Leveraging Segmentation to Improve Medical Image Registration

Sahar Almahfouz Nasser, Mohit Meena, Garweet Sresth, and Amit Sethi

***Abstract*— We investigate the impact of segmentation on medical image registration. We propose a novel approach called the "Weakly Supervised Semantic Attentive Medical Image Registration Network" (WSSAMNet++), which employs segmentation to guide the registration process. Specifically, our network utilizes the segmentation of the regions of interest to direct the attention of the registration network to anatomically-relevant features, enabling it to focus more effectively on the relevant parts of the images. We demonstrate the effectiveness of using segmentation and WSSAMNet++ through extensive experiments on various registration tasks, including single and multi-modal registration problems, on multiple datasets. Our approach does not require any input from the radiologist at test time and it improves the performance of the registration network in all the cases tested. In conclusion, our study highlights the importance of leveraging semantic information to aid the registration process and shows the effectiveness of the proposed method in achieving this goal.**

***Index Terms*— Attention, CT, Image-guided Surgery, MRI, Registration, Segmentation, Weakly Supervised**

I. Introduction

MEDICAL image registration refers to computing a displacement field to warp a so-called moving image so that its visual and anatomical features spatially align with the corresponding ones in a so-called fixed image. Accurate registration can enable monitoring of disease progression using longitudinally separated images, fusion of information for image-guided surgery (IGS) using images from multiple modalities, and measurement of deviation from an average anatomy using images from multiple patients [1]. Paradoxically, while the advantage of registering pairs of images separated in time, modality, or patient is more useful, it is also more challenging to find corresponding anatomical features and landmarks across disparate images. This is because images can differ in various characteristics, such as orientation, anatomical dimensions, imaging modalities, stages of treatment or disease, and spatial resolution.

Registration techniques currently used in commercial medical imaging systems require time-consuming and error-prone manual annotation of corresponding anatomical features in fixed and moving images irrespective of whether they represent the same or different modalities. These annotations can take the form of masks indicating regions of interest or landmarks representing anatomically significant keypoints. Improvements in registration accuracy (especially, across disparate images), level of automation, and computational efficiency are still needed for image registration methods to unlock more applications in diagnosis and IGS for improved patient outcomes.

We propose automating image registration and improving its accuracy by utilizing image segmentation to guide the registration process. That is, we propose not requiring the user to mark any keypoints; instead pre-trained neural networks will segment anatomical parts of the two images, while we will train another neural network to utilize the segmentation result for improving the quality of the registration. The registration network, once trained, will also not require any input from the user. Use of segmentation to improve registration is our novel idea.

To test this proposal, we devised and tested a deep neural network called weakly supervised semantic attentive medical image registration network (WSSAMNet) [2]. In this work, we further develop this idea, improve the network architecture and its training method, and rigorously assess the performance of our approach in multiple scenarios. These scenarios included inter-subject (single modality, MRI to MRI) registration, multiple modality (CT to MRI) registration, and temporal registration of images taken before and after surgery. Furthermore, we also tested our method on multiple organs and base neural network architectures. Our proposed approach showed improvement in registration accuracy for all the scenarios tested.

In the rest of the paper, we begin by discussing related work, where we provide a detailed overview of different categories of medical image registration. Following that, we delve into the specifics of our proposed method. Subsequently, we describe the experimental setup and present our results. In the conclusion, we discuss our findings, address the limitations of our method, and outline potential avenues for future research.

II. Related Work

The field of medical image registration has undergone significant advancements in recent years, with traditional methods giving way to those that utilize deep learning. Here we summarize this evolution. First, we provide a comprehensive review of traditional image registration methods, highlighting their limitations and the need for utilizing deep learning-based techniques for improved outcomes. Then, we categorize deep learning techniques into supervised, self-supervised, weakly

Work supported in part by Qualcomm Innovation Fellowship, India.
All authors were with the Indian Institute of Technology Bombay, Mumbai, India (e-mail: sahar.almahfouz.nasser@gmail.com).

supervised and unsupervised methods, and discuss their respective benefits and drawbacks, as well as the appropriate scenarios for their application. Finally, we examine the unique challenges associated with single modality and multi-modal image registration techniques.

Traditional techniques depend on pixel intensities, image features, interest points, or their combinations. For instance, notable techniques include Elastic [3], ANTs [4], NIFTYReg [5], and SIFT algorithm that is scale-invariant and can handle various types of distortions such as rotation, affine deformation, occlusion, clutter, and noise [6]. However, these techniques have serious limitations, such as failing in the presence of significant and nonrigid deformations, local minima problems, sensitivity to noise, the need for external data or assumptions, and time consumption.

To address the aforementioned limitations of traditional methods, deep learning-based methods were developed because they can learn complex nonlinear mappings from image data without explicit feature extraction or manual parameter tuning. These methods can handle nonrigid and large deformations as well as variations in scale, rotation, and viewpoint. Furthermore, these methods can be trained end-to-end for improved accuracy and efficiency. Deep learning-based image registration techniques can be mainly categorized based on the extent of supervision or the number of modalities used as described below.

1) Supervised: These methods train neural networks to reduce the error between an estimated deformation field output by them and a known deformation field (or a subset or parameters thereof). An example of the error to be reduced is the pixel-wise mean-square error between the two deformation vector fields (DVFs). For instance, in [7], a convolutional neural network (CNN) outputs a deformation field between different phases of patients' 4D-CT or 4D-cone beam CT scans. A rapid registration method was also introduced that relies on patch-wise prediction of the initial momentum parameterization in the large deformation diffeomorphic metric mapping (LDDMM) shooting formulation [8]. The obvious limitation of supervised methods is the requirement of ground truth deformation fields between the two images, which is time consuming and error-prone, for training and testing.

2) Self-supervised: When the deformation field is not available for supervised training, a single (fixed) image can be synthetically warped to produce a moving image, thus producing a known deformation field using self-supervised methods. For instance, three different categories of artificial DVFs were generated in [9] to represent the range of displacements that can be seen in real images.Another method called SAME breaks down the image registration process into three steps: affine transformation, coarse deformation, and deep deformable registration [10].In another method, to overcome the task of manually identifying transformation parameters for each image pair, a substantial amount of unlabelled data was utilized to create a synthetic dataset using affine transformations, enabling efficient training of the proposed model for 3D medical registration [11]. Self-supervised methods are limited by the range of simulated deformation fields and the similarity of characteristics between the original fixed and the synthetically generated moving images.

3) Unsupervised: In the absence of any known ground truth correspondence between the two images, alternative loss functions such as mean square error are employed in single-modality registration, while normalized cross correlation [12] or mutual information [13] are used in multi-modal registration. These losses are low when the pattern of pixel intensity and its spatial variation in a neighborhood of a point in one image can be predicted based on the same for the corresponding point in the other image. Unsupervised image registration methods include those mentioned by [14], who introduced a cGAN for retinal fundus and fluorescein angiography image registration, Guo et al. [15], who proposed an unsupervised method for multi-modal image registration using a mutual-information-based loss function, and Chen et al. [16], who introduced an unsupervised vision transformer (ViT) that uses patch-based processing for medical image registration. The ViT-V-Net combines ViT and ConvNet to facilitate volumetric medical image registration. The authors used it for unsupervised registration of T1 brain MRI scans.

A major limitation of unsupervised methods is the inability to objectively assess how good the final registration actually is.

4) Weakly supervised: Spatially sparse annotations, such as landmarks or keypoints, can be used for weak supervision to reduce the dependence on full supervision – the latter requires knowing the dense deformation field. These annotations are then used to measure the error between the registered image and the ground truth image. Additional supervision can come from the local alignment of visual features determined using image processing techniques across the two images. For instance, Hu et al. [17] presented a weakly supervised CNN for multimodal image registration. Their strategy is versatile in terms of training as it can use a variety of anatomical labels without requiring them to be identifiable in all training image pairs. In their study, two metrics were reported. The first one is the target registration error (TRE), which is calculated by taking the root-mean-square of the distances between the centers of mass of each pair of fixed and warped labels over all landmark pairs for each patient [18]. The second metric is the Dice similarity coefficient (DSC), which measures the overlap between the binary warped and fixed labels that represent the prostate glands [19]. [20] introduced ISTNs, which use a blend of image and spatial transformer networks for structure-guided registration. The proposed method employs a two-stage approach where first, an image transformer network is used to extract feature maps from the input images, and then a spatial transformer network is employed to register the images based on the extracted features. The authors introduce a new loss function that takes into account both the intensity-based similarity and structural similarity between the images. Gunnarsson et al. [21] presented an algorithm that uses a Laplacian pyramid for medical image registration which involves reducing the resolution of both the static and moving images at different feature map levels. At each level, a displacement field is calculated and then progressively improved within the network. The authors of [22] introduced a contrast-agnostic network based on GANs that is invariant to different types of contrasts

in medical images. Training the model on randomly generated shapes from noise distributions, eliminating the reliance on any type of collected data. Moreover, they demonstrated that using anatomical label maps, which are frequently accessible for the relevant anatomy, significantly enhances performance when generating images, while still avoiding the requirement for actual intensity images.

Although weakly supervised methods produce more accurate results compared to unsupervised methods, their feasibility is limited due to the requirement for annotations of regions of interest. The annotation process is often expensive and requires expert knowledge, resulting in limited and potentially noisy annotations. Therefore, in situations where such annotations are not available or not reliable, unsupervised training becomes the only feasible option, which is discussed next.

5) Single-modality and multi-modal image registration methods: Image registration techniques can be also classified based on the type of images being registered. If both images are from the same modality, it is called single-modality registration [14], whereas if they are from different modalities, it is called multi-modal registration [17]. While uni-modal registration has been extensively studied, multi-modal registration is more challenging due to the differences in image intensity, noise levels, and spatial resolutions. However, the practical utility of multi-modal registration makes it a worthwhile pursuit as it can provide complementary information from different modalities to improve diagnosis and treatment planning. Several studies have shown that multi-modal registration compared to uni-modal registration is more computationally expensive and requires more sophisticated algorithms [23].

This review of the existing method suggests that an accurate and flexible method that can work with single or multiple modalities and one that is automated such that does not require physicians to spend their valuable time annotating images at the time of diagnosis or surgery is still needed.

III. Proposed Method

Our key idea is to utilize the advances in automated segmentation of anatomical units using deep neural networks for weak supervision to train registration neural networks. Doing so requires no additional annotation from radiologists at the time of either training or testing the registration network, if a segmentation neural network is anyway available. Towards this end, we extend our preliminary work called "Weakly Supervised Semantic Attentive Medical Image Registration Network" (WSSAMNet) [2] with improved backbone architectures for segmentation and registration. We dub the improved method WSSAMNet++. We describe various components of the proposed method below.

A. WSSAMNet++

Previously, our team developed WSSAMNet [2], which is a network designed specifically for medical image registration. The network comprises two main components – a segmentation module and a registration module with an attention block in between. The segmentation module is composed of either one segmentation network for single-modal registration, or two parallel segmentation networks for multi-modal registration. The attention block is responsible for selectively enhancing the moving and fixed images by assigning different levels of attention to the segmented regions, depending on the requirements of the task at hand. See Fig. 1 for more details of the WSSAMNet architecture.

The revised version of WSSAMNet, referred to as WSSAMNet++, is presented in this paper. To enhance its segmentation capabilities, we substituted the original UNets [24] with more potent segmentation networks, such as Swin UNETR [25] and Multi-UNets [26]. Additionally, we tailored the attention blocks to the specific datasets we were analyzing and more rigorously tested this network.

For the experiments of inter-subject and intra-subject registration, we designed the attention block in the following manner. Initially, we utilized a Laplacian kernel to transform the predicted mask regions into contours, after which we calculated the resulting image according to the following equation 1

$$O = L \times M + I, \tag{1}$$

where I, L, M, and O are the input image, the Laplacian magnitude of the predicted mask, the predicted mask, and the output attention-driven (directed) image respectively.

In contrast to inter-subject or intra-subject registration problems, postoperative to preoperative registration presents a unique challenge. In this case, the objective is not to prioritize the segmented regions, specifically the tumor regions, as these regions are not common to both the moving and fixed volumes. To address this, a distinct attention mechanism was designed to redirect the focus of the registration network towards the shared regions. This attention mechanism differs from the one employed in inter-subject and intra-subject registration tasks. Thus for postoperative to preoperative registration, we calculated the resulting image according to 2

$$O = M \times I, \tag{2}$$

where I, M, and O are the input image, the binary predicted mask (tumor area has label 0 and rest of the brain has label 1), and the output (directed) image, respectively.

The registration module takes in a concatenation of the directed fixed and moving images, and uses it to predict the deformation field. This deformation field is then utilized to deform either the moving image, the corresponding mask, or both, so that they match the fixed image and its corresponding mask. Further details on the functionality of each component will be elaborated upon in the upcoming sections.

B. Segmentation

We conducted experiments with two distinct segmentation architectures. We now provide a detailed account of each network along with its number of parameters.

1) Swin UNETR: The segmentation model Swin UNETR [25], takes inspiration from the successful vision transformers and their variations. It aims to tackle the 3D brain tumor semantic segmentation task by converting it into a sequence. Swin UNETR employs a U-shaped network composed of a Swin Transformer as the encoder and a CNN-based decoder

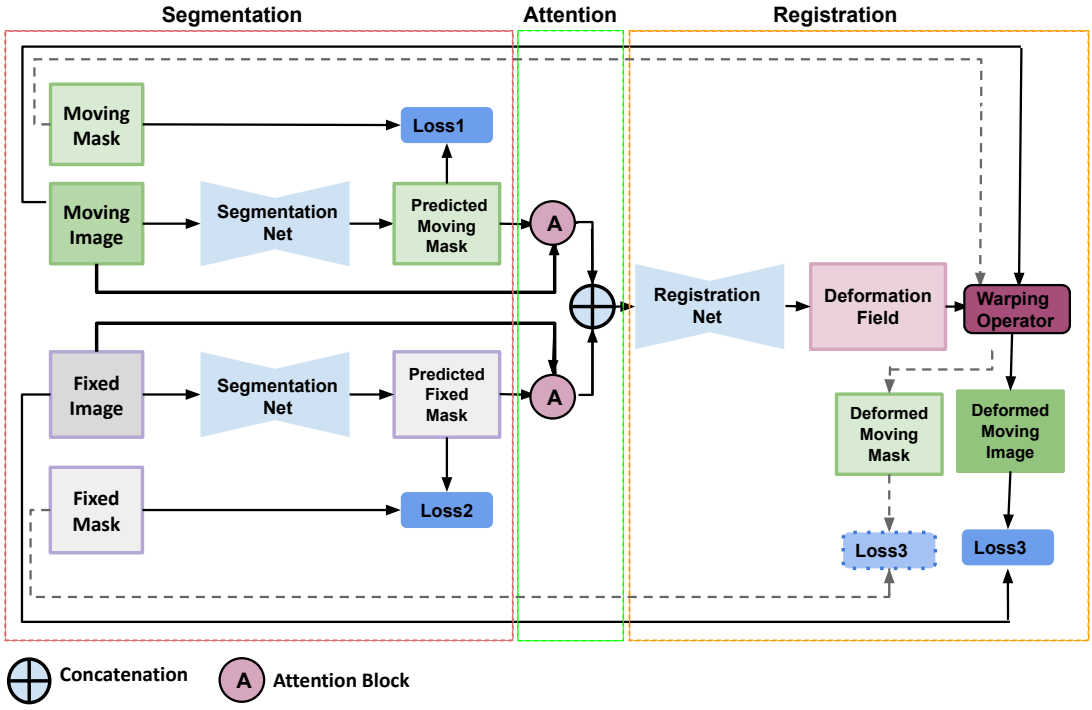

Fig. 1. Architecture of WSSAMNet comprises three parts: a segmentation network, an attention block, and a registration network.

connected at various resolutions through skip connections. In this method, the multi-modal input data is transformed into a 1D sequence of embedding, which serves as input to a hierarchical Swin Transformer acting as the encoder. The Swin Transformer encoder extracts features at five different resolutions by employing shifted windows to compute self-attention. These features are then connected to an FCNN-based decoder at each resolution through skip connections. The efficacy of Swin UNETR for multi-modal 3D brain tumor segmentation in the 2021 Multi-modal Brain Tumor Segmentation Challenge (BraTS) has also been demonstrated [27]. The overall model has 2.2 million parameters.

2) Multi-UNets: We utilized several 3D Convolutional Neural Networks, distributed independently in space, to segment the entire brain volume. Each volume has a resolution of $(224 \times 160 \times 192)$, which we divided into eight smaller non-overlapping volumes of size $(112 \times 80 \times 96)$. We trained a 3D UNet model [24] for each of the eight volumes to learn contextual information at a fixed spatial location. During inference, we combined the outputs of all eight models. The network was trained using pixel-wise cross entropy loss, and a single 3D UNet model contained 19.9 million trainable parameters. With eight such networks, the overall model has 159 million trainable parameters. The rationale for employing two distinct segmentation networks – Swin UNETR and Multi-UNets – for segmentation is that the Swin UNETR, being a complex architecture, requires a GPU with a large memory, particularly when employed for 35-label segmentation. To overcome this limitation, we introduce Multi-UNets as a viable alternative, which is better suited for GPUs with limited memory.

C. Registration

We experimented with two state-of-the-art registration networks – RegUNet [28] and TransMorph [29].

1) RegUNet: For RegUNet we used the deep learning toolkit for medical image registration (DeepReg) [28] in most of our experiments. The DeepReg toolkit consists of several modules that can be used to perform image registration tasks, including a deep neural network architecture, various loss functions, and optimization algorithms. The authors show that their approach outperforms state-of-the-art registration methods on several benchmark datasets and demonstrate the toolkit's versatility by applying it to a variety of medical imaging modalities, including MRI, CT, and ultrasound. They also provide open-source code for the DeepReg toolkit, making it widely accessible to the research community.

Our registration network follows a straightforward 3D UNet-like architecture, with the concatenated attentive source and target images serving as the input and the output being the deformation field. The network comprises three levels and does not include any skip connections. For additional details on the network's architecture, please refer to [28]. The total number of parameters is 0.37 million.

2) TransMorph: We also applied our technique to TransMorph [29], which is a state-of-the-art registration network based on its superior performance on a variety of registration problems and datasets, to demonstrate that our proposed method improves registration results regardless of the specific underlying registration network, even if it is already performing well. Unlike convolutional networks, vision transformers can account for long-range spatial relationships within an image. In recent years, vision transformers have demonstrated

state-of-the-art performance in various medical imaging tasks. Due to their substantially larger receptive field, transformers are a promising option for image registration as they can provide a more accurate understanding of the spatial correspondence between a moving and fixed image. As a hybrid Transformer-ConvNet model, TransMorph [29] utilizes these advantages to achieve volumetric medical image registration.

The initial step of the network's encoder is to divide the input volumes of the fixed and moving images into 3D patches with no overlap. Each of these patches is then flattened and treated as a token, which is projected to a feature representation by a linear projection layer. In this process, positional embedding is not used. Subsequently, patch merging and Swin Transformer blocks are applied in several consecutive stages. The decoder component is comprised of a series of upsampling and convolutional layers with a 3×3 kernel size. To incorporate skip connections, the decoder is also connected to the corresponding stages in the encoder. Finally, the network's output is the deformation field. The total number of parameters of this network is 107 million.

IV. Experiments and Results

This research investigates the benefits of directing the registration network's attention to regions of interest to improve registration accuracy. The study includes experiments on inter-subject, intra-subject, and postoperative to preoperative registration problems using single-modality and multi-modal datasets. By conducting extensive experiments with diverse datasets and registration tasks, this study highlights the significance of incorporating semantic information to enhance the registration process in medical image analysis. The proposed method effectively utilizes segmentation to guide and improve registration performance for both single-modality and multi-modal registration methods, outperforming existing approaches.

A. Datasets

Our experiments were conducted using three different datasets. The easiest challenge was posed by the OASIS dataset [30] to register healthy brains of various subjects. The challenge increased in the second unpaired dataset called Learn2Reg [31], which required multi-modal registration involving CT scans of abdominal organs and their corresponding MRI scans. Lastly, we used another challenging dataset from BraTS-Reg challenge [18] for the purpose of registering postoperative scans to pre-operative scans of patients who underwent treatment for glioma. More details about each of the datasets can be found below.

1) OASIS dataset: The OASIS dataset [30], which is used for inter-subject registration, comprises MRI data from 414 subjects. We partitioned this data into three sets: training, validation, and testing. These sets were divided in a ratio of 314:50:50, respectively. Each subject's data included original and normalized T1-weighted scans, as well as segmentation masks for different brain areas. The dataset contains three distinct types of brain segmentation: a four-label mask, a thirty-five label mask, and a twenty-four-label mask, which cover the major anatomical regions of the brain. Both the 3D T1-weighted scans and their corresponding masks have the following size: (160, 192, 224).

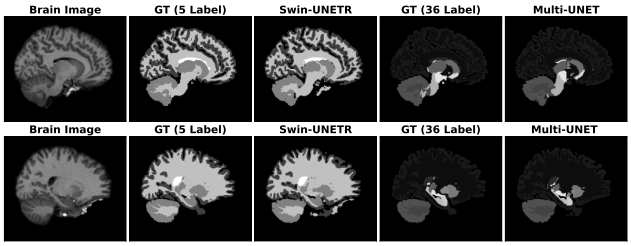

Fig. 2. OASIS dataset sample [30]: rows represent two different brain slices obtained from a subject illustrating the precision of the predicted masks compared to the ground truth (GT) masks. The Swin UNETR [25] and Multi-UNets [24] models were used for the 5-label and 36-label predictions, respectively.

Figure 2 shows two brain slices from a subject in the OASIS dataset, showcasing the accuracy of predicted masks in comparison to the ground truth masks. The Swin UNETR [25] and Multi-UNets [24] models were utilized for 4-label and 35-label predictions, respectively, revealing the quality of their performance.

2) Learn2Reg dataset: We utilized the unpaired CT-MR thorax-abdomen dataset from the Learn2reg collection [31] for intra-patient registration. Our objective was to demonstrate that segmentation aids in multi-modal registration also, where the source and target data domains differ significantly in resolution as well as visual patterns associated with each anatomical region, and the annotations are usually limited and noisy.

This dataset includes MRI and CT scans for each patient, and the goal is to align the CT scan to the MRI scan for the same patient or different patients. The dataset comprises 122 CT and a same number of MR scans. Only 16 of the CT and MR scans are paired, while 94 of each are unpaired. The unpaired data is divided into training, validation, and testing sets in a ratio of 66:20:8.

Fig. 3 contains two rows pertaining to two distinct cases, with each row comprising the CT scan, its corresponding segmentation, the MRI scan, and its corresponding segmentation. Please note that the CT and the MRI slice in each row are corresponding slices from paired CT and MRI scans.

3) BraTS-Reg and BraTS datasets: The objective proposed with the release of the BraTSReg 2022 dataset is to register follow-up MRI scans to their paired pre-operative baseline scans for 140 subjects who have undergone treatment for glioma [18]. The dataset is divided into training, validation, and testing sets in a ratio of 104:29:7. For each patient, the dataset provides pairs of pre-operative baseline and follow-up MRI brain scans along with landmarks. The multi-parametric MRI sequences at each time-point include native (T1), contrast-enhanced T1-weighted (T1-CE), T2-weighted, and T2 Fluid Attenuated Inversion Recovery (FLAIR). We only utilized the T1-CE scans in our experiments. The number of landmarks provided varies from 6 to 50 per scan, and we used them for evaluation purposes only.

It should be noted that the BraTSReg dataset does not

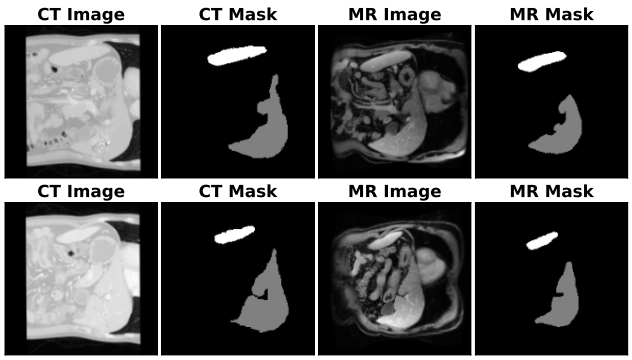

Fig. 3. Learn2Reg dataset sample [31]: Different cases (rows) from the Learn2Reg dataset are shown with their CT scan (first column), its segmentation (second column), the corresponding MRI scan (third column), and its segmentation (last column).

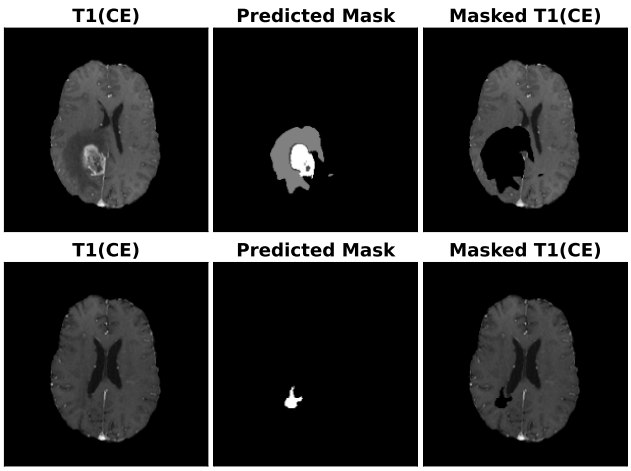

Fig. 4. BraTSReg dataset sample [18]: preoperative (top row) and postoperative (bottom row) slices of a single patient from the BraTSReg dataset are shown, and the columns represent the input slice, the predicted mask, and the masked image, respectively from left to right.

include segmentation masks. Therefore, we trained a segmentation network on the BraTS 2021 dataset [32], which contains MRI scans and corresponding tumor masks for 1251 subjects. Annotations for the GD-enhancing tumor, peritumoral edematous/invaded tissue, and necrotic tumor core for this dataset have been previously described [33] . The shape of the MRI scan and its mask in the BraTS2021 dataset is (155, 240, 240). Within our registration network, we employed preoperative and follow-up MRI scans as inputs. To concentrate our registration efforts on aligning the shared regions present in both images, we applied a masking technique to isolate the tumor regions from each scan.

Figure 4 illustrates the input slice, the predicted mask, and the masked image in three successive columns for two slices (rows). The slice in the top row is taken from a preoperative scan, while the one in the bottom row is taken from a postoperative scan. The masks were predicted using Swin UNETR trained on BraTS data and used in testing mode for the BraTSReg dataset.

B. Data preprocessing and experiments setup

For all the experiments, we used different techniques for normalizing the MRI and CT images. To normalize the MRI images, we used the min-max normalization method. On the other hand, for CT scans, we adopted a different normalization approach. First, we divided each pixel value in the CT scans by the maximum value that a pixel can take, which is usually 4095 for a 12-bit image. Then, we performed zero-centering to prevent gradient saturation, where we subtracted each pixel value from the average value of the pixels' subsample. Finally, we applied standardization to the CT scans to achieve a sample distribution with a mean of zero and a standard deviation of one.

The experimental setup for this study involves four different neural network models, namely Swin UNETR, Multi-UNets, TransMorph, and RegUNet. The optimizer used for all four models is Adam, with a learning rate of 10^{-4}. The batch sizes for the four models differ, with Swin UNETR having a batch size of 2, Multi-UNet having a batch size of 1, TransMorph having a batch size of 3, and RegUNet having a batch size of 4.

Additionally, the experiments were run on one Nvidia A100 GPU with 8 workers to ensure efficient processing of the large datasets involved. It is worth mentioning that the Multi-UNets were trained specifically on an NVIDIA GeForce RTX 2080 Ti 11 GB. These experimental parameters were chosen to strike a balance between computational efficiency and model performance, allowing for meaningful and accurate results.

C. Loss functions and evaluation metrics

For most of our experiments, we employed the Dice loss [19] as a training metric for our network. Figure 1 illustrates the different types of Dice loss we utilized. Specifically, loss1 represents the segmentation loss calculated by comparing the predicted and ground truth moving masks, loss2 is the Dice loss between the predicted and ground truth fixed masks, and loss3 represents the registration loss obtained by calculating the Dice loss between the deformed moving mask and the ground truth fixed mask. Additionally, we used a combination of normalized cross-correlation loss [12], mutual information loss [13] and bending energy loss [34] as the registration loss for the preoperative to postoperative registration network, as it was trained in an unsupervised manner.

The Dice score between the predicted mask p and the ground truth mask g is given by 3

$$\text{Dice Score} = \frac{2\sum_{i=1}^{n} p_i g_i}{\sum_{i=1}^{n} p_i + \sum_{i=1}^{n} g_i} \tag{3}$$

where n is the total number of labels. And Dice Loss $= 1 -$ Dice Score.

The local cross correlation between the fixed volume f and the moving volume m after deforming it by the deformation field Φ is given by 4 :

$$\text{cc(f,m}\circ\Phi) = \sum_{\boldsymbol{p}\in\boldsymbol{\Omega}} \frac{(\sum_{\boldsymbol{p}_i}(f(\boldsymbol{p})_i - \hat{f}(\boldsymbol{p}))([m\circ\Phi](\boldsymbol{p}_i) - [\hat{m}\circ\Phi](\boldsymbol{p})))^2}{(\sum_{\boldsymbol{p}_i}(f(\boldsymbol{p}_i) - \hat{f}(\boldsymbol{p}))^2)(\sum_{\boldsymbol{p}_i}([m\circ\Phi](\boldsymbol{p}_i) - [\hat{m}\circ\Phi](\boldsymbol{p}))^2)} \tag{4}$$

where $\hat{f}(\boldsymbol{p})$ and $[\hat{m}\circ\Phi](\boldsymbol{p})$ denote local mean intensity images.

The mutual information between f and $(m \circ \Phi)$ is given by 5:

$$I(f, m \circ \Phi) = \sum_{a \in f, b \in m \circ \Phi} p(a,b) log \left(\frac{p(a,b)}{p(a)p(b)} \right) \quad (5)$$

According to [35] the displacement field is smooth if it does not have severe hops. That is, it is smooth when there is a gradual change in the direction and the magnitude in a neighborhood. To ensure this, we use the L2 norm of the Laplacian of the deformation field as a loss as given in equation 6.

$$L_{smooth}(\Phi) = \sum_{\boldsymbol{p} \in \Omega} ||\nabla \boldsymbol{u}(\boldsymbol{p})||, \quad (6)$$

where $\nabla \boldsymbol{u}(\boldsymbol{p}) = (\frac{\partial \boldsymbol{u}(\boldsymbol{p})}{\partial x}, \frac{\partial \boldsymbol{u}(\boldsymbol{p})}{\partial y}, \frac{\partial \boldsymbol{u}(\boldsymbol{p})}{\partial z})$, and $\frac{\partial \boldsymbol{u}(\boldsymbol{p})}{\partial x} \approx \boldsymbol{u}(p_x + 1, p_y, p_z) - \boldsymbol{u}(p_x, p_y, p_z)$.

For the inter- and intra-subject registration tasks, we utilized the Dice score as in 3 as the evaluation metric, whereas for the postoperative to preoperative registration task, we employed the median absolute error (MAE) as the evaluation metric.

The MAE measures the registration error between the follow-up scan (F) and the preoperative scan (B) for a pair of scans (p) using the median absolute error of the landmarks [18], which is given by:

$$MAE = Median_{l \in L}(|x_l^B - \hat{x_l}^B|) \quad (7)$$

where x_l^B, x_l^F are the coordinates of corresponding landmarks of $l \in L$ the set of landmarks identified in both B and F.

D. Inter-subject registration

Inter-subject registration is a crucial technique in medical imaging that involves aligning two or more images of different subjects. It is an important step in many clinical applications such as treatment planning, image-guided therapy, and disease diagnosis. The process involves finding a transformation that aligns the anatomical structures of interest in one image to those in another image, while accounting for differences in anatomy, position, and orientation. Inter-patient registration is challenging due to the wide variability in anatomical structures among different patients, making it difficult to find a universal solution.

The OASIS dataset was utilized in our experiments, wherein we conducted three separate tests to demonstrate the significance of utilizing segmentation in a conscious manner. Our first step was to select two subjects at random and utilizing their moving and fixed T1CE (T1 contrast enhanced) scans as inputs. Secondly, we used WSSAMNet++ to segment the masks of the moving and fixed images, and then proceeded to use these masks in three different ways: directly as input to RegUNet to predict the deformation field, concatenating the masks with the images and then presenting them as input to RegUNet, and using the predicted masks to direct the attention of the registration network towards specific areas of interest.

Our results demonstrate that solely using the masks or concatenating them with the images resulted in a deterioration of performance compared to the performance of deep learning methods that utilized the images as inputs. However, directing the attention of the registration network towards specific regions of interest enhanced the performance of the registration network. Check Table I.

TABLE I
RESULTS OF INTER-SUBJECT REGISTRATION

Segmentation Network	Registration Network	Registration Input	Dice Score
None	RegUNet	Original Images	0.736±0.018
Swin UNETR	RegUNet	Predicted Masks	0.724±0.021
Swin UNETR	RegUNet	CAT(Images, Masks)	0.731±0.016
Swin UNETR	RegUNet	Images with Attention	**0.750±0.013**

CAT() stands for concatenation of moving and fixed images and their corresponding masks. In all of these experiments we used four-label masks. Each four-label mask in OASIS data contains the following regions: the cortex, the subcortical gray matter, the white matter, and the CSF.

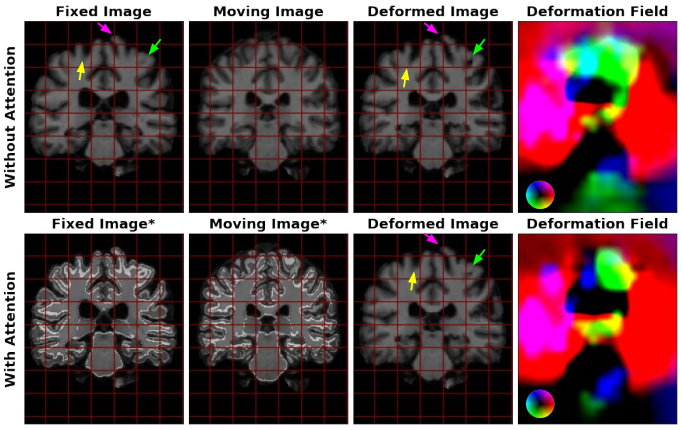

Fig. 5. RegUNet (top row) compared to WSSAMNet++ (bottom row) on OASIS dataset using a four-label mask: Asterisks indicate fixed and moving image with the Laplacian magnitude of the segmentation mask overlaid, and the red grid lines are overlaid to aid visual comparison along with color-matched arrows to highlight corresponding visual features.

Fig. 5 shows the results of RegUNet in the first row against the results of WSSAMNet++ in the second row when using four-label masks. Similarly, Fig. 6 shows the results of TransMorph against WSSAMNet using four-label masks. While Fig. 7 shows the results of TransMorph against WSSAMNet++ when using thirty five-label masks. WSSAMNet++ outperforms RegUNet and TransMorph in deforming the moving regions to achieve a higher level of semantic similarity with the corresponding regions in the fixed image, as evident in all the indicated areas across the figures.

Table II presents a comparison between WSSAMNet++ and TransMorph, indicating that WSSAMNet++ outperforms TransMorph in all experiments when utilizing both four-label (five-label if background is counted) and thirty-five-label (thirty-six-label if background is counted) predicted masks to generate Laplacian images for the registration network of WSSAMNet++. Furthermore, the table highlights that lower label counts result in better registration outcomes.

E. Intra-subject registration

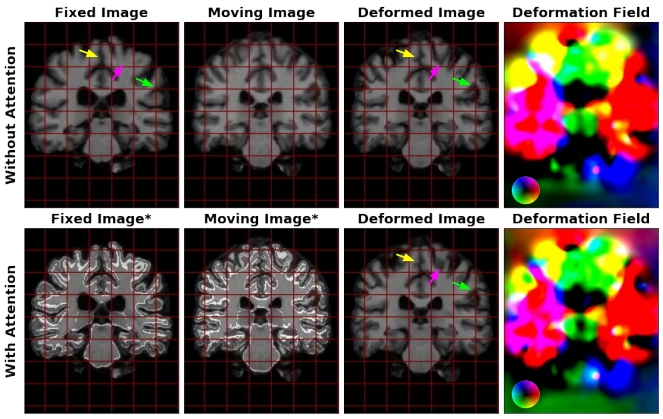

Fig. 6. TransMorph (top row) compared to WSSAMNet++ (bottom row) on OASIS dataset using a four-label mask: Asterisks indicate fixed and moving image with the Laplacian magnitude of the segmentation mask overlaid, and the red grid lines are overlaid to aid visual comparison along with color-matched arrows to highlight corresponding visual features.

TABLE II
COMPARISON BETWEEN WSSAMNET++ AND TRANSMORPH

Network	Number of Labels	Registration's Inputs	Dice Score
TransMorph	4-label Mask	Original Images	0.899±0.006
WSSAMNet++	4-label Mask	Laplacian Images	**0.901±0.004**
TransMorph	35-label Mask	Original Images	0.849±0.012
WSSAMNet++	35-label Mask	Laplacian Images	**0.852±0.011**

A comparison between the performance of WSSAMNet++ and TransMorph when using predicted masks of both four labels and thirty-five labels of OASIS Data.

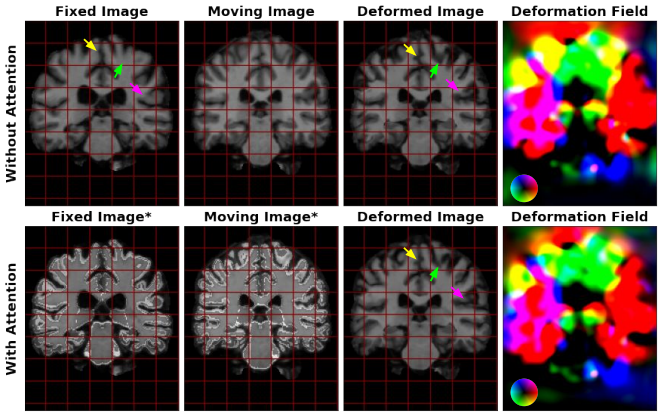

Fig. 7. TransMorph (top row) compared to WSSAMNet++ (bottom row) on OASIS dataset using a 35-label mask: Asterisks indicate fixed and moving image with the Laplacian magnitude of the segmentation mask overlaid, and the red grid lines are overlaid to aid visual comparison along with color-matched arrows to highlight corresponding visual features.

Intra-subject registration of CT to MRI of abdominal organs is a crucial technique in medical imaging that enables the integration of information from both modalities. This integration can provide valuable insights into the underlying anatomy and pathology of the abdominal organs. The process involves registering a CT image of the abdomen to an MRI image of the same patient, allowing for the accurate alignment of anatomical structures between the two modalities. This technique is particularly useful in cases where CT and MRI scans are acquired for the same patient at different times or for different clinical purposes. Intra-subject registration is a challenging task due to the differences in image acquisition parameters and contrast mechanisms between the two modalities.

TABLE III
RESULTS OF INTRA-SUBJECT REGISTRATION

Network	Registration's Inputs	Dice Score
RegUNet	Images	0.167±0.098
WSSAMNet++	Images with Attention	**0.241±0.074**

Dice score was computed between the four-label deformed moving mask and fixed mask of unpaired scans. Each mask of Learn2Reg data contains the following regions: the liver, the spleen, the left kidney, and the right kidney.

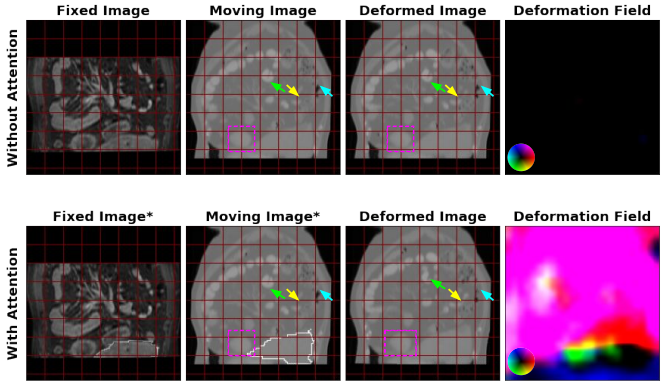

Fig. 8. RegUNet (top row) compared to WSSAMNet++ (bottom row) on Learn2Reg multi-modal dataset: Asterisks indicate fixed and moving image with the Laplacian magnitude of the segmentation mask overlaid, and the red grid lines are overlaid to aid visual comparison along with color-matched arrows to highlight corresponding visual features.

The Learn2Reg dataset was used in our experiments, wherein we employed WSSAMNet++ to identify the regions of interest in both MRI and CT scans, and subsequently directed the attention of RegUNet towards these regions. Our findings demonstrate a marked improvement in the performance of RegUNet compared to its performance when the original images were used without attention. See Table III

Fig. 8 shows the results of RegUNet in the first row against the results of WSSAMNet++ in the second row. WSSAMNet++ deforms the moving image effectively, even for large deformations, as evident from the deformation field. On the other hand, the deformation field of RegUNet exhibits high magnitude in the background, where no deformations should occur. Arrows highlight the disparities between the moving image and the output of WSSAMNet, whereas the moving image and output of RegUNet appear identical.

F. Postoperative to preoperative registration

Postoperative to preoperative registration of MRI scans in patients treated for glioma is a critical aspect of clinical care. Gliomas are malignant brain tumors that can be challenging to remove completely without causing damage to surrounding healthy tissues. Hence, postoperative MRI scans are crucial for evaluating the extent of tumor resection and assessing the need for further treatment. However, to accurately compare the postoperative images with the preoperative images and to determine the extent of resection, registration of these images is necessary.

TABLE IV
RESULTS OF POSTOPERATIVE TO PREOPERATIVE REGISTRATION

Evaluation Metric	Images as Input	Images with Attention as Input
Mean of Mean AE	1.587	**1.581**
Mean of Median AE	1.307	**1.293**
Median of Median AE	**1.494**	**1.494**
Mean of Median AE	1.698	**1.616**

AE stands for the absolute error between the registered landmarks and the ground truth landmarks.

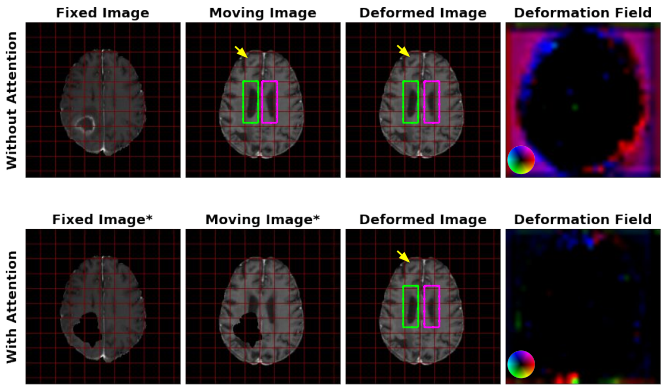

Fig. 9. RegUNet (top row) compared to WSSAMNet++ (bottom row) on BraTSReg dataset: Asterisks indicate fixed and moving image with the Laplacian magnitude of the segmentation mask overlaid, and the red grid lines are overlaid to aid visual comparison along with color-matched boxes to highlight corresponding visual features where the missing tissue is better shrunk by WSSAMNET++.

The postoperative to preoperative registration enables the integration of information from both scans, providing a more comprehensive view of the tumor and surrounding tissues. Accurate registration can help clinicians in determining the efficacy of the surgical procedure, planning further treatments, and assessing patient prognosis. Therefore, postoperative to preoperative registration of MRI scans in patients treated for glioma is crucial for improving patient outcomes and optimizing treatment strategies.

Our experiments were conducted using the BraTSReg dataset. However, as this dataset did not contain any masks, we trained WSSAMNet++'s segmentation network on the BraTS dataset to segment the tumor from the preoperative images. Once we obtained the tumor masks from the preoperative and postoperative images, we masked the corresponding areas in both of the MRI scans. This was done to address the challenge of RegUNet learning the deformation field, as the existence of the tumor in the preoperative scan and the cavity and/or tumor recurrence in the postoperative scan did not correspond to each other, which could potentially hinder the learning process. Therefore, we believed that masking these regions from the preoperative and postoperative scans would improve the accuracy of RegUNet's deformation field prediction.

The outcomes of our study demonstrate a noticeable enhancement in RegUNet's performance when utilizing the segmentation information and directing the network's attention towards the shared regions between the two scans,see Table IV.

Fig. 9 shows the results of RegUNet in the first row against the results of WSSAMNet++ in the second row. It is worth noting that the deformation field observed in RegUNet's output appears to have high magnitude in the background where the deformation should be zero, in contrast to the deformation field of WSSAMNet++

V. CONCLUSION

We proposed the use of segmentation in improving the medical image registration process. Through the introduction of the Semantic Attentive Medical Image Registration Network (WSSAMNet++), we have demonstrated the efficacy of using segmentation to guide the registration process, resulting in improved performance across various registration tasks and datasets. Additionally, the study demonstrates that our method consistently improves registration results across different backbones. Our study highlights the significant improvements achieved by our proposed method in multi-modal registration, while also demonstrating that an increased number of segmented regions does not necessarily lead to better registration performance. Overall, our work contributes to the development of more accurate and efficient medical image registration techniques, with the potential to benefit clinical practice and patient outcomes. The study's findings can potentially contribute to the development of new approaches for medical image analysis, resulting in improved accuracy for medical diagnosis and treatment.

REFERENCES

[1] W. E. L. Grimson, R. Kikinis, F. A. Jolesz, and P. M. Black, "Image-guided surgery," *Scientific American*, vol. 280, no. 6, pp. 62–69, 1999.
[2] S. A. Nasser, N. C. Kurian, S. Shamsi, M. Meena, and A. Sethi, "Ws-samnet: Weakly supervised semantic attentive medical image registration network," *arXiv preprint arXiv:2203.07114*, 2022.
[3] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim, "Elastix: a toolbox for intensity-based medical image registration," *IEEE transactions on medical imaging*, vol. 29, no. 1, pp. 196–205, 2009.
[4] T. X. Lin and H. H. Chang, "Medical image registration based on an improved ant colony optimization algorithm," *Int J Pharma Med Biol Sci*, vol. 5, no. 1, pp. 17–22, 2016.
[5] M. Modat, J. McClelland, and S. Ourselin, "Lung registration using the niftyreg package," *Medical image analysis for the clinic-a grand Challenge*, vol. 2010, pp. 33–42, 2010.
[6] G. Lowe, "Sift-the scale invariant feature transform," *Int. J*, vol. 2, no. 91-110, p. 2, 2004.
[7] X. Teng, Y. Chen, Y. Zhang, and L. Ren, "Respiratory deformation registration in 4d-ct/cone beam ct using deep learning," *Quantitative Imaging in Medicine and Surgery*, vol. 11, no. 2, p. 737, 2021.
[8] X. Yang, R. Kwitt, M. Styner, and M. Niethammer, "Quicksilver: Fast predictive image registration–a deep learning approach," *NeuroImage*, vol. 158, pp. 378–396, 2017.
[9] H. Sokooti, B. de Vos, F. Berendsen, M. Ghafoorian, S. Yousefi, B. P. Lelieveldt, I. Isgum, and M. Staring, "3d convolutional neural networks image registration based on efficient supervised learning from artificial deformations," *arXiv preprint arXiv:1908.10235*, 2019.
[10] K. Yan, J. Cai, D. Jin, S. Miao, D. Guo, A. P. Harrison, Y. Tang, J. Xiao, J. Lu, and L. Lu, "Sam: Self-supervised learning of pixel-wise anatomical embeddings in radiological images," *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2658–2669, 2022.
[11] E. Chee and Z. Wu, "Airnet: Self-supervised affine registration for 3d medical images using neural networks," *arXiv preprint arXiv:1810.02583*, 2018.
[12] J.-C. Yoo and T. H. Han, "Fast normalized cross-correlation," *Circuits, systems and signal processing*, vol. 28, pp. 819–843, 2009.
[13] J. X. Ji, H. Pan, and Z.-P. Liang, "Further analysis of interpolation effects in mutual information-based image registration," *IEEE Transactions on Medical Imaging*, vol. 22, no. 9, pp. 1131–1140, 2003.

[14] D. Mahapatra, B. Antony, S. Sedai, and R. Garnavi, "Deformable medical image registration using generative adversarial networks," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 1449–1453.

[15] C. K. Guo, "Multi-modal image registration with unsupervised deep learning," Ph.D. dissertation, Massachusetts Institute of Technology, 2019.

[16] J. Chen, Y. He, E. C. Frey, Y. Li, and Y. Du, "Vit-v-net: Vision transformer for unsupervised volumetric medical image registration," *arXiv preprint arXiv:2104.06468*, 2021.

[17] Y. Hu, M. Modat, E. Gibson, W. Li, N. Ghavami, E. Bonmati, G. Wang, S. Bandula, C. M. Moore, M. Emberton *et al.*, "Weakly-supervised convolutional neural networks for multimodal image registration," *Medical image analysis*, vol. 49, pp. 1–13, 2018.

[18] B. Baheti, D. Waldmannstetter, S. Chakrabarty, H. Akbari, M. Bilello, B. Wiestler, J. Schwarting, E. Calabrese, J. Rudie, S. Abidi *et al.*, "The brain tumor sequence registration challenge: establishing correspondence between pre-operative and follow-up mri scans of diffuse glioma patients," *arXiv preprint arXiv:2112.06979*, 2021.

[19] Y. Zhu, Z. Zhou Sr, G. Liao Sr, and K. Yuan, "New loss functions for medical image registration based on voxelmorph," in *Medical Imaging 2020: Image Processing*, vol. 11313. SPIE, 2020, pp. 596–603.

[20] M. C. Lee, O. Oktay, A. Schuh, M. Schaap, and B. Glocker, "Image-and-spatial transformer networks for structure-guided image registration," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer, 2019, pp. 337–345.

[21] N. Gunnarsson, J. Sjölund, and T. B. Schön, "Learning a deformable registration pyramid," in *Segmentation, Classification, and Registration of Multi-modality Medical Imaging Data: MICCAI 2020 Challenges, ABCs 2020, L2R 2020, TN-SCUI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 23*. Springer, 2021, pp. 80–86.

[22] M. Hoffmann, B. Billot, D. N. Greve, J. E. Iglesias, B. Fischl, and A. V. Dalca, "Synthmorph: learning contrast-invariant registration without acquired images," *IEEE transactions on medical imaging*, vol. 41, no. 3, pp. 543–558, 2021.

[23] L. A. Schwarz, "Non-rigid registration using free-form deformations," *Technische Universität München*, vol. 6, no. 8, 2007.

[24] N. Navab, J. Hornegger, W. M. Wells, and A. Frangi, *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III*. Springer, 2015, vol. 9351.

[25] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I*. Springer, 2022, pp. 272–284.

[26] Y. Huo, Z. Xu, K. Aboud, P. Parvathaneni, S. Bao, C. Bermudez, S. M. Resnick, L. E. Cutting, and B. A. Landman, "Spatially localized atlas network tiles enables 3d whole brain segmentation from limited data," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III 11*. Springer, 2018, pp. 698–705.

[27] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati *et al.*, "The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification," *arXiv preprint arXiv:2107.02314*, 2021.

[28] Y. Fu, N. M. Brown, S. U. Saeed, A. Casamitjana, Z. Baum, R. Delaunay, Q. Yang, A. Grimwood, Z. Min, S. B. Blumberg *et al.*, "Deepreg: a deep learning toolkit for medical image registration," *arXiv preprint arXiv:2011.02580*, 2020.

[29] J. Chen, E. C. Frey, Y. He, W. P. Segars, Y. Li, and Y. Du, "Transmorph: Transformer for unsupervised medical image registration," *Medical image analysis*, vol. 82, p. 102615, 2022.

[30] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults," *Journal of cognitive neuroscience*, vol. 19, no. 9, pp. 1498–1507, 2007.

[31] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle *et al.*, "The cancer imaging archive (tcia): maintaining and operating a public information repository," *Journal of digital imaging*, vol. 26, pp. 1045–1057, 2013.

[32] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," *arXiv preprint arXiv:1811.02629*, 2018.

[33] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[34] B. Fischer and J. Modersitzki, "Curvature based image registration," *Journal of Mathematical Imaging and Vision*, vol. 18, pp. 81–85, 2003.

[35] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain," *Medical image analysis*, vol. 12, no. 1, pp. 26–41, 2008.