Groupwise registration for longitudinal MRI analysis of glioma based on deep learning

Master thesis Claudia Chinea Hammecher - 4495306
Supervisors: Bo Li[1], Esther Bron[1], Frans Vos[2]

[1]BIGR, Department of Radiology and Nuclear Medicine, Erasmus MC, the Netherlands
[2]Department of Imaging Physics, Delft University of Technology, The Netherlands

Abstract

Glioma progression is monitored by routine MR scanning, enabling that tumor growth can be evaluated with respect to earlier time-points. This growth may present both as a mass effect and as an extension of abnormalities into previously healthy tissue. To accurately quantify tumor growth and tumor-induced deformations, longitudinal intrasubject image registration is often used. However, such registration in cases with large deformations and tissue change is highly challenging. Longitudinal image registration may benefit from groupwise strategies in which multiple images are concurrently aligned. This avoids introducing bias towards an a priori-selected reference image. However, existing learning-based methods for image registration mostly concern pair-wise approaches. Moreover, the few proposed learning-based methods for groupwise registration are designed for the analysis of images without pathologies and are prone to fail to register glioma images.

To bridge this gap, we present a learning-based method for the non-linear registration of longitudinal glioma images. We adapt an existing learning-based groupwise method to handle tumor infiltration by means of cost-function masking. The proposed method is able to register glioma images despite the presence of non-correspondences across the time-points by focusing on the normal-appearing tissue similarity. We train the framework both in one resolution and with a multi-stage strategy exploring multiple resolutions.

We evaluate on a dataset from the Glioma Longitudinal AnalySiS consortium and compare it to conventional groupwise registration methods. We achieve comparable Dice coefficients, with higher SSIM and more detailed registrations. These evaluation metrics are further improved when trained as a multi-stage method. The proposed framework preserves the diffeomorphic conditions and the geometric centrality of the deformation fields, while significantly reducing the runtime to under a minute. The proposed methods may serve as an alternative to conventional toolboxes to provide further insight into glioma growth.

1 Introduction

Glioma is the most common type of primary brain tumor, primarily occurring in the glial tissue [1]. Depending on their degree of malignant behavior, gliomas may be classified as grades I through IV according to the World Health Organization. These grades are associated with different prognoses: patients with the lowest grades face a median survival between 5 and 7 years [1], and patients with grade-IV gliomas have a median survival of under 15 months [2]. The current standard practice to treat gliomas includes a maximal safe surgical resection, followed by radiation therapy and chemotherapy [2]. Nevertheless, the highly invasive nature of glioma prevents their complete resection, causing more than half of patients to experience recurrence within the first five years after surgery [3]. Early detection of recurrence is critical for prognosis.

Glioma patients routinely undergo MRI scanning to monitor changes in tumor volume and tissue composition. Such changes may present as a focalized proliferation that compresses the surrounding structures (i.e. mass-effect), or as an infiltration of the tumor into previously normal-appearing tissue [4] (see Fig. 1). Clinical studies have found that the different tumor growth patterns and the compression of particular functional brain areas are correlated to different overall survival times [4, 5]. Accurate quantification of the longitudinal changes of the tumor and nearby tissues might, therefore, enhance our insight regarding how glioma develops and can be optimally treated.

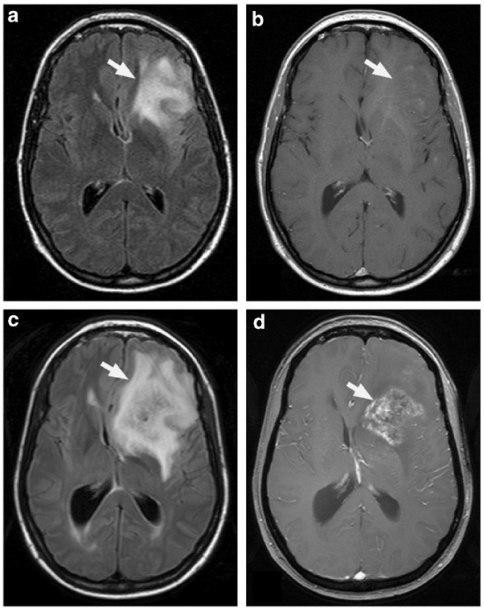

Figure 1: Axial plane MR scans of a woman with left frontal lobe glioma of grade III (a)(b), which progressed to grade IV (c)(d). (a) FLAIR imaging shows increased signal intensity, representing edema in response to infiltrating tumor cells. (b) T1-weighted image. (c) The tumor has further infiltrated as shown in the FLAIR image (arrow) and mass-effects in the ventricles and shift of the midline are visible. (d) The T1-weighted image with contrast agent further indicated the presence of necrosis (arrow) and thus a change in tumor tissue composition [6].

Image registration aims to find a transform that spatially aligns corresponding brain structures in two or more images [7]. The non-linear registration of longitudinal images results in transformations that may describe the anatomical changes between time-points. Hence, this facilitates the evaluation of the spatial patterns of the tumor over time. In longitudinal studies, follow-up images (i.e., moving image) are typically registered to the baseline scan (i.e., reference image). However, resampling these follow-up images and keeping the baseline untouched has been shown to introduce a bias that can affect any subsequent analyses [7]. An alternative is to perform groupwise registration, where all transformations are obtained simultaneously. Images are registered to a common mean-space, circumventing the need to choose a reference image and avoid introducing bias. Joshi et al. [8] proposed a groupwise registration method for the unbiased construction of atlases. The method finds a mean-space by iteratively registering all images in a set to their average image. The pairwise registrations are guided by the intensity mean squared error between a moving and average image. After the images are transformed, the average image is again defined and used as a new reference. Other grouwpwise approaches simultaneously register all

images, such as the method proposed by Huizinga et al. [9] based on Principal Component Analysis. The metric consists of a weighted sum of the eigenvalues. More weight is given to the higher eigenvalues so that their total magnitude shifts to the first few during registration. As the images are aligned, fewer eigenvalues contain most of the information shared across them.

The mentioned registration methods belong to so-called conventional strategies. They solve an independent optimization problem for each set of images. However, this optimization can be time-consuming, taking minutes to hours to perform the registration [10]. In recent years, learning-based approaches have gained increasing popularity in multiple medical image-processing topics, including groupwise image registration. Once trained, these models enable a significant acceleration in the application phase. Che et al. [11], proposed a method for the registration of multimodal images. A modified U-Net takes as input the stacked moving images and estimates their deformation fields. The images are warped and a PCA-based template is constructed. The similarity between the warped images and template guides the registration. Similarly, Zhang et al. [12] propose a method for dynamic MRI sequences. As before, a U-Net yields deformation vector fields for the stacked input images. The model is updated by maximizing the cross-correlation between the warped images and their mean. To ensure that the deformations meet at the mean-space, these methods introduce in their loss function a centrality or cyclic term. These terms minimize the sum of all the deformations or the difference between the transforms of consecutive images. However, this supposes a trade-off between centrality and the similarity metric. Alternatively, centrality may be enforced by subtracting the average from all the deformation fields [13]. A limit of these approaches is that the few proposed learning-based methods for groupwise registration are designed for the analysis of images without pathologies. Glioma infiltration introduces a local change in tissue composition, usually of very different intensity compared to the previously normal-appearing tissue. The similarity metric guiding the deformations will be low within this volume, and aiming to register the anatomical non-correspondence will lead to implausible deformations [14, 15]. We therefore propose a learning-based groupwise registration method to handle non-correspondences in images.

For the registration of images with glioma and other pathologies, several approaches have been proposed, which can be classified into different types. The first type consists of the estimation and registration of quasi-normal images derived from the input scans. These images may be obtained by means of low-rank decomposition [16, 17], but may result in a loss of detail and could affect the precision of the registration. To overcome this, other similar methods propose the use of the low-rank decomposition only within the tumor volume [18], or variational autoencoders optimized with the information of the normal-appearing volumes [19]. These solutions require tumor masks, and may detect large local mass-effects as irregularities to be smoothed out. In the second type, images with pathology can be registered to a healthy atlas. The atlas is seeded with an artificial tumor, which is expanded by a growth model to simulate the mass-effects induced by the tumor [20, 21, 22]. These sophisticated methods rely on accurate modeling of the observed tumor. This is of particular difficulty given the sparsity of glioma growth and the complexity in clinical practice, as the treatment may also have an influence on the tumor tissue composition and its shape change. The last type of methods consists of the registration of images with a cost-function masking strategy [23, 14, 24]. In these approaches, binary maps indicating the regions to exclude are used so that the similarity function can be computed only within the corresponding regions. The masked-out areas do not remain untransformed but are warped assuming continuity of the surrounding deformations. This rather straightforward concept shows significant improvements [23] and could be expanded to groupwise registration given a tumor segmentation.

In this paper, we propose a registration method for longitudinal brain MRI with glioma. We combined the unbiased benefits of groupwise registration and the acceleration of learning-based methods, expanding the existing method [13]. The proposed method was optimized and evaluated on a multi-center dataset from the Glioma Longitudinal AnalySiS consortium. Sets of three longitudinal images were extracted from 61 subjects. The registration performance was evaluated with Dice coefficients, the Structural Similarity Index, the standard deviation across the Jacobian determinant maps and their amount of negative values, and the geometric centrality of the resulting deformation fields. We compared our results to those of the state-of-the-art conventional groupwise methods available in toolboxes Elastix [25], NiftyReg [26] and ANTs [27].

2 Methods

We expanded an existing learning-based groupwise registration approach [13] to take into account tumor presence and growth over time. Given a set of n longitudinal three-dimensional images $\mathcal{I} = \{I_1, \ldots, I_n\}$ taken at n time-points, the proposed framework estimates a set of dense transformations

$\mathcal{T} = \{T_1, \ldots, T_n\}$. These transformations warp the images $\mathcal{I}$ from their native space to the geometrical mean space of the set.

The adopted transformation model is diffeomorphic, achieving smooth and invertible mappings. The diffeomorphic transforms $\mathcal{T}$ are obtained by integrating stationary vector field parameterizations $\mathcal{V} = \{\boldsymbol{v_1}, ..., \boldsymbol{v_n}\}$ over unit time [28, 13]. In this framework, a convolutional neural network models a function $\mathcal{G}_\Psi(\mathcal{I}) = \mathcal{V}$, where Ψ are the model parameters, estimating $\mathcal{V}$ as a set of three-dimensional vector fields.

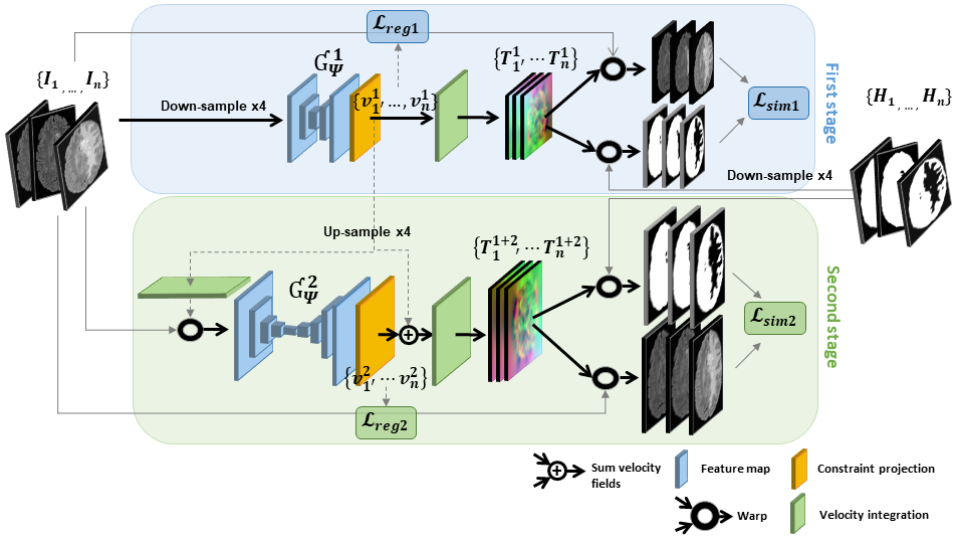

Figure 2: Schematic representation of the proposed framework. The first stage is trained at low resolution by taking as input down-sampled images in native space I_n. After training, the resulting stationary velocity fields v_n^1 are up-sampled to warp I_n and serve as an input to the second stage. Here, the residual displacement fields are obtained by summing up-sampled velocity fields v_n^1 and v_n^2 and trained on I_n instead of the warped images to reduce interpolation error. In the second stage, the learned parameters $\mathcal{G}_\Psi^1$ are fixed. During the optimization of both stages, the normal-appearance tissue masks H_n are warped and included in the loss functions.

2.1 Multi-resolution

We implemented the proposed method with a multi-stage and multi-resolution strategy. The lower amount of detail and relaxed regularization in the low-resolution stage may enable the registration of larger local mass-effects caused by gliomas. The high-resolution stage was expected to refine the deformations and register more detailed structures. A schematic representation is presented in Fig. 2. In the first stage, the larger brain structures were registered by training the proposed model with down-sampled input images. The images were stacked along the fourth dimension and passed to a neural network, in which the stationary velocity fields $\mathcal{V}^1$ were estimated. After integration over unit time, the resulting transformations $\mathcal{T}^1$ were used to warp the down-sampled input images, allowing the computation of a loss function to update the network parameters.

In the second stage, a second neural network learned the residual transformations at full resolution. After the training of the first stage, the transformations $\mathcal{T}^1$ were up-sampled and used to align the original images $\mathcal{I}$. These warped images served as input for the second stage. During its training, stationary velocity fields $\mathcal{V}^2$ were estimated. The two sets of stationary velocity fields were composed by summing the up-sampled $\mathcal{V}^1$ and $\mathcal{V}^2$. These were then integrated over unit time to obtain the final displacement fields $\mathcal{T}^{1+2}$. These final displacements transformed the input images $\mathcal{I}$ from their native space to their mean-space in one step. Thus, both stages were trained separately.

2.2 Centrality of deformations

To enforce the geometric centrality of the transformations, the average velocity field was subtracted from each estimated velocity field in $\mathcal{V}$ (Eq. 1), such that their sum was zero [13]. For this, an additional layer *constraint projection* was introduced between the output of the neural networks and the *velocity integration* layer (Fig. 2).

$$\hat{\boldsymbol{v}}_{\boldsymbol{i}} = \boldsymbol{v}_{\boldsymbol{i}} - \frac{1}{n}\sum_{j}^{n} \boldsymbol{v}_{\boldsymbol{j}} \quad ; \quad \sum_{i}^{n} \hat{\boldsymbol{v}}_i = 0 \tag{1}$$

2.3 Loss function

The framework was optimized with unsupervised loss (Eq. 3). This loss function consisted of a similarity metric $\mathcal{L}_{sim}$ that maximized the local cross-correlation [29] between the warped images $\mathcal{T} \circ \mathcal{I}$ and their average image $\bar{I} = \frac{1}{n}\sum_{n}^{i} T_i \circ I_i$. A regularization term $\mathcal{L}_{reg} = \frac{1}{n}\sum_{i}^{n} \|\nabla \mathbf{v_i}\|_2^2$ was introduced to encourage smooth and continuous transformations, penalizing high spatial gradients of $\mathcal{V}$ [13]. The influence of each term was balanced with a weight λ.

A loss-function masking strategy was implemented to compute the similarity only in regions of normal appearance across all images. The normal-appearance masks $\mathcal{H} = \{H_1, \ldots, H_n\}$ (brain mask minus tumor mask) of the input images were warped with the estimated transformations $\mathcal{T}$. These transformations should correct for mass-effects, and any residual misalignment of $\mathcal{T} \circ \mathcal{H}$ was asserted to be due to a non-correspondence. The intersection map S of all elements in $\mathcal{T} \circ \mathcal{H}$ was taken to exclude all tumor voxels from the loss functions (Fig. 2):

$$S = (T_1 \circ H_1) \cap (T_2 \circ H_2) \cap ... \cap (T_n \circ H_n) \tag{2}$$

$$Loss = \frac{1}{n}\sum_{i}^{n}((1 - S)(\mathcal{L}_{sim}(I_i, \bar{I})) + \lambda \cdot \frac{1}{n}\sum_{i}^{n} \mathcal{L}_{reg}(\hat{\boldsymbol{v}}_i) \tag{3}$$

3 Experiments

3.1 Dataset and data pre-processing

We used T2-weighted FLAIR MRI scans of 61 participants from the multi-center GLASS-NL study [30]. Participants were initially diagnosed with lower-grade (grade 2 or 3) IDH-mutant astrocytoma. All underwent 2 to 4 surgical resections, and adjuvant chemo- and radiotherapy. An average of 15.7 scans were available per subject, taken in an average time span of 7.93 years. Due to the clinical and multi-center nature of the data, many different acquisition parameters were used (i.e. 337 unique settings for the 572 images used, see example in Appendix A). Brain masks were obtained with HD-BET [31], tumors were automatically segmented using HD-GLIO [32, 33], and normal-appearing tissue segmentations were obtained with FAST as available in FSL [34]. FLAIR images were affinely aligned to the ICBM 2009a nonasymmetric atlas [35, 36] using Mutual Information [25]. Here, a loss-function masking approach was used to focus the registration on the normal-appearing tissue (i.e. brain minus tumor masks). Images were cropped to $176 \times 224 \times 176$ voxels, removing unnecessary background information while having sizes divisible by 2^4, where 4 is the downsampling factor applied in the neural networks (see subsection Implementation). Images were N4-bias field corrected with N4ITK [37], skull-stripped by setting to zero those voxels outside of the brain mask, and intensity standardized to zero mean and unit variance. All experiments in the presented study were performed with n=3 images per input. Therefore, for each subject, the available scans were grouped in all possible permutations of three images. Because we are focusing in the analysis of tumor growth, and not the brain changes due to surgery, images included in each set were taken in between resections. Moreover, those images acquired within the first 90 days after a surgery were excluded, as the brain is not yet settled and surgery-related edema might be present. The first and last images in a permutation were taken within a time interval of 2 months to 9 years. Fig. 3 shows the varying complexity of the dataset, quantified by a histogram of the maximum change in tumor-brain volume percentage in the permutation. The data was randomly split into 46:15 patients (3349:90 permutations) for training and testing.

3.2 Comparison methods

The performance of the proposed methods was evaluated against state-of-the-art conventional groupwise registration methods. These are implemented in the publicly available software packages Elastix [25], NiftyReg [26] and ANTs [27].

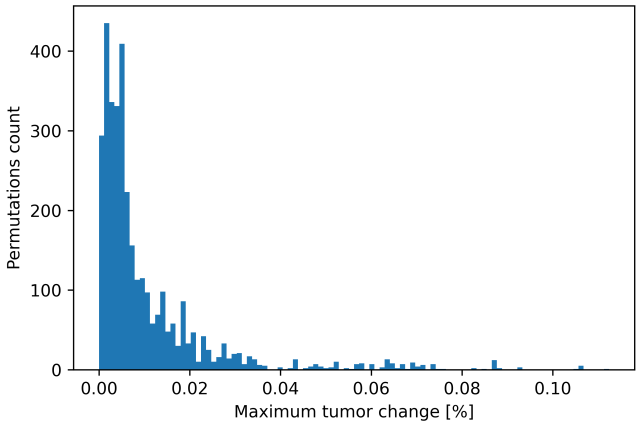

Figure 3: Histogram of the complexity of the dataset. Complexity is quantified as the maximum tumor change between two consecutive images in a permutation: $max\{abs(\%T_2 - \%T_1), abs(\%T_3 - \%T_2)\}$, where $\%T_n$ is the percentage of voxels classified as tumors in image n with respect to the number of voxels in the brain mask.

Elastix: The groupwise strategy presented by Huizinga et al. [9] registers all images in a set simultaneously. To ensure centrality, the resulting transformations are centered to zero mean. The proposed similarity metric is based on principal component analysis (PCA). This metric is a weighted sum of the eigenvalues obtained from the warped images. During registration, the total magnitude of the eigenvalues shifts and only a few contain most of the information shared across the images. The transformation model is free-form deformation with cubic B-spline: deformations are guided by a mesh of control points uniformly distributed on the image. B-spline curves define a continuous deformation field [38]. Parameters can be found in [1]. The method was implemented with Elastix v4.801.

NiftyReg: In the proposed groupwise strategy, the images of a set are initially registered to their first image in a pairwise manner. The average deformation is subtracted from the resulting deformations and applied to warp the input images. The average of the warped images is then computed and serves as a reference for the new iteration of pairwise registrations. This is repeated in a total of 10 iterations. The used similarity metric is mutual information, and registrations are performed with cubic B-splines. The optimization is regularized by a bending energy penalty term with a weight of 0.005. The used parameters and implementation are described in [2]. The jobs were computed using NiftyReg v1.5.58, submitting in parallel all the pairwise registrations.

ANTs: The images of a set are registered to the intensity average with a pairwise approach. After the images are warped, the intensity average image is again updated. The inverse of the obtained deformations are averaged and used to warp the latest average image toward the true mean-space [39]. This process is repeated in 4 iterations. The similarity metric guiding the registration is cross-correlation and the transformation model is a symmetric diffeomorphism. Deformation fields are smoothed with a Gaussian filter applying a variance of 3 voxels. The used script and parameters can be found in [3]. The groupwise registrations were run using ANTs v.2.3.5, computing in parallel all the pairwise registrations.

All conventional groupwise methods were evaluated on the purely non-linear registration of the test set. They were implemented with default parameters, except for the final grid-spacing of the control points in Elastix. Its parameter file was initially intended for larger structures (e.g. abdomen, heart) and so a final grid spacing of 5mm was found to yield a better performance. Elastix and NiftyReg allowed the input of masks and they were executed as such with the normal-appearance masks. All experiments were executed on a CPU cluster with: AMD Operon 6172, 2.1GHz (256GB RAM and 48 cores per node), and AMD Operon 6376, 2.3GHz (256GB RAM and 64 cores per node).

[1] `https://elastix.lumc.nl/modelzoo/par0039/`
[2] `http://cmictig.cs.ucl.ac.uk/wiki/index.php/Niftyreg_Groupwise`
[3] `https://github.com/ANTsX/ANTs/blob/master/Scripts/antsMultivariateTemplateConstruction2.sh`

3.3 Implementation

The proposed model was implemented both as a single stage and with the proposed multi-stage strategy. Experiments were conducted on Nvidia A40 48GB GPU and AMD EPYC 7742 CPU. Models were implemented in Python-3.7.4, using Keras-2.2.4 with Tensorflow-1.15.2 backend. All models were trained with the Adam optimizer with a learning rate of 1^{-3}, reducing it when the similarity metric of the validation set showed no significant improvement over 15 epochs. The order of the inputs was randomized and the batch size was set to 1. The hyperparameters of the loss function (Eq. 3) were empirically tuned based on the performance on the optimization dataset. This dataset consisted of a random subset of 600 permutations from the training dataset described in section 3.1. This subset was the same across all tuning experiments. For the proposed single-stage method, λ was set to 1.50. For the multi-stage approach, λ was set to 40 in the first stage, and 1.5 in the second stage (see Appendix B for the optimization of λ).

The convolutional neural network implemented in this study was a modified UNet [13]. Features of the concatenated input images were extracted with 3D convolutions with a kernel size of 3 and an output of k feature maps. For the one-stage-only implementation, we used k = [10, 20, 40, 80, 160, 80, 40, 20, 10]. Convolutions were followed by a leaky ReLu layer (a = 0.2). Four down-samplings were performed using max-pooling with a factor of 2 and four up-samplings with tri-linear interpolation.

For the first stage of the multi-stage setting the depth of the first UNet was reduced to two max-pooling and two up-sample layers, because the input images were already downsampled with a factor of four. The convolutional layers had output maps with feature numbers of k = [40, 80, 160, 80, 40]. The network of the second stage was equal to that of the one-resolution strategy.

3.4 Evaluation

The performance of the proposed method and conventional toolboxes were evaluated on the 90 permutations of the test set. The following metrics were computed to quantify the accuracy of the registration, the geometric centrality of the resulting deformations, and their smoothness. All metrics were computed within the normal-appearing volume across all images, i.e. the intersection of all the warped normal-appearing masks.

3.4.1 Dice coefficient

The anatomical correspondence between the registered images was assessed with the Dice coefficient (Eq. 4). It was computed for the normal-appearing tissue segmentation of cerebrospinal fluid (CSF), grey matter (GM), and white matter (WM), as well as for the tumor. Results for each permutation were obtained as the average Dice coefficient of all possible image pairs. Values by definition range between 0 (indicating poor alignment) and 1 (perfect overlap).

$$Dice(k) = \frac{1}{n}\sum_{i}^{n}\sum_{j}^{n}\frac{2\cdot|s_i^k \cap s_j^k|}{|s_i^k| + |s_j^k|}\ ,\ \forall i \neq j, \tag{4}$$

where n is the total number of images in the permutation. $\|s_i^k\|$ is the number of elements in warped segmentation of tissue-type k of the i^{th} image in the set.

3.4.2 Structural similarity index measure

The structural similarity index measure (SSIM) was computed to assess the perceptual similarity between the warped images [40]. To extend the metric to n images, the reported values were calculated as the average SSIM between every registered image I_i in the set and the average of the registered images $\bar{I}$ [41]. The metric is calculated in a sliding window w of size $7 \times 7 \times 7$ voxels:

$$SSIM = \frac{1}{n}\sum_{i}^{n}\frac{1}{X}\sum_{w}^{W}\frac{(2\mu_{I_{i,w}}\mu_{\bar{I}_w} + c_1) + (2\sigma_{I_{i,w},\bar{I}_w} + c_2)}{(\mu_{I_{i,w}}^2 + \mu_{\bar{I}_w}^2 + c_1)(\sigma_{I_{i,w}}^2 + \sigma_{\bar{I}_w}^2 + c_2)}, \tag{5}$$

where $\mu_{I_{i,w}}$ and $\mu_{\bar{I},w}$ are the mean intensity of the i^{th} image and the average image within window w. $\sigma_{I_{i,x}}^2$ and $\sigma_{\bar{I}}^2$ are their variances, and $\sigma_{I_{i,x},\bar{I}_x}$ their covariance. c_1 and c_2 are constants to ensure stability.

3.4.3 Centrality

The geometric centrality of the different methods was evaluated by averaging the obtained stationary velocity fields $\mathcal{V}$ in a permutation. If the resulting transformations meet in the mean-space, this average field should have a magnitude of zero everywhere. To quantify the centrality, the L2-norm of the average field is computed per voxel and averaged [42]:

$$Centrality = \frac{1}{X} \sum_{x} ||\bar{\mu}(x)||_2^2, \tag{6}$$

where $\bar{\mathbf{v}}(x)$ is the average stationary velocity field at each voxel x and X is the total number of voxels in the normal-appearing volume (i.e., S in Eq. 2).

3.4.4 Smoothness of transformations

The determinant of the Jacobian matrices $|J_{\mathcal{T}}|$ was computed [43] to capture the local properties of the deformation fields $\mathcal{T}$. Determinants of values larger than one characterize volume expansion in the vicinity of each voxel and values under one correspond to local compression. Negative values indicate foldings in the warped image and that the anatomical topology was not preserved [29]. Foldings were therefore quantified as the percentage of voxels x in the normal-appearing volume such that $|J_T(x)| \leq 0$. The smoothness of the deformations was further quantified by the standard deviations of their Jacobian determinant maps [9]. The mean value across all deformations in the set was reported:

$$SD|J_{\mathcal{T}}| = \frac{1}{n} \sum_{i}^{n} SD(|J_{T_i}|). \tag{7}$$

3.4.5 Statistical significance

To evaluate the statistical significance of different performances, the results of all permutations from the test set belonging to each subject were averaged. This resulted in 15 independent samples (i.e. 15 test subjects). The statistical significance was assessed with a Wilcoxon signed-rank test applying a significance level of $p \leq 0.05$. This threshold was adjusted per validation metric by the Bonferroni correction for multiple comparisons.

4 Results

The proposed method implemented with a single resolution (referred to as 'mask only') significantly improved the Dice coefficients of all brain structures between time points with respect to the affine-only registration (Fig. 4 left). The obtained Dice coefficients of the normal-appearing tissues were similar to those of the conventional methods. For the tumor segmentation, the proposed method yielded lower values. The obtained SSIM results (Fig. 4 right) had a higher mean than those achieved by the conventional methods, except for Elastix. The SSIM variance of all conventional methods was much larger, with some permutations obtaining results lower than with the affine registration only. The proposed multi-resolution strategy overall improved all these metrics with respect to the single-resolution implementation.

Elastix presented the best centrality, followed by our multi-resolution strategy (Table 1). In terms of the smoothness of the transformations, the proposed strategies showed a significantly lower percentage of folding. For the results of the proposed mask only, the foldings were consistently zero. In the multi-resolution strategy, two of the 90 permutations showed folding percentages in the range of 1^{-5}. The latter strategy also resulted in lowered standard deviation of the Jacobian maps with respect to all other methods. Particularly, Elastix and NiftyReg showed consistently less smooth deformations. The GPU implementation of our methods had inference runtimes of under a minute, significantly faster than the classical approaches, for which the longest average runtime was that of ANTs: 27.45 hours.

A qualitative example is presented in Fig. 5. Here, the affine registrations of the three time points of a permutation are presented next to the registration results of the conventional toolboxes and the proposed methods. An additional row depicts the average of the three warped images in the mean-space, and the overlap of their warped tumor segmentations. Here, Elastix and NiftyReg show more overlap of the segmentations across the images, which is not always a good sign. In the third time point (i.e., image I_3) the tumor has infiltrated into previously normal-appearing tissue, leading to a loss of anatomical

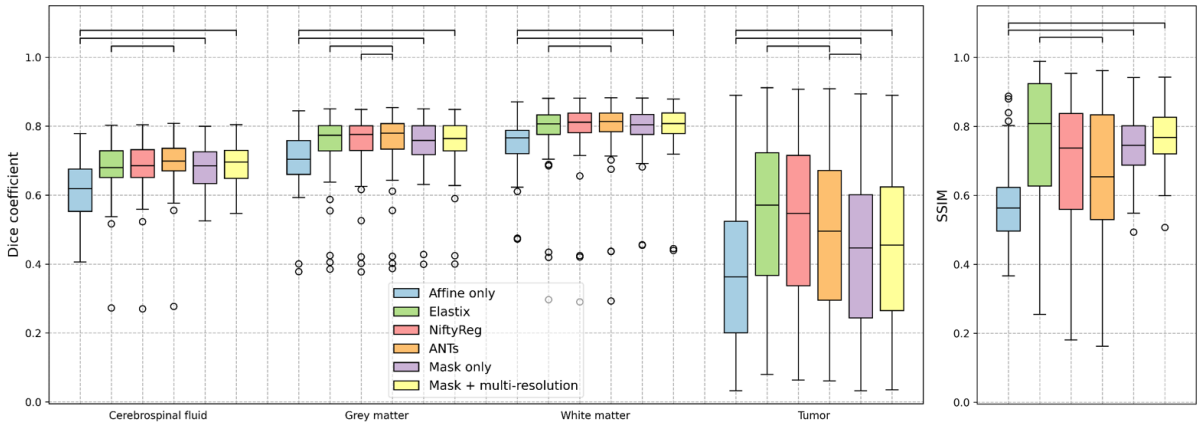

Figure 4: Registration accuracy for affine alignment only, Elastix, NiftyReg, ANTs, and our proposed method with and without multi-resolution implementation. Left: the boxplots of Dice coefficients for cerebrospinal fluid, grey matter, white matter, and tumor region. Right: the boxplots for average SSIM in normal-appearing tissue. The horizontal line within the boxplots represents the median value. Significant differences between two methods are indicated with bars over the plots, with significance threshold $p \leq 0.05/16$ (Bonferroni correction, 16 paired tests).

Table 1: Centrality and smoothness of the estimated deformations, and the required runtime in minutes using the methods Elastix, NiftyReg, ANTs, and the proposed framework with and without multi-resolution implementation. Results are computed within the normal appearance tissue and averaged over the test set. Significance reported for $p \leq 0.05/10$ (Bonferroni correction, 10 paired tests).

	↓ **Centrality**	↓ **Smoothness**		↓ **Run Time [min]**	
		$\lvert J_\mathcal{T}\rvert \leq 0$[%]	SD $\lvert J_\mathcal{T}\rvert$	GPU	CPU
Elastix	**1.008e-14 ****	6.665e-02	0.125	-	22.35
NiftyReg	4.828e-02	1.175e-02	0.159	-	32.58
ANTs	6.725e-02	3.161e-03	0.106	-	1647
Proposed (mask only)	1.033e-03	**0 ***	0.092	**0.029**	0.719 *
Proposed (mask + multi-resolution)	1.684e-03	5.999e-07 *	**0.085 ****	0.053	**0.114 ***

*Significant w.r.t classical methods
**Significant w.r.t all methods

correspondence between the images. In the case of Elastix, non-anatomically plausible deformations are seen near the tumor edge in I_3: the tumor has been compressed to match those of the previous images (see red arrow). Our proposed methods accurately register the normal-appearing tissue, but do not fully align the resection cavity (see blue arrows). The Jacobian maps corresponding to the results of these permutations are depicted in Fig. 6. The toolboxes Elastix and NiftyReg show much stronger deformations in the tumor and resection volumes, which is in line with the quantitative results in Table 1. The Jacobian maps of the proposed methods show more detailed registrations.

5 Discussion and conclusion

We presented a deep learning-based framework for the groupwise registration of longitudinal brain MRI with glioma. This method was able to register an image set to its mean-space despite the presence of non-correspondences by focusing on the normal-appearing tissue similarity. When implemented in a multi-resolution manner, registration accuracy improved compared to the single resolution registration. This potentially indicates that stronger mass-effects were better registered. The proposed framework achieved the Dice coefficients of CSF, GM, and WM comparable to those of state-of-the-art conventional toolboxes.

The Dice coefficient of the tumor segmentation showed large variances in the results of all methods (range: 0.05-0.90), which was mostly due to large fluctuations in the growth patterns of glioma. If the tumor is highly expansive, causing the surrounding tissue to deform from one time-point to another, the registration should be able to correct for such mass-effects. In this case, given the continuity and regularization of the deformation fields, we expect the registered tumor segmentations to have a high overlap. If on the contrary, the tumor largely infiltrates into previously normal-appearing tissue, a loss of

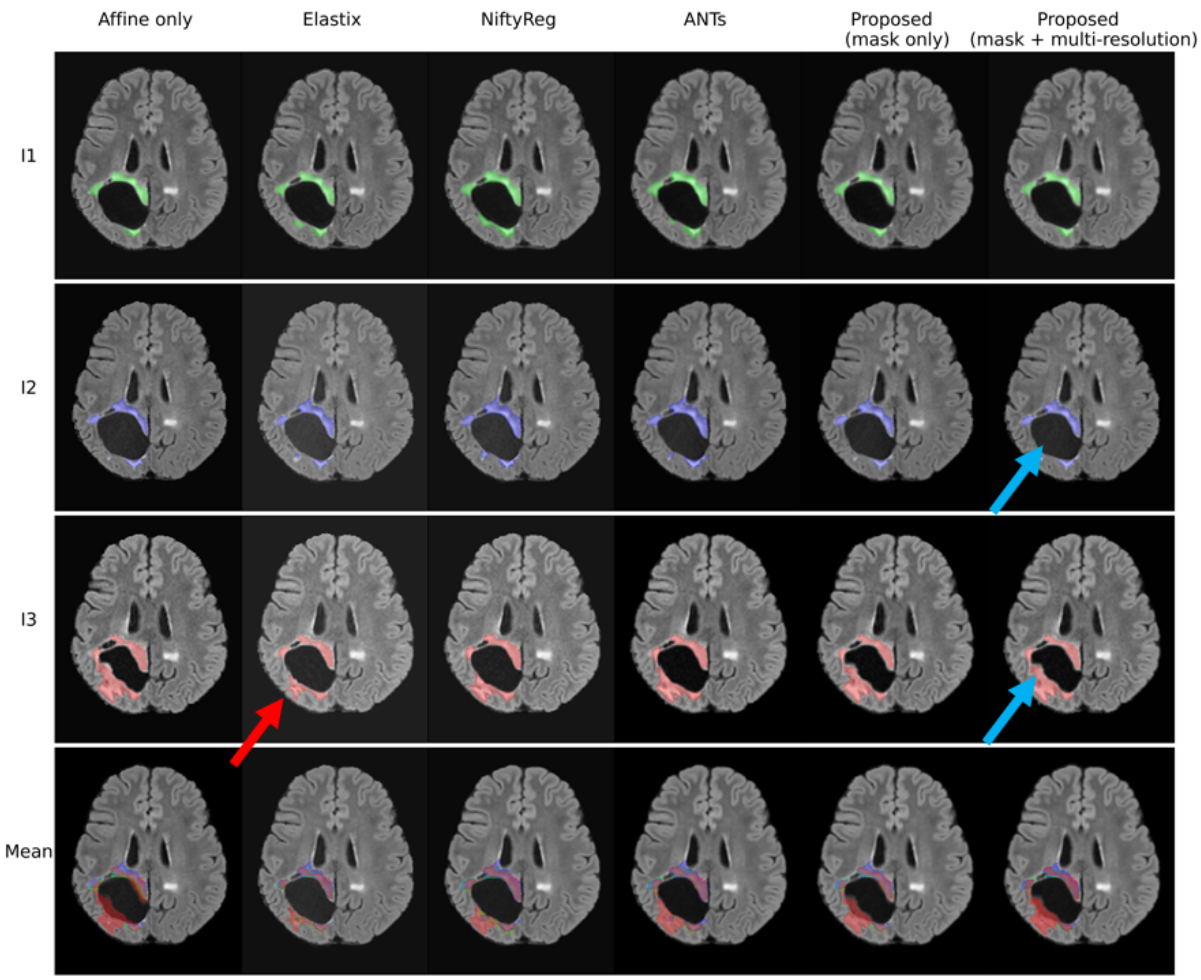

Figure 5: Results of one longitudinal permutation with images I1, I2, and I3 taken 3, 16, and 36 months after surgery. Overlaid on the axial slices the warped tumor segmentations. The last row shows the average image across the warped images and all tumor segmentations. Red arrow: excessive compression of the tumor. Blue arrows: resection cavity not aligned.

correspondence occurs. Since we focused the registration on the corresponding normal-appearing tissues rather than the tumor itself, correct registrations may not improve tumor overlap and can lead to lower Dice. This loss of correspondence also implies that the normal-appearing tissues cannot and should not obtain a perfect overlap. Another reason for the lowest tumor Dice values might be the volume size of the tumor. If their masks are small, their Dice coefficient is more susceptible to small misregistration. Therefore, this metric serves as an indicator of the registration performance but cannot be assessed without the context of glioma growth and the other metrics.

Elastix and NiftyReg showed the highest average tumor Dice. In qualitative results, we have observed that this high average was accompanied by stronger deformations around the tumor. When the tumors are mostly infiltrating this may lead to registrations that are not anatomically plausible (e.g. Fig. 5). The Jacobian maps of these methods showed strong local compression or expansion near the tumor, which was reflected in larger standard deviations and a larger number of foldings. These maps showed considerably less detailed structures and so we could expect less detailed registrations. Yet, the average values of SSIM of the aforementioned methods were high. This could reflect in a quantitative manner the over-registration of tumor infiltration: the large deformations that deform and smooth the surrounding tissue increased the visual similarity of the images, even if their registrations were incorrect. Interestingly, both methods were given normal-appearing masks (brain minus tumor masks) to ignore the tumors during the registration. Given the diffusive nature of glioma, automatic tumor segmentations were not always accurate. These conventional toolboxes may be more susceptible to the segmentation quality. Future

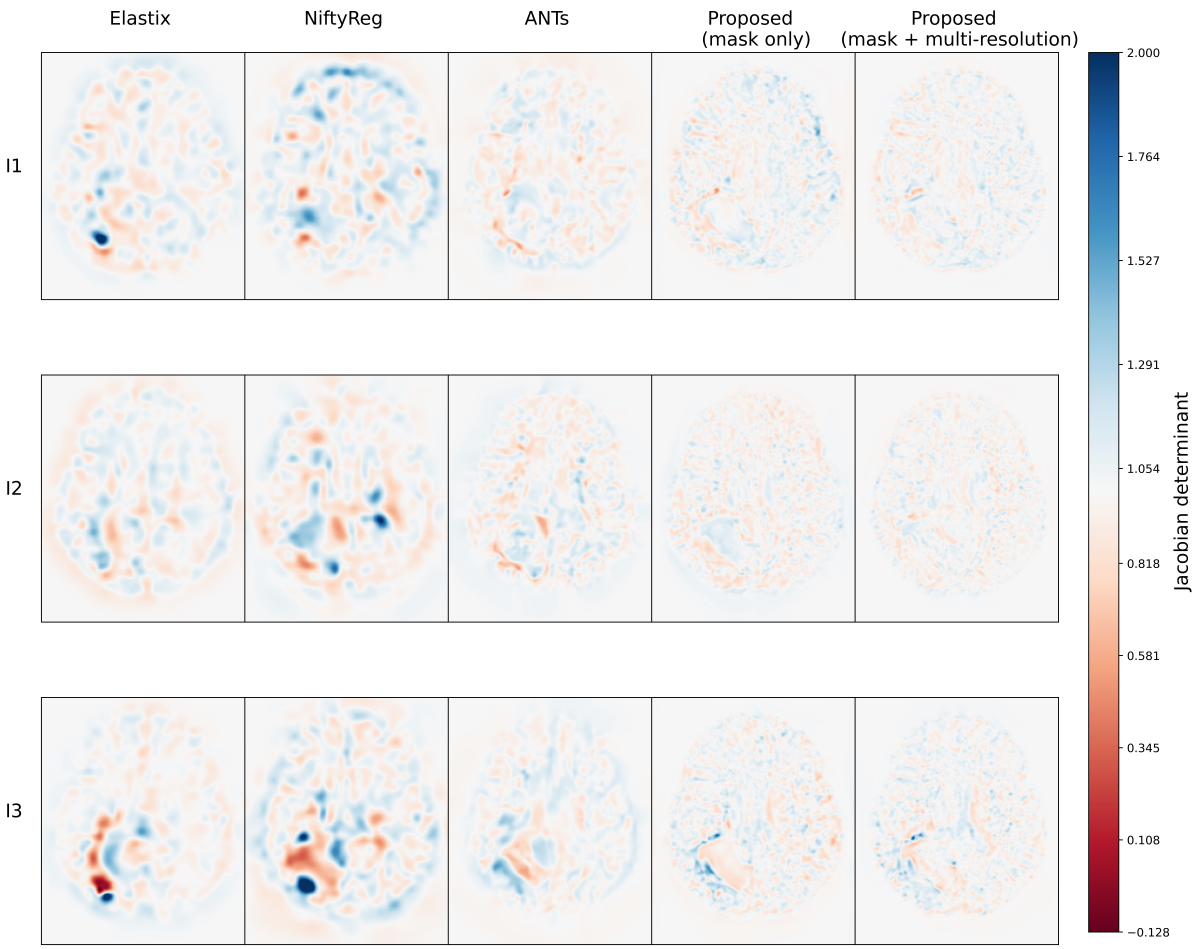

Figure 6: Example axial slices Jacobian maps, corresponding to the results observed in Figure 3. Expansions with respect to the mean-space are depicted in red while shrinking is in blue.

work could further analyze the effects of manual segmentations or expanding the volume of the currently available masks.

A peculiarity was found in the assessment of the Jacobian maps produced by our method. Given our cost-function masking strategy, we would expect the deformations within the tumor and the background voxels to be zero or a smooth continuation of the surrounding deformations. Although the deformations were more consistent within the tumor region, they were not completely smooth (see Fig. 7). This may be due to the variability of glioma: each patient presents a tumor in a different location of the brain and with varying appearance. The neural network learns these complex patterns over different subjects, thus introducing slight deformations within the tumors. Note however that we focus the registration on the corresponding normal-appearing tissue only. Therefore, we can disregard any small warpings within the tumor volume during the analysis of the deformations.

The proposed method achieved high SSIM, indicating that the registrations of the normal-appearing tissue were highly detailed, as was reflected in the Jacobian determinants maps. This is likely due to the diffeomorphic nature of the deformation fields, which specify voxelwise transformations. ANTs also used this transformation model, but applied a smoothing filter to the obtained fields, hence resulting in more consistent Jacobian maps and, possibly, the lower SSIM. Moreover, diffeomorphic transforms should by definition have no foldings in the registrations. The proposed model in one level of resolution obtained zero foldings, while only two permutations in the multi-stage approach had foldings. The larger amount in ANTs might be due to the numerical approximation of the integration of the velocity field over unit time. For the other two methods, the higher standard deviation and amount of folding are related to their B-spline transforms. A particular benefit of diffeomorphic deformations is that deformations are invertible. Therefore, if we want to analyze the tumor growth between any two points A and B in a permutation we can compute the required registration as the composition of the transformation of image A to the mean-space and the inverse field from mean-space to B: $T_{A\to B} = T_A \circ T_B^{-1}$.

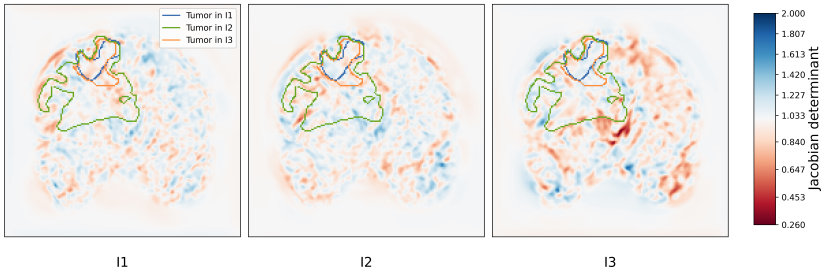

Figure 7: Resulting Jacobian determinant maps of one permutation with images I1, I2, and I3, taken 2, 11, and 17 months after surgery. The maps are not completely smooth within the area where all three tumors coincide.

Diffeomorphic transformations, however, come at a cost of higher run time. Our methods have the advantage of obtaining detailed and invertible transformations in a much faster inference time. The proposed model is able to achieve a performance similar to that of the conventional methods (in particular ANTs), while reducing the run time from tens of minutes, or even over a day, to just a fraction of a minute. This acceleration may make the model more attractive for further analysis of glioma growth.

5.1 Miss-registration of the resection cavity

A major difference between the proposed model and the conventional methods was observed in the example (Fig. 5, as our methods failed to align the resection cavity. Two possible reasons could explain this phenomenon:

Firstly, our method may have learned that the compression of the resection cavity was instead a tumor infiltration. Since there is no structure to register within the cavity, the model learned to not register this as a non-correspondence. To circumvent this, a new experiment was run including the CSF Dice coefficient in the loss function such that the registration would force the overlap of all CSF structures (e.g. resection cavities and ventricles). Multiple jobs were run with this new loss, but they did not result in an improvement in registration accuracy (Appendix C). A reason may be the quality of the CSF masks, as these contain narrower structures that can be highly affected by noise. Therefore, this interpretation of the resection cavity compression as an infiltration should be noted. While the resulting deformation fields can be used to quantify the mass-effects, the miss-alignment of the tumor masks, as seen in the bottom row of Fig. 5, represents tumor infiltration. The resection cavity compression would therefore be accounted for in this miss-alignment.

Secondly, the model may not have learned to perform large local registrations. Indeed when we apply a multi-resolution approach, we see an improvement in all Dice coefficients, including that of CSF and tumor. This strategy might be able to better capture the deformations caused by the tumor, but it should be further optimized to aim to correct for the largest mass-effects. Further experiments were run to analyze whether the regularization term was constraining large local transformations. We implemented again the described method in one stage only, but with no regularization of the deformation field within or near the tumor volume. The tumor masks were dilated by three voxels and warped to the mean-space. The loss term in Eq. 3 was then modified to only regularize the voxels outside of the overlap of these masks. However, no significant difference was found (see results in Appendix D). More likely, the distribution of the used dataset hindered the learning of large local deformations. As seen in Fig. 3, the distribution is shifted towards the small tumor changes. For reference, the qualitative example of Fig. 5 has a maximum tumor change of about 0.04%. Adding to the data complexity, the tumor may appear in different locations of the brain for each subject. The permutations of the 46 subjects included in the training set may be insufficient to capture the large variability and complexity of glioma location and growth.

5.2 Limitations and future work

A limitation of this project is the lack of parameter optimization of the conventional methods on our dataset. Further tuning of the methods might lead to better performances. Reducing the final grid-spacing in Elastix and NiftyReg, as well as decreasing the weight of the regularization in ANTs, might

result in more detailed transforms and higher SSIM. However, allowing less regularized transforms might also lead to the registration of noise and artifacts. This could cause worse smoothness of the Jacobian maps and possibly lower Dice coefficients. Therefore, further optimization should balance this trade-off.

The proposed framework achieved comparable performance to that of the state-of-the-art conventional methods. Yet, we have seen that the method may not sufficiently register large local mass-effects. When implemented in two stages, all Dice coefficients improved, likely due to the relaxed regularization in the lower resolution. Further optimization of this strategy might further improve the registrations. The model could be modified to include more levels of resolution (e.g. downsampling factors of two and eight). Moreover, as previously discussed, the complexity of the employed dataset is not uniform. We included a relatively high number of subjects with small tumor volume change, and they are more likely to be accompanied by more subtle mass-effects. Therefore, the model might not be able to learn large deformations. Future work could include more subjects in the dataset or compensate for the non-uniform distribution with a weighting of the permutation's influence on the training, giving more weight to the more complex cases (e.g., focal loss).

Lastly, the optimization of the model relies on tumor segmentations. Alternatively, an interesting potential for the proposed framework is the joint registration and detection of non-correspondences. Registration and segmentation are complementary tasks, as the segmentation is needed to know which volumes to register, while the non-correspondences after registrations define the segmentation. In our current implementation, an additional neural network could be jointly optimized to learn the tumor segmentations, which would then be used to mask the loss function of the registration network proposed here. This would therefore circumvent the need for tumor segmentations.

5.3 Conclusion

The proposed deep learning-based unbiased group-wise registration algorithm is able to register a set of longitudinal images to their mean-space despite the presence of non-correspondences. Results showed an accuracy similar to that of state-of-the-art conventional methods while obtaining more detailed registrations and significantly reducing the required run time. This method can therefore serve as an alternative to existing classical toolboxes for the analysis of glioma growth in longitudinal brain MRI.

References

[1] Elizabeth B. Claus, Kyle M. Walsh, John K. Wiencke, Annette M. Molinaro, Joseph L. Wiemels, Joellen M. Schildkraut, Melissa L. Bondy, Mitchel Berger, Robert Jenkins, and Margaret Wrensch. Survival and low-grade glioma: the emergence of genetic information. *Neurosurgical Focus*, 38(1):E6, 1 2015.

[2] Jianfeng Liang, Xiaomin Lv, Changyu Lu, Xun Ye, Xiaolin Chen, Jia Fu, Chenghua Luo, and Yuanli Zhao. Prognostic factors of patients with Gliomas – an analysis on 335 patients with Glioblastoma and other forms of Gliomas. *BMC Cancer*, 20(1):35, 12 2020.

[3] Chubei Teng, Yongwei Zhu, Yueshuo Li, Luohuan Dai, Zhouyang Pan, Siyi Wanggou, and Xuejun Li. Recurrence- and Malignant Progression-Associated Biomarkers in Low-Grade Gliomas and Their Roles in Immunotherapy. *Frontiers in Immunology*, 13, 5 2022.

[4] Fabio Raman, Elizabeth Scribner, Olivier Saut, Cornelia Wenger, Thierry Colin, and Hassan M. Fathallah-Shaykh. Computational Trials: Unraveling Motility Phenotypes, Progression Patterns, and Treatment Options for Glioblastoma Multiforme. *PLOS ONE*, 11(1):e0146617, 1 2016.

[5] Prateek Prasanna, Jhimli Mitra, Niha Beig, Ameya Nayate, Jay Patel, Soumya Ghose, Rajat Thawani, Sasan Partovi, Anant Madabhushi, and Pallavi Tiwari. Mass Effect Deformation Heterogeneity (MEDH) on Gadolinium-contrast T1-weighted MRI is associated with decreased survival in patients with right cerebral hemisphere Glioblastoma: A feasibility study. *Scientific Reports*, 9(1):1145, 12 2019.

[6] Daniel J Brat and Erwin G Van Meir. Vaso-occlusive and prothrombotic mechanisms associated with tumor hypoxia, necrosis, and accelerated growth in glioblastoma. *Laboratory Investigation*, 84(4):397–405, 4 2004.

[7] Martin Reuter, Nicholas J Schmansky, H Diana Rosas, and Bruce Fischl. Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage*, 61(4):1402–18, 7 2012.

[8] S. Joshi, Brad Davis, Matthieu Jomier, and Guido Gerig. Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage*, 23:S151–S160, 1 2004.

[9] Wyke Huizinga, Dirk HJ Poot, J-M Guyader, Remy Klaassen, Bram F Coolen, Matthijs van Kranenburg, RJM Van Geuns, André Uitterdijk, Mathias Polfliet, Jef Vandemeulebroucke, et al. Pca-based groupwise image registration for quantitative mri. *Medical image analysis*, 29:65–78, 2016.

[10] Arno Klein, Jesper Andersson, Babak A. Ardekani, John Ashburner, Brian Avants, Ming-Chang Chiang, Gary E. Christensen, D. Louis Collins, James Gee, Pierre Hellier, Joo Hyun Song, Mark Jenkinson, Claude Lepage, Daniel Rueckert, Paul Thompson, Tom Vercauteren, Roger P. Woods, J. John Mann, and Ramin V. Parsey. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage*, 46(3):786–802, 7 2009.

[11] Tongtong Che, Yuanjie Zheng, Xiaodan Sui, Yanyun Jiang, Jinyu Cong, Wanzhen Jiao, and Bojun Zhao. DGR-Net: Deep Groupwise Registration of Multispectral Images. pages 706–717. 2019.

[12] Yunlu Zhang, Xue Wu, H Michael Gach, Harold Li, and Deshan Yang. GroupRegNet: a groupwise one-shot deep learning-based 4D image registration method. *Physics in Medicine & Biology*, 66(4):045030, 2 2021.

[13] Bo Li, Wiro J Niessen, Stefan Klein, M Arfan Ikram, Meike W Vernooij, and Esther E Bron. Learning unbiased group-wise registration (lugr) and joint segmentation: evaluation on longitudinal diffusion mri. In *Medical Imaging 2021: Image Processing*, volume 11596, pages 136–144. SPIE, 2021.

[14] Julia Andresen, Timo Kepp, Jan Ehrhardt, Claus von der Burchard, Johann Roider, and Heinz Handels. Deep learning-based simultaneous registration and unsupervised non-correspondence segmentation of medical images with pathologies. *International Journal of Computer Assisted Radiology and Surgery*, 17(4):699–710, 4 2022.

[15] Stefan Bauer, Roland Wiest, Lutz-P Nolte, and Mauricio Reyes. A survey of MRI-based medical image analysis for brain tumor studies. *Physics in Medicine and Biology*, 58(13):R97–R129, 7 2013.

[16] Xiaoxiao Liu, Marc Niethammer, Roland Kwitt, Matthew McCormick, and Stephen Aylward. Low-rank to the rescue - atlas-based analyses in the presence of pathologies. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 17(Pt 3):97–104, 2014.

[17] Xiaoxiao Liu, Marc Niethammer, Roland Kwitt, Nikhil Singh, Matt McCormick, and Stephen Aylward. Low-Rank Atlas Image Analyses in the Presence of Pathologies. *IEEE transactions on medical imaging*, 34(12):2583–91, 12 2015.

[18] Zhenyu Tang, Yihong Wu, and Yong Fan. Groupwise registration of MR brain images with tumors. *Physics in Medicine & Biology*, 62(17):6853–6868, 8 2017.

[19] Xiao Yang, Xu Han, Eunbyung Park, Stephen Aylward, Roland Kwitt, and Marc Niethammer. Registration of Pathological Images. pages 97–107. 2016.

[20] A MOHAMED, E ZACHARAKI, D SHEN, and C DAVATZIKOS. Deformable registration of brain tumor images via a statistical model of tumor-induced deformation. *Medical Image Analysis*, 10(5):752–763, 10 2006.

[21] Ali Gooya, George Biros, and Christos Davatzikos. Deformable Registration of Glioma Images Using EM Algorithm and Diffusion Reaction Modeling. *IEEE Transactions on Medical Imaging*, 30(2):375–390, 2 2011.

[22] Cosmina Hogea, Christos Davatzikos, and George Biros. Brain–Tumor Interaction Biophysical Models for Medical Image Registration. *SIAM Journal on Scientific Computing*, 30(6):3050–3072, 1 2008.

[23] Matthew Brett, Alexander P. Leff, Chris Rorden, and John Ashburner. Spatial Normalization of Brain Images with Focal Lesions Using Cost Function Masking. *NeuroImage*, 14(2):486–500, 8 2001.

[24] Marc Niethammer, Gabriel L Hart, Danielle F Pace, Paul M Vespa, Andrei Irimia, John D Van Horn, and Stephen R Aylward. Geometric metamorphosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 639–646. Springer, 2011.

[25] Stefan Klein, Marius Staring, Keelin Murphy, Max A Viergever, and Josien PW Pluim. Elastix: a toolbox for intensity-based medical image registration. *IEEE transactions on medical imaging*, 29(1):196–205, 2009.

[26] Marc Modat, Gerard R Ridgway, Zeike A Taylor, Manja Lehmann, Josephine Barnes, David J Hawkes, Nick C Fox, and Sébastien Ourselin. Fast free-form deformation using graphics processing units. *Computer methods and programs in biomedicine*, 98(3):278–284, 2010.

[27] Brian B Avants, Nicholas J Tustison, Gang Song, Philip A Cook, Arno Klein, and James C Gee. A reproducible evaluation of ants similarity metric performance in brain image registration. *Neuroimage*, 54(3):2033–2044, 2011.

[28] John Ashburner. A fast diffeomorphic image registration algorithm. *Neuroimage*, 38(1):95–113, 2007.

[29] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019.

[30] Wies Rijan Vallentgoed, Anneke Niers, Karin van Garderen, Martin van den Bent, Erik van Dijk, Kaspar Draaisma, Paul van Eijk, Iris de Heer, Mathilde Kouwenhoven, Johan Kros, et al. Methylation analysis of matched primary and recurrent idh-mutant astrocytoma; an update from the glass-nl consortium. *Cancer Research*, 82(12_Supplement):4020–4020, 2022.

[31] Fabian Isensee, Marianne Schell, Irada Pflueger, Gianluca Brugnara, David Bonekamp, Ulf Neuberger, Antje Wick, Heinz-Peter Schlemmer, Sabine Heiland, Wolfgang Wick, et al. Automated brain extraction of multisequence mri using artificial neural networks. *Human brain mapping*, 40(17):4952–4964, 2019.

[32] Philipp Kickingereder, Fabian Isensee, Irada Tursunova, Jens Petersen, Ulf Neuberger, David Bonekamp, Gianluca Brugnara, Marianne Schell, Tobias Kessler, Martha Foltyn, et al. Automated quantitative tumour response assessment of mri in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *The Lancet Oncology*, 20(5):728–740, 2019.

[33] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[34] Yongyue Zhang, Michael Brady, and Stephen Smith. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57, 2001.

[35] Vladimir S Fonov, Alan C Evans, Robert C McKinstry, C Robert Almli, and DL Collins. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, (47):S102, 2009.

[36] Vladimir Fonov, Alan C Evans, Kelly Botteron, C Robert Almli, Robert C McKinstry, D Louis Collins, Brain Development Cooperative Group, et al. Unbiased average age-appropriate atlases for pediatric studies. *Neuroimage*, 54(1):313–327, 2011.

[37] Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4itk: improved n3 bias correction. *IEEE transactions on medical imaging*, 29(6):1310–1320, 2010.

[38] Kanwal K Bhatia, Joseph V Hajnal, Basant K Puri, A David Edwards, and Daniel Rueckert. Consistent groupwise non-rigid registration for atlas construction. In *2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821)*, pages 908–911. IEEE, 2004.

[39] Brian B Avants, Paul Yushkevich, John Pluta, David Minkoff, Marc Korczykowski, John Detre, and James C Gee. The optimal template effect in hippocampus studies of diseased populations. *Neuroimage*, 49(3):2457–2466, 2010.

[40] Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014.

[41] Nefeli Lamprinou, Nikolaos Nikolikos, and Emmanouil Z Psarakis. Groupwise image alignment via self quotient images. *Sensors*, 20(8):2325, 2020.

[42] Adrian Dalca, Marianne Rakic, John Guttag, and Mert Sabuncu. Learning conditional deformable templates with convolutional networks. *Advances in neural information processing systems*, 32, 2019.

[43] Ziv Yaniv, Bradley C Lowekamp, Hans J Johnson, and Richard Beare. Simpleitk image-analysis notebooks: a collaborative environment for education and reproducible research. *Journal of digital imaging*, 31(3):290–303, 2018.

[44] Jan Unkelbach, Bjoern H Menze, Ender Konukoglu, Florian Dittmann, Matthieu Le, Nicholas Ayache, and Helen A Shih. Radiotherapy planning for glioblastoma based on a tumor growth model: improving target volume delineation. *Physics in Medicine & Biology*, 59(3):747, 2014.

A Acquisition parameters of FLAIR images

The data was collected in three different medical centers and each patient underwent multiple scans with an average time span of 7.93 years. Due to the clinical nature of the data, many different sets of acquisition parameters were used. In the following image, the acquisition parameters of images of three random subjects are presented to show the parameter variety.

SUBJECT	Image	Inversion Time	Field Strength	Slice thickness	Repetition Time	Echo Time	Pixel Spacing x	Pixel Spacing y
A	I1	1987.0	1.5	1.3	6500.0	117.831	1.2109	1.2109
A	I2	1900.0	3.0	1.2	5600.0	451.0	0.5804	0.5804
A	I3	1900.0	3.0	1.2	5600.0	451.0	0.5804	0.5804
A	I4	2337.0	3.0	1.2	8000.0	131.087	0.9766	0.9766
A	I5	2337.0	3.0	1.2	8000.0	131.087	0.9766	0.9766
A	I6	1900.0	3.0	1.2	5600.0	451.0	0.5804	0.5804
A	I7	2337.0	3.0	1.2	8000.0	131.087	0.9766	0.9766
A	I8	2336.0	3.0	1.2	8000.0	131.618	0.9766	0.9766
A	I9	2337.0	3.0	1.2	8002.0	131.265	0.9766	0.9766
A	I10	2400.0	1.5	1.0	7600.0	431.0	0.4883	0.4883
B	I1	2340.0	3.0	1.2	8000.0	129.489	0.9766	0.9766
B	I2	1900.0	3.0	1.2	5600.0	451.0	0.5804	0.5804
B	I3	1650.0	3.0	1.12	4800.0	278.764	1.04166662693023	1.04166662693023
B	I4	1900.0	3.0	1.2	5600.0	451.0	0.5804	0.5804
B	I5	1650.0	3.0	1.12	4800.0	278.764	1.04166662693023	1.04166662693023
B	I6	2343.0	3.0	1.2	8000.0	128.062	0.9766	0.9766
B	I7	2342.0	3.0	1.2	8000.0	128.776	0.9766	0.9766
B	I8	2400.0	1.5	1.0	7600.0	431.0	0.4883	0.4883
B	I9	2400.0	3.0	1.0	7700.0	430.0	0.4883	0.4883
C	I1	2372.0	1.5	5.0	8000.0	82.0	0.359375	0.359375
C	I2	2370.0	1.5	4.0	8000.0	82.0	0.44921875	0.44921875
C	I3	2216.3	1.5	5.0	7000.0	82.0	0.359375	0.359375
C	I4	1900.0	3.0	1.2	5600.0	451.0	0.5804	0.5804
C	I5	2337.0	3.0	1.2	8000.0	131.087	0.9766	0.9766
C	I6	1650.0	3.0	1.12	4800.0	278.764	1.04166662693023	1.04166662693023

Figure 8: Acquisition parameters of the FLAIR images of subjects A, B, and C.

B Hyperparameter tuning

The weight λ of the regularization term in the loss function (Eq. 3) was tuned based on the performance of the optimization set. In Fig. 9, the average results of each evaluation metric are presented. The results of all Dice coefficients seem to be robust to the value of λ, with changes in the power of $1e-3$. Both the Dice coefficients of grey matter and white matter follow similar trends, peaking around λ=1.5. The plots of Dice of the cerebrospinal fluid and of the tumor decay with higher λ, with a smaller decrease between 1 and 1.5. Both these tissues have smaller volumes compared to GM and WM and thus their Dice coefficient is more susceptible to noise. Values of λ under 1 may indicate overfitting of the registration, i.e. the framework is registering noise or artifacts in the images. The plateau-like shape around 1 and 1.5 may then indicate an inflection point between the overfitting and insufficient registrations due to too strong regularization.

The SSIM decreased rather constantly as the regularization becomes stronger and fewer details in the images are registered. The standard deviation of the Jacobian, its negative values, and the centrality decay exponentially. The stronger regularization leads to smoother and less complex deformation fields that are better able to meet the condition of geometric centrality.

Given all these performances, the weight of λ=1.5 was chosen. Similar trends were observed for the multi-resolution strategy, where the loss function of the first stage was given λ=40 and the second stage again 1.5.

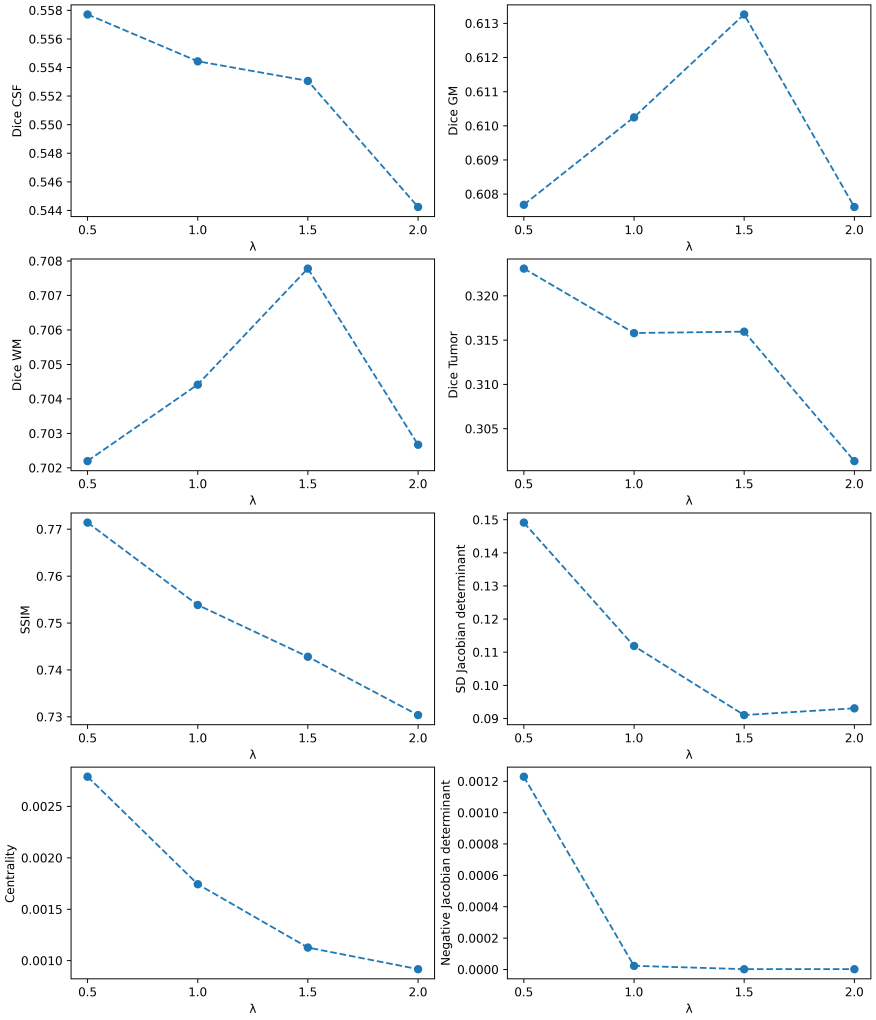

Figure 9: Performance of the proposed strategy in one resolution with varying weights λ of the regularization term (Eq. 3). Values averaged over the test set.

C New loss with dice CSf

Gliomas infiltrate into the white matter much faster than into the grey matter, while the cerebrospinal fluid (CSF), except for rare cases of ventricle seeding, acts as a barrier to migrating cells [44]. Therefore, CSF tissue should not lose correspondence between images. Based on this, the Dice coefficient of the warped CSF segmentation was included in the previous loss function (Eq. 3) such that the registration forces the overlap of all structures classified as CSF (e.g. resection cavities and ventricles).

$$Loss = \frac{1}{n}\sum_{i}^{n}((1-S)(\mathcal{L}_{sim}(I_i, \bar{I})) + \lambda_1 \cdot \frac{1}{n}\sum_{i}^{n}\mathcal{L}_{reg}(\hat{\boldsymbol{v}}_i) + \lambda_2 \cdot Dice(CSF), \tag{8}$$

where the last term corresponds to the average CSF Dice coefficient between all image pairs (see subsection Evaluation).

Multiple jobs were run with this new loss, but they did not result in an improvement in performance in any metric. A reason may be the quality of the CSF masks, as these contain narrow structures and can be highly affected by noise.

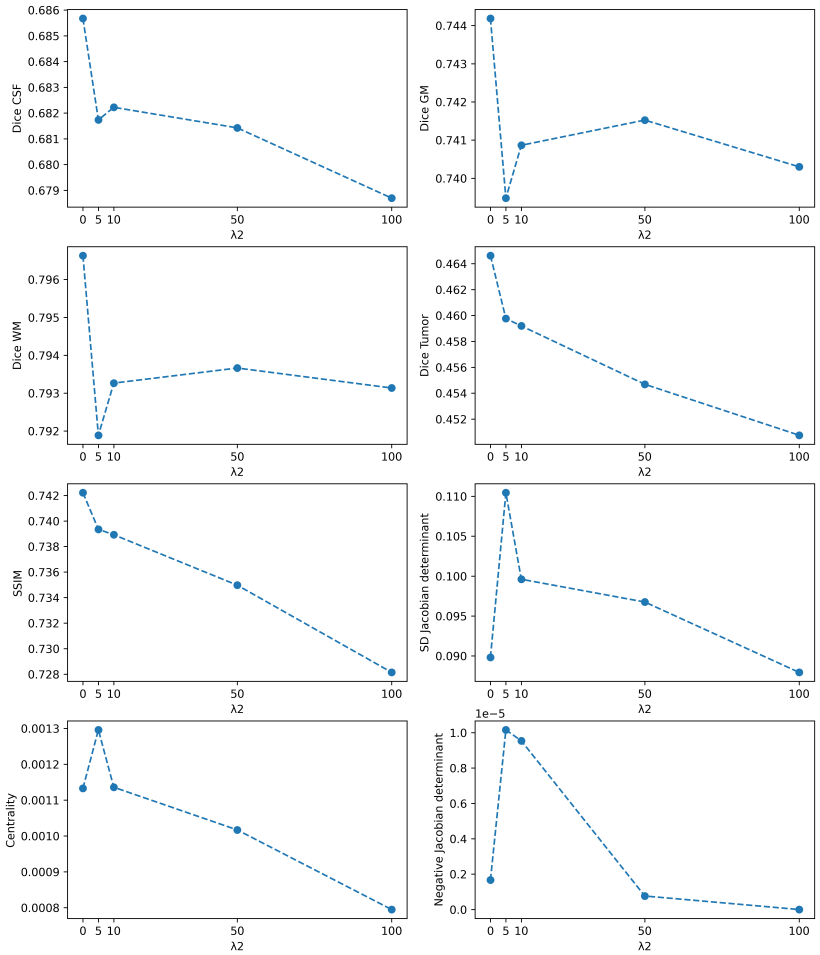

Figure 10: Performance of the proposed strategy in one resolution with the new loss function (Eq. 8). Values averaged over the optimization set.

D No regularization within tumor

To analyze whether the required larger deformations near the tumor were hindered by the regularization term, we trained again the single-stage method but relaxed the regularization term in and near the tumor. The tumor masks were dilated by three voxels and the intersection of the warped masks was calculated (similar to Eq. 2). The regularization term was then calculated only over those voxels outside of this intersection (i.e. the corresponding normal-appearing tissue, minus the dilation of the tumor). The results of the single-stage method with the loss function in Eq. 3, and of the single-stage method with this new loss function are presented in Fig. 11. The performance did not change significantly over any of the metrics.

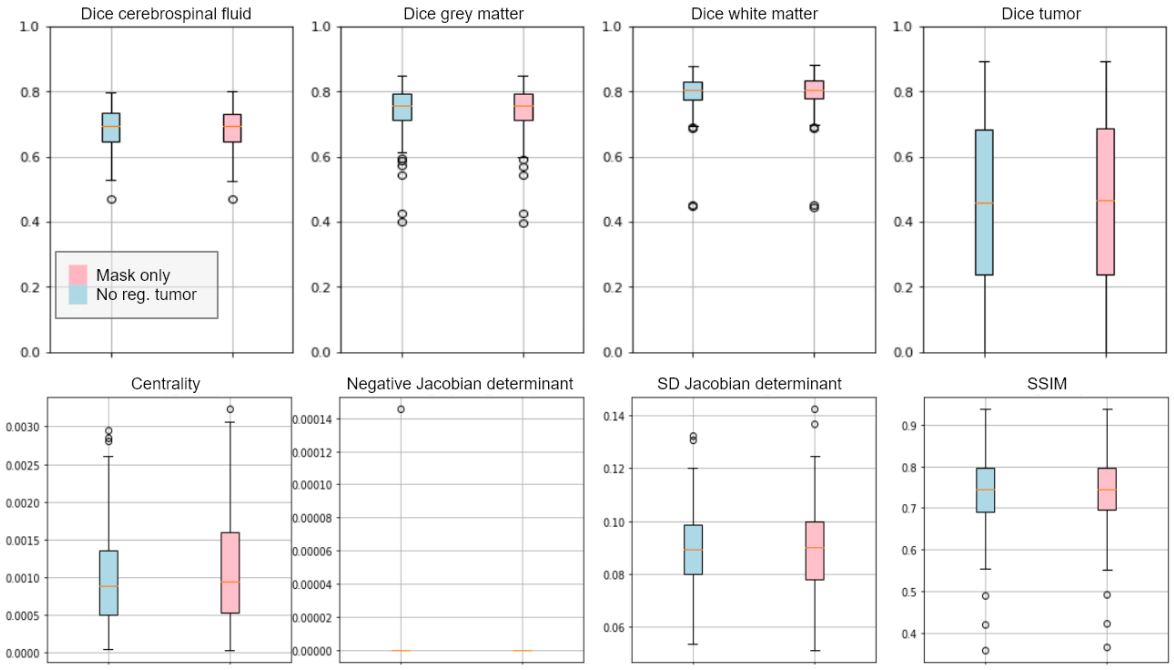

Figure 11: Performance of the proposed strategy in one resolution with no regularization of the deformation fields within and near tumor volume. Values averaged over the optimization set.