Phi_l(\mathbf{x})\}$ are L functions of space, typically radial basis functions (for instance, Gaussian functions) defined on a regular grid, and $w_l \in W$ are weights considered as random variables. Thus the prior probabilities of two geometrically close voxels are related to each other through the smoothness of the function $f(\mathbf{x}_n) = \sum_{l=1}^{L} \Phi_l(\mathbf{x}_n) w_l = \Phi_n^T W$, writing $\Phi_n^T = [\Phi_1(x_n), \cdots, \Phi_L(x_n)]$.

The smoothness of the label prior $\sigma\left(f(\mathbf{x}_n)\right)$ depends on the choice of the L basis functions $\{\Phi_l(\mathbf{x})\}$ which are commonly uniformly spread over the image domain. The key parameters are the spacing between the basis centers, the standard deviations (or radii) r of the Gaussian functions and the position of the origin basis. Together, they influence the amount of smoothing brought by the label prior, large spacing and standard deviations leading to smoother prior probability maps.

To obtain a robust description, the weight vector $W = [w_1, \ldots, w_L]^T$ is fitted with a zero mean Gaussian prior parameterized by the diagonal precision matrix $\alpha \mathbf{I}_L$: $p(W) = \mathcal{N}(0, \alpha^{-1}\mathbf{I}_L)$. Finally, a non-informative prior is chosen for α, $p(\alpha) \propto 1$. In contrast to the MRF formulation, a Bayesian inference of the hyperparameter is possible here, as shown in section 4.2.

3.2.3. Finite Difference Spatial Prior

As a third regularization strategy, we introduce in this paper the Finite Difference Spatial Prior (FDSP). The prior probability $p(Z_n = 1)$ is again defined as a Bernoulli distribution whose parameter belongs to a spatially smooth random field:

$$p(Z_n = 1|W) = \sigma\left(w_n\right) , \tag{4}$$

where $\sigma(u)$ is once more the sigmoid function. The smoothness of the label field is caused by a prior applied to the vector $W = [w_1, \ldots, w_n]^T$ penalizing the

squared norm of its derivatives of order p:

$$p(W|\alpha) = \frac{1}{T(\alpha)} \exp\left(-\alpha \sum_{n=1}^{N} \|\Delta_p(w_n)\|^2\right), \tag{5}$$

where $\Delta_p(w_n)$ is the p order central finite difference operator at w_n and $T(\alpha)$ is the normalization factor. The quantity $\Delta_p(w_n)$ is a tensor of order p approximating the p-order derivatives of the scalar field defined by w_n. Since the function $h(x) = \|\Delta_p(x)\|^2$ is 2-homogenous, we know that the normalization factor has the form $T(\alpha) = c\alpha^{-N/2}$ where c is constant independent of α (Pereyra et al., 2015). One can easily show that $p(W|\alpha)$ is a zero mean Gaussian distribution whose precision matrix consists of difference operators. The value of the parameter α controls the amount of the spatial regularization applied to the weights W.

In this paper, we consider only first order derivatives ($p = 1$) corresponding to the discretization of the Dirichlet energy. In that case Eq. 5 is written:

$$p(W|\alpha) = c\alpha^{\frac{N}{2}} \exp\left(-\frac{\alpha}{4} \sum_{n=1}^{N} \sum_{d=1}^{D} (w_{\delta_d(n+1)} - w_{\delta_d(n-1)})^2\right), \tag{6}$$

where $\delta_d(n+i)$ represents the neighbor of index i of voxel n in the dimension d.

The graphical models of the different segmentation frameworks are shown in Figure 2.

3.3. Implementation

The second step is to compute a variational approximation of the posterior $P(Z_n|I_n)$ which involves solving an inference problem depending on the choice of spatial prior. After convergence of the probabilistic model, a new segmentation M is generated by thresholding the posterior $p(Z_n|I_n)$ at the level 0.5.

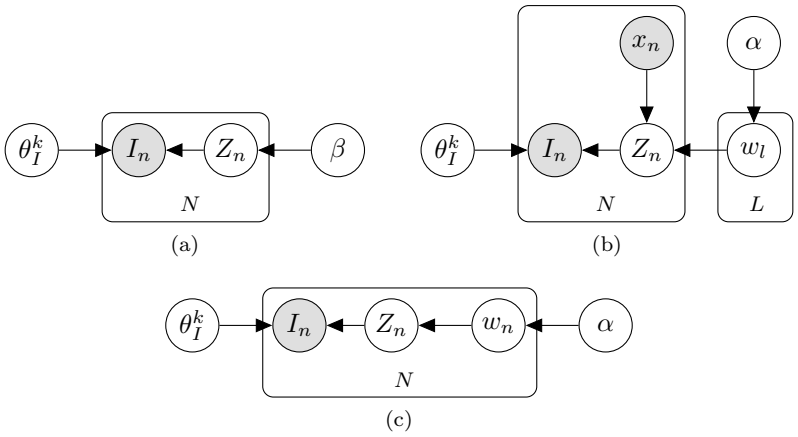


Figure 2: Graphical model of the framework with a discrete MRF prior (2a), a GLSP prior (2b) or a FDSP prior (2c).

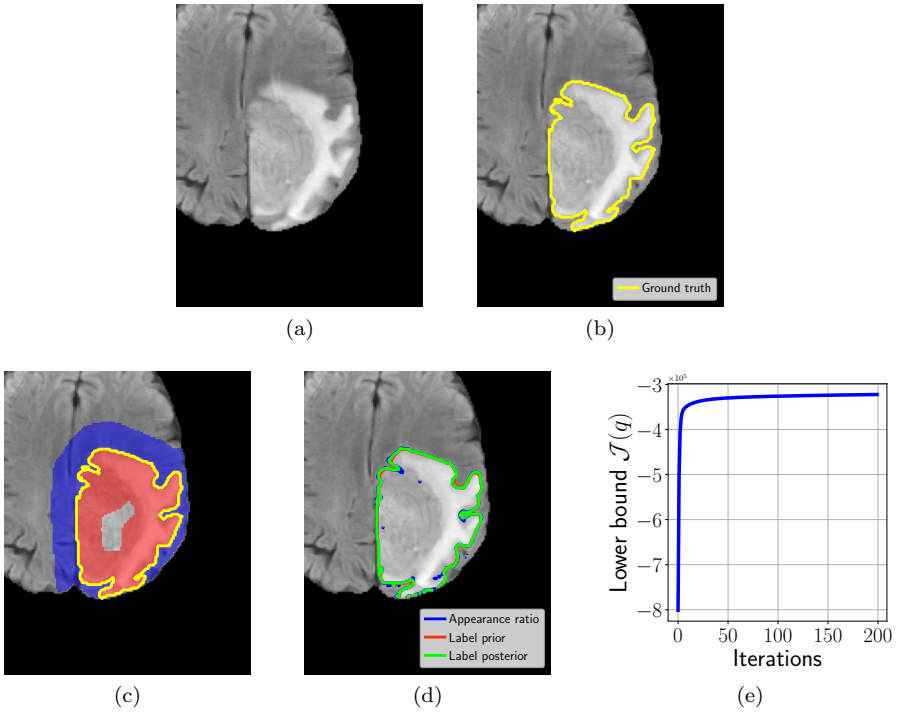


Figure 3: Quality control workflow on a glioblastoma segmentation from the BRATS 2017 dataset with FDSP regularization. (3a) Original image. (3b) Input segmentation S. (3c) Narrow band along the ground truth boundary with foreground (in red) and background (in blue) regions. (3d) Appearance ratio, label spatial prior and label posterior. (3e) Evolution of the lower bound $\mathcal{J}(q)$.

4. Probabilistic inference

4.1. MRF regularization

A classical way to maximize the log likelihood $\log p(I)$ with an MRF prior is to use variational inference with a mean field approximation (Ambroise and Govaert, 1998; Roche et al., 2011). The label posterior distribution $q(Z)$ is assumed to factorize as $\prod_i q_i(Z_i)$, which leads to the following fixed-point equation for voxel i at iteration $m+1$:

$$q_{ip}^{m+1} = \frac{r_i \exp\{\beta \sum_{j \in \delta_i} q_{jp}^m\}}{\sum_{k=0}^{1} r_i^k (1-r_i)^{1-k} \exp\{\beta \sum_{j \in \delta_i} q_{jk}^m\}}\,, \tag{7}$$

where $p \in \{0, 1\}$, q_{ik} represents $q_i(Z_i = k)$, r_i is the appearance probability ratio for voxel i and β is fixed by the user.

4.2. GLSP regularization

A type-II maximum likelihood approach is used to estimate the model parameters. A Gaussian approximation for the weights posterior distribution is found by computing a Laplace approximation through iterative reweighted least squares. The parameter α is then updated by maximizing the marginal likelihood. We refer to the original paper for more details (Audelan and Delingette, 2019).

The algorithm requires several steps of covariance matrix inversion which can be prohibitive for large images. The problem was addressed in Audelan and Delingette (2019) by splitting the narrow band into smaller overlapping patches that were then merged. In this paper, the code was improved and the decomposition into patches is no longer needed.

4.3. FDSP regularization

We propose a variational inference scheme to estimate prior and hyperprior parameters $U = \{Z, W, \alpha\}$. Variational inference approximates the true posterior $p(U|I)$ by a chosen family of distributions $q(U)$. Maximizing the data log

likelihood $\log p(I)$ implies minimizing the Kullbach-Leibler divergence between $q(U)$ and $p(U|I)$ or equivalently maximizing the lower bound $\mathcal{L}(q)$:

$$\log p(I) = \underbrace{\int_U q(U) \log \frac{p(I,U)}{q(U)} dU}_{\mathcal{L}(q)} + \mathrm{KL}\left[q(U)||p(U|I)\right] . \tag{8}$$

We assume that the approximation of the posterior can be factorized as $q(U) = q_Z(Z)q_W(W)q_\alpha(\alpha)$. The lower bound can thus be re-written as:

$$\begin{aligned} \log p(I) \geqslant \mathcal{L}(q) = & \sum_Z \int_\alpha \int_W q_Z(Z)q_W(W)q_\alpha(\alpha) \\ & \log \frac{p(I|Z)p(Z|W)p(W|\alpha)}{q_Z(Z)q_W(W)q_\alpha(\alpha)} dW d\alpha \, . \end{aligned} \tag{9}$$

We can further expand the factors defining the joint probability: $p(I|Z) = \prod_n r_n^{Z_n}(1-r_n)^{1-Z_n}$. The spatial prior $p(Z_n|W)$ can be likewise written as $[\sigma(w_n)]^{Z_n}[\sigma(-w_n)]^{1-Z_n}$ and the weights prior $p(W|\alpha)$ is given by (6) for first order derivatives.

However, the right hand side of (9) is intractable because the spatial prior does not belong to the exponential family (due to the sigmoid function). As an alternative to the Laplace approximation, we use a local variational bound as introduced in Jaakkola and Jordan (2000) in the context of logistic regression. In this case, we replace the sigmoid function with a well-chosen lower bound: $\sigma(x) \geqslant g(x,\xi) = \sigma(\xi)\exp\left[(x-\xi)/2 - \lambda(\xi)(x^2-\xi^2)\right]$. ξ is a variational parameter and $\lambda(\xi) = \tanh(\xi/2)/(4\xi)$. The spatial prior $p(Z|W)$ can thus be approximated by $F(Z,W,\xi) = \prod_n \left[g(w_n,\xi_n)\right]^{Z_n}\left[g(-w_n,\xi_n)\right]^{1-Z_n}$. This approximation leads to a new lower bound $\mathcal{J}(q)$ on the lower bound $\mathcal{L}(q)$:

$$\begin{aligned} \log p(I) \geqslant \mathcal{L}(q) \geqslant \mathcal{J}(q) = & \sum_Z \int_\alpha \int_W q_Z(Z)q_W(W)q_\alpha(\alpha) \\ & \log \frac{p(I|Z)F(Z,W,\xi)p(W|\alpha)}{q_Z(Z)q_W(W)q_\alpha(\alpha)} dW d\alpha \, . \end{aligned} \tag{10}$$

This new lower bound $\mathcal{J}(q)$ is now tractable and the optima q^* for each of the

variational posteriors can be derived by variational calculus (See Appendix B for details of the derivations). $q_Z^*(Z)$ is therefore given by $q_Z^*(Z) = \prod_n \eta_{n1}^{Z_n} \eta_{n0}^{1-Z_n}$ with $\eta_{nk} = \rho_{nk} / \sum_k \rho_{nk}$ for $k \in \{0, 1\}$ and:

$$\begin{aligned} \rho_{nk} = r_n^k (1 - r_n)^{1-k} \sigma(\xi_n) \exp \Bigg[(-1)^{1-k} \frac{\mathbb{E}[w_n]}{2} \\ - \frac{\xi_n}{2} - \lambda(\xi_n)(\mathbb{E}[w_n^2] - \xi_n^2) \Bigg] . \end{aligned} \tag{11}$$

By further assuming that $q_W(W) = \prod_n q_{w_n}(w_n)$, the variational optimization for $q_{w_n}(w_n)$ yields a normal distribution of the form $q_{w_n}^*(w_n) = \mathcal{N}(\boldsymbol{\mu}_{w_n}, \boldsymbol{\Sigma}_{w_n})$. A fixed-point equation is found for updating the mean. For first order derivatives, we have:

$$\boldsymbol{\Sigma}_{w_n} = \left[2\lambda(\xi_n) + 2 \sum_d \frac{\alpha}{2} \right]^{-1} , \tag{12}$$

$$\boldsymbol{\mu}_{w_n} = \boldsymbol{\Sigma}_{w_n} \left[\eta_{n1} - \frac{1}{2} + \frac{\alpha}{2} \sum_d \left(\boldsymbol{\mu}_{w_{\delta_d(n+2)}} + \boldsymbol{\mu}_{w_{\delta_d(n-2)}} \right) \right] . \tag{13}$$

The variational posterior $q_\alpha(\alpha)$ is assumed to be a Dirac distribution which leads to the following update:

$$\alpha^{-1} = \frac{1}{2N} \sum_n \sum_{d=1}^{D} \mathbb{E} \left[(w_{\delta_d(n+1)} - w_{\delta_d(n-1)})^2 \right] . \tag{14}$$

Finally, following Bishop (2006), maximizing (10) with respect to ξ_n gives an update formula of the form:

$$\xi_n^2 = \mathbb{E}[w_n^2] . \tag{15}$$

To compute (11), (14) and (15), we need the expectations $\mathbb{E}[w_n]$, $\mathbb{E}\left[(w_{\delta_d(n+1)} - w_{\delta_d(n-1)})^2\right]$ and $\mathbb{E}[w_n^2]$ with respect to the variational distribution q_{w_n}. They can be easily evaluated to give $\mathbb{E}[w_n] = \boldsymbol{\mu}_{w_n}$, $\mathbb{E}[w_n^2] = \boldsymbol{\Sigma}_{w_n} + \boldsymbol{\mu}_{w_n}^2$ and $\mathbb{E}\left[(w_{\delta_d(n+1)} - w_{\delta_d(n-1)})^2\right] = \boldsymbol{\mu}_{w_{\delta_d(n+1)}}^2 + \boldsymbol{\mu}_{w_{\delta_d(n-1)}}^2 - 2\boldsymbol{\mu}_{w_{\delta_d(n+1)}} \boldsymbol{\mu}_{w_{\delta_d(n-1)}} + \boldsymbol{\Sigma}_{w_{\delta_d(n+1)}} + \boldsymbol{\Sigma}_{w_{\delta_d(n-1)}}$.

After convergence, the variational distribution $q_Z(Z)$ gives an approxima-

tion to the posterior label probability $p(Z_n = 1|I, W)$, which combines prior and intensity likelihoods. Finally, the maximum a posteriori estimate of the segmented structure is obtained as the isosurface $p(Z_n = 1|I, W) = 0.5$.

This approach has some advantages in comparison with the first two. First, it allows an automatic estimation of all its parameters. For the MRF, the user needs to fix β and for the GLSP, the layout of the basis functions and their radii are also user-defined. Moreover, a lower bound (Fig. 3e) on the marginal likelihood can be computed in this case, which can be used to monitor the convergence and is helpful to compare segmentation results. The computation of the lower bound is given in Appendix C.

5. Results

5.1. Datasets

The proposed method was evaluated on four publicly available datasets: the BRATS 2017 training and validation datasets (Menze et al., 2015), the LIDC dataset (Armato III et al., 2011), the training data from the MSSEG challenge (Commowick et al., 2018) and finally the COCO 2017 validation dataset (Lin et al., 2014).

The BRATS 2017 datasets consist of multisequence preoperative MR images of patients diagnosed with malignant brain tumors. It includes 285 patients for the training dataset and 46 for the validation set. Four MR sequences are available for each patient: T1-weighted, post-contrast (gadolinium) T1-weighted, T2-weighted and FLAIR. All the images have been pre-processed: skull-stripped, registered to the same anatomical template and re-sampled to 1 mm^3 resolution. Ground truth segmentations of the brain tumors are provided only for the training set.

The LIDC dataset comprises 1018 pulmonary CT scans with 0.6 mm to 5.0 mm slice thickness. The in-plane pixel size ranges from 0.461 mm to 0.977 mm. Each scan was reviewed by 4 radiologists who annotated lesions of sizes ranging from 3 mm to 30 mm. Annotations include localization and manual delineations

of the nodules. Up to 4 segmentations can be available for the same nodule, depending on the number of radiologists who considered the lesion to be a nodule. In this paper, all scans were first re-sampled to 1 mm^3 resolution as pre-processing step, and we restrict the analysis to nodules of diameter above 20 mm, i.e. 309 segmentations.

The MSSEG training dataset contains MR data from 15 multiple sclerosis (MS) patients. Manual delineations of lesions were performed on the FLAIR sequence by seven experts.

Finally, COCO is a large-scale object detection and segmentation dataset of real world images. The 2017 validation set contains 5000 images with 80 object categories, ground truth object classification, object localization and segmentation. To annotate such a large number of images, the authors resorted to a crowd-sourcing annotation pipeline.

5.2. Unsupervised indices

As discussed in section 1, different indices have been proposed in prior works for unsupervised segmentation evaluation. We selected 4 of them in order to provide a qualitative and quantitative comparison with our approach. They all involve the computation of 2 metrics, the former measuring the intra-region uniformity while the latter gives an estimate of the inter-region disparity.

Three out of the four indices are taken from Zhang et al. (2008): Zeb, η and F_{RC}. The last one was introduced in Johnson and Xie (2011) and is denoted by GS in this paper. Formula are given in Appendix A.

5.3. Setting hyperparameters

5.3.1. Width of the narrow band

As noted in section 3.3, the analysis is restricted to a narrow band alongside the input segmentation's contour. The width of this narrow band controls the extent of the region taken into account for learning the appearance models for both background and foreground and fitting the regularization model.

We assessed the sensitivity of the results to this hyperparameter on the BRATS training set and LIDC dataset. We applied the algorithm for several narrow band widths using FDSP as a spatial prior. Different ASE values were obtained for each segmentation depending on the narrow band setting. We then analysed the stability of the sets made by the 40 segmentations with the largest ASE values by computing pairwise intersection over union (IoU) coefficients. A value of 1.0 indicates that the 40 images are the same for a pair of narrow band widths. The outcome is shown in Fig. 4.

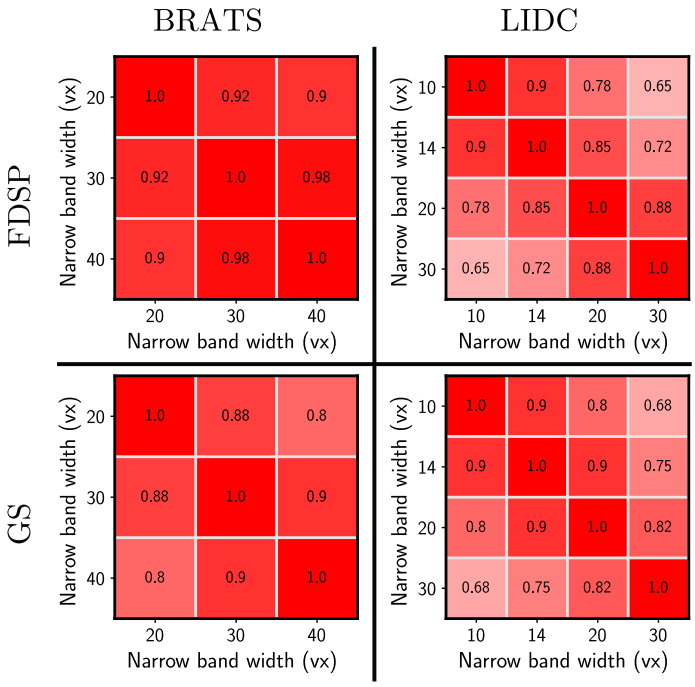


Figure 4: Sensitivity analysis of the narrow band width for the BRATS and LIDC datasets with our approach using FDSP as a spatial prior (first row) or with the unsupervised indicator *GS* (second row). Matrices show IoU scores computed between the sets made of the 40 segmentations with largest ASE (first row) or largest *GS* score (second row).

While the sets from the BRATS training set are rather stable, those from the LIDC dataset show some variability. An explanation of the sensitivity of LIDC segmentations to the narrow band width can be found in Fig. 5. If the narrow band is too wide, the high intensity differences between the pleura and

the lung parenchyma lead appearance models of nodules close to the pleura to leak outside the lung.

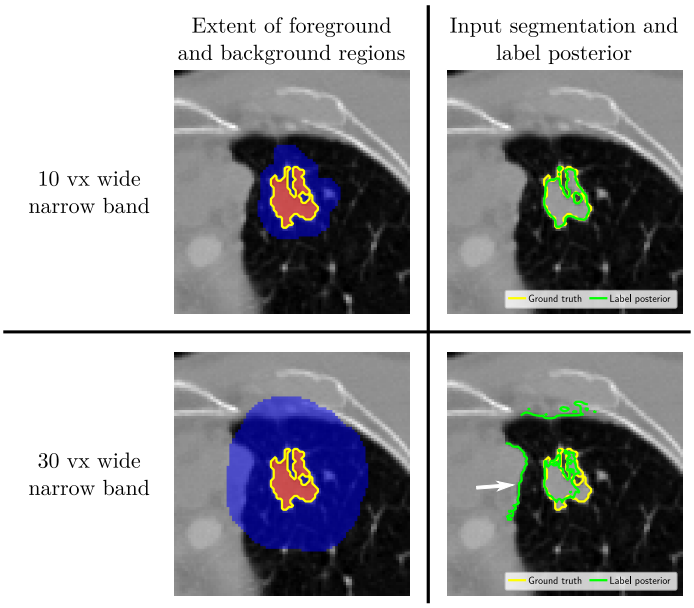


Figure 5: Example of a nodule segmentation from LIDC where the result of the quality assessment is different depending on the narrow band width. If too large, the appearance model of the foreground leaks inside the pleura leading to an irrelevant result.

In brief, the sensitivity of the algorithm with respect to the narrow band's width varies from case to case. As the computation time is not a bottleneck for FDSP regularization, we propose in practice to perform the analysis with different width settings and then choose the one leading to the most stable and reasonable results.

We would also like to underline that previously published unsupervised indices are likewise sensitive to the width of the narrow band. An example is shown in Fig. 4 for the indicator *GS*. In order to provide a fair comparison between approaches, the computations of the selected unsupervised indices are always performed on the same narrow band as the one used for our method.

5.3.2. Other hyperparameters

Among the parameters that need to be defined by the user is the number of components for the mixtures of multivariate Student's t-distributions. It is fixed to 7 in all our experiments. This parameter is not so sensitive as unnecessary components will be pruned by the Dirichlet prior and removed from the model.

The number of remaining parameters depends on the chosen spatial prior. For an MRF prior, the user needs to provide a value for β, which controls the strength of the regularization. We tested 3 values for this hyperparameter throughout our experiments: 0.2, 1 and 3. For a GLSP regularization, the user has to define a dictionary of basis functions whose key parameters are the step between each basis function and their radii. They likewise control the amount of regularization. In this paper, we set the step to 6 vx and the radius to 17 vx, except for the LIDC dataset for which the step was set to 4 vx and the radius to 12 vx. Finally, for the regularization using an FDSP prior, no further parameter needs to be set by the user as the model's hyperparameters are all learnt automatically, which is a great advantage in comparison with the first two approaches.

5.4. Qualitative analysis

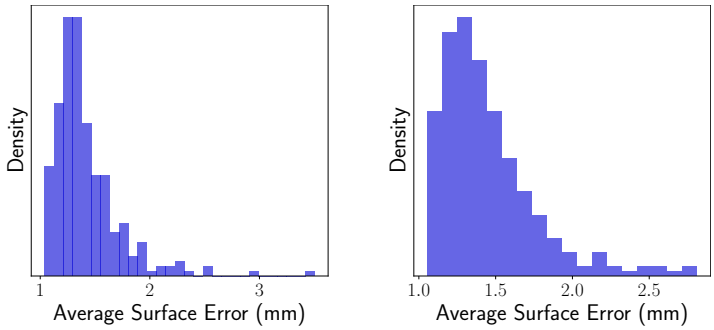


Figure 6: ASE distributions for the analysis of ground truth segmentations from the BRATS (6a) and LIDC datasets (6b). Samples from the left tail are identified as explained by the model while samples from the right tail are classified as challenging.

In the case of segmentations produced by human raters, possibly with the

help of interactive annotation tools, it is very useful to be able to rank segmentations, highlight potentially difficult segmentations and track possible errors in large databases.

In this section we present some results from two datasets of medical images, whole brain tumor segmentations from the BRATS 2017 training set and pulmonary nodule segmentations from the LIDC dataset. On average, one minute is required to complete the quality control workflow for a 3D image from the BRATS dataset using an MRF or FDSP regularization. The inference time increases to 4 minutes for a model with a GLSP prior. The computation time of course also depends on the size of the segmented structure and on the extent of the narrow band.

Computation of the ASE for each segmentation allows the distribution for the whole dataset to be drawn. Histograms obtained with FDSP regularization are shown in Fig. 6. They present a similar shape, with a short left tail, a single peak and a heavier right tail. Cases in the right tail isolated from the rest of the distribution are atypical and possibly include errors. Samples from the left and right tail are shown in Figs. 7 and 8, respectively. For both datasets, cases with larger ASE are clearly more challenging than the cases taken from the left tail.

Furthermore, one can see that contours in the right tail samples from BRATS are more irregular and that intensity variations in some regions are very weak making their accuracy questionable. Those contours were probably extracted through thresholding instead of being manually drawn as was permitted in the annotation process (Jakab, 2012; Menze et al., 2015). Similarly, some contours in the right tail samples from LIDC cross regions of uniform intensity and therefore require other priors like shape to be explained. Yet, the contours are far from obvious in some areas in comparison with the left tail samples. Therefore, our approach fulfills its role of extracting challenging, possibly suspicious, cases within a dataset.

We present in Fig. 9 a qualitative comparison between the spatial priors proposed for our approach and the unsupervised indices presented in section 5.2.

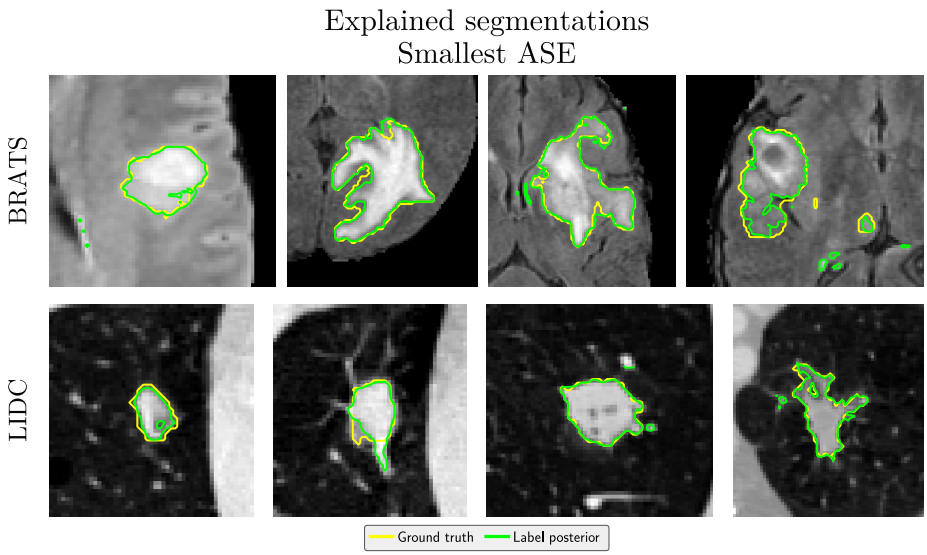


Figure 7: Segmentations with the smallest ASE taken from the left tail of the distributions. Cases are ranked according to their ASE value (Largest values to the right) and slices with largest ground truth area are shown. The width of the narrow band is 30 vx for BRATS and 10 vx for LIDC.

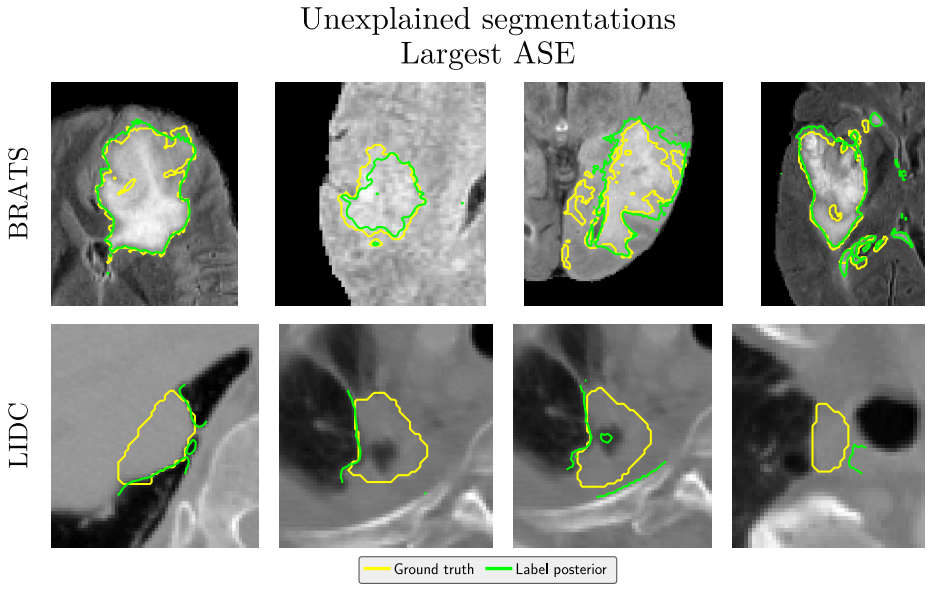


Figure 8: Segmentations with the largest ASE taken from the right tail of the distributions. Cases are ranked according to their ASE value (Largest values to the right) and slices with largest ground truth area are shown. The width of the narrow band is 30 vx for BRATS and 10 vx for LIDC.

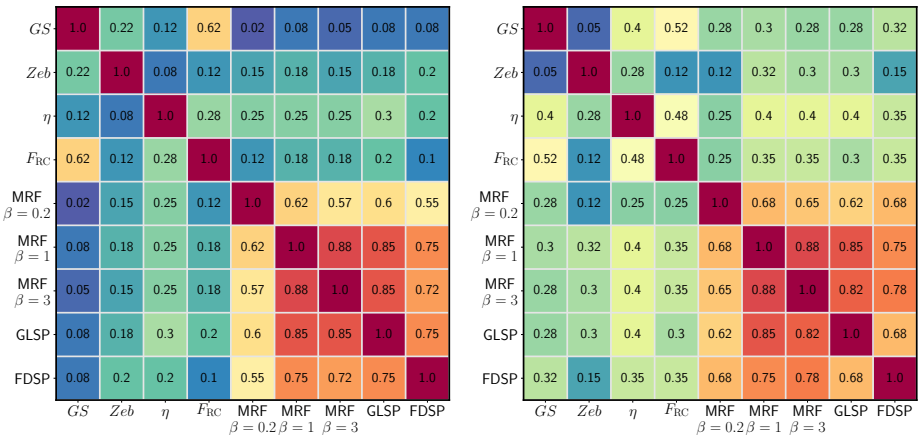


Figure 9: Comparison of different approaches on the BRATS training set (9a) and LIDC (9b). The 40 segmentations with largest ASE or indicator score are compared using IoU. The width of the narrow band is 30 vx for BRATS and 10 vx for LIDC.

Each dataset was sorted according to those indices and the 40 segmentations with largest ASE/score were extracted. The variability of this set of suspicious segmentations across unsupervised methods was studied by computing the pairwise IoU.

First, we note that our approaches and the unsupervised indices yield different sets of suspicious segmentations, as the IoU score is always less than 0.4. Furthermore, the unsupervised indices lead to inconsistent results on both datasets which make their performance highly unreliable on medical images. One possible explanation is that those methods were designed for 2D color images with large contrast and may not scale well to 3D medical images.

If we now compare the different regularization strategies proposed for our approach, we observe that the level of regularization has some impact. Indeed, there is a significant variability of results with the value of β for the MRF prior. This observation supports using the last regularization strategy proposed, the FDSP prior, as in this case all hyperparameters are learnt in a data-driven way.

5.5. Quantitative analysis

In order to perform a quantitative comparison of the different methods, we need to have a grading of the quality of all segmentations. As they are easier to obtain for real world images than medical images, we propose to conduct this quantitative assessment on the COCO dataset which contains real world pictures with a large variability among them, with grayscale or color images, segmented structures of variable sizes and large ranges of noise level.

5.5.1. Quality grading process

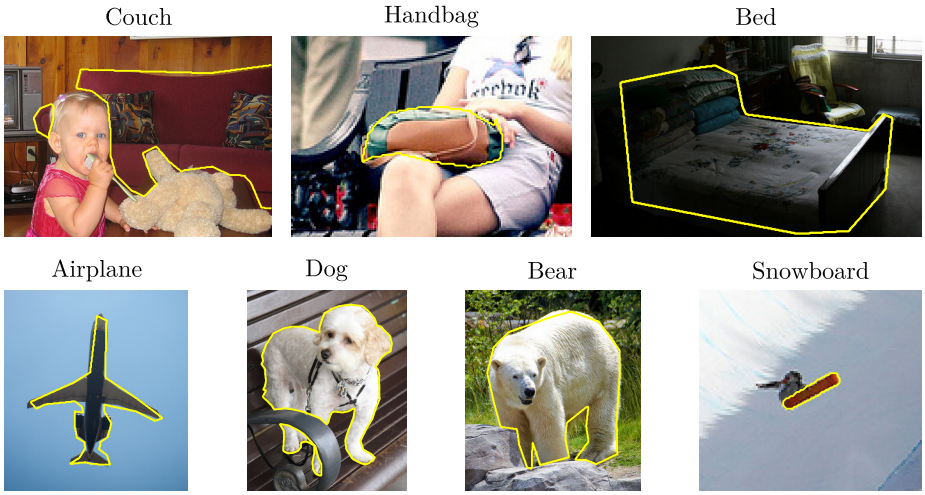


Figure 10: Examples of ground truth segmentations from the COCO dataset representing the 7 selected object categories.

Seven object categories from the COCO dataset were selected for the quantitative assessment: airplane, bear, dog, snowboard, couch, bed and handbag (see Fig 10). Each segmentation was ranked according to the different methods, leading to 9 distributions for each object category: FDSP, GLSP, MRF with 3 values of β and the 4 unsupervised indices. The width of the narrow band was set to 30 px (pixels) for all approaches. Since grading the entire set of images would have been too time-consuming, we chose to focus on the right tails of the distributions (segmentations with largest ASE or indicator score) where the suspicious cases are expected to lie.

For each distribution, we extracted segmentations from the right tail corresponding to 20% of the total distribution. If the number of extracted segmentations was larger than 40, only the 40 cases with largest ASE/score were retained. All segmentations were pooled leading to a total of 703 delineations.

Six raters were then recruited in order to grade each segmented image as good or poor through a custom application presenting the segmentations in a random order. Raters were asked to repeat the annotation twice in order to estimate the intra-rater variability. The intra-rater variability was found to be slightly lower than the inter-rater variability, with a mean rate of identical responses of 83% for the former and of 73% for the latter.

5.5.2. Performance comparison

The objective is to compare the segmentation quality among the right tails of the distributions given by the different approaches. These tails correspond to segmentations with the largest ASE or index score. The percentage of cases rated as poor by the raters strongly depends on the size of the set of segmentations extracted from the right tail of each distribution. Yet, it is useful to compare two quality control algorithms since a better algorithm is expected to have a greater proportion of segmentations annotated as poor by the raters than a worse one.

Each segmentation was assigned to a quality category, good or poor, after taking the mean across the raters' responses. Proportions of poor segmentations per approach were then derived for each object category. Distributions of these proportions over the 7 object categories are shown in Fig. 11. Two observations can be made. First, our approaches show competitive results as they lead to higher mean and median proportions of poor segmentations than the unsupervised indices. Second, no regularization strategy seems better than the others. In particular, variations of the value of β do not affect the results very much for the MRF.

Fig. 12 is obtained after pooling all object categories. To assess the robustness of the results, different thresholds are used to select the segmentations

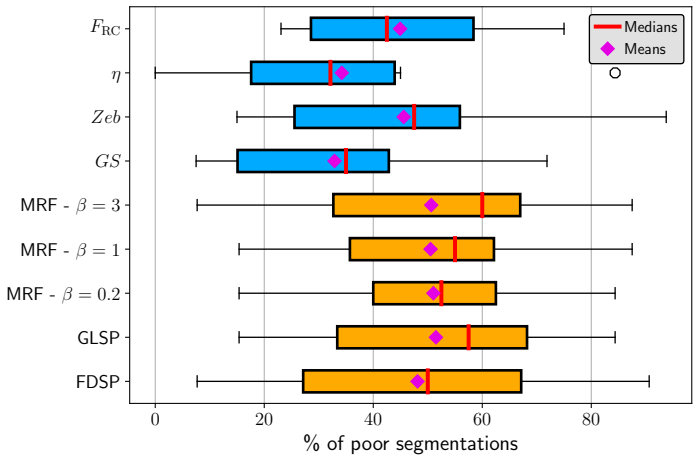


Figure 11: Distribution of the proportion of poor segmentations over the 7 object categories for each approach. The mean over the raters is taken as the final label for each segmentation.

taken into account, depending on the level of agreement among raters' responses. Fisher's exact test is used to assess the difference between the proportion for a given approach and the one obtained with an FDSP prior. Three unsupervised indices η, Zeb and GS, are found to give significantly different results than our approach with FDSP regularization, regardless of the threshold. More generally, our approaches always lead to a higher percentage of poor segmentations than the indices. Again, all regularization strategies seem to be appropriate. The results seem to be stable with respect to the level of regularization enforced by β.

5.5.3. Comparison with inter-rater variability

Assessing the quality of segmentations inside a medical imaging dataset is difficult without any expert knowledge. However, some datasets provide several segmentations of the same image produced by different experts. For instance, up to four segmentations are available for each nodule in the LIDC dataset and MS lesions in the MSSEG training dataset were delineated by seven radiologists. The inter-rater variability measures the level of agreement between the experts. It is reasonable to assume that images for which there is a low level of agreement

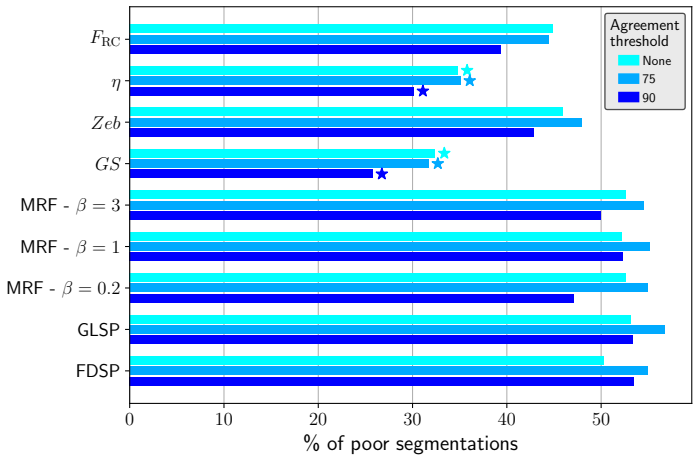


Figure 12: Proportion of poor segmentations after pooling all object categories. Only segmentations with a raters' agreement above a given threshold are retained in the computation of the proportion. Results found to be significantly different from the ones given by the FDSP prior with Fisher's exact test and p-value 0.05 are marked with star symbols $\star$.

between the experts are more challenging than others. Therefore we study in this section the relationship between inter-rater variability and the score produced by our unsupervised model.

Tabs. 1 and 2 show the correlation coefficient between the inter-rater variability and the Dice score or ASE produced by our model on the LIDC and MSSEG datasets, respectively. We also compare with the four unsupervised indices selected earlier. The inter-rater variability was quantified in three manners: by computing the average Dice score between all pairs of experts, the average pairwise Hausdorff distance (HD) and the average 95% percentile of the pairwise Hausdorff distance (95% HD). It is compared to the average score computed on the different raters' segmentations for each unsupervised method. For the LIDC dataset, we discarded all nodules annotated by a single radiologist, leaving a total of 87 nodules.

Better correlations are achieved on the MSSEG dataset than on the LIDC dataset. Furthermore, our approach differs significantly from the other unsupervised indices with much larger correlation values. The others (except Zeb) exhibit indeed coefficients close to zero.

	Inter-rater variability		
	Avg Dice score	Avg HD	Avg 95% HD
Avg F_{RC}	0.11	0.05	-0.03
Avg Zeb	-0.34	0.12	0.2
Avg η	0.13	-0.18	-0.12
Avg GS	0.01	0.03	0.05
Avg Dice score between S and M	**0.47**	**-0.32**	**-0.39**
Avg ASE between S and M	0	0.05	0.03

Table 1: Values of the correlation coefficient between the inter-rater variability and the average score given by different methods on the LIDC dataset. The width of the narrow band is 10 vx and FDSP was used as a spatial prior for our model.

	Inter-rater variability		
	Avg Dice score	Avg HD	Avg 95% HD
Avg F_{RC}	0.17	-0.21	-0.36
Avg Zeb	-0.72	0.14	0.44
Avg η	-0.55	0.03	0.32
Avg GS	0.06	-0.19	-0.25
Avg Dice score between S and M	**0.81**	**-0.49**	**-0.7**
Avg ASE between S and M	-0.64	0.47	0.67

Table 2: Values of the correlation coefficient between the inter-rater variability and the average score given by different methods on the MSSEG dataset. The width of the narrow band is 20 vx and FDSP was used as a spatial prior for our model.

We further analyse the link with inter-rater variability by showing some examples from both datasets on Fig. 13. The first row presents results on the MSSEG dataset, where the correlation is quite good (0.81). Case A has a high inter-rater variability and is labelled as challenging by our model (low average Dice between the inputs and the model). Indeed, only three raters out of seven considered that some lesions were visible on the slice presented in Fig. 13b. Moreover, the low intensity contrast does not help to understand the segmentations. On the other hand, case B is better explained by the model with a good agreement between the experts, as shown in Fig. 13c.

The bottom row shows poorer results on the LIDC dataset. Two contradictory cases are highlighted. The first one, case C, has a low inter-rater variability but is predicted as challenging by our model (low average Dice score between S and M). The two radiologists are indeed giving close contours (Fig. 13e) but it is also clear that the case is challenging according to the assumptions of our model. In the image regions highlighted by the arrows, the contours are indeed crossing areas of uniform intensity distribution, which make them more difficult to understand. On the other hand, case D is a typical case illustrating the limitations of our model (Fig. 13f). Raters disagree about the extent of the nodule, but all segmentations correspond to visible boundaries and match the assumptions of our model. One possible explanation for the poorer correlation obtained on the LIDC dataset is that the annotations were made in two stages, the second stage allowing radiologists to see the annotations made by the other experts in the first stage. This may have led to a decrease in inter-rater variability.

This analysis shows that in some cases, the inter-rater variability may not be a good surrogate of the difficulty of a segmentation. Raters may provide similar segmentations despite the fact that they are not close to visible boundaries (Case C in Fig. 13e) in the image. In that case, a low inter-rater variability is associated with a difficult segmentation.

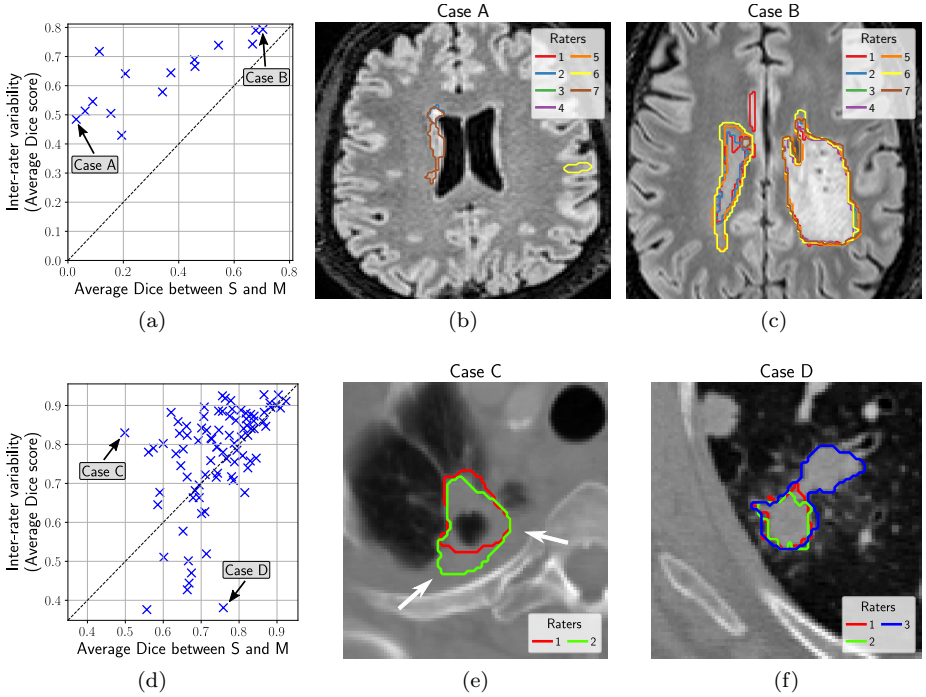
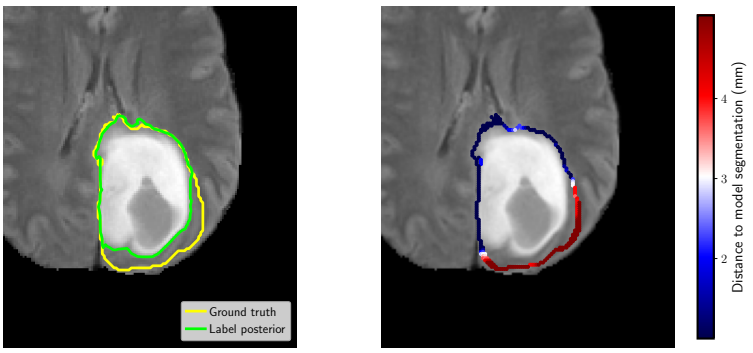


Figure 13: Correlation between the inter-rater variability and the difficulty of a segmentation as predicted by our model on the MSSEG dataset (top row) and LIDC dataset (bottom row).

5.6. Results interpretability

The previous section demonstrated how well our approach performed in extracting suspicious segmentations from a dataset in an unsupervised manner. However, it also differs from approaches proposed in the literature regarding the output of the algorithm. For instance, the unsupervised indices output only a scalar score as a ratio of 2 metrics measuring the intra-region homogeneity and the inter-region dissimilarity. In our case, the output of the algorithm is a new segmentation used as a comparison tool. Although this segmentation must not be seen as a surrogate ground truth, it can help to visually understand why a segmentation is considered atypical, that is, has a large ASE, which is not possible with the indices.

Voxels lying on the input segmentation border can thus be colored depending on their distance to the model segmentation contour, as shown in Fig. 14. When dealing with 3D medical images with a large number of slices, it is useful to be able to retrieve quickly the most problematic regions according to the model. Identifying the most suspicious slices is not possible with approaches outputting a simple score. Last but not least, the model segmentation could also be used as a guide for the correction of poor cases.



(a) (b)

Figure 14: Interpretability of the result given by our approach on a brain tumor segmentation from BRATS. (14a) Ground truth segmentation and label posterior given by the probabilistic model with FDSP regularization. (14b) Coloring of voxels lying on the ground truth border depending on their distance to the ouput of the probabilistic model.

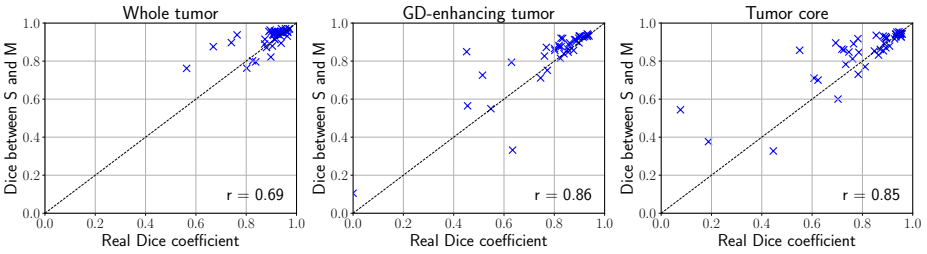


Figure 15: Real Dice coefficient versus Dice score between the prediction S of the CNN and the probabilistic segmentation M with FDSP prior exhibiting good correlation. Results are shown for a narrow band width of 30 vx on 3 tumor compartments.

5.7. Surrogate segmentation performance

In this section we investigate if metrics estimated by our segmentation quality assessment algorithm can be correlated with the overall segmentation performance of an algorithm. In particular, we consider the segmentations generated by a convolutional neural network (CNN) detailed in Mlynarski et al. (2019) on 46 test images of the BRATS 2017 challenge. The Dice score computed between the predicted segmentation S and the one obtained by thresholding the posterior map, M, is then compared to the true Dice index obtained by uploading the generated segmentation on the evaluation website of the challenge. In other words, we want to assess if the Dice score between S and M can be predictive of the real segmentation performance of the algorithm.

Correlations obtained with an FDSP prior on a narrow band of width 30 vx are given in Fig. 15 for the 3 different tumor compartments and are all above 0.69 with few outliers. Fig. 16 present correlation coefficients with all regularization strategies and for different values of the narrow band width. The coefficients are very similar across the approaches and are little affected by the variations of the narrow band width.

However, we do not find that this approach always predicts the performance of segmentation algorithms well. For instance, we have noticed poor predictions for most categories in the COCO dataset. This can be explained by the fact that good performance predictions can only be obtained when the segmented structure follows the model assumptions, that is, the background and foreground

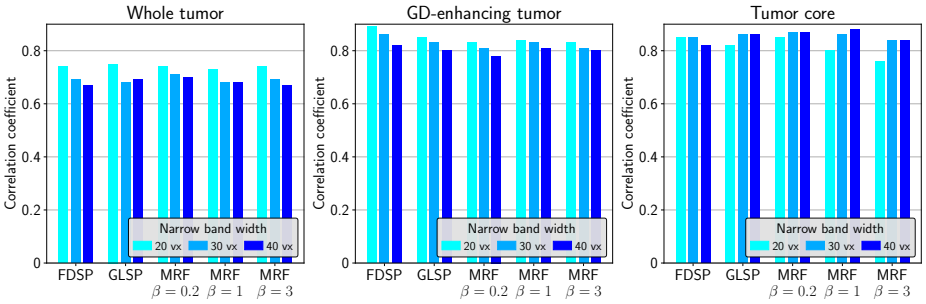


Figure 16: Values of the correlation coefficient between the real Dice and estimated Dice score for different regularization strategies and different widths of the narrow band.

regions have different mixtures of Student's t-distributions.

5.8. Discussion

The proposed unsupervised quality control method was shown to efficiently and automatically isolate challenging or atypical segmentation cases from a whole dataset. It was shown to outperform four previously introduced segmentation quality indices on the COCO dataset. Furthermore, those four indices do not provide stable results on the LIDC and BRATS medical datasets. The proposed algorithm does not produce a classification between good or poor segmentations but rather a ranking between cases within a dataset.

The genericity of the algorithm allows it to work on any type of object category or image (2D RGB or 3D grayscale images). We demonstrated the ability of the method to handle a wide range of segmentations, from small structures (lung nodules) to large brain tumor delineations. Yet, the approach is not suited for very tiny objects since a reasonable size is required to have a reliable estimation of the intensity parameters. Also, the spatial prior is likely to wipe out the segmentation if its area is really too small. Furthermore, the genericity of the algorithm may also be considered as a limitation when focusing on a specific structure of interest. For instance, if we aim at segmenting objects from the car category on the COCO dataset, a contour perfectly following intensity boundaries but around another object category would not be identified as atypical. To this end, one would need to also monitor several specific features of

that structure such as its color, size or shape, which amounts to performing a supervised quality control as in Xu et al. (2009). This limitation is shared by all unsupervised quality control methods.

Another limitation is the difficulty to distinguish boundaries in areas with low intensity contrast. Our method is based on mixtures of Student's t-distributions, which is already a far more general assumption than some previous unsupervised approaches that hypothesize a unique Gaussian distribution in each region (Zhang et al., 2008). Furthermore, our Bayesian formulation integrates intensity and smoothness assumptions into a single probabilistic model, as opposed to previous unsupervised methods, which require weighting of the heterogeneity and homogeneity terms.

Different spatial regularization strategies are proposed and tested in this paper. Quantitative assessment on COCO seems to indicate that all approaches lead to similar results. However, the FDSP prior based on derivative penalization does not require any hyperparameters to be set while keeping the computation time low, supporting its use in preference to the others.

Finally, compared to learning-based approaches such as Kohlberger et al. (2012) or Robinson et al. (2018) and also to previous unsupervised indices which only output a score, our method provides an explanation for the mismatches between the posterior probabilities M and the input segmentation S. This is a major advantage considering the growing importance of providing interpretable models.

6. Conclusion

Image segmentation is an important task in medical image analysis and computer vision. Quality control assessment of segmentations is therefore crucial, but the trend towards the generation of large databases makes any human-based monitoring onerous if not impossible. This paper introduces a new framework for generic quality control assessment which relies on a simple and unsupervised model. It has the advantage of not requiring a priori any knowledge about

the segmented objects nor a subset of trusted images to be extracted. This is especially suited to the monitoring of manually created segmentations, where potential errors can be found, as shown by our results. Its application to segmentations generated by algorithms is also of great interest and in some cases can be used as a surrogate for segmentation performance.

The proposed generic segmentation model produces contours of variable smoothness that are mostly aligned with visible boundaries in the image. Three regularization strategies were proposed in this paper and produced similar results. However, the prior based on derivative penalization has the great advantage of allowing an automated estimation of all hyperparameters with variational Bayesian inference, which is not possible within the classical MRF framework.

Extensive testing has been performed on different datasets containing various types of images and segmented structures, showing the ability of the method to isolate atypical cases and therefore to perform quality control assessment. Comparison with unsupervised indices from the literature proved our approach to be effective and competitive. Coping with multiple foreground labels may be an interesting extension to process multiple regions of interest jointly rather than sequentially. Finally, an interactive use of the proposed algorithm during the manual delineation of structures in images is an exciting perspective to help reduce the inter-rater variability in the context of crowdsourcing.

Appendix A. Unsupervised indices

We give in this section the formula used to compute the unsupervised indices. We denote by R the number of regions inside an image (typically 2 here, for the foreground and background regions). R_j denotes the set of voxels in region j and $|R_j|$ is the number of voxels in region j. Each indicator requires the computation of an intra-region uniformity metric IU and an inter-region disparity metric ID.

Appendix A.1. Zeb (Zhang et al., 2008)

$$\mathrm{IU}_j = \frac{1}{|R_j|} \sum_{s \in R_j} \max \left\{ \mathrm{contrast}(s,t), t \in W(s) \cap R_j \right\} , \tag{A.1}$$

where $W(s)$ is the neighborhood of voxel s and:

$$\mathrm{contrast}(s,t) = \frac{1}{\nu} \sum_{i=1}^{\nu} \left| I_s^i - I_t^i \right| . \tag{A.2}$$

$$\mathrm{ID}_j = \frac{1}{|b(R_j)|} \sum_{s \in b(R_j)} \max \left\{ \mathrm{constrast}(s,t), t \in W(s), t \notin R_j \right\} , \tag{A.3}$$

where $b(R_j)$ is the set of pixels on the border of R_j.

The final indicator is given by:

$$Zeb = \frac{\mathrm{IU}}{\mathrm{ID}} = \frac{\sum_j \mathrm{IU}_j}{\sum_j \mathrm{ID}_j} . \tag{A.4}$$

Appendix A.2. F_{RC} (Zhang et al., 2008)

$$\mathrm{IU} = \frac{1}{R} \sum_{j=1}^{R} \frac{|R_j|}{N} e^2(R_j) , \tag{A.5}$$

where:

$$e^2(R_j) = \frac{1}{\nu} \sum_{i=1}^{\nu} \sum_{s \in R_j} \left(I_s^i - \hat{I}_{R_j}^i \right)^2 . \tag{A.6}$$

$\hat{I}_{R_j}^i$ is defined for $1 \leq i \leq \nu$ by:

$$\hat{I}_{R_j}^i = \frac{1}{|R_j|} \sum_{s \in R_j} I_s^i . \tag{A.7}$$

$$\mathrm{ID} = \frac{1}{R}\sum_{j=1}^{R}\frac{|R_j|}{N}\left(\frac{1}{|W(R_j)|}\sum_{t\in W(R_j)} D(R_j,R_t)\right)\,, \tag{A.8}$$

where $W(R_j)$ is the set of neighboring regions of R_j and:

$$D(R_j,R_t) = \frac{1}{\nu}\sum_i |\hat{I}^i_{R_j} - \hat{I}^i_{R_t}|\,. \tag{A.9}$$

The final indicator is given by:

$$F_{\mathrm{RC}} = \mathrm{IU} - \mathrm{ID}\ . \tag{A.10}$$

Appendix A.3. η (Zhang et al., 2008)

The background is denoted here by b, while f denotes the foreground.

$$\mathrm{IU} = \frac{N_b}{N}e^2(R_b) + \frac{N_f}{N}e^2(R_f)\,, \tag{A.11}$$

where N_b and N_f are the number of voxels in the background and foreground, respectively, and $e^2(R_j)$ is defined as previously.

$$\mathrm{ID} = \frac{N_b N_f}{N^2}\left(\hat{I}_{R_f} - \hat{I}_{R_b}\right)^2\,, \tag{A.12}$$

where $\hat{I}_{R_j} = \frac{1}{\nu}\sum_{i=1}^{\nu}\hat{I}^i_{R_j}$.

The final indicator is given by:

$$\eta = \frac{\mathrm{IU}}{\mathrm{ID}}\,. \tag{A.13}$$

Appendix A.4. GS (Johnson and Xie, 2011)

$$\mathrm{IU} = \frac{\sum_j |R_j| V_j}{\sum_j |R_j|}\,, \tag{A.14}$$

where V_j is the variance of region j.

The inter-region disparity metric used is the Global Moran's I, defined as:

$$\mathrm{ID} = \frac{R}{\sum_i \sum_{j \neq i} w_{ij}} \frac{\sum_{i=1}^{R} \sum_{j=1}^{R} w_{ij}(y_i - \overline{y})(y_j - \overline{y})}{\sum_{i=1}^{R}(y_i - \overline{y})^2}\,, \tag{A.15}$$

where $w_{ii} = 0$, $w_{ij} = 1$ if R_i and R_j are neighbors and 0 otherwise. y_i is the mean intensity value of region R_i and $\overline{y}$ is the mean intensity value of the image.

The final indicator is given by:

$$\mathrm{GS} = \mathrm{IU} + \mathrm{ID}\ . \tag{A.16}$$

Appendix B. FDSP prior - variational inference

We present in this section the derivation of the variational update formula (11), (12), (13), (14) and (15). The likelihood of the model $p(I, Z, W, \alpha)$ factorizes as $p(I|Z)p(Z|W)p(W|\alpha)p(\alpha)$.

Appendix B.1. Update of $q_Z^*(Z)$

$$\begin{aligned} \log q_Z^*(Z) &= \mathbb{E}_{W,\alpha}[\log p(I|Z) + \log p(Z|W)] + cst\,, \\ &\geqslant \mathbb{E}_{W,\alpha}[\log p(I|Z) + \log F(Z, W, \xi)]\,. \end{aligned} \tag{B.1}$$

Recalling that $p(I|Z) = \prod_n r_n^{Z_n}(1 - r_n)^{1 - Z_n}$, we have:

$$\mathbb{E}_{W,\alpha}[\log p(I|Z)] = \sum_n Z_n \log r_n + (1 - Z_n)\log(1 - r_n)\,. \tag{B.2}$$

The prior $p(Z|W) = \prod_n [\sigma(w_n)]^{Z_n} [\sigma(-w_n)]^{1-Z_n}$ is lower bounded by $F(Z, W, \xi)$ to give:

$$\begin{aligned}\mathbb{E}_{W,\alpha}[\log F(Z, W, \xi)] = \sum_n Z_n [\log \sigma(\xi_n) + (\mathbb{E}[w_n] \\ - \xi_n)/2 - \lambda(\xi_n)(\mathbb{E}[w_n^2] - \xi_n^2)] + (1 - Z_n)[\log \sigma(\xi_n) \\ - (\mathbb{E}[w_n] + \xi_n)/2 - \lambda(\xi_n)(\mathbb{E}[w_n^2] - \xi_n^2)] \,.\end{aligned} \tag{B.3}$$

Summing (B.2) and (B.3) and taking the exponential, we have $q_Z^*(Z) \propto \prod_n \rho_{n0}^{1-Z_n} \rho_{n1}^{Z_n}$ where the expressions of ρ_{n0} and ρ_{n1} are given by (11). With the normalization constraint, we finally obtain $q_Z^*(Z) = \prod_n \eta_{n1}^{Z_n} \eta_{n0}^{1-Z_n}$ with $\eta_{nk} = \rho_{nk} / \sum_k \rho_{nk}$ for $k \in \{0, 1\}$.

Appendix B.2. Update of $q_W^*(W)$

$$\begin{aligned}\log q_W^*(W) &= \mathbb{E}_{Z,\alpha}[\log p(Z|W) + \log p(W|\alpha)] + cst \,, \\ &\geqslant \mathbb{E}_{Z,\alpha}[\log F(Z, W, \xi) + \log p(W|\alpha)] \,.\end{aligned} \tag{B.4}$$

With the expression of $p(W|\alpha)$ given in (5) and assuming that $q_W(W) = \prod_n q_{w_n}(w_n)$, we obtain:

$$\begin{aligned}\log q_{w_n}^*(w_n) = -\frac{1}{2}\left[2\lambda(\xi_n)\left(w_n - \frac{1}{2\lambda(\xi_n)}\left(\eta_{n1} - \frac{1}{2}\right)\right)^2\right] \\ - \frac{1}{2}\mathbb{E}_{w_j,\, j \neq n}\left[\frac{\alpha}{2}\sum_d (w_n - w_{\delta_d(n-2)})^2 + (w_n - w_{\delta_d(n+2)})^2\right] \,.\end{aligned} \tag{B.5}$$

By identifying the quadratic and linear terms in w_n, we obtain the formula for $\boldsymbol{\Sigma}_{w_n}$ and $\boldsymbol{\mu}_{w_n}$ given in (12) and (13).

Appendix B.3. Update of $q^*_\alpha(\alpha)$

$$\begin{aligned} \log q^*_\alpha(\alpha) &= \mathbb{E}_W[\log p(W|\alpha)] + cst\,, \\ &= \mathbb{E}_W\left[\frac{N}{2}\log\alpha - \frac{\alpha}{4}\sum_n\sum_{d=1}^{D}(w_{\delta_d(n+1)} - w_{\delta_d(n-1)})^2\right] + cst\,. \end{aligned} \tag{B.6}$$

Assuming $q^*_\alpha(\alpha)$ to be a Dirac distribution, we take the derivative of (B.6) with respect to α which leads to the update formula given in (14).

Appendix B.4. Update of $q^*_{\xi_n}(\xi_n)$

$$\begin{aligned} \log q^*_{\xi_n}(\xi_n) &= \mathbb{E}_{Z,W}[\log F(Z,W,\xi)] + cst\,, \\ &= \log\sigma(\xi_n) - \frac{\xi_n}{2} - \lambda(\xi_n)(\mathbb{E}[w_n^2] - \xi_n^2) + cst\,. \end{aligned} \tag{B.7}$$

Taking the derivative with respect to ξ_n and setting it equal to zero gives $\lambda'(\xi_n)(\mathbb{E}[w_n^2] - \xi_n^2) = 0$. As $\lambda'(\xi_n) \geqslant 0$, we finally obtain the formula reported in (15).

Appendix C. FDSP prior - lower bound

The lower bound on the log-likelihood is used as a stopping criterion. To compute $\mathcal{J}(q)$, we need to evaluate the right hand side of (10):

$$\begin{aligned} \mathcal{J}(q) = \mathbb{E}[\log p(I|Z)] + \mathbb{E}[\log F(Z,W,\xi)] + \mathbb{E}[\log p(W|\alpha)] \\ - \mathbb{E}[\log q_Z(Z)] - \mathbb{E}[\log q_W(W)]\,. \end{aligned} \tag{C.1}$$

The values of the different expectations can be computed and are reported below.

$$\mathbb{E}[\log p(I|Z)] = \sum_n \eta_{n1} \log r_n + \eta_{n0} \log(1 - r_n) \,. \tag{C.2}$$

$$\begin{aligned} \mathbb{E}[\log F(Z, W, \xi)] = & \sum_n \eta_{n1} \mathbb{E}[w_n] + \log \sigma(\xi_n) \\ & - \frac{\mathbb{E}[w_n] + \xi_n}{2} - \lambda(\xi_n)(\mathbb{E}[w_n^2] - \xi_n^2) \,. \end{aligned} \tag{C.3}$$

$$\mathbb{E}\left[\log p(W|\alpha)\right] = \frac{N}{2} \log \alpha - \frac{N}{2} \,. \tag{C.4}$$

$$\mathbb{E}\left[\log q_Z(Z)\right] = \sum_n \eta_{n1} \log \eta_{n1} + \eta_{n0} \log \eta_{n0} \,. \tag{C.5}$$

$$\mathbb{E}\left[\log q_W(W)\right] = -\frac{1}{2} \sum_n \log \boldsymbol{\Sigma}_{w_n} \,. \tag{C.6}$$

Acknowledgments

This work was partially funded by the French government, through the UCA$^{\text{JEDI}}$ and 3IA Côte d'Azur "Investments in the Future" projects managed by the National Research Agency (ANR) with the reference numbers ANR-15-IDEX-01 and ANR-19-P3IA-0002 and supported by the Inria Sophia Antipolis - Méditerranée "NEF" computation cluster.

References

Ambroise, C., Govaert, G., 1998. Convergence of an EM-type algorithm for spatial clustering. Pattern Recognition Letters 19, 919 – 927.

Arbelle, A., Elul, E., Riklin Raviv, T., 2019. QANet – Quality Assurance Network for Image Segmentation. arXiv e-prints , arXiv:1904.08503arXiv:1904.08503.

Archambeau, C., Verleysen, M., 2007. Robust Bayesian clustering. Neural Networks 20, 129 – 138.

Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., et al., 2011. The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans. Medical Physics 38, 915–931.

Audelan, B., Delingette, H., 2019. Unsupervised Quality Control of Image Segmentation Based on Bayesian Learning, in: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Springer International Publishing, Cham. pp. 21–29.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg.

Chabrier, S., Emile, B., Rosenberger, C., Laurent, H., 2006. Unsupervised Performance Evaluation of Image Segmentation. EURASIP Journal on Advances in Signal Processing 2006, 096306.

Commowick, O., Istace, A., Kain, M., Laurent, B., et al., 2018. Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure. Scientific Reports 8, 13650.

DeVries, T., Taylor, G.W., 2018. Leveraging Uncertainty Estimates for Predicting Segmentation Quality. arXiv e-prints , arXiv:1807.00502arXiv:1807.00502.

Gao, H., Tang, Y., Jing, L., Li, H., Ding, H., 2017. A Novel Unsupervised Segmentation Quality Evaluation Method for Remote Sensing Images. Sensors 17, 2427.

Huang, C., Wu, Q., Meng, F., 2016. QualityNet: Segmentation quality evaluation with deep convolutional networks, in: 2016 Visual Communications and Image Processing (VCIP), pp. 1–4.

Hui Zhang, Cholleti, S., Goldman, S.A., Fritts, J.E., 2006. Meta-Evaluation of Image Segmentation Using Machine Learning, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), pp. 1138–1145.

Jaakkola, T.S., Jordan, M.I., 2000. Bayesian parameter estimation via variational methods. Statistics and Computing 10, 25–37.

Jakab, A., 2012. Segmenting Brain Tumors with the Slicer 3D Software. `http://www2.imm.dtu.dk/projects/BRATS2012/Jakab_TumorSegmentation_Manual.pdf`. Accessed 22 June 2020.

Johnson, B., Xie, Z., 2011. Unsupervised image segmentation evaluation and refinement using a multi-scale approach. ISPRS Journal of Photogrammetry and Remote Sensing 66, 473 – 483.

Jungo, A., Reyes, M., 2019. Assessing Reliability and Challenges of Uncertainty Estimations for Medical Image Segmentation, in: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Springer International Publishing, Cham. pp. 48–56.

Keshavan, A., Datta, E., McDonough, I.M., Madan, C.R., Jordan, K., Henry, R.G., 2018. Mindcontrol: A web application for brain segmentation quality control. NeuroImage 170, 365 – 372. Segmenting the Brain.

Kohlberger, T., Singh, V., Alvino, C., Bahlmann, C., et al., 2012. Evaluating Segmentation Error without Ground Truth, in: Ayache, N., Delingette, H., Golland, P., Mori, K. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 528–536.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., et al., 2014. Microsoft COCO: Common Objects in Context, in: Computer Vision – ECCV 2014, Springer International Publishing, Cham. pp. 740–755.

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., et al., 2015. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Transactions on Medical Imaging 34, 1993–2024.

Mlynarski, P., Delingette, H., Criminisi, A., Ayache, N., 2019. 3D convolutional neural networks for tumor segmentation using long-range 2D context. Computerized Medical Imaging and Graphics 73, 60 – 72.

Pereyra, M., Bioucas-Dias, J.M., Figueiredo, M.A.T., 2015. Maximum-a-posteriori estimation with unknown regularisation parameters, in: 2015 23rd European Signal Processing Conference (EUSIPCO), pp. 230–234.

Robinson, R., Oktay, O., Bai, W., Valindria, V.V., et al., 2018. Real-Time Prediction of Segmentation Quality, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Springer International Publishing, Cham. pp. 578–585.

Robinson, R., Valindria, V.V., Bai, W., Oktay, O., et al., 2019. Automated quality control in image segmentation: application to the UK Biobank cardiovascular magnetic resonance imaging study. Journal of Cardiovascular Magnetic Resonance 21, 18.

Roche, A., Ribes, D., Bach-Cuadra, M., Krüger, G., 2011. On the convergence of EM-like algorithms for image segmentation using Markov random fields. Medical Image Analysis 15, 830 – 839.

Rosenberger, C., Chabrier, S., Laurent, H., Emile, B., 2006. Unsupervised and supervised image segmentation evaluation. Advances in Image and Video Segmentation, IGI Global , 365–393.

Rother, C., Kolmogorov, V., Blake, A., 2004. "GrabCut": Interactive Foreground Extraction Using Iterated Graph Cuts. ACM Trans. Graph. 23, 309–314.

Roy, A.G., Conjeti, S., Navab, N., Wachinger, C., 2019. Bayesian QuickNAT: Model uncertainty in deep whole-brain segmentation for structure-wise quality control. NeuroImage 195, 11 – 22.

Shamir, R.R., Bomzon, Z., 2019. Evaluation of head segmentation quality for treatment planning of tumor treating fields in brain tumors. arXiv e-prints , arXiv:1906.11014arXiv:1906.11014.

Shi, W., Meng, F., Wu, Q., 2017. Segmentation quality evaluation based on multi-scale convolutional neural networks, in: 2017 IEEE Visual Communications and Image Processing (VCIP), pp. 1–4.

Valindria, V.V., Lavdas, I., Bai, W., Kamnitsas, K., et al., 2017. Reverse Classification Accuracy: Predicting Segmentation Performance in the Absence of Ground Truth. IEEE Transactions on Medical Imaging 36, 1597–1606.

Visser, M., Müller, D., van Duijn, R., Smits, M., Verburg, N., Hendriks, E., et al., 2019. Inter-rater agreement in glioma segmentations on longitudinal MRI. NeuroImage: Clinical 22, 101727.

Xu, Y., Kavanagh, P., Fish, M., Gerlach, J., et al., 2009. Automated Quality Control for Segmentation of Myocardial Perfusion SPECT. Journal of Nuclear Medicine 50, 1418–1426.

Zhang, H., Fritts, J.E., Goldman, S.A., 2008. Image segmentation evaluation: A survey of unsupervised methods. Computer Vision and Image Understanding 110, 260 – 280.