A Deep Learning Based Framework for Joint Image Registration and Segmentation of Brain Metastases on Magnetic Resonance Imaging

Jay Patel[1] JBPATEL@ALUM.MIT.EDU
Syed Rakin Ahmed[1,2,3] SYEDRAKIN_AHMED@FAS.HARVARD.EDU
Ken Chang[4] CHANGKEN1@GMAIL.COM
Praveer Singh[5] PRAVEER.SINGH@CUANSCHUTZ.EDU
Mishka Gidwani[1] MXG393@CASE.EDU
Katharina Hoebel[1] KHOEBEL@MGH.HARVARD.EDU
Albert Kim[1,6] AKIM46@MGH.HARVARD.EDU
Christopher Bridge[1,7] CBRIDGE@MGH.HARVARD.EDU
Chung-Jen Teng[8,9] CJTENG.TW@GMAIL.COM
Xiaomei Li[10] SDULIXIAOMEI@163.COM
Gongwen Xu[10] XUGONGWEN@163.COM
Megan McDonald[10] MMCDON19@BU.EDU
Ayal Aizer[11] AYAL_AIZER@DFCI.HARVARD.EDU
Wenya Linda Bi[12] WBI@BWH.HARVARD.EDU
Ina Ly[1,6] ILY@MGB.ORG
Bruce Rosen[1] BRROSEN@MGH.HARVARD.EDU
Priscilla Brastianos[6] PBRASTIANOS@MGH.HARVARD.EDU
Raymond Huang[10] RYHUANG@BWH.HARVARD.EDU
Elizabeth Gerstner[1,6] EGERSTNER@MGH.HARVARD.EDU
Jayashree Kalpathy-Cramer[5] JKALPATHY-CRAMER@MGH.HARVARD.EDU

[1] *Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Boston, MA, USA*
[2] *Harvard Graduate Program in Biophysics, Harvard Medical School, Cambridge, MA, USA*
[3] *Geisel School of Medicine at Dartmouth, Dartmouth College, Hanover, NH, USA*
[4] *Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA*
[5] *Department of Ophthalmology, University of Colorado School of Medicine, Aurora, CO, USA*
[6] *Division of Neuro-Oncology, Massachusetts General Hospital, Boston, MA, USA*
[7] *MGH and BWH Center for Clinical Data Science, Massachusetts General Hospital, Boston, MA, USA*
[8] *National Yang Ming Chiao Tung University School of Medicine, Taipei, Taiwan*
[9] *Far Eastern Memorial Hospital, New Taipei City, Taiwan*
[10] *Department of Radiology, Brigham and Women's Hospital, Boston, MA, USA*
[11] *Department of Radiation Oncology, Brigham and Women's Hospital, Dana-Farber Cancer Institute, Boston, MA, USA*
[12] *Department of Neurosurgery, Brigham and Women's Hospital, Dana-Farber Cancer Institute, Boston, MA, USA*

Abstract

Manual segmentation of brain metastases (BM) is a laborious and time-consuming task for expert clinicians, especially in the setting of longitudinal patient imaging. Although automated deep learning (DL) approaches can segment larger lesions effectively, they suffer from poor sensitivity of lesion detection for micro-metastases. Moreover, these approaches segment all patient imaging independently of each other, ignoring relevant information from prior time-points. In order to utilize prior time-point information, we propose SPIRS, a joint image registration and segmentation method. Given a prior time-point image and segmentation mask (which are readily available in a routine clinical environment), we affinely and deformably register these to a new time-point image. This warped prior image and mask are then used to enhance and improve the segmentation of the new time-point. We apply SPIRS to a large retrospectively acquired single institution dataset and show that it outperforms current registration approaches on BM imaging and that it significantly improves segmentation performance for micro-metastatic lesions.

1. Introduction

Brain metastases (BM) are the most common form of intracranial tumors in adults with an annual incidence of 170,000 in the United States (Tabouret et al. (2012)), which is expected to increase as systemic treatments for primary tumors improve (Fabi et al. (2011); Sperduto et al. (2020)). Given rising incidence and limited treatment, BM are an unmet need in modern oncology, with median survival post diagnosis of ranging from only 2.7 to 24 months (Li et al. (2023); Cagney et al. (2017)). Alongside monitoring changes in clinical metrics such as performance status and cognitive function, neuroradiologists can assess the efficacy of a given treatment regimen by tracking individual lesion sizes across T1-weighted contrast-enhanced (T1-CE) magnetic resonance (MR) imaging time-points, noting whether tumor burden is decreasing, stable, or increasing (Vogelbaum et al. (2022)). If BM enlarge or new BM appear over time, a different treatment option may be necessary. However, manual determination of tumor boundaries needed for lesion tracking can be challenging in the presence of heterogeneous contrast enhancement, diffuse tumor boundaries from surrounding edema, or blunted contrast relative to surrounding normal brain due to treatment effects. Moreover, patients can present with anywhere from a single lesion to upwards of one hundred lesions, varying in volume from as small as a few mm^3 to as large as $10000mm^3$. These lesions can exhibit varied shapes/structures (from spherical to highly irregular) and can be situated across every region of the brain parenchyma. In addition to being a highly time-consuming and costly task, manual segmentation is subject to significant inter- and intra-rater variability for the aforementioned factors (Growcott et al. (2020)). As such, there has been much interest in developing reproducible automated methods for segmentation.

While there is minimal work published in the automated segmentation of *metastatic* brain tumors, the segmentation of *primary* brain tumors is a well-researched field, mainly due to the availability of the large multi-institutional publicly available BraTS dataset (Baid et al. (2021); Menze et al. (2014); Bakas et al. (2017)). In recent years, 3D U-Net architectures (Ronneberger et al. (2015)) have consistently dominated the BraTS leaderboards and are the current state-of-the-art method for brain tumor segmentation (Patel et al. (2021); Isensee et al. (2020); Futrega et al. (2021)). Guided by these approaches, most researchers also choose to use 3D U-Nets for BM segmentation. Due to the fact that BM can vary in

size and number of lesions, *metastatic* brain tumor segmentation is a more difficult task than *primary* brain tumor segmentation. More specifically, the difficulty is attributed to the presence of micro-metastases, which are lesions with a diameter no greater than 5mm (Cheng et al. (2019); Nomoto et al. (1994)). These brain micro-metastases (especially dural lesions located at the peripheries of the brain) can share similarity in shape, size, and MR intensity to small blood vessels. In such cases, it can be challenging to confidently label a small focus of enhancement as a metastases or a blood vessel until it grows on a subsequent time-point. In other cases, micro-metastases can present with little to no contrast enhancement, especially in the setting of lower quality scans. These factors together make the automated detection and segmentation of micro-metastatic brain lesions a challenging machine learning problem.

Past approaches in published literature on the automated segmentation of BM have included a variety of architectural modifications and different loss functions, but they all report a sensitivity of detection of micro-metastases well under 50% (Grøvik et al. (2019); Ottesen et al. (2023); Rudie et al. (2021)). Studies that have looked at the inter-rater variability for detection of micro-metastases have concluded that current deep learning (DL) based approaches are inferior to that of expert neuroradiologists and might not be ready for clinical deployment just yet (Rudie et al. (2021)). That being said, even though current models are imperfect, especially for detection of small lesions, they have the potential to improve workflow efficiency for radiologists by reducing the amount of manual segmentation that must be done by a clinician. Specifically, a clinician can now simply correct a label map by adding in missed detections or removing false positives, a process that can save significant amounts of time relative to needing to segment the whole volume from scratch.

While the process of correcting mistakes is acceptable in the setting of single patient visits, it can become cumbersome and tedious to the clinician in the setting of longitudinal patient data. For instance, a metastasis that is missed in the baseline scan is likely to be missed again in all future time-points assuming it does not substantially change in size or appearance. Such a scenario requires the clinician to manually fix the same mistake repeatedly on all time-points, creating unnecessary annotation burden for the clinician. A better system would entail the clinician fixing the mistake only once on the first time-point, with the neural network carrying forward that prior information to subsequent time-points. To the best of our knowledge, no published work has assessed the utility of using longitudinal imaging data for the purpose of improving BM segmentation quality.

In this work, we propose a novel DL based approach to jointly register and segment BM on T1-CE MR imaging, a method we call Sequential and Pyramidal Image Registration and Segmentation (SPIRS). More specifically, given a prior time-point and a new time-point image, we train a *Siamese* style convolutional neural network (CNN) to first affinely (i.e. linearly) and then deformably (i.e. non-linearly) register the pair of images. This registration transform is parameterized as a dense displacement vector field (DVF), and it maps the offset from the prior time-point onto the new time-point image. Assuming we already have a prior time-point segmentation mask that has been manually edited by a clinician (which will be the case in a routine clinical environment), we can then use the found DVF to transform this prior segmentation mask onto the new time-point. This warped prior mask can then be used to enhance and improve the segmentation of the new time-point (figure 1).

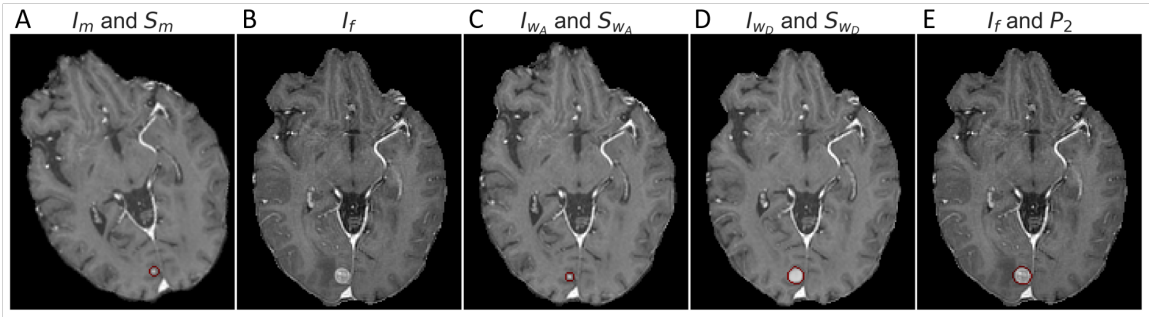

Figure 1: An example prior time-point image I_m and tumor mask S_m outlined in red (A) and new time-point image I_f (without tumor mask) (B). The priors are first affinely registered (creating I_{w_A} and S_{w_A}) (C) and then deformably registered (creating I_{w_D} and S_{w_D}) (D) to the new time-point. Note how the prior time-point tumor is warped to match the location and size of the new time-point tumor. These warped priors I_{w_D} and S_{w_D} are used to aid and improve the segmentation of the new time-point (E).

Generalizable Insights about Machine Learning in the Context of Healthcare

- **We show that SPIRS outperforms other methods for the specific task of registration of MR imaging with BM.** Many registration methods exist, but they are trained and validated on normal brain anatomy. Not only do we develop a novel architecture for combined affine and deformable registration, we show that by training a task specific model, we can improve performance over current baseline methods.
- **Our approach utilizes readily available prior time-point information in order to improve the segmentation of micro-metastases.** Despite the fact that longitudinal imaging data is readily available, most approaches segment scans independently of each other. Our method shows that using prior time-point information can significantly improve the detection rate of micro-metastases on follow-up imaging, which will reduce annotation burden for clinicians and improve performance of downstream clinical tasks.

2. Related Work

There is extensive literature in medical image registration and it can be broadly be split into two categories: non-learning based and learning-based. We provide a brief overview of some relevant methodology below.

2.1. Non-Learning Based Registration

Given a fixed and a moving image, classical registration approaches perform a gradient descent based numerical optimization to iteratively align pixels from the moving image

onto the fixed image to improve a chosen similarity metric (e.g. mean squared error (MSE), normalized cross correlation (NCC)). The learned transformation can be either linear or non-linear, depending on one's use case. If the transform is non-linear, certain constraints can be placed on the outputted DVF to encourage a spatially smooth transform. To alleviate this numerical optimization problem (which even for linear transforms can get stuck at poor local minimas for anatomically complex images), classical methods often employ a *sequential* and *pyramidal* hierarchy. *Sequential* refers to solving lower complexity transforms before higher complexity transforms. In other words, a purely affine transformation is computed first before solving for the deformable transformation. *Pyramidal* refers to a multi-scale approach wherein the transformation is first computed at a coarser image scale and is progressively updated at finer image scales (Adelson et al. (1984)). We note that due to the iterative nature of these classical algorithms, they can be quite computationally intensive. Indeed, deformable registration of 3D brain MR imaging can take upwards of one to two hours on CPU per image pair. While there are many classical registration algorithms for deformable registration, including but not limited to B-splines (Rueckert et al. (1999)), Demons (Thirion (1998)), and Large Diffeomorphic Distance Metric Mapping (LDDMM) (Glaunes et al. (2008)), the current gold standard is generally accepted to be Symmetric Normalization (SyN) (Avants et al. (2007)) from the Advanced Normalization Tools (ANTs) package (Avants et al. (2010); Klein et al. (2009)). ANTs can also be used for highly performant affine registration of imaging.

2.2. Learning Based Registration

Newer methods utilize neural networks to learn a function for (affine and/or deformable) registration. This can be advantageous because each image pair can be fully registered with one forward pass of the network, which will take only a few seconds on GPU. DL based affine registration networks are usually formulated as a supervised regression problem. Chee and Wu (2018) use a *Siamese* style encoder to directly predict the affine transform matrix. Islam et al. (2021) uses a similar regression approach, but focuses on cross-modality registration. DL based deformable registration networks can either be trained in a supervised or unsupervised manner. While earlier approaches like that from Sokooti et al. (2017) required ground truth DVFs to train the network, newer approaches tend to be fully unsupervised. Dalca et al. (2019) proposed a U-Net based diffeomorphic registration model they named VoxelMorph (VXM). Building off this approach, Mok and Chung (2020) utilized a pyramidal architecture to improve the quality of the registration. However, they did not incorporate feature sharing at the different levels of the pyramid, resulting in redundant parameters. de Vos et al. (2019) devised a network to sequentially perform affine and deformable registration, but their deformable registration was based only on b-spline grids. Christodoulidis et al. (2018) also performed both affine and deformable registration, however their approach was neither sequential nor pyramidal.

2.3. Joint Frameworks for Registration and Segmentation

While registration and segmentation are two of the largest and most researched areas of computer vision for medical applications, there is significantly less research in how the coupling of these two tasks may improve one or both tasks. Such joint methods are mainly used

in areas where longitudinal imaging data is widely available. For instance, segmentation of cardiac MR imaging is usually done only on end-diastolic and end-systolic frames, with information from other frames not exploited. By using a joint framework, more data may be incorporated during model training and can improve performance of both cardiac motion estimation and atrial/ventricular segmentation (Qin et al. (2018); Upendra et al. (2021)). Joint approaches have also been shown to be effective in low-annotation settings, where only a fraction of the whole dataset has ground truth segmentations (Xu and Niethammer (2019)). When compared to the baseline of not having any annotated data, weak supervision from a small sample of ground truth annotations can improve registration performance due to the incorporation of an anatomy similarity loss. Indeed, joint methodological approaches have been shown to outperform independently optimized networks on tasks such as cardiac, knee, and brain (Xu and Niethammer (2019); Chen et al. (2022)). While these existing works have applied their respective methods to normal anatomy, to the best of our knowledge no work has focused specifically on improving the registration quality and segmentation performance of BM on T1-CE MR.

3. Methods

We let I_f, I_m, S_f, S_m denote the fixed image, moving image, fixed segmentation, and moving segmentation, respectively. $\hat{T}_A$ and $\hat{T}_D$ denote the predicted affine and deformable registration transforms, respectively. Here, $\hat{T}_A$ and $\hat{T}_D$ are mappings such that $I_m \circ \hat{T} = I_w \approx I_f$. To warp image I_m with transformation $\hat{T}$, we use a fully differentiable spatial transformer module, which allows for gradient backpropagation during network optimization (Jaderberg et al. (2016)). Our proposed architecture consists of three successive blocks: 1) the affine registration network $\mathcal{F}_A$, 2) the deformable registration network $\mathcal{F}_D$, and 3) the segmentation network $\mathcal{F}_S$. A diagram showing this sequential structure is visualized in figure 2. We note that these three blocks all share the joint *Siamese* style feature encoder, which helps prevent overfitting towards any one task, since the model must learn encoded feature representations that are useful for all three tasks. We describe in detail the full CNN architecture and the training optimization strategies in the following sections.

3.1. Shared Encoder Architecture

In lieu of training three separate task specific encoders for each of the sub-networks, we instead train a single encoder which is shared between the tasks. This encoder is composed of 5 blocks, where each block consists of a batch normalization operation (Ioffe and Szegedy (2015)), ReLU activation (Nair and Hinton (2010)), and kernel size 3 convolution with a stride of 1. To ensure our encoder learns robust yet powerful representations of the input data, we use 64 filters for the convolution in the first block, and double this number of filters as we go deeper into the network. Feature map downsampling occurs after each block and is accomplished through a max pooling operation (Nagi et al. (2011)) with a kernel size and stride both equal to 2.

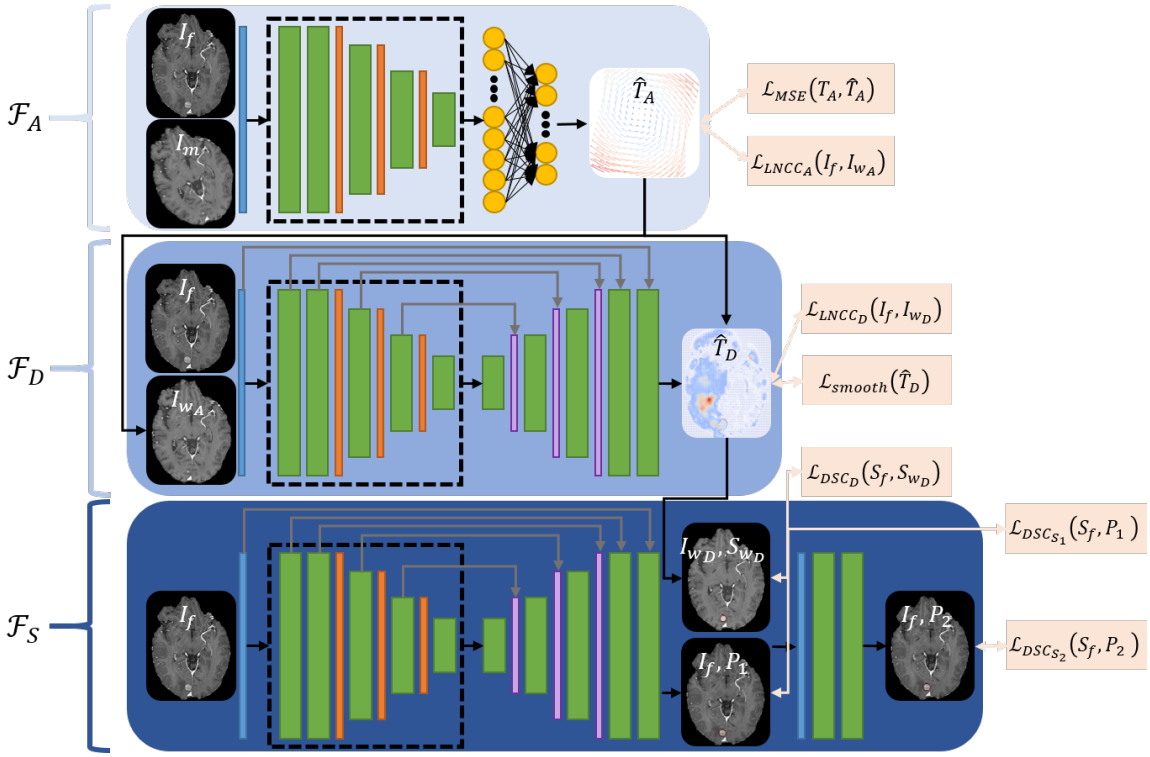

Figure 2: Schematic showing the sequential structure of our proposed framework for joint image registration and segmentation. First, the fixed image I_f, moving image I_m, and moving label S_m are passed to the affine registration network $\mathcal{F}_A$, which outputs the affinely warped image I_{w_A} and label S_{w_A}. Next, the fixed image I_f, affinely warped image I_{w_A}, and affinely warped label S_{w_A} are passed to the deformable registration network $\mathcal{F}_D$, which outputs the deformably warped image I_{w_D} and label S_{w_D}. Finally, the segmentation network $\mathcal{F}_S$ is used to segment I_f with the help of I_{w_D} and label S_{w_D} to output predicted segmentations P_1 (which does not use any prior time-point information) and P_2 (which uses prior time-point information).

3.2. Affine Registration Network $\mathcal{F}_A$ Architecture

Given input images I_f, I_m and S_m, the affine registration network $\mathcal{F}_A$ outputs $\hat{T}_A$, I_{w_A}, $S_{w_A} = \mathcal{F}_A(I_f, I_m, S_m)$, where $\hat{T}_A \in \mathbb{R}^{3\times4}$ represents the 3D affine transformation and $I_{w_A} = I_m \circ \hat{T}_A$ and $S_{w_A} = S_m \circ \hat{T}_A$ are the affinely warped moving image and label, respectively. $\mathcal{F}_A$ is composed of an opening convolution operation, the shared encoder, and a specialized affine transform decoding module. I_f and I_m are passed to the opening convolution, which uses a kernel size 7 with stride of 1. Increasing the kernel size from 3 to 7 for this opening convolution helps increase the effective receptive field (ERF) of the network, allowing

for larger transformations to be learned. The decoding module is composed of two fully connected layers with dropout (Srivastava et al. (2014)) of 0.15.

To train $\mathcal{F}_A$, we use a combination of two losses. First, we take the MSE loss between the true affine matrix T_A and the predicted affine matrix $\hat{T}_A$.

$$\mathcal{L}_{MSE}(T_A, \hat{T}_A) = \|T_A - \hat{T}_A\|_2^2 \tag{1}$$

Second, we utilize the unsigned local normalized cross correlation (LNCC) to measure the similarity between I_f and I_{w_A} (Avants et al. (2007); Dalca et al. (2019)). We define the local mean centered image $\overline{I}_f = I_f - \mu_{I_f}$, where μ_{I_f} is the convolved output of I_f and a kernel size 9 box filter. $\overline{I}_{w_A}$ is defined similarly. The LNCC is then given by:

$$LNCC(I_f, I_{w_A}) = \frac{\langle I_f - \mu_{I_f}, I_{w_A} - \mu_{I_{w_A}} \rangle^2}{\langle I_f - \mu_{I_f}, I_f - \mu_{I_f} \rangle \langle I_{w_A} - \mu_{I_{w_A}}, I_{w_A} - \mu_{I_{w_A}} \rangle} = \frac{\langle \overline{I}_f, \overline{I}_{w_A} \rangle^2}{\langle \overline{I}_f, \overline{I}_f \rangle \langle \overline{I}_{w_A}, \overline{I}_{w_A} \rangle} \tag{2}$$

where $\langle \cdot, \cdot \rangle$ is the Frobenius inner product. LNCC ranges from 0 to 1, with 1 representing perfectly aligned images. To use as a loss function, we take the negative of the value.

$$\mathcal{L}_{LNCC_A}(I_f, I_{w_A}) = -LNCC(I_f, I_{w_A}) \tag{3}$$

3.3. Deformable Registration Network $\mathcal{F}_D$ Architecture

Given input images I_f, I_{w_A}, and S_{w_A}, the deformable registration network $\mathcal{F}_D$ outputs $\hat{T}_D$, I_{w_D}, $S_{w_D} = \mathcal{F}_D(I_f, I_{w_A}, S_{w_A})$, where $\hat{T}_D \in \mathbb{R}^{h \times w \times d \times 3}$ represents the deformable transformation and $I_{w_D} = I_{w_A} \circ \hat{T}_D$ and $S_{w_D} = S_{w_A} \circ \hat{T}_D$ are the (affinely and) deformably warped moving image and label, respectively. $\mathcal{F}_D$ is composed of an opening convolution operation, the shared encoder, and a specialized deformable transform decoding module. The decoding module is the inverse of the shared encoder, with trilinear upsampling layers in lieu of max pooling. Following standard U-Net approaches, we interleave skip connections from the encoder to the decoder.

To train $\mathcal{F}_D$, we use a combination of three losses. First, we measure the similarity between I_f and I_{w_D} as follows:

$$\mathcal{L}_{LNCC_D}(I_f, I_{w_D}) = -LNCC(I_f, I_{w_D}) \tag{4}$$

Next, to encourage the predicted DVF to be spatially smooth, we use the following second order bending energy penalty (Rueckert et al. (1999)):

$$\begin{aligned} \mathcal{L}_{smooth}(\hat{T}_D) = \sum \Bigg(& \left\| \frac{\partial^2 \hat{T}_D}{\partial x} \right\|_2^2 + \left\| \frac{\partial^2 \hat{T}_D}{\partial y} \right\|_2^2 + \left\| \frac{\partial^2 \hat{T}_D}{\partial z} \right\|_2^2 \\ & + 2 \left\| \frac{\partial^2 \hat{T}_D}{\partial xy} \right\|_2^2 + 2 \left\| \frac{\partial^2 \hat{T}_D}{\partial xz} \right\|_2^2 + 2 \left\| \frac{\partial^2 \hat{T}_D}{\partial yz} \right\|_2^2 \Bigg) \end{aligned} \tag{5}$$

where spatial gradients are approximated via a second order finite difference. If we place too much weight on this penalty, the predicted DVF will be over-smoothed and will

not adequately align I_f and I_{w_D}. Conversely, if we do not penalize enough, we may see physiologically unrealistic transformations such as folding or other discontinuities.

Finally, to add extra incentive to the network to learn how to accurately shrink or enlarge tumors (which will improve our downstream segmentation performance), we utilize the Dice Score Coefficient (DSC) (Dice (1945)). Given two label maps p and q, the DSC measures how well they overlap as follows:

$$DSC(p, q) = \frac{2 \sum pq}{\sum p + \sum q} \tag{6}$$

DSC ranges from 0 to 1, with 1 representing perfect overlap. To use as a loss function, we take the negative of the value.

$$\mathcal{L}_{DSC_D}(S_f, S_{w_D}) = -DSC(S_f, S_{w_D}) \tag{7}$$

3.4. Segmentation Network $\mathcal{F}_S$ Architecture

Given input images I_f, I_{w_D}, and S_{w_D}, the segmentation network $\mathcal{F}_S$ outputs P_1, $P_2 = \mathcal{F}_S(I_f, I_{w_D}, S_{w_D})$, where P_1 and P_2 are pixelwise probability maps for likely brain metastases. $\mathcal{F}_S$ is composed of an opening convolution operation, the shared encoder, and a specialized segmentation module. This module works slightly differently from the prior two in that the input to the opening convolution is solely the fixed image. The output of the segmentation decoding module, which follows the same structure as the deformable decoding module, is P_1. As this part of the segmentation module is run solely using the fixed image, it does not incorporate any prior time-point information at this point. The second part of the segmentation module is a residual block (He et al. (2015)) which fuses information from the current time-point (I_f and P_1) with information from the prior time-point (I_{w_D} and S_{w_D}) to output the final enhanced segmentation P_2.

To train $\mathcal{F}_S$, we apply DSC loss to both P_1 and P_2 as follows:

$$\mathcal{L}_{DSC_{S_1}}(S_f, P_1) = -DSC(S_f, P_1) \tag{8}$$

$$\mathcal{L}_{DSC_{S_2}}(S_f, P_2) = -DSC(S_f, P_2) \tag{9}$$

3.5. Pyramidal Architecture

In this section, we will briefly describe the pyramidal structure of our network architecture, a schematic of which is shown in figure 3. To begin, we use a L-level pyramid framework for both our affine and deformable registration networks, where we set $L = 3$ for this paper. For level $i \in \{1, 2, 3\}$ in the pyramid, the input images are downsampled by a factor 0.5^{L-i}. A forward pass through the pyramidal structure entails iteratively registering the images at from the coarsest scale (level $i = 1$) to the finest scale (level $i = 3$). More specifically, at pyramid level $i = 1$, we downsample images I_f and I_m by a factor of $0.5^{L-i} = 4$ to obtain coarse images I_{f_1} and I_{m_1}. These are passed through $\mathcal{F}_{A_1}$ to output a coarse affine transformation $\hat{T}_{A_1}$. At pyramid levels $i > 1$, we downsample images I_f and I_m by the appropriate scale factor to obtain images I_{f_i} and I_{m_i} and we warp I_{m_i} with the previously

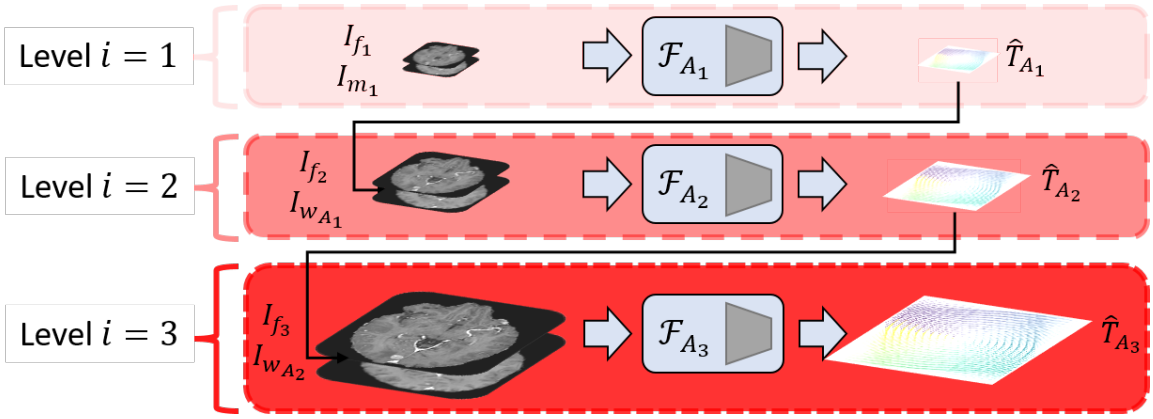

Figure 3: Schematic showing the pyramidal structure of our proposed framework for affine image registration. The predicted affine transformation $\hat{T}_A$ is iteratively refined by registering the images from the coarsest scale (level $i = 1$) to the finest scale (level $i = 3$). The pyramidal structure for deformable registration follows the same approach.

computed transform $\hat{T}_{A_{i-1}}$ to make $I_{w_{A_{i-1}}}$. I_f and $I_{w_{A_{i-1}}}$ are passed through $\mathcal{F}_{A_i}$ to output a refined affine transformation $\hat{T}_{A_i}$. The pyramidal structure for deformable registration follows the same approach.

As stated previously, the pyramidal sub-networks $\mathcal{F}_{A_i}$ and $\mathcal{F}_{D_i}$ for $i \in \{1, 2, 3\}$ share the same trainable parameters. The only difference between sub-networks is that level $i = 1$ uses two fewer and level $i = 2$ uses one fewer non-trainable max pooling and trilinear upsampling layer than does level $i = 3$, respectively. By carefully removing down-sampling and up-sampling layers for the coarser image resolutions, we ensure that all input and output dimensions match up.

3.6. Total Pyramidal Loss Function

To balance losses coming from different levels in the pyramid, we use the following:

$$\mathcal{L}_{total} = \sum_{i=1}^{L} \gamma^{L-i} (\mathcal{L}_{MSE} + \mathcal{L}_{LNCC_A} + \mathcal{L}_{LNCC_D} + \lambda_1 \mathcal{L}_{smooth} + \lambda_2 \mathcal{L}_{DSC_D} + \mathcal{L}_{DSC_{S_1}} + \mathcal{L}_{DSC_{S_2}}) \tag{10}$$

where γ controls how much to decrease the loss at coarser image scales, λ_1 controls the strength of $\mathcal{L}_{smooth}$, and λ_2 is a weighting hyperparameter to prevent $\mathcal{L}_{DSC_D}$ from overpowering $\mathcal{L}_{smooth}$ and resulting in spatially discontinuous deformations at tumor boundaries.

4. Cohort

We acquired a cohort of patients with clinically diagnosed BM from a retrospective database from the Brigham and Women's Hospital radiation oncology clinic. We selected adult patients with newly diagnosed BM who were undergoing stereotactic radiosurgery treatment from April 2004 to November 2014. This yields 148 patients with 885 time-points total. To train our registration and segmentation model, we divided this cohort on the patient level into training (100 patients; 617 time-points), validation (25 patients; 139 time-points), and testing (23 patients; 129 time-points) sets. To better understand how model performance varies as a function of lesion volume, each set was sub-divided into groups of consisting of micro ($< 25\text{mm}^3$), small ($\geq 25\text{mm}^3$ and $< 125\text{mm}^3$), medium ($\geq 125\text{mm}^3$ and $< 1000\text{mm}^3$), and large ($\geq 1000\text{mm}^3$) lesions. Table 2 in Appendix A details demographical statistics of these selected cohorts. All patient examinations were viewed and annotated in Slicer3D (Fedorov et al. (2012)). Segmentations were first manually segmented by a neuro-oncologist and then manually edited by a board-certified neuro-radiologist with 16 years' experience. These segmentations serve as the ground truth for our experiments.

5. Results

Our goal is A) to accurately and efficiently compute affine and deformable registrations for brain tumor imaging data, and B) to improve segmentation of brain metastases by using prior time-point information. In this section, we quantitatively evaluate these goals.

5.1. Evaluation Approach/Study Design

We will evaluate registration performance as follows:

1. By computing LNCC between the fixed image I_f and the moved images I_{w_A} and I_{w_D}.
2. By computing DSC between the fixed label S_f and the moved labels S_{w_A} and S_{w_D}.

Since ANTS and VXM are the most well-known and most rigorously validated classical and DL based registration methods, respectively, we will compare our method SPIRS to these two baselines. Namely, we will employ the Wilcoxon signed-rank test (Wilcoxon (1992)), the non-parametric analog to the paired t-test. Following recommendation from Dalca et al. (2019), we change the ANTS default parameters to use a step size of 0.25, Gaussian parameters of (9.0, 0.2), and three levels in the pyramid.

We will evaluate segmentation performance as follows:

1. By computing DSC, 95th percentile of Hausdorff distance (HD95) (Huttenlocher et al. (1993)), and sensitivity of lesion detection between the fixed label I_f and the predicted segmentations P_1 and P_2.

We hypothesize that using segmentation labels generated using prior time-point information (P_2) will significantly improve sensitivity of lesion detection when compared to using segmentation labels generated without any prior time-point information (P_1). To verify this, we will employ McNemar's test, a type of chi-squared test for paired data.

5.2. Implementation and Training Details

Our model SPIRS was implemented in DeepNeuro (Beers et al. (2020)) with Tensorflow 2.10 backend (Abadi et al. (2015)). The three hyperparameters γ, λ_1, and λ_2 in our total pyramidal loss function $\mathcal{L}_{total}$ are set to 0.5, 0.5, and 0.1, respectively. Further training details are included in Appendix B.

5.3. Image Registration Results

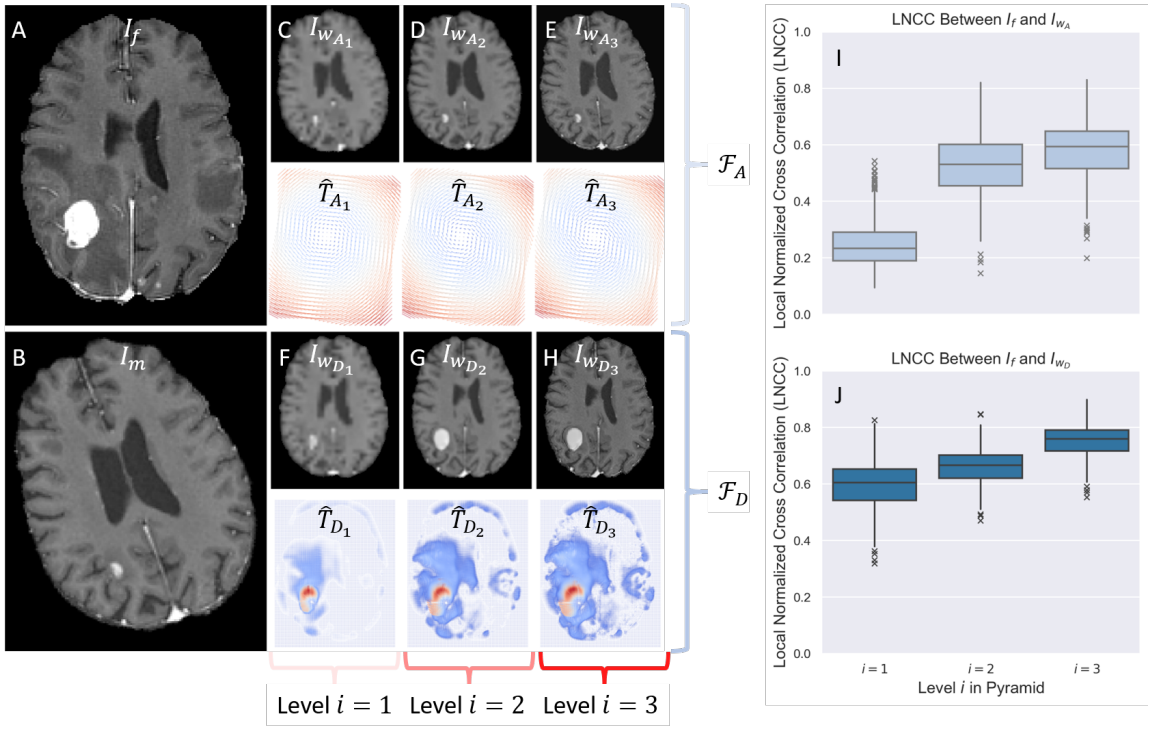

Figure 4: A fixed image I_f (A) and a moving image I_m (B) from the test set are registered together via SPIRS. (C-E) show the results of pyramidal affine registration and (F-H) shows the results of pyramidal deformable registration. (I,J) shows the LNCC of all intra-patient pairs of images in our test set for affine and deformable registration, respectively.

To begin, at each level $i \in \{1, 2, 3\}$ in the pyramid, we compute the LNCC between the fixed image I_{f_i} and the moved images $I_{w_{A_i}}$ and $I_{w_{D_i}}$. We find that LNCC markedly improves as we pass through the pyramid structures for both $\mathcal{F}_{A_i}$ and $\mathcal{F}_{D_i}$ (figure 4(I, J)). An example registration between two test set images is shown in figure 4(A-H). Panel (A) shows the fixed image I_f and panel (B) shows the initially misaligned moving image I_m, which have an initial LNCC of 0.064. We note that the affine misalignment is purely 2D for the purpose of this figure, but emphasize that our network can register images fully

in 3D. Panels (C-E) show the results of affine registration from the coarsest to the finest resolution along with the predicted transformations (which are visualized as a DVF). Due to the size of the figure, it is difficult to see the minute differences between $\hat{T}_{A_1}$, $\hat{T}_{A_2}$, and $\hat{T}_{A_3}$. Nevertheless, we note that the registration quality improved as evidenced by the LNCC, which rises to 0.110, 0.160, and 0.163 at levels 1, 2, and 3, respectively. Panels (F-H) show the results of deformable registration. Here, we can see striking differences between $\hat{T}_{D_1}$, $\hat{T}_{D_2}$, and $\hat{T}_{D_3}$, with LNCC rising to 0.398, 0.525, and 0.614, respectively. Qualitatively, we can see that the major deformations that occur include enlarging the tumor and shrinking the ventricles.

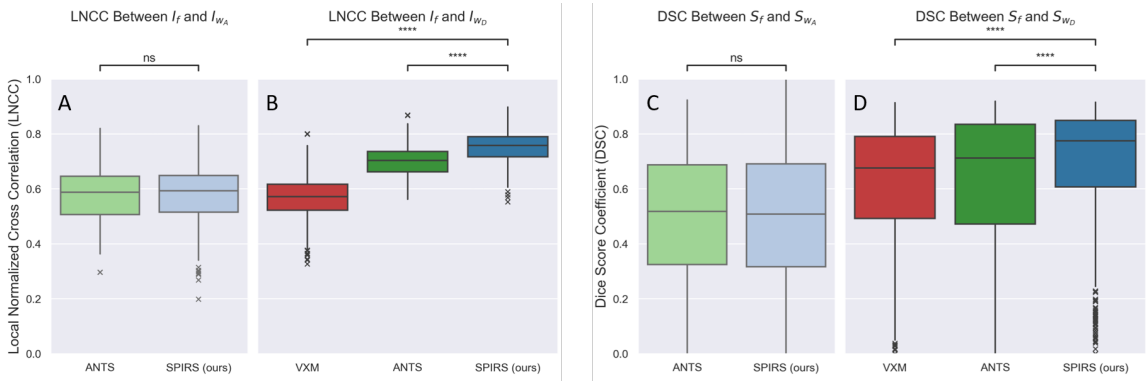

Figure 5: Quantitative comparison between our method SPIRS and baseline methods VXM and ANTS using LNCC (A,B) and DSC (C,D).

Next, we compare our method SPIRS against baseline registration methods ANTS and VXM (figure 5). Panels (A,C) compare LNCC and DSC between ANTS and SPIRS for affine registration; panels (B,D) compare LNCC and DSC between VXM, ANTS, and SPIRS for deformable registration. Since VXM only performs deformable registration (thus requiring images to be affinely aligned as a pre-processessing step), we affinely align all images via SPIRS before testing the three deformable registration algorithms to ensure fair comparison. Using LNCC and DSC as our metrics, we observe that SPIRS performs similarly to ANTS for affine registration, and performs better than VXM and ANTS for deformable registration ($p < 0.0001$). In figure 6, we show example deformable registrations between two pairs of test set images using VXM, ANTS, and SPIRS. We note that while all three methods can deformably register the large lesion in the first row fairly well (though SPIRS subjectively does the best), we can see that only SPIRS can correctly deformably warp the smaller lesion in the second row.

Finally, to understand the effect of the pyramidal component of our registration network, we ran a small ablation study (table 3 in Appendix C). When removing the pyramidal component of the network, we observe a decrease in median LNCC of 0.45 ($p < 0.0001$) and 0.04 ($p < 0.0001$) for affine and deformable registration, respectively. This highlights the importance of using a multi-scale approach in order to guarantee optimal results.

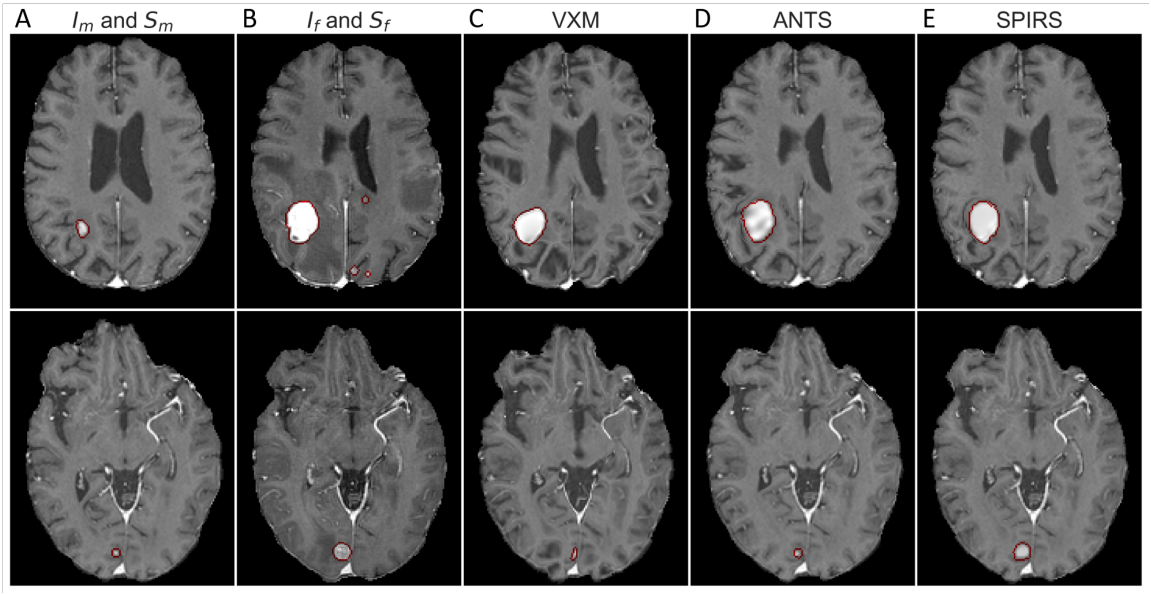

Figure 6: Comparison of deformable registration quality between a fixed image (A) and moving image (B) for VXM (C), ANTS (D), and SPIRS (E). Lesion outlined in red.

5.4. Image Segmentation Results

We report results for the segmentation of BM with and without using prior time-point information in table 1. We observe significant improvement in sensitivity of lesion detection for micro and small sized lesions when incorporating prior time-point information ($p < 0.0001$). Specifically, we detect 15% more micro-lesions (amounting to 92 fewer missed detections), and we detect 10% more small-lesions (amounting to 45 fewer missed detections). The sensitivity for medium and large sized lesions remains unchanged. We note similar trends for DSC and HD95. Examples of newly detected micro-metastatic lesions (all with ground truth volume $< 10\text{mm}^3$) are shown in figure 7.

6. Discussion

Automated segmentation of BM is challenging machine learning task, with current segmentation algorithms exhibiting poor performance for micro-metastatic lesions. Patients undergoing active treatment will require regular follow-up imaging scans for the purpose of treatment response assessment, and current segmentation approaches are likely to make the same mistakes repeatedly (e.g. micro-metastatic lesion missed at baseline is missed again at time-point 1). Instead of segmenting each image independently of each other, we propose to utilize the prior time-point imaging as a means to improve segmentation of the new time-point. To that end, we developed SPIRS, a method to affinely and deformably register a prior time-point image (with known ground truth segmentation) to a new time-point image

Table 1: Median DSC, HD95, and sensitivity with and without using prior time-point information split by lesion size.

Lesion Size	Prior Info	DSC†	HD95†	Sensitivity*
micro		0.0 (0.0 - 0.21)	inf (2.12 – inf)	27 (24 - 31)
micro	✓	0.0 (0.0 - 0.44)	inf (1.41 – inf)	42 (38 - 46)
small		0.53 (0.0 – 0.68)	1.78 (1.0 – inf)	70 (66 - 74)
small	✓	0.57 (0.16 – 0.69)	1.41 (1.0 – 3.98)	80 (76 - 84)
medium		0.78 (0.74 – 0.84)	1.0 (1.0 – 1.41)	97 (92 - 99)
medium	✓	0.79 (0.74 – 0.85)	1.0 (1.0 – 1.41)	97 (92 - 99)
large		0.90 (0.88 – 0.92)	1.0 (1.0 – 1.41)	100 (93 - 100)
large	✓	0.91 (0.89 – 0.92)	1.0 (1.0 – 1.41)	100 (93 - 100)

* Data in parantheses are percentages.
† Data in parantheses are IQRs.

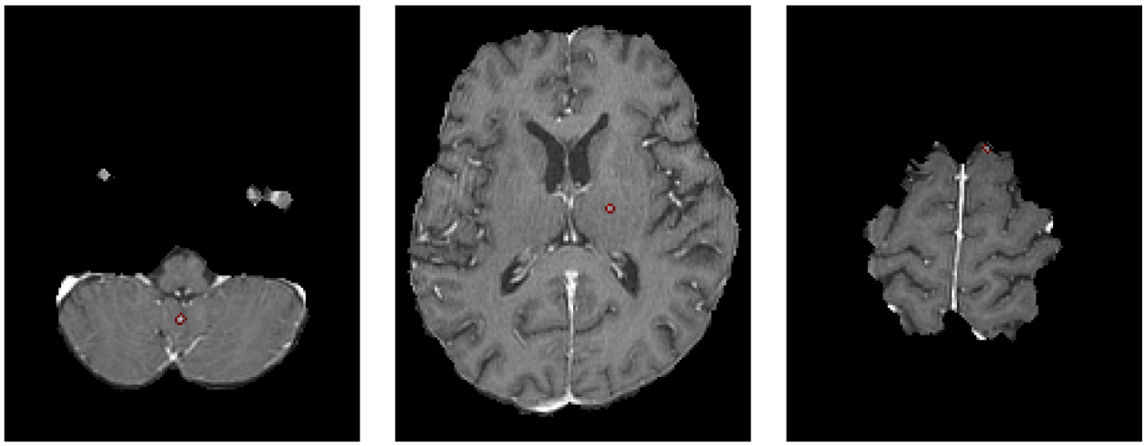

Figure 7: Three different test set patients with automatically segmented micro-metastases (outlined in red) that are missed if not using prior time-point information.

for the purpose of improving segmentation performance on this new image. While we focus on BM for this paper, we emphasize that our model architecture is generalizable to other challenging medical segmentation problems where longitudinal imaging data is present.

In our experimental studies, we compare registration via SPIRS to registration via ANTS and VXM and the comparative analysis indicates that SPIRS performs equivalently for affine registration and performs better for deformable registration. We attribute this improvement in performance to the fact that most algorithms are developed for normal anatomy and are not well equipped to model large radial deformations like we see for tumor

growth/shrinkage. Using our method, we subsequently show that we can drastically decrease the number of missed detections of BM when we utilize prior time-point information. This has numerous clinical implications.

First, we can reduce the annotation burden on clinicians by decreasing the number of mistakes that must be corrected each time a patient comes in for follow-up imaging. This will help streamline clinical workflows and enable the clinician to spend more time working on important downstream tasks such as treatment response assessment. Next, our approach can help make longitudinal patient analysis more consistent. Due to a multitude of factors, patient follow-up imaging will occasionally be read and interpreted by a different radiologist. Not only can this can lead to inter-rater variability, where a lesion may be identified on one visit but not the other, but it may also have significant effects on the categorization of treatment response (e.g. accidentally assigning partial response (PR) instead of stable disease (SD)). Our method can help prevent such issues by ensuring that lesions that were identified in prior time-points are correctly carried forwards to new time-points. Third, many BM specific clinical trials are run independently at differing institutions (Tawbi et al. (2018); Goldberg et al. (2020); Brastianos et al. (2021)). In order to accurately assess treatment efficacy across multiple institutions and clinical trials, a standardized non-volumetric measurement system known as the response assessment in neuro-oncology (RANO) criteria is used (Lin et al. (2015)). The RANO criteria has been shown to lead to higher amounts of inter- and intra-rater variability and has significantly poorer repeatability and consistency compared to true volumetric measurements (Chang et al. (2019); Peng et al. (2022)). Showing promising results for segmentation of BM, our approach is a step towards replacing RANO with volumetric tumor burden.

Limitations There are a few limitations of our work. First, our dataset was collected retrospectively from a single institution. Model performance when used in a prospective manner is currently uncertain. Moreover, variations in scanner settings and MRI parameters between institutions can affect performance, and future work will entail validating our approach on a larger multi-site dataset. Second, our approach relies on the existence of high quality prior time-point segmentations. If the radiologist that interpreted the prior time-point missed a lesion or segmented a false positive, these mistakes will most likely be inadvertently carried through to the new time-point. Finally, we did not run any exhaustive grid searches for the hyper-parameters in our approach. Minor improvements to both the registration and segmentation may be achieved through sophisticated hyper-parameter tuning.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on

heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.

Edward H. Adelson, Peter J. Burt, Charles H. Anderson, Joan M. Ogden, and James R. Bergen. Pyramid methods in image processing. 1984.

B B Avants, C L Epstein, M Grossman, and J C Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med Image Anal*, 12(1):26–41, June 2007.

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, September 2010.

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C. Kitamura, Sarthak Pati, Luciano M. Prevedello, Jeffrey D. Rudie, Chiharu Sako, Russell T. Shinohara, Timothy Bergquist, Rong Chai, James Eddy, Julia Elliott, Walter Reade, Thomas Schaffter, Thomas Yu, Jiaxin Zheng, Ahmed W. Moawad, Luiz Otavio Coelho, Olivia McDonnell, Elka Miller, Fanny E. Moron, Mark C. Oswood, Robert Y. Shih, Loizos Siakallis, Yulia Bronstein, James R. Mason, Anthony F. Miller, Gagandeep Choudhary, Aanchal Agarwal, Cristina H. Besada, Jamal J. Derakhshan, Mariana C. Diogo, Daniel D. Do-Dai, Luciano Farage, John L. Go, Mohiuddin Hadi, Virginia B. Hill, Michael Iv, David Joyner, Christie Lincoln, Eyal Lotan, Asako Miyakoshi, Mariana Sanchez-Montano, Jaya Nath, Xuan V. Nguyen, Manal Nicolas-Jilwan, Johanna Ortiz Jimenez, Kerem Ozturk, Bojan D. Petrovic, Chintan Shah, Lubdha M. Shah, Manas Sharma, Onur Simsek, Achint K. Singh, Salil Soman, Volodymyr Statsevych, Brent D. Weinberg, Robert J. Young, Ichiro Ikuta, Amit K. Agarwal, Sword C. Cambron, Richard Silbergleit, Alexandru Dusoi, Alida A. Postma, Laurent Letourneau-Guillon, Gloria J. Guzman Perez-Carrillo, Atin Saha, Neetu Soni, Greg Zaharchuk, Vahe M. Zohrabian, Yingming Chen, Milos M. Cekic, Akm Rahman, Juan E. Small, Varun Sethi, Christos Davatzikos, John Mongan, Christopher Hess, Soonmee Cha, Javier Villanueva-Meyer, John B. Freymann, Justin S. Kirby, Benedikt Wiestler, Priscila Crivellaro, Rivka R. Colen, Aikaterini Kotrotsou, Daniel Marcus, Mikhail Milchenko, Arash Nazeri, Hassan Fathallah-Shaykh, Roland Wiest, Andras Jakab, Marc-Andre Weber, Abhishek Mahajan, Bjoern Menze, Adam E. Flanders, and Spyridon Bakas. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification, 2021.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci Data*, 4:170117, September 2017.

Andrew Beers, James Brown, Ken Chang, Katharina Hoebel, Jay Patel, K Ina Ly, Sara M Tolaney, Priscilla Brastianos, Bruce Rosen, Elizabeth R Gerstner, and Jayashree Kalpathy-Cramer. DeepNeuro: an open-source deep learning toolbox for neuroimaging. *Neuroinformatics*, 2020. ISSN 1559-0089. doi: 10.1007/s12021-020-09477-5. URL https://doi.org/10.1007/s12021-020-09477-5.

Priscilla K Brastianos, Albert E Kim, Nancy Wang, Eudocia Q Lee, Jennifer Ligibel, Justine V Cohen, Ugonma N Chukwueke, Maura Mahar, Kevin Oh, Michael D White, Helen A Shih, Deborah Forst, Justin F Gainor, Rebecca S Heist, Elizabeth R Gerstner, Tracy T Batchelor, Donald Lawrence, David P Ryan, A John Iafrate, Anita Giobbie-Hurder, Sandro Santagata, Scott L Carter, Daniel P Cahill, and Ryan J Sullivan. Palbociclib demonstrates intracranial activity in progressive brain metastases harboring cyclin-dependent kinase pathway alterations. *Nature Cancer*, 2(5):498–502, 2021. ISSN 2662-1347. doi: 10.1038/s43018-021-00198-5. URL https://doi.org/10.1038/s43018-021-00198-5.

Daniel N Cagney, Allison M Martin, Paul J Catalano, Amanda J Redig, Nancy U Lin, Eudocia Q Lee, Patrick Y Wen, Ian F Dunn, Wenya Linda Bi, Stephanie E Weiss, Daphne A Haas-Kogan, Brian M Alexander, and Ayal A Aizer. Incidence and prognosis of patients with brain metastases at diagnosis of systemic malignancy: a population-based study. *Neuro Oncol*, 19(11):1511–1521, October 2017.

Ken Chang, Andrew L. Beers, Harrison X. Bai, James M. Brown, K. Ina Ly, Xuejun Li, Joeky T. Senders, Vasileios K. Kavouridis, Alessandro Boaro, Chang Su, Wenya Linda Bi, Otto Rapalino, Weihua Liao, Qin Shen, Hao Zhou, Bo Xiao, Yinyan Wang, Paul J. Zhang, Marco C. Pinho, Patrick Y. Wen, Tracy T. Batchelor, Jerrold L. Boxerman, Omar Arnaout, Bruce R. Rosen, Elizabeth R. Gerstner, Li Yang, Raymond Y. Huang, and Jayashree Kalpathy-Cramer. Automatic assessment of glioma burden: A deep learning algorithm for fully automated volumetric and bidimensional measurement. *Neuro-Oncology*, 2019. ISSN 15235866. doi: 10.1093/neuonc/noz106.

Evelyn Chee and Zhenzhou Wu. Airnet: Self-supervised affine registration for 3d medical images using neural networks, 2018.

Xiang Chen, Yan Xia, Nishant Ravikumar, and Alejandro F Frangi. Joint segmentation and discontinuity-preserving deformable registration: Application to cardiac cine-mr images, 2022.

Vinton W.T. Cheng, Manuel Sarmiento Soto, Alexandre A. Khrapitchev, Francisco Perez-Balderas, Rasheed Zakaria, Michael D. Jenkinson, Mark R. Middleton, and Nicola R. Sibson. VCAM-1–targeted MRI Enables Detection of Brain Micrometastases from Different Primary Tumors. *Clinical Cancer Research*, 25(2):533–543, 01 2019. ISSN 1078-0432. doi: 10.1158/1078-0432.CCR-18-1889. URL https://doi.org/10.1158/1078-0432.CCR-18-1889.

Stergios Christodoulidis, Mihir Sahasrabudhe, Maria Vakalopoulou, Guillaume Chassagnon, Marie-Pierre Revel, Stavroula Mougiakakou, and Nikos Paragios. Linear and deformable image registration with 3d convolutional neural networks, 2018.

Adrian V. Dalca, Guha Balakrishnan, John Guttag, and Mert R. Sabuncu. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical Image Analysis*, 57:226–236, oct 2019. doi: 10.1016/j.media.2019.07.006.

Bob D. de Vos, Floris F. Berendsen, Max A. Viergever, Hessam Sokooti, Marius Staring, and Ivana Išgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical Image Analysis*, 52:128–143, feb 2019. doi: 10.1016/j.media.2018.11.010.

Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26 (3):297–302, 1945. doi: https://doi.org/10.2307/1932409. URL https://esajournals.onlinelibrary.wiley.com/doi/abs/10.2307/1932409.

Alessandra Fabi, Alessandra Felici, Giulio Metro, Alessandra Mirri, Emilio Bria, Stefano Telera, Luca Moscetti, Michelangelo Russillo, Gaetano Lanzetta, Giovanni Mansueto, Andrea Pace, Marta Maschio, Antonello Vidiri, Isabella Sperduti, Francesco Cognetti, and Carmine M Carapella. Brain metastases from solid tumors: disease outcome according to type of treatment and therapeutic resources of the treating center. *J Exp Clin Cancer Res*, 30(1):10, January 2011.

Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean-Christophe Fillion-Robin, Sonia Pujol, Christian Bauer, Dominique Jennings, Fiona Fennessy, Milan Sonka, John Buatti, Stephen Aylward, James V Miller, Steve Pieper, and Ron Kikinis. 3D slicer as an image computing platform for the quantitative imaging network. *Magn Reson Imaging*, 30(9):1323–1341, July 2012.

Michał Futrega, Alexandre Milesi, Michal Marcinkiewicz, and Pablo Ribalta. Optimized u-net for brain tumor segmentation, 2021.

Joan Glaunes, Anqi Qiu, Michael I Miller, and Laurent Younes. Large deformation diffeomorphic metric curve mapping. *Int J Comput Vis*, 80(3):317–336, December 2008.

Sarah B Goldberg, Kurt A Schalper, Scott N Gettinger, Amit Mahajan, Roy S Herbst, Anne C Chiang, Rogerio Lilenbaum, Frederick H Wilson, Sacit Bulent Omay, James B Yu, Lucia Jilaveanu, Thuy Tran, Kira Pavlik, Elin Rowen, Heather Gerrish, Annette Komlo, Richa Gupta, Hailey Wyatt, Matthew Ribeiro, Yuval Kluger, Geyu Zhou, Wei Wei, Veronica L Chiang, and Harriet M Kluger. Pembrolizumab for management of patients with NSCLC and brain metastases: long-term results and biomarker analysis from a non-randomised, open-label, phase 2 trial. *Lancet Oncol*, 21(5):655–663, April 2020.

Endre Grøvik, Darvin Yi, Michael Iv, Elizabeth Tong, Daniel Rubin, and Greg Zaharchuk. Deep learning enables automatic detection and segmentation of brain metastases on multisequence MRI. *J Magn Reson Imaging*, 51(1):175–182, May 2019.

S. Growcott, T. Dembrey, R. Patel, D. Eaton, and A. Cameron. Inter-observer variability in target volume delineations of benign and metastatic brain tumours for stereotactic radiosurgery: Results of a national quality assurance programme. *Clinical Oncology*, 32 (1):13–25, 2020. ISSN 0936-6555. doi: https://doi.org/10.1016/j.clon.2019.06.015. URL https://www.sciencedirect.com/science/article/pii/S0936655519302766.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

D.P. Huttenlocher, G.A. Klanderman, and W.J. Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15 (9):850–863, 1993. doi: 10.1109/34.232073.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015.

Fabian Isensee, Paul F. Jaeger, Peter M. Full, Philipp Vollmuth, and Klaus H. Maier-Hein. nnu-net for brain tumor segmentation, 2020.

Kh Tohidul Islam, Sudanthi Wijewickrema, and Stephen O'Leary. A deep learning based framework for the registration of three dimensional multi-modal medical images of the head. *Scientific Reports*, 11(1):1860, January 2021.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks, 2016.

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, 2009. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2008.12.037. URL https://www.sciencedirect.com/science/article/pii/S1053811908012974.

Alyssa Y. Li, Karolina Gaebe, Amna Zulfiqar, Grace Lee, Katarzyna J. Jerzak, Arjun Sahgal, Steven Habbous, Anders W. Erickson, and Sunit Das. Association of Brain Metastases With Survival in Patients With Limited or Stable Extracranial Disease: A Systematic Review and Meta-analysis. *JAMA Network Open*, 6(2):e230475–e230475, 02 2023. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2023.0475. URL https://doi.org/10.1001/jamanetworkopen.2023.0475.

Nancy U Lin, Eudocia Q Lee, Hidefumi Aoyama, Igor J Barani, Daniel P Barboriak, Brigitta G Baumert, Martin Bendszus, Paul D Brown, D Ross Camidge, Susan M Chang, Janet Dancey, Elisabeth G E de Vries, Laurie E Gaspar, Gordon J Harris, F Stephen Hodi, Steven N Kalkanis, Mark E Linskey, David R Macdonald, Kim Margolin, Minesh P Mehta, David Schiff, Riccardo Soffietti, John H Suh, Martin J van den Bent, Michael A Vogelbaum, Patrick Y Wen, and Response Assessment in Neuro-Oncology (RANO) group. Response assessment criteria for brain metastases: proposal from the RANO group. *Lancet Oncol*, 16(6):e270–8, May 2015.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization, 2017.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-André Weber, Tal Arbel, Brian B Avants, Nicholas

Ayache, Patricia Buendia, D Louis Collins, Nicolas Cordier, Jason J Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Çağatay Demiralp, Christopher R Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M Iftekharuddin, Raj Jena, Nigel M John, Ender Konukoglu, Danial Lashkari, José Antonió Mariz, Raphael Meier, Sérgio Pereira, Doina Precup, Stephen J Price, Tammy Riklin Raviv, Syed M S Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A Silva, Nuno Sousa, Nagesh K Subbanna, Gabor Szekely, Thomas J Taylor, Owen M Thomas, Nicholas J Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imaging*, 34(10):1993–2024, December 2014.

Tony C. W. Mok and Albert C. S. Chung. Large deformation diffeomorphic image registration with laplacian pyramid networks, 2020.

Jawad Nagi, Frederick Ducatelle, Gianni A. Di Caro, Dan Cireşan, Ueli Meier, Alessandro Giusti, Farrukh Nagi, Jürgen Schmidhuber, and Luca Maria Gambardella. Max-pooling convolutional neural networks for vision-based hand gesture recognition. In *2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pages 342–347, 2011. doi: 10.1109/ICSIPA.2011.6144164.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 807–814, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

Y Nomoto, T Miyamoto, and Y Yamaguchi. Brain metastasis of small cell lung carcinoma: comparison of Gd-DTPA enhanced magnetic resonance imaging and enhanced computerized tomography. *Jpn J Clin Oncol*, 24(5):258–262, October 1994.

Jon André Ottesen, Darvin Yi, Elizabeth Tong, Michael Iv, Anna Latysheva, Cathrine Saxhaug, Kari Dolven Jacobsen, Åslaug Helland, Kyrre Eeg Emblem, Daniel L Rubin, Atle Bjørnerud, Greg Zaharchuk, and Endre Grøvik. 2.5d and 3D segmentation of brain metastases with deep learning on multinational MRI data. *Front Neuroinform*, 16:1056068, January 2023.

Jay Patel, Ken Chang, Katharina Hoebel, Mishka Gidwani, Nishanth Arun, Sharut Gupta, Mehak Aggarwal, Praveer Singh, Bruce R. Rosen, Elizabeth R. Gerstner, and Jayashree Kalpathy-Cramer. Segmentation, survival prediction, and uncertainty estimation of gliomas from multimodal 3d mri using selective kernel networks. In Alessandro Crimi and Spyridon Bakas, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 228–240, Cham, 2021. Springer International Publishing. ISBN 978-3-030-72087-2.

Jian Peng, Daniel D Kim, Jay B Patel, Xiaowei Zeng, Jiaer Huang, Ken Chang, Xinping Xun, Chen Zhang, John Sollee, Jing Wu, Deepa J Dalal, Xue Feng, Hao Zhou, Chengzhang Zhu, Beiji Zou, Ke Jin, Patrick Y Wen, Jerrold L Boxerman, Katherine E

Warren, Tina Y Poussaint, Lisa J States, Jayashree Kalpathy-Cramer, Li Yang, Raymond Y Huang, and Harrison X Bai. Deep learning-based automatic tumor burden assessment of pediatric high-grade gliomas, medulloblastomas, and other leptomeningeal seeding tumors. *Neuro-oncology*, 24(2):289–299, feb 2022. ISSN 1523-5866 (Electronic). doi: 10.1093/neuonc/noab151.

Chen Qin, Wenjia Bai, Jo Schlemper, Steffen E. Petersen, Stefan K. Piechnik, Stefan Neubauer, and Daniel Rueckert. Joint learning of motion estimation and segmentation for cardiac mr image sequences, 2018.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

Jeffrey D. Rudie, David A. Weiss, John B. Colby, Andreas M. Rauschecker, Benjamin Laguna, Steve Braunstein, Leo P. Sugrue, Christopher P. Hess, and Javier E. Villanueva-Meyer. Three-dimensional u-net convolutional neural network for detection and segmentation of intracranial metastases. *Radiology: Artificial Intelligence*, 3(3):e200204, 2021. doi: 10.1148/ryai.2021200204. URL https://doi.org/10.1148/ryai.2021200204.

D Rueckert, L I Sonoda, C Hayes, D L Hill, M O Leach, and D J Hawkes. Nonrigid registration using free-form deformations: application to breast MR images. *IEEE Trans Med Imaging*, 18(8):712–721, August 1999.

Hessam Sokooti, Bob de Vos, Floris Berendsen, Boudewijn P. F. Lelieveldt, Ivana Isgum, and Marius Staring. Nonrigid image registration using multi-scale 3d convolutional neural networks. In Maxime Descoteaux, Lena Maier-Hein, Alfred Franz, Pierre Jannin, D. Louis Collins, and Simon Duchesne, editors, *Medical Image Computing and Computer Assisted Intervention MICCAI 2017*, pages 232–239, Cham, 2017. Springer International Publishing. ISBN 978-3-319-66182-7.

Paul W. Sperduto, Shane Mesko, Jing Li, Daniel Cagney, Ayal Aizer, Nancy U. Lin, Eric Nesbit, Tim J. Kruser, Jason Chan, Steve Braunstein, Jessica Lee, John P. Kirkpatrick, Will Breen, Paul D. Brown, Diana Shi, Helen A. Shih, Hany Soliman, Arjun Sahgal, Ryan Shanley, William A. Sperduto, Emil Lou, Ashlyn Everett, Drexell H. Boggs, Laura Masucci, David Roberge, Jill Remick, Kristin Plichta, John M. Buatti, Supriya Jain, Laurie E. Gaspar, Cheng-Chia Wu, Tony J.C. Wang, John Bryant, Michael Chuong, Yi An, Veronica Chiang, Toshimichi Nakano, Hidefumi Aoyama, and Minesh P. Mehta. Survival in patients with brain metastases: Summary report on the updated diagnosis-specific graded prognostic assessment and definition of the eligibility quotient. *Journal of Clinical Oncology*, 38(32):3773–3784, 2020. doi: 10.1200/JCO.20.01255. URL https://doi.org/10.1200/JCO.20.01255. PMID: 32931399.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

Emeline Tabouret, Olivier Chinot, Philippe Metellus, Agnès Tallet, Patrice Viens, and Anthony Gonçalves. Recent trends in epidemiology of brain metastases: an overview. *Anticancer Res*, 32(11):4655–4662, November 2012.

Hussein A. Tawbi, Peter A. Forsyth, Alain Algazi, Omid Hamid, F. Stephen Hodi, Stergios J. Moschos, Nikhil I. Khushalani, Karl Lewis, Christopher D. Lao, Michael A. Postow, Michael B. Atkins, Marc S. Ernstoff, David A. Reardon, Igor Puzanov, Ragini R. Kudchadkar, Reena P. Thomas, Ahmad Tarhini, Anna C. Pavlick, Joel Jiang, Alexandre Avila, Sheena Demelo, and Kim Margolin. Combined nivolumab and ipilimumab in melanoma metastatic to the brain. *New England Journal of Medicine*, 379(8):722–730, 2018. doi: 10.1056/NEJMoa1805453. URL https://doi.org/10.1056/NEJMoa1805453. PMID: 30134131.

J.-P. Thirion. Image matching as a diffusion process: an analogy with maxwell's demons. *Medical Image Analysis*, 2(3):243–260, 1998. ISSN 1361-8415. doi: https://doi.org/10.1016/S1361-8415(98)80022-4. URL https://www.sciencedirect.com/science/article/pii/S1361841598800224.

Roshan Reddy Upendra, Richard Simon, and Cristian A Linte. Joint deep learning framework for image registration and segmentation of late gadolinium enhanced MRI and cine cardiac MRI. *Proc SPIE Int Soc Opt Eng*, 11598, February 2021.

Michael A. Vogelbaum, Paul D. Brown, Hans Messersmith, Priscilla K. Brastianos, Stuart Burri, Dan Cahill, Ian F. Dunn, Laurie E. Gaspar, Na Tosha N. Gatson, Vinai Gondi, Justin T. Jordan, Andrew B. Lassman, Julia Maues, Nimish Mohile, Navid Redjal, Glen Stevens, Erik Sulman, Martin van den Bent, H. James Wallace, Jeffrey S. Weinberg, Gelareh Zadeh, and David Schiff. Treatment for brain metastases: Asco-sno-astro guideline. *Journal of Clinical Oncology*, 40(5):492–516, 2022. doi: 10.1200/JCO.21.02314. URL https://doi.org/10.1200/JCO.21.02314. PMID: 34932393.

Frank Wilcoxon. *Individual Comparisons by Ranking Methods*, pages 196–202. Springer New York, New York, NY, 1992. ISBN 978-1-4612-4380-9. doi: 10.1007/978-1-4612-4380-9_16. URL https://doi.org/10.1007/978-1-4612-4380-9_16.

Zhenlin Xu and Marc Niethammer. Deepatlas: Joint semi-supervised learning of image registration and segmentation, 2019.

Appendix A.

Table 2: Patient demographic information for the selected brain metastases cohort.

Characteristic	Training Set	Validation Set	Test Set
Patient Characteristics			
No. of patients	100	25	23
No. of women*	68 (68)	13 (52)	13 (57)
Median age†	60 (52 - 66)	61 (54 - 71)	67 (60 - 73)
Examination Characteristics			
No. of MR examinations	617	139	129
No. of micro lesions	594	82	637
No. of small lesions	1245	138	465
No. of medium lesions	1070	141	156
No. of large lesions	522	128	72
Primary Cancer Type			
Lung*	51 (51)	10 (40)	16 (70)
Breast*	21 (21)	6 (24)	1 (4)
Melanoma*	17 (17)	7 (28)	3 (13)
Gastrointestinal*	3 (3)	1 (4)	2 (9)
Renal*	4 (4)	1 (4)	0 (0)
Other/Unknown*	4 (4)	0 (0)	1 (4)

* Data in parantheses are percentages.
† Data in parantheses are IQRs.

Appendix B.

Training is done using the SGD optimizer with decoupled weight decay (Loshchilov and Hutter (2017)) and we progressively decrease the learning rate via the following cosine decay schedule:

$$\eta_t = \eta_{min} + 0.5(\eta_{max} - \eta_{min})(1 + cos(\pi T_{curr}/T)) \tag{11}$$

where η_{max} is our initial learning rate (set to 1e-2), η_{min} is our final learning rate (set to 4e-5), T_{curr} is the current iteration counter, and T is the total number of iterations to train for (set to 250 epochs).

To mitigate overfitting, we apply weight decay of 4e-5 to all convolutional kernel parameters, leaving biases and scales un-regularized. The same cosine decay schedule is applied, where we set the final weight decay to be 2e-7. Furthermore, we apply real-time data augmentation during the training process. Specifically, we utilize random mirror axis flips about all three axes along with anisotropic scaling (0.9 to 1.1), rotations ($-20°$ to $20°$), shearing (-0.05 to 0.05), and translations (-20 to 20 pixels). Intensity augmentation in the form of gamma correction (.75 to 1.25) is used as well.

End-to-end training for our network is unstable due to the difficulty in balancing losses computed at different levels in the pyramid. To overcome this issue, we use a coarse-to-fine training approach. Since network parameters are shared between all levels in the pyramid, we instead begin by training only the coarsest level and successively add the other levels each time the model converges. In particular, we start by training the affine, deformable, and segmentation modules only at level $i = 1$. Since our 3D images are downsampled by a factor of 4 at this scale, we can fit a batch size of 32 into GPU memory, which allows us to train with batch normalization. After convergence, we fine-tune this model with levels $i = 1$ and $i = 2$ together. Since our images are now larger (only downsampled by a factor of 2 at this scale), we drop the batch size to 2 to ensure our model still fits into memory and we freeze batch statistics. This process is repeated one more time at the final level of the pyramid. This training process took around 48 hours on a NVIDIA Tesla V100 32GB GPU.

Appendix C.

Table 3: Ablation study to assess effect of pyramidal registration scheme on median LNCC and DSC.

Pyramidal	Transformation Type	LNCC†	DSC†
	Affine	0.14 (0.11 – 0.17)	0.26 (0.11 – 0.44)
✓	Affine	0.59 (0.52 – 0.65)	0.51 (0.32 – 0.69)
	Deformable	0.68 (0.64 – 0.71)	0.76 (0.72 – 0.79)
✓	Deformable	0.72 (0.53 – 0.82)	0.78 (0.61 – 0.85)

† Data in parantheses are IQRs.