frontiers in Neurology

ORIGINAL RESEARCH
published: 14 July 2021
doi: 10.3389/fneur.2021.653375

Multi-Modal Segmentation of 3D Brain Scans Using Neural Networks

Jonathan Zopes, Moritz Platscher, Silvio Paganucci and Christian Federau*

Institute for Biomedical Engineering, ETH Zürich, Zurich, Switzerland

Anatomical segmentation of brain scans is highly relevant for diagnostics and neuroradiology research. Conventionally, segmentation is performed on T_1-weighted MRI scans, due to the strong soft-tissue contrast. In this work, we report on a comparative study of automated, learning-based brain segmentation on various other contrasts of MRI and also computed tomography (CT) scans and investigate the anatomical soft-tissue information contained in these imaging modalities. A large database of in total 853 MRI/CT brain scans enables us to train convolutional neural networks (CNNs) for segmentation. We benchmark the CNN performance on four different imaging modalities and 27 anatomical substructures. For each modality we train a separate CNN based on a common architecture. We find average Dice scores of 86.7 ± 4.1% (T_1-weighted MRI), 81.9 ± 6.7% (fluid-attenuated inversion recovery MRI), 80.8 ± 6.6% (diffusion-weighted MRI) and 80.7 ± 8.2% (CT), respectively. The performance is assessed relative to labels obtained using the widely-adopted FreeSurfer software package. The segmentation pipeline uses dropout sampling to identify corrupted input scans or low-quality segmentations. Full segmentation of 3D volumes with more than 2 million voxels requires <1 s of processing time on a graphical processing unit.

Keywords: brain imaging (CT and MRI), anatomical segmentation, multi-modal, convolutional neural networks, dropout sampling

OPEN ACCESS

Edited by:
Volker Rasche,
University of Ulm, Germany

Reviewed by:
Alessia Sarica,
University of Magna Graecia, Italy
Yuankai Huo,
Vanderbilt University, United States

****Correspondence:***
Jonathan Zopes
zopes@biomed.ee.ethz.ch

Specialty section:
This article was submitted to Applied Neuroimaging, a section of the journal Frontiers in Neurology

Received: *14 January 2021*
Accepted: *17 June 2021*
Published: *14 July 2021*

Citation:
Zopes J, Platscher M, Paganucci S and Federau C (2021) Multi-Modal Segmentation of 3D Brain Scans Using Neural Networks. Front. Neurol. 12:653375. doi: 10.3389/fneur.2021.653375

1. INTRODUCTION

Anatomical segmentation of magnetic resonance imaging (MRI) or computed tomography (CT) scans is important for clinical diagnostics and scientific research. In particular, quantitative volumetric measures of anatomical structures can be derived from accurate segmentation labels, which can then be used to identify and monitor the progression of degenerative diseases, such as Alzheimer's disease, which is characterized by atrophy of the hippocampus and the medial temporal lobe (1), Huntington disease, which results in athrophy of the striatum (2), and frontotemporal lobar degeneration, which causes atrophy of the frontal and temporal lobes (3).

Manual brain segmentation, however, requires expert knowledge of radiologists, is extremely tedious and time consuming, and is therefore limited to small datasets or simply not available. An alternative approach is to automatize segmentation, which sparked the development of various segmentation software packages. In brain imaging these include e.g., FreeSurfer (4), BrainSuite (5), FSL (6), and ANTS (7). These tools apply a set of complex transformations and thresholding procedures to the input volume (8) and are typically tailored toward T_1-weighted scans. As a consequence, direct segmentation of highly relevant MRI contrasts like FLAIR (fluid-attenuated inversion recovery) or DWI (diffusion-weighted imaging) remain unsupported. The same statement is true for CT volumes.

Although the recent literature contains attempts to automatize segmentation on FLAIR (9–11), DWI (12, 13), or CT volumes (14, 15), a comparative study on the achievable segmentation quality on the different imaging modalities is, to the best of our knowledge, still outstanding. We attribute this in part to the lack of structured databases that contain several paired imaging modalities for the same patient. Further, the limited flexibility of conventional segmentation tools, that require careful fine-tuning of parameters, might be a second contributing factor.

In our work, we present a broad study on the segmentation performance achievable on T_1-weighted MRI, FLAIR, DWI, and CT scans for a wide range of 27 anatomical classes. The analysis is based on two large databases with in total 853 MRI/CT scans and with several imaging modalities per patient. To implement a flexible segmentation pipeline, which can be quickly adapted to the different imaging modalities, we leverage the flexibility and performance of convolutional neural networks (CNNs).

The recent success of CNNs in computer vision tasks (16) provided a strong impetus for applying CNNs in brain segmentation (17–21). CNNs can be rapidly adjusted to segment on a given contrast, merely by adjusting the weights of the neural network via training. This eliminates the need for additional human fine-tuning and enables us to benchmark the segmentation performance for a common network architecture (see **Figure 1**). Further, CNN segmentation tools recently exceeded conventional processing tools in performance (23, 24) and due to their efficient implementation on graphical processing units (GPUs), achieve full segmentation of 3D volumes almost in real-time. This is orders of magnitude faster than with conventional methods (25).

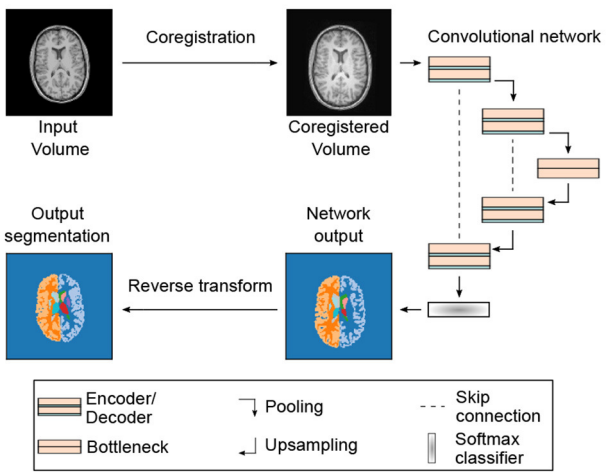

FIGURE 1 | Segmentation pipeline and neural network architecture: 3D MRI or CT input volumes are coregistered to a reference volume with an affine transformation. By proper resampling the pixel dimensions of the registered volume are adjusted to the input shape of the neural network. In addition, the pixel intensities are normalized to the interval $I = [0, 100]$. Neural networks are based on the U-Net architecture (22) with 3D convolutions in the encoder and decoder blocks. Each encoder and decoder block contains two consecutive convolution, batch normalization and rectified linear activation operations. The encoder and decoder blocks are followed by a dropout layer (see **Table 1** and text for details). The softmax output of the network is converted into a segmentation map with 28 labels (including background). The segmentation map is finally registered back to the input volume using the inverse affine transformation of the initial coregistration.

2. METHODS

2.1. Segmentation Pipeline

In **Figure 1**, we show a schematic of our segmentation pipeline. The input MRI/CT volume is first coregistered to a reference volume with an affine transformation. The reference volume was selected from our data set by optimizing signal-to-noise ratio and by ensuring the absence of imaging artifacts. For coregistration we use the registration tool elastix 4.8 (26). The coregistered volume is resampled using spline interpolation to match the input dimensions of the segmentation CNN. The coregistration procedure increases the performance of the segmentation network and further allows for arbitrarily shaped input volumes due to resampling.

For segmentation we use a fully-convolutional neural network (F-CNN) based on the U-Net architecture (22). A schematic of the network architecture is displayed in **Figure 1** and further details on network training and parameters are discussed in the subsequent sections. The network outputs a softmax quasi-probability map $P_s(x)$ for each segmentation class $s \in \mathcal{S}$. Each individual map has the same dimension as the input image. The list of segmented classes $\mathcal{S}$ follows reference (23) and comprises in total 27 structures. All segmented classes are listed in **Supplementary Table 1**.

The softmax output P of the network is converted to a hard segmentation mask S using the *arg* max function:

$$S(x) = \arg\max_s P_s(x). \quad (1)$$

Subsequently, the hard segmentation mask S is registered back to the input volume. For this purpose the initial affine coregistration transformation is inverted. After applying the inverse transformation the mask is resampled using nearest-neighbor sampling with the dimensions defined by the initial input volume.

2.2. Neural Networks and Training

As mentioned before, we use a U-Net based network architecture for segmentation. Following the findings in (27), we make only minor modifications to the original implementation in (22, 28). The network consists of an encoder-decoder structure with skip connections (see **Figure 1**). In each encoder and decoder block we apply two repetitions of convolutional layers, with kernel size $K = (3, 3, 3)$. Each convolutional layer is followed by batch normalization and non-linear activation with rectified linear units. The initial number of feature maps, after the first convolutional layer, was fixed to $F = 32$ for all models and after each encoder (decoder) block the number of feature maps is doubled (halved).

We use dropout layers after the encoders and decoders to prevent overfitting and to perform dropout sampling for uncertainty quantification (see section 3.3). Max pooling after

TABLE 1 | Parameters of training and test datasets and segmentation scores on all four imaging modalities using all available training samples.

Modality	N_{train}	N_{val}	N_{test}	Volume shape	Average $\mathcal{D}_A$	Weighted $\mathcal{D}_V$	$\mathcal{C}(\mathcal{D}_A, CV)$	ASSD [mm]
MPRAGE	465	51	6	[128, 128, 128]	(86.7 ± 4.1)%	(88.7 ± 0.8)%	−0.91	(0.67 ± 0.23)
FLAIR	107	11	6	[128, 128, 128]	(81.9 ± 6.7)%	(83.7 ± 1.7)%	−0.87	(0.82 ± 0.26)
DWI	142	15	6	[160, 160, 32]	(80.8 ± 6.6)%	(82.0 ± 5.9)%	−0.87	(0.70 ± 0.27)
CT	34	5	5	[96, 128, 96]	(80.7 ± 8.2)%	(77.9 ± 5.2)%	−0.97	(1.12 ± 0.40)

Each dataset contains N_{train} training volumes, N_{val} validation volumes and N_{test} test samples. The voxel dimension of all volumes in the training and test dataset is fixed to the reported volume shape. Average and weighted Dice scores are reported according to Equations (4) and (5), respectively. The Pearson correlation coefficient between average Dice score and the uncertainty metric CV, obtained from dropout sampling, is listed for each imaging modality. The average symmetric surface distance (ASSD) in millimeters is reported as mean ± s.d., where the mean and standard deviation (s.d.) are computed over all samples from the test dataset and averaged over all anatomical structures.

each encoder block halves the feature map dimensions. Likewise, upsampling with transpose convolutions after the decoder blocks doubles the feature map dimensions and finally restores the initial dimensions at the output.

The number of max pooling operations defines the depth D of the U-Net architecture, which we fixed to $D = 4$ for all trained models. The bottleneck block restricts information flow from encoder to decoder and consists of two convolutional layers, each followed by batch normalization and rectified linear activation. In contrast to the encoder and decoder blocks, we do not use dropout layers in the bottleneck block (29).

CNNs are implemented in tensorflow 2.2.0 (30) and training is performed on a single GPU (Nvidia Titan RTX 24GB). Due to memory constraints the input brain volumes are limited to about 2 million voxels, which we typically distribute evenly among the imaging dimensions. The input dimensions for each network are listed in **Table 1**. We train the network using the Adam optimizer with initial learning rates of 0.001. During training, we apply a set of random transformations, e.g., translations, rotations, or cropping, to the volumes for data augmentation. As the loss function, we use a combination of the Dice score, summed over all class labels, and the categorical cross-entropy function:

$$\mathcal{L} = -\sum_{s\in\mathcal{S}} \left(\frac{2\sum_x P_s(x)T_s(x)}{\sum_x P_s(x) + T_s(x)} - \sum_x T_s(x)\log(P_s(x)) \right). \quad (2)$$

Here, $P_s(x)$ is the softmax output of the network at voxel position x and $T_s(x)$ is the ground truth at the same position. We use the categorical cross-entropy loss to alleviate convergence problems when using solely the Dice loss (27). In principle, the influence of cross-entropy and Dice loss can be additionally weighted, but we found little influence on performance and therefore omit additional weighting. We train the CNNs for up to 400 epochs and abort the training process, if the validation loss does not improve for 100 epochs. The performance of the models is evaluated on separate test datasets.

2.3. MRI/CT Databases: Preprocessing and Label Generation

For training of the CNNs we use two large database of MRI and CT brain scans acquired on healthy patients. The acquisition parameters are listed in **Supplementary Table 3**. The first database contains 530 scans of healthy patients for which MPRAGE, FLAIR and DWI scans are available. The MPRAGE contrast was used to generate training labels using FreeSurfer 6.0 (4). The FreeSurfer labels were mapped to 27 segmentation classes using the mapping strategy described in (25). The resulting labels are in the following considered the ground truth and subsequently coregistered to the corresponding FLAIR and DWI scans.

After coregistration we manually checked for a proper alignment of the segmentation masks to the FLAIR or DWI volume. Out of the initial database with 530 cases, we select 124 (FLAIR) and 163 (DWI) volumes for training, validation and testing. We thus removed a large fraction of cases from the database. This is due to the limited fidelity of the coregistration process and because we observe that a smaller, yet higher quality database leads to better segmentation performance. For the MPRAGE contrast no further coregistration was necessary and we therefore manually selected a large fraction of 522 out of 530 volumes, with high-quality FreeSurfer segmentations, for training.

The second database contains 60 healthy patients for which both MPRAGE and CT brain scans are available. Again we coregister MPRAGE and CT volumes to each other and use FreeSurfer on the MPRAGE scans to obtain training labels for both imaging modalities. By manually checking the alignment of the segmentation mask to the CT volume we selected 41 volumes for training, validation, and testing. Here, we also manually corrected minor coregistration errors to keep most of the available samples for training.

In order to evaluate the achievable segmentation performance on the different imaging modalities, we perform two separate studies, which are presented in the results section. In the first study, we use all available samples on each modality for training, validation and testing and in the second study we remove scans from the larger databases to ensure equally-sized training datasets. The number of resulting training, validation and test samples for all modalities is listed in **Tables 1**, **2** for both studies. In order to ensure comparability among modalities, we decided to use the same patients/volumes for testing on the MRI modalities. This constrained the size of the test dataset to the intersection of the three MRI datasets, which includes in total six patients/volumes.

2.4. Segmentation Performance Metrics

We use several metrics to quantify the segmentation performance of our models. As an overlap-based metric, we use the Dice score $\mathcal{D}_s$, associated with the anatomical structure $s \in \mathcal{S}$, as the

TABLE 2 | Parameters of training and test datasets and segmentation scores on all four imaging modalities using approximately equally-sized training sets.

Modality	N_{train}	N_{val}	N_{test}	Volume shape	Average $\mathcal{D}_A$	Weighted $\mathcal{D}_V$	ASSD [mm]
MPRAGE	39	3	6	[128, 128, 128]	(83.6 ± 7.2)%	(86.6 ± 2.0) %	(0.81 ± 0.31)
FLAIR	39	3	6	[128, 128, 128]	(80.0 ± 8.8)%	(82.6 ± 2.1) %	(0.94 ± 0.38)
DWI	39	3	6	[160, 160, 32]	(78.9 ± 7.8)%	(80.7 ± 5.8) %	(0.71 ± 0.28)
CT[a]	34	5	5	[96, 128, 96]	(80.7 ± 8.2)%	(77.9 ± 5.2) %	(1.12 ± 0.40)

*Each dataset contains N_{train} training volumes, N_{val} validation volumes, and N_{test} test samples. The voxel dimension of all volumes in the training and test dataset is fixed to the reported volume shape. Average and weighted Dice scores are reported according to Equations (4) and (5), respectively. The average symmetric surface distance (ASSD) in millimeters is reported as mean±s.d., where the mean and standard deviation (s.d.) are computed over all samples from the test dataset and averaged over all anatomical structures. [a]Same model as reported in **Table 1**.*

performance metric:

$$\mathcal{D}_s = \frac{2\sum_x S_s(x)T_s(x)}{\sum_x S_s(x) + T_s(x)}. \tag{3}$$

Here, $S_s(x)$ is the hard segmentation mask, given in Equation (1), in one-hot encoding format. To compare the overall performance, we introduce two additional metrics: The average Dice score:

$$\mathcal{D}_A = \sum_{s\in\mathcal{S}} \mathcal{D}_s, \tag{4}$$

and a volume-weighted Dice score:

$$\mathcal{D}_V = \frac{1}{V}\sum_{s\in\mathcal{S}} \mathcal{V}_s\mathcal{D}_s. \tag{5}$$

Here, $\mathcal{V}_s$ is the volume of the structure s and V is the total volume of all anatomical structures $V = \sum_{s\in\mathcal{S}} \mathcal{V}_s$. The background label is not included in the average and the volume-weighted Dice score and the volumes are computed from the segmentation masks.

In addition to the overlap-based Dice similarity metric, we also report the average symmetric surface distance (ASSD) in millimeters:

$$ASSD(A,B) = \frac{1}{|A|+|B|}\left(\sum_{a\in A} D_B(a) + \sum_{b\in B} D_A(b)\right). \tag{6}$$

Here, A and B are the surfaces of ground-truth and predicted anatomical features, respectively. $D_B(a)$ is the minimal distance between surface B and a given surface voxel $a \in A$ and $D_A(b)$ is the minimal distance between surface A and a given surface voxel $b \in B$. Further, $|A|$ and $|B|$ are the number of surface voxels.

We compute the ASSD, for each anatomical structure, using the hard segmentation mask $S_s(x)$ and the ground truth mask $T_s(x)$ with the python module Scipy (31). Apart from the individual ASSDs for each anatomical structure, we also report the average value of the ASSD to reduce the metric into a single quantity.

2.5. Uncertainty Quantification

A common challenge for automatic segmentation tools is uncertainty quantification or quality control of the segmentation output. Low quality segmentation can occur, for example, due to corrupted input volumes, acquisition artifacts, unrecognized pathologies, or in general due to input volumes outside the training distribution. The incorporation of a direct quality control method, into the segmentation process, is therefore highly desirable.

The softmax output of neural networks, however, does not directly provide credible information on the certainty associated with the assigned labels (32). Instead the authors of (32) proposed to use the dropout layers of the network during prediction to make the network output stochastic. By switching some nodes off at random, we can generate a set of N Monte Carlo (MC) samples $P_s^1, \ldots, P_s^N$ from the network output. The distribution of the MC samples can subsequently be used to gauge the certainty of the assigned labels. Recently, this approach has been successfully applied to brain segmentation on T_1-weighted MRI scans in (29) and we follow their methodology to equip our segmentation pipeline with a credibility metric.

To integrate dropout sampling into our segmentation pipeline we keep the dropout layers of the networks active after training. We generate $N = 15$ MC segmentation samples using the, now stochastic, output of the network. The dropout rate is here fixed to $r = 0.2$ for all neural networks. The final segmentation map is obtained by adding the softmax outputs of all MC samples and then applying the argmax function:

$$S(x) = \arg\max_s \sum_{i=1}^{N} P_s^i(x). \tag{7}$$

To gauge the quality of the segmentation, we use the coefficient of variation CV_s of anatomical volumes over the MC samples. This metric was introduced in (29) and reads:

$$CV_s = \frac{\sigma_s}{\mu_s}. \tag{8}$$

Here, σ_s is the variance of the anatomical volumes between MC samples and μ_s is the mean volume. To reduce the uncertainty measure to a single quantity CV we additionally average the coefficient of variation over all segmented structures:

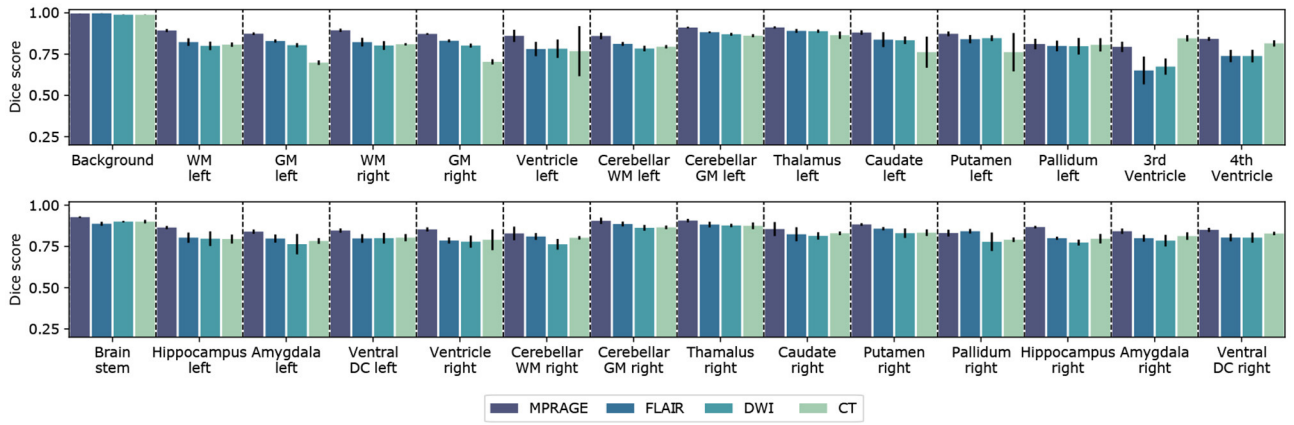

FIGURE 2 | Segmentation performance for different imaging modalities. Barplot of the Dice score for all 27 segmented anatomical structures and including background. For each imaging modality a separate neural network was trained and evaluated. Error bars extend from the lower to upper quartile values of the data. The Dice scores were computed from the test datasets for the corresponding imaging modality, which included 6 (MPRAGE), 6 (DWI), 6 (FLAIR), and 5 (CT) samples, respectively. All parameters of the trained CNNs are summarized in **Table 1**.

$$CV = \sum_{s \in \mathcal{S}} CV_s. \tag{9}$$

3. RESULTS

3.1. Segmentation Performance

3.1.1. Modality-Dependent Segmentation Performance Using All Training Samples

In **Figure 2**, we compare the performance of the segmentation networks on the different imaging modalities for all 27 labeled structures and the background. The reported Dice scores represent the mean over all scans from the test set and the error bars extend from the lower to the upper quartile of values. We further collect all resulting metrics for the different imaging modalities in **Table 1**. This table also includes the ASSD, averaged over all anatomical structures.

We find that the best segmentation results are obtained for T_1-weighted, MPRAGE scans for almost all investigated anatomical structures. This is also expressed by the best average Dice score $\mathcal{D}_A(\text{MPRAGE}) = (86.7 \pm 4.1)\%$ and the best volume-weighted Dice score $\mathcal{D}_V(\text{MPRAGE}) = (88.7 \pm 0.8)\%$. Second-best performance is achieved on the FLAIR contrast. Here, the average Dice score is $\mathcal{D}_A(\text{FLAIR}) = (81.9 \pm 6.7)\%$ and the volume-weighted Dice score is $\mathcal{D}_V(\text{FLAIR}) = (83.7 \pm 1.7)\%$. However, the difference to the performance on the DWI contrast with $\mathcal{D}_A(\text{DWI}) = (80.8 \pm 6.6)\%$ and $\mathcal{D}_V(\text{DWI}) = (82.0 \pm 5.9)\%$ is small.

For CT scans, we find that the segmentation performance is strongly structure-dependent: The low signal contrast between gray and white matter limits to some extent the accuracy of the segmentation, especially of the gray matter regions. At the same time, the segmentation of structures like e.g. ventricles, the putamen or the hippocampus can be performed with high accuracy. We find an average Dice score $\mathcal{D}_A(\text{CT}) = (80.7 \pm 8.2)\%$ on the CT dataset and the volume-weighted Dice score is $\mathcal{D}_V(\text{CT}) = (77.9 \pm 5.2)\%$. The significant reduction in the volume-weighted score is due to the large volume fraction of gray and white matter.

In terms of the surface distances, we find the lowest ASSD, averaged over all anatomical structures, for the MPRAGE scans with 0.67 ± 0.23 mm. The ASSD for the FLAIR scans is slightly higher with 0.82 ± 0.26 mm and exceeds the value for the DWI scans with 0.70 ± 0.27 mm. Again, the lowest performance is found for the CT scans with an ASSD of 1.12 ± 0.40 mm. In general, both similarity and distance-based metrics lead to a consistent quality assessment of the segmentation performance.

3.1.2. Modality-Dependent Segmentation Performance Using Equally-Sized Training Sets

In the previous section, we analyzed the achievable segmentation performance for the different imaging modalities by using all available scans in our database. As a consequence, the number of training samples N_{train} is significantly imbalanced same for the different modalities (see **Table 1**). In order to compare segmentation performance under equally-sized training data sets, we retrained all MRI-related models with a common number of samples ($N_{\text{train}} = 39$, $N_{\text{val}} = 3$, $N_{\text{test}} = 6$). Consequently, for the MPRAGE, FLAIR and DWI scans the number of training samples is reduced by 426, 68, 103 samples, respectively. We use scans from the same subjects out of the first database to make the comparison as fair as possible. We compare the segmentation performance to the existing CT model, which has a similar number of training samples s($N_{\text{train}} = 34$).

In **Table 2**, we list the resulting average and volume-weighted Dice scores of the new models on the test sets. We observe, that due to the reduction of training samples the model performance is consistently reduced by approximately $1 - 2\%$. Nevertheless, the ranking of model performance on different modalities remains as in the previous section.

3.1.3. Dependence of Segmentation Performance on the Number of Training Samples

To further quantify the relation between model performance and the number of available training samples, we trained in total six networks, with varying numbers of training samples $N_{\text{train}} \in \{4, 8, 20, 258, 465\}$, on the MPRAGE contrast. The training samples were chosen randomly from the training database. In **Figure 3**, we show the average Dice score on the test set as a function of the number of training samples ($N_{\text{test}} = 6$ for all models). We find that the test score improves monotonically with the number of available training samples and increases from $(79.0 \pm 8.7\%)$ for the smallest training dataset ($N_{\text{train}} = 4$) to $(86.7 \pm 4.1\%)$ for the largest training dataset ($N_{\text{train}} = 489$). When training with $N_{\text{train}} > 100$, the gain in performance levels off significantly, but does not reach saturation. All models were trained with identical hyperparameters, as described in the Methods section.

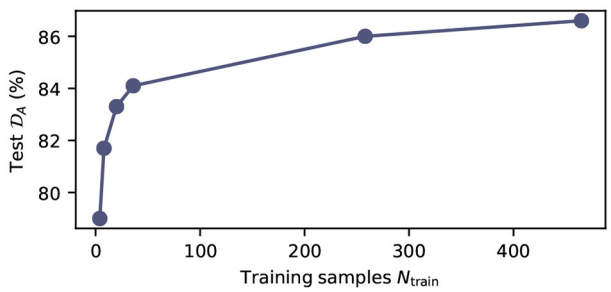

FIGURE 3 | Model performance as function of the number of training samples. Average test Dice score $\mathcal{D}_A$ on the MPRAGE dataset ($N_{\text{test}} = 6$) as function of training samples N_{train}. All models were trained with the same hyperparameters as summarized in **Supplementary Table 3**. The exact number of training samples for the plotted data points is $N_{\text{train}} \in \{4, 8, 20, 258, 465\}$.

3.2. Example Segmentations

In **Figure 4**, we show exemplary input slices, ground-truth labels and the prediction of our segmentation networks for each of the four imaging modalities in the axial view. For the MPRAGE, DWI and FLAIR modalities the segmentation was performed on the same patient and approximately the same slice location is displayed. Exact overlapping of slices is not possible, because the scans are not coregistered to each other. For the CT scan a separate patient was selected from the test dataset of the second database. All predictions are taken from models, which were trained with all available training samples.

The example segmentation clearly show that the gray and white matter boundaries are captured best on the T_1-weighted MPRAGE contrast. Here, even fine structures are properly distinguished. On the DWI and FLAIR contrast gray and white matter are segmented with lower level of detail and with lower fidelity. Due to the significantly reduced signal contrast, the gray and white matter segmentation on the CT scans displays a further reduction in performance. In terms of anatomical structures other than gray and white matter, the CT segmentation provides excellent results. This is especially the case for the ventricles, which are segmented more precisely than on the FLAIR and DWI scans.

3.3. Uncertainty Quantification

In **Figure 5**, we show the relationship between uncertainty metric CV and the average Dice score $\mathcal{D}_A$, derived from the

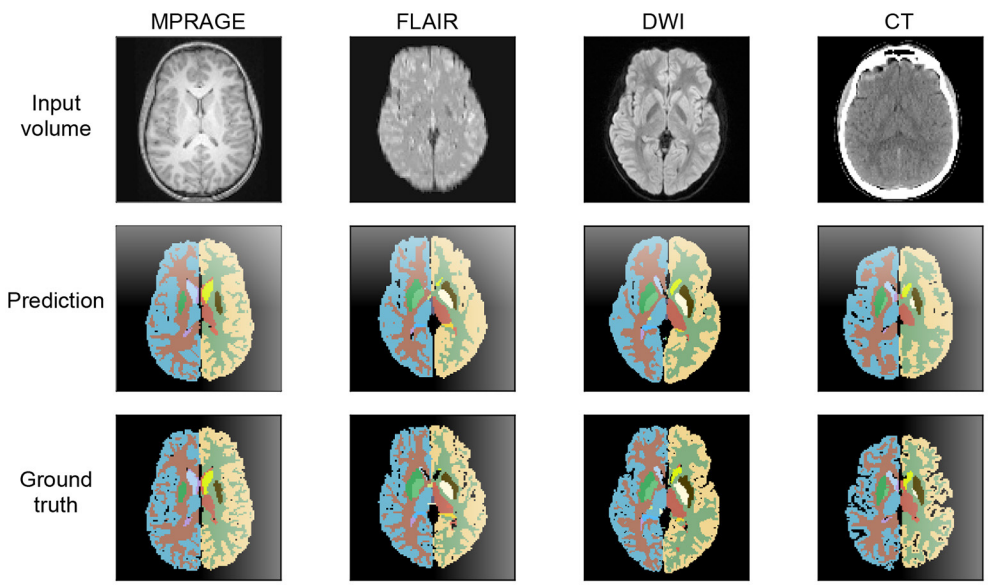

FIGURE 4 | Axial view at the level of the basal ganglia of brain segmentations on MPRAGE, FLAIR, DWI, and CT. The MRI were obtained from the same patient, the CT image stems from a different patient. The thalamus, the nucleus lentiformis, the nucleus caudatus, and the cortical ribbon are well-demarcated on all contrast. The segmentation of the cortical ribbon on CT and DWI, where the white matter–gray matter (WM-GM) contrast is low, is less detailed compared to MPRAGE, but still of good quality.

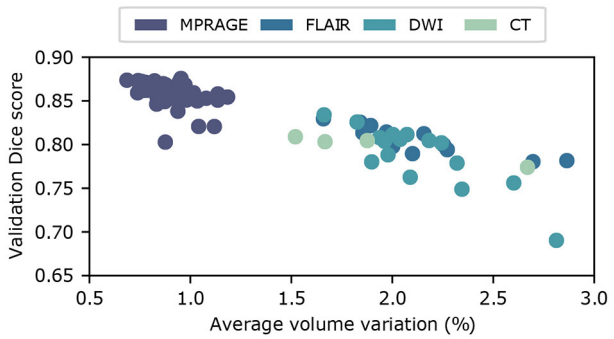

FIGURE 5 | Uncertainty estimation using dropout sampling. Scatter plot of the average Dice score $\mathcal{D}_A$ vs. the coefficient of variation *CV*. Each scatter point corresponds to a input volume from the test dataset. Coefficients of variation are obtained from $N = 15$ MC samples. The strong correlation between $\mathcal{D}_A$ and *CV* demonstrates that *CV* is a good measure of segmentation quality. Pearson correlation coefficients, derived separately for each imaging modality, are summarized in **Table 1**.

ground truth labels, for volumes from the test set. Here, we combine the results for all imaging modalities. We clearly observe a strong correlation between CV_s and Dice scores, which indicates that *CV* is in fact a good metric to gauge the quality of the segmentation. The Pearson correlation coefficients are $C_{\text{MPRAGE}} = -0.91$, $C_{\text{FLAIR}} = -0.87$, $C_{\text{DWI}} = -0.85$, and $C_{\text{CT}} = -0.98$, for the corresponding imaging modalities. As a consequence, we integrate the dropout sampling as an optional processing step into our pipeline, which warns the user if the coefficient of variation *CV* for the requested segmentation exceeds 1.0% for MPRAGE and 2.5% for FLAIR, DWI and CT contrasts, respectively.

4. DISCUSSION

In this work, we investigated the segmentation quality that can be achieved using convolutional neural networks on various MR/CT imaging modalities. In a first study with in total 853 MR/CT scans, we find that T_1-weighted images provide the best segmentation results. This finding agrees with our naive expectation, because the T_1-weighted MPRAGE scans provide the best gray-to-white matter contrast and the ground truth labels were generated on this contrast. Further, the largest dataset for training was available for this contrast. Nevertheless, also FLAIR, DWI and even CT scans can be segmented with excellent results when using current state of the art deep neural networks. In case of CT scans, we observe that segmentation quality is dependent on the anatomical structure. While gray and white matter segmentation is challenging, due to low signal contrast, the performance on ventricles, putamen, pallidum, and brain stem reaches or exceeds the performance achieved on the MRI contrasts.

Because our database of available training samples in the first study is imbalanced across the different modalities, we performed a second study with approximately equally-sized training sets to facilitate a fairer comparison. We found that the general ranking of the achievable segmentation performance remains the same, but that the difference between the performances reduces. Additionally, we investigated the scaling behavior of model performance with the number of available training samples on the MPRAGE contrast.

Finally, we implemented an uncertainty estimator, using dropout sampling as introduced in (29), to gauge the quality of the generated segmentation labels. We observe a strong correlation between our uncertainty metric, the coefficient of volume variation *CV*, and the quality of the segmentation derived from the ground truth labels. This is the case for all imaging modalities. Consequently, we incorporate the uncertainty metric *CV* in our segmentation pipeline to identify faulty input volumes or low-quality segmentation.

Based upon the presented results, several improvements and further investigations that extend the applicability of our segmentation pipeline can be envisioned: In our work, we use input volumes with ∼2 million voxels, which is limited by the available memory on our GPU. In the future, it would be desirable to scale this number up by one order of magnitude in order to directly process entire 3D brain scans with an isotropic resolution of 1 mm, which results in input volumes with approximately 20 million voxels. This could be achieved by a combination of more memory-efficient network architectures, ensembles of smaller segmentation networks, which only address a subset of labels, or via improvements in GPU hardware.

In addition, our investigation on the achievable segmentation performance on different imaging modalities is only the first step toward reliable segmentation tools for DWI, FLAIR, and CT volumes. As a next step, the segmentation performance should also be quantified relative to manual label maps from human experts. This could be done by performing the manual annotation on MPRAGE scans and transferring the labels over to the other modalities, as it was done in our study with FreeSurfer labels. Alternatively, the annotations could also be directly added to the DWI, FLAIR, and CT volumes. This was recently done for CT scans (33) and would alleviate the systematic errors introduced by non-perfect coregistration.

Based on performance, flexibility, and processing speed, CNNs already now represent a valuable tool for automated anatomical segmentation. In our view, however, the most significant obstacle to the broad applicability of segmentation CNNs is the limited generalizability to different acquisition parameters and MRI/CT scanners. To train networks that generalize very well, the generation and distribution of large structured databases of MRI and CT scans, acquired on various scanners and imaging contrasts, is highly desirable. In addition, further research on the combination or improvement of methods, such as lifelong learning (34) or advanced data augmentation (35) is necessary. In terms of data augmentation, generative models, such as generative adversarial networks (GANs) or variational autoencoders (VAEs) could be used to generate large databases

of synthetic MRI/CT scans. These databases could subsequently be used to enhance training.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: datasets are not public due to data protection. Request to access these datasets should be directed to federau@biomed.ee.ethz.ch.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethikkommission Nordwest- und Zentralschweiz. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

JZ and SP implemented and trained the neural networks. JZ wrote the manuscript with input from all other authors. All authors were involved in the preparation of the dataset.

FUNDING

This work was supported by a SPARK award of the Swiss National Science Foundation and a University of Zurich Forschungskredit (Spark Grant: CRSK-3_190697).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fneur.2021.653375/full#supplementary-material

REFERENCES

1. Scahill RI, Schott JM, Stevens JM, Rossor MN, Fox NC. Mapping the evolution of regional atrophy in Alzheimer's disease: unbiased analysis of fluid-registered serial MRI. *Proc Natl Acad Sci USA*. (2002) 99:4703–7. doi: 10.1073/pnas.052587399
2. Halliday GM, McRitchie DA, Macdonald V, Double KL, Trent RJ, McCusker E. Regional specificity of brain atrophy in Huntington's disease. *Exp Neurol*. (1998) 154:663–72. doi: 10.1006/exnr.1998.6919
3. Lu PH, Mendez MF, Lee GJ, Leow AD, Lee HW, Shapira J, et al. Patterns of brain atrophy in clinical variants of frontotemporal lobar degeneration. *Dement Geriatr Cogn Disord*. (2013) 35:34–50. doi: 10.1159/000345523
4. Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*. (2002) 33:341–55. doi: 10.1016/S0896-6273(02)00569-X
5. Shattuck DW, Leahy RM. BrainSuite: an automated cortical surface identification tool. *Med Image Anal*. (2002) 6:129–42. doi: 10.1016/S1361-8415(02)00054-3
6. Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM. FSL. *NeuroImage*. (2012) 62:782–90. doi: 10.1016/j.neuroimage.2011.09.015
7. Avants BB, Tustison NJ, Stauffer M, Song G, Wu B, Gee JC. The Insight ToolKit image registration framework. *Front Neuroinformatics*. (2014) 8:44. doi: 10.3389/fninf.2014.00044
8. Dale AM, Fischl B, Sereno MI. Cortical surface-based analysis: I. Segmentation and surface reconstruction. *NeuroImage*. (1999) 9:179–94. doi: 10.1006/nimg.1998.0395
9. Korfiatis P, Kline TL, Erickson BJ. Automated segmentation of hyperintense regions in FLAIR MRI using deep learning. *Tomography*. (2016) 2:334–40. doi: 10.18383/j.tom.2016.00166
10. Gibson E, Gao F, Black SE, Lobaugh NJ. Automatic segmentation of white matter hyperintensities in the elderly using FLAIR images at 3T. *J Magn Reson Imaging*. (2010) 31:1311–22. doi: 10.1002/jmri.22004
11. Duong MT, Rudie JD, Wang J, Xie L, Mohan S, Gee JC, et al. Convolutional neural network for automated FLAIR lesion segmentation on clinical brain MR imaging. *Am J Neuroradiol*. (2019) 40:1282–90. doi: 10.3174/ajnr.A6138
12. Cheng H, Newman S, Afzali M, Fadnavis SS, Garyfallidis E. Segmentation of the brain using direction-averaged signal of DWI images. *Magn Reson Imaging*. (2020) 69:1–7. doi: 10.1016/j.mri.2020.02.010
13. Ciritsis A, Boss A, Rossi C. Automated pixel-wise brain tissue segmentation of diffusion-weighted images via machine learning. *NMR Biomed*. (2018) 31:e3931. doi: 10.1002/nbm.3931
14. Irimia A, Maher AS, Rostowsky KA, Chowdhury NF, Hwang DH, Law EM. Brain segmentation from computed tomography of healthy aging and geriatric concussion at variable spatial resolutions. *Front Neuroinformatics*. (2019) 13:9. doi: 10.3389/fninf.2019.00009
15. Hu Q, Qian G, Aziz A, Nowinski WL. Segmentation of brain from computed tomography head images. In: *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*. Shanghai (2005). p. 3375–8. doi: 10.1109/IEMBS.2005.1617201
16. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS'12*. Red Hook, NY; Lake Tahoe, CA: Curran Associates Inc. (2012). p. 1097–105.
17. Moeskops P, Viergever MA, Mendrik AM, de Vries LS, Benders MJNL, Išgum I. Automatic segmentation of MR brain images with a convolutional neural network. *IEEE Trans Med Imaging*. (2016) 35:1252–61. doi: 10.1109/TMI.2016.2548501
18. Mehta R, Majumdar A, Sivaswamy J. BrainSegNet: a convolutional neural network architecture for automated segmentation of human brain structures. *J Med Imaging*. (2017) 4:1–11. doi: 10.1117/1.JMI.4.2.024003
19. Huo Y, Xu Z, Xiong Y, Aboud K, Parvathaneni P, Bao S, et al. 3D whole brain segmentation using spatially localized atlas network tiles. *NeuroImage*. (2019) 194:105–19. doi: 10.1016/j.neuroimage.2019.03.041
20. Coupé P, Mansencal B, Clément M, Giraud R, de Senneville BD, Ta VT, et al. AssemblyNet: a novel deep decision-making process for whole brain MRI segmentation. In: Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, et al., editors. *Medical Image Computing and Computer Assisted Intervention-MICCAI 2019*. Cham: Springer International Publishing (2019). p. 466–74. doi: 10.1007/978-3-030-32248-9_52
21. Lee M, Kim J, EY Kim R, Kim HG, Oh SW, Lee MK, et al. Split-attention U-Net: a fully convolutional network for robust multi-label segmentation from brain MRI. *Brain Sci*. (2020) 10:974. doi: 10.3390/brainsci10120974
22. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Vol. 9351. Munich: Springer (2015). p. 234–41. doi: 10.1007/978-3-319-24574-4_28
23. Wachinger C, Reuter M, Klein T. DeepNAT: Deep convolutional neural network for segmenting neuroanatomy. *NeuroImage*. (2018) 170:434–45. doi: 10.1016/j.neuroimage.2017.02.035
24. Henschel L, Conjeti S, Estrada S, Diers K, Fischl B, Reuter M. FastSurfer - A fast and accurate deep learning based neuroimaging pipeline. *NeuroImage*. (2020) 219:117012. doi: 10.1016/j.neuroimage.2020.117012

25. Roy] AG, Conjeti S, Navab N, Wachinger C. QuickNAT: A fully convolutional network for quick and accurate segmentation of neuroanatomy. *NeuroImage*. (2019) 186:713–27. doi: 10.1016/j.neuroimage.2018.11.042
26. Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW. elastix: a toolbox for intensity-based medical image registration. *IEEE Trans Med Imaging*. (2010) 29:196–205. doi: 10.1109/TMI.2009.2035616
27. Isensee F, Kickingereder P, Wick W, Bendszus M, Maier-Hein KH. No new-net. In: Crimi A, Bakas S, Kuijf H, Keyvan F, Reyes M, van Walsum T, editors. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Cham: Springer International Publishing (2019). p. 234–44. doi: 10.1007/978-3-030-11726-9_21
28. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin S, Joskowicz L, Sabuncu MR, Unal G, Wells W, editors. *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016*. Cham: Springer International Publishing (2016). p. 424–32. doi: 10.1007/978-3-319-46723-8_49
29. Roy AG, Conjeti S, Navab N, Wachinger C. Bayesian QuickNAT: Model uncertainty in deep whole-brain segmentation for structure-wise quality control. *NeuroImage*. (2019) 195:11–22. doi: 10.1016/j.neuroimage.2019.03.042
30. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: A system for large-scale machine learning. In: *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)* Savannah, GA (2016). p. 265–83.
31. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat Methods*. (2020) 17:261–72. doi: 10.1038/s41592-019-0686-2
32. Gal Y, Ghahramani Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML'16.* (2016). p. 1050–9.
33. Cai JC, Akkus Z, Philbrick KA, Boonrod A, Hoodeshenas S, Weston AD, et al. Fully automated segmentation of head CT neuroanatomy using deep learning. *Radiology*. (2020) 2:e190183. doi: 10.1148/ryai.2020190183
34. Karani N, Chaitanya K, Baumgartner C, Konukoglu E. A lifelong learning approach to brain MR segmentation across scanners and protocols. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G, editors. *Medical Image Computing and Computer Assisted Intervention-MICCAI 2018*. Cham: Springer International Publishing (2018). p. 476–84. doi: 10.1007/978-3-030-00928-1_54
35. Zhao A, Balakrishnan G, Durand F, Guttag JV, Dalca AV. Data augmentation using learned transforms for one-shot medical image segmentation. *CoRR*. (2019) abs/1902.09383. doi: 10.1109/CVPR.2019.00874

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.