IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 15, 2019, accepted April 5, 2019, date of publication April 11, 2019, date of current version April 29, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2910348

Multiclass Brain Tissue Segmentation in 4D CT Using Convolutional Neural Networks

SIL C. VAN DE LEEMPUT, MIDAS MEIJS, AJAY PATEL, FREDERICK J. A. MEIJER, BRAM VAN GINNEKEN, AND RASHINDRA MANNIESING

Diagnostic Image Analysis Group, Department of Radiology and Nuclear Medicine, Radboud University Medical Center, 6525 Nijmegen, The Netherlands

Corresponding author: Sil C. van de Leemput (sil.vandeleemput@radboudumc.nl)

This work was supported in part by the Netherlands Organization for Scientific Research (NWO), The Netherlands, and in part by the Canon Medical Systems Corporation, Japan.

ABSTRACT 4D CT imaging has a great potential for use in stroke workup. A fully convolutional neural network (CNN) for 3D multiclass segmentation in 4D CT is presented, which can be trained end-to-end from sparse 2D annotations. The CNN was trained and validated on 42 4D CT acquisitions of the brain of patients with suspicion of acute ischemic stroke. White matter, gray matter, cerebrospinal fluid, and vessels were annotated by two trained observers. The mean Dice coefficients, contour mean distances, and absolute volume differences were, respectively, 0.87 ± 0.04, 0.52 ± 0.47 mm, and 11.78 ± 9.55 % on a separate test set of five patients, which were similar to the average interobserver variability scores of 0.88 ± 0.03, 0.72 ± 0.93 mm, and 8.86 ± 7.65 % outperforming the current state of the art. The proposed method is, therefore, a promising deep neural network for multiclass segmentation in 4D spatiotemporal imaging data.

INDEX TERMS Deep learning, convolutional neural network (CNN), segmentation, sparse annotations, brain, stroke, computed tomography (CT), CT perfusion (CTP), 3D, 4D.

I. INTRODUCTION

Computed tomography (CT) is at the core of modern acute stroke workup [19]. CT is cheap, widely available, and fast compared to other imaging modalities like magnetic resonance imaging (MRI). Additionally, modern CT scanners can cover the whole brain with high temporal and spatial resolution. From a head CT scan tissue densities can be derived, which enables detecting pathology like hemorrhages. Additionally, acquiring a head CT shortly after injection of contrast agent enables the visualization of the cerebral vasculature and hemodynamics. CT angiography (CTA) and CT perfusion (4D CT) are two such post-contrast techniques, which are respectively a single 3D CT scan and a series of 3D CT scans over time. This work focuses on the later type of acquisition, since we expect 4D CT to be the future image modality for stroke. Essentially, 4D CT contains more temporal information and the CTA can be derived from the 4D CT by a maximum intensity projection [39].

4D CT imaging will become increasingly important in the clinical workup of acute stroke. It can be used to assess penumbra, infarct core and collateral flow, which can be used for selecting stroke patients for reperfusion therapy [32]. A recent prospective clinical trial showed that 4D CT imaging helps in identifying patients who will benefit from endovascular treatment beyond the recommended time window of six hours [28]. Segmentation of soft tissue is important because it enables tissue dependent perfusion analysis, potentially refining the identification of infarction core and penumbra [2]. Segmentation of the cerebral vasculature is important for many applications [19], [34], [43]. We have demonstrated that it can be used to visualize vascular flow disturbances reducing the time to detect abnormalities such as vascular occlusion and arterio-venous malformations [23]. Despite the potential uses of 4D CT imaging for stroke, little work has been done on automatic segmentation of tissues from 4D CT data using computer algorithms.

Only one related method was found for 4D CT [22] which was based on a traditional pattern recognition approach. Although Manniesing et al. [22] does provide a coarse segmentation for cerebrospinal fluid (CSF) and vessels, the quantitative evaluation was only done for white matter (WM) and gray matter (GM), and only in axial direction of slices at specific brain locations. To our knowledge, a full

The associate editor coordinating the review of this manuscript and approving it for publication was Ge Wang.

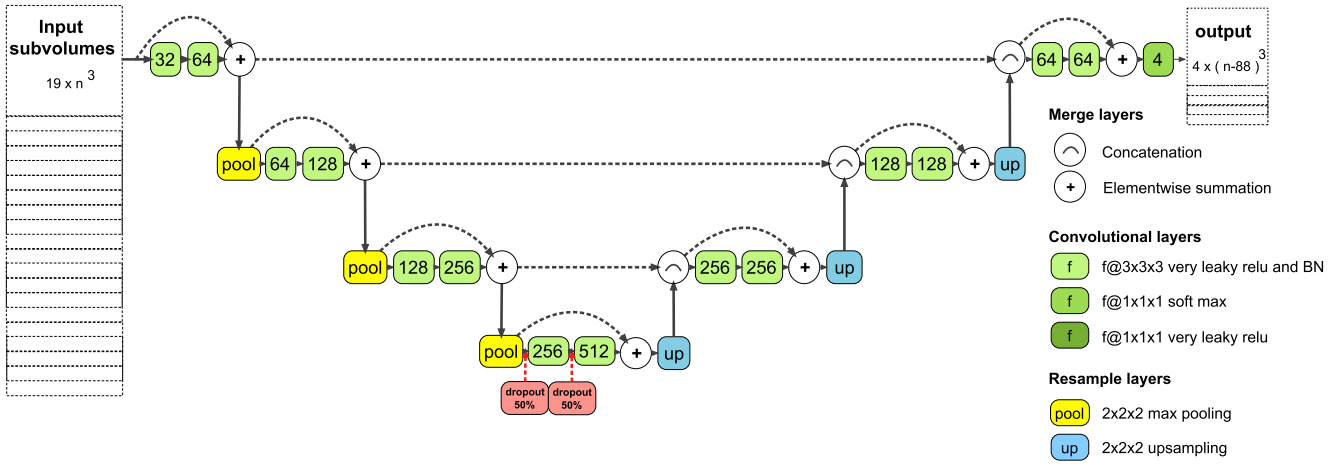

FIGURE 1. Our model, a CNN architecture for multiresolution volumetric segmentation from 4D data. Input data were 4D CT subvolumes consisting of 19 timepoints and a input size of $n \times n \times n$ voxels, with $n \in \{92, 100, 108, 116, \ldots\}$. The network produced volumetric class probability maps for the four segmentation classes with the same size as the input minus the size of the receptive field of 88 × 88 × 88 voxels. BN is an abbreviation for batch normalization.

multiclass 3D segmentation method that includes WM, GM, CSF and vessels in 4D CT and that has been quantitatively evaluated for all classes, is currently nonexistent.

In this work we present a method for 3D multiclass segmentation in 4D CT using a multiresolution fully convolutional neural network (CNN) which is able to learn end-to-end from 2D sparse annotations. The CNN is applied to 4D CT images of acute ischemic stroke patients for segmentation of WM, GM, CSF, and vessels.

Medical imaging has witnessed a sharp rise of applications based on convolutional neural networks (CNNs) in a few years of time [21]. CNNs are feed forward artificial neural networks consisting of multiple convolutional layers successively encoding higher abstract representations. A powerful trait of CNNs is that representations can be directly learned from data without the need for manually creating or selecting features.

However, many deep learning approaches avoid learning from high dimensional data because of practical limitations, i.e. higher GPU memory requirements and increased number of computations. For example, [6], [11], [26], [36], [37], [40], [47] propose a 2.5D approach in which multiple 2D patches are sampled in different orientations around a center voxel in 3D, and are then fed individually into a 2D CNN for predicting the output class at the intersection. This approach is suboptimal since 3D context outside of the sampled planes is ignored.

Full 3D approaches have been proposed to a lesser extent. Most provide fully convolutional approaches that include multiresolution contextual information by processing downscaled versions of the input and integrating the lower resolution images later in the network at the original voxel resolution [1], [5], [9], [16], [18], [25]. 3D U-Net [5] is a fully convolutional network which processes 3D input at four different image resolutions and provides a voxel weighting scheme and smooth deformation field data augmentation to be able to learn from sparsely annotated data. Other 3D segmentation approaches try to leverage recurrent operations [4], [31], [41], [46]. Some segmentation approaches [3] utilize CNNs for processing multi-channel 3D data, but the channels represent data from different modalities, whereas 4D spatiotemporal data represents multiple acquisitions using the same modality over time. The distinction is useful, since the voxel intensities encode for similar physical phenomena in the latter case, hence calculating statistics (e.g. averages, variance) over the temporal dimension becomes sensible. For example, consider carefully registered temporal images, taking a temporal average yields a meaningful image, since its voxel intensities are approximately similar. However, for multi-model data, for example MR T0, T1 and Flair images, averaging over its channels is less meaningful, since the voxel intensities do not correspond between channels. Only a single work was found in the literature that addressed 4D spatiotemporal data [38] for automatic multi-organ detection in MR using unsupervised deep learning techniques. However, the resulting segmentations have limited precision and class overlap.

II. METHODS

A. MODEL ARCHITECTURE

CNNs can be represented as directed graphs, where a node (hereinafter referred to as layer) indicates an operation on volumetric feature maps, incoming edges indicate what feature maps are fed to a layer, and outgoing edges represent the feature maps produced by a layer. Figure 1 shows such a representation of our model. It consists of 15 convolution layers (green), 3 max-pool layers (yellow), and 3 upscale layers (blue). All solid arrows form the path through the network that visits all the layers exactly once, whereas the dotted arrows skip several layers within the network (shortcuts). In addition to the shortcuts from the original U-Net [35], shortcuts were added over every two consecutive

3^3 convolutions as these were found to speed up convergence and increase overall performance in combination with the other shortcuts [10].

The model uses concatenation or elementwise summation to merge two sets of feature maps at a layer into a single set of feature maps. The concatenation layer joined the two sets, resulting in a larger set of feature maps. To perform elementwise summation, both sets are required to have the same number of feature maps. If this was not the case, all feature maps from the first set A (at the start of the curved arrows in Figure 1) were (repeatedly) iterated and concatenated to a new set C until the number of feature maps between set A and the second set B were the same. Next, the new set C was used instead of the original first set A to perform elementwise summation. For example, let $A = (a, b, c)$ (3 feature maps) and $B = (f_1, f_2, f_3, \ldots, f_{64})$ (64 feature maps). Now a new set $C = (a, b, c, a, b, c, \ldots)$ is created from set A by iterating its elements until it has the same 64 feature maps as set B. Finally, the feature map sets B and C are merged by summation per feature map at the summation layer $D = (f_1 + a, f_2 + b, f_3 + c, f_4 + a, f_5 + b, \ldots, f_{63} + c, f_{64} + a)$. Note that swapping the contents of sets A and B, yields $D = (f_1 + a, f_2 + b, f_3 + c)$. The feature maps in both sets were cropped around the center to the smallest input feature map size for both merge layers to resolve size mismatches.

The network was inspired by the 3D U-Net architecture. Feature extraction at each voxel resolution was achieved by two subsequent 3^3 convolution layers with batch normalization [15]. Each of these convolution was followed by a leaky variant of a rectified linear activation unit (very leaky ReLU) as defined in [13]:

$$f(x) = \begin{cases} x, & \text{if } x \geq 0 \\ x/3, & \text{otherwise} \end{cases} \tag{1}$$

The very leaky ReLU was preferred over the normal ReLU since it emits similar behavior, but prevents 'dying ReLU'. This problem refers to a unit which only produces zeros for any given input and which is unlikely to break out of that state during training, which make these units no longer useful. Downscaling the input by a factor of 2 was achieved by a 2^3 max-pool layer. For this architecture, there are four voxel resolutions at which features were extracted: the original resolution of 0.5 mm^3/voxel, 1.0 mm^3/voxel, 2.0 mm^3/voxel and 4.0 mm^3/voxel. To synthesize the output class probability maps from the lower resolution feature maps, the lower resolution feature maps were first upsampled at each upsampling layer by a factor 2 using nearest neighbor interpolation and were then concatenated with the feature maps acquired earlier at similar resolution (depicted by the horizontal striped arrows in Figure 1). This upscaling operation was preferred to the deconvolution operations in Çiçek et al. [5], since the latter is thought to introduce artificial checkerboard patterns in the output [29]. This data integration process was repeated from lowest to the highest resolution until feature maps at the original voxel resolution were retrieved. Finally, the output at the last layer was passed through a soft-max activation function.

The network architecture was fixed for training and evaluation and therefore introduced a fixed relation between network input size and network output size. For instance, at each 3^3 convolution layer the size of the input feature maps is reduced by 2 voxels, whereas a 2^3 max-pooling layer halves the number of voxels and a 2^3 upsampling layer doubles the spatial voxel size for the output feature maps. As the feature maps were passed from layer to layer through the network, it finally produced an output size which had 88 voxels less than the input size in every spatial dimension. In this particular case, the size difference equals the size of the receptive field of the network, where the size of the receptive field of the network is the spatial extent of the input voxels (subvolume) which contribute to the activation of a single output unit, i.e. to the output class probability for an individual voxel.

The network input were batches consisting of 4D CT subvolumes. Each subvolume could be varied in size (number of voxels per spatial dimension) and could be varied in batch size (number of subvolumes per batch), but should always have a fixed number of time points. Selecting the subvolume size and the batch size have practical implications on the required GPU RAM and on training performance. Valid input size values are $n \in \{92, 100, 108, 116, \ldots\}$, since n must be bigger than the size of the receptive field of the network ($n > 88$) and the input size should produce even sized feature maps before each pooling layer to preserve voxel correspondence at each resolution. For the experiments in this work, we fixed the number of time points to 19, since it matched the number of time points for each 4D CT acquisition collected for this study (see section III). For network training, we put the subvolume size to $n = 124$ voxels for each spatial dimension and employed a batch size of 2. This gave an output class probability map per segmentation class with 36 voxels ($124 - 88$) for each spatial dimension.

The full size final prediction segmentations were obtained following a similar strategy as described by Çiçek *et al.* [5] and Ronneberger *et al.* [35], by repeatedly shifting and applying a CNN on the input data until all input voxels had their corresponding predictions. First, the input data was zero padded with a border half the size of the receptive field of 44 voxels for all spatial dimensions. Next, the model was repeatedly applied until all voxels within the input data had corresponding brain tissue predictions.

B. MODEL TRAINING

In deep learning, training of the architecture is at least as important as the design of the architecture. In this work, a training strategy was used similar to the work of Çiçek et al. [5], consisting of a categorical cross-entropy objective function adapted for sparse data; this was minimized using default stochastic gradient descent optimizer with Nestorov momentum [27]. Training was done on sparse annotations, that is, annotations in 2D cross sections of 3D

volumes derived from 4D data (See section III). In this section, we describe the objective function and parameter regularization, data sampling and augmentation, parameter initialization, optimizer, and other technical details. The reported hyperparameters in this section were experimentally selected.

1) OBJECTIVE AND REGULARIZATION

The training objective is to find the set of weight parameters Θ for our model that minimize the loss function $L(\Theta \mid t, w)$, given the reference standard t and voxel weights w. The loss function was constructed from the weighted categorical cross entropy $WCCE(\cdot)$, and L_1-norm $L_1(\cdot)$ and L_2-norm $L_2(\cdot)$ weight regularization terms, as follows:

$$L(\Theta \mid t, w) = \lambda_0 WCCE(\Theta \mid t, w) + \lambda_1 L_1(\Theta) + \lambda_2 L_2(\Theta) \quad (2)$$

where $\lambda_0 = 1$, $\lambda_1 = 1e^{-6}$, and $\lambda_2 = 1e^{-5}$. The $WCCE(\cdot)$ is the weighted categorical cross-entropy loss function, which calculates the weighted mean over the categorical cross entropy $CCE_i(\cdot)$ per voxel i with weights w_i. The $CCE_i(\cdot)$ defines the error between output $p_{i,j}(\Theta)$ of the soft-max activation function at the last layer of our model given the weight parameters Θ and the reference standard $t_{i,j}$ for each voxel i and segmentation class j:

$$WCCE(\Theta \mid t, w) = \frac{\sum_i w_i CCE_i(\Theta \mid t)}{\sum_i w_i}$$
$$CCE_i(\Theta \mid t) = -\sum_j t_{i,j} log(p_{i,j}(\Theta)) \quad (3)$$

The weights w were set to an annotation mask by setting the weights w_i to 1 if annotations were present for voxel i and to 0 otherwise, thereby only learning from labeled voxels.

Dropout was applied during training before the 3^3 convolutions by setting 50% randomly selected voxels to zero at the coarsest image resolution (Figure 1) for each processed batch.

2) SAMPLING

All the annotated voxels within the cranial cavity formed the sampling candidates. The cranial cavity is defined as the space containing all soft tissues and CSF, including the meninges, cerebrum, ventricles, cerebellum, and brain stem, and was segmented using the method of Patel et al. [30]. Each subvolume selected during training was centered on a single sampling candidate in world coordinates. All subvolume voxels which were sampled outside of the input data were set to zero value.

Each CNN model was shown 60k subvolumes, which were processed in batches of 2 subvolumes during training. For every 400 subvolumes an equal number of candidates were sampled uniformly per tissue type from the set of sampling candidates.

3) AUGMENTATIONS

Five types of augmentations were used during training to artificially enlarge the sparsely annotated dataset. The use of augmentations have shown to prevent overfitting, improve generalization, and introduce invariance to the augmentations used [5], [20].

For each subvolume in the training data, one of the five following augmentations was assigned with equal probability: identity-, mirroring-, rotation-, uniform scaling-, or elastic deformation. Only one augmentation was computed per subvolume to keep the computation time low. The identity transformation reproduces the original signal. Mirroring flips the input along the sagittal axis only. Rotation is expressed as a 3D Euler rotation in degrees around the center of the subvolume where the x, y, and z rotations are individually sampled form the continuous uniform distribution $\mathcal{U}(-8, 8)$. Uniform scaling is defined as an affine transform which rescales the input uniformly by a scalar over all axis, which is sampled from the continuous uniform distribution $\mathcal{U}(1.01, 1.25)$. Scaling down was omitted from the scaling augmentation since it could potentially remove small vascular structures in the input. The elastic deformation applies a 3D linear interpolation of the input subvolume where each individual corner point of the bounding box of the subvolume was given a different randomized offset in voxels for the x, y, and z coordinates drawn from the normal distribution $\mathcal{N}(0, 6)$, resulting in warped subvolumes.

A selected transformation was calculated once and then applied to the input subvolume, annotation labels, and annotation mask, with interpolation orders 1, 0, and 0, respectively.

4) WEIGHT INITIALIZATION

At the start of training, all weights in the model were initialized using a He initialization scheme [14], which was adjusted for the very leaky ReLU activation function (equation 1). That is, at each layer, the weights were sampled from the following normal distribution:

$$\mathcal{N}\left(0, \sqrt{9/(5fan_{in})}\right) \quad (4)$$

where fan_{in} is defined as the number of feature maps being input to the layer multiplied by the size of the convolution kernel. To keep the initialization constant across different experiments, the random generators were seeded with the same constant.

5) OPTIMIZER AND IMPLEMENTATION

A stochastic gradient descent optimizer was used starting with a learning rate of 0.1, which was decreased by a factor of 10 after having processed every 20k subvolumes. Momentum was used and was kept constant at 0.9.

The model was implemented in Theano/Lasagne [8], [44] and training was performed on an NVidia Titan X graphics card with 12 GB of video memory.

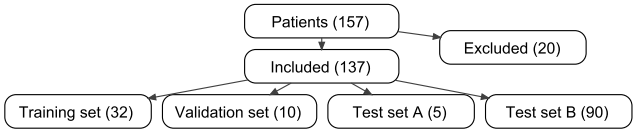

FIGURE 2. Data flow diagram for the 4D CT data distribution over the training set, the validation set, test set A, and test set B.

III. DATA

A. PATIENT INCLUSION AND IMAGE ACQUISITION

This retrospective study was approved by the institutional ethics committee and informed consent was waived. In total, 157 patients (age 63 $\pm$ 14 years, 61% male) with a suspicion of stroke in 2015 or 2016 at the Radboud University Medical Center, Nijmegen, the Netherlands, were included. 4D CT were acquired using a 320-row CT scanner (Toshiba Aquilion ONE, Japan) consisting of 19 volumetric scans with different exposures per time point. Patients received 80 mL of contrast agent (Iomeron) injected in the cephalic vein at the start of the first acquisition. Image reconstruction was done using a FC41 smooth convolution kernel, resulting in $512 \times 512 \times 320$ voxels with a voxel size of $0.47 \times 0.47 \times 0.5$ mm. One full 4D CT acquisition took in total less than a minute to complete using a strict protocol with fixed time intervals between each of the 19 volumetric scans. No preprocessing or motion correction were performed during the acquisition.

Twenty patients were excluded because of the presence of large pathology (bleedings, infarcts, and excessive liquor) or because of imaging artifacts (e.g., clips, drains, patient motion, or beam hardening). Test set B was formed from ninety patients cases. The remaining 47 patients were randomly split into a training set of 32, a validation set of 10, and test set A of 5 patients. Test set A was also used to assess the observer variability. Figure 2 summarizes the data selection.

B. PREPROCESSING

Time points $t > 0$ were rigidly registered to the first time point ($t = 0$), to correct for potential head movement during acquisition. The registration was performed with Elastix [17] using the steps and parameter settings as described by Manniesing et al. [22].

Cerebral soft tissue has a limited intensity range in CT, approximately 20 HU to 65 HU [33]. Intensity values outside this range, for example bone which starts from 700 HU, may complicate training of CNNs and limit the optimal achievable performance. Therefore, the registered 4D CT was first clipped within the range $[-50, 400]$ then linearly scaled to $[0, 1]$. A broader clipping range was used rather than the defined soft tissue HU ranges to preserve more spatial contextual information.

C. REFERENCE STANDARD

The reference standard was obtained by manually annotating the WM, GM, CSF, and vessels in a single 2D cross section per patient imaging data. Annotations were carried out by two medical students, who were trained and supervised by an experienced neuroradiologist with more than 10 years of experience. A 3D annotation tool, called VCAST (volumetric cluster annotation and segmentation tool [45]) was developed in-house specifically for this task. VCAST provides normal annotation capabilities (like brushes for annotating voxels in a cross section) and, in addition, provides supervoxel grids of various sizes to add or remove annotations. Other capabilities of the tool include instant navigation to the cross sections requiring annotations and preset keys for the window levels (center/width was 30/80 HU for CSF, 50/50 HU for WM/GM, and 60/60 for vessels).

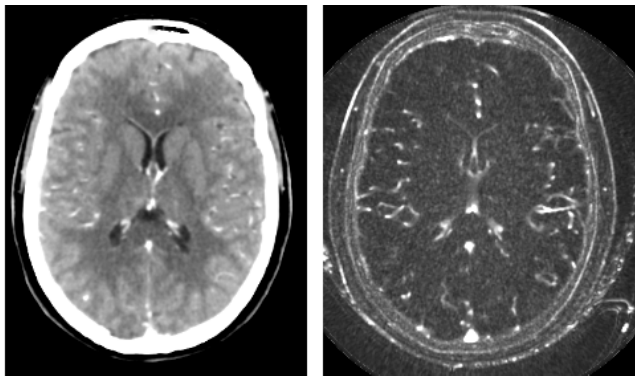

FIGURE 3. Example axial cross section for the derived images of a single 4D CT image used for annotation. Left: the temporal average for WM, GM, and CSF segmentation. Right: the temporal variance for vessel segmentation.

4D CT data is hard to interpret by human readers, which results in long annotation times and an increased likelihood of error. To facilitate the human readers, two images were derived for each 4D CT, by merging the temporal information. The weighted temporal average (WTA) [22] for annotating the WM, GM, and CSF because it has the highest signal-to-noise ratio and best soft tissue contrast and the weighted temporal variance (WTV) [24] for annotating the vessels because of its sensitivity to contrast variations. However, even in the WTV image, manually annotating vessel structures is complex and time consuming because of their varying shapes and sizes and the partial volume effects. Therefore, vessels were first pre-segmented by an automated segmentation algorithm based on local histogram features and a random forest classifier [24]. This segmentation was then presented within VCAST for further manual refinements. See Figure 3 for an example of the derived images.

The areas selected for annotation are indicated in blue in Figure 4. The cerebellum was insufficiently detailed for an experienced reader to reliably derive WM and GM annotations from — mainly because of the limitations of CT imaging — and was therefore excluded from all cross sections. The falx cerebri and the tentorium cerebelli were left out because these structures do not contain any of the four tissue types used in this study.

The method of Patel et al. [30] was used to segment all intracranial soft tissue, which was then manually adjusted

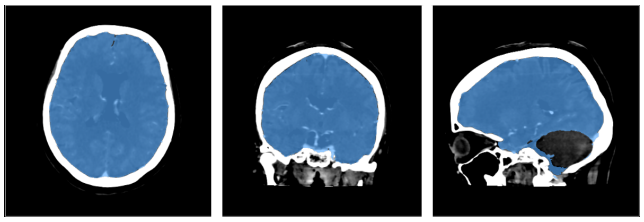

FIGURE 4. Three cross sections (axial, coronal, sagittal) of an exemplar 4D CT case. Blue areas were selected for annotation by the observers, other areas were not annotated.

to reflect masks similar to Figure 4. For each cross section, the orthogonal plane was randomly selected, after which the cross section to be annotated was extracted from the selected plane. All cross sections consisting of less than 10% of the mask voxels were excluded from selection. Six patients had 2D cross sections for all three orthogonal planes, each plane was selected using previously described method.

During annotations, four small densely connected voxel subareas were found within the to be annotated cross sections for which the soft tissue labels could not reliably be determined by the observers; these areas were ignored during training (three areas) and evaluation (one area). The areas had an average size of 76.4 mm^3, were less than 0.1% of all annotated voxels and the effect on the evaluation measures was assessed to be insignificant. After the observers annotated all patients once, two qualitative inspections were performed by the radiologist to assess overall annotation quality. Errors detected during these inspections were subsequently corrected.

In total over 410 hours were spent by the observers in creating the reference standard.

IV. EXPERIMENTS

A. OBSERVER VARIABILITY

Observer variability was estimated on five 4D CT data from test set A, which were annotated in two subsequent series by both observers. When observers were unsure about their annotations, they were asked to leave those voxels out. Only voxels annotated twice by both observers were used for calculating the estimation. Intraobserver variability was reported for both observers and interobserver variability was reported for the first series of annotations. The Dice Similarity Coefficient (DSC) [7], contour mean distance (CMD), absolute volume difference (AVD), and mean volume difference (MVD) were used as measures of evaluation.

The CMD between two non-zero pixel sets A and B is defined as: $CMD(A,B) = max(h(A,B), h(B,A))$ where $h(A,B) = \text{mean}_{a\in A}\min_{b\in B}||a-b||$.It defines the mean distance between boundaries of non-zero pixel regions. The MVD between two non-zero-pixel sets A and B is defined as: $MVD(A,B) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|A_i - B_j|$, which computes the volume difference in mm^3. The AVD is then defined as: $AVD(A,B) = MVD(A,B)/\frac{1}{n}\sum_{i=1}^{n}A_i$, computing the relative volume difference between A and B in %.

B. MODEL EVALUATION

Our model was compared with 3D U-Net [5], which is a state-of-the-art CNN model for volumetric image segmentation, on DSC for the segmentations. The models are similar except that our model has additional shortcuts over every pair of two 3^3 convolutions, uses very leaky ReLU instead of ReLU as an activation function throughout the architecture, and uses nearest neighbor upsampling instead of deconvolution. With respect to the training parameters, our model uses modified He initialization instead of Xavier initialization [12], uses a batch size of 2 instead of 1, and has additional L_1 and L_2 regularization terms on the weights.

For a fair comparison, the 3D U-Net was trained and evaluated in the same manner as described in section II-B, but used the architecture and weight initialization scheme from the original work. Additionally, the two models shared most of the hyper parameters, which were experimentally tuned on the validation set. We have kept the batch size (2) and subvolume size (124^3 voxels) the same. Also, upsampling layers instead of deconvolution layers were used since we wanted to avoid checkerboard artifacts [29]. Furthermore, we used the same optimizer, learning rate scheme, momentum term, L1 and L2-norm weighting, and augmentations from section II-B. Essentially, the only differences between the models were: the weight initialization, use of additional shortcuts over every pair of two 3^3 convolutions, and the activation function. All other hyper parameters were kept the same.

DSC, CMD, AVD, and MVD were used as evaluation measure in all experiments. Each model was trained on 60k subvolumes randomly drawn from the annotated voxels of the training set of 32 registered and normalized cases (section III-B). After every 400 subvolumes processed the DSC for each model was calculated on the separate validation set of 10 cases. After completing all iterations, the model with the highest average DSC for all classes on the validation set was evaluated one last time on test set A of five cases to assess the final performance. The average of all the test cases were reported and specified per tissue type and observer.

C. STATE-OF-THE-ART COMPARISON

Our best performing model from section IV-B was compared to Manniesing et al. [22] which is the current state-of-the-art for WM/GM segmentation in 4D CT. The latter method is based on feature extraction and support vector machine (SVM) classification. It was evaluated on a different dataset with 22 different patients than the 32 patients in section III, but the data was obtained with the same scanner. The dataset had more annotated cross sections 87 than our training dataset 40. The method was cross-validated on selected axial cross sectional slices in 22 patients.

For a fair comparison, only the voxels within the WM and GM classes defined by their reference standard were used for evaluation, since the competing method provided coarse unevaluated segmentation classes for CSF and vessels.

TABLE 1. The observer variability across five cases. Measures used are the Dice coefficient (DSC), contour mean distance (CMD) in mm, and absolute volume difference (AVD), and mean volume difference (MVD) in mm^3.

		CSF	WM	GM	vessel	mean
DSC	**intraobs. 1**	0.94 ± 0.01	0.94 ± 0.01	0.95 ± 0.01	0.95 ± 0.01	0.94 ± 0.01
	intraobs. 2	0.85 ± 0.07	0.87 ± 0.02	0.91 ± 0.02	0.90 ± 0.03	0.88 ± 0.05
	interobs.	0.86 ± 0.04	0.86 ± 0.02	0.89 ± 0.02	0.89 ± 0.03	0.88 ± 0.03
CMD	**intraobs. 1**	0.13 ± 0.10	0.10 ± 0.02	0.13 ± 0.06	0.08 ± 0.02	0.11 ± 0.07
(mm)	**intraobs. 2**	1.16 ± 1.26	0.24 ± 0.08	0.37 ± 0.13	0.42 ± 0.17	0.55 ± 0.73
	interobs.	1.46 ± 1.61	0.35 ± 0.13	0.48 ± 0.09	0.59 ± 0.20	0.72 ± 0.93
AVD	**intraobs. 1**	2.48 ± 1.62	2.56 ± 2.55	2.35 ± 2.95	4.52 ± 2.64	2.98 ± 2.65
(%)	**intraobs. 2**	6.89 ± 6.87	6.98 ± 4.28	3.61 ± 2.88	3.78 ± 2.65	5.31 ± 4.78
	interobs.	17.28 ± 9.59	6.81 ± 2.50	6.81 ± 3.61	4.52 ± 4.94	8.86 ± 7.65
MVD	**intraobs. 1**	17 ± 9	61 ± 76	48 ± 71	10 ± 7	34 ± 57
(mm^3)	**intraobs. 2**	68 ± 63	120 ± 64	61 ± 48	7 ± 5	64 ± 65
	interobs.	152 ± 97	129 ± 56	116 ± 66	11 ± 12	102 ± 85

We compared the output segmentations of our method with that of Manniesing et al. [22] using DSC, CMD, AVD, and computation time. The best performing model (full model, trained on 4D data from scratch) was selected and applied to the entire dataset from Manniesing et al. [22] without preprocessing or fine-tuning. Similar as in Manniesing et al. [22], the statistics were first calculated per slice and then averaged over all slices.

D. EXTENDED EVALUATION

Our method was applied to all cases from test set B and the resulting segmentations were qualitatively inspected. For ten of these cases a single cross section was annotated by a single observer using the same selection and annotation procedures as in section III-C. For this annotated subset, segmentations from our method and 3D U-Net were quantitatively scored versus on DSC, CMD, AVD, and MVD.

E. ABLATION EXPERIMENTS

Ablation experiments were performed to assess the contribution of the He initialization scheme versus Xavier initialization, the addition of the short shortcut connections, and the replacement of the ReLU by the very leaky ReLU activation function. We used our best performing model architecture and training scheme as a basis and trained three new models. For the first model we replaced the modified He initialization scheme with Xavier initialization. For the second model we left out the additional short shortcut connections, and for the third model we replaced the very leaky ReLU functions with normal ReLU functions. All models were reinitialized at the beginning of training and were trained as described in section II-B. The best models were selected by taking the highest average DSC performance on the validation set. The best models were evaluated on the ten annotated cases from previous experiment.

V. RESULTS

A. OBSERVER VARIABILITY

The observer results are summarized in Table 1. The average DSC intra- and interobserver agreements were equal or greater than 0.85 for all tissue types for both observers, with most classes having over 0.90 overlap. Interobserver agreement had average DSC scores equal or greater than 0.86 and overall were slightly lower than the intraobserver agreement. Paired t-test showed statistically significant differences ($p < 0.05$) between the two observers for all tissue types.

B. MODEL EVALUATION

The evaluation results are summarized in Table 2. In general, high degrees of overlap with our model and the reference standard were found for all classes, with average DSC in the range of $[0.85, 0.88]$ for observer 1 and $[0.82, 0.84]$ for observer 2. Paired t-tests over all experiments and classes showed significant differences between observers ($p < 0.05$). Paired t-tests showed that the segmentation results from our model and that of 3D U-Net differed significantly ($p < 0.05$). The training time for each of the models were approximately 4 days.

Figure 5 shows the results on test set A of five patients obtained from the best performing model.

C. STATE-OF-THE-ART COMPARISON

The comparison results are summarized in Table 3. Paired t-test showed significant differences for all three evaluation measures for GM and computation time ($p < 0.05$) and a significant difference for WM on CMD ($p < 0.05$). In general, our model outperforms the pattern recognition SVM method by Manniesing et al. [22] on DSC, AVD, CMD and computation time.

D. EXTENDED EVALUATION

The segmentations from our method show good differentiation of the WM, GM, CSF, and Vessels, with slight overestimation of the GM. The method makes more mistakes around imaging artifacts, like streaking and metal artifacts, but overall these errors appear minor. The quantitative results on the extended evaluation set are listed in Table 4.Paired t-tests showed significant differences between our model and 3D U-Net, for all tissue types and all metrics ($p < 0.05$). Overall, our model outperforms 3D U-Net on all metrics.

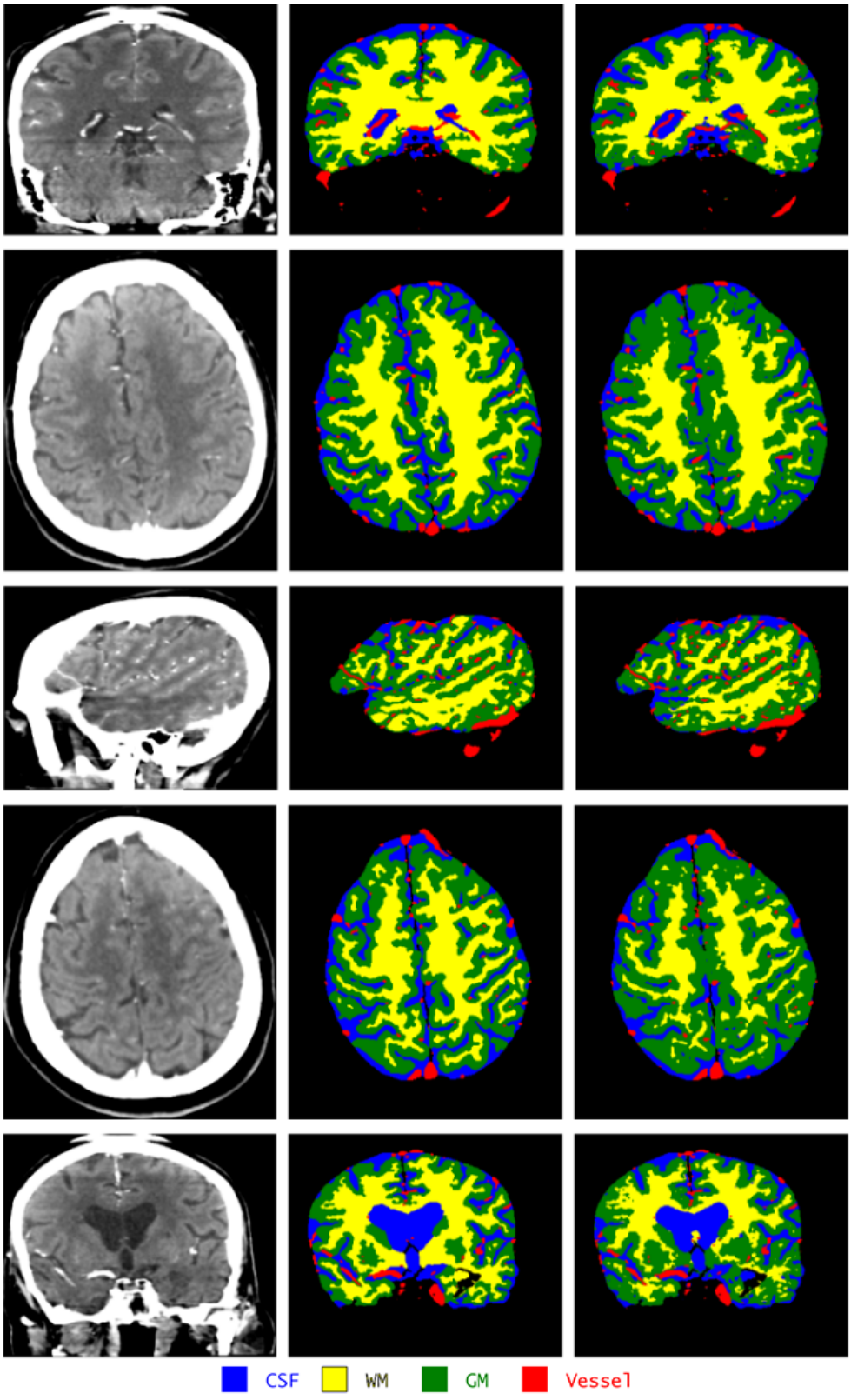

FIGURE 5. Qualitative results produced by our model on test set A. From left to right: temporal average, reference standard (observer 2), model prediction; each row represents an annotated cross section from the five test cases. The cerebellar area in the top row was unlabeled.

TABLE 2. Quantitative segmentation results on the observer reference standards for our model and 3D U-Net. The Dice coefficient (DSC), contour mean distance (CMD) in mm, absolute volume difference (AVD) in %, and mean volume difference (MVD) in mm^3 were used for which the mean and standard deviation were calculated for all five cases in test set A per tissue type and per observer (obs 1 and obs 2). For comparison we have added the interobserver variability. Paired t-tests showed significant differences between observers $p < 0.05$ and between models $p < 0.05$, see section V-B for details. Bold values indicate best performance between models per metric, per class, and per observer.

		interobs.	our model		3D U-Net	
			vs obs 1	vs obs 2	vs obs 1	vs obs 2
DSC	**CSF**	0.86 ± 0.04	$\mathbf{0.85 \pm 0.05}$	$\mathbf{0.81 \pm 0.06}$	0.76 ± 0.10	0.75 ± 0.09
	WM	0.86 ± 0.02	$\mathbf{0.88 \pm 0.04}$	$\mathbf{0.86 \pm 0.03}$	0.88 ± 0.03	0.86 ± 0.03
	GM	0.89 ± 0.02	$\mathbf{0.88 \pm 0.02}$	$\mathbf{0.84 \pm 0.02}$	0.85 ± 0.03	0.81 ± 0.03
	vessel	0.89 ± 0.03	$\mathbf{0.86 \pm 0.03}$	$\mathbf{0.83 \pm 0.03}$	0.65 ± 0.11	0.64 ± 0.10
	mean	0.88 ± 0.03	$\mathbf{0.87 \pm 0.04}$	$\mathbf{0.84 \pm 0.04}$	0.78 ± 0.12	0.77 ± 0.11
CMD (mm)	**CSF**	1.46 ± 1.61	$\mathbf{0.82 \pm 0.80}$	$\mathbf{0.65 \pm 0.45}$	1.68 ± 1.45	1.22 ± 0.93
	WM	0.35 ± 0.13	$\mathbf{0.49 \pm 0.15}$	$\mathbf{0.70 \pm 0.14}$	0.65 ± 0.21	0.76 ± 0.13
	GM	0.48 ± 0.09	$\mathbf{0.38 \pm 0.24}$	$\mathbf{0.38 \pm 0.08}$	0.43 ± 0.20	0.40 ± 0.15
	vessel	0.59 ± 0.20	$\mathbf{0.39 \pm 0.16}$	$\mathbf{0.58 \pm 0.23}$	4.17 ± 1.69	4.20 ± 1.65
	mean	0.72 ± 0.93	$\mathbf{0.52 \pm 0.47}$	$\mathbf{0.58 \pm 0.29}$	1.73 ± 1.86	1.64 ± 1.78
AVD (%)	**CSF**	17.28 ± 9.59	$\mathbf{12.95 \pm 12.24}$	$\mathbf{12.15 \pm 7.49}$	48.90 ± 40.78	35.76 ± 29.74
	WM	6.46 ± 2.15	$\mathbf{10.62 \pm 6.97}$	13.22 ± 4.85	13.03 ± 4.74	$\mathbf{12.36 \pm 6.64}$
	GM	6.48 ± 3.46	9.28 ± 7.27	12.61 ± 6.58	$\mathbf{3.58 \pm 3.69}$	$\mathbf{5.41 \pm 3.63}$
	vessel	4.35 ± 4.76	$\mathbf{14.26 \pm 9.93}$	$\mathbf{14.50 \pm 11.15}$	27.41 ± 14.63	30.10 ± 14.34
	mean	8.86 ± 7.65	$\mathbf{11.78 \pm 9.55}$	$\mathbf{13.12 \pm 7.91}$	23.23 ± 27.75	20.91 ± 21.01
MVD (mm^3)	**CSF**	152 ± 97	$\mathbf{74 \pm 41}$	$\mathbf{109 \pm 82}$	267 ± 66	226 ± 79
	WM	129 ± 56	$\mathbf{165 \pm 110}$	208 ± 61	209 ± 72	$\mathbf{188 \pm 85}$
	GM	116 ± 66	192 ± 149	267 ± 161	$\mathbf{61 \pm 55}$	$\mathbf{111 \pm 73}$
	vessel	11 ± 12	$\mathbf{27 \pm 19}$	$\mathbf{28 \pm 18}$	50 ± 23	59 ± 26
	mean	102 ± 85	$\mathbf{114 \pm 116}$	153 ± 133	147 ± 110	$\mathbf{146 \pm 95}$

TABLE 3. Comparison between the our model and Manniesing et al. [22] on Dice coefficient (DSC), contour mean distance (CMD) in mm, and absolute volume difference (AVD) in %. First three rows show segmentation scores for white matter (WM) and gray matter (GM). The final row shows average computation time of the segmentation per model. * indicates a $p < 0.05$. Bold values indicate best performance between models per metric and per class.

	Ours		Manniesing et al. [22]	
	WM	**GM**	**WM**	**GM**
DSC	$\mathbf{0.82 \pm 0.07}$	$\mathbf{0.81 \pm 0.04}^*$	0.81 ± 0.04	0.79 ± 0.05
CMD (mm)	$\mathbf{0.94 \pm 0.40}^*$	$\mathbf{0.57 \pm 0.30}^*$	1.35 ± 0.26	0.74 ± 0.19
AVD (%)	$\mathbf{13.54 \pm 11.83}$	$\mathbf{9.80 \pm 6.62}^*$	15.83 ± 10.85	16.48 ± 11.16
Time	$\pm$**5 mins***		$\pm$60 mins	

E. ABLATION EXPERIMENTS

The ablation results are listed in Table 5.Paired t-tests showed significant differences, for all tissue types and all metrics, between our model and our model without additional shortcuts over 3^3 convolution pairs and between our model and our model with ReLU instead of very leaky ReLU activation functions ($p < 0.05$). However, the tests showed no significant differences between our model and our model with Xavier initialization instead of He initialization ($p > 0.05$).

VI. DISCUSSION

We have presented a fully convolutional multiclass deep learning architecture for 3D segmentation which can learn end-to-end from sparsely annotated 4D data. The method gives high quality segmentations of WM, GM, CSF, and vessels in 4D CT, approximating the interobserver agreement and outperforms the current state-of-the-art.

The experimental results (Table 2 and Table 4) highlight that our model significantly outperforms 3D U-Net with respect to the DSC and CMD. This is likely to be the combined contribution of the additional shortcuts, and the very leaky ReLU activation function (Table 5). The additional shortcuts are thought to simplify learning by allowing information to directly skip the 2^3 convolutions pairs throughout the network. The very leaky ReLU activation function is thought to work better since it avoids 'dead' ReLU, which is a state of a normal ReLU which always outputs zero and is unlikely to break out of this state. Finally, He initialization was thought to work better than Xavier initialization, since it has been optimized for the ReLU function, which we use throughout the network. However, there was no significant improvement found from the ablation experiments (Table 5).

Our model slightly, but significantly, outperforms the current state-of-the-art method by Manniesing et al. [22] on WM and GM segmentation with respect to the DSC, AVD, and CMD (see Table 3), without any training or optimization on the dataset used to train the competing method and with our model being trained on less annotated training slices. Despite these disadvantages our model significantly outperformed the competing state-of-the-art method. If such measures were taken the model is expected to perform even better. Furthermore, at prediction time, our model can be run on a GPU

TABLE 4. Quantitative segmentation results on the extended reference standard for our model and 3D U-Net. The Dice coefficient (DSC), contour mean distance (CMD) in mm, absolute volume difference (AVD) in %, and mean volume difference (MVD) in mm^3 were used for which the mean and standard deviation were calculated for the ten annotated cases in test set B per tissue type. Paired *t*-tests showed significant differences between models $p < 0.05$. Bold values indicate best performance between models per metric, and per class.

		our model	3D U-Net
DSC	CSF	$\mathbf{0.82 \pm 0.07}$	0.67 ± 0.14
	WM	$\mathbf{0.85 \pm 0.02}$	0.78 ± 0.03
	GM	$\mathbf{0.82 \pm 0.02}$	0.75 ± 0.03
	vessel	$\mathbf{0.81 \pm 0.05}$	0.51 ± 0.17
	mean	$\mathbf{0.82 \pm 0.04}$	0.68 ± 0.15
CMD (mm)	CSF	$\mathbf{0.61 \pm 0.27}$	1.73 ± 1.01
	WM	$\mathbf{0.92 \pm 0.31}$	1.88 ± 0.54
	GM	$\mathbf{0.66 \pm 0.19}$	0.88 ± 0.29
	vessel	$\mathbf{0.72 \pm 0.39}$	6.82 ± 2.47
	mean	$\mathbf{0.73 \pm 0.30}$	2.83 ± 2.71
AVD (%)	CSF	$\mathbf{10.57 \pm 4.71}$	62.68 ± 48.45
	WM	$\mathbf{13.23 \pm 7.60}$	23.82 ± 12.62
	GM	19.29 ± 11.81	$\mathbf{14.56 \pm 11.27}$
	vessel	$\mathbf{15.00 \pm 9.41}$	44.18 ± 21.60
	mean	$\mathbf{14.52 \pm 8.69}$	36.31 ± 31.90
MVD (mm^3)	CSF	$\mathbf{108 \pm 61}$	509 ± 418
	WM	$\mathbf{237 \pm 118}$	487 ± 358
	GM	307 ± 198	$\mathbf{210 \pm 135}$
	vessel	$\mathbf{18 \pm 12}$	51 ± 23
	mean	$\mathbf{168 \pm 159}$	314 ± 327

TABLE 5. Ablation experiment segmentation results on Dice coefficient (DSC), contour mean distance (CMD) in mm, absolute volume difference (AVD) in %, and mean volume difference (MVD) in mm^3. Reported values are averages over all tissue classes. From left to right: our best performing model (Ours), our model with Xavier instead of He initialization (Ours-Xavier), our model with additional shortcuts over 3^3 convolution pairs removed (Ours-no skip), and our model with ReLU instead of very leaky ReLU activation functions (Ours-ReLU). * indicates a significant difference between the average metric score for our model and the average metric score of the ablated model ($p < 0.05$). Bold values indicate best performance between models per metric.

	Ours	Ours-Xavier	Ours-no skip	Ours-ReLU
DSC	$\mathbf{0.82 \pm 0.04}$	0.80 ± 0.11	$0.81 \pm 0.06^*$	$0.70 \pm 0.12^*$
CMD	$\mathbf{0.73 \pm 0.30}$	1.27 ± 1.29	$0.93 \pm 0.51^*$	$1.55 \pm 1.22^*$
AVD	$\mathbf{14.52 \pm 8.69}$	17.23 ± 18.29	$21.94 \pm 11.59^*$	$43.77 \pm 52.79^*$
MVD	168 ± 159	$\mathbf{99 \pm 74}$	$268 \pm 253^*$	$469 \pm 564^*$

within 5 minutes whereas the competing method, which can not be easily GPU optimized, takes hours to compute on multiple CPUs. Additionally, our model provides CSF and vessels segmentations that were learned directly from 4D CT data opposed to unevaluated segmentation methods based on simple heuristics.

The quantitative results (see Table 2) approximate the interobserver overlap for the model compared to observer 1. However, while still having good overlap, the results are significantly inferior for the model compared to observer 2. This difference might be a result of the fact that two out of three training cases were annotated by observer 1. In other words there was an observer imbalance of the training set. Another explanation is that observer 1 had significantly lower intraobserver variability, which may have resulted in easier cases for the model to generalize to.

The scores from the extended evaluation on the test set B (see Table 4) are overall in line with the findings on test set A (see Table 2).3D U-Net significantly performed worse than our model. However, the results on test set B are worse on average than those on test set A. We suspect that to be the case because the test set B had more difficult cases than test set A. Because of this, the previously obtained interobserver overlap on test set A cannot be fairly compared with the new scores, since the previous indications are expected to be overly optimistic regarding the more difficult cases. Furthermore, we cannot compute new indications based on a single observer.

We emphasize that annotating brain tissues of 4D CT is a very difficult task for humans. Even though the observers had access to 3D data while annotating, in practice they tend to focus mainly on a single 2D cross section, which may introduce an annotation bias. Visualization of test set A predictions (see Figure 5) generally resulted in a good approximation of the reference standard. In the axial cross sections, some GM oversegmentation occurred with respect to the reference standard, but the model predictions seem to better match the underlying anatomy presented in the temporal average image.

The dataset used for the evaluation consisted exclusively of normal appearing brain tissues without pathology or foreign objects, which are seen in everyday clinical practice. The data was collected as such to focus on testing the feasibility of segmentation of WM/GM/CSF and vessels in 4D CT using deep learning, which is traditionally the domain of MR imaging. This implies that the method likely must be trained on cases with pathology or foreign objects and at least be evaluated on such cases, before it can be used in practice. However, we argue that our method provides a valuable first step towards this goal in the next paragraph.

In principle, the architecture is not limited to normal tissues. It can easily be extended to include tissue classes for pathology or foreign objects, such as core, penumbra, bleedings, clips, drains, calcifications, and bone if sufficiently annotated data for each class is collected. Furthermore, The presented method can be easily applied in a semi-supervised way on a cohort of pathology cases to get novel segmentations, which would yield most likely correct segmentations in healthy tissue areas but would have many errors near pathology or artifacts. Hence, expert observers could subsequently refine the segmentations for use as a novel reference standard, which in turn can be used to train and improve the model to reliably and verifiably deal with pathology as well. We intend to address these issues in future work by adding more annotated patient scans to the dataset using the method described above, which will include pathology and foreign objects and will be from different scanners and acquisition protocols.

Our model has a straightforward design expecting a fixed number of time points, which works well for data from standardized acquisition protocols. However, dealing with variable number of time points might be desirable in some cases. In this case interpolation could be used, or recurrent

layers could be used. While our model can be expanded with additional layers, filters and shortcut connections to enable learning a richer set of problem relevant feature maps. However, this remains technically challenging due to the total GPU RAM memory requirements for these experiments. Although these changes may improve the stability of the training process, it was not possible to increase the batch size to more than 2 or to increase our input spatial dimensions (picking a larger value for n, Figure 1) much further without altering the network. In the future, we would like to distribute our computations across multiple GPUs to cope with the memory requirements and scale to larger networks, which might involve switching to other deep learning frameworks.

Many aspects of the network architecture contributed to a successful deep learning model, like the number of multi-resolution levels, the number of feature maps and the size of the filter kernels. our model has three max-pooling and upscaling operations, which provide feature extraction at four different resolution levels. The number of resolution levels can be changed by removing or adding a pair of max-pooling and upscaling operations and a long shortcut connection at a particular resolution. Generally, more resolution levels result in a bigger receptive field and hence each voxel can infer its class from a broader volume of surrounding context, but the minimal required input size increases. For example, increasing the resolution level from four to five results in respectively minimum required input subvolume sizes of 92^3 and 188^3, which would also increase the memory requirement by more than a factor of eight and would no longer fit in GPU RAM.

The number of feature maps could only be slightly increased due to memory limitations, but early experiments did not give significant performance increase. Increasing the feature maps at the earlier layers is especially troublesome, since the resulting intermediate feature maps take a lot of GPU memory. This problem is less expressed at the lower resolution layers, where each feature map uses approximately eight times less memory than a feature map at a previous resolution level. The minimal required feature maps for achieving similar performance was not investigated due to the required computation times, but it is expected that reducing the number of feature maps will at some point have a big effect on the performance.

The size of the filter kernels can be varied, but can be difficult to optimize, since it holds a close relation to the receptive field and therefore also the minimum input size for the network. Increasing all filter sizes from 3^3 to 5^3 for example requires much larger input subvolumes, which would not fit in GPU RAM anymore. Another approach would be to replace every pair of 3^3 filters by a single 5^5 filter, which effectively leaves the receptive field the same and reduces intermediate feature map computations at the cost of an extra non-linearity. Doing this properly involved lowering the initial learning rate.

Setting the training hyper parameters – like batch size, input size, optimizer choice, and optimizer parameters – were found to be at least as important as the network architecture to achieve good model performance. In our experience, changing the batch size and input size, had a great impact on training and final model performance. Generally taking the batch size and input size as big as possible while still being able to fit in GPU RAM memory worked best. For the optimizer we have only experimented with default stochastic gradient descent with momentum. We did not test with other optimizers, but they might require some tweaking. We do not expect big performance increases from using different optimizer. Tweaking of the optimizer parameters, like the learning rate and momentum factor in our experience can have a big impact on training and final model performance.

Whereas predicting with our model is relatively fast (approximately 5 minutes for a full 4D CT case), the end-to-end training of the network could take several days. Hence, only a limited amount of experiments could be performed for this study. We parallelized our experiments over multiple Titan X GPUs to speed up training. Additionally, we split our training and validation computations per experiment and distributed these over different GPUs. We simultaneously used the CPU to prepare subvolumes while training on another subvolume on the GPU, to ensure the best possible continuity of input data. Furthermore, for validation we increased the input subvolume size to better utilize the GPU memory and predict slightly larger subvolumes, which also sped up the process significantly. It might be possible to further reduce training times through deep supervision approaches [42], by reducing some of the complexity of the model or by implementing more efficient data sampling schemes.

There is not much literature on CNNs with respect to handling 4D or higher dimensional data; yet, it is the opinion of the authors that deep learning approaches which are able to cope with high dimensional data will become increasingly important as datasets increase in size and incorporate more dimensions. Hence, the competitive segmentation results achieved by our proposed method, which was directly learned from 4D CT input, suggests potential application of the method beyond the application of stroke imaging.

ACKNOWLEDGMENT

The authors would like to thank Lisan Kaal, BSc. and Loes Vos, BSc. (medical students at Radboud UMC, the Netherlands) for annotating. They would also like to thank Anton Schreuder, MSc. MD and Jonas Teuwen PhD, for proofreading this manuscript.

REFERENCES

[1] T. Brosch, L. Y. W. Tang, Y. Yoo, D. K. B. Li, A. Traboulsee, and R. Tam, "Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1229–1239, May 2016.

[2] C. Chen, A. Bivard, L. Lin, C. R. Levi, N. J. Spratt, and M. W. Parsons, "Thresholds for infarction vary between gray matter and white matter in acute ischemic stroke: A CT perfusion study," *J. Cereb. Blood Flow Metab.* to be published.

[3] H. Chen, Q. Dou, X. Wang, J. Qin, J. C. Cheng, and P.-A. Heng, "3D fully convolutional networks for intervertebral disc localization and segmentation," in *Proc. Int. Conf. Med. Imag. Virt. Real.*, Aug. 2016, pp. 375–382.

[4] J. Chen, L. Yang, Y. Zhang, M. S. Alber, and D. Z. Chen. (2016). "Combining fully convolutional and recurrent neural networks for 3D biomedical image segmentation." [Online]. Available: https://arxiv.org/abs/1609.01006
[5] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. (2016). "3D U-Net: Learning dense volumetric segmentation from sparse annotation." [Online]. Available: https://arxiv.org/abs/1606.06650
[6] F. Ciompi *et al.*, "Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box," *Med. Image Anal.*, vol. 26, no. 1, pp. 195–202, Dec. 2015.
[7] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.
[8] S. Dieleman *et al.*, "Lasagne: First release," Zenodo, Geneva, Switzerland, Tech. Rep., Aug. 2015. doi: 10.5281/zenodo.27878.
[9] Q. Dou *et al.*, "Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1182–1195, May 2016.
[10] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal. (2016). "The importance of skip connections in biomedical image segmentation." [Online]. Available: https://arxiv.org/abs/1608.04117
[11] M. Ghafoorian *et al.*, "Non-uniform patch sampling with deep convolutional neural networks for white matter hyperintensity segmentation," in *Proc. IEEE Int. Symp. Biomed. Imaging*, Apr. 2016, pp. 1414–1417.
[12] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, May 2010, pp. 249–256.
[13] B. Graham. (2014). "Spatially-sparse convolutional neural networks." [Online]. Available: https://arxiv.org/abs/1409.6070
[14] K. He, X. Zhang, S. Ren, and J. Sun. (2015). "Deep residual learning for image recognition." [Online]. Available: https://arxiv.org/abs/1512.03385
[15] S. Ioffe and C. Szegedy. (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift." [Online]. Available: https://arxiv.org/abs/1502.03167
[16] K. Kamnitsas *et al.*, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Med. Image Anal.*, vol. 36, pp. 61–78, Feb. 2017.
[17] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim, "Elastix: A toolbox for intensity-based medical image registration," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 196–205, Jan. 2010.
[18] R. Korez, B. Likar, F. Pernuš, and T. Vrtovec, "Model-based segmentation of Vertebral bodies from MR images with 3D CNNs," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, Oct. 2016, pp. 433–441.
[19] D. B. Larson, L. W. Johnson, B. M. Schnell, S. R. Salisbury, and H. P. Forman, "National trends in CT use in the emergency department: 1995–2007," *Radiology*, vol. 258, no. 1, pp. 164–173, 2011.
[20] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[21] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.
[22] R. Manniesing *et al.*, "White matter and gray matter segmentation in 4D computed tomography," *Sci. Rep.*, vol. 7, no. 1, p. 119, Mar. 2017.
[23] M. Meijs, S. Pegge, M. Prokop, B. van Ginneken, F. J. A. Meijer, and R. Manniesing, "Detection of vessel occlusion in acute stroke is facilitated by color-coded 4D-CTA," in *Proc. Eur. Congr. Radiol.*, 2017.
[24] M. Meijs *et al.*, "Robust segmentation of the full cerebral vasculature in 4D CT of suspected stroke patients," *Sci. Rep.*, vol. 7, no. 1, p. 15622, 2017.
[25] F. Milletari, N. Navab, and S. Ahmadi. (2016). "V-Net: Fully convolutional neural networks fors, volumetric medical image segmentation." [Online]. Available: https://arxiv.org/abs/1606.04797
[26] P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. de Vries, M. J. N. L. Benders, and I. Išgum, "Automatic segmentation of MR brain images with a convolutional neural network," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1252–1261, May 2016.
[27] Y. Nesterov, "A method of solving a convex programming problem with convergence rate $O(1/k^2)$," *Soviet Math. Doklady*, vol. 27, no. 2, pp. 372–376, 1983.
[28] R. G. Nogueira *et al.*, "Thrombectomy 6 to 24 hours after stroke with a mismatch between deficit and infarct," *New England J. Med.*, vol. 378, no. 1, pp. 11–21, 2017.
[29] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, p. e3, 2016.
[30] A. Patel, B. van Ginneken, F. J. A. Meijer, E. J. van Dijk, M. Prokop, and R. Manniesing, "Robust cranial cavity segmentation in CT and CT perfusion images of trauma and suspected stroke patients," *Med. Image Anal.*, vol. 36, pp. 216–228, Feb. 2016.
[31] R. P. Poudel, P. Lamata, and G. Montana. (2016). "Recurrent fully convolutional neural networks for multi-slice MRI cardiac segmentation." [Online]. Available: https://arxiv.org/abs/1608.03974
[32] W. J. Powers *et al.*, "2015 American heart association/American stroke association focused update of the 2013 guidelines for the early management of patients with acute ischemic stroke regarding endovascular treatment," *Stroke*, vol. 46, no. 10, pp. 3020–3035, 2015.
[33] M. Prokop, M. Galanski, and C. Schaefer-Prokop, *Spiral Multislice Computed Tomography Body*. Stuttgart, Germany: Thieme Medical, 2003.
[34] K. A. Quaday, J. G. Salzman, and B. D. Gordon, "Magnetic resonance imaging and computed tomography utilization trends in an academic ED," *Amer. J. Emergency Med.*, vol. 32, no. 6, pp. 524–528, Jun. 2014.
[35] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2015, pp. 234–241.
[36] A. A. A. Setio *et al.*, "Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks," *IEEE Trans Med. Imag.*, vol. 35, no. 5, pp. 1160–1169, May 2016.
[37] M. Shakeri *et al.*, "Sub-cortical brain structure segmentation using F-CNN'S," in *Proc. IEEE Int. Symp. Biomed. Imaging*, Apr. 2016, pp. 269–272.
[38] H.-C. Shin, M. R. Orton, D. J. Collins, S. J. Doran, and M. O. Leach, "Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1930–1943, Aug. 2013.
[39] E. J. Smit *et al.*, "Timing-invariant imaging of collateral vessels in acute ischemic stroke," *Stroke*, vol. 44, no. 8, pp. 2194–2199, Aug. 2013.
[40] Y. Song, L. Zhang, S. Chen, D. Ni, B. Lei, and T. Wang, "Accurate segmentation of cervical cytoplasm and nuclei based on multiscale convolutional network and graph partitioning," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 10, pp. 2421–2433, Oct. 2015.
[41] M. F. Stollenga, W. Byeon, M. Liwicki, and J. Schmidhuber. (2015). "Parallel multi-dimensional LSTM, with application to fast biomedical volumetric image segmentation." [Online]. Available: https://arxiv.org/abs/1506.07452
[42] C. Szegedy *et al.* (2014). "Going deeper with convolutions." [Online]. Available: https://arxiv.org/abs/1409.4842
[43] A. Talwalkar and S. Uddin, *Trends Emergency Dept. Visits for Ischemic Stroke Transient Ischemic Attack: United States, 2001-2011*. Hyattsvill, MA, USA: National Center for Health Statistics, 2015, p. 194.
[44] T. D. Team. (2016). "Theano: a Python framework for fast computation of mathematical expressions." [Online]. Available: https://arxiv.org/abs/1605.02688
[45] S. C. van de Leemput, F. J. A. Meijer, M. Prokop, and R. Manniesing, "Cerebral white matter, gray matter and cerebrospinal fluid segmentation in CT using VCAST: A, volumetric cluster annotation and segmentation tool," in *Proc. Eur. Congr. Radiol.*, 2017, pp. 1–10.
[46] Y. Xie, Z. Zhang, M. Sapkota, and L. Yang, "Spatial clockwork recurrent neural network for muscle perimysium segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, vol. 9901, 2016, pp. 185–193. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-46723-8_22
[47] L. Zhao and K. Jia, "Multiscale CNNs for brain tumor segmentation and diagnosis," *Comput. Math. Methods Med.*, vol. 2016, Aug. 2016, Art. no. 123569.

SIL C. VAN DE LEEMPUT received the B.Sc. and M.Sc. degrees (Hons.) in artificial intelligence from the Radboud University Medical Center, Nijmegen, in 2013 and 2015, respectively, where he is currently pursuing the Ph.D. degree with the Diagnostic Image Analysis Group with a focus on computer-aided diagnosis in acute stroke.

His research interests include stroke imaging, machine learning, and computer-aided diagnosis.

MIDAS MEIJS received the B.Sc. degree in biomedical engineering and the M.Sc. degree in medical engineering from the Eindhoven University of Technology (TU/e). He is currently pursuing the Ph.D. degree with the Diagnostic Image Analysis Group, Radboud University Medical Center, with a focus on computer-aided diagnosis in acute stroke.

AJAY PATEL received the B.Sc. degree in biomedical sciences and the M.Sc. degree in biomedical image sciences from the University of Utrecht. He is currently pursuing the Ph.D. degree with the Diagnostic Image Analysis Group, Radboud University Medical Center, with a focus on computer-aided diagnosis in acute stroke.

FREDERICK J. A. MEIJER has been a Radiologist with the Department of Radiology, Radboud University Medical Center, since 2010, where he is involved in neuro-, head, and neck and emergency radiology. He participates in clinical and research projects in imaging in neurovascular and neurodegenerative diseases.

BRAM VAN GINNEKEN received the degrees in physics from the Eindhoven University of Technology and Utrecht University, and the Ph.D. degree from the Image Sciences Institute on Computer-Aided Diagnosis in Chest Radiography. He is currently a Professor of medical image analysis with the Radboud University Medical Center, where also chairs the Diagnostic Image Analysis Group. He is also with Fraunhofer MEVIS, Bremen, Germany, and is the Founder of Thirona, a company that develops software and provides services for medical image analysis. He pioneered the concept of challenges in medical image analysis. He has authored/coauthored over 200 publications in international journals. He is an Associate Editor of the IEEE TRANSACTIONS ON MEDICAL IMAGING and a member of the Editorial Board of *Medical Image Analysis*.

RASHINDRA MANNIESING received the master's degree in electrical engineering from the Delft University of Technology, in 1999, and the Ph.D. degree in radiology from Utrecht University, University Medical Center Utrecht, in 2006. He is currently an Assistant Professor leading the Neuroimaging Group, Diagnostic Image Analysis Group, Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Nijmegen, The Netherlands. His current research interests include deep learning in stroke and trauma imaging.