NeuroImage

Longitudinal diffusion MRI analysis using Segis-Net: a single-step deep-learning framework for simultaneous segmentation and registration

--Manuscript Draft--

Manuscript Number:	NIMG-20-2952
Article Type:	VSI: Longitudinal Neuroimaging
Section/Category:	Computational modelling and analysis
Abstract:	This work presents a single-step deep-learning framework for longitudinal image analysis, coined Segis-Net. To optimally exploit information available in longitudinal data, this method concurrently learns a multi-class segmentation and nonlinear registration. Segmentation and registration are modeled using a convolutional neural network and optimized simultaneously for their mutual benefit. An objective function that optimizes spatial correspondence for the segmented structures across time-points is proposed. We applied Segis-Net to the analysis of white matter tracts from N=8045 longitudinal brain MRI datasets of 3249 elderly individuals. Segis-Net approach showed a significant increase in registration accuracy, spatio-temporal segmentation consistency, and reproducibility comparing with two multistage pipelines. This also led to a significant reduction in the sample-size that would be required to achieve the same statistical power in analyzing tract-specific measures. Thus, we expect that Segis-Net can serve as a new reliable tool to support longitudinal imaging studies to investigate macro- and microstructural brain changes over time.

We would like to thank the reviewers for their review and the positive evaluation of our manuscript. They appreciate that

- the manuscript is well written and the method is clearly described,
- the proposed method is a novel solution to longitudinal neuroimaging analysis,
- the aim of the study is rigorously evaluated with a set of well-conceived experiments, i.e., comparing the proposed Segis-Net of simultaneous optimization with two sequential methods consisting of either classical or deep-learning-based components,
- the performance is promising.

Nevertheless, the reviewers also raised concerns relating to the novelty of the proposed simultaneous optimization strategy in comparison with similar works involving segmentation and registration, for instance, VoxelMorph (Balakrishnan et al., 2019) and U-ReSNet (Estienne et al., 2019).

For the current version of the manuscript, we made an effort to address all suggestions made by the reviewers, and specifically to the comparison with related methods. In the following section, we provide a point-by-point response to all reviewers' comments.

Reviewer #2:

Summary:

The authors present a generic CNN framework for simultaneous segmentation and registration and show its application in longitudinal diffusion MRI analysis. The framework consists of two parallel network modules, one for segmentation and the other for deformable registration, and optimizes their loss function jointly. The authors compare the proposed network with two-step methods and the results show superior performance of the proposed method in terms of registration accuracy, spatio-temporal consistency and reproducibility. Overall, the paper is well written and the method is clear.

We thank the reviewer for the positive comments.

General comments:

1. *There are other CNNs for simultaneous segmentation and registration. What's the difference of the proposed method with those, e.g. Estienne et al.?*

Please clarify the novelty.

The proposed method differs from the existing methods in a number of ways, the major one being the way in which the two components are coupled.

For the coupling method, Estienne et al. (2019) used a shared CNN encoder for segmentation and registration and the same paired-input was required for both tasks, namely, the input for segmentation is a pair of rigid-aligned images. That strategy limits the use of task-specific input, and for cross-sectional tasks at test stage it would be complicated to segment one image while requiring two registered images as input. Our proposed framework uses parallel network modules for the two tasks and only aligns the predicted segmentation, which makes it more generally applicable and flexible. It can be applied to task-specific network architecture and use optimized input for each task, e.g., in the present study we use diffusion tensor image for segmentation and use FA map for registration. This would not be possible for

the architecture as used in Estienne et al. (2019) due to the required interpolation of the diffusion tensor images.

We made a revision to the paper to explain the difference of our method to existing simultaneous works, and we explain why we hypothesize improved performance of the proposed method. Also, experiments in which the methods are compared have been performed to test this hypothesis; they are described in the newly added section (4.6 Comparison with related methods) and discussed as follows:

"In the comparison with related methods, Segis-Net had a better segmentation performance than U-ReSNet, an existing simultaneous method (Table 1). We expect this improved performance of Segis-Net because of two reasons: 1) the method allows the use of diffusion tensor images for tract segmentation, as we use parallel network modules and only align the predicted segmentation; this circumvents the need to interpolate the tensor images. In other words, task-specific inputs can be used; and 2) the sub-branches in the segmentation stream (Figure 8) are designed for the prediction of white matter tracts which can overlap with each other, unlike the exclusive tissue labels focused by other works."

2. *The proposed method is compared with two-step methods including both conventional and deep-learning methods, and overperform them. Did the authors compare to single-step method in this application setting?*

To the best of our knowledge, the proposed method is the first single-step method that is directly applicable to this diffusion imaging setting. Therefore, we focused on the comparison with two-step methods.

Other single-step methods developed for structural imaging, such as Estienne et al. (2019), are expected to have suboptimal performance for segmentation in this setting, because it requires the same scalar input image for both the registration and the segmentation task. However, for white matter tract segmentation the information in available scalar images such as FA maps is expected to be insufficient.

We made a revision to the paper to add a direct comparison of the proposed method and Estienne et al. (2019) in the newly added section 4.6.

Specific comments:

1. *For deformable registration, is the affine transformation also treated as an input to the network? Is it also determined by optimizing the SSE? Can the whole registration be an end-to-end learning instead of introducing preprocessing step? In Figure 9, the network takes Is and It as the input while Figure 1 the affine-aligned source image is used instead of Is*.

- For deformable registration, the affine transformation is indeed treated as an additional input to the network and is optimized using mutual information. In the revised version, we added the metric used for affine preprocessing (section 3.1.3): "The affine matrix of each image pair was estimated by optimizing the mutual information of FA images using Elastix software".

- We agree with the reviewer that the notation of the inputs in supplementary Figure 9 can be misinterpreted. Therefore, we made a revision to the paper to correct the notation of source

image in Figure 9 and added a description to the figure caption: “The target image and affine-aligned source image are used as the input to predict non-rigid deformation, which can subsequently lead to a composite displacement field as shown in Figure 1”.

- The whole registration can be learnt in an end-to-end fashion, which was demonstrated in our preliminary work (Li et al., 2019). In the present method, we learn only non-rigid deformation, which highlights those changes in the brain are due to aging.

We made a revision to introduce this difference in registration strategy: (in introduction) “The registration task within the framework is updated to learn only local deformations rather than an end-to-end composite including rigid transformation, as brain local changes over time is a focus in longitudinal imaging studies.”

2. *In individual Reg-Net, what is the degree of regularization? Same as the proposed network or experimentally determined?*

In Reg-Net, the weight of regularization loss is 0.01 times the weight of the intensity similarity (MSE) term. This was experimentally determined, and was the same as the weights used for the proposed framework. Specifically, for the hyperparameters in loss function, we first optimized the individual Reg-Net (i.e., 1 for Lreg, β/α for Ldef); and subsequently for the proposed network we optimized the relative weights between the segmentation and registration term (i.e., 1 for Lseg, α for Lreg); as a last step, we optimized the composite term for the proposed network (i.e., 1 for Lseg, γ for Lcom).

We made a revision to add the hyperparameters used in individual Reg-Net in implementation (section 3.6): “Loss function hyperparameters were optimized based on segmentation and registration performance on the validation dataset (search range: [10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}]); and were set to α=10, β=0.1 for Reg-Net and subsequently set to γ=1 for Segis-Net”.

Reviewer #3:

Summary:

The paper “Longitudinal diffusion MRI analysis using Segis-Net: a single-step dep-learning framework for simultaneous segmentation and registration” presents a novel framework to combine registration and segmentation into a unified training scheme. The method uses two networks, a "segmentation network" and a registration network, which are not interdependent, i.e. can be used independently in the end, but are trained in conjunction, i.e. they introduce a loss term that aims to include both networks in its term.

Within this framework, networks themselves do not present a novelty in their design. The networks are a standard U-Nets, the registration network is significantly derived from Voxelmorph. The methodological contribution then is the "segmentation consistency loss" term.

Since we consider the key novel element of this work the introduction of a generic optimization framework, we did not emphasize novel elements in network architecture. Nevertheless, the used networks have been optimized for application to diffusion MR images. For instance, in comparison with a

"standard U-Net" (Ronneberger et al., 2015), we used LeakyReLU instead of ReLU, used linear up-sampling instead of convolution transpose, and for multi-class segmentation we used sub-branches with Sigmoid function instead of one Softmax.

In the revised paper, we added a description on the sub-branch design of the network in method section 2.3.1, "For the segmentation stream, we split the output layer into sub-branches to facilitate multi-class classification for voxels with multiple labels. The final layer of the sub-branches consisted of a (1, 1, 1) convolution and a sigmoid activation."

The paper heavily builds on top of a MICCAI accepted conference paper with a significantly extended evaluation (and an extension to 6 tracts instead of 1). This evaluation includes comparison of with vs, without the new loss term, i.e. theirs vs. separate training of individual networks, for both the segmentation and the registration. Further the spatio-temporal consistency, reproducibility and increase in statistical power are evaluated.

The reviewer correctly summarized the novel extensions of this work compared to the MICCAI conference paper.

We made a revision to the paper to emphasize the change in extension of multiple tracts segmentation in introduction: "In this paper, we extend the tract-specific method by enabling concurrent segmentation of multiple tracts, which is a non-trivial task as a voxel can belong to multiple tracts. This also solves the problem of inconsistencies in deformations because of tract-specific ROIs".

Strengths

1. The paper is extremely well written and very consistent in its presentation (some exceptions, see minor)

2. The rigorous evaluation with a very appreciated set of experiments is well conceived

3. Results are promising

4. Code will be published as open source

5. Introducing a neuroimaging solution for longitudinal deep learning for diffusion images (application novelty afaik)

We thank the reviewer for recognizing the strength of our work.

Weaknesses

1. *Innovation is limited to the consecutive training of two networks and the introduction of one loss term*

As stated by the authors, registration networks have been regularized by a segmentation network loss and vice-versa. Since the (methodological) contribution of this work is in the loss "alone" a comparison to registration plus segmentation auxiliary loss is missing.

A main innovative aspect of the work is indeed the introduction of a new optimization framework by using parallel network modules, and designing a new loss function that make use of the simultaneous

optimization and mutual benefit of two tasks. Importantly, this innovation makes this the first single-step method for analysis of longitudinal diffusion MR imaging data.

As the reviewer suggested, in the revised paper, we added a comparison of registration accuracy between Segis-Net and VoxelMorph (Balakrishnan et al., 2019) on our dataset (section 4.6.2); and included a discussion on the results.

2. *Comparison with State of the art*

A direct (in-figure) and "on equal footing" comparison of segmentation performance with State-of-the-art methods is missing (e.g. Wasserthals work or Neuro4Neuro).

While this would also be nice for registration, (Grigorescu et al., WBIR 2020) or similar

The segmentation performance of Neuro4Neuro and the preliminary (tract-specific) Segis-Net was compared in the pilot conference work (Li et al., MICCAI 2019), in which Segis-Net achieved a higher accuracy in the segmentation of forceps minor tract.

In the revised version, we performed additional experiments to compare segmentation performance with a recently published method U-ReSNet (Estienne et al., MICCAI 2019), and compare registration performance with VoxelMorph (Balakrishnan et al., 2019) and U-ReSNet in the newly added section 4.6.

We made a revision to discuss the registration method mentioned by the reviewer as a potential research direction: “Deformable registration of diffusion tensor images is known to be challenging due to the directional components contained in voxels. Despite developments in classical methods for tensor reorientation during the optimization (Cao et al., 2006; Zhang et al., 2007), for learning-based registration it still largely remains unexplored. With the promising results of diffusion tensor interpolation as shown by Grigorescu et al. (2020), Segis-Net based on solely tensor images would be an interesting direction to explore.”

3. *Training on probtrackx*

It is unclear in how far training on probtrackx labels is biasing. At a minimum, this should be part of the discussion. Ideally, a second reference/evaluation scheme for tract segmentations would be included (such as manual labels).

In particular how does this interact with the classical pipeline (I am assuming you use probtrackx here as well)

The reviewer is correct that the used reference segmentations, which indeed is based on probtrackx combined with atlas information, might have biased the training of the method. However, no better ground truth is available since manual labels are not available for white matter tracts in-vivo (Crick and Jones, Nature 1993).

To assess any bias of the proposed approach, we evaluated performance on a number of aspects, e.g., the accuracy of volumetric segmentation, the reproducibility of segmentation and of the derived imaging

biomarkers, as well as the clinical value in the reduction of sample-size, in which the method showed an overall improved performance.

We made a revision to explain the situation that manual annotations are unavailable in the present study (introduction section): “Segmentation of WM tracts is however non-trivial because tracts cannot be identified directly from diffusion MRI, i.e., there is no in-vivo “gold standard” for tract (Crick and Jones, 1993)”.

4. *The "new loss term" (Naming, etc.)*

The authors call the new loss term "segmentation consistency loss". I disagree with this wording and would propose to name it "longitudinal composite loss".

"Consistency" implies that a result could be obtained in two different ways and they are then the same (consistent) -- at least in learning it does.

In particular, the construction of the following loss term would, in my opinion "consistency" --> \Ell_Dice(\F_theta (I_s) \circ \hat varphi, \F_theta(I_s \circ \hat varphi)) [observe the ordering of the mapping and the segmentation].

In their implementation, on the other hand, the loss between a reference segmentation I_t and the interpolated Segmentation map only implies composition of registration and segmentation makes sense.

In fact, with the current formulation, I would understand this loss to work similar to a data augmentation scheme (basically introducing a second ground truth by composition with the registration).

Since segmentation results are basically the same, one could argue all the advantage to this scheme really lies in the registration being regularized a bit better.

The proposed solution neatly circumvents the problem of "interpolation in tensor images/components", this could strengthen the argument for the proposed solution.

We agree with the reviewer to replace the naming of “segmentation consistency loss” by “longitudinal composite loss”, and accordingly made a revision to the paper and changed the notation of L_{cons} to L_{com}: (section 2.3.2) “The loss function is composed of four terms that measure segmentation accuracy, intensity similarity between registered images, deformation field smoothness, and longitudinal composite of registration and segmentation (L_{com}, Eq. 11)”; “We quantify the longitudinal composite loss term using the average Dice coefficient over all K structures”.

5. *Resampling of the Diffusion data*

It is unclear, whether there would be a negative bias from resampling the diffusion data. Some explanation of the reasoning for the upsampling would be required. In general the information provided on the Diffusion data seems minimal.

The diffusion data was resampled for the ball and stick diffusion model required for the probabilistic tractography, since the original resolution of our clinical quality data is rather low and it is suggested that with increasing spatial resolution more fiber bundles will be visible (Behrens et al., NeuroImage 2007).

We made a revision to explain the reason for up-sampling of the diffusion data: (section 3.2) “The voxel size was resampled from 3.3x2.2x3.5mm^3 to 1 mm^3 as required for probabilistic tractography (Behrens et al., 2007)”.

Minor

1. *varphi and phi should be consistent across text and figures - Figure 1 vs. Equations and text*

We made a revision to replace the /phi in Figure 1 by /varphi to be consistent with the text.

2. In 2.3.1, authors claim memory economy is an advantage of the design. This is still a bit unclear, in particular as interpolation on the GPU is very quick. But even more, you don’t have to keep this data in memory, just put it onto an SSD in HDF5

The use of “memory” indeed can be unclear, as we meant to be economic in disk storage instead of physical memory. Reducing the required storage space by a factor 2 makes quite a difference for large datasets like the one used in this study. We made a revision to replace “memory” by “disk storage”.

3. In 2.3.1, "probabilistic segmentations" could be defined a bit more robustly: e.g. are you predicting logits or probabilities?

Probabilities; we provide a probability map with a value between 0 and 1 for each possible class of white matter tract. For performance evaluation, we applied a threshold of p>0.5 on the probabilities to obtain a binary segmentation. This is explained in section 2.1: “the goal of CNN-based segmentation is to automatically infer, for each voxel $x \in R^n$, its probability of belonging to the structure $k \in [1, K]$”.

4. In 3.1.2, it is not clear, what the relevance of T1w and FLAIR data is. In particular, the Diffusion data is significantly more important, but completely disappears in the contrast.

The brain tissue mask we applied to the diffusion data was obtained from structural imaging, for which reason we provided the acquisition parameters of the T1w and FLAIR data.

We made a revision to the paper to omit the acquisition parameters for the structural imaging and refer readers to Ikram et al. (2011) for details: (section 3.2) “The acquisition parameters for structural and diffusion MRI can be found in Ikram et al. (2011). Specifically, diffusion MRI was scanned with the following parameters (…)”.

5. In 3.1.6 (and before that), Reasoning to why tensors are used as input are missing Neuro4Neuro evaluated tract segmentation on the basis of different input data.

We added the reason for using diffusion tensor images as input in section 3.6, “segmentation was based on the diffusion tensor image, as it contains directional information of fiber population and was shown to be optimal in this application setting (Li et al., NeuroImage 2020)”.

6. Suppl. Mat. Graphics: It is difficult to understand the conventions for the diagrams (Fig. 8+9), please describe and provide better legends (i.e. naming/clear assignment/etc.)

We simplified the legends for convolution unites by omitting the number of input feature channels, and added descriptions for the legend in captions of Figure 8 and 9: “The number of convolution kernels used in Conv Unit 01 is denoted as [K1, K2], and that for Conv Unit 02 as [K1].”

Conclusion

I do quite like the work and would have liked to propose acceptance after a minor edit. However, after close inspection and review, I do believe a major revision is necessary to substantiate the (limited) innovation and provide evidence of the superiority of the approach.

In particular, comparison to composition of "out of the box" methods for the individual tasks is missing and comparison to "regularization of individual networks by additional loss" is also missing. (i.e. train networks independently, but use an auxiliary loss).

In response to the reviewer's suggestion, we added additional experiments to compare with two related methods (section 4.6): 1) compare with an “out-of-the-box” registration method that uses an auxiliary loss (VoxelMorph, Balakrishnan et al., 2019), and 2) compare segmentation and registration performance with an existing single-step method which uses a shared encoder and separate decoders (U-ReSNet, Estienne et al., MICCAI 2019). In addition, a comparison to an “out-of-the-box” method for segmentation was provided in our preliminary work (Li et al., MICCAI 2019). The proposed approach showed overall superior performance in the above comparisons. In the revised paper, we also added a discussion on the innovation of the proposed method compared to an existing singe-step method U-ReSNet (see also response to the first comment of Reviewer #2).

Reviewer #4:

Summary:

The current paper presents an extended evaluation of an already published framework by the authors (Li et al., MICCAI 2019) for joint learning of image segmentation and registration termed Segis-Net. In the current submission, the authors extend the number of analyzed WM tracts from 1 to 6 and add reproducibility and tract-specific measures.

Strengths:

The paper is generally very well written with a clear outline and a good explanation of the basic concepts of the deep-learning aspects. The authors give a detailed comparison between their jointly trained framework (Segmentation 3D UNet + Registration 3D UNet trained together), the sequential application of the components of the framework (Segmentation 3D UNetSegNet and Registration 3D UNet trained separately) and a classical approach (an atlas-based segmentation approach + elastix for image registration).

We thank the reviewer for recognizing the strengths of our work.

Weakness:

1. *However, given that the general method is already published and the current submission is supposed to depict an extended validation of the framework, the paper unfortunately lacks a detailed comparison to other highly relevant state-of-the-art deep-learning methods. Specifically, comparison to Voxelmorph (Balakrishnan et al., 2019) and U-ReSNet (Estiennet et al., 2019 and 2020) should be included. This would specifically help to show the benefit of jointly training a separate 3D UNet for the segmentation task instead of a.) only including segmentations as an auxiliary Data Loss function as done in Voxelmorph and b.) a joint encoder and separate decoder for the registration and segmentation task as done in U-ResNet.*

In response to the reviewer's suggestion we performed additional experiments to compare our approach with VoxelMorph on the registration accuracy and compare with U-ReSNet. The comparison results (section 4.6) showed that

a) In the analysis of diffusion MR imaging, our approach of joint training with a segmentation task and a composite loss term led to a better registration performance than the VoxelMorph approach that only includes existing labels in an auxiliary loss term (Table 2);

b) For the segmentation task, the proposed Segis-Net yielded a better accuracy than the shared-encoder-based U-ReSNet for all six tracts. For the registration performance of VoxelMorph, U-ReSNet and Segis-Net, Segis-Net achieved the best correspondence in warped reference-segmentations (DC) and the least mean squared error (MSE); in the evaluation of correlation similarity, U-ReSNet was the best for three tracts, and Segis-Net was the best for two tracts and overall with the least standard deviation. Both methods led to a better performance than VoxelMorph (Table 1, 2).

In addition, comparison to other classical image-registration methods like ANTs (Avants 2011) and NiftyReg (Modat et al., 2010) would be beneficial (or at least a motivation why only comparison to elastix is done).

Elastix was adopted in the paper as a competing classical method since it has been widely used for these data and thereby an optimal parameter setting can be applied for performance comparison, whereas a default setting would be used for other methods.

In the revised paper, we added the explanation to the method section of Classical pipeline (section 3.5), “Elastix was adopted as a competing classical method since it has been widely used on our dataset and thereby an optimal parameter setting can be applied for performance comparison”.

2. *In addition, there are a few other aspects that are in my opinion improvable.*

Given the high similarity to Voxelmorph, it would be nice if the authors could clearly differentiate what exactly is their novel contribution with respect to the loss function. From what I see the main difference is the inclusion of the additional Lseg loss for the joint segmentation network.

Regarding the loss function, the proposed method has two main differences with VoxelMorph: 1) an addition of Lseg term for the segmentation accuracy, and 2) the composite term Lcom. For Lcom, the

VoxelMorph auxiliary loss term is a function of registration parameters only, in which the segmentation labels remain constant during the optimization. In contrast, for the proposed method, the Lcom term drives the optimization of both the registration and segmentation parameters. These two loss terms result in a key difference in the objective of both methods. VoxelMorph belongs to the category of methods that use existing segmentation labels to improve registration. While the proposed method belongs to the category of simultaneous optimization that optimize both segmentation and registration based on their correlation, similar to conventional unified methods like Ashburner and Friston (2005).

This is explained in the introduction:

"Various studies have shown that combining segmentation and registration at the stage of algorithm optimization can lead to improved performance. A popular combination strategy is to use the output of one task to optimize the other. Registration can be improved by using segmentation-level correspondences as input for deformation initialization (Dai and Khorram, 1999; Postelnicu et al., 2008) and optimization (De Groot et al., 2013b; Roh´e et al., 2017; Hu et al., 2018; Balakrishnan et al., 2019; Bastiaansen et al., 2020; Zhu et al., 2020)."

3. *Furthermore, it is unclear how the authors decided on the hyperparameters for balancing/regularizing the different components of the loss function (alpha, beta and gamma; Paragraph 3.1.6). It would be great if the authors could include a short overview of the tested ranges (if applicable) and resulting accuracy on the validation set*.

The hyperparameters in the loss function were experimentally optimized on the validation data.

We made a revision to the paper to include how we experimentally optimized the hyperparameters in the loss function. In section 3.6 we added: "Loss function hyperparameters were optimized based on segmentation and registration performance on the validation dataset (search range: [10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}]); and were set to α=10, β=0.1 for Reg-Net and subsequently set to γ=1 for Segis-Net".

4. *A short explanation why the initial learning rate for Reg-Net is different from Seg-Net and Segis-Net as well as specification which epoch/iteration resulted in the best model for each network (smallest error with respect to validation set) would be nice*.

The initial learning rates for three learning-based methods were separately optimized on validation data.

We made a revision to indicate the experimental optimization of initial learning rates in section 3.6: "The initial learning rates were experimentally optimized on the validation dataset and set to $1e^{-4}$, $1e^{-3}$ and $1e^{-3}$ for Reg-Net, Seg-Net and Segis-Net".

5. *Validation of the quality and spatio-temporal consistency of the segmentation is further solely based on the Dice score, which is problematic as it often fails to accurately capture differences with regard to boundaries on small segments and preservation of volume similarity (Taha and Hanbury, 2015). Adding distance-based metrics (i.e., surface-based Hausdorff distance, Average Surface Distance) and a volumetric similarity would be beneficial for the evaluation*.

We agree with the reviewer that it is insufficient to validate on the Dice score only. Therefore, we evaluated the method on a number of metrics, including tract-specific volume reproducibility (Figure 6b) which is equivalent to the suggested volumetric similarity. Additionally, the reproducibility of tract-specific diffusion measures (Figure 6 c; d) further indicates the reliability of the method in practical epidemiological and clinical imaging studies.

6. *The improved performance of registration network by inclusion of segmentation task was already reported by others and should be added in the discussion (joint learning of segmentation and registration tasks are highly correlated which helps in boosting performance, was also shown in e.g. Voxelmorph (Balakrishnan et al., 2019) , NeurReg (Zhu et al., 2020), U-ResNet (Estiennet 2020)).*

In the revised version, we added the mentioned papers to the introduction, where we explain the concepts of segmentation-based registration (Balakrishnan et al., 2019; Zhu et al., 2020) and single-step methods (Estienne et al., 2019, 2020). Also, a direct comparison with two of those methods was added to section 4.6 (Balakrishnan et al., 2019; Estienne et al., 2019) and discussed on the results.

Minor:

1. *Include citation of DeepAtlas (Xu et al., 2019) in the introduction ("… several methods for simultaneous segmentation and deformable registration have been proposed.."). Maybe also add this to the discussion, as the network is nearly identical to the Segis-Net (published at same MICCAI conference).*

We added the mentioned paper to the introduction, where we explain the concept of joint estimation: "2) "joint estimation" alternately updates (separate) models in a multi-step optimization. Although the initialization and robustness of joint estimation can be influenced by the selection of the order to optimize and the criteria to switch tasks, this approach is preferred as it requires less computation power and allows to use task-specific training datasets (Yezzi et al., 2003; Wyatt and Noble, 2003; Ashburner and Friston, 2005; Pohl et al., 2006; Parisot et al., 2014; Gooya et al., 2011; Cheng et al., 2017; **Xu and Niethammer, 2019**)".

2. *What is n=3 in paragraph 2.3.1 (...we focus on the analysis of 3D images (n=3)...)?*

We denote the dimension of image as n (section 2.1, 2.2), in which n=3 indicates 3D images.

We made a revision to omit "(n=3)", since "we focus on the analysis of 3D images" is indeed more clear.

3. *Typo: non-linerities (→ non-linearities) in paragraph 2.3.1 (...and a leaky ReLU layer for modelling non-linearities...)*

We corrected the typo to "non-linearity".

4. *Figure 8 and 9 should have extended descriptions: explain the three numbers in each ConvUnit (what is K0, K1, K2→ number of input feature maps, number of filters for convolution step 1 and number of filters for convolution step 2), add a header for the conv unit blocks in the legend (white = Conv Unit 1, gray = Conv Unit 2, blue = Conv Unit 3?)*

We made a revision to include the suggested changes to add a header and description for the conv units in the legend. We additionally added a description to the figure caption: “The number of convolution kernels used in Conv Unit 01 is denoted as [K1, …, Kn], and that for Conv Unit 02 as[K1]”.

5. *Paragraph 3.1.6 Implementation - In my opinion, it would be better to motivate the use of the different input modalities for the segmentation and registration network here instead of adding it in the discussion (because you do not do any comparison with regard to different input modalities) → Relevant part from the discussion: "...The diffusion tensor image contains more directional information of fiber population thereby being favored for the segmentation task, in which we expect to be challenging because of the complex tract geometry, fiber crossings and the clinical-quality resolution….")*

We moved the motivation of the use of the different input imaging modalities to the implementation section: “Specifically, segmentation was based on the diffusion tensor image, as it contains directional information of fiber populations and has been shown to be optimal in the present setting of clinical-quality resolution (Li et al., 2020a).”

6. *Overview Tables with the results for Fig. 3-6 would be helpful (in appendix maybe) as the differences are hard to see in the graphs for some evaluations.*

We added a table with the results of Fig. 3-6 in the supplementary files.

References:

Li et al., 2019: "A hybrid deep learning framework for integrated segmentation and registration: evaluation on longitudinal white matter tract changes." in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 645-653 (2019).

Balakrishnan et al., 2019: "VoxelMorph: a learning framework for deformable medical image registration.", in IEEE Trans. Med. Imag. (2019).

Estiennet et al., 2019: "U-ReSNet: Ultimate coupling of registration and segmentation with deep nets.", in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 310-319 (2019).

Estiennet et al., 2020: "Deep Learning-Based Concurrent Brain Registration and Tumor Segmentation", in Frontiers in Computational Neuroscience, vol 14, p 17 (2020).

Avants et al., 2011: "A reproducible evaluation of ANTs similarity metric performance in brain image registration", in NeuroImage, vol 53, pp. 2033-2044 (2011).

Modat et al., 2010: "Fast free-form deformation using graphics processing units in Computer Methods And Programs.", in Biomedicine, vol 98, pp. 278-284 (2010).

Taha and Hanbury. "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool." BMC medical imaging 15.1 (2015): 29.

Zhu et al., 2020: "NeurReg: Neural Registration and Its Application to Image Segmentation.", in WACV 2020.

Xu and Niethammer, 2019: "DeepAtlas: Joint Semi-Supervised Learning ofImage Registration and Segmentation.", in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 420-429 (2019).

Longitudinal diffusion MRI analysis using Segis-Net: a single-step deep-learning framework for simultaneous segmentation and registration

Bo Li[a], Wiro J. Niessen[a,b], Stefan Klein[a], Marius de Groot[a,c], M. Arfan Ikram[a,c,d], Meike W. Vernooij[a,c], Esther E. Bron[a]

[a]*Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, the Netherlands*
[b]*Imaging Physics, Applied Sciences, Delft University of Technology, the Netherlands*
[c]*Department of Epidemiology, Erasmus MC, Rotterdam, the Netherlands*
[d]*Department of Neurology, Erasmus MC, Rotterdam, the Netherlands*

Abstract

This work presents a single-step deep-learning framework for longitudinal image analysis, coined Segis-Net. To optimally exploit information available in longitudinal data, this method concurrently learns a multi-class segmentation and nonlinear registration. Segmentation and registration are modeled using a convolutional neural network and optimized simultaneously for their mutual benefit. An objective function that optimizes spatial correspondence for the segmented structures across time-points is proposed. We applied Segis-Net to the analysis of white matter tracts from N=8045 longitudinal brain MRI datasets of 3249 elderly individuals. Segis-Net approach showed a significant increase in registration accuracy, spatio-temporal segmentation consistency, and reproducibility comparing with two multistage pipelines. This also led to a significant reduction in the sample-size that would be required to achieve the same statistical power in analyzing tract-specific measures. Thus, we expect that Segis-Net can serve as a new reliable tool to support longitudinal imaging studies to investigate macro- and microstructural brain changes over time.

Keywords: Segmentation, Registration, Diffusion MRI, Deep Learning, CNN, Longitudinal, White Matter Tract

1. Introduction

The increasing availability of longitudinal imaging data is expanding our ability to capture and characterize progressive anatomical changes, ranging from normal changes in the life span, to responses along disease trajectories or therapeutic actions. Compared to cross-sectional studies, longitudinal imaging studies have the advantage of allowing to trace the order of events at the individual level and to correct for the confounding effect of time-invariant individual differences (Van der Krieke et al., 2017). They are thus considered to be more accurate and sensitive in capturing subtle changes over time. To analyze spatio-temporal changes from longitudinal imaging data, a tailored framework that involves both segmentation and registration is required to segment the structures-of-interest and to register temporal frames. This can be achieved by directly combining two existing segmentation and registration tools, which are often designed for cross-sectional studies. However, the information offered in longitudinal data remains underutilized.

Various studies have shown that combining segmentation and registration at the stage of algorithm optimization can lead to improved performance. A popular combination strategy is to use the output of one task to optimize the other. Registration can be improved by using segmentation-level correspondences as input for deformation initialization (Dai and Khorram, 1999; Postelnicu et al., 2008) and optimization (De Groot et al., 2013b; Rohé et al., 2017; Hu et al., 2018; Balakrishnan et al., 2019; Bastiaansen et al., 2020; Zhu et al., 2020). Likewise, segmentation can benefit from registration by propagating anatomical information to subsequent frames, as has been shown in classical multi-atlas based segmentation methods (Fischl et al., 2002; Vakalopoulou et al., 2018) and in recent data-augmentation techniques which introduce labels to support unsupervised (Pathak et al., 2017) and weakly-supervised segmentation (Bortsova et al., 2019; Vlontzos and Mikolajczyk, 2018).

Other approaches combine the optimization of parameters from both tasks on a deeper level. Wyatt and Noble (2003) subgrouped these methods into two types according to the way in which they update their parameters: 1) "simultaneous estimation" that updates both the class labels and the transformations in a single-step optimization, and 2) "joint estimation" that alternately updates (separate) models in a multi-step optimization. Although the initialization and robustness of joint estimation can be influenced by the selection of the order to optimize and the criteria to switch tasks, this approach is preferred as it requires less computation power and allows to use task-specific training datasets (Yezzi et al., 2003; Wyatt and Noble, 2003; Ashburner and Friston, 2005; Pohl et al., 2006; Parisot et al., 2014; Gooya et al., 2011; Cheng et al., 2017; Xu and Niethammer, 2019). Simultaneous estimation is expected to be more ac-

Email address: b.li@erasmusmc.nl (Bo Li)

Abbreviations: MRI, Magnetic Resonance Imaging; DTI, Diffusion Tensor Imaging; FA, Fractional Anisotropy; MD, Mean Diffusivity; TE, Echo Time; TR, Repetition Time

curate, as it fully exploits the conditional correlations between two tasks that can be discounted in sequential processing (Ashburner and Friston, 2005). In addition, simultaneous estimation can explicitly optimize performances that rely on both tasks. We expect that this advantage has a large potential in improving the reliability of analysis of longitudinal imaging data, for instance by optimizing the spatio-temporal consistency of the segmentation. With the growing capability of modeling and computation by deep learning techniques, several simultaneous methods have been proposed and coupled segmentation with deformable registration in different ways, either for 2D (Qin et al., 2018) or 3D images (Li et al., 2019; Estienne et al., 2019, 2020).

Diffusion magnetic resonance imaging (MRI) is a non-invasive imaging technique that measures the diffusion of water in-vivo and can be used to quantitatively characterize white matter (WM) microstructure. The quantitative nature of diffusion MRI makes its derived measures, such as diffusion tensor imaging (DTI) metrics (Le Bihan et al., 2001), very suitable for longitudinal analysis. In addition, diffusion measures are likely to be more sensitive than structural measures in the early detection of changes in WM, and are therefore promising for the identification of subtle changes that relate to the early stages of the disease (Niessen, 2016), for instance in studying dementia subtypes (Meijboom et al., 2019). Longitudinal diffusion MRI has been widely studied at various levels, i.e., from regions-of-interest (Sullivan et al., 2010; Keihaninejad et al., 2013), to tract level (Lebel and Beaulieu, 2011; Yendiki et al., 2016; Meijboom et al., 2019; Dimond et al., 2020), and voxel level (Barrick et al., 2010; Farbota et al., 2012; De Groot et al., 2016). Since WM tracts are functionally grouped axonal fibers and thought to subserve particular brain functions, tract-specific investigation may highlight categorical differences in vulnerability to neurodegeneration and bridge the interpretation of imaging biomarkers with clinical symptoms.

Segmentation of WM tracts is however non-trivial because tracts cannot be identified directly from diffusion MRI, i.e., there is no in-vivo "gold standard" for tract (Crick and Jones, 1993), and because their anatomy can be complex. WM tracts are commonly segmented based on diffusion tractography by reconstruction of potential fiber pathways (Conturo et al., 1999). Recently, deep learning based methods, in particular using convolutional neural networks (CNN), have emerged and showed promising accuracy and efficiency in segmenting WM tracts (Li et al., 2018, 2020a; Wasserthal et al., 2018).

In the present work we focus on a CNN-based framework for longitudinal analysis of WM tracts, i.e., Segis-Net, and investigate the value of simultaneous optimization of segmentation and registration in this setting. In Li et al. (2019), we introduced a generic framework for simultaneous optimization, in which increased accuracies of both tasks were observed in a pilot analysis of a single tract (forceps minor; FMI). In this paper, we extend the tract-specific method by enabling concurrent segmentation of multiple tracts, which is a non-trivial task as a voxel can belong to multiple tracts. This also solves the problem of inconsistencies in deformations because of tract-specific ROIs. The registration task within the framework is updated to learn only local deformations rather than an end-to-end composite including rigid transformation, as brain local changes over time is a focus in longitudinal imaging studies. In addition, we compare the performance of Segis-Net to two multistage pipelines based on both classical and deep learning algorithms. The segmentation accuracy, registration accuracy, spatio-temporal consistency of segmentation, and reproducibility of segmentation and tract-specific measures of the three pipelines are quantitatively evaluated. Also, we evaluate the sample-size reduction that can be achieved in the imaging analysis of WM tracts to provide insight into the practical value of the methods in clinical applications.

2. Method

In this section, we first describe how the segmentation and registration tasks are individually modeled using CNN-based approaches. Subsequently, we present the proposed Segis-Net that integrates both tasks in a single-step CNN framework.

2.1. CNN-based image segmentation

Given a n-D image I which can be described by either intensity values, multi-channel features or directional tensors, the goal of CNN-based segmentation is to automatically infer, for each voxel $x \in \mathbb{R}^n$, its probability of belonging to the structure $k \in [1, K]$, i.e., voxel-wise classification. The CNN model can be interpreted as a parameterized mapping function $\mathcal{F}_{\mathbf{\Theta}}$ such that the segmented structures spatially correspond to a segmentation ground truth with multiple channels $\mathcal{S} = \{S_1, ..., S_K\}$. The estimation of the segmentation is formulated as:

$$\hat{\mathcal{S}} = \mathcal{F}_{\mathbf{\Theta}}(I). \tag{1}$$

$\mathcal{F}_{\mathbf{\Theta}}$ is commonly modeled by a nested series of convolutions, non-linearity, normalization, and re-sampling operations embedded in the network architecture. $\mathbf{\Theta}$ indicate trainable parameters, i.e., weights inside convolution kernels.

The procedure of estimating parameter $\mathbf{\Theta}$ is then defined as an optimization with respect to a loss function $\mathcal{L}_{seg}$, aiming at minimizing the classification error over all the N pairs of training samples $\{(\mathcal{S}^i, I^i)\}_{i=1}^N$, i.e.,

$$\mathbf{\Theta} \leftarrow \underset{\mathbf{\Theta}}{\operatorname{argmin}} \sum_{i=1}^{N} \mathcal{L}_{seg}\left(\mathcal{S}^i, \mathcal{F}_{\mathbf{\Theta}}(I^i)\right). \tag{2}$$

The loss function comprises metrics that quantify the difference between the prediction and the ground truth. In this study, $\mathcal{L}_{seg}$ is the average Dice coefficient (Dice, 1945) over all K structures:

$$\mathcal{L}_{seg}(\mathcal{S}, \hat{\mathcal{S}}) = -\frac{2}{K} \sum_{k=1}^{K} \frac{\sum_x S_k \hat{S}_k}{\sum_x (S_k)^2 + \sum_x (\hat{S}_k)^2}. \tag{3}$$

After estimation of the map function $\mathcal{F}_{\mathbf{\Theta}}$, a probabilistic prediction for the structures of interest $\hat{\mathcal{S}}$ in a given image can be inferred (Eq.1).

2.2. CNN-based deformable registration

Let us consider a pair of n-D images, I_s in the source space $\Omega_s \subset \mathbb{R}^n$, and I_t in the target space $\Omega_t \subset \mathbb{R}^n$, which contain a common structure to be aligned. The spatial correspondence between images can be established by estimating a dense displacement field $\boldsymbol{\phi}$, such that $I_s \circ \boldsymbol{\phi}$ and I_t correspond spatially.

In line with the hierarchical optimization scheme of classical registration algorithms, most existing learning-based registration methods use affine alignment as a prepossessing, in which case the displacement field denotes the composition of affine and deformable transform, i.e., $\boldsymbol{\phi} = \boldsymbol{\phi}_A \circ \boldsymbol{\phi}_D$. To estimate $\boldsymbol{\phi}_D$, the CNN model can be interpreted as a shared domain-invariant mapping function $\mathcal{G}_{\boldsymbol{\Psi}}$ such that for any unseen pair of images a most likely transformation between them can be inferred without pair-specific optimization, i.e.,

$$\hat{\boldsymbol{\phi}}_D = \mathcal{G}_{\boldsymbol{\Psi}}(I_t, I_s \circ \boldsymbol{\phi}_A) \tag{4}$$

$$\implies \quad I_s \circ \hat{\boldsymbol{\phi}} \leftarrow \mathcal{G}_{\boldsymbol{\Psi}}(I_t, I_s, \boldsymbol{\phi}_A). \tag{5}$$

The parameters $\boldsymbol{\Psi}$ of the mapping function are optimized based on a registration dissimilarity loss $\mathcal{L}_{reg}$, aimed at minimizing the registration error. Meanwhile, to penalize large deviations of deformation and preserve anatomical topology during transformations, a deformation smoothness term $\mathcal{L}_{def}$ is commonly included in the loss function. In this work, we use the mean squared error based on intensities for $\mathcal{L}_{reg}$ and the average spatial gradients of the displacement field for $\mathcal{L}_{def}$, i.e.,

$$\mathcal{L}_{reg}(I_t, I_s \circ \hat{\boldsymbol{\phi}}) = \frac{1}{|\Omega_t|} \|I_t - I_s \circ \hat{\boldsymbol{\phi}}\|_2^2, \tag{6}$$

$$\mathcal{L}_{def}(\hat{\boldsymbol{\phi}}_D) = \frac{1}{|\Omega_t|} \|\nabla \hat{\boldsymbol{\phi}}_D\|_2^2. \tag{7}$$

Combining Eq. (4), (5), (6) and (7), the estimation of $\boldsymbol{\Psi}$ over all the N training samples $\{(I_t^i, I_s^i, \boldsymbol{\phi}_A^i)\}_{i=1}^N$ can be formulated as:

$$\boldsymbol{\Psi} \leftarrow \underset{\boldsymbol{\Psi}}{\operatorname{argmin}} \sum_{i=1}^{N} \mathcal{L}_{reg}(I_t^i, \mathcal{G}_{\boldsymbol{\Psi}}(I_t^i, I_s^i, \boldsymbol{\phi}_A^i)) + \mathcal{L}_{def}(\mathcal{G}_{\boldsymbol{\Psi}}(I_t^i, I_s^i \circ \boldsymbol{\phi}_A^i)). \tag{8}$$

2.3. Simultaneous estimation of segmentation and registration

In this work, we aim to simultaneously estimate the parameters for segmentation ($\boldsymbol{\Theta}$) and for registration ($\boldsymbol{\Psi}$) in a single-step optimization. For this purpose, we integrate the segmentation and registration function $\mathcal{F}_{\boldsymbol{\Theta}}$ and $\mathcal{G}_{\boldsymbol{\Psi}}$ using an end-to-end optimization with the Segis-Net. The loss function of the Segis-Net is designed to meet the joint objective of both tasks and meanwhile to optimize the spatio-temporal consistency of segmentation which rely on both tasks. The overview of the proposed framework is illustrated in Fig. 1. We describe the framework architecture and loss function in the following paragraphs.

2.3.1. Segis-Net framework

In the present study, we focus on the analysis of 3D images and utilize 3D convolutions for the Segis-Net framework. The framework involves function $\mathcal{F}_{\boldsymbol{\Theta}}$ and $\mathcal{G}_{\boldsymbol{\Phi}}$ as two parallel streams that interact on their outputs. In order to eliminate the loss in image quality caused by multiple interpolations, Segis-Net warps source images with only the composite displacement fields ($\boldsymbol{\phi}$) by taking as input the original source image (I_s) and pre-estimated affine matrix ($\boldsymbol{\phi}_A$). This design has additional advantages over existing methods that prepare all ordered pairs of affine-aligned images in disk storage, as only up to half the storage is needed and as it can be flexibly applied to related images in the same space such as the DTI metrics.

$\mathcal{F}_{\boldsymbol{\Theta}}$ outputs a set of probabilistic segmentations ($\hat{\mathcal{S}}_s$) of the source image. $\mathcal{G}_{\boldsymbol{\Psi}}$ outputs a dense local displacement $\hat{\boldsymbol{\phi}}_D$ along the x, y, and z axes. The source image and its segmentations are subsequently warped into the target space using the composed displacement field. The warp operation is implemented by a computational layer with differentiable trilinear interpolation (Jaderberg et al., 2015; Balakrishnan et al., 2019). The segmentation and registration streams have independent network architectures which are only connected by the output, i.e., the transformed source-segmentation to the target space ($\hat{\mathcal{S}}_s \circ \hat{\boldsymbol{\phi}}$). Thus, they can be applied separately after taking advantage of the simultaneous optimization. The Segis-Net framework gives four outputs during training:

1. The segmentation of the structures of interest from the source image ($\hat{\mathcal{S}}_s$),
2. A local displacement field between the source and target images ($\hat{\boldsymbol{\phi}}_D$),
3. The warped source image in the target space ($I_s \circ \hat{\boldsymbol{\phi}}$),
4. The warped source segmentations in the target space ($\hat{\mathcal{S}}_s \circ \hat{\boldsymbol{\phi}}$).

We propose a generic framework where the architecture of each stream can be adapted based on specific applications. For the particular network used in this study, we encoded two streams with a U-Net architecture, that was modified as detailed below (Ronneberger et al., 2015). In short, each stream was composed of an encoder and decoder path with skip connections of feature pyramid at multiple scales in order to merge coarse- and fine-convolved features, similar to the multi-resolution strategy used in classical algorithms to increase robustness. The encoder paths with max-pooling operation between convolution layers gradually extract abstract features for the target anatomy ($\mathcal{F}_{\boldsymbol{\Theta}}$) and global transformation between images ($\mathcal{G}_{\boldsymbol{\Psi}}$). Subsequently, the decoder paths restore the details in segmentations ($\mathcal{F}_{\boldsymbol{\Theta}}$) and refine local deformations ($\mathcal{G}_{\boldsymbol{\Psi}}$) by linear up-sampling the feature maps and concatenating them with the coarse counterpart at the same scale. The convolution layers produce a set of feature maps by individually convolving inputs with 3D kernels of size $(3, 3, 3)$, followed by batch normalization (Ioffe and Szegedy, 2015) and a leaky ReLu layer ($a = 0.2$) for modeling non-linearity (Maas et al., 2013). For the segmentation stream ($\mathcal{F}_{\boldsymbol{\Theta}}$), we split the output layer into sub-branches to facilitate multi-class classification for voxels

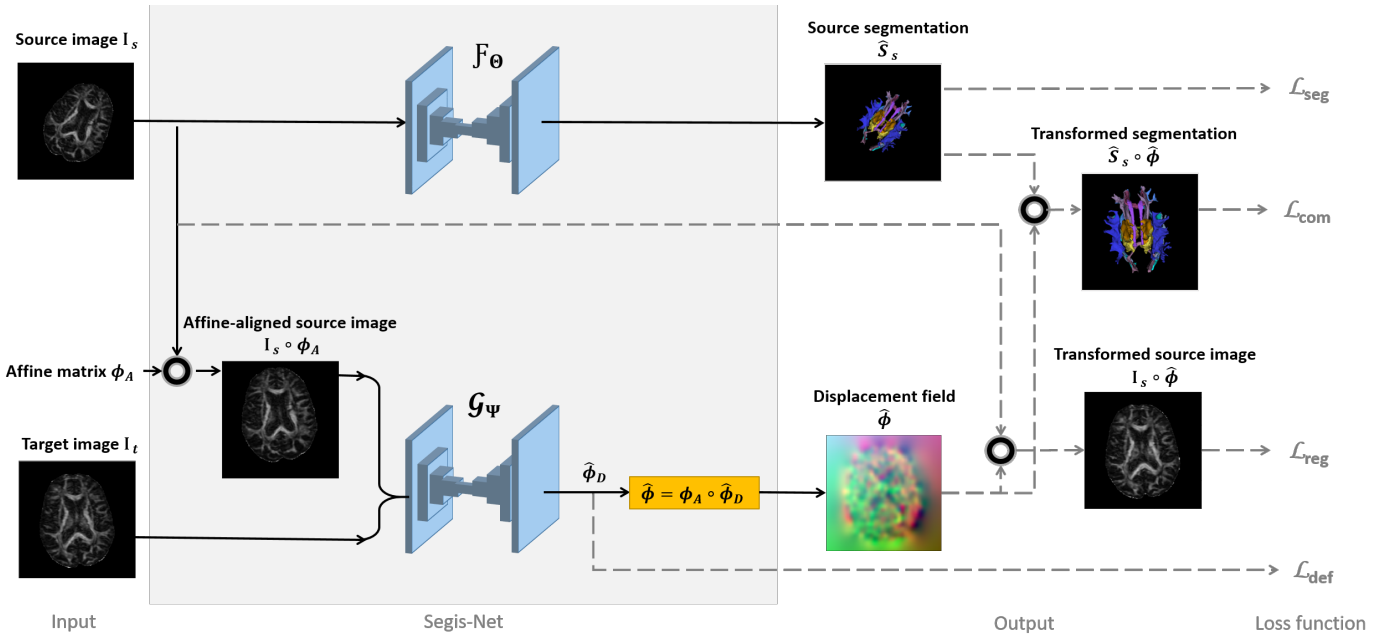


Figure 1: Overview of the Segis-Net framework. Θ and Ψ denote the parameters of the segmentation ($\mathcal{F}_\Theta$) and registration ($\mathcal{G}_\Psi$) function, respectively. Black circle indicates spatial warp with affine matrix (ϕ_A) or the composite displacement field ($\hat{\phi}$). The concatenation of the affine-aligned images is used as the input for $\mathcal{G}_\Psi$. Loss function consists of $\mathcal{L}_{seg}$, $\mathcal{L}_{com}$, $\mathcal{L}_{reg}$ and $\mathcal{L}_{def}$ terms. Solid lines indicate the primary workflow of the method; dash lines indicate that only implemented during training or that can be adapted for applications.

with multiple labels. The final layer of the sub-branches consisted of a $(1, 1, 1)$ convolution and a sigmoid activation. For the registration stream, the output layer was a convolution layer with three kernels that yielded the local displacement $\hat{\phi}_D$. We provide detailed implementation of the network architecture in the supplementary material (Figure 8 and 9).

2.3.2. *Segis-Net loss function*

The loss function of Segis-Net is composed of four terms that measure segmentation accuracy ($\mathcal{L}_{seg}$, Eq. 3), intensity similarity between registered images ($\mathcal{L}_{reg}$, Eq. 6), deformation field smoothness ($\mathcal{L}_{def}$, Eq. 7), and longitudinal composite of registration and segmentation ($\mathcal{L}_{com}$, Eq. 11). It is formulated as:

$$\begin{aligned}\mathcal{L} = &\mathcal{L}_{seg}(\mathcal{S}_s, \hat{\mathcal{S}}_s) \\ &+ \alpha\mathcal{L}_{reg}(I_t, I_s \circ \hat{\phi})) + \beta\mathcal{L}_{def}(\hat{\phi}_D) \\ &+ \gamma\mathcal{L}_{com}(\mathcal{S}_t, \hat{\mathcal{S}}_s \circ \hat{\phi}),\end{aligned} \tag{9}$$

and optimized for Θ and Ψ over all N training samples $\{(\mathcal{S}_t^i, \mathcal{S}_s^i, I_t^i, I_s^i, \phi_A^i)\}_{i=1}^N$:

$$\begin{aligned}\Theta, \Psi \leftarrow \underset{\Theta,\Psi}{\operatorname{argmin}} &\sum_{i=1}^{N} \mathcal{L}_{seg}\left(\mathcal{S}_s^i, \mathcal{F}_\Theta(I_s^i)\right) \\ &+ \alpha\mathcal{L}_{reg}(I_t^i, \mathcal{G}_\Psi(I_t^i, I_s^i, \phi_A^i)) + \beta\mathcal{L}_{def}(\mathcal{G}_\Psi(I_t^i, I_s^i \circ \phi_A^i)) \\ &+ \gamma\mathcal{L}_{com}(\mathcal{S}_t^i, \mathcal{F}_\Theta(I_s^i)), \mathcal{G}_\Psi(I_t^i, I_s^i, \phi_A^i)).\end{aligned} \tag{10}$$

We quantify the longitudinal composite loss term using the average Dice coefficient over all K structures:

$$\mathcal{L}_{com}(\mathcal{S}_t, \hat{\mathcal{S}}_s \circ \hat{\phi}) = -\frac{2}{K}\sum_{k=1}^{K} \frac{\sum_{x\in\Omega_t} S_t^k(\hat{S}_s^k \circ \hat{\phi})}{\sum_{x\in\Omega_t}(S_t^k)^2 + \sum_{x\in\Omega_t}(\hat{S}_s^k \circ \hat{\phi})^2}. \tag{11}$$

In longitudinal imaging studies, the spatial correspondence in segmentation depends on the performance of both the segmentation and the registration procedure. Besides an explicit optimization of correspondence, the $\mathcal{L}_{com}$ term also exploits longitudinal information to boost both tasks, which introduces some degree of augmentation and regularization for registration and on the other hand constraints and prior knowledge for segmentation.

The hyperparameters α and γ balance the loss magnitude of segmentation, registration, and their interdependent composite. The degree of regularization is described by β (or β/α). The procedure of simultaneous optimization is summarized with pseudo code in supplementary material (Algorithm 1).

3. Application to diffusion MRI

The performance of Segis-Net is demonstrated by analyzing white matter tracts in a large diffusion MRI dataset, and compared to that of two multi-stage pipelines, in which segmentation and registration are independently optimized. Performance is evaluated in a longitudinal setting where multiple time-points from the same individual are available.

3.1. Dataset

The Rotterdam Study is a prospective and population-based study targeting causes and consequences of age-related diseases (Ikram et al., 2020). For the present analysis, we included 3249 individuals who underwent diffusion MRI scanning twice or more often, resulting in $N = 8045$ scans. The mean age at first scan was 61.2±9.4 years (range: 45.7−91.1 years). The number of female participants was 1780 (54.8%). A flowchart for the inclusion, exclusion, and split of the datasets is shown in Figure 2. We split the data into two subsets. The larger subset was repeatedly acquired in a time interval of 1–5 years ($N = 7770$ scans from 3166 individuals). In these long time-interval scans, it is expected that brain microstructure changes due to aging exist. By matching any two time-points from the same individual regardless of the visiting order, these long time-interval scans can be grouped into 6043 pairs. We used 5175 pairs of scans as training data, 200 pairs as validation data to tune the hyperparameters, monitor the decay of learning rate and select the optimal epoch, and used an independent cohort of 668 pairs for testing. The remaining scans from the smaller subset were from 97 individuals who were scanned twice within a month. No changes in brain macro- and microstructure were expected within such a short time-interval. We used these scans for evaluation of reproducibility of the algorithm. The data split was based on the participants, namely, we made sure that scans from the same participant ended up in either training, validation, or test dataset.

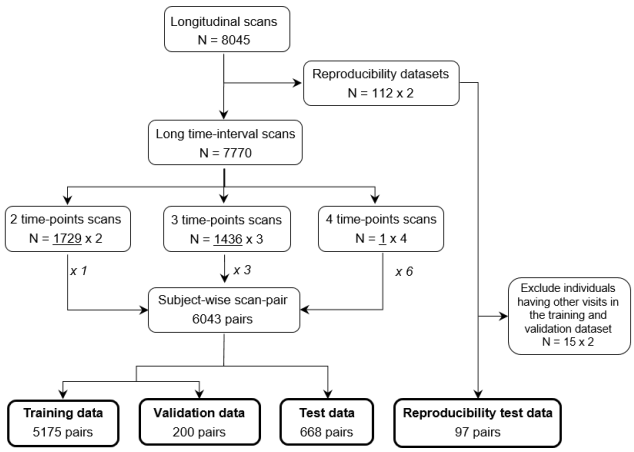


Figure 2: A flowchart for the inclusion, exclusion, and split of the datasets.

3.2. MRI acquisition

Scans were acquired on a 1.5T MRI scanner (GE Signa Excite). The acquisition parameters for structural and diffusion MRI can be found in Ikram et al. (2011). Specifically, diffusion MRI was scanned with the following parameters: TR/TE =8575ms/82.6ms, imaging matrix of 64×96, FOV=21×21cm^2, 35 contiguous slices with slice thickness 3.5 mm, 25 diffusion weighted volumes with a b-value of 1000s/mm^2 and 3 non-weighted volumes (b-value = 0s/mm^2). The voxel size was resampled from $3.3 \times 2.2 \times 3.5mm^3$ to $1mm^3$ as required for probabilistic tractography (Behrens et al., 2007).

3.3. Image preprocessing

Diffusion data were preprocessed using a standardized pipeline (Koppelmans et al., 2014). In short, motion and eddy currents were corrected by affine co-registration of all diffusion weighted volumes to the averaged b0 volumes, including correction of gradient vector directions using Elastix software (Klein et al., 2010). Diffusion tensors were estimated with a Levenberg-Marquard non-linear least-squares optimization algorithm (Leemans et al., 2009). We subsequently computed DTI measures: fractional anisotropy (FA) and mean diffusivity (MD). Due to noise, tensor estimation failed in a small proportion of voxels, resulting in significant outliers. Outlier voxels with a tensor norm (Frobenius norm) larger than $0.1mm^2/s$ were set to zero (Zhang et al., 2007). Brain tissue masks including WM and gray matter segmentations were obtained based on structural imaging (Vrooman et al., 2007) and applied to the diffusion tensor images. In this study, we used a ROI of $112 \times 208 \times 112$ voxels to analyze six WM tracts, including left and right cingulate gyrus part of cingulum (CGC), left and right parahippocampal part of cingulum (CGH), forceps major (FMA) and forceps minor (FMI). Diffusion tensor image (with six components) was image-wise normalized to zero-mean and standard deviation of one. The affine matrix ($\boldsymbol{\phi}_A$) of each image pair was estimated by optimizing the mutual information of FA images using Elastix software.

3.4. Reference segmentations

The segmentation labels for model training and evaluation were generated using a probabilistic tractography and atlas-based segmentation method by De Groot et al. (2015). The resulting tract-density images for each tract were normalized by division with the total number of tracts in the tractography run. Finally, tract-specific thresholds for the normalized density images were established by maximizing the reproducibility of FA measures on a subset of 30 participants, which were included in our reproducibility dataset as well (De Groot et al., 2013b).

3.5. Baseline multi-stage pipelines

We compared the performance of the proposed Segis-Net with two multi-stage pipelines that consist of either non-learning-based or learning-based algorithms.

First, a non-learning-based *Classical* pipeline was built using an existing tractography-based segmentation algorithm as detailed in Section 3.4, and deformable registration algorithm Elastix (Klein et al., 2010). Elastix was adopted as a competing classical registration method since it has been widely used on our dataset and thereby an optimal parameter setting can be applied for performance comparison. Elastix is designed to run in a cascade of resolutions, and offers the choice between multiple objective functions and multiple optimizers including an efficient adaptive stochastic gradient descent optimizer (Klein et al., 2009). For Elastix (version 4.8), we used a rigid, affine, and B-spline transformation model consecutively by maximizing mutual information between images. The B-spline transformation of spline order 3 was implemented using a multi-resolution framework with isotropic control-point spacing of

24, 12, and 6 *mm* in three-level resolutions. The maximum number of iterations was 1024.

Second, we built a learning-based *CNN* pipeline using components from the proposed Segis-Net to evaluate the sole contribution of simultaneous optimization. In this pipeline, we split the integrated segmentation $\mathcal{F}_{\mathbf{\Theta}}$ and registration stream $\mathcal{G}_{\mathbf{\Phi}}$ into two separate neural networks for independent optimization. Subsequently, the segmented images and estimated transformations were combined. The segmentation network had the same architecture as that for $\mathcal{F}_{\mathbf{\Theta}}$, except being independently optimized using the segmentation accuracy $\mathcal{L}_{seg}$ term. As this is a typical setting for CNN-based segmentation approaches (Li et al., 2018; Ronneberger et al., 2015), we denote it as *Seg-Net*. Similarly, the registration network denoted as *Reg-Net* had the same setting as that for $\mathcal{G}_{\mathbf{\Psi}}$, except being independently optimized using registration similarity $\mathcal{L}_{reg}$ and regularization $\mathcal{L}_{def}$ terms (Balakrishnan et al., 2019). We ensured that the training dataset for the *Seg-Net* and *Reg-Net* was the same as that used for the Segis-Net framework.

3.6. Implementation

For this diffusion MRI application, the segmentation ($\mathcal{F}_{\mathbf{\Theta}}$) and registration ($\mathcal{G}_{\mathbf{\Psi}}$) components of Segis-Net used different input images. Specifically, segmentation was based on the diffusion tensor image, as it contains directional information of fiber populations and was shown to be optimal in the present setting of clinical-quality resolution (Li et al., 2020a). For spatial alignment, we adopted the input the commonly used scalar-value FA map derived from diffusion tensor imaging.

To mitigate class imbalance and to improve computational efficiency, we combined the reference segmentation for the six tracts (Section 3.3) into a three-channel map for using it as the segmentation ground truth S. This combination was possible since only few crossing fibers are expected between codirectional WM tracts (e.g., FMI and FMA). To evaluate performance on individual tracts after training, we extracted the two largest components from each of the three channels of the probabilistic prediction and subsequently identified the left and right (for CGC and CGH) or the anterior and posterior (for FMI and FMA) tract based on coordinates.

The experiments of model training and evaluation were performed on an NVIDA 1080Ti GPU and an AMD 1920X CPU. CNN-based methods were implemented using Keras-2.2.0 with a Tensorflow-1.4.0 backend and the Adam optimizer (Kingma and Ba, 2014). Weights of convolution kernels were initialized with the Glorot uniform distribution (Glorot and Bengio, 2010). In each training epoch, input images were fed in random batches (size = 1). Loss function hyperparameters were optimized based on segmentation and registration performance on the validation dataset (search range: $[10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}]$); and were set to $\alpha = 10$, $\beta = 0.1$ for *Reg-Net* and subsequently set to $\gamma = 1$ for Segis-Net. The initial learning rates were experimentally optimized on the validation dataset and set to $1e^{-4}$, $1e^{-3}$ and $1e^{-3}$ for *Reg-Net*, *Seg-Net* and Segis-Net, which were decayed with a factor of 0.8 if the validation loss stopped decreasing for 10 epochs (decay condition, Algorithm 1). We stopped the training procedure at the point that the validation loss showed consecutive increases, i.e., early stopping (Bishop, 2006). The parameters of the model with the smallest error with respect to the validation dataset were used.

4. Experiments and results

We applied the proposed Segis-Net and two baseline multi-stage pipelines to analyze six WM tracts on the test and reproducibility datasets. To assess whether difference of performance between approaches was statistically significant, paired t-tests with $\alpha = 0.05$ and Bonferroni correction for controlling the family-wise error of multiple testings were performed. In addition, we compared Segis-Net with two relevant methods in literature.

4.1. Segmentation accuracy

The segmentation accuracy of the proposed Segis-Net was compared with that obtained by the *CNN* pipeline, i.e., *Seg-Net*. We quantified segmentation accuracy with respect to the reference segmentation using the Dice coefficient on the test dataset.

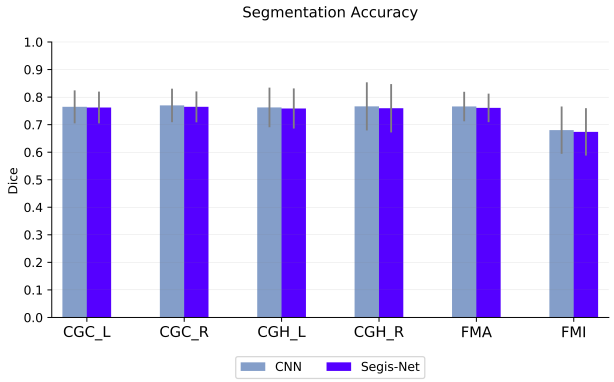


Figure 3: Segmentation accuracies of the *CNN* pipeline and Segis-Net for different tracts. Error bars indicate standard deviations.

The segmentation accuracies of the *CNN* pipeline and Segis-Net were similar for the six tracts (Figure 3). Both methods achieved relatively high accuracy in segmenting cingulum, i.e., the accuracy of left and right CGC and CGH tracts was around 0.76 ± 0.07. The accuracy was lowest for FMI (*CNN*: 0.68 ± 0.09; Segis-Net: 0.67 ± 0.09), which is a thin and arch-shaped tract that is known to be more difficult to segment. Correcting for 6 tests resulted in an adjusted P-value threshold of 8.3×10^{-3}. There was no significant differences in segmentation accuracy between methods.

4.2. Registration accuracy

Registration accuracy of the three pipelines was evaluated with the spatial correlation (SC) similarity on the test dataset. According to the procedure in De Groot et al. (2013b) the estimated transformation was applied to the continuous density

maps of individual tracts obtained from probabilistic tractography, subsequently, the SC similarity between warped density maps was computed as follow:

$$SC_k = \frac{\sum_{x\in\Omega_t} J_t^k (J_s^k \circ \hat{\boldsymbol{\phi}})}{\left(\sum_{x\in\Omega_t} \sqrt{(J_t^k)^2}\right)\left(\sum_{x\in\Omega_t} \sqrt{(J_s^k \circ \hat{\boldsymbol{\phi}})^2}\right)}, \quad (12)$$

where J_t^k and J_s^k indicate intensity of the target and source density image of the tract k.

Despite a lot of intensity variation in the tract density maps across scans due to the probabilistic nature of tractography, higher intensity in general indicates more support for the tract while lower intensity conversely indicates increased uncertainty. Therefore, we assume that SC reflects the spatial correspondence of tracts.

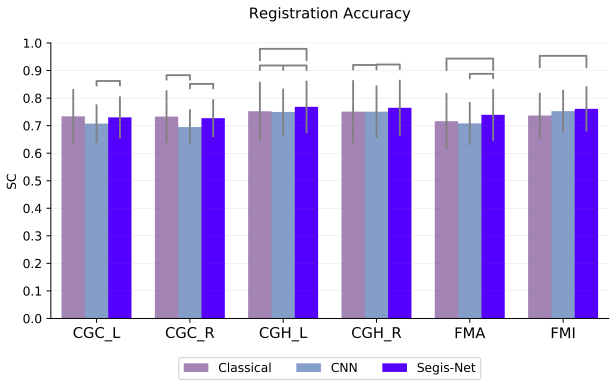


Figure 4: Registration accuracies of the *Classical*, *CNN*, and Segis-Net pipeline as quantified by spatial correlation (SC) of the registered tract density maps. Error bars indicate standard deviations. The bracket hat indicates a significant difference between two methods (t-test, $p < 2.8 \times 10^{-3}$).

The SC in all six tracts was overall highest for the Segis-Net, followed by the *Classical* pipeline (Figure 4). Correcting for 18 tests resulted in Bonferroni adjusted P-value threshold of 2.8×10^{-3}. Segis-Net results yielded a significantly better spatial correspondence than the *Classical* pipeline in the left CGH (Segis-Net vs *Classical* = 0.77 ± 0.09 vs 0.75 ± 0.11), FMA (0.74 ± 0.09 vs 0.72 ± 0.10), and FMI (0.76 ± 0.08 vs 0.74 ± 0.08) tract. Significantly higher registration accuracy over the *CNN* pipeline was observed in the left CGC (Segis-Net vs *CNN* = 0.73 ± 0.08 vs 0.71 ± 0.07), right CGC (0.73 ± 0.07 vs 0.69 ± 0.06), left CGH (0.77 ± 0.09 vs 0.75 ± 0.08), right CGH (0.76 ± 0.10 vs 0.75 ± 0.10), and FMA (0.74 ± 0.09 vs 0.71 ± 0.08) tract. In general, the proposed Segis-Net approach achieved a better spatial correspondence than the two independently optimized registration algorithms using classical and learning-based techniques.

4.3. Spatio-temporal consistency of segmentation

To evaluate the spatio-temporal consistency of segmentation (STCS) for each of the three pipelines, we measured the correspondence between warped segmentation results across time-points using the Dice coefficient. The consistency of each tract was averaged over two directions by reversing the target and source image, which for the tract k can be formulated as:

$$STCS_k = \frac{1}{2}\left(\frac{2|\hat{S}_t^k \cap \hat{S}_s^k \circ \hat{\boldsymbol{\phi}}|}{|\hat{S}_t^k| + |\hat{S}_s^k \circ \hat{\boldsymbol{\phi}}|} + \frac{2|\hat{S}_s^k \cap \hat{S}_t^k \circ \hat{\boldsymbol{\phi}}^{-1}|}{|\hat{S}_s^k| + |\hat{S}_t^k \circ \hat{\boldsymbol{\phi}}^{-1}|}\right). \quad (13)$$

Each pipeline was evaluated as a whole, that is, 1) in *Classical* pipeline the reference segmentation was warped by Elastix algorithm, 2) in *CNN* pipeline the prediction of *Seg-Net* was warped by the predicted transformation of the *Reg-Net*, and 3) in Segis-Net framework the segmentation prediction in native space and the segmentation warped from another time-point were available after a bidirectional test.

The proposed Segis-Net overall showed higher segmentation consistency than the *CNN* and the *Classical* pipeline (Figure 5). Correcting for 18 tests resulted in an adjusted P-value threshold of 2.8×10^{-3}. In comparison with the *CNN* pipeline, Segis-Net results yielded significantly higher spatio-temporal consistency in left CGC (Segis-Net vs *CNN* = 0.83 ± 0.04 vs 0.82 ± 0.04), right CGC (0.83 ± 0.04 vs 0.82 ± 0.06), left CGH (0.82 ± 0.05 vs 0.81 ± 0.05), FMA (0.87 ± 0.02 vs 0.84 ± 0.03), and FMI (0.81 ± 0.05 vs 0.77 ± 0.07) tract. In all six tracts, Segis-Net significantly outperformed the *Classical* pipeline, i.e., in left CGC (Segis-Net vs *Classical* = 0.83 ± 0.04 vs 0.68 ± 0.06), right CGC (0.83 ± 0.04 vs 0.68 ± 0.06), left CGH (0.82 ± 0.05 vs 0.66 ± 0.08), right CGH (0.81 ± 0.05 vs 0.66 ± 0.09), FMA (0.87 ± 0.02 vs 0.69 ± 0.06), and FMI (0.81 ± 0.05 vs 0.57 ± 0.09) tract.

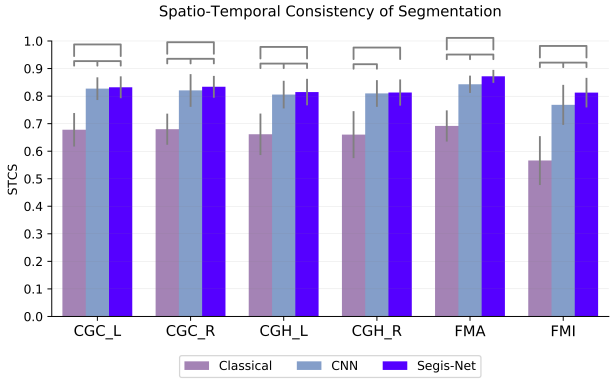


Figure 5: Spatio-temporal consistency of segmentation (STCS) with the *Classical*, *CNN*, and Segis-Net pipeline. Error bars indicate standard deviations. The bracket hat indicates a significant difference between two methods (t-test, $p < 2.8 \times 10^{-3}$).

4.4. Reproducibility of segmentation and measurements

Reproducibility of tract-specific segmentations, volumes, and diffusion metrics of the three pipelines was evaluated using the reproducibility dataset. We quantified voxel-wise agreement between segmentations of repeated scans using Cohen's kappa coefficient (κ). The segmentations $(\hat{S}_t, \hat{S}_s)$ were obtained in the native space, and subsequently aligned $(\hat{S}_s \circ \hat{\boldsymbol{\phi}})$. Kappa κ of the tract k is defined as:

$$\kappa_k = \frac{p_o(\hat{S}_t^k, \hat{S}_s^k \circ \hat{\boldsymbol{\phi}}) - p_e(\hat{S}_t^k, \hat{S}_s^k \circ \hat{\boldsymbol{\phi}})}{1 - p_e(\hat{S}_t^k, \hat{S}_s^k \circ \hat{\boldsymbol{\phi}})}, \quad (14)$$

in which $p_o(\hat{S}_t^k, \hat{S}_s^k \circ \hat{\boldsymbol{\phi}})$ is the observed agreement between $\hat{S}_t^k$ and $\hat{S}_s^k \circ \hat{\boldsymbol{\phi}}$, p_e is the hypothetical probability of the agreement. Given $|\Omega_t|$ being the total number of voxels in the target image, $|S|$ and $|\Omega_t| - |S|$ being the number of tract and non-tract voxels, the observed agreement (i.e., accuracy) is computed as:

$$p_o(\hat{S}_t^k, \hat{S}_s^k \circ \hat{\boldsymbol{\phi}}) = \frac{|\hat{S}_t^k \cap (\hat{S}_s^k \circ \hat{\boldsymbol{\phi}})| + |(1 - \hat{S}_t^k) \cap (1 - (\hat{S}_s^k \circ \hat{\boldsymbol{\phi}}))|}{|\Omega_t|}, \tag{15}$$

the hypothetical probability of the agreement can be formulated as:

$$p_e(\hat{S}_t^k, \hat{S}_s^k \circ \hat{\boldsymbol{\phi}}) = \frac{1}{|\Omega_t|^2}(|\hat{S}_t^k| \times |\hat{S}_s^k \circ \hat{\boldsymbol{\phi}}| + (|\Omega_t| - |\hat{S}_t^k|) \times (|\Omega_t| - |\hat{S}_s^k \circ \hat{\boldsymbol{\phi}}|)). \tag{16}$$

Typically, a $\kappa > 0.60$ indicates "substantial" agreement, and a $\kappa > 0.80$ indicates "almost perfect" agreement (Landis and Koch, 1977).

Similarly, to evaluate the reproducibility of tract-specific measurements, we computed the FA, MD and volume in image native space, and subsequently assessed relative differences in paired scan-rescan measures (m_t , m_s) as an indicator of measurement error (ϵ), i.e.,

$$\epsilon = \frac{2|m_s - m_t|}{(m_s + m_t)} \times 100\%. \tag{17}$$

For FA and MD, the tract-specific measures were quantified as the median of non-zero values within the segmented images. A lower ϵ indicates a better reproducibility.

Figure 6 presents the reproducibility of tract-specific segmentation and measures determined with the three pipelines. The proposed Segis-Net achieved the best segmentation reproducibility, followed by the *CNN* pipeline (Figure 6 (a)); in all six tracts, κ was around 0.80 or higher, indicating "almost perfect" agreements between segmentations of repeated scans. Correcting for 18 tests for each metric resulted in an adjusted P-value threshold of 2.8×10^{-3}, resulting in overall statistically significant improvement by Segis-Net over the *Classical* pipeline. For two tracts, voxel-wise agreement of Segis-Net was significantly higher than that of the *CNN* pipeline, i.e., FMA (Segis-Net vs *CNN* = 0.87 ± 0.03 vs 0.85 ± 0.03) and FMI (0.82 ± 0.06 vs 0.79 ± 0.08).

Additionally, in the evaluation of the reproducibility in tract-specific volume measures, Segis-Net showed the smallest error in all six tracts (Figure 6 (b)). The error of Segis-Net was significantly smaller than the *Classical* pipeline in left CGC (Segis-Net vs *Classical* = $4.8 \pm 4.1\%$ vs $11 \pm 8.8\%$), right CGC ($4.5 \pm 3.9\%$ vs $11 \pm 8.9\%$), left CGH ($7.3 \pm 5.6\%$ vs $11 \pm 9.8\%$), and FMA ($3.4 \pm 2.6\%$ vs $6.6 \pm 5.9\%$) tract. This outperformed the *CNN* pipeline significantly in the FMA tract (Segis-Net vs *CNN* = $3.4 \pm 2.6\%$ vs $4.9 \pm 3.6\%$). Reproducibility of FA and MD measurements was similar for the three methods (Figure 6 (c, d)). For the CGH and left CGC tracts, the reproducibility of FA using the *CNN* pipeline was significantly higher than that of the *Classical* pipeline. Segis-Net outperformed the FA reproducibility of the *Classical* pipeline only in the left CGC tract. For MD, no significant improvement over the *Classical* pipeline was observed. A table (Table 3) with the results of Figure 3-6 is provided in the supplementary files.

4.5. Sample-size reduction

An implication of the reduced measurement error (ϵ) is that fewer participants or time-points would be required to achieve the same statistical power, i.e., a smaller sample size. We followed Diggle et al. (2002) and Reuter et al. (2012) to estimate the percentage of the sample sizes (P) that would be required for each of the three pipelines:

$$P_{ij} = \frac{\sigma_i^2 \times (1 - \rho_i)}{\sigma_j^2 \times (1 - \rho_j)} \times 100\%, \tag{18}$$

where σ_i and σ_j are standard deviations in the measurements determined with the pipeline i and j, and ρ_i and ρ_j are the correlation coefficients between the repeated measurements determined with the two pipelines.

Figure 7 presents the percentage of sample-size reduction that could be achieved by the *CNN* and the proposed Segis-Net compared to the *Classical* pipeline. In line with the reproducibility results, the data analyzed with Segis-Net would overall require the least sample-size to achieve the same statistical power. The percentage of reduction was especially remarkable in volume measures, in which on average only 33.0% of data would be required. The average percentage of reduction was 60.5% for FA and 57.0% for MD. Several percentages of the *CNN* pipeline were smaller than those of the Segis-Net, e.g., in FA measures of CGH and FMA tract (Figure 7 (b)), but its performance showed to be less stable across tracts than the Segis-Net, which in all settings consistently decreased in the required samples over the *Classical* pipeline. The percentage of reduction was generally similar for left/right homologous tracts except for the MD measure in the left of CGH (Figure 7 (c)). This large reduction could be related to the MD reproducibility of the *Classical* pipeline, in which the left CGH tract had a much higher variation in errors comparing with that of the other tracts (Figure 6 (d)).

4.6. Comparison with related methods

In this section, we compared the proposed Segis-Net with two previously published methods that also involve segmentation and registration: 1) U-ReSNet for simultaneous optimization that used a shared feature encoder (Estienne et al., 2019), and 2) VoxelMorph for image registration that used an auxiliary loss term of correspondence in existing segmentation labels (Balakrishnan et al., 2019).

Since the same input for both segmentation and registration tasks was required by the shared feature-encoder of U-ReSNet, i.e., a pair of scalar images, the FA map was used as input. This was different from the FA-based registration and diffusion tensor-based segmentation used in the proposed Segis-Net (section 3.6). For a fair comparison, we applied affine registration to the input image pairs as a pre-processing step. Also, we tuned the hyperparameters on the validation dataset and obtained improved performance by using an initial learning rate of 0.0005 and by clipping the warped segmentation predictions into the range of $[10^{-7}, 1 - 10^{-7}]$.

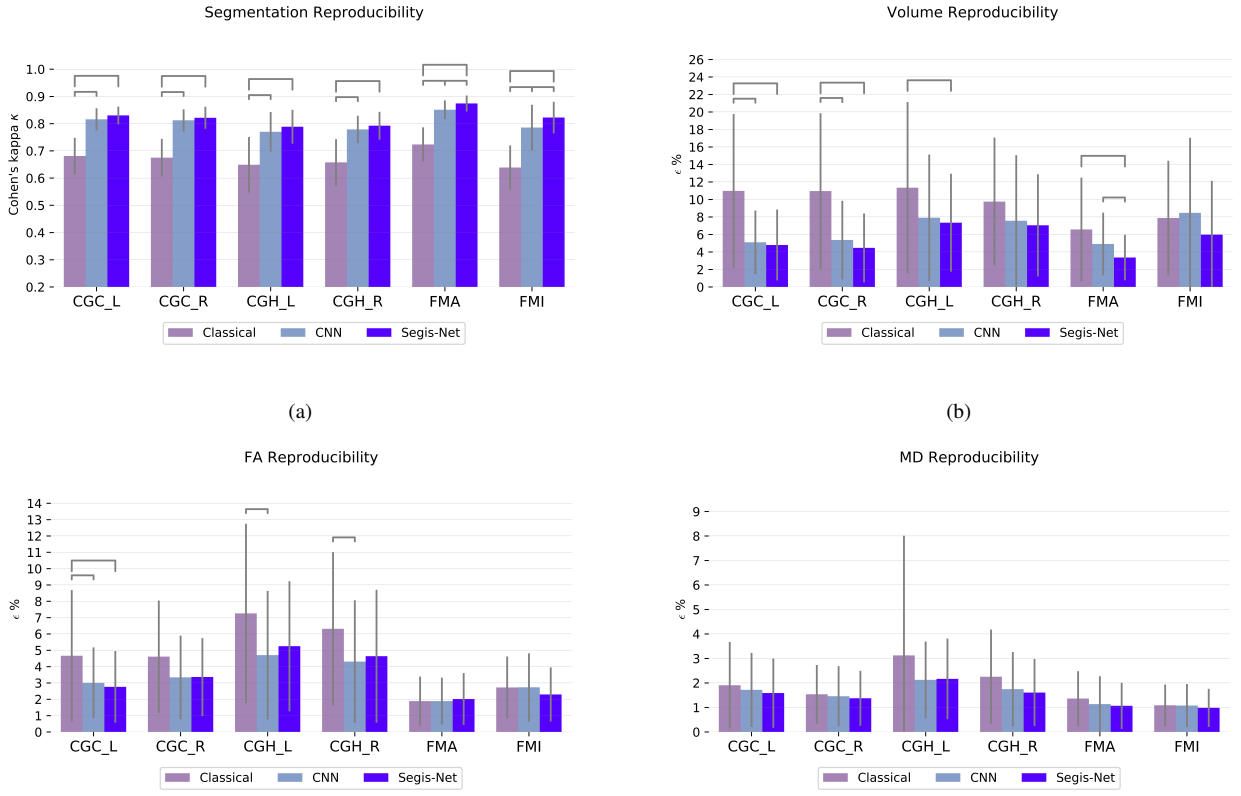
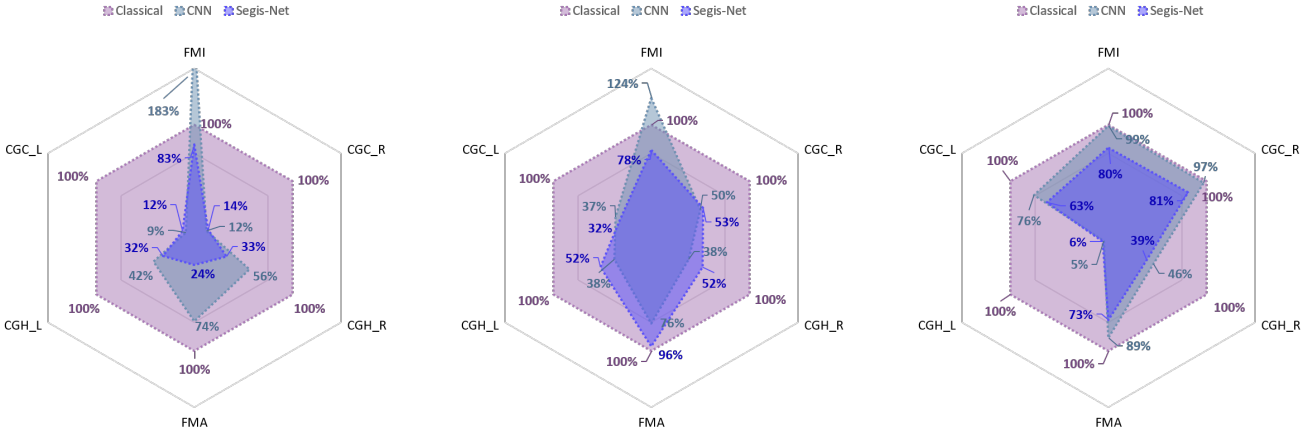


Figure 6: Reproducibility of tract-specific measures with the *Classical*, *CNN*, and Segis-Net pipeline. Error bars indicate standard deviations. The bracket hat indicates a significant difference between two methods (t-test, $p < 2.8{\times}10^{-3}$). In figure (a), a higher Cohen's kappa coefficient (κ) indicates a better reproducibility. In figure (b-d), a lower error ($\epsilon\%$) indicates a better reproducibility. Volume: tract-specific volume (ml), FA: fractional anisotropy, MD: mean diffusivity ($10^{-3}mm^2/s$).



(a) Volume Measures (b) FA Measures (c) MD Measures

Figure 7: The percentage of sample-size that would be required in tract measures of volume, FA, and MD with the *CNN* pipeline and Segis-Net. The sample size required for the *Classical* pipeline is used as the reference (100%).

VoxelMorph presents an approach that uses correspondence in existing segmentation labels to boost registration. To investigate the benefit of this approach in comparison with the proposed simultaneous optimization, we applied VoxelMorph to the registration of longitudinal FA images on our dataset. The implementation as detailed by Balakrishnan et al. (2019) was used directly.

4.6.1. Assessment of segmentation performance

We compared segmentation accuracy between U-ReSNet and the proposed Segis-Net on the test dataset, which was quantified by the Dice coefficient with respect to the reference segmentation (section 3.6).

For all six tracts, the segmentation accuracy of Segis-Net was higher than that of U-ReSNet by a margin of around 10% and with a smaller standard deviation (Table 1).

	U-ReSNet	Segis-Net
CGC_L	0.69 ± 0.06	**0.76 ± 0.06**
CGC_R	0.69 ± 0.07	**0.76 ± 0.06**
CGH_L	0.67 ± 0.08	**0.76 ± 0.07**
CGH_R	0.67 ± 0.09	**0.76 ± 0.09**
FMA	0.69 ± 0.06	**0.76 ± 0.05**
FMI	0.60 ± 0.08	**0.67 ± 0.09**

Table 1: Segmentation Dice coefficient of U-ReSNet and Segis-Net. The bold value indicates a better performance between methods.

4.6.2. Assessment of registration performance

We compared registration performance between U-ReSNet, VoxelMorph and the proposed Segis-Net on the test dataset by evaluating the spatial correlation (SC) similarity (Eq. 12) of registered density map of tracts, the Dice coefficient (DC) of registered reference segmentation of tracts, and the mean squared error (MSE) of registered FA maps.

Generally, U-ReSNet and Segis-Net yielded better SC similarity than that of VoxelMorph; Segis-Net achieved the best DC and MSE (Table 2); the standard deviation of Segis-Net was overall smallest for all three metrics, except that of DC in three tracts which were smallest for VoxelMorph. In the evaluation of SC similarity, U-ReSNet led the highest similarity in the left and right CGC, and the right of CGH tract; Segis-Net was the highest for FMA and FMI tract; for the left CGH tract, a similar high SC was observed for U-ReSNet and Segis-Net, although the variation was smaller in Segis-Net. For all six tracts, the DC of Segis-Net were higher than that of U-ReSNet, followed by VoxelMorph.

		U-ReSNet	VoxelMorph	Segis-Net
SC	CGC_L	**0.77 ± 0.11**	0.72 ± 0.09	0.73 ± 0.08
	CGC_R	**0.77 ± 0.11**	0.71 ± 0.09	0.73 ± 0.07
	CGH_L	**0.77 ± 0.11**	0.75 ± 0.10	**0.77 ± 0.10**
	CGH_R	**0.77 ± 0.12**	0.75 ± 0.11	0.76 ± 0.10
	FMA	0.73 ± 0.11	0.73 ± 0.10	**0.74 ± 0.09**
	FMI	0.75 ± 0.09	0.74 ± 0.08	**0.76 ± 0.08**
DC	CGC_L	0.69 ± 0.07	0.65 ± 0.06	**0.74 ± 0.06**
	CGC_R	0.70 ± 0.07	0.65 ± 0.06	**0.74 ± 0.05**
	CGH_L	0.67 ± 0.08	0.64 ± 0.07	**0.71 ± 0.08**
	CGH_R	0.67 ± 0.10	0.64 ± 0.08	**0.71 ± 0.09**
	FMA	0.70 ± 0.06	0.68 ± 0.06	**0.72 ± 0.06**
	FMI	0.57 ± 0.10	0.56 ± 0.07	**0.60 ± 0.09**
MSE ($\times 10^{-2}$)		0.47 ± 0.38	0.19 ± 0.88	**0.13 ± 0.10**

Table 2: Registration performance of U-ReSNet, VoxelMorph and Segis-Net, as quantified by the spatial correlation (SC) similarity, the Dice coefficient (DC), and the mean squared error (MSE). The bold value indicates a better performance between methods.

5. Discussion

We developed a single-step deep learning framework, coined Segis-Net, for simultaneous optimization of segmentation and registration. The method was applied to analyze changes in WM tracts from a large set of longitudinal diffusion MRI images. To evaluate the performance of the method, we compared it with two multistage pipelines consisting of independent segmentation and registration components, i.e., the *Classical* and *CNN* pipeline. Segis-Net showed improved performances in registration accuracy, spatio-temporal consistency of segmentation, and reproducibility of segmentation and tract-specific measures. We evaluated the practical value of the improved performance in terms of sample-size reduction that could be achieved when employing the method. The tract-specific mesures analyzed with Segis-Net would only require 33.0% – 60.5% sample-size of the data for achieving the same effect size as the *Classical* pipeline.

To date most developments in longitudinal analysis frameworks have focused on unbiased ways of registering image time series (Metz et al., 2011; Keihaninejad et al., 2013), in which a multistage approach combing independent segmentation and registration components is often used (De Groot et al., 2013a; Yendiki et al., 2016). In this paper, we aimed to investigate a different way to improve the performance of the longitudinal framework by using a single-step CNN that optimizes both tasks simultaneously. The sole value of simultaneous optimization was demonstrated by the comparison with the *CNN* pipeline. There was no benefit observed for segmentation alone, but for registration, spatio-temporal consistency of segmentation, and reproducibility, simultaneous optimization led to significantly improved performance.

In the segmentation evaluation, similar accuracies for the *CNN* and Segis-Net framework was observed for the six tracts (Figure 3). Relative segmentation accuracy between individual tracts were in line with those reported in literature (Wasserthal et al., 2018; Li et al., 2020a). For instance, a small and thin object like the FMI tract tended to have a lower Dice coefficient than the larger cingulum and FMA tracts. In the task of registration alone, Segis-Net overall yielded the best accuracy among the three pipelines, significantly outperforming the *Classical* pipeline for three tracts and the *CNN* pipeline for five tracts (Figure 4). This is an important observation as, first, it showed that simultaneous optimization was beneficial to one of the individual tasks, and second, it is non-trivial to improve registration accuracy over a classical algorithm, in which the transformation is pair-wise optimized on the test images.

In all six tracts, we observed substantially higher spatio-temporal consistency of segmentation and reproducibility of segmentation with Segis-Net than with the two multistage pipelines (Figure 5, 6). The spatio-temporal consistency of segmentation as quantified by the Dice coefficient ranged 0.81 – 0.87 for Segis-Net, significantly outperforming the *Classical* pipeline for all the six tracts (range: 0.57 – 0.69) and the *CNN* pipeline for five tracts (range: 0.77 – 0.84). The segmentation reproducibility as quantified by Cohen's kappa ranged 0.79 – 0.87 for Segis-Net, significantly higher than the *Classi-*

cal pipeline for all the six tracts (range: 0.64 – 0.72) and the *CNN* pipeline for two tracts (range: 0.77 – 0.85). These results indicate that Segis-Net can serve as a reliable alternative to the *Classical* pipeline in spatially capturing macro-structural brain changes over time.

In addition, more significant improvements were observed for the reproducibility of tract-specific volume assessment, but not for the FA and MD measures. For volume reproducibility, Segis-Net yielded the least error in the measurements of scan and re-scan, followed by the *CNN* pipeline (Figure 6). For the FA and MD measures, we observed relatively similar reproducibility for the three methods, in which significant difference was only observed in FA reproducibility of CGH and left CGC tract. This suggests that diffusion measures are quite robust to variations in the geometry of the segmented tract.

These improved performances have practical values in a power analysis, where both the *CNN* pipeline and Segis-Net showed to be able to reduce the required sample-size to achieve the same statistical power as the *Classical* pipeline. The data processed with Segis-Net would require on average 33.0% of the sample-size for volume measures, 60.5% for FA, and 57.0% for MD measures, requiring consistently a decreased sample-size for all the settings. The averaged percentages for the *CNN* pipeline were 62.7%, 60.5% and 68.7%. For FMI tract, it would, however, require 183% and 124% of the sample-size for the volume and FA measures. The observed dispersion of sample-size reduction with the *CNN* pipeline may suggest that simultaneous optimization was beneficial to the robustness of the method across the concurrently segmented tracts.

In the comparison with related methods, Segis-Net had a better segmentation performance than U-ReSNet, an existing simultaneous method (Table 1). We expect this improved performance of Segis-Net because of two reasons: 1) the method allows the use of diffusion tensor images for tract segmentation, as we use parallel network modules and only align the predicted segmentation; this circumvents the need to interpolate tensor images. In other words, task-specific inputs can be used; and 2) the sub-branches in the segmentation stream (Figure 8) are designed for the prediction of white matter tracts which can overlap with each other, unlike the exclusive tissue labels focused by other works.

During the comparison of registration performance, we observed two interesting results (Table 2). First, VoxelMorph was the only method that directly optimized on the DC metric, but it led to a least DC score. This can be due to the fact that the segmentation labels used in diffusion imaging studies are often independently obtained for each image, which is much less correlated to the registration performance than is the case for atlas-based segmentation (Balakrishnan et al., 2019). As a result, the alignment of "imperfect" segmentation labels can be an obstructive loss term instead. Second, although the MSE of U-ReSNet was almost four times that of the Segis-Net, it achieved a good SC similarity, especially in small structures like CGC and CGH. This can be attributed to the formulation of their registration loss as the sum of local-SC and MSE.

Whereas the method is generic, we specifically implemented and optimized it for longitudinal study in diffusion MRI data. In diffusion MRI application, we adopted the commonly used scalar-value FA map as the input for registration. Deformable registration of diffusion tensor images is known to be challenging due to the directional components contained in voxels. Despite developments in classical methods for tensor reorientation during the optimization (Cao et al., 2006; Zhang et al., 2007), for learning-based registration it still largely remains unexplored. With the promising results of diffusion tensor interpolation as shown by Grigorescu et al. (2020), Segis-Net based on solely tensor images would be an interesting direction to explore.

The Segis-Net framework presented in the current study is limited to two time-points. This is because learning-based registration algorithms currently only support pairwise transformations (Balakrishnan et al., 2019). One limitation of our method is therefore that it does not allow for analysis of arbitrary number of time-points. In the present study, we grouped the available triple time-points from the same participant into orderless image-pairs for bidirectional analysis. A future possible improvement of the method could be extending the registration component of Segis-Net to enable learning-based group-wise analysis of a set of time-points (Li et al., 2020b).

Beyond the current application, we expect that this work could be extended to other imaging sequences and for example for segmentation of lesion images. For future work, we plan to adapt the proposed method to analyze brain diseases with large and progressive changes. For instance, registration of brains with lesions due to cortical infarct may benefit from a simultaneous segmentation of infarct regions.

6. Conclusion

We proposed a single-step deep learning framework for longitudinal diffusion MRI analysis, in which segmentation and deformable registration were integrated for simultaneous optimization. The comparison with two multistage approaches showed that the proposed Segis-Net can be applied as a reliable tool to support spatio-temporal analysis of WM tracts from longitudinal diffusion MRI imaging. Besides the improved performances, a two-in-one framework for concurrent segmentation and registration also enables a light-weight way of fast quantification of brain changes overtime. This may lead to a more prominent role for tract-specific biomarkers in applications where tract segmentation and registration are subject to time constraints. With the increasing availability of longitudinal diffusion data, we expect future studies investigating progressive neurodegeneration can greatly benefit from the improved reliability and efficiency of Segis-Net.

Acknowledgments

This work was sponsored through grants of the Medical Delta Diagnostics 3.0: Dementia and Stroke, the EU Horizon 2020 project EuroPOND (666992), the Netherlands CardioVascular Research Initiative (Heart-Brain Connection: CVON2012-06, CVON2018-28), and the Dutch Heart Foundation (PPP Allowance, 2018B011).

Data availability

The datasets analyzed during the current study are not publicly available. Due to the sensitive nature of the data used in this study, participants were assured raw data would remain confidential and would not be shared.

References

Ashburner, J., Friston, K.J., 2005. Unified segmentation. Neuroimage 26, 839–851.

Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2019. VoxelMorph: a learning framework for deformable medical image registration. IEEE Trans. Med. Imag. .

Barrick, T.R., Charlton, R.A., Clark, C.A., Markus, H.S., 2010. White matter structural decline in normal ageing: a prospective longitudinal study using tract-based spatial statistics. Neuroimage 51, 565–577.

Bastiaansen, W.A., Rousian, M., Steegers-Theunissen, R.P., Niessen, W.J., Koning, A., Klein, S., 2020. Towards segmentation and spatial alignment of the human embryonic brain using deep learning for atlas-based registration, in: International Workshop on Biomedical Image Registration, Springer. pp. 34–43.

Behrens, T.E., Berg, H.J., Jbabdi, S., Rushworth, M.F., Woolrich, M.W., 2007. Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? Neuroimage 34, 144–155.

Bishop, C.M., 2006. Pattern recognition and machine learning. springer.

Bortsova, G., Dubost, F., Hogeweg, L., Katramados, I., de Bruijne, M., 2019. Semi-supervised medical image segmentation via learning consistency under transformations, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 810–818.

Cao, Y., Miller, M.I., Mori, S., Winslow, R.L., Younes, L., 2006. Diffeomorphic matching of diffusion tensor images, in: 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06), IEEE. pp. 67–67.

Cheng, J., Tsai, Y.H., Wang, S., Yang, M.H., 2017. Segflow: Joint learning for video object segmentation and optical flow, in: Proceedings of the IEEE international conference on computer vision, pp. 686–695.

Conturo, T.E., Lori, N.F., Cull, T.S., Akbudak, E., Snyder, A.Z., Shimony, J.S., McKinstry, R.C., Burton, H., Raichle, M.E., 1999. Tracking neuronal fiber pathways in the living human brain. Proceedings of the National Academy of Sciences 96, 10422–10427.

Crick, F., Jones, E., 1993. Backwardness of human neuroanatomy. Nature 361, 109–110.

Dai, X., Khorram, S., 1999. A feature-based image registration algorithm using improved chain-code representation combined with invariant moments. IEEE Transactions on Geoscience and Remote Sensing 37, 2351–2362.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26, 297–302.

Diggle, P., Diggle, P.J., Heagerty, P., Liang, K.Y., Heagerty, P.J., Zeger, S., et al., 2002. Analysis of longitudinal data. Oxford University Press.

Dimond, D., Rohr, C.S., Smith, R.E., Dhollander, T., Cho, I., Lebel, C., Dewey, D., Connelly, A., Bray, S., 2020. Early childhood development of white matter fiber density and morphology. NeuroImage 210, 116552.

Estienne, T., Lerousseau, M., Vakalopoulou, M., Alvarez Andres, E., Battistella, E., Carré, A., Chandra, S., Christodoulidis, S., Sahasrabudhe, M., Sun, R., et al., 2020. Deep learning-based concurrent brain registration and tumor segmentation. Frontiers in Computational Neuroscience 14, 17.

Estienne, T., Vakalopoulou, M., Christodoulidis, S., Battistela, E., Lerousseau, M., Carre, A., Klausner, G., Sun, R., Robert, C., Mougiakakou, S., et al., 2019. U-ReSNet: Ultimate coupling of registration and segmentation with deep nets, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 310–319.

Farbota, K.D., Bendlin, B.B., Alexander, A.L., Rowley, H.A., Dempsey, R.J., Johnson, S.C., 2012. Longitudinal diffusion tensor imaging and neuropsychological correlates in traumatic brain injury patients. Frontiers in human neuroscience 6, 160.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., et al., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341–355.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp. 249–256.

Gooya, A., Pohl, K.M., Bilello, M., Biros, G., Davatzikos, C., 2011. Joint segmentation and deformable registration of brain scans guided by a tumor growth model, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 532–540.

Grigorescu, I., Uus, A., Christiaens, D., Cordero-Grande, L., Hutter, J., Edwards, A.D., Hajnal, J.V., Modat, M., Deprez, M., 2020. Diffusion tensor driven image registration: a deep learning approach, in: International Workshop on Biomedical Image Registration, Springer. pp. 131–140.

De Groot, M., Cremers, L.G., Ikram, M.A., Hofman, A., Krestin, G.P., van der Lugt, A., Niessen, W.J., Vernooij, M.W., 2016. White matter degeneration with aging: longitudinal diffusion MR imaging analysis. Radiology 279, 532–541.

De Groot, M., Ikram, M.A., Akoudad, S., Krestin, G.P., Hofman, A., van der Lugt, A., Niessen, W.J., Vernooij, M.W., 2015. Tract-specific white matter degeneration in aging: the Rotterdam Study. Alzheimer's & Dementia 11, 321–330.

De Groot, M., Verhaaren, B.F., De Boer, R., Klein, S., Hofman, A., van der Lugt, A., Ikram, M.A., Niessen, W.J., Vernooij, M.W., 2013a. Changes in normal-appearing white matter precede development of white matter lesions. Stroke 44, 1037–1042.

De Groot, M., Vernooij, M.W., Klein, S., Ikram, M.A., Vos, F.M., Smith, S.M., Niessen, W.J., Andersson, J.L., 2013b. Improving alignment in tract-based spatial statistics: evaluation and optimization of image registration. Neuroimage 76, 400–411.

Hu, Y., Modat, M., Gibson, E., Ghavami, N., Bonmati, E., Moore, C.M., Emberton, M., Noble, J.A., Barratt, D.C., Vercauteren, T., 2018. Label-driven weakly-supervised learning for multimodal deformable image registration, in: 15th ISBI, IEEE. pp. 1070–1074.

Ikram, M.A., Brusselle, G., Ghanbari, M., Goedegebure, A., Ikram, M.K., Kavousi, M., Kieboom, B.C., Klaver, C.C., de Knegt, R.J., Luik, A.I., et al., 2020. Objectives, design and main findings until 2020 from the rotterdam study. European Journal of Epidemiology , 1–35.

Ikram, M.A., van der Lugt, A., Niessen, W.J., Krestin, G.P., Koudstaal, P.J., Hofman, A., Breteler, M.M., Vernooij, M.W., 2011. The rotterdam scan study: design and update up to 2012. European journal of epidemiology 26, 811–824.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 .

Jaderberg, M., Simonyan, K., Zisserman, A., et al., 2015. Spatial transformer networks, in: Advances in neural information processing systems, pp. 2017–2025.

Keihaninejad, S., Zhang, H., Ryan, N.S., Malone, I.B., Modat, M., Cardoso, M.J., Cash, D.M., Fox, N.C., Ourselin, S., 2013. An unbiased longitudinal analysis framework for tracking white matter changes using diffusion tensor imaging with application to Alzheimer's disease. Neuroimage 72, 153–163.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Klein, S., Pluim, J.P., Staring, M., Viergever, M.A., 2009. Adaptive stochastic gradient descent optimisation for image registration. Int J Comput Vis. 81, 227.

Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P., 2010. Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imag. 29, 196–205.

Koppelmans, V., de Groot, M., de Ruiter, M.B., Boogerd, W., Seynaeve, C., Vernooij, M.W., Niessen, W.J., Schagen, S.B., Breteler, M.M., 2014. Global and focal white matter integrity in breast cancer survivors 20 years after adjuvant chemotherapy. Hum Brain Mapp. 35, 889–899.

Van der Krieke, L., Blaauw, F.J., Emerencia, A.C., Schenk, H.M., Slaets, J.P., Bos, E.H., de Jonge, P., Jeronimus, B.F., 2017. Temporal dynamics of health and well-being: A crowdsourcing approach to momentary assessments and automated generation of personalized feedback. Psychosomatic medicine 79, 213–223.

Landis, J.R., Koch, G.G., 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics , 363–374.

Le Bihan, D., Mangin, J.F., Poupon, C., Clark, C.A., Pappata, S., Molko, N., Chabriat, H., 2001. Diffusion tensor imaging: concepts and applications.

Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine 13, 534–546.
Lebel, C., Beaulieu, C., 2011. Longitudinal development of human brain wiring continues from childhood into adulthood. Journal of Neuroscience 31, 10937–10947.
Leemans, A., Jeurissen, B., Sijbers, J., Jones, D., 2009. ExploreDTI: a graphical toolbox for processing, analyzing, and visualizing diffusion MR data, in: 17th annual meeting of intl soc mag reson med, p. 3537.
Li, B., de Groot, M., Steketee, R.M., Meijboom, R., Smits, M., Vernooij, M.W., Ikram, M.A., Liu, J., Niessen, W.J., Bron, E.E., 2020a. Neuro4Neuro: A neural network approach for neural tract segmentation using large-scale population-based diffusion imaging. NeuroImage , 116993.
Li, B., de Groot, M., Vernooij, M.W., Ikram, M.A., Niessen, W.J., Bron, E.E., 2018. Reproducible white matter tract segmentation using 3D U-Net on a large-scale DTI dataset, in: International Workshop on MLMI, Springer. pp. 205–213.
Li, B., Niessen, W.J., Klein, S., de Groot, M., Ikram, M.A., Vernooij, M.W., Bron, E.E., 2019. A hybrid deep learning framework for integrated segmentation and registration: evaluation on longitudinal white matter tract changes, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 645–653.
Li, B., Niessen, W.J., Klein, S., Ikram, M.A., Vernooij, M.W., Bron, E.E., 2020b. Learning unbiased registration and joint segmentation: evaluation on longitudinal diffusion MRI. arXiv preprint arXiv:2011.01869 .
Maas, A.L., Hannun, A.Y., Ng, A.Y., 2013. Rectifier nonlinearities improve neural network acoustic models, in: Proc. icml, p. 3.
Meijboom, R., Steketee, R., Ham, L., Mantini, D., Bron, E., van der Lugt, A., van Swieten, J., Smits, M., 2019. Exploring quantitative group-wise differentiation of Alzheimer's disease and behavioural variant frontotemporal dementia using tract-specific microstructural white matter and functional connectivity measures at multiple time points. European radiology 29, 5148–5159.
Metz, C., Klein, S., Schaap, M., van Walsum, T., Niessen, W.J., 2011. Nonrigid registration of dynamic medical imaging data using nD+ t B-splines and a groupwise optimization approach. Med Image Anal. 15, 238–249.
Niessen, W.J., 2016. MR brain image analysis in dementia: From quantitative imaging biomarkers to ageing brain models and imaging genetics.
Parisot, S., Wells III, W., Chemouny, S., Duffau, H., Paragios, N., 2014. Concurrent tumor segmentation and registration with uncertainty-based sparse non-uniform graphs. Med Image Anal. 18.
Pathak, D., Girshick, R., Dollár, P., Darrell, T., Hariharan, B., 2017. Learning features by watching objects move, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2701–2710.
Pohl, K.M., Fisher, J., Grimson, W.E.L., Kikinis, R., Wells, W.M., 2006. A Bayesian model for joint segmentation and registration. NeuroImage 31, 228–239.
Postelnicu, G., Zollei, L., Fischl, B., 2008. Combined volumetric and surface registration. IEEE transactions on medical imaging 28, 508–522.
Qin, C., Bai, W., Schlemper, J., Petersen, S.E., Piechnik, S.K., Neubauer, S., Rueckert, D., 2018. Joint learning of motion estimation and segmentation for cardiac mr image sequences, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 472–480.
Reuter, M., Schmansky, N.J., Rosas, H.D., Fischl, B., 2012. Within-subject template estimation for unbiased longitudinal image analysis. Neuroimage 61, 1402–1418.
Rohé, M.M., Datar, M., Heimann, T., Sermesant, M., Pennec, X., 2017. SVF-Net: Learning deformable image registration using shape matching, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 266–274.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on MICCAI, Springer. pp. 234–241.
Sullivan, E.V., Rohlfing, T., Pfefferbaum, A., 2010. Longitudinal study of callosal microstructure in the normal adult aging brain using quantitative DTI fiber tracking. Developmental neuropsychology 35, 233–256.
Vakalopoulou, M., Chassagnon, G., Bus, N., Marini, R., Zacharaki, E.I., Revel, M.P., Paragios, N., 2018. Atlasnet: Multi-atlas non-linear deep networks for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 658–666.
Vlontzos, A., Mikolajczyk, K., 2018. Deep segmentation and registration in X-Ray angiography video. arXiv preprint arXiv:1805.06406 .
Vrooman, H.A., Cocosco, C.A., van der Lijn, F., Stokking, R., Ikram, M.A., Vernooij, M.W., Breteler, M.M., Niessen, W.J., 2007. Multi-spectral brain tissue segmentation using automatically trained k-nearest-neighbor classification. Neuroimage 37, 71–81.
Wasserthal, J., Neher, P., Maier-Hein, K.H., 2018. TractSeg: Fast and accurate white matter tract segmentation. NeuroImage 183, 239–253.
Wyatt, P.P., Noble, J.A., 2003. MAP MRF joint segmentation and registration of medical images. Medical Image Analysis 7, 539–552.
Xu, Z., Niethammer, M., 2019. DeepAtlas: Joint semi-supervised learning of image registration and segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 420–429.
Yendiki, A., Reuter, M., Wilkens, P., Rosas, H.D., Fischl, B., 2016. Joint reconstruction of white-matter pathways from longitudinal diffusion MRI data with anatomical priors. Neuroimage 127, 277–286.
Yezzi, A., Zöllei, L., Kapur, T., 2003. A variational framework for integrating segmentation and registration through active contours. Medical image analysis 7, 171–185.
Zhang, H., Avants, B.B., Yushkevich, P.A., Woo, J.H., Wang, S., McCluskey, L.F., Elman, L.B., Melhem, E.R., Gee, J.C., 2007. High-dimensional spatial normalization of diffusion tensor images improves the detection of white matter differences: an example study using amyotrophic lateral sclerosis. IEEE transactions on medical imaging 26, 1585–1597.
Zhu, W., Myronenko, A., Xu, Z., Li, W., Roth, H., Huang, Y., Milletari, F., Xu, D., 2020. Neurreg: Neural registration and its application to image segmentation, in: The IEEE Winter Conference on Applications of Computer Vision, pp. 3617–3626.

Supplementary Material

The network architectures of $\mathcal{F}_\Theta$ and $\mathcal{G}_\Psi$ are illustrated in Figure 8 and Figure 9, respectively. The image size of the input and output for $\mathcal{F}_\Theta$ used in the present study were $(112 \times 208 \times 112 \times 6)$ and $(112 \times 208 \times 112 \times 3)$ voxels. That of $\mathcal{G}_\Psi$ were $(112 \times 208 \times 112 \times 2)$ and $(112 \times 208 \times 112 \times 3)$ voxels.

The procedure of simultaneous optimization is summarized with the pseudo code in Algorithm 1.

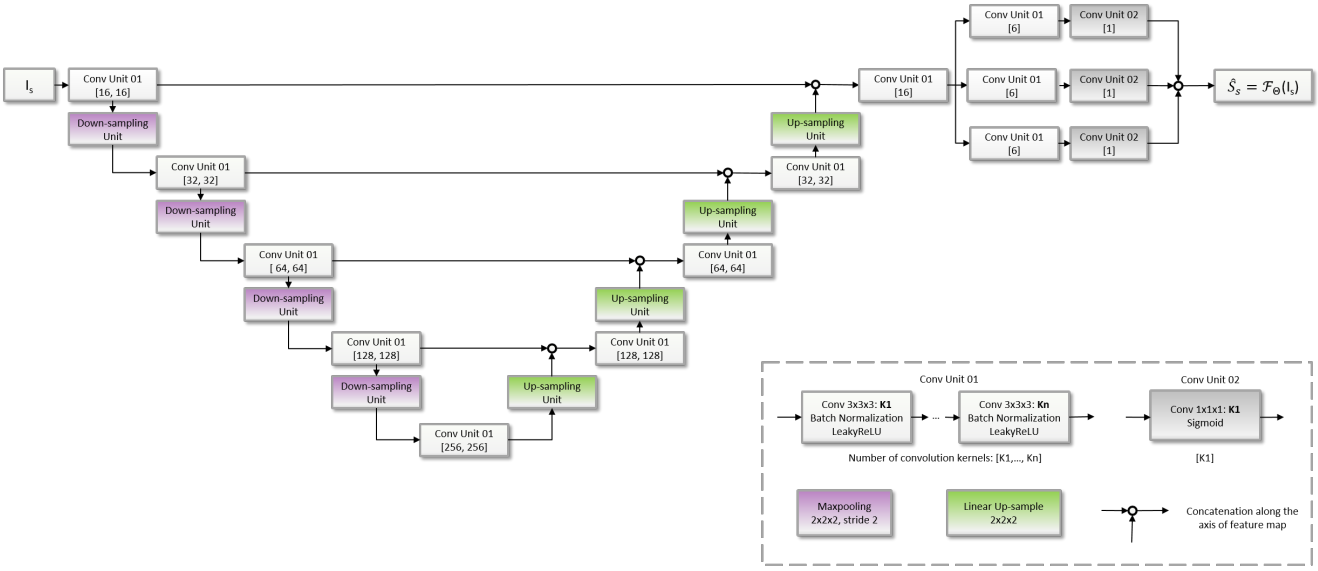


Figure 8: The network architectures of $\mathcal{F}_\Theta$ module used in the present study. The number of convolution kernels used in Conv Unit 01 is denoted as [K1, ..., Kn], and that for Conv Unit 02 as [K1].

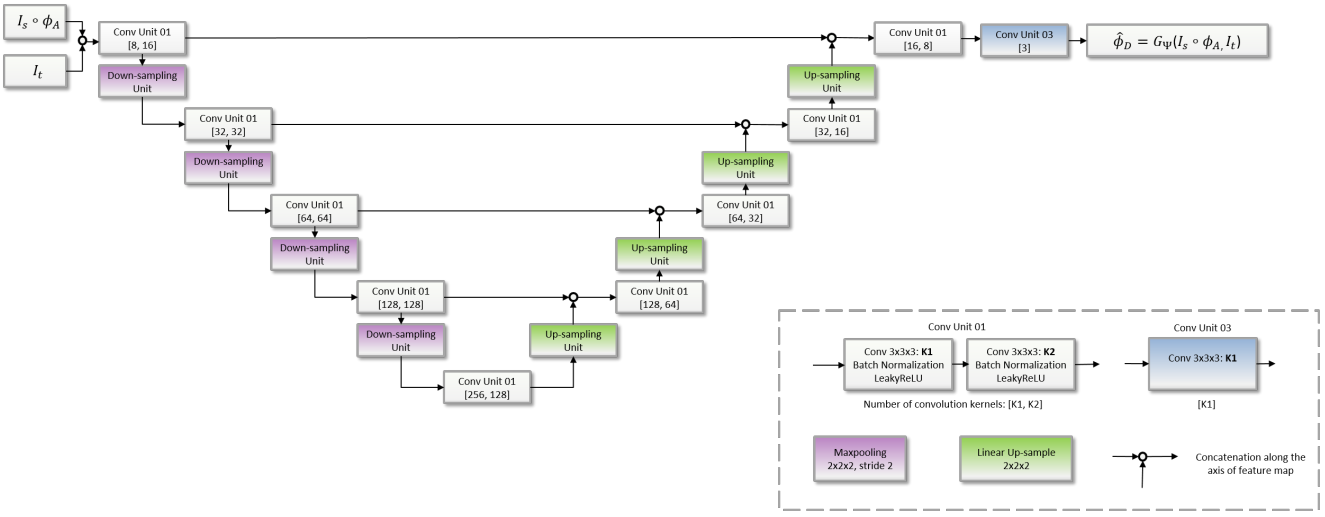


Figure 9: The network architectures of $\mathcal{G}_\Psi$ module used in the present study. The target image (I_t) and affine-aligned source image ($I_s \circ \boldsymbol{\phi}_A$) are used as the input to predict non-rigid deformation ($\boldsymbol{\phi}_D$), which can subsequently lead to a composite displacement field ($\boldsymbol{\phi}$) as shown in Figure 1. The number of convolution kernels used in Conv Unit 01 is denoted as [K1, K2], and that for Conv Unit 03 as [K1].

Algorithm 1: Simultaneous optimization

Input: $\{\mathcal{S}_s^i,\ \mathcal{S}_t^i,\ I_s^i, I_t^i, \boldsymbol{\phi}_A^i\}_{i=1}^N$
Parameters: $\boldsymbol{\Theta}, \boldsymbol{\Psi}$
Output: $\{\hat{\mathcal{S}}_s^i,\ \hat{\mathcal{S}}_s^i \circ \hat{\boldsymbol{\phi}}^i,\ I_s^i \circ \hat{\boldsymbol{\phi}}^i,\ \hat{\boldsymbol{\phi}}^i,\ \hat{\boldsymbol{\phi}}^i{}_D\}_{i=1}^N$
Initialization:
lr, decay_factor `// Initial and decay ratio of learning rate`
$\boldsymbol{\Theta}, \boldsymbol{\Psi} \sim$ GlorotUniform `// Kernel initialization`
for number of training iterations **do**
  shuffle(**Input**)
  for $i = 0$ **to** N **do**
    $\hat{\mathcal{S}}_s^i = \mathcal{F}_\Theta(I_s^i)$ `// Segmentation, Eq.1`
    $\hat{\boldsymbol{\phi}}_D^i = \mathcal{G}_\Psi(I_t^i,\ I_s^i \circ \boldsymbol{\phi}_A^i)$ `// Registration, Eq.4`
    $\hat{\boldsymbol{\phi}}^i = \boldsymbol{\phi}_A^i \circ \hat{\boldsymbol{\phi}}_D^i$ `// Transform composition`
    $\hat{I}_t^i = I_s^i \circ \hat{\boldsymbol{\phi}}^i$ `// Image warp`
    `/* Dependency on both tasks */`
    $\hat{\mathcal{S}}_t^i = \hat{\mathcal{S}}_s^i \circ \hat{\boldsymbol{\phi}}^i$ `// Segmentation warp`
    $\mathcal{L}^i = \mathcal{L}_{seg}(\mathcal{S}_s^i,\ \hat{\mathcal{S}}_s^i) + \alpha\mathcal{L}_{reg}(I_t^i,\ \hat{I}_t^i) + \beta\mathcal{L}_{def}(\hat{\boldsymbol{\phi}}_D^i)$
     $+\gamma\mathcal{L}_{com}(\mathcal{S}_t^i,\ \hat{\mathcal{S}}_t^i)$ `// Segis-Net objective, Eq.9`
    `/* Simultaneous optimization of parameters */`
    $\boldsymbol{\Theta}, \boldsymbol{\Psi} \leftarrow$ **Adam** $(\mathcal{L}^i, \text{lr}, \boldsymbol{\Theta}, \boldsymbol{\Psi})$
  end
  `/* Custom condition of learning rate decay */`
  if learning rate decay **then**
    $lr \leftarrow lr \times$ decay_factor
  end
  return $\boldsymbol{\Theta}, \boldsymbol{\Psi}$ `// Return parameters per epoch`
end

		Classical	CNN	Segis-Net
Figure 3	CGC_L	-	0.76 ± 0.06	0.76 ± 0.06
	CGC_R	-	0.76 ± 0.06	0.76 ± 0.06
	CGH_L	-	0.76 ± 0.07	0.76 ± 0.07
	CGH_R	-	0.77 ± 0.09	0.76 ± 0.09
	FMA	-	0.76 ± 0.05	0.76 ± 0.05
	FMI	-	0.68 ± 0.09	0.67 ± 0.09
Figure 4	CGC_L	0.73 ± 0.10	0.71 ± 0.07	0.73 ± 0.08
	CGC_R	0.73 ± 0.10	0.69 ± 0.06	0.73 ± 0.07
	CGH_L	0.75 ± 0.11	0.75 ± 0.10	0.77 ± 0.09
	CGH_R	0.75 ± 0.11	0.75 ± 0.10	0.76 ± 0.10
	FMA	0.72 ± 0.10	0.71 ± 0.08	0.74 ± 0.09
	FMI	0.74 ± 0.08	0.75 ± 0.08	0.76 ± 0.08
Figure 5	CGC_L	0.68 ± 0.06	0.82 ± 0.04	0.83 ± 0.04
	CGC_R	0.68 ± 0.06	0.82 ± 0.06	0.83 ± 0.04
	CGH_L	0.66 ± 0.08	0.81 ± 0.05	0.82 ± 0.05
	CGH_R	0.66 ± 0.09	0.81 ± 0.05	0.81 ± 0.05
	FMA	0.69 ± 0.06	0.84 ± 0.03	0.87 ± 0.02
	FMI	0.57 ± 0.09	0.77 ± 0.07	0.81 ± 0.05
Figure 6(a)	CGC_L	0.68 ± 0.07	0.82 ± 0.04	0.83 ± 0.03
	CGC_R	0.68 ± 0.07	0.81 ± 0.04	0.82 ± 0.04
	CGH_L	0.65 ± 0.10	0.77 ± 0.07	0.79 ± 0.06
	CGH_R	0.66 ± 0.09	0.78 ± 0.05	0.79 ± 0.05
	FMA	0.72 ± 0.06	0.85 ± 0.03	0.87 ± 0.03
	FMI	0.64 ± 0.08	0.79 ± 0.08	0.82 ± 0.06
Figure 6(b)	CGC_L	11 ± 8.8%	5.1 ± 3.6%	4.8 ± 4.1%
	CGC_R	11 ± 8.9%	5.4 ± 4.5%	4.5 ± 3.9%
	CGH_L	11 ± 9.8%	7.9 ± 7.2%	7.3 ± 5.6%
	CGH_R	9.8 ± 7.3%	7.6 ± 7.5%	7.0 ± 5.8%
	FMA	6.6 ± 5.9%	4.9 ± 3.6%	3.4 ± 2.6%
	FMI	7.9 ± 6.5%	8.5 ± 8.6%	6.0 ± 6.2%
Figure 6(c)	CGC_L	4.7 ± 4.0%	3.0 ± 2.2%	2.8 ± 2.2%
	CGC_R	4.6 ± 3.4%	3.3 ± 2.6%	3.4 ± 2.4%
	CGH_L	7.3 ± 5.5%	4.7 ± 3.9%	5.2 ± 4.0%
	CGH_R	6.3 ± 4.7%	4.3 ± 3.8%	4.6 ± 4.1%
	FMA	1.9 ± 1.5%	1.9 ± 1.4%	2.0 ± 1.6%
	FMI	2.7 ± 1.9%	2.7 ± 2.1%	2.3 ± 1.7%
Figure 6(d)	CGC_L	1.9 ± 1.8%	1.7 ± 1.5%	1.6 ± 1.4%
	CGC_R	1.5 ± 1.2%	1.5 ± 1.2%	1.4 ± 1.1%
	CGH_L	3.1 ± 4.9%	2.1 ± 1.6%	2.2 ± 1.6%
	CGH_R	2.3 ± 1.9%	1.7 ± 1.5%	1.6 ± 1.4%
	FMA	1.4 ± 1.1%	1.1 ± 1.1%	1.1 ± 0.9%
	FMI	1.1 ± 0.8%	1.1 ± 0.9%	1.0 ± 0.8%

Table 3: Results overview for Figure 3-5. Figure 3, Segmentation accuracy; Figure 4, Spatial correlation (SC) similarity; Figure 5, Spatio-temporal consistency of segmentation (STCS); Figure 6 (a), Segmentation reproducibility; Figure 6 (b), Volume Reproducibility; Figure 6 (c), FA reproducibility; Figure 6 (d), MD reproducibility.

Longitudinal diffusion MRI analysis using Segis-Net: a single-step deep-learning framework for simultaneous segmentation and registration

Bo Li[a], Wiro J. Niessen[a,b], Stefan Klein[a], Marius de Groot[a,c], M. Arfan Ikram[a,c,d], Meike W. Vernooij[a,c], Esther E. Bron[a]

[a]*Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, the Netherlands*
[b]*Imaging Physics, Applied Sciences, Delft University of Technology, the Netherlands*
[c]*Department of Epidemiology, Erasmus MC, Rotterdam, the Netherlands*
[d]*Department of Neurology, Erasmus MC, Rotterdam, the Netherlands*

Abstract

This work presents a single-step deep-learning framework for longitudinal image analysis, coined Segis-Net. To optimally exploit information available in longitudinal data, this method concurrently learns a multi-class segmentation and nonlinear registration. Segmentation and registration are modeled using a convolutional neural network and optimized simultaneously for their mutual benefit. An objective function that optimizes spatial correspondence for the segmented structures across time-points is proposed. We applied Segis-Net to the analysis of white matter tracts from N=8045 longitudinal brain MRI datasets of 3249 elderly individuals. Segis-Net approach showed a significant increase in registration accuracy, spatio-temporal segmentation consistency, and reproducibility comparing with two multistage pipelines. This also led to a significant reduction in the sample-size that would be required to achieve the same statistical power in analyzing tract-specific measures. Thus, we expect that Segis-Net can serve as a new reliable tool to support longitudinal imaging studies to investigate macro- and microstructural brain changes over time.

Keywords: Segmentation, Registration, Diffusion MRI, Deep Learning, CNN, Longitudinal, White Matter Tract

1. Introduction

The increasing availability of longitudinal imaging data is expanding our ability to capture and characterize progressive anatomical changes, ranging from normal changes in the life span, to responses along disease trajectories or therapeutic actions. Compared to cross-sectional studies, longitudinal imaging studies have the advantage of allowing to trace the order of events at the individual level and to correct for the confounding effect of time-invariant individual differences (Van der Krieke et al., 2017). They are thus considered to be more accurate and sensitive in capturing subtle changes over time. To analyze spatio-temporal changes from longitudinal imaging data, a tailored framework that involves both segmentation and registration is required to segment the structures-of-interest and to register temporal frames. This can be achieved by directly combining two existing segmentation and registration tools, which are often designed for cross-sectional studies. However, the information offered in longitudinal data remains underutilized.

Various studies have shown that combining segmentation and registration at the stage of algorithm optimization can lead to improved performance. A popular combination strategy is to use the output of one task to optimize the other. Registration can be improved by using segmentation-level correspondences as input for deformation initialization (Dai and Khorram, 1999; Postelnicu et al., 2008) and optimization (De Groot et al., 2013b; Rohé et al., 2017; Hu et al., 2018; Balakrishnan et al., 2019; Bastiaansen et al., 2020; Zhu et al., 2020). Likewise, segmentation can benefit from registration by propagating anatomical information to subsequent frames, as has been shown in classical multi-atlas based segmentation methods (Fischl et al., 2002; Vakalopoulou et al., 2018) and in recent data-augmentation techniques which introduce labels to support unsupervised (Pathak et al., 2017) and weakly-supervised segmentation (Bortsova et al., 2019; Vlontzos and Mikolajczyk, 2018).

Other approaches combine the optimization of parameters from both tasks on a deeper level. Wyatt and Noble (2003) subgrouped these methods into two types according to the way in which they update their parameters: 1) "simultaneous estimation" that updates both the class labels and the transformations in a single-step optimization, and 2) "joint estimation" that alternately updates (separate) models in a multi-step optimization. Although the initialization and robustness of joint estimation can be influenced by the selection of the order to optimize and the criteria to switch tasks, this approach is preferred as it requires less computation power and allows to use task-specific training datasets (Yezzi et al., 2003; Wyatt and Noble, 2003; Ashburner and Friston, 2005; Pohl et al., 2006; Parisot et al., 2014; Gooya et al., 2011; Cheng et al., 2017; Xu and Niethammer, 2019). Simultaneous estimation is expected to be more ac-

Email address: b.li@erasmusmc.nl (Bo Li)

Abbreviations: MRI, Magnetic Resonance Imaging; DTI, Diffusion Tensor Imaging; FA, Fractional Anisotropy; MD, Mean Diffusivity; TE, Echo Time; TR, Repetition Time

curate, as it fully exploits the conditional correlations between two tasks that can be discounted in sequential processing (Ashburner and Friston, 2005). In addition, simultaneous estimation can explicitly optimize performances that rely on both tasks. We expect that this advantage has a large potential in improving the reliability of analysis of longitudinal imaging data, for instance by optimizing the spatio-temporal consistency of the segmentation. With the growing capability of modeling and computation by deep learning techniques, several simultaneous methods have been proposed and coupled segmentation with deformable registration in different ways, either for 2D (Qin et al., 2018) or 3D images (Li et al., 2019; Estienne et al., 2019, 2020).

Diffusion magnetic resonance imaging (MRI) is a non-invasive imaging technique that measures the diffusion of water in-vivo and can be used to quantitatively characterize white matter (WM) microstructure. The quantitative nature of diffusion MRI makes its derived measures, such as diffusion tensor imaging (DTI) metrics (Le Bihan et al., 2001), very suitable for longitudinal analysis. In addition, diffusion measures are likely to be more sensitive than structural measures in the early detection of changes in WM, and are therefore promising for the identification of subtle changes that relate to the early stages of the disease (Niessen, 2016), for instance in studying dementia subtypes (Meijboom et al., 2019). Longitudinal diffusion MRI has been widely studied at various levels, i.e., from regions-of-interest (Sullivan et al., 2010; Keihaninejad et al., 2013), to tract level (Lebel and Beaulieu, 2011; Yendiki et al., 2016; Meijboom et al., 2019; Dimond et al., 2020), and voxel level (Barrick et al., 2010; Farbota et al., 2012; De Groot et al., 2016). Since WM tracts are functionally grouped axonal fibers and thought to subserve particular brain functions, tract-specific investigation may highlight categorical differences in vulnerability to neurodegeneration and bridge the interpretation of imaging biomarkers with clinical symptoms.

Segmentation of WM tracts is however non-trivial because tracts cannot be identified directly from diffusion MRI, i.e., there is no in-vivo "gold standard" for tract (Crick and Jones, 1993), and because their anatomy can be complex. WM tracts are commonly segmented based on diffusion tractography by reconstruction of potential fiber pathways (Conturo et al., 1999). Recently, deep learning based methods, in particular using convolutional neural networks (CNN), have emerged and showed promising accuracy and efficiency in segmenting WM tracts (Li et al., 2018, 2020a; Wasserthal et al., 2018).

In the present work we focus on a CNN-based framework for longitudinal analysis of WM tracts, i.e., Segis-Net, and investigate the value of simultaneous optimization of segmentation and registration in this setting. In Li et al. (2019), we introduced a generic framework for simultaneous optimization, in which increased accuracies of both tasks were observed in a pilot analysis of a single tract (forceps minor; FMI). In this paper, we extend the tract-specific method by enabling concurrent segmentation of multiple tracts, which is a non-trivial task as a voxel can belong to multiple tracts. This also solves the problem of inconsistencies in deformations because of tract-specific ROIs. The registration task within the framework is updated to learn only local deformations rather than an end-to-end composite including rigid transformation, as brain local changes over time is a focus in longitudinal imaging studies. In addition, we compare the performance of Segis-Net to two multistage pipelines based on both classical and deep learning algorithms. The segmentation accuracy, registration accuracy, spatio-temporal consistency of segmentation, and reproducibility of segmentation and tract-specific measures of the three pipelines are quantitatively evaluated. Also, we evaluate the sample-size reduction that can be achieved in the imaging analysis of WM tracts to provide insight into the practical value of the methods in clinical applications.

2. Method

In this section, we first describe how the segmentation and registration tasks are individually modeled using CNN-based approaches. Subsequently, we present the proposed Segis-Net that integrates both tasks in a single-step CNN framework.

2.1. CNN-based image segmentation

Given a n-D image I which can be described by either intensity values, multi-channel features or directional tensors, the goal of CNN-based segmentation is to automatically infer, for each voxel $x \in \mathbb{R}^n$, its probability of belonging to the structure $k \in [1, K]$, i.e., voxel-wise classification. The CNN model can be interpreted as a parameterized mapping function $\mathcal{F}_{\mathbf{\Theta}}$ such that the segmented structures spatially correspond to a segmentation ground truth with multiple channels $\mathcal{S} = \{S_1, ..., S_K\}$. The estimation of the segmentation is formulated as:

$$\hat{\mathcal{S}} = \mathcal{F}_{\mathbf{\Theta}}(I). \tag{1}$$

$\mathcal{F}_{\mathbf{\Theta}}$ is commonly modeled by a nested series of convolutions, non-linearity, normalization, and re-sampling operations embedded in the network architecture. $\mathbf{\Theta}$ indicate trainable parameters, i.e., weights inside convolution kernels.

The procedure of estimating parameter $\mathbf{\Theta}$ is then defined as an optimization with respect to a loss function $\mathcal{L}_{seg}$, aiming at minimizing the classification error over all the N pairs of training samples $\{(\mathcal{S}^i, I^i)\}_{i=1}^N$, i.e.,

$$\mathbf{\Theta} \leftarrow \underset{\mathbf{\Theta}}{\operatorname{argmin}} \sum_{i=1}^{N} \mathcal{L}_{seg}\left(\mathcal{S}^i, \mathcal{F}_{\mathbf{\Theta}}(I^i)\right). \tag{2}$$

The loss function comprises metrics that quantify the difference between the prediction and the ground truth. In this study, $\mathcal{L}_{seg}$ is the average Dice coefficient (Dice, 1945) over all K structures:

$$\mathcal{L}_{seg}(\mathcal{S}, \hat{\mathcal{S}}) = -\frac{2}{K} \sum_{k=1}^{K} \frac{\sum_x S_k \hat{S}_k}{\sum_x (S_k)^2 + \sum_x (\hat{S}_k)^2}. \tag{3}$$

After estimation of the map function $\mathcal{F}_{\mathbf{\Theta}}$, a probabilistic prediction for the structures of interest $\hat{\mathcal{S}}$ in a given image can be inferred (Eq.1).

2.2. CNN-based deformable registration

Let us consider a pair of n-D images, I_s in the source space $\Omega_s \subset \mathbb{R}^n$, and I_t in the target space $\Omega_t \subset \mathbb{R}^n$, which contain a common structure to be aligned. The spatial correspondence between images can be established by estimating a dense displacement field $\boldsymbol{\phi}$, such that $I_s \circ \boldsymbol{\phi}$ and I_t correspond spatially.

In line with the hierarchical optimization scheme of classical registration algorithms, most existing learning-based registration methods use affine alignment as a prepossessing, in which case the displacement field denotes the composition of affine and deformable transform, i.e., $\boldsymbol{\phi} = \boldsymbol{\phi}_A \circ \boldsymbol{\phi}_D$. To estimate $\boldsymbol{\phi}_D$, the CNN model can be interpreted as a shared domain-invariant mapping function $\mathcal{G}_{\boldsymbol{\Psi}}$ such that for any unseen pair of images a most likely transformation between them can be inferred without pair-specific optimization, i.e.,

$$\hat{\boldsymbol{\phi}}_D = \mathcal{G}_{\boldsymbol{\Psi}}(I_t, I_s \circ \boldsymbol{\phi}_A) \tag{4}$$

$$\implies \quad I_s \circ \hat{\boldsymbol{\phi}} \leftarrow \mathcal{G}_{\boldsymbol{\Psi}}(I_t, I_s, \boldsymbol{\phi}_A). \tag{5}$$

The parameters $\boldsymbol{\Psi}$ of the mapping function are optimized based on a registration dissimilarity loss $\mathcal{L}_{reg}$, aimed at minimizing the registration error. Meanwhile, to penalize large deviations of deformation and preserve anatomical topology during transformations, a deformation smoothness term $\mathcal{L}_{def}$ is commonly included in the loss function. In this work, we use the mean squared error based on intensities for $\mathcal{L}_{reg}$ and the average spatial gradients of the displacement field for $\mathcal{L}_{def}$, i.e.,

$$\mathcal{L}_{reg}(I_t, I_s \circ \hat{\boldsymbol{\phi}}) = \frac{1}{|\Omega_t|} \|I_t - I_s \circ \hat{\boldsymbol{\phi}}\|_2^2, \tag{6}$$

$$\mathcal{L}_{def}(\hat{\boldsymbol{\phi}}_D) = \frac{1}{|\Omega_t|} \|\nabla \hat{\boldsymbol{\phi}}_D\|_2^2. \tag{7}$$

Combining Eq. (4), (5), (6) and (7), the estimation of $\boldsymbol{\Psi}$ over all the N training samples $\{(I_t^i, I_s^i, \boldsymbol{\phi}_A^i)\}_{i=1}^N$ can be formulated as:

$$\boldsymbol{\Psi} \leftarrow \underset{\boldsymbol{\Psi}}{\operatorname{argmin}} \sum_{i=1}^{N} \mathcal{L}_{reg}(I_t^i, \mathcal{G}_{\boldsymbol{\Psi}}(I_t^i, I_s^i, \boldsymbol{\phi}_A^i)) + \mathcal{L}_{def}(\mathcal{G}_{\boldsymbol{\Psi}}(I_t^i, I_s^i \circ \boldsymbol{\phi}_A^i)). \tag{8}$$

2.3. Simultaneous estimation of segmentation and registration

In this work, we aim to simultaneously estimate the parameters for segmentation ($\boldsymbol{\Theta}$) and for registration ($\boldsymbol{\Psi}$) in a single-step optimization. For this purpose, we integrate the segmentation and registration function $\mathcal{F}_{\boldsymbol{\Theta}}$ and $\mathcal{G}_{\boldsymbol{\Psi}}$ using an end-to-end optimization with the Segis-Net. The loss function of the Segis-Net is designed to meet the joint objective of both tasks and meanwhile to optimize the spatio-temporal consistency of segmentation which rely on both tasks. The overview of the proposed framework is illustrated in Fig. 1. We describe the framework architecture and loss function in the following paragraphs.

2.3.1. Segis-Net framework

In the present study, we focus on the analysis of 3D images and utilize 3D convolutions for the Segis-Net framework. The framework involves function $\mathcal{F}_{\boldsymbol{\Theta}}$ and $\mathcal{G}_{\boldsymbol{\Phi}}$ as two parallel streams that interact on their outputs. In order to eliminate the loss in image quality caused by multiple interpolations, Segis-Net warps source images with only the composite displacement fields ($\boldsymbol{\phi}$) by taking as input the original source image (I_s) and pre-estimated affine matrix ($\boldsymbol{\phi}_A$). This design has additional advantages over existing methods that prepare all ordered pairs of affine-aligned images in disk storage, as only up to half the storage is needed and as it can be flexibly applied to related images in the same space such as the DTI metrics.

$\mathcal{F}_{\boldsymbol{\Theta}}$ outputs a set of probabilistic segmentations ($\hat{\mathcal{S}}_s$) of the source image. $\mathcal{G}_{\boldsymbol{\Psi}}$ outputs a dense local displacement $\hat{\boldsymbol{\phi}}_D$ along the x, y, and z axes. The source image and its segmentations are subsequently warped into the target space using the composed displacement field. The warp operation is implemented by a computational layer with differentiable trilinear interpolation (Jaderberg et al., 2015; Balakrishnan et al., 2019). The segmentation and registration streams have independent network architectures which are only connected by the output, i.e., the transformed source-segmentation to the target space ($\hat{\mathcal{S}}_s \circ \hat{\boldsymbol{\phi}}$). Thus, they can be applied separately after taking advantage of the simultaneous optimization. The Segis-Net framework gives four outputs during training:

1. The segmentation of the structures of interest from the source image ($\hat{\mathcal{S}}_s$),
2. A local displacement field between the source and target images ($\hat{\boldsymbol{\phi}}_D$),
3. The warped source image in the target space ($I_s \circ \hat{\boldsymbol{\phi}}$),
4. The warped source segmentations in the target space ($\hat{\mathcal{S}}_s \circ \hat{\boldsymbol{\phi}}$).

We propose a generic framework where the architecture of each stream can be adapted based on specific applications. For the particular network used in this study, we encoded two streams with a U-Net architecture, that was modified as detailed below (Ronneberger et al., 2015). In short, each stream was composed of an encoder and decoder path with skip connections of feature pyramid at multiple scales in order to merge coarse- and fine-convolved features, similar to the multi-resolution strategy used in classical algorithms to increase robustness. The encoder paths with max-pooling operation between convolution layers gradually extract abstract features for the target anatomy ($\mathcal{F}_{\boldsymbol{\Theta}}$) and global transformation between images ($\mathcal{G}_{\boldsymbol{\Psi}}$). Subsequently, the decoder paths restore the details in segmentations ($\mathcal{F}_{\boldsymbol{\Theta}}$) and refine local deformations ($\mathcal{G}_{\boldsymbol{\Psi}}$) by linear up-sampling the feature maps and concatenating them with the coarse counterpart at the same scale. The convolution layers produce a set of feature maps by individually convolving inputs with 3D kernels of size $(3, 3, 3)$, followed by batch normalization (Ioffe and Szegedy, 2015) and a leaky ReLu layer ($a = 0.2$) for modeling non-linearity (Maas et al., 2013). For the segmentation stream ($\mathcal{F}_{\boldsymbol{\Theta}}$), we split the output layer into sub-branches to facilitate multi-class classification for voxels

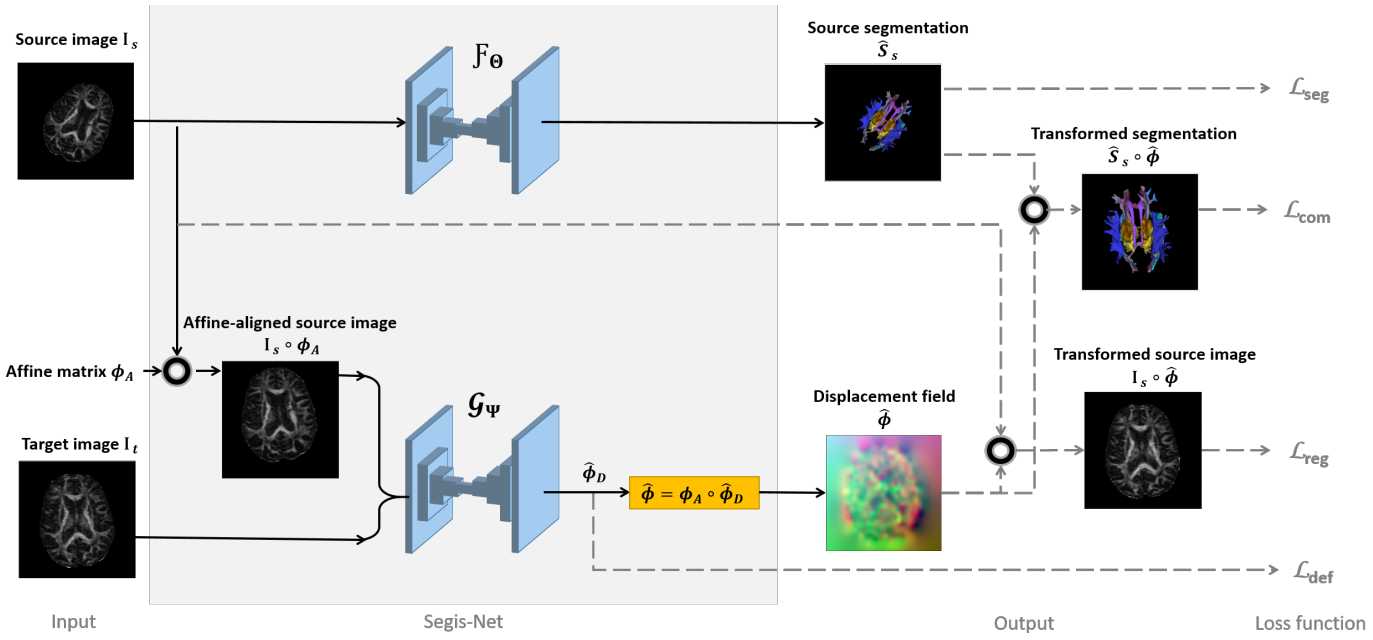


Figure 1: Overview of the Segis-Net framework. $\mathbf{\Theta}$ and $\mathbf{\Psi}$ denote the parameters of the segmentation ($\mathcal{F}_\Theta$) and registration ($\mathcal{G}_\Psi$) function, respectively. Black circle indicates spatial warp with affine matrix (ϕ_A) or the composite displacement field ($\hat{\phi}$). The concatenation of the affine-aligned images is used as the input for $\mathcal{G}_\Psi$. Loss function consists of $\mathcal{L}_{seg}$, $\mathcal{L}_{com}$, $\mathcal{L}_{reg}$ and $\mathcal{L}_{def}$ terms. Solid lines indicate the primary workflow of the method; dash lines indicate that only implemented during training or that can be adapted for applications.

with multiple labels. The final layer of the sub-branches consisted of a $(1, 1, 1)$ convolution and a sigmoid activation. For the registration stream, the output layer was a convolution layer with three kernels that yielded the local displacement $\hat{\boldsymbol{\phi}}_D$. We provide detailed implementation of the network architecture in the supplementary material (Figure 8 and 9).

2.3.2. *Segis-Net loss function*

The loss function of Segis-Net is composed of four terms that measure segmentation accuracy ($\mathcal{L}_{seg}$, Eq. 3), intensity similarity between registered images ($\mathcal{L}_{reg}$, Eq. 6), deformation field smoothness ($\mathcal{L}_{def}$, Eq. 7), and longitudinal composite of registration and segmentation ($\mathcal{L}_{com}$, Eq. 11). It is formulated as:

$$\begin{aligned}\mathcal{L} = &\mathcal{L}_{seg}(\mathcal{S}_s, \hat{\mathcal{S}}_s) \\ &+ \alpha\mathcal{L}_{reg}(I_t, I_s \circ \hat{\boldsymbol{\phi}})) + \beta\mathcal{L}_{def}(\hat{\boldsymbol{\phi}}_D) \\ &+ \gamma\mathcal{L}_{com}(\mathcal{S}_t, \hat{\mathcal{S}}_s \circ \hat{\boldsymbol{\phi}}),\end{aligned} \tag{9}$$

and optimized for $\mathbf{\Theta}$ and $\mathbf{\Psi}$ over all N training samples $\{(\mathcal{S}_t^i, \mathcal{S}_s^i, I_t^i, I_s^i, \boldsymbol{\phi}_A^i)\}_{i=1}^N$:

$$\begin{aligned}\mathbf{\Theta}, \mathbf{\Psi} \leftarrow \underset{\mathbf{\Theta},\mathbf{\Psi}}{\operatorname{argmin}} &\sum_{i=1}^{N} \mathcal{L}_{seg}\left(\mathcal{S}_s^i, \mathcal{F}_\Theta(I_s^i)\right) \\ &+ \alpha\mathcal{L}_{reg}(I_t^i, \mathcal{G}_\Psi(I_t^i, I_s^i, \boldsymbol{\phi}_A^i)) + \beta\mathcal{L}_{def}(\mathcal{G}_\Psi(I_t^i, I_s^i \circ \boldsymbol{\phi}_A^i)) \\ &+ \gamma\mathcal{L}_{com}(\mathcal{S}_t^i, \mathcal{F}_\Theta(I_s^i), \mathcal{G}_\Psi(I_t^i, I_s^i, \boldsymbol{\phi}_A^i)).\end{aligned} \tag{10}$$

We quantify the longitudinal composite loss term using the average Dice coefficient over all K structures:

$$\mathcal{L}_{com}(\mathcal{S}_t, \hat{\mathcal{S}}_s \circ \hat{\boldsymbol{\phi}}) = -\frac{2}{K}\sum_{k=1}^{K}\frac{\sum_{x\in\Omega_t} S_t^k(\hat{S}_s^k \circ \hat{\boldsymbol{\phi}})}{\sum_{x\in\Omega_t}(S_t^k)^2 + \sum_{x\in\Omega_t}(\hat{S}_s^k \circ \hat{\boldsymbol{\phi}})^2}. \tag{11}$$

In longitudinal imaging studies, the spatial correspondence in segmentation depends on the performance of both the segmentation and the registration procedure. Besides an explicit optimization of correspondence, the $\mathcal{L}_{com}$ term also exploits longitudinal information to boost both tasks, which introduces some degree of augmentation and regularization for registration and on the other hand constraints and prior knowledge for segmentation.

The hyperparameters α and γ balance the loss magnitude of segmentation, registration, and their interdependent composite. The degree of regularization is described by β (or β/α). The procedure of simultaneous optimization is summarized with pseudo code in supplementary material (Algorithm 1).

3. Application to diffusion MRI

The performance of Segis-Net is demonstrated by analyzing white matter tracts in a large diffusion MRI dataset, and compared to that of two multi-stage pipelines, in which segmentation and registration are independently optimized. Performance is evaluated in a longitudinal setting where multiple time-points from the same individual are available.

3.1. Dataset

The Rotterdam Study is a prospective and population-based study targeting causes and consequences of age-related diseases (Ikram et al., 2020). For the present analysis, we included 3249 individuals who underwent diffusion MRI scanning twice or more often, resulting in $N = 8045$ scans. The mean age at first scan was 61.2±9.4 years (range: 45.7−91.1 years). The number of female participants was 1780 (54.8%). A flowchart for the inclusion, exclusion, and split of the datasets is shown in Figure 2. We split the data into two subsets. The larger subset was repeatedly acquired in a time interval of 1–5 years ($N = 7770$ scans from 3166 individuals). In these long time-interval scans, it is expected that brain microstructure changes due to aging exist. By matching any two time-points from the same individual regardless of the visiting order, these long time-interval scans can be grouped into 6043 pairs. We used 5175 pairs of scans as training data, 200 pairs as validation data to tune the hyperparameters, monitor the decay of learning rate and select the optimal epoch, and used an independent cohort of 668 pairs for testing. The remaining scans from the smaller subset were from 97 individuals who were scanned twice within a month. No changes in brain macro- and microstructure were expected within such a short time-interval. We used these scans for evaluation of reproducibility of the algorithm. The data split was based on the participants, namely, we made sure that scans from the same participant ended up in either training, validation, or test dataset.

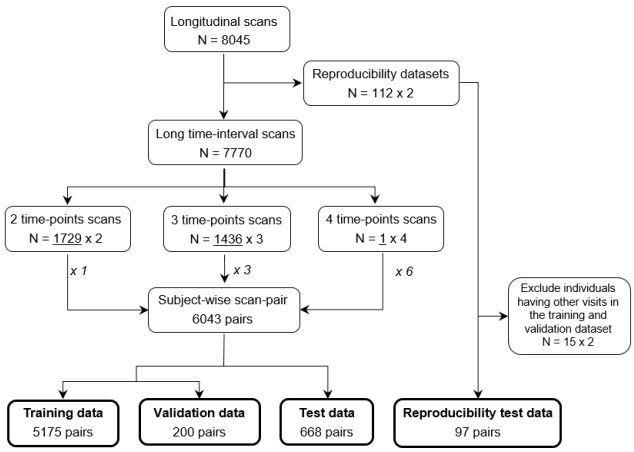


Figure 2: A flowchart for the inclusion, exclusion, and split of the datasets.

3.2. MRI acquisition

Scans were acquired on a 1.5T MRI scanner (GE Signa Excite). The acquisition parameters for structural and diffusion MRI can be found in Ikram et al. (2011). Specifically, diffusion MRI was scanned with the following parameters: TR/TE =8575ms/82.6ms, imaging matrix of 64×96, FOV=21×21cm^2, 35 contiguous slices with slice thickness 3.5 mm, 25 diffusion weighted volumes with a b-value of 1000s/mm^2 and 3 non-weighted volumes (b-value = 0s/mm^2). The voxel size was resampled from $3.3 \times 2.2 \times 3.5mm^3$ to $1mm^3$ as required for probabilistic tractography (Behrens et al., 2007).

3.3. Image preprocessing

Diffusion data were preprocessed using a standardized pipeline (Koppelmans et al., 2014). In short, motion and eddy currents were corrected by affine co-registration of all diffusion weighted volumes to the averaged b0 volumes, including correction of gradient vector directions using Elastix software (Klein et al., 2010). Diffusion tensors were estimated with a Levenberg-Marquard non-linear least-squares optimization algorithm (Leemans et al., 2009). We subsequently computed DTI measures: fractional anisotropy (FA) and mean diffusivity (MD). Due to noise, tensor estimation failed in a small proportion of voxels, resulting in significant outliers. Outlier voxels with a tensor norm (Frobenius norm) larger than $0.1mm^2/s$ were set to zero (Zhang et al., 2007). Brain tissue masks including WM and gray matter segmentations were obtained based on structural imaging (Vrooman et al., 2007) and applied to the diffusion tensor images. In this study, we used a ROI of $112 \times 208 \times 112$ voxels to analyze six WM tracts, including left and right cingulate gyrus part of cingulum (CGC), left and right parahippocampal part of cingulum (CGH), forceps major (FMA) and forceps minor (FMI). Diffusion tensor image (with six components) was image-wise normalized to zero-mean and standard deviation of one. The affine matrix ($\boldsymbol{\phi}_A$) of each image pair was estimated by optimizing the mutual information of FA images using Elastix software.

3.4. Reference segmentations

The segmentation labels for model training and evaluation were generated using a probabilistic tractography and atlas-based segmentation method by De Groot et al. (2015). The resulting tract-density images for each tract were normalized by division with the total number of tracts in the tractography run. Finally, tract-specific thresholds for the normalized density images were established by maximizing the reproducibility of FA measures on a subset of 30 participants, which were included in our reproducibility dataset as well (De Groot et al., 2013b).

3.5. Baseline multi-stage pipelines

We compared the performance of the proposed Segis-Net with two multi-stage pipelines that consist of either non-learning-based or learning-based algorithms.

First, a non-learning-based *Classical* pipeline was built using an existing tractography-based segmentation algorithm as detailed in Section 3.4, and deformable registration algorithm Elastix (Klein et al., 2010). Elastix was adopted as a competing classical registration method since it has been widely used on our dataset and thereby an optimal parameter setting can be applied for performance comparison. Elastix is designed to run in a cascade of resolutions, and offers the choice between multiple objective functions and multiple optimizers including an efficient adaptive stochastic gradient descent optimizer (Klein et al., 2009). For Elastix (version 4.8), we used a rigid, affine, and B-spline transformation model consecutively by maximizing mutual information between images. The B-spline transformation of spline order 3 was implemented using a multi-resolution framework with isotropic control-point spacing of

24, 12, and 6 *mm* in three-level resolutions. The maximum number of iterations was 1024.

Second, we built a learning-based *CNN* pipeline using components from the proposed Segis-Net to evaluate the sole contribution of simultaneous optimization. In this pipeline, we split the integrated segmentation $\mathcal{F}_{\mathbf{\Theta}}$ and registration stream $\mathcal{G}_{\mathbf{\Phi}}$ into two separate neural networks for independent optimization. Subsequently, the segmented images and estimated transformations were combined. The segmentation network had the same architecture as that for $\mathcal{F}_{\mathbf{\Theta}}$, except being independently optimized using the segmentation accuracy $\mathcal{L}_{seg}$ term. As this is a typical setting for CNN-based segmentation approaches (Li et al., 2018; Ronneberger et al., 2015), we denote it as *Seg-Net*. Similarly, the registration network denoted as *Reg-Net* had the same setting as that for $\mathcal{G}_{\mathbf{\Psi}}$, except being independently optimized using registration similarity $\mathcal{L}_{reg}$ and regularization $\mathcal{L}_{def}$ terms (Balakrishnan et al., 2019). We ensured that the training dataset for the *Seg-Net* and *Reg-Net* was the same as that used for the Segis-Net framework.

3.6. Implementation

For this diffusion MRI application, the segmentation ($\mathcal{F}_{\mathbf{\Theta}}$) and registration ($\mathcal{G}_{\mathbf{\Psi}}$) components of Segis-Net used different input images. Specifically, segmentation was based on the diffusion tensor image, as it contains directional information of fiber populations and was shown to be optimal in the present setting of clinical-quality resolution (Li et al., 2020a). For spatial alignment, we adopted the input the commonly used scalar-value FA map derived from diffusion tensor imaging.

To mitigate class imbalance and to improve computational efficiency, we combined the reference segmentation for the six tracts (Section 3.3) into a three-channel map for using it as the segmentation ground truth S. This combination was possible since only few crossing fibers are expected between codirectional WM tracts (e.g., FMI and FMA). To evaluate performance on individual tracts after training, we extracted the two largest components from each of the three channels of the probabilistic prediction and subsequently identified the left and right (for CGC and CGH) or the anterior and posterior (for FMI and FMA) tract based on coordinates.

The experiments of model training and evaluation were performed on an NVIDA 1080Ti GPU and an AMD 1920X CPU. CNN-based methods were implemented using Keras-2.2.0 with a Tensorflow-1.4.0 backend and the Adam optimizer (Kingma and Ba, 2014). Weights of convolution kernels were initialized with the Glorot uniform distribution (Glorot and Bengio, 2010). In each training epoch, input images were fed in random batches (size = 1). Loss function hyperparameters were optimized based on segmentation and registration performance on the validation dataset (search range: $[10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}]$); and were set to $\alpha = 10$, $\beta = 0.1$ for *Reg-Net* and subsequently set to $\gamma = 1$ for Segis-Net. The initial learning rates were experimentally optimized on the validation dataset and set to $1e^{-4}$, $1e^{-3}$ and $1e^{-3}$ for *Reg-Net*, *Seg-Net* and Segis-Net, which were decayed with a factor of 0.8 if the validation loss stopped decreasing for 10 epochs (decay condition, Algorithm 1). We stopped the training procedure at the point that the validation loss showed consecutive increases, i.e., early stopping (Bishop, 2006). The parameters of the model with the smallest error with respect to the validation dataset were used.

4. Experiments and results

We applied the proposed Segis-Net and two baseline multi-stage pipelines to analyze six WM tracts on the test and reproducibility datasets. To assess whether difference of performance between approaches was statistically significant, paired t-tests with $\alpha = 0.05$ and Bonferroni correction for controlling the family-wise error of multiple testings were performed. In addition, we compared Segis-Net with two relevant methods in literature.

4.1. Segmentation accuracy

The segmentation accuracy of the proposed Segis-Net was compared with that obtained by the *CNN* pipeline, i.e., *Seg-Net*. We quantified segmentation accuracy with respect to the reference segmentation using the Dice coefficient on the test dataset.

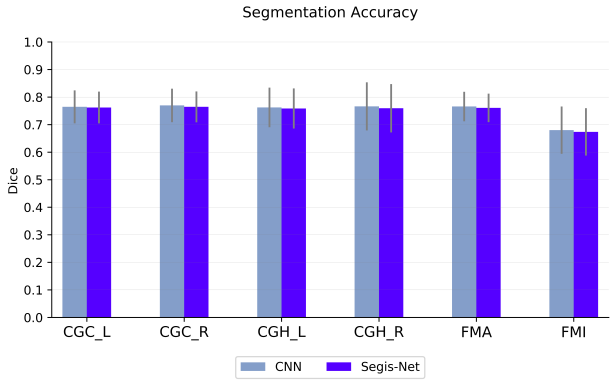


Figure 3: Segmentation accuracies of the *CNN* pipeline and Segis-Net for different tracts. Error bars indicate standard deviations.

The segmentation accuracies of the *CNN* pipeline and Segis-Net were similar for the six tracts (Figure 3). Both methods achieved relatively high accuracy in segmenting cingulum, i.e., the accuracy of left and right CGC and CGH tracts was around 0.76 ± 0.07. The accuracy was lowest for FMI (*CNN*: 0.68 ± 0.09; Segis-Net: 0.67 ± 0.09), which is a thin and arch-shaped tract that is known to be more difficult to segment. Correcting for 6 tests resulted in an adjusted P-value threshold of $8.3{\times}10^{-3}$. There was no significant differences in segmentation accuracy between methods.

4.2. Registration accuracy

Registration accuracy of the three pipelines was evaluated with the spatial correlation (SC) similarity on the test dataset. According to the procedure in De Groot et al. (2013b) the estimated transformation was applied to the continuous density

maps of individual tracts obtained from probabilistic tractography, subsequently, the SC similarity between warped density maps was computed as follow:

$$SC_k = \frac{\sum_{x\in\Omega_t} J_t^k (J_s^k \circ \hat{\boldsymbol{\phi}})}{\left(\sum_{x\in\Omega_t} \sqrt{(J_t^k)^2}\right)\left(\sum_{x\in\Omega_t} \sqrt{(J_s^k \circ \hat{\boldsymbol{\phi}})^2}\right)}, \quad (12)$$

where J_t^k and J_s^k indicate intensity of the target and source density image of the tract k.

Despite a lot of intensity variation in the tract density maps across scans due to the probabilistic nature of tractography, higher intensity in general indicates more support for the tract while lower intensity conversely indicates increased uncertainty. Therefore, we assume that SC reflects the spatial correspondence of tracts.

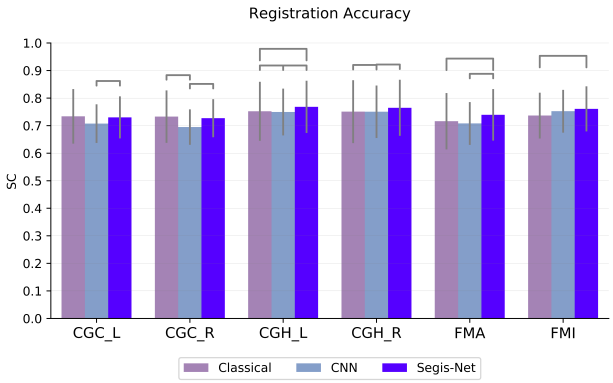


Figure 4: Registration accuracies of the *Classical*, *CNN*, and Segis-Net pipeline as quantified by spatial correlation (SC) of the registered tract density maps. Error bars indicate standard deviations. The bracket hat indicates a significant difference between two methods (t-test, $p < 2.8 \times 10^{-3}$).

The SC in all six tracts was overall highest for the Segis-Net, followed by the *Classical* pipeline (Figure 4). Correcting for 18 tests resulted in Bonferroni adjusted P-value threshold of 2.8×10^{-3}. Segis-Net results yielded a significantly better spatial correspondence than the *Classical* pipeline in the left CGH (Segis-Net vs *Classical* = 0.77 ± 0.09 vs 0.75 ± 0.11), FMA (0.74 ± 0.09 vs 0.72 ± 0.10), and FMI (0.76 ± 0.08 vs 0.74 ± 0.08) tract. Significantly higher registration accuracy over the *CNN* pipeline was observed in the left CGC (Segis-Net vs *CNN* = 0.73 ± 0.08 vs 0.71 ± 0.07), right CGC (0.73 ± 0.07 vs 0.69 ± 0.06), left CGH (0.77 ± 0.09 vs 0.75 ± 0.08), right CGH (0.76 ± 0.10 vs 0.75 ± 0.10), and FMA (0.74 ± 0.09 vs 0.71 ± 0.08) tract. In general, the proposed Segis-Net approach achieved a better spatial correspondence than the two independently optimized registration algorithms using classical and learning-based techniques.

4.3. Spatio-temporal consistency of segmentation

To evaluate the spatio-temporal consistency of segmentation (STCS) for each of the three pipelines, we measured the correspondence between warped segmentation results across time-points using the Dice coefficient. The consistency of each tract was averaged over two directions by reversing the target and source image, which for the tract k can be formulated as:

$$STCS_k = \frac{1}{2}\left(\frac{2|\hat{S}_t^k \cap \hat{S}_s^k \circ \hat{\boldsymbol{\phi}}|}{|\hat{S}_t^k| + |\hat{S}_s^k \circ \hat{\boldsymbol{\phi}}|} + \frac{2|\hat{S}_s^k \cap \hat{S}_t^k \circ \hat{\boldsymbol{\phi}}^{-1}|}{|\hat{S}_s^k| + |\hat{S}_t^k \circ \hat{\boldsymbol{\phi}}^{-1}|}\right). \quad (13)$$

Each pipeline was evaluated as a whole, that is, 1) in *Classical* pipeline the reference segmentation was warped by Elastix algorithm, 2) in *CNN* pipeline the prediction of *Seg-Net* was warped by the predicted transformation of the *Reg-Net*, and 3) in Segis-Net framework the segmentation prediction in native space and the segmentation warped from another time-point were available after a bidirectional test.

The proposed Segis-Net overall showed higher segmentation consistency than the *CNN* and the *Classical* pipeline (Figure 5). Correcting for 18 tests resulted in an adjusted P-value threshold of 2.8×10^{-3}. In comparison with the *CNN* pipeline, Segis-Net results yielded significantly higher spatio-temporal consistency in left CGC (Segis-Net vs *CNN* = 0.83 ± 0.04 vs 0.82 ± 0.04), right CGC (0.83 ± 0.04 vs 0.82 ± 0.06), left CGH (0.82 ± 0.05 vs 0.81 ± 0.05), FMA (0.87 ± 0.02 vs 0.84 ± 0.03), and FMI (0.81 ± 0.05 vs 0.77 ± 0.07) tract. In all six tracts, Segis-Net significantly outperformed the *Classical* pipeline, i.e., in left CGC (Segis-Net vs *Classical* = 0.83 ± 0.04 vs 0.68 ± 0.06), right CGC (0.83 ± 0.04 vs 0.68 ± 0.06), left CGH (0.82 ± 0.05 vs 0.66 ± 0.08), right CGH (0.81 ± 0.05 vs 0.66 ± 0.09), FMA (0.87 ± 0.02 vs 0.69 ± 0.06), and FMI (0.81 ± 0.05 vs 0.57 ± 0.09) tract.

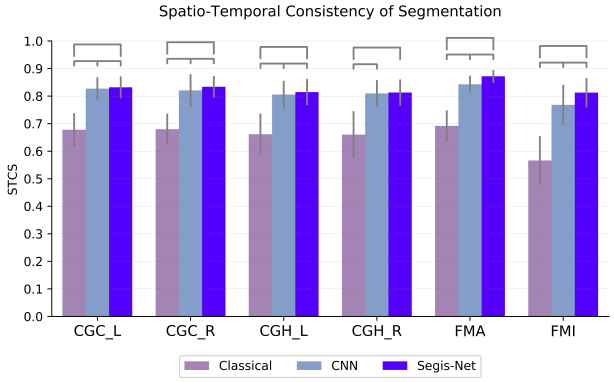


Figure 5: Spatio-temporal consistency of segmentation (STCS) with the *Classical*, *CNN*, and Segis-Net pipeline. Error bars indicate standard deviations. The bracket hat indicates a significant difference between two methods (t-test, $p < 2.8 \times 10^{-3}$).

4.4. Reproducibility of segmentation and measurements

Reproducibility of tract-specific segmentations, volumes, and diffusion metrics of the three pipelines was evaluated using the reproducibility dataset. We quantified voxel-wise agreement between segmentations of repeated scans using Cohen's kappa coefficient (κ). The segmentations $(\hat{S}_t, \hat{S}_s)$ were obtained in the native space, and subsequently aligned $(\hat{S}_s \circ \hat{\boldsymbol{\phi}})$. Kappa κ of the tract k is defined as:

$$\kappa_k = \frac{p_o(\hat{S}_t^k, \hat{S}_s^k \circ \hat{\boldsymbol{\phi}}) - p_e(\hat{S}_t^k, \hat{S}_s^k \circ \hat{\boldsymbol{\phi}})}{1 - p_e(\hat{S}_t^k, \hat{S}_s^k \circ \hat{\boldsymbol{\phi}})}, \quad (14)$$

in which $p_o(\hat{S}_t^k, \hat{S}_s^k \circ \hat{\boldsymbol{\phi}})$ is the observed agreement between $\hat{S}_t^k$ and $\hat{S}_s^k \circ \hat{\boldsymbol{\phi}}$, p_e is the hypothetical probability of the agreement. Given $|\Omega_t|$ being the total number of voxels in the target image, $|S|$ and $|\Omega_t| - |S|$ being the number of tract and non-tract voxels, the observed agreement (i.e., accuracy) is computed as:

$$p_o(\hat{S}_t^k, \hat{S}_s^k \circ \hat{\boldsymbol{\phi}}) = \frac{|\hat{S}_t^k \cap (\hat{S}_s^k \circ \hat{\boldsymbol{\phi}})| + |(1 - \hat{S}_t^k) \cap (1 - (\hat{S}_s^k \circ \hat{\boldsymbol{\phi}}))|}{|\Omega_t|}, \tag{15}$$

the hypothetical probability of the agreement can be formulated as:

$$p_e(\hat{S}_t^k, \hat{S}_s^k \circ \hat{\boldsymbol{\phi}}) = \frac{1}{|\Omega_t|^2}(|\hat{S}_t^k| \times |\hat{S}_s^k \circ \hat{\boldsymbol{\phi}}| + (|\Omega_t| - |\hat{S}_t^k|) \times (|\Omega_t| - |\hat{S}_s^k \circ \hat{\boldsymbol{\phi}}|)). \tag{16}$$

Typically, a $\kappa > 0.60$ indicates "substantial" agreement, and a $\kappa > 0.80$ indicates "almost perfect" agreement (Landis and Koch, 1977).

Similarly, to evaluate the reproducibility of tract-specific measurements, we computed the FA, MD and volume in image native space, and subsequently assessed relative differences in paired scan-rescan measures (m_t , m_s) as an indicator of measurement error (ϵ), i.e.,

$$\epsilon = \frac{2|m_s - m_t|}{(m_s + m_t)} \times 100\%. \tag{17}$$

For FA and MD, the tract-specific measures were quantified as the median of non-zero values within the segmented images. A lower ϵ indicates a better reproducibility.

Figure 6 presents the reproducibility of tract-specific segmentation and measures determined with the three pipelines. The proposed Segis-Net achieved the best segmentation reproducibility, followed by the *CNN* pipeline (Figure 6 (a)); in all six tracts, κ was around 0.80 or higher, indicating "almost perfect" agreements between segmentations of repeated scans. Correcting for 18 tests for each metric resulted in an adjusted P-value threshold of 2.8×10^{-3}, resulting in overall statistically significant improvement by Segis-Net over the *Classical* pipeline. For two tracts, voxel-wise agreement of Segis-Net was significantly higher than that of the *CNN* pipeline, i.e., FMA (Segis-Net vs *CNN* = 0.87 ± 0.03 vs 0.85 ± 0.03) and FMI (0.82 ± 0.06 vs 0.79 ± 0.08).

Additionally, in the evaluation of the reproducibility in tract-specific volume measures, Segis-Net showed the smallest error in all six tracts (Figure 6 (b)). The error of Segis-Net was significantly smaller than the *Classical* pipeline in left CGC (Segis-Net vs *Classical* = $4.8 \pm 4.1\%$ vs $11 \pm 8.8\%$), right CGC ($4.5 \pm 3.9\%$ vs $11 \pm 8.9\%$), left CGH ($7.3 \pm 5.6\%$ vs $11 \pm 9.8\%$), and FMA ($3.4 \pm 2.6\%$ vs $6.6 \pm 5.9\%$) tract. This outperformed the *CNN* pipeline significantly in the FMA tract (Segis-Net vs *CNN* = $3.4 \pm 2.6\%$ vs $4.9 \pm 3.6\%$). Reproducibility of FA and MD measurements was similar for the three methods (Figure 6 (c, d)). For the CGH and left CGC tracts, the reproducibility of FA using the *CNN* pipeline was significantly higher than that of the *Classical* pipeline. Segis-Net outperformed the FA reproducibility of the *Classical* pipeline only in the left CGC tract. For MD, no significant improvement over the *Classical* pipeline was observed. A table (Table 3) with the results of Figure 3-6 is provided in the supplementary files.

4.5. Sample-size reduction

An implication of the reduced measurement error (ϵ) is that fewer participants or time-points would be required to achieve the same statistical power, i.e., a smaller sample size. We followed Diggle et al. (2002) and Reuter et al. (2012) to estimate the percentage of the sample sizes (P) that would be required for each of the three pipelines:

$$P_{ij} = \frac{\sigma_i^2 \times (1 - \rho_i)}{\sigma_j^2 \times (1 - \rho_j)} \times 100\%, \tag{18}$$

where σ_i and σ_j are standard deviations in the measurements determined with the pipeline i and j, and ρ_i and ρ_j are the correlation coefficients between the repeated measurements determined with the two pipelines.

Figure 7 presents the percentage of sample-size reduction that could be achieved by the *CNN* and the proposed Segis-Net compared to the *Classical* pipeline. In line with the reproducibility results, the data analyzed with Segis-Net would overall require the least sample-size to achieve the same statistical power. The percentage of reduction was especially remarkable in volume measures, in which on average only 33.0% of data would be required. The average percentage of reduction was 60.5% for FA and 57.0% for MD. Several percentages of the *CNN* pipeline were smaller than those of the Segis-Net, e.g., in FA measures of CGH and FMA tract (Figure 7 (b)), but its performance showed to be less stable across tracts than the Segis-Net, which in all settings consistently decreased in the required samples over the *Classical* pipeline. The percentage of reduction was generally similar for left/right homologous tracts except for the MD measure in the left of CGH (Figure 7 (c)). This large reduction could be related to the MD reproducibility of the *Classical* pipeline, in which the left CGH tract had a much higher variation in errors comparing with that of the other tracts (Figure 6 (d)).

4.6. Comparison with related methods

In this section, we compared the proposed Segis-Net with two previously published methods that also involve segmentation and registration: 1) U-ReSNet for simultaneous optimization that used a shared feature encoder (Estienne et al., 2019), and 2) VoxelMorph for image registration that used an auxiliary loss term of correspondence in existing segmentation labels (Balakrishnan et al., 2019).

Since the same input for both segmentation and registration tasks was required by the shared feature-encoder of U-ReSNet, i.e., a pair of scalar images, the FA map was used as input. This was different from the FA-based registration and diffusion tensor-based segmentation used in the proposed Segis-Net (section 3.6). For a fair comparison, we applied affine registration to the input image pairs as a pre-processing step. Also, we tuned the hyperparameters on the validation dataset and obtained improved performance by using an initial learning rate of 0.0005 and by clipping the warped segmentation predictions into the range of $[10^{-7}, 1 - 10^{-7}]$.

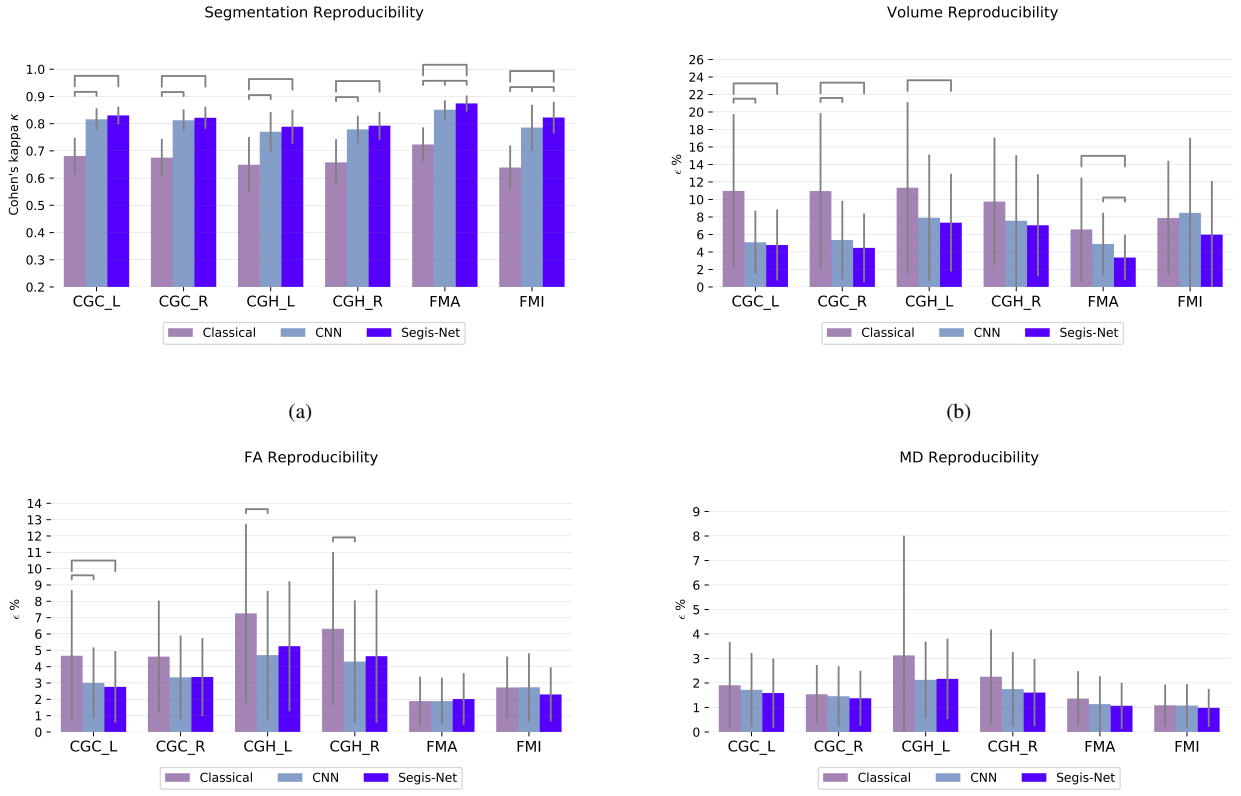
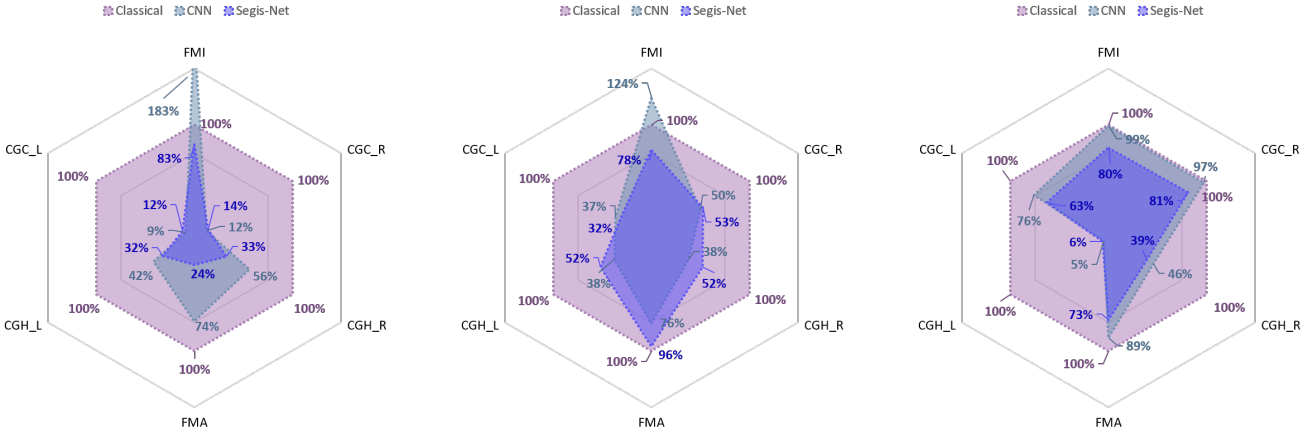


Figure 6: Reproducibility of tract-specific measures with the *Classical*, *CNN*, and Segis-Net pipeline. Error bars indicate standard deviations. The bracket hat indicates a significant difference between two methods (t-test, $p < 2.8{\times}10^{-3}$). In figure (a), a higher Cohen's kappa coefficient (κ) indicates a better reproducibility. In figure (b-d), a lower error ($\epsilon\%$) indicates a better reproducibility. Volume: tract-specific volume (ml), FA: fractional anisotropy, MD: mean diffusivity ($10^{-3}mm^2/s$).



(a) Volume Measures (b) FA Measures (c) MD Measures

Figure 7: The percentage of sample-size that would be required in tract measures of volume, FA, and MD with the *CNN* pipeline and Segis-Net. The sample size required for the *Classical* pipeline is used as the reference (100%).

VoxelMorph presents an approach that uses correspondence in existing segmentation labels to boost registration. To investigate the benefit of this approach in comparison with the proposed simultaneous optimization, we applied VoxelMorph to the registration of longitudinal FA images on our dataset. The implementation as detailed by Balakrishnan et al. (2019) was used directly.

4.6.1. Assessment of segmentation performance

We compared segmentation accuracy between U-ReSNet and the proposed Segis-Net on the test dataset, which was quantified by the Dice coefficient with respect to the reference segmentation (section 3.6).

For all six tracts, the segmentation accuracy of Segis-Net was higher than that of U-ReSNet by a margin of around 10% and with a smaller standard deviation (Table 1).

	U-ReSNet	Segis-Net
CGC_L	0.69 ± 0.06	**0.76 ± 0.06**
CGC_R	0.69 ± 0.07	**0.76 ± 0.06**
CGH_L	0.67 ± 0.08	**0.76 ± 0.07**
CGH_R	0.67 ± 0.09	**0.76 ± 0.09**
FMA	0.69 ± 0.06	**0.76 ± 0.05**
FMI	0.60 ± 0.08	**0.67 ± 0.09**

Table 1: Segmentation Dice coefficient of U-ReSNet and Segis-Net. The bold value indicates a better performance between methods.

4.6.2. Assessment of registration performance

We compared registration performance between U-ReSNet, VoxelMorph and the proposed Segis-Net on the test dataset by evaluating the spatial correlation (SC) similarity (Eq. 12) of registered density map of tracts, the Dice coefficient (DC) of registered reference segmentation of tracts, and the mean squared error (MSE) of registered FA maps.

Generally, U-ReSNet and Segis-Net yielded better SC similarity than that of VoxelMorph; Segis-Net achieved the best DC and MSE (Table 2); the standard deviation of Segis-Net was overall smallest for all three metrics, except that of DC in three tracts which were smallest for VoxelMorph. In the evaluation of SC similarity, U-ReSNet led the highest similarity in the left and right CGC, and the right of CGH tract; Segis-Net was the highest for FMA and FMI tract; for the left CGH tract, a similar high SC was observed for U-ReSNet and Segis-Net, although the variation was smaller in Segis-Net. For all six tracts, the DC of Segis-Net were higher than that of U-ReSNet, followed by VoxelMorph.

		U-ReSNet	VoxelMorph	Segis-Net
SC	CGC_L	**0.77 ± 0.11**	0.72 ± 0.09	0.73 ± 0.08
	CGC_R	**0.77 ± 0.11**	0.71 ± 0.09	0.73 ± 0.07
	CGH_L	**0.77 ± 0.11**	0.75 ± 0.10	**0.77 ± 0.10**
	CGH_R	**0.77 ± 0.12**	0.75 ± 0.11	0.76 ± 0.10
	FMA	0.73 ± 0.11	0.73 ± 0.10	**0.74 ± 0.09**
	FMI	0.75 ± 0.09	0.74 ± 0.08	**0.76 ± 0.08**
DC	CGC_L	0.69 ± 0.07	0.65 ± 0.06	**0.74 ± 0.06**
	CGC_R	0.70 ± 0.07	0.65 ± 0.06	**0.74 ± 0.05**
	CGH_L	0.67 ± 0.08	0.64 ± 0.07	**0.71 ± 0.08**
	CGH_R	0.67 ± 0.10	0.64 ± 0.08	**0.71 ± 0.09**
	FMA	0.70 ± 0.06	0.68 ± 0.06	**0.72 ± 0.06**
	FMI	0.57 ± 0.10	0.56 ± 0.07	**0.60 ± 0.09**
MSE ($\times 10^{-2}$)		0.47 ± 0.38	0.19 ± 0.88	**0.13 ± 0.10**

Table 2: Registration performance of U-ReSNet, VoxelMorph and Segis-Net, as quantified by the spatial correlation (SC) similarity, the Dice coefficient (DC), and the mean squared error (MSE). The bold value indicates a better performance between methods.

5. Discussion

We developed a single-step deep learning framework, coined Segis-Net, for simultaneous optimization of segmentation and registration. The method was applied to analyze changes in WM tracts from a large set of longitudinal diffusion MRI images. To evaluate the performance of the method, we compared it with two multistage pipelines consisting of independent segmentation and registration components, i.e., the *Classical* and *CNN* pipeline. Segis-Net showed improved performances in registration accuracy, spatio-temporal consistency of segmentation, and reproducibility of segmentation and tract-specific measures. We evaluated the practical value of the improved performance in terms of sample-size reduction that could be achieved when employing the method. The tract-specific mesures analyzed with Segis-Net would only require 33.0% – 60.5% sample-size of the data for achieving the same effect size as the *Classical* pipeline.

To date most developments in longitudinal analysis frameworks have focused on unbiased ways of registering image time series (Metz et al., 2011; Keihaninejad et al., 2013), in which a multistage approach combing independent segmentation and registration components is often used (De Groot et al., 2013a; Yendiki et al., 2016). In this paper, we aimed to investigate a different way to improve the performance of the longitudinal framework by using a single-step CNN that optimizes both tasks simultaneously. The sole value of simultaneous optimization was demonstrated by the comparison with the *CNN* pipeline. There was no benefit observed for segmentation alone, but for registration, spatio-temporal consistency of segmentation, and reproducibility, simultaneous optimization led to significantly improved performance.

In the segmentation evaluation, similar accuracies for the *CNN* and Segis-Net framework was observed for the six tracts (Figure 3). Relative segmentation accuracy between individual tracts were in line with those reported in literature (Wasserthal et al., 2018; Li et al., 2020a). For instance, a small and thin object like the FMI tract tended to have a lower Dice coefficient than the larger cingulum and FMA tracts. In the task of registration alone, Segis-Net overall yielded the best accuracy among the three pipelines, significantly outperforming the *Classical* pipeline for three tracts and the *CNN* pipeline for five tracts (Figure 4). This is an important observation as, first, it showed that simultaneous optimization was beneficial to one of the individual tasks, and second, it is non-trivial to improve registration accuracy over a classical algorithm, in which the transformation is pair-wise optimized on the test images.

In all six tracts, we observed substantially higher spatio-temporal consistency of segmentation and reproducibility of segmentation with Segis-Net than with the two multistage pipelines (Figure 5, 6). The spatio-temporal consistency of segmentation as quantified by the Dice coefficient ranged 0.81 – 0.87 for Segis-Net, significantly outperforming the *Classical* pipeline for all the six tracts (range: 0.57 – 0.69) and the *CNN* pipeline for five tracts (range: 0.77 – 0.84). The segmentation reproducibility as quantified by Cohen's kappa ranged 0.79 – 0.87 for Segis-Net, significantly higher than the *Classi-*

cal pipeline for all the six tracts (range: 0.64 – 0.72) and the *CNN* pipeline for two tracts (range: 0.77 – 0.85). These results indicate that Segis-Net can serve as a reliable alternative to the *Classical* pipeline in spatially capturing macro-structural brain changes over time.

In addition, more significant improvements were observed for the reproducibility of tract-specific volume assessment, but not for the FA and MD measures. For volume reproducibility, Segis-Net yielded the least error in the measurements of scan and re-scan, followed by the *CNN* pipeline (Figure 6). For the FA and MD measures, we observed relatively similar reproducibility for the three methods, in which significant difference was only observed in FA reproducibility of CGH and left CGC tract. This suggests that diffusion measures are quite robust to variations in the geometry of the segmented tract.

These improved performances have practical values in a power analysis, where both the *CNN* pipeline and Segis-Net showed to be able to reduce the required sample-size to achieve the same statistical power as the *Classical* pipeline. The data processed with Segis-Net would require on average 33.0% of the sample-size for volume measures, 60.5% for FA, and 57.0% for MD measures, requiring consistently a decreased sample-size for all the settings. The averaged percentages for the *CNN* pipeline were 62.7%, 60.5% and 68.7%. For FMI tract, it would, however, require 183% and 124% of the sample-size for the volume and FA measures. The observed dispersion of sample-size reduction with the *CNN* pipeline may suggest that simultaneous optimization was beneficial to the robustness of the method across the concurrently segmented tracts.

In the comparison with related methods, Segis-Net had a better segmentation performance than U-ReSNet, an existing simultaneous method (Table 1). We expect this improved performance of Segis-Net because of two reasons: 1) the method allows the use of diffusion tensor images for tract segmentation, as we use parallel network modules and only align the predicted segmentation; this circumvents the need to interpolate tensor images. In other words, task-specific inputs can be used; and 2) the sub-branches in the segmentation stream (Figure 8) are designed for the prediction of white matter tracts which can overlap with each other, unlike the exclusive tissue labels focused by other works.

During the comparison of registration performance, we observed two interesting results (Table 2). First, VoxelMorph was the only method that directly optimized on the DC metric, but it led to a least DC score. This can be due to the fact that the segmentation labels used in diffusion imaging studies are often independently obtained for each image, which is much less correlated to the registration performance than is the case for atlas-based segmentation (Balakrishnan et al., 2019). As a result, the alignment of "imperfect" segmentation labels can be an obstructive loss term instead. Second, although the MSE of U-ReSNet was almost four times that of the Segis-Net, it achieved a good SC similarity, especially in small structures like CGC and CGH. This can be attributed to the formulation of their registration loss as the sum of local-SC and MSE.

Whereas the method is generic, we specifically implemented and optimized it for longitudinal study in diffusion MRI data. In diffusion MRI application, we adopted the commonly used scalar-value FA map as the input for registration. Deformable registration of diffusion tensor images is known to be challenging due to the directional components contained in voxels. Despite developments in classical methods for tensor reorientation during the optimization (Cao et al., 2006; Zhang et al., 2007), for learning-based registration it still largely remains unexplored. With the promising results of diffusion tensor interpolation as shown by Grigorescu et al. (2020), Segis-Net based on solely tensor images would be an interesting direction to explore.

The Segis-Net framework presented in the current study is limited to two time-points. This is because learning-based registration algorithms currently only support pairwise transformations (Balakrishnan et al., 2019). One limitation of our method is therefore that it does not allow for analysis of arbitrary number of time-points. In the present study, we grouped the available triple time-points from the same participant into orderless image-pairs for bidirectional analysis. A future possible improvement of the method could be extending the registration component of Segis-Net to enable learning-based group-wise analysis of a set of time-points (Li et al., 2020b).

Beyond the current application, we expect that this work could be extended to other imaging sequences and for example for segmentation of lesion images. For future work, we plan to adapt the proposed method to analyze brain diseases with large and progressive changes. For instance, registration of brains with lesions due to cortical infarct may benefit from a simultaneous segmentation of infarct regions.

6. Conclusion

We proposed a single-step deep learning framework for longitudinal diffusion MRI analysis, in which segmentation and deformable registration were integrated for simultaneous optimization. The comparison with two multistage approaches showed that the proposed Segis-Net can be applied as a reliable tool to support spatio-temporal analysis of WM tracts from longitudinal diffusion MRI imaging. Besides the improved performances, a two-in-one framework for concurrent segmentation and registration also enables a light-weight way of fast quantification of brain changes overtime. This may lead to a more prominent role for tract-specific biomarkers in applications where tract segmentation and registration are subject to time constraints. With the increasing availability of longitudinal diffusion data, we expect future studies investigating progressive neurodegeneration can greatly benefit from the improved reliability and efficiency of Segis-Net.

Acknowledgments

This work was sponsored through grants of the Medical Delta Diagnostics 3.0: Dementia and Stroke, the EU Horizon 2020 project EuroPOND (666992), the Netherlands CardioVascular Research Initiative (Heart-Brain Connection: CVON2012-06, CVON2018-28), and the Dutch Heart Foundation (PPP Allowance, 2018B011).

Data availability

The datasets analyzed during the current study are not publicly available. Due to the sensitive nature of the data used in this study, participants were assured raw data would remain confidential and would not be shared.

References

Ashburner, J., Friston, K.J., 2005. Unified segmentation. Neuroimage 26, 839–851.

Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2019. VoxelMorph: a learning framework for deformable medical image registration. IEEE Trans. Med. Imag. .

Barrick, T.R., Charlton, R.A., Clark, C.A., Markus, H.S., 2010. White matter structural decline in normal ageing: a prospective longitudinal study using tract-based spatial statistics. Neuroimage 51, 565–577.

Bastiaansen, W.A., Rousian, M., Steegers-Theunissen, R.P., Niessen, W.J., Koning, A., Klein, S., 2020. Towards segmentation and spatial alignment of the human embryonic brain using deep learning for atlas-based registration, in: International Workshop on Biomedical Image Registration, Springer. pp. 34–43.

Behrens, T.E., Berg, H.J., Jbabdi, S., Rushworth, M.F., Woolrich, M.W., 2007. Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? Neuroimage 34, 144–155.

Bishop, C.M., 2006. Pattern recognition and machine learning. springer.

Bortsova, G., Dubost, F., Hogeweg, L., Katramados, I., de Bruijne, M., 2019. Semi-supervised medical image segmentation via learning consistency under transformations, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 810–818.

Cao, Y., Miller, M.I., Mori, S., Winslow, R.L., Younes, L., 2006. Diffeomorphic matching of diffusion tensor images, in: 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06), IEEE. pp. 67–67.

Cheng, J., Tsai, Y.H., Wang, S., Yang, M.H., 2017. Segflow: Joint learning for video object segmentation and optical flow, in: Proceedings of the IEEE international conference on computer vision, pp. 686–695.

Conturo, T.E., Lori, N.F., Cull, T.S., Akbudak, E., Snyder, A.Z., Shimony, J.S., McKinstry, R.C., Burton, H., Raichle, M.E., 1999. Tracking neuronal fiber pathways in the living human brain. Proceedings of the National Academy of Sciences 96, 10422–10427.

Crick, F., Jones, E., 1993. Backwardness of human neuroanatomy. Nature 361, 109–110.

Dai, X., Khorram, S., 1999. A feature-based image registration algorithm using improved chain-code representation combined with invariant moments. IEEE Transactions on Geoscience and Remote Sensing 37, 2351–2362.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26, 297–302.

Diggle, P., Diggle, P.J., Heagerty, P., Liang, K.Y., Heagerty, P.J., Zeger, S., et al., 2002. Analysis of longitudinal data. Oxford University Press.

Dimond, D., Rohr, C.S., Smith, R.E., Dhollander, T., Cho, I., Lebel, C., Dewey, D., Connelly, A., Bray, S., 2020. Early childhood development of white matter fiber density and morphology. NeuroImage 210, 116552.

Estienne, T., Lerousseau, M., Vakalopoulou, M., Alvarez Andres, E., Battistella, E., Carré, A., Chandra, S., Christodoulidis, S., Sahasrabudhe, M., Sun, R., et al., 2020. Deep learning-based concurrent brain registration and tumor segmentation. Frontiers in Computational Neuroscience 14, 17.

Estienne, T., Vakalopoulou, M., Christodoulidis, S., Battistela, E., Lerousseau, M., Carre, A., Klausner, G., Sun, R., Robert, C., Mougiakakou, S., et al., 2019. U-ReSNet: Ultimate coupling of registration and segmentation with deep nets, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 310–319.

Farbota, K.D., Bendlin, B.B., Alexander, A.L., Rowley, H.A., Dempsey, R.J., Johnson, S.C., 2012. Longitudinal diffusion tensor imaging and neuropsychological correlates in traumatic brain injury patients. Frontiers in human neuroscience 6, 160.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., et al., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341–355.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp. 249–256.

Gooya, A., Pohl, K.M., Bilello, M., Biros, G., Davatzikos, C., 2011. Joint segmentation and deformable registration of brain scans guided by a tumor growth model, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 532–540.

Grigorescu, I., Uus, A., Christiaens, D., Cordero-Grande, L., Hutter, J., Edwards, A.D., Hajnal, J.V., Modat, M., Deprez, M., 2020. Diffusion tensor driven image registration: a deep learning approach, in: International Workshop on Biomedical Image Registration, Springer. pp. 131–140.

De Groot, M., Cremers, L.G., Ikram, M.A., Hofman, A., Krestin, G.P., van der Lugt, A., Niessen, W.J., Vernooij, M.W., 2016. White matter degeneration with aging: longitudinal diffusion MR imaging analysis. Radiology 279, 532–541.

De Groot, M., Ikram, M.A., Akoudad, S., Krestin, G.P., Hofman, A., van der Lugt, A., Niessen, W.J., Vernooij, M.W., 2015. Tract-specific white matter degeneration in aging: the Rotterdam Study. Alzheimer's & Dementia 11, 321–330.

De Groot, M., Verhaaren, B.F., De Boer, R., Klein, S., Hofman, A., van der Lugt, A., Ikram, M.A., Niessen, W.J., Vernooij, M.W., 2013a. Changes in normal-appearing white matter precede development of white matter lesions. Stroke 44, 1037–1042.

De Groot, M., Vernooij, M.W., Klein, S., Ikram, M.A., Vos, F.M., Smith, S.M., Niessen, W.J., Andersson, J.L., 2013b. Improving alignment in tract-based spatial statistics: evaluation and optimization of image registration. Neuroimage 76, 400–411.

Hu, Y., Modat, M., Gibson, E., Ghavami, N., Bonmati, E., Moore, C.M., Emberton, M., Noble, J.A., Barratt, D.C., Vercauteren, T., 2018. Label-driven weakly-supervised learning for multimodal deformable image registration, in: 15th ISBI, IEEE. pp. 1070–1074.

Ikram, M.A., Brusselle, G., Ghanbari, M., Goedegebure, A., Ikram, M.K., Kavousi, M., Kieboom, B.C., Klaver, C.C., de Knegt, R.J., Luik, A.I., et al., 2020. Objectives, design and main findings until 2020 from the rotterdam study. European Journal of Epidemiology , 1–35.

Ikram, M.A., van der Lugt, A., Niessen, W.J., Krestin, G.P., Koudstaal, P.J., Hofman, A., Breteler, M.M., Vernooij, M.W., 2011. The rotterdam scan study: design and update up to 2012. European journal of epidemiology 26, 811–824.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 .

Jaderberg, M., Simonyan, K., Zisserman, A., et al., 2015. Spatial transformer networks, in: Advances in neural information processing systems, pp. 2017–2025.

Keihaninejad, S., Zhang, H., Ryan, N.S., Malone, I.B., Modat, M., Cardoso, M.J., Cash, D.M., Fox, N.C., Ourselin, S., 2013. An unbiased longitudinal analysis framework for tracking white matter changes using diffusion tensor imaging with application to Alzheimer's disease. Neuroimage 72, 153–163.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Klein, S., Pluim, J.P., Staring, M., Viergever, M.A., 2009. Adaptive stochastic gradient descent optimisation for image registration. Int J Comput Vis. 81, 227.

Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P., 2010. Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imag. 29, 196–205.

Koppelmans, V., de Groot, M., de Ruiter, M.B., Boogerd, W., Seynaeve, C., Vernooij, M.W., Niessen, W.J., Schagen, S.B., Breteler, M.M., 2014. Global and focal white matter integrity in breast cancer survivors 20 years after adjuvant chemotherapy. Hum Brain Mapp. 35, 889–899.

Van der Krieke, L., Blaauw, F.J., Emerencia, A.C., Schenk, H.M., Slaets, J.P., Bos, E.H., de Jonge, P., Jeronimus, B.F., 2017. Temporal dynamics of health and well-being: A crowdsourcing approach to momentary assessments and automated generation of personalized feedback. Psychosomatic medicine 79, 213–223.

Landis, J.R., Koch, G.G., 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics , 363–374.

Le Bihan, D., Mangin, J.F., Poupon, C., Clark, C.A., Pappata, S., Molko, N., Chabriat, H., 2001. Diffusion tensor imaging: concepts and applications.

Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine 13, 534–546.
Lebel, C., Beaulieu, C., 2011. Longitudinal development of human brain wiring continues from childhood into adulthood. Journal of Neuroscience 31, 10937–10947.
Leemans, A., Jeurissen, B., Sijbers, J., Jones, D., 2009. ExploreDTI: a graphical toolbox for processing, analyzing, and visualizing diffusion MR data, in: 17th annual meeting of intl soc mag reson med, p. 3537.
Li, B., de Groot, M., Steketee, R.M., Meijboom, R., Smits, M., Vernooij, M.W., Ikram, M.A., Liu, J., Niessen, W.J., Bron, E.E., 2020a. Neuro4Neuro: A neural network approach for neural tract segmentation using large-scale population-based diffusion imaging. NeuroImage , 116993.
Li, B., de Groot, M., Vernooij, M.W., Ikram, M.A., Niessen, W.J., Bron, E.E., 2018. Reproducible white matter tract segmentation using 3D U-Net on a large-scale DTI dataset, in: International Workshop on MLMI, Springer. pp. 205–213.
Li, B., Niessen, W.J., Klein, S., de Groot, M., Ikram, M.A., Vernooij, M.W., Bron, E.E., 2019. A hybrid deep learning framework for integrated segmentation and registration: evaluation on longitudinal white matter tract changes, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 645–653.
Li, B., Niessen, W.J., Klein, S., Ikram, M.A., Vernooij, M.W., Bron, E.E., 2020b. Learning unbiased registration and joint segmentation: evaluation on longitudinal diffusion MRI. arXiv preprint arXiv:2011.01869 .
Maas, A.L., Hannun, A.Y., Ng, A.Y., 2013. Rectifier nonlinearities improve neural network acoustic models, in: Proc. icml, p. 3.
Meijboom, R., Steketee, R., Ham, L., Mantini, D., Bron, E., van der Lugt, A., van Swieten, J., Smits, M., 2019. Exploring quantitative group-wise differentiation of Alzheimer's disease and behavioural variant frontotemporal dementia using tract-specific microstructural white matter and functional connectivity measures at multiple time points. European radiology 29, 5148–5159.
Metz, C., Klein, S., Schaap, M., van Walsum, T., Niessen, W.J., 2011. Nonrigid registration of dynamic medical imaging data using nD+ t B-splines and a groupwise optimization approach. Med Image Anal. 15, 238–249.
Niessen, W.J., 2016. MR brain image analysis in dementia: From quantitative imaging biomarkers to ageing brain models and imaging genetics.
Parisot, S., Wells III, W., Chemouny, S., Duffau, H., Paragios, N., 2014. Concurrent tumor segmentation and registration with uncertainty-based sparse non-uniform graphs. Med Image Anal. 18.
Pathak, D., Girshick, R., Dollár, P., Darrell, T., Hariharan, B., 2017. Learning features by watching objects move, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2701–2710.
Pohl, K.M., Fisher, J., Grimson, W.E.L., Kikinis, R., Wells, W.M., 2006. A Bayesian model for joint segmentation and registration. NeuroImage 31, 228–239.
Postelnicu, G., Zollei, L., Fischl, B., 2008. Combined volumetric and surface registration. IEEE transactions on medical imaging 28, 508–522.
Qin, C., Bai, W., Schlemper, J., Petersen, S.E., Piechnik, S.K., Neubauer, S., Rueckert, D., 2018. Joint learning of motion estimation and segmentation for cardiac mr image sequences, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 472–480.
Reuter, M., Schmansky, N.J., Rosas, H.D., Fischl, B., 2012. Within-subject template estimation for unbiased longitudinal image analysis. Neuroimage 61, 1402–1418.
Rohé, M.M., Datar, M., Heimann, T., Sermesant, M., Pennec, X., 2017. SVF-Net: Learning deformable image registration using shape matching, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 266–274.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on MICCAI, Springer. pp. 234–241.
Sullivan, E.V., Rohlfing, T., Pfefferbaum, A., 2010. Longitudinal study of callosal microstructure in the normal adult aging brain using quantitative DTI fiber tracking. Developmental neuropsychology 35, 233–256.
Vakalopoulou, M., Chassagnon, G., Bus, N., Marini, R., Zacharaki, E.I., Revel, M.P., Paragios, N., 2018. Atlasnet: Multi-atlas non-linear deep networks for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 658–666.
Vlontzos, A., Mikolajczyk, K., 2018. Deep segmentation and registration in X-Ray angiography video. arXiv preprint arXiv:1805.06406 .
Vrooman, H.A., Cocosco, C.A., van der Lijn, F., Stokking, R., Ikram, M.A., Vernooij, M.W., Breteler, M.M., Niessen, W.J., 2007. Multi-spectral brain tissue segmentation using automatically trained k-nearest-neighbor classification. Neuroimage 37, 71–81.
Wasserthal, J., Neher, P., Maier-Hein, K.H., 2018. TractSeg: Fast and accurate white matter tract segmentation. NeuroImage 183, 239–253.
Wyatt, P.P., Noble, J.A., 2003. MAP MRF joint segmentation and registration of medical images. Medical Image Analysis 7, 539–552.
Xu, Z., Niethammer, M., 2019. DeepAtlas: Joint semi-supervised learning of image registration and segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 420–429.
Yendiki, A., Reuter, M., Wilkens, P., Rosas, H.D., Fischl, B., 2016. Joint reconstruction of white-matter pathways from longitudinal diffusion MRI data with anatomical priors. Neuroimage 127, 277–286.
Yezzi, A., Zöllei, L., Kapur, T., 2003. A variational framework for integrating segmentation and registration through active contours. Medical image analysis 7, 171–185.
Zhang, H., Avants, B.B., Yushkevich, P.A., Woo, J.H., Wang, S., McCluskey, L.F., Elman, L.B., Melhem, E.R., Gee, J.C., 2007. High-dimensional spatial normalization of diffusion tensor images improves the detection of white matter differences: an example study using amyotrophic lateral sclerosis. IEEE transactions on medical imaging 26, 1585–1597.
Zhu, W., Myronenko, A., Xu, Z., Li, W., Roth, H., Huang, Y., Milletari, F., Xu, D., 2020. Neurreg: Neural registration and its application to image segmentation, in: The IEEE Winter Conference on Applications of Computer Vision, pp. 3617–3626.

Supplementary Material

The network architectures of $\mathcal{F}_\Theta$ and $\mathcal{G}_\Psi$ are illustrated in Figure 8 and Figure 9, respectively. The image size of the input and output for $\mathcal{F}_\Theta$ used in the present study were $(112 \times 208 \times 112 \times 6)$ and $(112 \times 208 \times 112 \times 3)$ voxels. That of $\mathcal{G}_\Psi$ were $(112 \times 208 \times 112 \times 2)$ and $(112 \times 208 \times 112 \times 3)$ voxels.

The procedure of simultaneous optimization is summarized with the pseudo code in Algorithm 1.

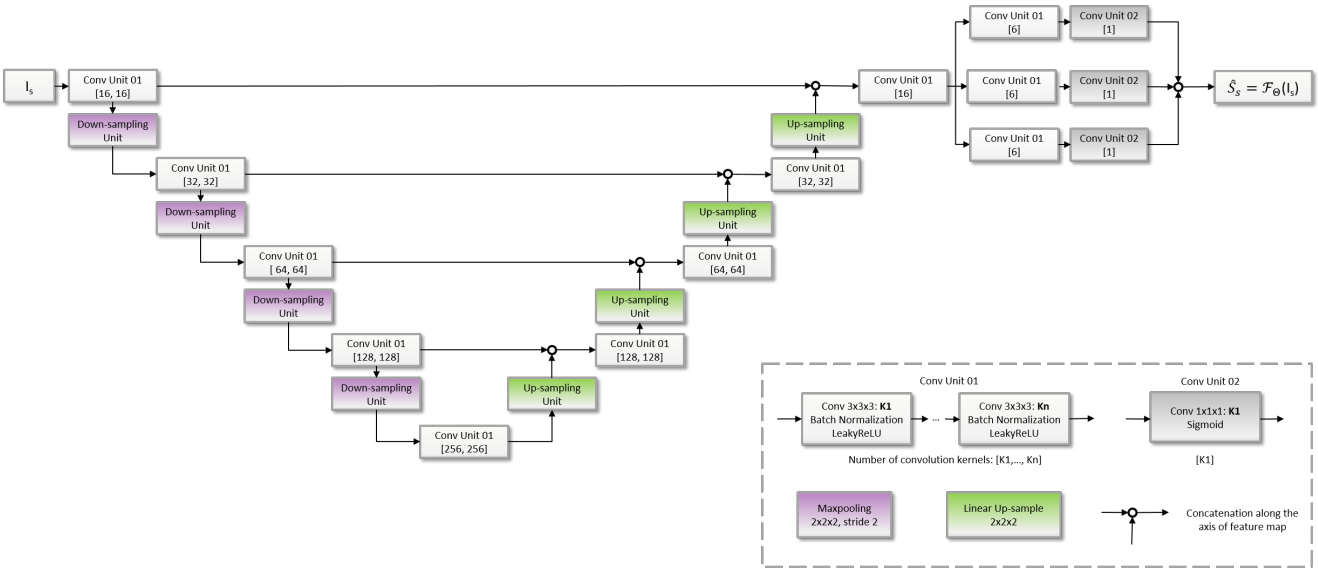


Figure 8: The network architectures of $\mathcal{F}_\Theta$ module used in the present study. The number of convolution kernels used in Conv Unit 01 is denoted as [K1, ..., Kn], and that for Conv Unit 02 as [K1]

.

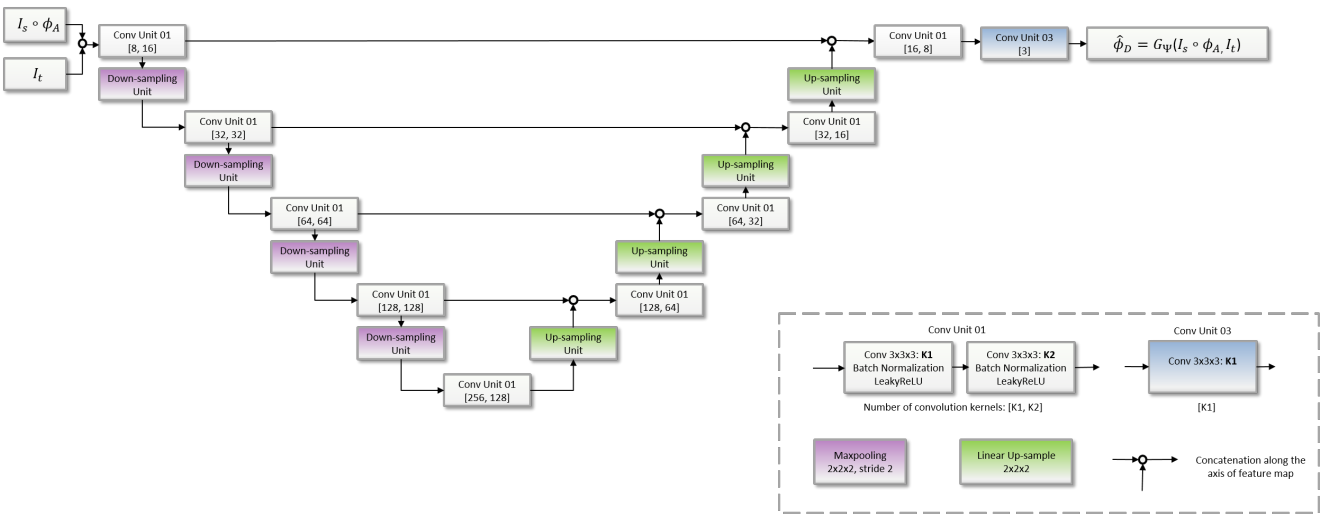


Figure 9: The network architectures of $\mathcal{G}_\Psi$ module used in the present study. The target image (I_t) and affine-aligned source image ($I_s \circ \boldsymbol{\phi}_A$) are used as the input to predict non-rigid deformation ($\boldsymbol{\phi}_D$), which can subsequently lead to a composite displacement field ($\boldsymbol{\phi}$) as shown in Figure 1. The number of convolution kernels used in Conv Unit 01 is denoted as [K1, K2], and that for Conv Unit 03 as [K1].

Algorithm 1: Simultaneous optimization

Input: $\{\mathcal{S}_s^i,\ \mathcal{S}_t^i,\ I_s^i, I_t^i, \boldsymbol{\phi}_A^i\}_{i=1}^N$
Parameters: $\boldsymbol{\Theta}, \boldsymbol{\Psi}$
Output: $\{\hat{\mathcal{S}}_s^i,\ \hat{\mathcal{S}}_s^i \circ \hat{\boldsymbol{\phi}}^i,\ I_s^i \circ \hat{\boldsymbol{\phi}}^i,\ \hat{\boldsymbol{\phi}}^i,\ \hat{\boldsymbol{\phi}}^i{}_D\}_{i=1}^N$
Initialization:
lr, decay_factor `// Initial and decay ratio of learning rate`
$\boldsymbol{\Theta}, \boldsymbol{\Psi} \sim$ GlorotUniform `// Kernel initialization`
for number of training iterations **do**
  shuffle(**Input**)
  for $i = 0$ **to** N **do**
    $\hat{\mathcal{S}}_s^i = \mathcal{F}_\Theta(I_s^i)$ `// Segmentation, Eq.1`
    $\hat{\boldsymbol{\phi}}_D^i = \mathcal{G}_\Psi(I_t^i,\ I_s^i \circ \boldsymbol{\phi}_A^i)$ `// Registration, Eq.4`
    $\hat{\boldsymbol{\phi}}^i = \boldsymbol{\phi}_A^i \circ \hat{\boldsymbol{\phi}}_D^i$ `// Transform composition`
    $\hat{I}_t^i = I_s^i \circ \hat{\boldsymbol{\phi}}^i$ `// Image warp`
    `/* Dependency on both tasks */`
    $\hat{\mathcal{S}}_t^i = \hat{\mathcal{S}}_s^i \circ \hat{\boldsymbol{\phi}}^i$ `// Segmentation warp`
    $\mathcal{L}^i = \mathcal{L}_{seg}(\mathcal{S}_s^i,\ \hat{\mathcal{S}}_s^i) + \alpha\mathcal{L}_{reg}(I_t^i,\ \hat{I}_t^i) + \beta\mathcal{L}_{def}(\hat{\boldsymbol{\phi}}_D^i)$
     $+\gamma\mathcal{L}_{com}(\mathcal{S}_t^i,\ \hat{\mathcal{S}}_t^i)$ `// Segis-Net objective, Eq.9`
    `/* Simultaneous optimization of parameters */`
    $\boldsymbol{\Theta}, \boldsymbol{\Psi} \leftarrow$ **Adam** $(\mathcal{L}^i, \text{lr}, \boldsymbol{\Theta}, \boldsymbol{\Psi})$
  end
  `/* Custom condition of learning rate decay */`
  if learning rate decay **then**
    $lr \leftarrow lr \times$ decay_factor
  end
  return $\boldsymbol{\Theta}, \boldsymbol{\Psi}$ `// Return parameters per epoch`
end

		Classical	CNN	Segis-Net
Figure 3	CGC_L	-	0.76 ± 0.06	0.76 ± 0.06
	CGC_R	-	0.76 ± 0.06	0.76 ± 0.06
	CGH_L	-	0.76 ± 0.07	0.76 ± 0.07
	CGH_R	-	0.77 ± 0.09	0.76 ± 0.09
	FMA	-	0.76 ± 0.05	0.76 ± 0.05
	FMI	-	0.68 ± 0.09	0.67 ± 0.09
Figure 4	CGC_L	0.73 ± 0.10	0.71 ± 0.07	0.73 ± 0.08
	CGC_R	0.73 ± 0.10	0.69 ± 0.06	0.73 ± 0.07
	CGH_L	0.75 ± 0.11	0.75 ± 0.10	0.77 ± 0.09
	CGH_R	0.75 ± 0.11	0.75 ± 0.10	0.76 ± 0.10
	FMA	0.72 ± 0.10	0.71 ± 0.08	0.74 ± 0.09
	FMI	0.74 ± 0.08	0.75 ± 0.08	0.76 ± 0.08
Figure 5	CGC_L	0.68 ± 0.06	0.82 ± 0.04	0.83 ± 0.04
	CGC_R	0.68 ± 0.06	0.82 ± 0.06	0.83 ± 0.04
	CGH_L	0.66 ± 0.08	0.81 ± 0.05	0.82 ± 0.05
	CGH_R	0.66 ± 0.09	0.81 ± 0.05	0.81 ± 0.05
	FMA	0.69 ± 0.06	0.84 ± 0.03	0.87 ± 0.02
	FMI	0.57 ± 0.09	0.77 ± 0.07	0.81 ± 0.05
Figure 6(a)	CGC_L	0.68 ± 0.07	0.82 ± 0.04	0.83 ± 0.03
	CGC_R	0.68 ± 0.07	0.81 ± 0.04	0.82 ± 0.04
	CGH_L	0.65 ± 0.10	0.77 ± 0.07	0.79 ± 0.06
	CGH_R	0.66 ± 0.09	0.78 ± 0.05	0.79 ± 0.05
	FMA	0.72 ± 0.06	0.85 ± 0.03	0.87 ± 0.03
	FMI	0.64 ± 0.08	0.79 ± 0.08	0.82 ± 0.06
Figure 6(b)	CGC_L	11 ± 8.8%	5.1 ± 3.6%	4.8 ± 4.1%
	CGC_R	11 ± 8.9%	5.4 ± 4.5%	4.5 ± 3.9%
	CGH_L	11 ± 9.8%	7.9 ± 7.2%	7.3 ± 5.6%
	CGH_R	9.8 ± 7.3%	7.6 ± 7.5%	7.0 ± 5.8%
	FMA	6.6 ± 5.9%	4.9 ± 3.6%	3.4 ± 2.6%
	FMI	7.9 ± 6.5%	8.5 ± 8.6%	6.0 ± 6.2%
Figure 6(c)	CGC_L	4.7 ± 4.0%	3.0 ± 2.2%	2.8 ± 2.2%
	CGC_R	4.6 ± 3.4%	3.3 ± 2.6%	3.4 ± 2.4%
	CGH_L	7.3 ± 5.5%	4.7 ± 3.9%	5.2 ± 4.0%
	CGH_R	6.3 ± 4.7%	4.3 ± 3.8%	4.6 ± 4.1%
	FMA	1.9 ± 1.5%	1.9 ± 1.4%	2.0 ± 1.6%
	FMI	2.7 ± 1.9%	2.7 ± 2.1%	2.3 ± 1.7%
Figure 6(d)	CGC_L	1.9 ± 1.8%	1.7 ± 1.5%	1.6 ± 1.4%
	CGC_R	1.5 ± 1.2%	1.5 ± 1.2%	1.4 ± 1.1%
	CGH_L	3.1 ± 4.9%	2.1 ± 1.6%	2.2 ± 1.6%
	CGH_R	2.3 ± 1.9%	1.7 ± 1.5%	1.6 ± 1.4%
	FMA	1.4 ± 1.1%	1.1 ± 1.1%	1.1 ± 0.9%
	FMI	1.1 ± 0.8%	1.1 ± 0.9%	1.0 ± 0.8%

Table 3: Results overview for Figure 3-5. Figure 3, Segmentation accuracy; Figure 4, Spatial correlation (SC) similarity; Figure 5, Spatio-temporal consistency of segmentation (STCS); Figure 6 (a), Segmentation reproducibility; Figure 6 (b), Volume Reproducibility; Figure 6 (c), FA reproducibility; Figure 6 (d), MD reproducibility.

Data availability:

The datasets analyzed during the current study are not publicly available. Due to the sensitive nature of the data used in this study, participants were assured raw data would remain confidential and would not be shared.

Code availability:

The third party code (mainly from Elastix software) has been stated and cited appropriately in the manuscript. The code for the proposed method will be made available upon acceptance of the manuscript.