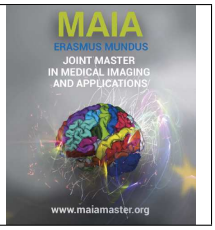
Medical Imaging and Applications

Master Thesis, September 2020



Brain Image Analysis using Spatially Localized Neural Networks

Ahmed Gouda, Corné Hoogendoorn

Canon Medical Research Europe Ltd., Edinburgh, United Kingdom

Abstract

Parcellation of brain MRI is a powerful tools for characterization of normal and pathological tissues. Recently, three-dimensional Deep Convolution Neural Network (CNN) have quickly turned into the state-of-the-art in a variety of brain image parcellation applications. However, it brings with it various challenges. Specifically, these are posed by GPU memory limitations for high resolution volumes, the complexity of the segmentation task and low numbers of manually annotated training data. These challenges have been addressed in this thesis through two main brain MRI segmentation approaches. The first approach proposes an alternative strategy for Spatially Localized Atlas Network Tiles (SLANT). It simplified the brain segmentation task into registered localized sub-volumes, leveraging the brains symmetry and voxel spatial locations. The scarcity of annotated data is addressed through another semi-supervised proposed approach by combining a feature matching GAN model with SLANT. The second approach is an implementation for unsupervised deep learning approach combining atlas base segmentation and deep learning-based registration, without the need for any annotated volumes during training. The proposed models are trained on a collection of datasets from various sources to classify 31 anatomical structures. The SLANT approach using UNet model achieves the best segmentation results, reaching percentage Dice similarity scores ranging from 79.4% to 96.6% on selected parcels of interest for neurodegenerative diseases.

Keywords:
Brain Segmentation, Network Tiles, Deep CNN, GAN, Atlas Segmentation

1. Introduction

Automated imaging segmentation has opened new horizons in brain imaging applications, as it is an essential stage for measuring and visualizing anatomical structures of tissue-volumes derived from Magnetic Resonance Images (MRI). MRI quantitative and qualitative analysis has been used extensively for analysis of brain disorders, which helped clinical specialists to diagnose, monitor progression and therapeutic response for various neurodegenerative and neurodevelopmental disorders, tumors and psychiatric disorders. Moreover, segmentation is used extensively in intervention planning and guidance.

Delivering critical information about the shapes and volumes of brain structures is a very challenging segmentation tasks. Manual delineation for Brain lobes provides very precise brain parcels but it is time-consuming, complex and a lack of reproducibility process. Enormous progression in brain MR imaging has contributed to generating high quality MR volumes that makes manual delineation not feasible for clinical use. Consequently, many computerized-based segmentation algorithms, both semi-automated and fully automated, have been proposed to facilitate the delineation process.

Semi-automated segmentation requires medical specialist intervention to guide initialization and/or interaction. In interactive methods, the generated labels after non-precise computerized segmentation are corrected manually. By contrast, manually initialized methods require manually initialized seed points or contour that roughly represents the boundary of a target brain structure. Manually initialized methods can be divided into two primary sub-categories which are region-based and boundary-based. In region-based approaches, each voxel is assigned to membership according to homogeneity of the adjacent voxels, as in region growing and merging algorithm (Zhu and Yuille,

1996). Boundary-based approaches attempt to deform the initialized boundaries seeds around the objects by minimizing the energy function that measures the variation in gradient features near to the boundary, such as snake and balloon algorithms (Kass et al., 1988), (Staib and Duncan, 1992), (McInemey and Terzopoulos, 1999). Although semi-automated segmentation approaches facilitates the delineation process, they have been deployed in small-scale medical applications.

Fully-automated segmentation is the preferred technique in medical imaging fields, as it does not require human intervention through the segmentation stages, and it is therefore easy to be deployed in clinical application. Classical fully-automated unsupervised clustering algorithms such Fuzzy c-mean (Zhang and Chen, 2004), k-mean (Dhanachandra et al., 2015) and Expectation Maximization (EM) (Zhang et al., 2001) have been widely used for MRI brain segmentation. These algorithms are effective to classify a group of tissues with similar pixels without accounting the spatial location. Hence, they have been used to segment the main brain tissue classes with significant intensity differences, which are White Matter (WM), Gray Matter (GM), and CerebroSpinal Fluid (CSF).

Supervised segmentation approaches are applied to segment some specific brain anatomical structures, steered by a model of the shape and/or appearance of these structures like Active Shape Model-based approaches (Van Ginneken et al., 2002), and level sets segmentation based approaches (Baillard et al., 2001) (Wang et al., 2014).

Image artifacts such as bias field and partial volume effects present important challenges for fully-automated segmentation due to the variety of anatomical brain structures that may share the same tissue contrast. Therefore, probabilistic atlas-based (Aljabar et al., 2009) segmentation algorithms are widely used as they exploit prior anatomical information to make the segmentation task more robust. In this approach, previously delineated labels for reference MRI images (atlases) are manipulated as a prior knowledge to segment target image. The image segmentation problem is cast as a registration problem by registering the reference atlas images to the domain of the target image. The relevant atlas labels are then propagated onto the target image. Single-atlas (Guimond et al., 2000) (Wu et al., 2007) is the basic framework in atlas-based segmentation approaches, as it uses single atlas image. However, its performance degrades in high anatomy variation between the atlas image and the target image. Therefore, multi-atlas (Rohlfing et al., 2004) (Heckemann et al., 2006) segmentation approaches address this problem by registering multiple atlases images, while the conflict between propagated label are harmonized using multi-atlas label fusion techniques (Warfield et al., 2004) (Heckemann et al., 2006) (Wang et al., 2012) (Asman and Landman, 2013) (Iglesias and Sabuncu, 2015). The different atlas based segmentation approaches have been regarded as robust brain volume segmentation standards, and they are still used in many recent medical parcellation tools (Mikhael and Pernet, 2019).

Recently, deep Convolutional Neural Networks (CNN) approaches have been deployed in large-scale for computer-aided medical imaging systems. Recent advances in semantic segmentation using three-dimensional kernels have enabled to segment three-dimensional brain structure. Though semantic segmentation achieves state-of-the-art volume segmentation results. It includes specific challenges that need to be addressed, such as the scarcity of labelled data, the high class imbalance found in the ground truth labels and the memory limitation problems for three-dimensional images. In this thesis works, we compare three recent MRI volume segmentation approaches to analyze brain structures based on supervised, semi-supervised and unsupervised learning methods. These approaches employ CNN segmentation and registration models to leveraging some advantages of regionalized network specialization while mitigating the memory limitations for high resolution images.

2. State of the art

In the last decade, deep CNN have outperformed other machine learning approaches in many visual recognition tasks (Bengio et al., 2013). Although convolutional networks have already existed for a long time (Lawrence et al., 1997) (LeCun, 1998), their success was restricted due to implementation scale of the networks size which is limited by he lack of computational power, and the complexity of the problem that could be addressed. The increased availability and power of GPU technology allowed the applicability of deep CNNs for large scale medical imaging applications.

The state-of-the-art CNN models for supervised image segmentation are variants of conventional encoder-decoder architecture like UNet (Ronneberger et al., 2015) (Çiçek et al., 2016) and V-Net (Milletari et al., 2016). In the encoder part, the input image is down-sampled into the latent space using strided convolution and max pooling layers. In the decoder part, the compressed image is up-sampled and concatenated to the same level encoding layers via skip connections, in order to make the decoder output follow the spatial structure of the input. These skip connections constrain the reconstruction process at the same-scale feature maps of the encoder and decoder layers.

An alternative semi-supervised learning approach utilizes Generative Adversarial Networks (GANs) to use a very few annotated training examples (Mondal et al., 2018). The segmentation model is trained with labeled and unlabeled scans by extracting few-shot patches from these volumes. The adversarial networks consists of a generator and discriminator. The generator network

tries to produce realistic fake patches, while the discriminator tries to distinguish the generated fake patches from the true patches.

Another recent unsupervised volume segmentation approach combines a conventional Bayesian probabilistic atlas-based segmentation with deep learning (Dalca et al., 2019). This approach comprises a deformable medical image registration framework using VoxelMorph (Balakrishnan et al., 2018), a probabilistic atlas and the global statistics of image intensity classes to efficiently estimate the deformation field and the scan-specific likelihood intensity parameters for the input image. One of the major advantage of this approach that the segmentation model does not require any ground truth data during training.

The basic technique to apply a full volume brain segmentation is to fit the complete MRI volume to a 3D CNN. Despite the decent results of this technique, it faces a GPU memory limitation for high resolution 3D brain volumes. Therefore, other approaches have been proposed to solve this problem by performing 2D slice segmentation of the 3D volume (Dong et al., 2017). The predicted output for each slice separately is then combined into a volume. This approach may have limited accuracy because it does not consider the inter-slice information between the neighboring slices. Hence, different 2.5D[1] segmentation approaches (Roth et al., 2014) (Angermann and Haltmeier, 2019) have been proposed to address the missing information in 2D segmentation and memory limitations of 3D segmentation. Although the 2.5D approaches can provide good segmentation results for some medical case studies, it is not standard technique to express volume images.

Patch-based segmentation is one of proposed solutions to deal with memory and computational requirements. In this approach, the brain region of interest is divided into similar-sized 3D overlapped sub-regions (patches). All the generated patches are then used to train a 3D CNN model. The testing volumes are predicted through non-overlapped 3D patches that cover the whole brain region. This approach also addresses the problem of labeled database limitation through few-shot learning (Fei-Fei et al., 2006). However it is not robust to segment a high number of anatomical parcels for full volume brain structure, since each patch will cover only a subset of brain regions, and this can seriously exacerbate the class imbalance problem.

Spatially Localized Atlas Network Tiles (SLANT) (Huo et al., 2018) (Huo et al., 2019) approach addresses the problem of limited GPU memory and simplifies the complex problem of high number of anatomical labels segmentation into simpler problems, better suited to limited training data. In this approach, the issue of arbitrary per-patch coverage of the brain regions is addressed by registering the training and testing volumes into a standard template. Then, the whole volume is divided into overlapped fixed sub-volumes (tiles), each one being processed by a different UNet.

[1] 2.5D is a general term for methods that conceptually lie somewhere between 2D and 3D.

3. Material and methods

The systems pipelines for this work are based on two existing systems. The first pipeline system proposes an alternative strategy for SLANT approach. This approach comprises an implementation for two sub-approaches using UNet model as a supervised segmentation, and feature matching GAN model as semi-supervised segmentation. The second pipeline system builds on unsupervised volume segmentation approach, combining VoxelMorph registration network with Bayesian probabilistic atlas.

3.1. Dataset Description and Pre-processing

The input images for these pipeline systems are collected from various resources with different isotropic spatial resolution for single modality MRI T1 weighted brain scan. The dataset is composed of raw clinical MRI brain images from different scanners and MRI brain datasets from different challenges. Whole dataset is divided with fixed proportion into Training set and Testing set, as shown in Table 1.

Dataset Name	Training Set	Testing Set
BGM Atlas	64	19
ADNI MCI	34	10
MRICloud	33	10
MICCAI 2012 MR	25	8
ADNI AD	21	7
Volunteers (Canon)	19	6
IBSR MR	14	4
Total	210	64

Table 1: Number of training and testing T1w Brain MRI volumes per dataset (Wu et al., 2016) (de Vent et al., 2016) (Mori et al., 2016) (Landman and Warfield, 2012) (Frazier et al., 2007).

In order to unify all disparate scans, they are affinely registered using Elastix toolbox (Klein et al., 2009) to MNI ICBM 152 space (Lancaster et al., 2007). Then, N4ITK was applied to suppress the bias field noise (Tustison et al., 2010). Since the acquired scans from MRI devices are non-scaled, the scans intensities are not harmonized over different scanners, and even different scans across the same scanner. Therefore, each scan is normalized by truncating the intensities outside the percentile range 5% to 95%. This harmonization technique eliminate the outliers intensities, and stretches the major brain intensities information over the histogram. Furthermore, it is a relatively simple approach in comparison to existing approaches (Madabhushi and Udupa, 2006) (Schaap et al., 2009) (Simkó et al., 2019).

3.2. Ground Truth Generation

The ground truth labels have been generated using brain parcellation application (Murphy et al., 2014). It was developed based on traditional image analysis techniques and comprises five stages. In the first stage, the left-right volume direction is rotated towards the mid-sagittal plane. In the second stage, 99 atlas volumes are affinely registered and propagated to align the input volume. In the following stage, mutual information metric is computed be between the registered atlases and input volume in order to select the best aligned atlases to proceed to the fourth stage. Then, the selected atlases are non-rigidly registered to the input volume space, and the relevant labels are propagated using the affine and non-rigid deformation fields. In the last stage, the priory probability distribution are generated over the structure labels, and they are refined using the EM algorithm for final assignment.

This brain parcellation application generates 290 brain parcels including the background region. The parcellation protocol defines five levels of increasing refinement. The first level includes main divisions of the forebrain, midbrain and hindbrain, as well as CSF and the skull, while the fifth level include a detailed brain classes that was generated by the parcellation application, as shown in Table 2. Merging the left and right counterparts into a single label leverages the brain symmetry and can lead to greater memory savings.

Parcellation Level	Without L/R Merging	With L/R Merging
Level 1	9	8
Level 2	23	14
Level 3	57	31
Level 4	139	72
Level 5	290	149

Table 2: Number of ground truth parcels for each level, with and without left-side and the right-side labels merging.

3.3. Brain Atlas Generation

VoxelMorph based atlas segmentation approach requires generating the probabilistic atlas priors for all brain parcels. As illustrated in Figure 1, an affine registration followed by B-splines registration have been preformed using Elastix toolbox to propagate the training volumes to the MNI ICBM 152 space. Then, bias field noise was removed using N4ITK. Using the brain parcellation application, left-side and right-side combined labels were generated for L3 parcels. The prior probabilities of observing the propagated parcels for the training volumes are firstly computed. Afterwards, the left-side and right-side prior probabilities for the axial direction are mirrored and combined to generate symmetric probabilistic atlas.

3.4. Spatially Localized Atlas Network Tiles (SLANT)

The proposed SLANT pipeline system is shown in Figure 2. This pipeline is fed with the pre-processed volumes in the MNI space, and the corresponding generated ground truth.

3.4.1. Network Tiles

The brain region of interest of the MNI space is constrained within a bounding-box (160, 192, 160). The entire bounding box region is covered by $k_x \times k_y \times k_z$ equally spaced, equally sized 3D tiles, extending $d\%$ of the bounding box size along each side. The amount of overlap between tiles can be varied through combinations of k and d.

In the original SLANT system (Huo et al., 2018) (Huo et al., 2019), each tile subspace trained with specific CNN model using 3D UNet architecture. We propose exploiting the symmetry of the brain across the midsagittal plane. The right-side tiles are horizontally mirrored and augmented with the left-side tiles. However, medial common tiles are horizontally mirrored and augmented with themselves to keep number of training cases balanced as in the side tiles. This technique boosts the accuracy, and reduces the number of the trainable tiles.

3.4.2. UNet Network Model

Besides dedicating a UNet model to each individual tile, another technique is proposed in this research by training all the SLANT tiles using a single UNet model. The 3D coordinates for each voxel can be added as

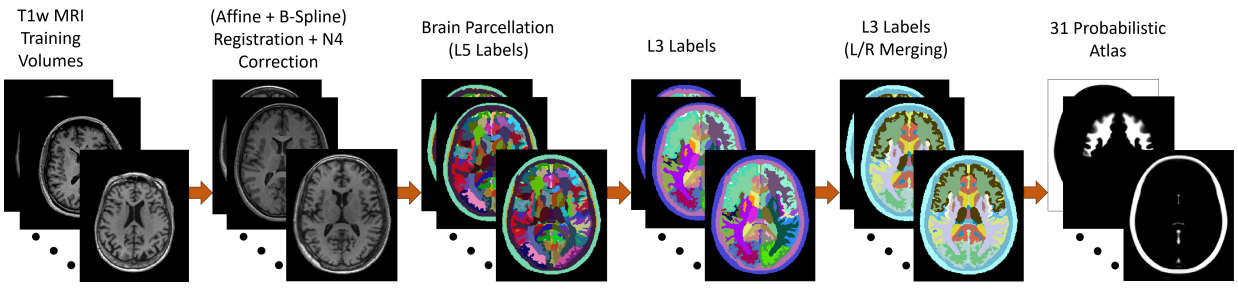


Figure 1: Probabilistic atlas generation for 31 brain parcels.

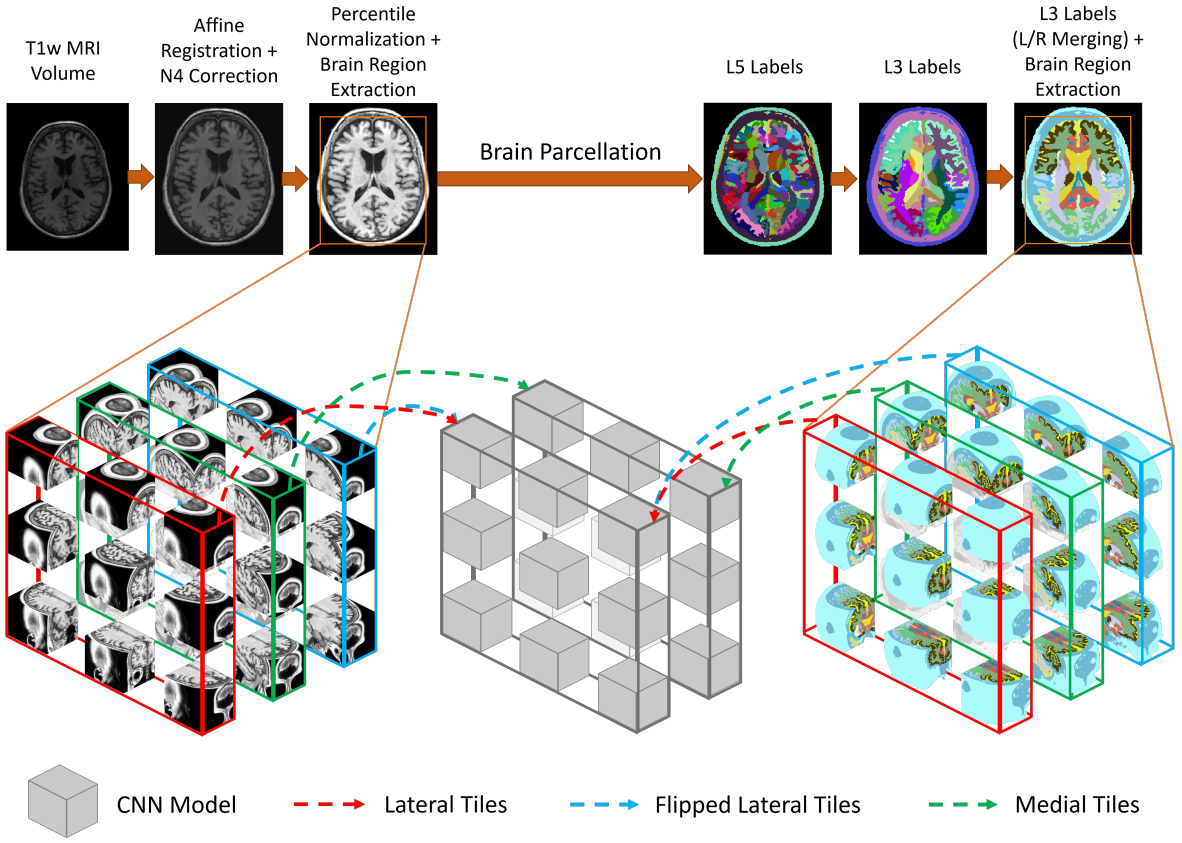


Figure 2: The proposed SLANT($3, \frac{2}{3}$) pipeline system includes canonical medical image pre-processing and ground truth generation. Each tile covers 66.67% of the bounding-box region that covers the brain volume.

spatial feature, as shown in Figure 3. In order to decrease the model complexity is this configuration, the spatial feature map size is downsampled by a factor of 2. The volume image is centred by normalizing the spatial features between -1 and 1. These 3D coordinates feature map is divided into subspace relevant to each tile, and concatenated with the second level of the UNet model. This proposed architecture exploit the advantages of few-shot learning SLANT approach. In addition, it decreases the number of training models. The UNet network model in the original SLANT paper uses voxel-wise Dice loss function between predicted parcels A_i and the ground truth parcels B_i, ignoring the background. Using M as the total number of parcels, the DSC and its derived loss function are defined as

$$DSC_i = \frac{2\,|A_i \bigcap B_i|}{|A_i| + |B_i|} \quad i = 1, ., ., M \tag{1}$$

and

$$L_{dice} = 1 - \frac{\sum_{i=1}^{M} DSC_i}{M} \tag{2}$$

3.4.3. *Feature Matching GAN Network Model*

This model was built based on existing work (Mondal et al., 2018) as illustrated in Figure 4. It proposes a novel combination between SLANT system architecture

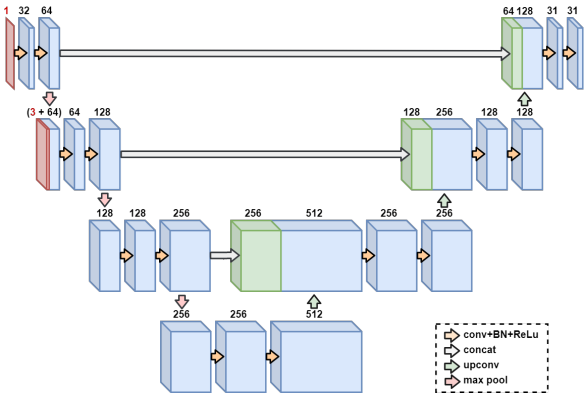


Figure 3: 3D UNet architecture with 4 inputs represented by red boxes. The input volume image in the first level and the three coordinates in the second level. Green boxes represent copied feature maps from the encoder and concatenated to the decoder at the same level.

and semi-supervised adversarial deep learning. To implement this model, the training set is split into labeled and unlabeled volumes, and each volume is divided into SLANT tiles. The labeled, unlabeled and the generated fake tiles are included during the training process.

The conventional GANs algorithmic architectures uses two CNNs, pitting one against the other. The gen-

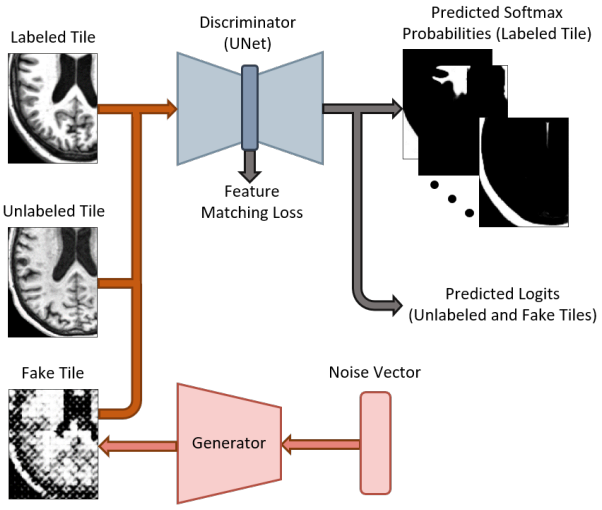


Figure 4: Feature matching GAN Model is an adversarial setup consisting of Generator, and Discriminator in UNet architecture. Both networks are trained simultaneously.

erator G_{θ_G} is trained to map a randomly generated noise vector $z \in \mathbb{R}^d$ with uniform distribution into a synthetic image vector $\tilde{x} = G(z)$. Meanwhile, the discriminator D_{θ_D} is trained to differentiate between real tiles $x \sim p_{data(x)}$ and synthesized tiles $\tilde{x} \sim p_{G(z)}$. Both of the generator and discriminator networks are two players in a min-max optimization game, as shown in the following function $V(D_{\theta_D}, G_{\theta_G})$.

$$\min_{G_{\theta_G}} \max_{D_{\theta_D}} \mathbb{E}_{x\sim p_{data(x)}}[logD_{\theta_D}] + \mathbb{E}_{z\sim noise}[1 - D_{\theta_D}(G_{\theta_G}(z))] \tag{3}$$

The labeled tiles in sub-volume space $x_{H\times W\times D}$ are trained using standard 3D UNet segmentation model to the output space $y_{H\times W\times D}$ with M logit classes $[l_{i,1}, ..., l_{i,M}]$. Using the Softmax function, the output can be represented by class probabilities.

$$p_{model}(y_i = j|x) = \frac{exp(l_{i,j})}{\sum_{m=1}^{M} exp(l_{i,m})} \tag{4}$$

A voxel-wise Dice loss function $L_{labeled}$ are computed between the predicted segmentation probabilities $p_{model}(y_i = j|x)$ using Softmax function and the ground truth of the labeled tile. Since the generator model G predict the realistic synthesized tile, the discriminator D requires an additional class to distinguish if the generated fake tile is true. This additional class can be recast back within the M classes by maximizing the following equation.

$$\mathbb{E}_{x\sim p_{data(x)}} \sum_{i=1}^{H\times W\times D} log\ p_{modal}(y_i \in 1, ..., M|x) \tag{5}$$

From the last equation, the unlabeled and fake tiles can be also trained using the same UNet model. Using the normalized logits strategy in (Salimans et al., 2016), the loss functions for the fake and unlabeled tiles $L_{unlabeled}$ and L_{fake} can be directly calculated by employing the normalized logits in the Softmax function of Equation (4), where $Z_i(x) = \sum_{m=1}^{M} exp[l_{i,k}(x)]$, as shown in the following equations.

$$L_{unlabeled} = -\mathbb{E}_{x\sim p_{data(x)}} \sum_{i=1}^{H\times W\times D} log\left[\frac{Z_i(x)}{Z_i(x)+1}\right] \tag{6}$$

$$L_{fake} = -\mathbb{E}_{x\sim p_{data(x)}} \sum_{i=1}^{H\times W\times D} log\left[\frac{1}{Z_i(G_{\theta_G}(z))+1}\right] \tag{7}$$

The yield Dice loss functions from labeled tiles $L_{labeled}$ are weighted by parameter α to stimulate the UNet segmentation predictions, and it is combined with obtained loss output from fake and unlabeled tiles in a discriminator loss function.

$$L_{discriminator} = \alpha L_{labeled} + L_{unlabeled} + L_{fake} \tag{8}$$

The generator uses a feature matching (FM) strategy for calculating the loss, which aims to match the expected values of features $f(x)$ in an intermediate layer of the discriminator. $f(x)$ is the output from the second last of the UNet encoder models, as it provides a higher performance than using the last layer (Mondal et al., 2018).

$$L_{generator} = \left\|\mathbb{E}_{x\sim p_{data(x)}}(x)f(x) - \mathbb{E}_{z\sim noise}f(G_{\theta_G}(z))\right\|_2^2 \tag{9}$$

Following the FM GAN authors' steps, the 3D UNet architecture is modified to adapt the GAN framework in order to make the training more stable. The weight normalization is used instead of the batch normalization. Also, Leaky ReLUs is used for activation functions because it is robust to small negative outputs (logits), as they will still provide a gradient whereas standard ReLU would not. Since sparse gradients are induced by max pooling which are not good for GANs, the authors replaced them with average pooling.

3.4.4. Predicted Labels Reconstruction from Network Tiles

During the reconstruction process for the predicted labels, the common voxels within the overlapped regions would be segmented more than once from multiple models. Since the predicted probabilities from the softmax activation function of the model output layer represents the membership of each voxel towards all classes, the robust trained models may provide high probability variance between the predicted classes. Therefore, in this work the different predicted probabilities for the overlapped voxels are summed together.

This technique provides a soft decision for the maximum argument classification of each class. The reconstructed function for the 3D image in the MNI space S_{MNI} at voxel point (i_x, i_y, i_z) is shown in the following equation, where S_t is the sub-space tile at voxel point (j_x, j_y, j_z), and T the total number of tiles.

$$S_{MNI}(i_x, i_y, i_z) = \underset{l\in\{0,1,\ldots,M-1\}}{argmax} \sum_{t=1}^{T} p\left(l \mid S_t(j_x, j_y, j_z)\right) \tag{10}$$

The predicted probabilities addition technique is better than the majority voting method which is not sensitive to the predicted probability outputs form the more robust trained models. In addition, it is less complex because it does not require an additional classification stage for the overlapped regions.

3.5. VoxelMorph Based Atlas Segmentation (VMBAS)

VoxelMorph Based Atlas Segmentation approach is carried out utilizing Bayes' rule for segmentation, and merging it with an unsupervised deep learning-based registration framework. The network model architecture $g_{\theta_C}(I, A) = (\theta_S, \theta_I) = (v, \mu, \sigma^2)$ of this system was designed using the 3D UNet based architecture for VoxelMorph (Balakrishnan et al., 2018). As demonstrated in Figure 5, the model reads two inputs: an MRI volume I and the probabilistic atlas A. This UNet includes 32 convolutional filters in both the encoder and decoder stages using a kernel size of 3, stride of 2, and LeakyReLu activation functions. At the end point of the UNet, a pair of convolutional layers are attached. The first convolutional layer to output stationary velocity field v. This layer is followed by scaling and squaring (Arsigny, 2006) (Dalca et al., 2018) (Krebs et al., 2019) integration layer to calculate the deformation field $\phi = exp(v)$, which yield the diffeomorphic flow loss parameter. Then, the probabilistic atlas A is warped using an additional spatial transform layer. The second convolutional layer to output the Gaussian intensity parameters μ, σ^2, which is combined with input MRI volume I to provide the likelihood maps. These maps with warped atlas enable computation of the data loss term.

From the generated probabilistic atlas, the prior probability A for each ground truth label l at the spatial voxel location $x_j \in \Omega$ is given by $A(l, x)$. The probabilistic atlas map is deformed by the diffeomorphic transform function ϕ_v, and by using the stationary velocity field parameter v which parametrizes the prior $\theta_s = v$. The S_j represents the segmentation at each voxel j, as shown in the following equation.

$$p(S|\theta_s; A) = p(S|v; A) = \prod_{j\in\Omega} A\left(S_j, \phi_v(x_j)\right) \tag{11}$$

The spatial location of the voxels j in image I are displaced by deformation field ϕ_v by the spatial gradient ∇u_v, where $\phi_v = Id + u_v$. The deformation term in the loss equation is weighted by the parameter λ.

$$p(\theta_S; \lambda) = p(v; \lambda) \propto exp\left[-\lambda \|\nabla u_v\|^2\right] \tag{12}$$

The likelihood parameters $\theta_I = \{\mu, \sigma^2\}$ are represented by Gaussian distribution function $\mathcal{N}(.; \mu_{S_j}, \sigma^2_{S_j})$ for the voxel intensity I at location j, where μ and σ^2 are the means and variances of voxels intensities under each class.

$$p(I|S, \theta_I) = p(I|S, \mu, \sigma^2) = \prod_{j\in\Omega} \mathcal{N}(I_j; \mu_{S_j}, \sigma^2_{S_j}) \tag{13}$$

Since the number of training volumes are limited, data was augmented to N volumes by horizontal flipping

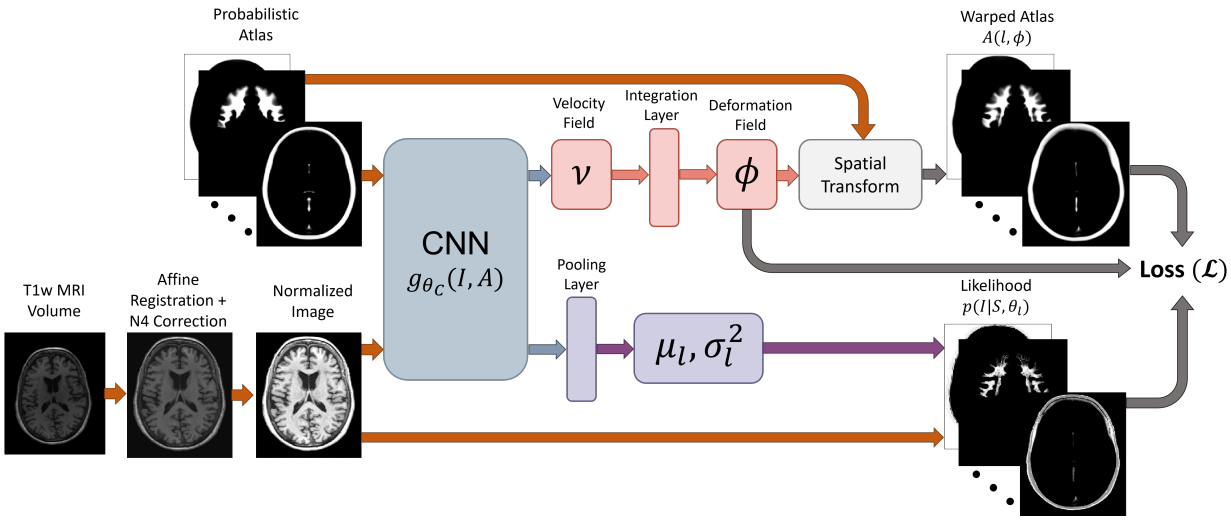


Figure 5: VoxelMorph Based Atlas segmentation pipeline. The network model $g_{\theta_C(.,.)}$ provides the deformation velocity field v which propagates the atlas to the input image space. and the intensity likelihood parameters μ, σ^2 that resample the likelihood map per each parcel.

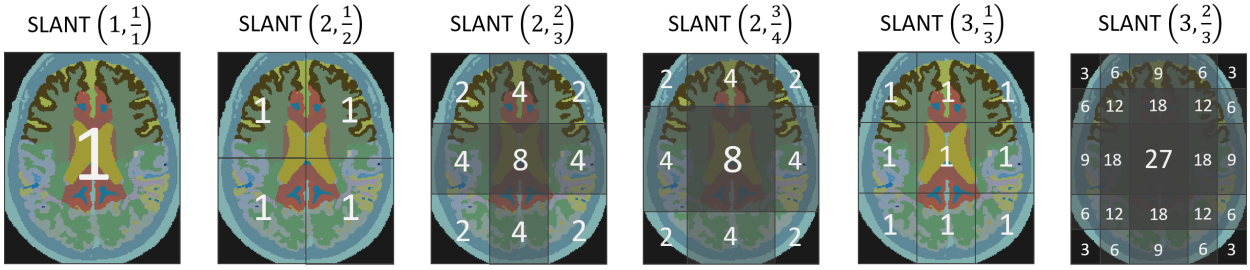


Figure 6: Different experimental setup for the SLANT approach. It show the number tiles of overlays in the axial cross sectional center.

of the images. The total loss function can be expressed as shown in the following equation. The log-partition function $K(\lambda)$ is controlled by the hyper-parameter λ, keeps the probability distribution with a proper values without affecting the optimization.

$$\begin{aligned}\mathcal{L}(A,I) &= -\sum_{n=1}^{N} log\ p(v^n,\mu^n,[\sigma^2]^n I^n;A,\lambda)\\ &= -\sum_{n=1}^{N}\sum_{j\in\Omega} log\left[\sum_{l=1}^{M}\mathcal{N}(I_j^n;,\mu_l^n,[\sigma_l^2]^n)A\left(l,\phi_{v_m}(x_j)\right)\right]\\ &\quad -K(\lambda)+const\end{aligned} \tag{14}$$

This approach does not require the ground truth during the training. Consequently, it is contrast adaptive to MRI volumes with unobserved contrast. Given a new testing volume and the probabilistic atlas, the trained model predicts the deformation field $\hat{v}_t$ and the intensity parameters $\hat{\theta}_t$. The optimal segmentation can be computed from the maximum argument of the wrapped atlas and the likelihood values according to the following equation.

$$\hat{S}_j = \underset{l}{argmax}\mathcal{N}(I_j;\hat{\mu}_t;\hat{\sigma}_t^2)A(l,\phi_{\hat{v}}(x_j)) \tag{15}$$

3.6. Implementation Details

Figure 6 shows the different experimental setup for SLANT approach using UNet model which addresses three different SLANT cases. The first case SLANT$(1,\frac{1}{1})$ uses the entire volume, which is equivalent to not applying SLANT. The second case is without overlapping tiles, as illustrated in SLANT$(2,\frac{1}{2})$ and SLANT$(3,\frac{1}{3})$. The last case is by using overlapping tiles with different sizes. Two additional setups were implemented in SLANT$(3,\frac{2}{3})$ configuration, using a single model with and without the three dimensional spatial features. In order to make fair analysis for these various settings, the same hyper-parameters tuning was set for all experiments using batch size = 1, optimizer = "Adam" (Kingma and Ba, 2015) and learning rate = 0.0001. In addition, all the network models are trained over 30 epochs in all experiments. Besides the initial normalization technique using percentile for the entire volume, another normalisation was applied for the extracted sub-volume tiles before training using mean and standard deviation.

The plot in Figure 7 illustrates the model loss per epoch in the case of SLANT$(3,\frac{2}{3})$ with UNet configuration. It shows lower training and validation Dice loss for the center tile 2_2_2 more than the corner tile 1_1_1. Since the loss function does not account the background portion loss which having a large portion of background, it may lead to higher loss values in absolute terms. However, it shows a higher training and validation accuracy for the same corner tile more than the center tile as shown in Figure 8.

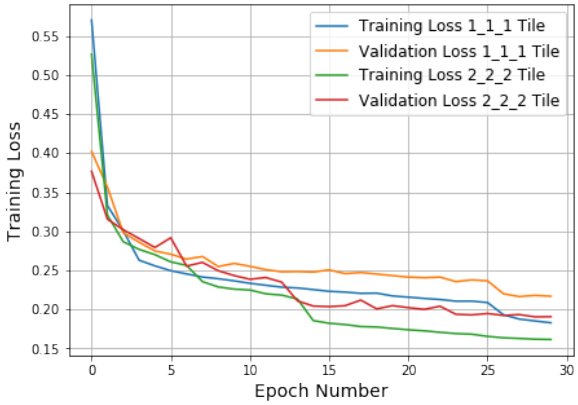


Figure 7: The training and validation Dice loss of the first fold cross validation for SLANT$(3,\frac{2}{3})$, and using UNet. The tiles 1_1_1 and 2_2_2 are located at the corner and the center respectively.

The hyper-parameters configuration for feature matching GAN uses the same patch size as in UNet, the "Adam" optimizer configures the generator and discriminator with learning rate 0.0001 and a momentum of 0.5. The labeled loss weight parameter α is set to 15. Two experimental configuration during the training stage based on the ratio of labeled and unlabeled volumes. The training set is divided into two equivalent portions of labeled and unlabeled volumes in the first configuration. The number of the training epochs in this case is set to 120 with 70 training volumes per epoch. In the second configuration, the training set is divided as quarter for labeled and three-quarter for unlabeled.

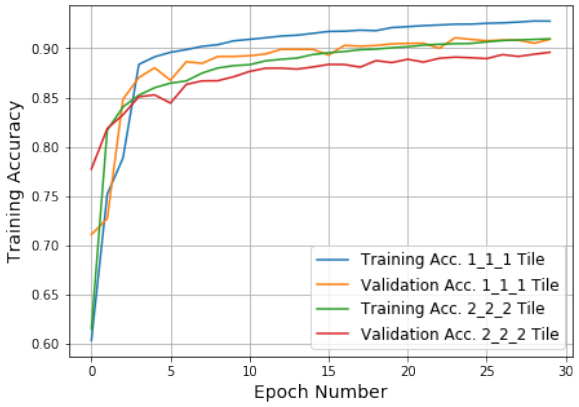


Figure 8: The training and validation accuracy of the first fold cross validation for SLANT($3,\frac{2}{3}$), and using UNet. The tiles 1_1_1 and 2_2_2 are located at the corner and the center respectively.

Since the number of training volume samples per epoch increased to 105 in the second configuration, the number of the training epochs is decreased to 80.

As shown in the loss plot in Figure 9, the Dice loss for labeled image is decreasing smoothly while the unlabeled and fake losses overall appear unstable. The center tile 2_2_2 shows lower labeled loss than the corner tile 1_1_1. However, Figure 10 shows higher accuracy for the corner tile 1_1_1 than the center tile 2_2_2.

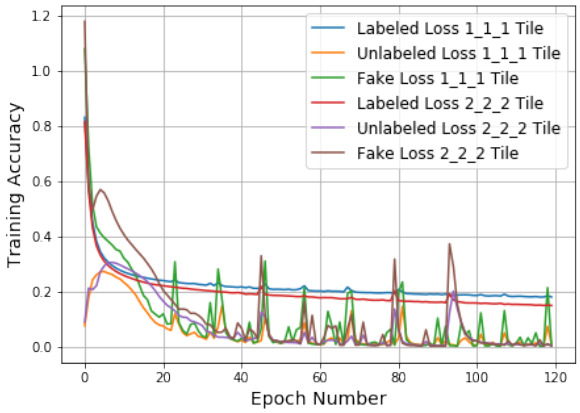


Figure 9: The training Dice losses of the first fold cross validation for SLANT($3,\frac{2}{3}$) using a half to half labeled and unlabeled data, and using FM GAN. The tiles 1_1_1 and 2_2_2 are located at the corner and the center respectively.

Atlas based VoxelMorph experiments are also configured using the same SLANT hyper-parameters. Meanwhile, the registration parameter λ is set to 10 which is set empirically. Figure 11 shows a decay in the data loss during training the model, while diffusion loss increasing. On the other hand, Figure 12 shows unstable and fast overfitting for the validation accuracy because the network model was trained using loss function to estimate the deformation field, without using spatial segmentation loss function between the predicted and the ground truth images.

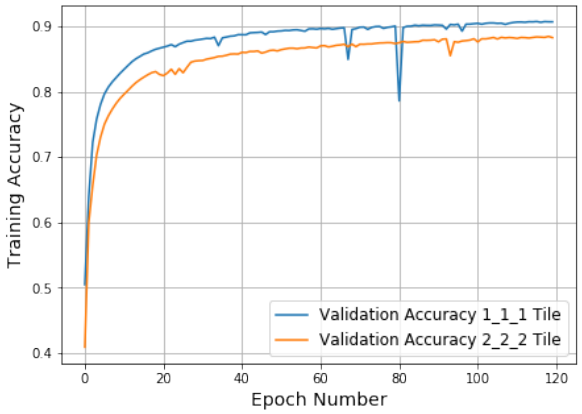


Figure 10: The validation accuracy of the first fold cross validation for SLANT($3,\frac{2}{3}$), and using FM GAN. The tiles 1_1_1 and 2_2_2 are located at the corner and the center respectively.

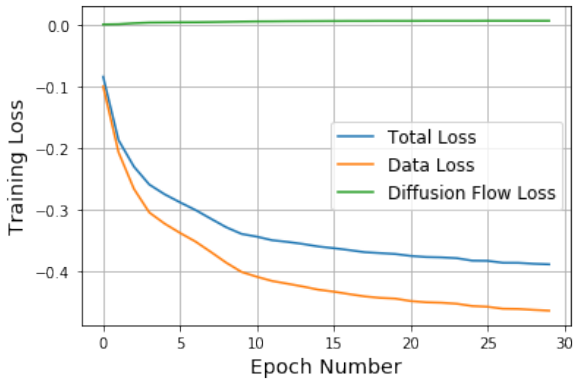


Figure 11: The training loss of the first fold cross validation for Atlas Voxelmorph. Total Loss = Data Loss + λ (Diffusion Flow Loss) , where $\lambda = 10$.

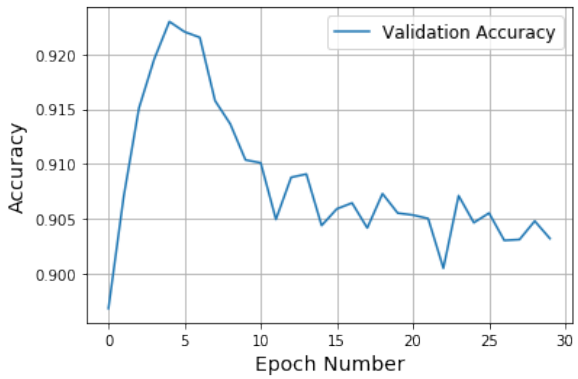


Figure 12: The validation accuracy of the first fold cross validation for VMBAS.

All the pre-processing and registration methods are kept the same for all SLANT and VMBAS experiments.

All training and testing was done on an Nvidia DGX-1[2] machine with Tesla V100 SXM2 GPUs 32GB memory.

3.7. Evaluation Criteria

In order to obtain robust evaluation for the predictive models results, the training set is shuffled in fixed seed, and then split into 3 equal portions for 3-fold cross-validation. The model which provides the greatest validation accuracy in each fold is selected for predicting the testing volumes. Then, the affine registration parameters from the pre-processing stage are inverted, and both of the ground truth and the predicted images are propagated from the MNI space to the original volume space. In the testing stage, post-processing techniques have been applied on the predicted images to sparse false predicted labels due to the noisy background. This technique is carried out in object-level by keeping the biggest connected component which represents the brain region, and discarding the tiny disconnected components.

All the experimental results have been evaluated using Dice Similarity Coefficient (DSC) for voxel-level metric functions according to Equation (1). Symmetric Hausdorff Distance (HD) is another applied metric function which is carried out by calculating the highest of all the distances from a point a in the predicted segmentation parcel A of specific parcel to the closest point b in the relevant ground truth parcel B. The Symmetric Hausdorff Distance is measured in the patient space.

$$HD = max(D(A, B), D(B, A)) \tag{16}$$

where

$$D(A, B) = max_{a \in A} min_{b \in B} \|a - b\| \tag{17}$$

4. Results

Quantitative and qualitative analysis have been performed to evaluate the segmentation results. The testing volumes are predicted three times from the best selected models in each cross validation fold. Using the metric functions, the predicted images are evaluated numerically over all parcels collectively and individually. Then, the three predicted results for each testing image are averaged. The box-plot diagram in Figure 13 provides high-level summaries of the performance of all the models. In order to obtain a precise analysis for the brain region, the background is not considered during computing the DSC, as it changes from image to another. Overall, SLANT approach outperforms VMBAS approach in DSC.

The segmentation performance increases by using overlapped tiles. It is affected by two factors which are the number of overlaid models and overlapped region size. In non-overlapped SLANT models, using a bigger tile size in SLANT($2,\frac{1}{2}$) model achieves higher performance than SLANT($3,\frac{1}{3}$). However, the segmentation performance degrades by fitting the brain volume model in one model as in SLANT($1,\frac{1}{1}$). In addition, it provides poor segmentation performance in many testing image cases. Increasing the number of overlaid models makes the overlapped regions to be predicted from multiple models at once, and it boosts the segmentation accuracy as in SLANT($3,\frac{2}{3}$). Meanwhile, increasing overlapped region size offers the possibility for bigger regions to be trained and predicted from different models.

The single SLANT model is trained on a higher number of tiles and it is less computationally and resource expensive in comparison to multi-model SLANT. Moreover, it achieves higher evaluation performance than SLANT($1,\frac{1}{1}$) because the model is trained on relatively few number of training examples, and it see the overlapping parts more often during training. Despite these benefits, the segmentation accuracy of a single SLANT model fails to attain the multi-model-SLANT. Adding the spatial location features produces a more robust segmentation performance but it still cannot exceed the multi-model SLANT.

The FM GAN model is a semi-supervised segmentation approach (contrasting with SLANT using UNet model, which is fully supervised), as it uses a partially labeled training set. Consequently, it provides lower segmentation performance in comparison to the UNet model using the same SLANT($3,\frac{2}{3}$) configuration.

In the case of using quarter to three-quarter labeled to unlabeled data, it gives a very slightly elevated results than half to half data. Hence, this method is robust to a reduction from 50% labeled to 25% labeled. Using three-quarter of unlabeled data makes the model trained on more various unlabeled examples in each epoch. Accordingly, it drives the model to better generalize to unseen instances.

Per-parcel analysis is reported for a set of parcels of specific interest for neurodegenerative disorders. These parcels are divided under six groups of Level 1, and they are listed in Tables 3 through 8. All experiments provide high Dice scores within bigger size parcels in general, while the implemented systems sometimes fail to detect small parcels. VMBAS approach shows the lowest DSC output in the major parcels. On the other hand, this approach and SLANT($3,\frac{2}{3}$) configuration provide relatively precise HD scores, in comparison to other SLANT configurations.

As illustrated in Table 3, the segmented parcels under Telencephalon group achieves the best DSC score in SLANT($3,\frac{2}{3}$) UNet configuration. However, Basal Ganglia parcel shows a slightly higher score with single model configuration, and it is the highest score among Telencephalon parcels. Conversely, Limbic White Matter shows the lowest DSC score. In Table 4,

[2]https://www.nvidia.com/en-gb/data-center/dgx-1/

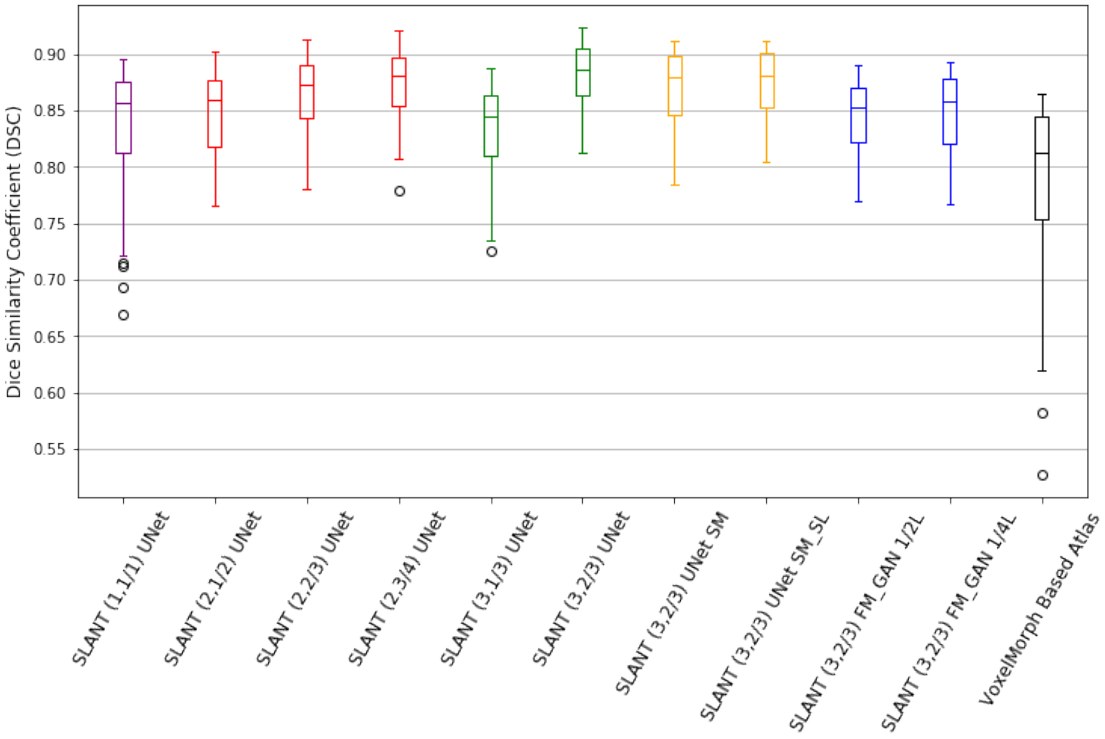


Figure 13: Statistical analysis using Dice Similarity metric function for the entire brain region. SM: Single Model, SL: Spatial Location and L: Labeled Volumes.

SLANT$(3,\frac{2}{3})$ UNet configuration obtain the lowest HD only Limbic White Matter parcel, while the same configuration with FM GAN model and half labeled data has the lowest score in Insula and Basel Ganglia parcels. However, the rest of Telencephalon parcels attain the lowest HD with VMBAS. SLANT$(1,\frac{1}{1})$ UNet configuration is unable to predict the Insula parcel in the two of cross validation models.

Table 5 lists the average DSC scores for four Brain tissue groups, which are Diencephalon, Mesencephalon, Metencephalon and Myelencephalon. SLANT$(3,\frac{2}{3})$ UNet provides the highest DSC scores in the parcels under the all four groups except the Thalamus parcel under Diencephalon group. The highest DSC score for Thalamus parcel achieved in single model configuration in SLANT$(3,\frac{2}{3})$. Meanwhile, this configuration fails to detect the Medulla parcel in one of cross validation models. In Table 6, the lowest HD values in Basal Forebrain, Midbrain and the Pons parcels are obtained in SLANT$(3,\frac{2}{3})$ UNet, while Cerebellum parcel achieves the lowest HD in VMBAS. Medulla parcel shows the lowest HD in single model SLANT$(3,\frac{2}{3})$ with spatial features. The lowest HD score among all the parcels is obtained in the Thalamus parcel by using $(3,\frac{2}{3})$ FM_GAN $\frac{1}{2}$L configuration.

SLANT$(3,\frac{2}{3})$ UNet configuration carries out the highest DSC scores in the CSF tissue group, as shown in Table 7. It also provides the lowest HD values in III_Ventricle and IV_Ventricle parcels as shown in Table 8, while the lowest HD value in LateralVentricle is attained by VMBAS.

Figures 14 and 15 illustrate the parcellation results for a predicted testing image sample. Comparing the predicted images visually with the ground truth reveals some segmentation challenges in the different experimental cases. As shown in the predicted images using the SLANT approach, some parcels are bleeding as they are falsely segmented as the neighboring parcels. Meanwhile, the predicted parcels in the overlapped regions are detected more precisely. The predicted parcels boundaries from the VMBAS are less accurate. Moreover, they completely fail to classify some brain voxels.

5. Discussion

Many sources of error have an impact on the segmentation. For instance, the delineation error from the parcellation application that generates a noisy pseudo ground truth. It affects the training model and the prediction results, and it may underestimate the tested methods due to the noise that pseudo GT inevitably. Therefore, an iterative certainty metrics such as cross-fold validation can be applied to reduce the pseudo GT error impact, and to obtain more meaningful results.

Affine registration process can has an indirect effect on the segmentation process, which may be correlated with similarity between the registered image and the MNI template image. Consequently, the distribution of

Experiment	Telencephalon							
	Parietal	**Limbic**	**Insula**	**BasalGangl**	**InferiorWM**	**Frontal**	**Temporal**	**LimbicWM**
$(1,\frac{1}{1})$ UNet	77.00	80.99	*26.79	85.53	84.85	82.72	83.74	72.18
$(2,\frac{1}{2})$ UNet	71.84	84.34	81.22	86.20	87.75	84.70	86.49	75.51
$(2,\frac{2}{3})$ UNet	79.39	85.49	88.50	89.94	88.60	85.61	87.60	72.89
$(2,\frac{3}{4})$ UNet	80.05	85.84	88.92	90.14	88.92	85.94	87.82	77.88
$(3,\frac{1}{3})$ UNet	75.53	84.13	87.16	83.69	84.41	84.12	86.16	75.42
$(3,\frac{2}{3})$ UNet	**81.00**	**86.99**	**89.84**	91.47	**89.93**	**86.98**	**88.62**	**79.39**
$(3,\frac{2}{3})$ UNet SM	78.82	86.39	89.17	**91.57**	89.08	86.02	87.14	79.10
$(3,\frac{2}{3})$ UNet SM_SL	79.59	86.66	89.16	91.12	89.43	86.03	87.47	78.84
$(3,\frac{2}{3})$ FM_GAN $\frac{1}{2}$L	75.66	81.05	83.91	89.89	87.82	81.61	81.54	74.46
$(3,\frac{2}{3})$ FM_GAN $\frac{1}{4}$L	75.90	80.06	82.68	88.13	86.90	82.68	82.48	73.12
VMBAS	70.12	76.32	78.44	76.64	79.36	78.37	80.28	60.89

Table 3: Percentage average Dice Similarity Coefficient for Telencephalon parcels group. BaselGangl: Basel Ganglia and WM: White Matter. (*) Failed to calculate Dice score in two of the cross fold experiments.

Experiment	Telencephalon							
	Parietal	**Limbic**	**Insula**	**BasalGangl**	**InferiorWM**	**Frontal**	**Temporal**	**LimbicWM**
$(1,\frac{1}{1})$ UNet	54.05	34.77	–	27.02	26.73	72.17	51.48	18.31
$(2,\frac{1}{2})$ UNet	68.38	28.31	44.60	21.15	29.47	62.30	41.64	19.82
$(2,\frac{2}{3})$ UNet	32.00	24.15	10.13	14.69	22.05	28.09	39.30	13.87
$(2,\frac{3}{4})$ UNet	20.17	19.90	9.49	8.09	16.18	26.68	20.60	14.85
$(3,\frac{1}{3})$ UNet	36.84	32.92	56.15	49.08	38.88	35.95	48.74	23.71
$(3,\frac{2}{3})$ UNet	10.33	10.24	6.00	7.10	11.28	13.18	13.66	**10.73**
$(3,\frac{2}{3})$ UNet SM	30.91	22.00	11.50	10.12	14.11	30.51	31.74	13.89
$(3,\frac{2}{3})$ UNet SM_SL	25.21	23.77	11.32	8.13	17.66	20.15	27.08	12.16
$(3,\frac{2}{3})$ FM_GAN $\frac{1}{2}$L	9.34	9.54	**4.74**	**6.70**	12.03	11.37	9.44	14.58
$(3,\frac{2}{3})$ FM_GAN $\frac{1}{4}$L	11.55	11.64	5.43	6.84	12.79	11.66	11.74	11.53
VMBAS	**8.71**	**8.38**	5.73	7.04	**11.19**	**8.68**	**7.96**	11.49

Table 4: Average Hausdorff Distance for Telencephalon parcels group. BaselGangl: Basel Ganglia and WM: White Matter. (–) Failed to calculate HD in two of the cross fold experiments.

Experiment	Diencephalon		Mesencephalon	Metencephalon		Myelencephalon
	Thalamus	**BasalForebrain**	**Midbrain**	**Pons**	**Cerebellum**	**Medulla**
$(1,\frac{1}{1})$ UNet	89.21	79.79	91.84	94.54	94.15	92.89
$(2,\frac{1}{2})$ UNet	91.87	77.78	92.57	94.93	94.96	75.65
$(2,\frac{2}{3})$ UNet	91.45	84.59	93.11	94.91	96.20	91.61
$(2,\frac{3}{4})$ UNet	92.12	86.01	94.32	95.85	96.01	94.05
$(3,\frac{1}{3})$ UNet	71.90	76.89	88.69	93.45	94.85	87.76
$(3,\frac{2}{3})$ UNet	92.99	**87.39**	**94.58**	**96.27**	**96.56**	**94.94**
$(3,\frac{2}{3})$ UNet SM	**93.57**	87.28	94.10	95.41	95.55	*62.62
$(3,\frac{2}{3})$ UNet SM_SL	93.38	86.63	94.53	95.67	95.93	93.90
$(3,\frac{2}{3})$ FM_GAN $\frac{1}{2}$L	93.14	83.73	93.24	94.64	92.71	93.09
$(3,\frac{2}{3})$ FM_GAN $\frac{1}{4}$L	91.87	82.45	93.03	94.82	94.80	92.89
VMBAS	83.60	65.10	80.97	86.71	90.02	83.94

Table 5: Percentage average Dice Similarity Coefficient for Diencephalon, Mesencephalon, Metencephalon and Myelencephalon parcels groups. (*) Failed to calculate Dice score in one of the cross fold experiments.

the training data will be centered on the template at least as far as the spatial arrangement of the anatomy is concerned. In addition, it changes the different parcels size.

Another challenging issue is the lack of differentiability between the intensity distributions/textures of neighboring parcels which cause inaccurate segmentation boundaries between the parcels or completely false segmented parcels. Therefore, strong spatial priors can overcome this problem as in VMBAS approach. In addition, the non-brain tissue classes especially that include a combined different types of face and head tissues in a single parcel. This creates a class with variable features that interfere with the multiple primary features of other classes. Therefore, the non-brain tissue classes

Experiment	Diencephalon		Mesencephalon	Metencephalon		Myelencephalon
	Thalamus	BasalForebrain	Midbrain	Pons	Cerebellum	Medulla
$(1,\frac{1}{1})$ UNet	12.87	32.90	8.31	12.71	51.82	26.37
$(2,\frac{1}{2})$ UNet	29.63	75.48	49.23	35.12	97.62	105.98
$(2,\frac{2}{3})$ UNet	10.92	14.07	30.00	6.05	21.86	93.12
$(2,\frac{3}{4})$ UNet	5.45	7.33	5.68	9.79	28.60	46.77
$(3,\frac{1}{3})$ UNet	56.24	56.77	62.26	56.56	42.46	100.39
$(3,\frac{2}{3})$ UNet	4.35	**4.75**	**5.05**	**4.59**	8.30	51.54
$(3,\frac{2}{3})$ UNet SM	10.70	10.69	6.90	5.99	14.46	–
$(3,\frac{2}{3})$ UNet SM_SL	21.65	8.33	14.31	6.08	14.10	**3.70**
$(3,\frac{2}{3})$ FM_GAN $\frac{1}{2}$L	**3.24**	5.37	9.73	17.61	12.61	25.47
$(3,\frac{2}{3})$ FM_GAN $\frac{1}{4}$L	3.47	5.97	10.17	6.91	12.53	14.65
VMBAS	4.63	5.21	5.18	5.16	**7.74**	4.41

Table 6: Average Hausdorff Distance for Diencephalon, Mesencephalon, Metencephalon and Myelencephalon parcels groups. (–) Failed to calculate HD in one of the cross fold experiments.

Experiment	CSF		
	LateralVentricle	III_Ventricle	IV_Ventricle
$(1,\frac{1}{1})$ UNet	90.35	86.28	90.55
$(2,\frac{1}{2})$ UNet	92.73	87.04	90.71
$(2,\frac{2}{3})$ UNet	93.90	90.30	76.46
$(2,\frac{3}{4})$ UNet	94.10	90.17	92.35
$(3,\frac{1}{3})$ UNet	91.69	40.35	85.25
$(3,\frac{2}{3})$ UNet	**94.56**	**91.25**	**92.75**
$(3,\frac{2}{3})$ UNet SM	94.21	90.14	91.24
$(3,\frac{2}{3})$ UNet SM_SL	94.13	90.37	91.29
$(3,\frac{2}{3})$ FM_GAN $\frac{1}{2}$L	93.25	89.22	90.86
$(3,\frac{2}{3})$ FM_GAN $\frac{1}{4}$L	92.81	89.88	90.55
VMBAS	85.42	74.98	79.54

Table 7: Percentage average Dice Similarity Coefficient for CSF parcels group. III: The Third and IV: The Fourth

Experiment	CSF		
	LateralVentricle	III_Ventricle	IV_Ventricle
$(1,\frac{1}{1})$ UNet	22.49	12.48	20.17
$(2,\frac{1}{2})$ UNet	25.03	25.23	41.75
$(2,\frac{2}{3})$ UNet	20.96	11.54	30.07
$(2,\frac{3}{4})$ UNet	17.42	7.23	4.42
$(3,\frac{1}{3})$ UNet	31.43	63.68	88.38
$(3,\frac{2}{3})$ UNet	16.16	**6.22**	**3.42**
$(3,\frac{2}{3})$ UNet SM	20.26	14.65	4.35
$(3,\frac{2}{3})$ UNet SM_SL	16.29	9.86	7.66
$(3,\frac{2}{3})$ FM_GAN $\frac{1}{2}$L	15.72	5.45	3.91
$(3,\frac{2}{3})$ FM_GAN $\frac{1}{4}$L	18.58	8.22	4.73
VMBAS	**13.35**	6.85	5.02

Table 8: Average Hausdorff Distance for CSF parcels group. III: The Third and IV: The Fourth

are scripted from the database in many implementations before the processing.

Furthermore, unbalanced parcels size makes detecting and segmenting tasks very hard for small parcels. Thus, this problem can be handled by using a loss function during the training stage which normalize each parcel loss based on the its size, such as in Dice Similarity loss function, or multiplying each parcel loss by a certain weight. The drawback of these techniques that it can make a single false predicted pixel in tiny parcels can have the same effect as missing nearly a whole large parcels.

In the SLANT approach, the network segmentation model suffers from disperse false predicted segmentation voxels, and accordingly they deteriorate the HD error. These errors are decreased in the high number of

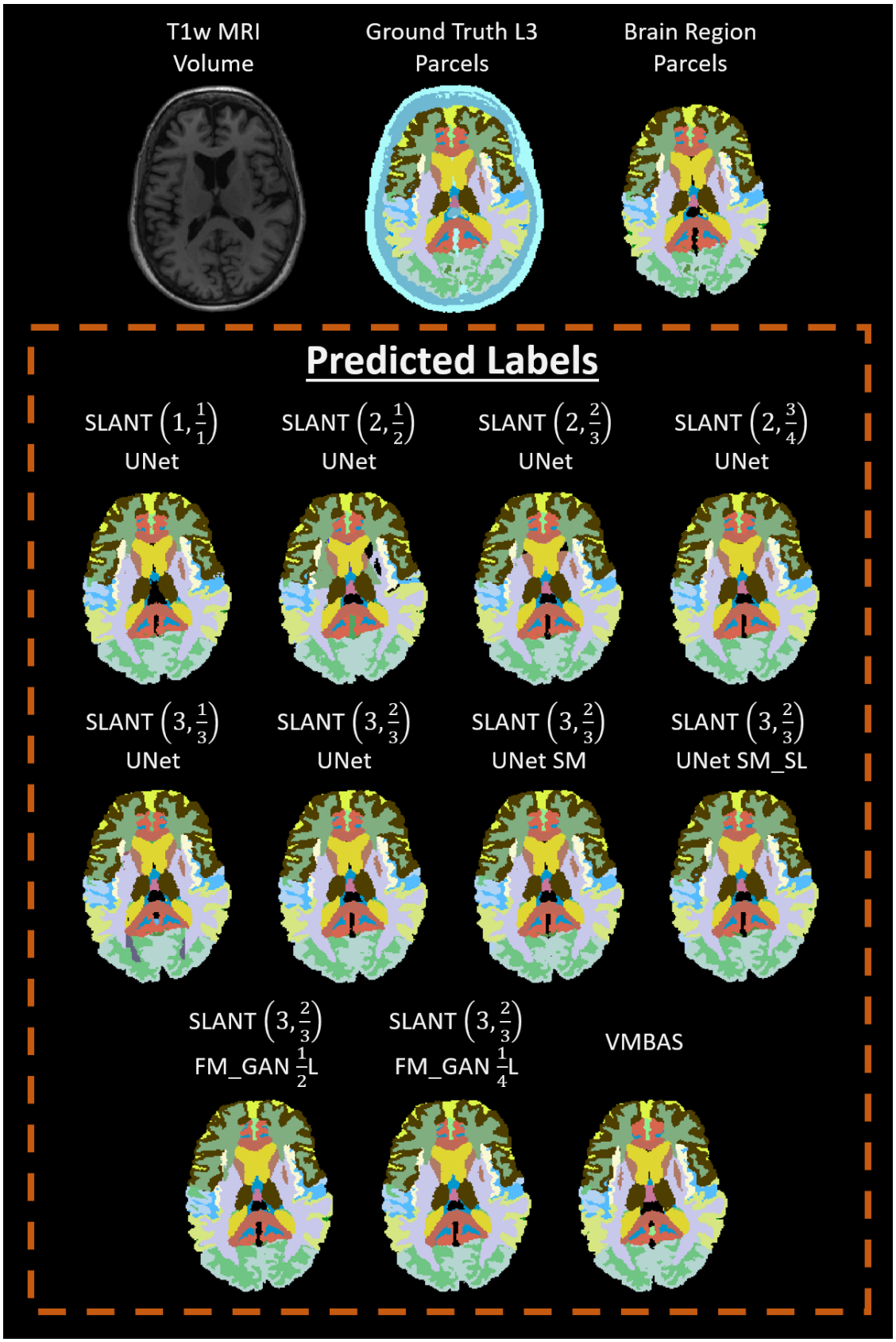


Figure 14: Axial position for selected medium quality testing image. The parcels are predicted using the best selected models from the same cross validation.

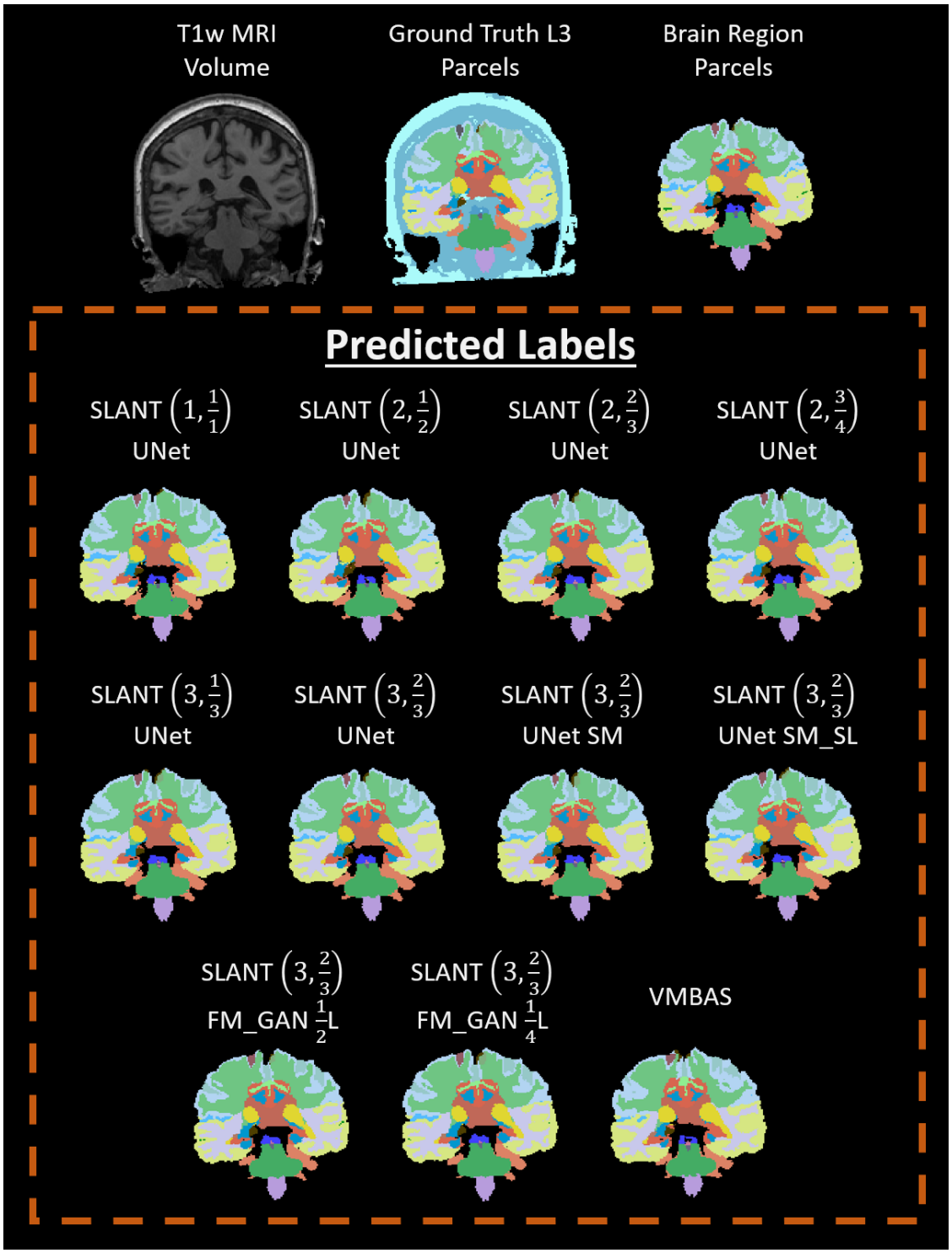


Figure 15: Coronal position for selected medium quality testing image. The parcels are predicted using the best selected models from the same cross validation.

overlapped regions as they are predicted from different models that corrects the combined prediction probabilities. Therefore, the parcels in the overlapped regions towards the image center as in SLANT$(3,\frac{2}{3})$ configuration have smaller HD error.

The depth of the generator and the discriminator in the FM GAN model are not adaptive to the variation of the nonuniform tiles size. Thus, it generates more pixeled fake images when the tile size increases. Using more deep network can overcome this problem as shown in Figure 16, however it can make the model over-fits very fast, and it requires more GPU memory

size. Moreover, having a precise generator can actually degenerate the training performances since in this case the model will not be able to distinguish between unlabeled and fake tiles.

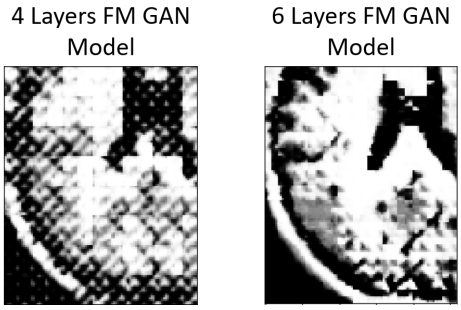


Figure 16: The generated fake tiles using different depth of discriminator and generator networks.

The predicted parcels from VMBAS are spatially constrained with deformed atlas probabilities which prevents the segmentation from having dispersed false predicted segmentation voxels, and therefore it has comparatively low HD error. On the other hand, this approach is reliant on intensity variation for likelihood parameters which makes it not robust to segment high numbers of parcels with similar intensity characteristics. Furthermore, its level of supervision during training does not consider the class size imbalance between parcels. Consequently, it provides comparatively low Dice scores.

6. Conclusions

This research employs two recent approaches which combine medical image processing in MNI space with Deep Learning for full brain volume segmentation. The first approach is substitutional implementation for the SLANT approach by merging the left side and right side tiles while training each model, exploiting the symmetry property of the brain. In addition, preforming analyses using model single model for all tiles with and without feeding the training model with the 3D spatial location feature. All the annotated training data are used for fully supervised training technique using UNet model. Another semi-supervised learning technique is proposed to use a small portion of the annotated training data, combining the tiles-based method in the SLANT approach with the FM GAN model. The second approach is an implementation for unsupervised segmentation principled approach based on combining VoxelMorph registration network with Probabilistic atlas priors. The experimental works are performed on a combination of MRI datasets from several medical resources, and the ground truth are annotated using an automatic parcellation application.

SLANT approach using UNet model has the strongest supervision during training, and it is therefore have the highest segmentation performance. Semi-supervised learning model using FM GAN can be trained on very few number of labeled data and providing comparably acceptable segmentation results. The overlapped regions in SLANT improves segmentation, partly overcoming the lack of a strong spatial prior. Increasing this overlaid regions provided more robust results. Using single model and adding the spatial features requires Less computational resources. However, it does not outperform the original SLANT configuration. VMBAS has the lowest segmentation performance, since it is unsupervised segmentation approach.

References

Aljabar, P., Heckemann, R.A., Hammers, A., Hajnal, J.V., Rueckert, D., 2009. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. Neuroimage 46, 726–738.

Angermann, C., Haltmeier, M., 2019. Random 2.5 D U-Net for fully 3D segmentation, in: Machine Learning and Medical Engineering for Cardiovascular Health and Intravascular Imaging and Computer Assisted Stenting. Springer, pp. 158–166.

Arsigny, V., 2006. Processing data in Lie groups: an algebraic approach. application to non-linear registration and diffusion tensor MRI. Ph.D. thesis.

Asman, A.J., Landman, B.A., 2013. Non-local statistical label fusion for multi-atlas segmentation. Medical image analysis 17, 194–208.

Baillard, C., Hellier, P., Barillot, C., 2001. Segmentation of brain 3D MR images using level sets and dense registration. Medical image analysis 5, 185–194.

Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2018. An unsupervised learning model for deformable medical image registration, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 9252–9260.

Bengio, Y., Courville, A., Vincent, P., 2013. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence 35, 1798–1828.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 424–432.

Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R., 2018. Unsupervised learning for fast probabilistic diffeomorphic registration, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 729–738.

Dalca, A.V., Yu, E., Golland, P., Fischl, B., Sabuncu, M.R., Iglesias, J.E., 2019. Unsupervised deep learning for Bayesian brain MRI segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 356–365.

Dhanachandra, N., Manglem, K., Chanu, Y.J., 2015. Image segmentation using k-means clustering algorithm and subtractive clustering algorithm. Procedia Computer Science 54, 764–771.

Dong, H., Yang, G., Liu, F., Mo, Y., Guo, Y., 2017. Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks, in: annual conference on medical image understanding and analysis, Springer. pp. 506–517.

Fei-Fei, L., Fergus, R., Perona, P., 2006. One-shot learning of object categories. IEEE transactions on pattern analysis and machine intelligence 28, 594–611.

Frazier, J., Caviness, V., Kennedy, D., Worth, A., Haselgrove, C., Caplan, D., Makris, N., 2007. Internet brain segmentation repository (IBSR) 1.5 mm dataset. Collections 10, C6RC85.

Guimond, A., Meunier, J., Thirion, J.P., 2000. Average brain models: A convergence study. Computer vision and image understanding 77, 192–210.

Heckemann, R.A., Hajnal, J.V., Aljabar, P., Rueckert, D., Hammers, A., 2006. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage 33, 115–126.

Huo, Y., Xu, Z., Aboud, K., Parvathaneni, P., Bao, S., Bermudez, C., Resnick, S.M., Cutting, L.E., Landman, B.A., 2018. Spatially localized atlas network tiles enables 3D whole brain segmentation from limited data, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 698–705.

Huo, Y., Xu, Z., Xiong, Y., Aboud, K., Parvathaneni, P., Bao, S., Bermudez, C., Resnick, S.M., Cutting, L.E., Landman, B.A., 2019. 3D whole brain segmentation using spatially localized atlas network tiles. NeuroImage 194, 105–119.

Iglesias, J.E., Sabuncu, M.R., 2015. Multi-atlas segmentation of biomedical images: a survey. Medical image analysis 24, 205–219.

Kass, M., Witkin, A., Terzopoulos, D., 1988. Snakes: Active contour models. International journal of computer vision 1, 321–331.

Kingma, D.P., Ba, J.L., 2015. Adam: A method for stochastic gradient descent, in: ICLR: International Conference on Learning Representations.

Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P., 2009. Elastix: a toolbox for intensity-based medical image registration. IEEE transactions on medical imaging 29, 196–205.

Krebs, J., Delingette, H., Mailhé, B., Ayache, N., Mansi, T., 2019. Learning a probabilistic model for diffeomorphic registration. IEEE transactions on medical imaging 38, 2165–2176.

Lancaster, J.L., Tordesillas-Gutiérrez, D., Martinez, M., Salinas, F., Evans, A., Zilles, K., Mazziotta, J.C., Fox, P.T., 2007. Bias between MNI and Talairach coordinates analyzed using the ICBM-152 brain template. Human brain mapping 28, 1194–1205.

Landman, B., Warfield, S., 2012. MICCAI 2012 multi-atlas labeling challenge, in: MICCAI 2012 Workshop on Multi-Atlas Labeling, pp. 1–164.

Lawrence, S., Giles, C.L., Tsoi, A.C., Back, A.D., 1997. Face recognition: A convolutional neural-network approach. IEEE transactions on neural networks 8, 98–113.

LeCun, Y., 1998. The mnist database of handwritten digits. http://yann. lecun. com/exdb/mnist/ .

Madabhushi, A., Udupa, J.K., 2006. New methods of mr image intensity standardization via generalized scale. Medical physics 33, 3426–3434.

McInemey, T., Terzopoulos, D., 1999. Topology adaptive deformable surfaces for medical image volume segmentation. IEEE transactions on medical imaging 18, 840–850.

Mikhael, S.S., Pernet, C., 2019. A controlled comparison of thickness, volume and surface areas from multiple cortical parcellation packages. BMC bioinformatics 20, 55.

Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 fourth international conference on 3D vision (3DV), IEEE. pp. 565–571.

Mondal, A.K., Dolz, J., Desrosiers, C., 2018. Few-shot 3D multimodal medical image segmentation using generative adversarial learning. arXiv preprint arXiv:1810.12241 .

Mori, S., Wu, D., Ceritoglu, C., Li, Y., Kolasny, A., Vaillant, M.A., Faria, A.V., Oishi, K., Miller, M.I., 2016. MRICloud: delivering high-throughput mri neuroinformatics as cloud-based software as a service. Computing in Science & Engineering 18, 21–35.

Murphy, S., Mohr, B., Fushimi, Y., Yamagata, H., Poole, I., 2014. Fast, simple, accurate multi-atlas segmentation of the brain, in: International Workshop on Biomedical Image Registration, Springer. pp. 1–10.

Rohlfing, T., Russakoff, D.B., Maurer, C.R., 2004. Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. IEEE transactions on medical imaging 23, 983–994.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Roth, H.R., Lu, L., Seff, A., Cherry, K.M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., Summers, R.M., 2014. A new 2.5 D representation for lymph node detection using random sets of deep convolutional neural network observations, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 520–527.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training gans, in: Advances in neural information processing systems, pp. 2234–2242.

Schaap, M., Metz, C.T., van Walsum, T., van der Giessen, A.G., Weustink, A.C., Mollet, N.R., Bauer, C., Bogunović, H., Castro, C., Deng, X., et al., 2009. Standardized evaluation methodology and reference database for evaluating coronary artery centerline extraction algorithms. Medical image analysis 13, 701–714.

Simkó, A., Löfstedt, T., Garpebring, A., Nyholm, T., Jonsson, J., 2019. A generalized network for MRI intensity normalization. arXiv preprint arXiv:1909.05484 .

Staib, L.H., Duncan, J.S., 1992. Boundary finding with parametrically deformable models. IEEE Transactions on Pattern Analysis & Machine Intelligence , 1061–1075.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: improved N3 bias correction. IEEE transactions on medical imaging 29, 1310–1320.

Van Ginneken, B., Frangi, A.F., Staal, J.J., ter Haar Romeny, B.M., Viergever, M.A., 2002. Active shape model segmentation with optimal features. IEEE transactions on medical imaging 21, 924–933.

de Vent, N.R., Agelink van Rentergem, J.A., Schmand, B.A., Murre, J.M., Huizenga, H.M., Consortium, A., et al., 2016. Advanced Neuropsychological Diagnostics Infrastructure (ANDI): A normative database created from control datasets. Frontiers in Psychology 7, 1601.

Wang, H., Suh, J.W., Das, S.R., Pluta, J.B., Craige, C., Yushkevich, P.A., 2012. Multi-atlas segmentation with joint label fusion. IEEE transactions on pattern analysis and machine intelligence 35, 611–623.

Wang, L., Shi, F., Li, G., Gao, Y., Lin, W., Gilmore, J.H., Shen, D., 2014. Segmentation of neonatal brain MR images using patch-driven level sets. NeuroImage 84, 141–158.

Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE transactions on medical imaging 23, 903–921.

Wu, D., Ma, T., Ceritoglu, C., Li, Y., Chotiyanonta, J., Hou, Z., Hsu, J., Xu, X., Brown, T., Miller, M.I., et al., 2016. Resource atlases for multi-atlas brain segmentations with multiple ontology levels based on T1-weighted MRI. Neuroimage 125, 120–130.

Wu, M., Rosano, C., Lopez-Garcia, P., Carter, C.S., Aizenstein, H.J., 2007. Optimum template selection for atlas-based segmentation. NeuroImage 34, 1612–1618.

Zhang, D.Q., Chen, S.C., 2004. A novel kernelized fuzzy c-means algorithm with application in medical image segmentation. Artificial intelligence in medicine 32, 37–50.

Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE transactions on medical imaging 20, 45–57.

Zhu, S.C., Yuille, A., 1996. Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. IEEE transactions on pattern analysis and machine intelligence 18, 884–900.