Neuroinformatics (2019) 17:563–582
https://doi.org/10.1007/s12021-019-09417-y

ORIGINAL ARTICLE

# Fast and Precise Hippocampus Segmentation Through Deep Convolutional Neural Network Ensembles and Transfer Learning

**Dimitrios Ataloglou[1] · Anastasios Dimou[1] · Dimitrios Zarpalas[1] · Petros Daras[1]**

Published online: 15 March 2019


**Abstract**

Automatic segmentation of the hippocampus from 3D magnetic resonance imaging mostly relied on multi-atlas registration methods. In this work, we exploit recent advances in deep learning to design and implement a fully automatic segmentation method, offering both superior accuracy and fast result. The proposed method is based on deep Convolutional Neural Networks (CNNs) and incorporates distinct segmentation and error correction steps. Segmentation masks are produced by an ensemble of three independent models, operating with orthogonal slices of the input volume, while erroneous labels are subsequently corrected by a combination of Replace and Refine networks. We explore different training approaches and demonstrate how, in CNN-based segmentation, multiple datasets can be effectively combined through transfer learning techniques, allowing for improved segmentation quality. The proposed method was evaluated using two different public datasets and compared favorably to existing methodologies. In the EADC-ADNI HarP dataset, the correspondence between the method's output and the available ground truth manual tracings yielded a mean Dice value of 0.9015, while the required segmentation time for an entire MRI volume was 14.8 seconds. In the MICCAI dataset, the mean Dice value increased to 0.8835 through transfer learning from the larger EADC-ADNI HarP dataset.

**Keywords** Hippocampus segmentation · Convolutional neural networks · Deep learning · Error correction · Transfer learning · Magnetic resonance imaging

## Introduction

Medical studies have proven that there is a close relationship between the hippocampus and memory function (Scoville and Milner 2000). Volume reduction and morphological degeneration of the hippocampus have been associated with the existence of neurological diseases, such as the Alzheimer's disease and other forms of dementia (Du et al. 2001). Additionally, patients with extensive hippocampal damage may suffer from depression (Bremner et al. 2000), epilepsy (Bernasconi et al. 2003) or schizophrenia (Harrison 2004).

Structural and volumetric analysis of the hippocampus can aid clinicians in the diagnosis and early detection of related pathologies (Jack et al. 2011). Thus, the hippocampus has been the subject of several longitudinal studies and medical research projects (Leung et al. 2010; Bateman et al. 2012). Analysis of the hippocampus is usually performed using magnetic resonance imaging (MRI) of the brain (Fig. 1). While manual segmentation of the hippocampus by specially trained human raters is considered to be the gold standard, it is also laborious and expensive. Furthermore, manual segmentations are susceptible to inter-rater and intra-rater variability.

The aforementioned limitations of manual tracing highlight the need for automated segmentation methods, especially when dealing with large datasets. Over the last years, automatic segmentation of the hippocampus attracted great scientific and research interest. As a result, many

✉ Dimitrios Ataloglou
ataloglou@iti.gr

Anastasios Dimou
dimou@iti.gr

Dimitrios Zarpalas
zarpalas@iti.gr

Petros Daras
daras@iti.gr

[1] Information Technologies Institute (ITI), Centre for Research and Technology HELLAS, 1st km Thermi - Panorama, 57001, Thessaloniki, Greece

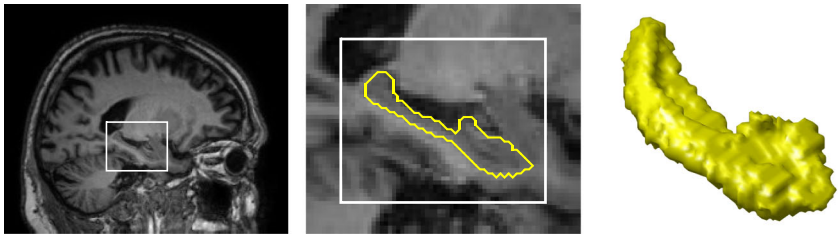

**Fig. 1** A sagittal brain MRI slice (left), zoomed hippocampal region with the outline of the left hippocampus depicted in yellow (middle) and 3D reconstruction of the left hippocampus (right)

different approaches have been proposed, which can be classified into different categories.

The most popular one is based on multi-atlas registration and fusion techniques. First, multiple atlas images are registered (usually non-linearly) to the new image. The computed transformations are then applied to the manual segmentation masks, producing one separate segmentation for each used atlas. Methods differ in terms of total number of used atlases and their selection method. Individual segmentations are combined to a final result using a variety of fusion techniques, such as majority or average voting (Collins and Pruessner 2010), use of global (Langerak et al. 2010) or local (Coupé et al. 2010) weights, joint label fusion (Wang et al. 2013) and accuracy maps (Sdika 2010). Multi-atlas methods are robust to anatomical variability, but their performance depends on registration quality and segmentation time increases linearly with the number of registrations.

Another category is based on Active Contour Models (ACM), which evolve according to the intensities of the image. Their performance depends largely on the existence of clear edges at the boundaries of the segmented object. To produce better segmentations, Active Shape Models incorporate prior knowledge about the shape of the structure to the evolution of the contour (Shen et al. 2002). In Yang et al. (2004), the shapes of both the hippocampus and neighbouring structures were modeled using Principal Component Analysis and a set of manually segmented atlases.

A combination of the multi-atlas framework with ACM was proposed in Zarpalas et al. (2014a). The method was based on 3D Optimal Local Maps (OLMs), which locally control the influence that image information and prior knowledge should have at a voxel level. The OLMs were built using an extended multi-atlas concept. In Zarpalas et al. (2014b), the ACM evolution was controlled again through a voxel-level map. Blending of image information and prior knowledge was based on three-phase Gradient Distribution on Boundary maps, having one phase for the strong edge boundary parts, where image gradient information is to be trusted, a second one for the blurred/noisy boundaries, where image regional information is to be trusted, and a third one for the missing boundaries, where shape prior knowledge should take the lead to influence the overall ACM.

The last category of segmentation methods is based on machine learning (Morra et al. 2010; Tong et al. 2013). Typical methods of this category include the training of a classifier, based on a training set of atlases. Conventional machine learning methods usually extract a set of hand-crafted features from each training instance to construct a training dataset, which is then used to optimize the classifier's parameters. The same set of features is then extracted from each new image and fed to the trained classifier, which produces the segmentation mask.

Convolutional Neural Networks (CNNs) have been proposed as a new type of classifier (LeCun et al. 1998). In correspondence to conventional neural networks, CNNs also consist of interconnected neurons, organized in successive layers. However, in CNNs, neurons of a single layer share trainable parameters and are usually connected only with a subset of the previous layer's neurons, thus having a limited field of view. The potential value of CNNs became evident when Krizhevsky et al. (2012) won the 2012 ImageNet Large Scale Visual Recognition Challenge (Russakovsky et al. 2015). Since then, CNNs have been utilized to address a variety of image processing problems. Advances in the field allowed the training of much deeper networks with better distinctive capabilities (He et al. 2016).

Recently, deep learning methods have been applied in the medical domain, including various applications in structural brain MRI. More specifically, CNNs have been utilized for the segmentation of brain tissue (Moeskops et al. 2016; Zhang et al. 2015; Chen et al. 2016), tumor (Havaei et al. 2017; Pereira et al. 2016) and lesions (Kamnitsas et al. 2017; Brosch et al. 2016). Segmentation of anatomical structures using CNNs have been studied in Choi and Jin (2016) for the striatum and in Shakeri et al. (2016) and Dolz et al. (2017) for the thalamus, caudate, putamen and pallidum. CNN-based segmentation of basal ganglia (including the hippocampus) has been also explored in Milletari et al. (2017), Wachinger et al. (2017), and Kushibar et al. (2017). Finally, CNNs have been used for brain extraction (Kleesiek et al. 2016) and full brain segmentation (de Brébisson and Montana 2015; Mehta et al. 2017).

In this work, we leverage the unique properties of CNNs to design and train a fully automatic segmentation method, aiming to provide superior segmentation accuracy compared to previous methods, while substantially reducing the required segmentation time. Compared to existing CNN-based segmentation methods, the proposed method differs in various ways. A comparative analysis is presented in the remainder of this section, focusing mainly on methods that segment the hippocampus and other anatomical brain structures.

Most previous methods performed the segmentation in a single step (de Brebisson and Montana 2015; Mehta et al. 2017; Kushibar et al. 2017). In contrast, we split the processing pipeline into distinct stages and include an error correction mechanism to improve the overall performance. In the first stage, a segmentation mask of both hippocampi is computed. Erroneous labels are subsequently corrected by an independent module. The processing pipeline in Choi and Jin (2016) also involved two stages. However, in that case the first stage operated in lower resolution and only performed an approximate localization of the structure, which was segmented in the second stage. In contrast, our first stage completes both tasks at once and provides high quality segmentation masks to the error correction module.

Contrary to methods that use 3D patches as input to a single CNN (Wachinger et al. 2017), in each processing stage of the proposed method, the MRI volume is decomposed into orthogonal slices, which are fed to an ensemble of three independent CNNs. The final 3D segmentation is obtained by fusing the individual segmentations. Orthogonal slices have also been used in de Brébisson and Montana (2015), Mehta et al. (2017), and Kushibar et al. (2017), but fusion in these cases was performed within the CNN, before the final classification layers. Using a different approach, Milletari et al. (2017) stacked the three orthogonal 2D patches to form a 3-channel input, which was then processed by a single CNN. In our architecture, we perform a late fusion of the outputs of independent CNNs, which we train separately, allowing them to optimize better for each slicing operation. Furthermore, by using model ensembles we manage to improve the segmentation quality and eliminate spatial inconsistencies without the need of complicated post-processing, as was the case in Shakeri et al. (2016), Wachinger et al. (2017), and Milletari et al. (2017), where Conditional Random Fields and Hough voting where utilized for such purposes.

In general, previously proposed CNN-based medical image segmentation methods used shallower networks, with fewer convolutional layers (de Brébisson and Montana 2015; Wachinger et al. 2017; Mehta et al. 2017). While easier to train, these lower capacity models exhibit inferior distinctive capabilities and may prove inadequate for the segmentation of complex structures, such as the hippocampus. Deeper CNNs can generalize better to unseen cases, but their training can suffer from vanishing or exploding gradients (Glorot and Bengio 2010). To overcome such difficulties, we make extensive use of Residual Blocks (He et al. 2016) and Batch Normalization layers (Ioffe and Szegedy 2015) in all CNNs. This enables us to design and effectively train deeper networks, achieving higher segmentation quality.

Lastly, we explore and validate the importance of transfer learning in the medical domain, where large annotated datasets are rare. In contrast to other segmentation methods, CNNs inherently require a sufficiently large training dataset, in order to be able to properly generalize to previously unseen cases (Sun et al. 2017). When sufficiently large datasets are not available, the performance of CNNs can be improved by exploiting transfer learning techniques (Razavian et al. 2014; Yosinski et al. 2014). In our case, we were able to utilise the CNNs trained with the EADC-ADNI Harp dataset and fine-tune them to the smaller MICCAI dataset. Despite the fact that different manual segmentation protocols were used (HarP and brainCOLOR respectively), such combination of multiple datasets proved notably advantageous, leading to the conclusion that such strategies could greatly benefit the community, by offering robust and extendible automatic segmentation mechanisms.

## Materials

Two different open access datasets were used for training and testing, with a total of 135 atlases. Those datasets include cases with a wide variety of ages, medical diagnoses and MR scanners, therefore assisting in the demonstration of the robustness of our method.

### EADC-ADNI HarP Dataset

We used the preliminary version of the dataset provided by the EADC-ADNI Harmonized Protocol project (Boccardi et al. 2015), consisting of 100 3D MP-RAGE T1-weighted MRIs and the corresponding segmentation masks of both hippocampi. This dataset includes MRIs from elderly people, with four different medical diagnoses: healthy (Normal), Mild Cognitive Impairment (MCI), Late Mild Cognitive Impairment (LMCI) and Alzheimer's Disease (AD). A total of 15 different MR scanners (1.5 and 3.0 Tesla) were used for the acquisition of the images. The distribution of participants per clinical and demographic characteristic is presented in Table 1.

All MRIs were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (Jack et al. 2008) and have the same dimensions of 197 × 233 × 183

**Table 1** Distribution of demographic characteristics per medical diagnosis in the HarP dataset

| Diagnosis | N | Gender | | Age | | | Scanner | |
|---|---|---|---|---|---|---|---|---|
| | | M | F | 60-70 | 70-80 | 80-90 | 1.5T | 3.0T |
| Normal | 29 | 16 | 13 | 4 | 17 | 8 | 17 | 12 |
| MCI | 21 | 13 | 8 | 7 | 6 | 8 | 13 | 8 |
| LMCI | 13 | 7 | 6 | 4 | 7 | 2 | 8 | 5 |
| AD | 37 | 20 | 17 | 13 | 14 | 10 | 20 | 17 |
| Total | 100 | 56 | 44 | 28 | 44 | 28 | 58 | 42 |

voxels, with a voxel size of 1 × 1 × 1 mm. Manual segmentations of both hippocampi were carried out by specially trained tracers using the Harmonized Protocol (HarP). Segmentation masks are freely available at http://www.hippocampal-protocol.net.

### MICCAI Dataset

The MICCAI dataset consists of 35 atlases and was first used at the MICCAI 2012 Grand Challenge on Multi-atlas Labeling (Landman and Warfield 2012). During the challenge, 15 of the atlases were available for training purposes, while the remaining 20 were used only for testing. The mean participant age is 23 years (ranging from 19 to 34) for the training set and 45 years (ranging from 18 to 90) for the test set. All participants were considered to be healthy.

All MRIs were obtained from the Open Access Series of Imaging Studies repository (Marcus et al. 2007). The original 3D MP-RAGE T1-weighted MRIs were acquired using a Siemens Vision MR scanner (1.5 Tesla). The MRIs used in the challenge have a voxel size of 1 × 1 × 1 mm and dimensions up to 256 × 334 × 256 voxels. Manual segmentations of 143 brain structures were carried out according to the brainCOLOR labeling protocol. The MICCAI dataset is publicly available at https://my.vanderbilt.edu/masi/workshops/.

### Preprocessing

Unlike multi-atlas segmentation methods, the proposed method does not require any additional registration of the MRIs, thus making it significantly lighter in terms of required computational time. The preprocessing procedure consists of three separate steps, which are common for both datasets.

First, brain extraction was performed using the Brain Extraction Tool (Smith 2002), which is available as part of the FMRIB Software Library. Besides the fact that non-brain information is not essential for the segmentation of internal structures, such as the hippocampus, this step was also necessary as specific non-brain regions contained voxels with arbitrarily high brightness values, which negatively affected the normalization process in the final preprocessing step.

Then, brain regions were corrected for intensity inhomogeneity with the N3 package of the MINC toolkit (Sled et al. 1998). Although inhomogeneity correction is carried out internally by software in most modern MR scanners, we include this step to account for older scanners and to further improve the preprocessing outcome.

Finally, we normalized the intensity of each voxel by subtracting the mean and dividing with the standard deviation. The mean and standard deviation values were calculated separately for each MRI, taking into account only the brain region. Voxels outside the brain region were assigned a zero value.

## Method

A top level diagram of the proposed architecture is presented in Fig. 2. The proposed method is composed of three separate modules. In the segmentation module, the input, which is a pre-processed 3D MRI, passes through a group of CNNs and a segmentation mask of the hippocampus is obtained. Subsequently, a wider region around the hippocampus is cropped, both form the mask produced by the segmentation module and the input MRI. Finally, the cropped MRI volume and segmentation mask are given as inputs to the error correction module, which corrects the erroneous labels and produces the final, error corrected mask.

### Segmentation

A schematic diagram of the segmentation module is presented in Fig. 3. We opted for an ensemble of three independent segmentation CNNs, operating with orthogonal

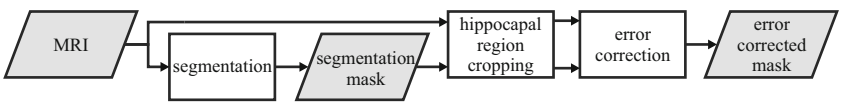


**Fig. 2** Top level architecture of the proposed method. The processing pipeline consists of three main modules: segmentation, hippocampal region cropping and error correction

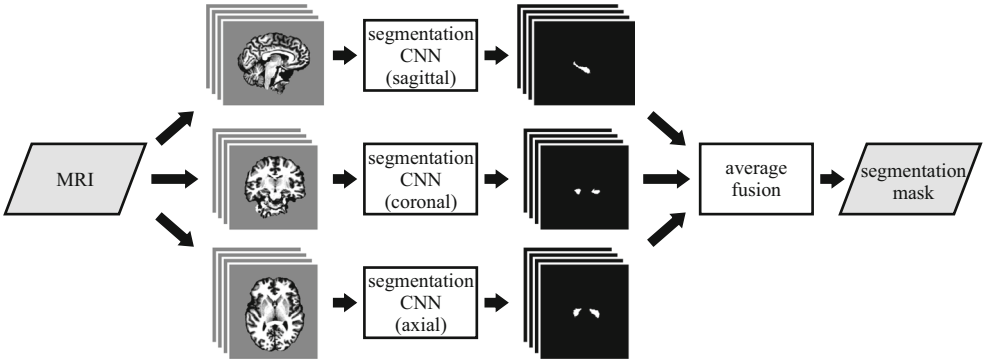


**Fig. 3** Segmentation module architecture. The 3D MRI is decomposed into orthogonal slices. Each type of slices is processed by an independent segmentation CNN. Average fusion combines all single slice outputs to a final 3D segmentation mask

slices of the input MRI volume, followed by a late fusion of their outputs. Both the input and output of the segmentation module are 3D images. However, all three segmentation CNNs used inside this module receive 2D slices as input and produce the corresponding 2D segmentations. In particular, the 3D MRI is decomposed into sagittal, coronal and axial slices. Each type of slices is then fed to an independent segmentation CNN, which was trained using only slices of the same type. The 2D outputs of each segmentation CNN are stacked along the third dimension to form a 3D segmentation mask. The output of the segmentation module is obtained by performing a voxel-wise average fusion of the individual 3D segmentation masks.

Orthogonal patches were used in combination with 3D patches in de Brébisson and Montana (2015). Authors claimed that using three 2D orthogonal patches is preferable over a single 3D patch. In their architecture, different input types were provided to separate branches within a single CNN and the intermediate results were concatenated inside the CNN, before the final fully-connected layers. In contrast, we use full orthogonal slices to train three independent segmentation CNNs and form an ensemble of models, each receiving a different input. Model ensembles have shown consistent performance benefits in other visual tasks (He et al. 2016; Szegedy et al. 2017), which we found to be also true for hippocampus segmentation.

Forming such a model ensemble would not be possible if the 3D MRI volumes were used as input to the CNNs, which is the main reason for preferring 2D CNN inputs. Other reasons supporting the preference of 2D inputs concern the training process. In particular, 3D inputs would significantly limit the number of independent training examples, while GPU memory requirements would impose a constraint on the maximum CNN depth, especially when working with complete images instead of patches (refer to "Training with the HarP Dataset" for more details). Also, the combination of multiple complete 3D volumes into training batches, which ensure smoother training, would only have been possible for very shallow CNNs with inferior performance.

### Segmentation CNNs

Sagittal, coronal and axial segmentation CNNs share a common internal structure, which is presented in Fig. 4. The input of each segmentation CNN is a 2D image (MRI slice), with spatial dimensions of $d_1 \times d_2$ pixels and one channel (grayscale image). The output is an image of equal dimensions with values in the range of [0, 1], which correspond to the probability of each pixel belonging to the hippocampus.

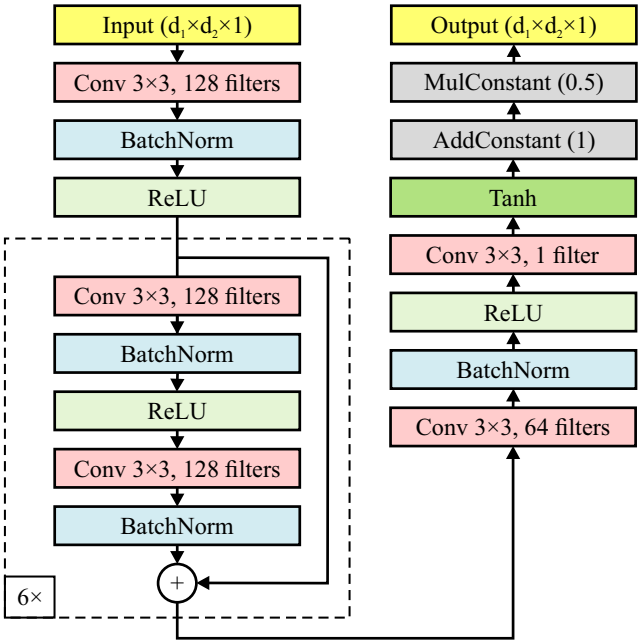


**Fig. 4** Internal structure of sagittal, coronal and axial segmentation CNNs. Each CNN is 15 convolutional layers deep and contains six consecutive residual blocks at its core

The core network consists of six consecutive Residual Blocks, depicted with a dashed bounding box in Fig. 4. Each block is composed of two parallel branches. The first branch includes two convolution layers. The second is a simple identity shortcut, forwarding the input of the block and adding it to the output of the first branch. Residual networks have been proven to be easier to train, as shortcut connections help to deal with the problem of vanishing gradients.

Each segmentation CNN has a total of 15 spatial convolutional layers, with $3 \times 3$ filters and additional bias parameters. Convolutions are always performed with single stride and zero-padding to maintain the spatial dimensions of the layer's input. Since no fully connected layers are used, segmentation CNNs are fully convolutional, which makes their evaluation with different input sizes much more efficient (Long et al. 2015).

In contrast to the U-Net architecture, which is commonly used in medical segmentation tasks (Ronneberger et al. 2015; Çiçek et al. 2016), our segmentation CNNs do not include any pooling layers. As a result, the spatial dimensions of the input image remain unaltered throughout the CNN. U-Net like models with equal depth have been evaluated at the initial stage of our research, but they produced consistently inferior results compared to the selected CNN structure ($-3\%$ in terms of Dice for individual segmentation CNNs when using two downsampling and upsampling operation, even when learned upsampling was used). We attribute this behavior to the nature of pooling operations, which suppress the input information into a more coarse representation, combined with the morphology of the structure of interest, which contains a high level of detail. While the U-Net has shown good performance in segmenting more arbitrarily shaped structures like tumors, the proposed CNN structure appears to be better suited for hippocampus segmentation.

With the exception of the last one, every convolutional layer is followed by a spatial batch normalization layer (BatchNorm). These layers improve the gradient flow, allow the usage of higher learning rates, minimize the effect of parameter initialization and act as a form of regularization. Batch normalization is skipped only after the last convolutional layer, since we do not want to alter the output's distribution.

A ReLU activation (Nair and Hinton 2010) is added after most BatchNorm layers, which is defined as $f(x) = max(0, x)$. In comparison with the sigmoid function, ReLU activations do not limit the output's range, are computationally cheaper and lead to faster convergence rates during training. The last convolutional layer is followed by a Tanh activation layer and subsequent addition and multiplication with constants to transfer the output's range to [0, 1].

Each segmentation CNN has a total of 1.86 million trainable parameters. The Field of View (FoV) at the output layer is $31 \times 31$ pixels, meaning that the value of each output pixel depends on a $31 \times 31$ region of the input, centered at that specific location (Fig. 5). Adding more residual blocks and therefore increasing the FoV size did not increase the performance, leading to the conclusion that the selected FoV is sufficiently large to capture all useful anatomical information around the hippocampus.

## Hippocampal Region Cropping

While significantly reducing the training and testing times of subsequent modules, the introduction of the hippocampal region cropping module to the processing pipeline leaves the overall segmentation accuracy unaffected, due to the carefully selected cropped region size. In order for the performance of subsequent CNNs to remain constant, the whole FoV must be covered, even at the endpoints of the hippocampus. As error correction CNNs have a similar structure to segmentation CNNs, with the same FoV size of $31 \times 31$ pixels, we adjust the cropped region to include at least 15 more voxels at every direction from the boundary of the hippocampus. Also, we crop a single region including both hippocampi, as separate regions usually overlapped, unnecessarily increasing the overall processing time.

During training, the calculation of the cropped region coordinates is based on the ground truth. Since the ground truth masks are available, the optimal crop position and size

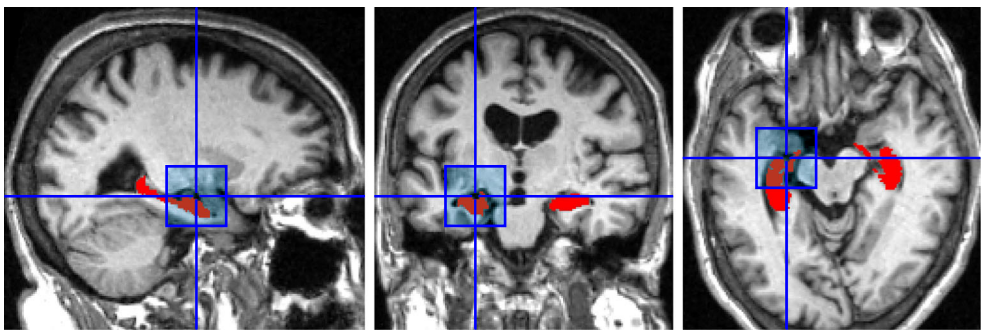

**Fig. 5** Field of View (FoV) size at the output layer of each segmentation CNN (blue square) in relation to a medium sized hippocampus (shown in red). Best viewed in color

for each training atlas can be calculated. However, we crop all training atlases to a common size, accounting for the largest hippocampus size in each dimension, in order to be able to combine slices from different MRIs to the same mini-batch during training.

In contrast, during testing the cropping procedure is based on the segmentation masks produced by the segmentation module of the proposed method. First, we calculate the weight center of the segmentation mask. Then, we crop a region of 120 × 100 × 100 voxels (along the sagittal, coronal and axial axes) around that center point, which is both sufficiently large to contain both hippocampi and the required area around them and small enough to lead to substantial performance benefits in terms of total processing time.

## Error Correction

Errors in automatic segmentation methods can be categorized to random and systematic. While random errors can be caused by noise or anatomical differences, systematic errors originate from the segmentation method itself and are repeated under specific conditions. Thus, systematic errors can be corrected using machine learning techniques. For example, a classifier may be built to identify the conditions under which systematic errors occur, estimate the probability of error for segmentations produced by a host method and correct the erroneous labels. Wang et al. (2011) proposed an error correction method for hippocampus segmentation, using a multi-atlas registration and fusion host method. Combined with joint label fusion, this error correction method won the MICCAI 2012 Grand Challenge on Multi-Atlas Segmentation.

CNNs have been successfully utilized for error correction purposes in the fields of human pose estimation (Carreira et al. 2016), saliency detection (Wang et al. 2016) and semantic image segmentation (Li et al. 2016). There are two different alternatives regarding the output of error correction CNNs. Replace CNNs calculate new labels, which substitute the labels computed by the host method. On the contrary, Refine CNNs calculate residual correction values and their output is added to the host method's output. According to Gidaris and Komodakis (2017), each variant has its own shortcomings. Replace CNNs must learn to operate as unitary functions in case the initial labels are correct, which is challenging for deeper networks. Refine CNNs can more easily learn to output a zero value for correct initial labels, but face greater difficulties in calculating big residual correction in the case of hard mistakes.

The proposed method incorporates an error correction module targeted to systematic errors originating from the base segmentation algorithm. The detailed architecture is presented in Fig. 6. Orthogonal slices are extracted from the cropped MRIs and segmentation masks and are fed to independent CNNs, followed by a late average fusion of their outputs. A combination of Replace and Refine CNNs is used at each of the three branches. A Replace CNN is placed first, followed by the corresponding Refine CNN. The extended use of residual blocks in Replace CNNs minimizes the aforementioned problem of them having difficulties operating as unitary function when needed and lets them focus on correcting hard mistakes. With hard mistakes already corrected, Refine CNNs are used to only make fine adjustments to the final labels. Thus, in the proposed architecture, we keep only the advantages of each

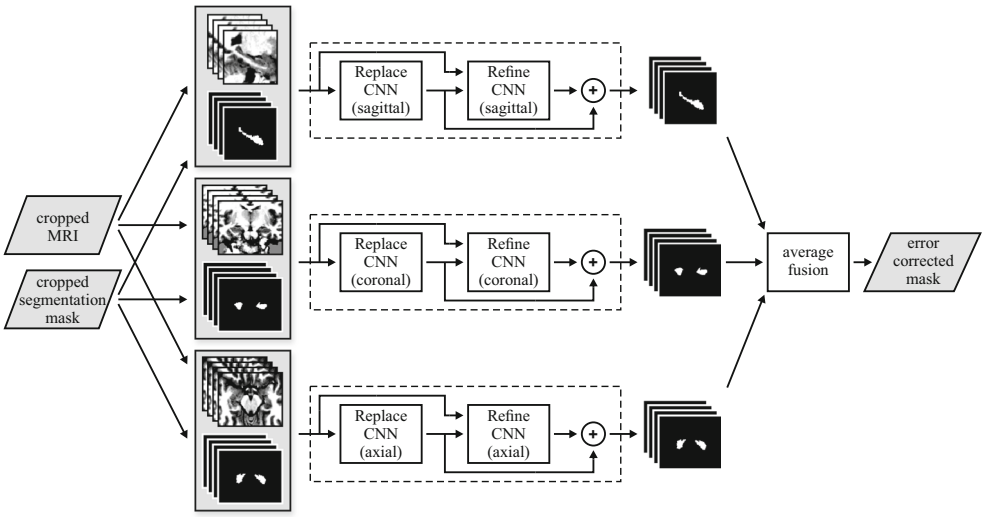


**Fig. 6** Error correction module architecture. The cropped MRI and segmentation mask are decomposed into orthogonal slices. Each type of slices is processed by a separate chain of error correction CNNs. Average fusion combines all single slice outputs to a final 3D and error corrected segmentation mask

error correction CNN variant and efficiently correct both hard and soft segmentation mistakes.

#### Error Correction CNNs

Each Replace and Refine CNN in Fig. 6 has the CNN structure presented in Fig. 4, with two minor differences. The first is the number of channels in the input layer of error correction CNNs, which is equal to the number of inputs in each case (2 for Replace and 3 for Refine CNNs). Filter depth at the first convolutional layer is modified accordingly. The second difference only applies to Refine CNNs. Layers after the last convolutional layer are omitted, since we do not need to explicitly restrict the output to a specific range in the case of residual corrections.

The Replace and Refine CNNs in each branch, along with the addition that follows were implemented as a single, deeper network and were trained in an end-to-end way, allowing their parameters to be co-adapted. On a technical level, this module consists of only three deeper error correction CNNs, denoted with dotted bounding boxes in Fig. 6.

## Implementation

We used Torch7 (Collobert et al. 2011) for the implementation and training of the proposed method. We also utilized the cuDNN Library (Chetlur et al. 2014) to accelerate the training and inference processes. All experiments were performed on a computer equipped with a NVIDIA GeForce GTX 1080, Ubuntu 16.04 LTS, CUDA v9.0.176 and cuDNN v7.0.3. The code and trained models will be available upon the acceptance of this paper.

### Training with the HarP Dataset

To facilitate the training and evaluation processes, the dataset must be first split into training and test sets. Aiming to obtain meaningful results during evaluation, we ensured that these sets were mutually exclusive in all of our experiments. Only the training set was utilized to train the proposed method, while the separate test set was subsequently used to measure the performance.

After pre-processing, only a small amount of voxels corresponded to the region of the brain. This imbalance created significant problems, such as loss oscillations, that disrupted the training process. To limit the training data to the brain region, we used only slices that contained part of the brain and cropped them to smaller dimensions, maintaining the whole area of even the largest brain. Cropping dimensions were common for each slicing operation, to facilitate the combination of slices in mini-batches. These steps where applied only to the inputs of segmentation CNNs and only during training. At test time, the whole MRI volume was provided as input to the segmentation module. Inputs provided to error correction CNNs were already cropped, as was described in "Hippocampal Region Cropping".

Training the proposed method consists in training six different CNNs, three for the segmentation and three for the error correction of sagittal, coronal and axial slices, respectively. We preferred full slices over patches as input to all CNNs. This led to better efficiency and lower processing times, as each output was calculated with a single forward pass and unnecessary calculations for overlapping patch regions were avoided (Long et al. 2015). Slices from all MRIs belonging to the training set were fed to the networks in random order, which also changed after the completion of each training epoch. The mini-batch size was set to eight slices, due to memory constraints.

Initializing the trainable parameters of each CNN was based on the Xavier method (Glorot and Bengio 2010), but with random values obtained from a uniform distribution in the $[-1, 1]$ range. Trainable parameters were updated using the Adam optimizer (Kingma and Ba 2014), with $beta1 = 0.9$ and $beta2 = 0.999$. The loss function used during training was the mean square error between the output and the ground truth segmentation masks. Training of each CNN always lasted for 40 complete epochs and the model saved at the end of the training process was used later to evaluate the performance. Initial learning rate was set to $5 \times 10^{-5}$ and was exponentially decayed after every iteration according the formula:

$$learning\_rate = 5 \times 10^{-5} e^{-0.175 epoch} \quad (1)$$

where $epoch$ is the exact number of completed epochs. Learning rate was decreased by a factor of 1000 until the end of training.

Training of each segmentation and error correction CNN required on average 7.5 and 4.2 hours respectively. Although error correction CNNs are twice as deep, as they include both the Replace and Refine networks, their training time was lower, due to the smaller input dimensions after cropping.

### Transfer Learning to the MICCAI Dataset

Multi-atlas based segmentation methods can provide accurate segmentations using a small number of very similar atlases. When large datasets are available, a selection method is necessary to extract the most similar atlases, as using the whole dataset deteriorates the segmentation quality (Aljabar et al. 2009). On the contrary, CNNs inherently require sufficiently large training sets. In segmentation tasks, CNN performance increases logarithmically with the

size of the training dataset (Sun et al. 2017). The required dataset size is in general proportional to the capacity of the used model. Deeper CNNs can generalize better to unseen cases, but at the same time require more data to train properly. The MICCAI dataset, which contains only 15 training atlases, is not directly applicable to the proposed method. To overcome this inefficiency, we exploited the unique properties of CNNs, which allow efficient transfer learning from larger datasets.

In practice, CNNs are rarely trained from scratch, either due to insufficiently large training sets or to reduce the required training time. Instead, it is common practice to use a pre-trained network, which was usually trained with a much larger dataset and to fine-tune it to the new, smaller dataset, to account for this dataset's specific characteristics. Features already learned by a CNN can be reused to address a new problem, using a different dataset. Transfer learning is more efficient when data from different datasets are of similar nature. In this way, features already learned are more appropriate for the new task and less fine-tuning is required.

The conditions for applying transfer learning between the datasets used in this study are favorable, as they both contain T1-weighted MRIs of the brain. The transfer learning procedure is presented in Fig. 7. First, all CNNs were pre-trained from scratch using the HarP dataset. Then, the pre-trained CNNs were fine-tuned utilizing only the 15 atlases from the MICCAI training set. Finally, the fine-tuned models were evaluated with the MICCAI test set. The same procedure was followed for both segmentation and error correction CNNs.

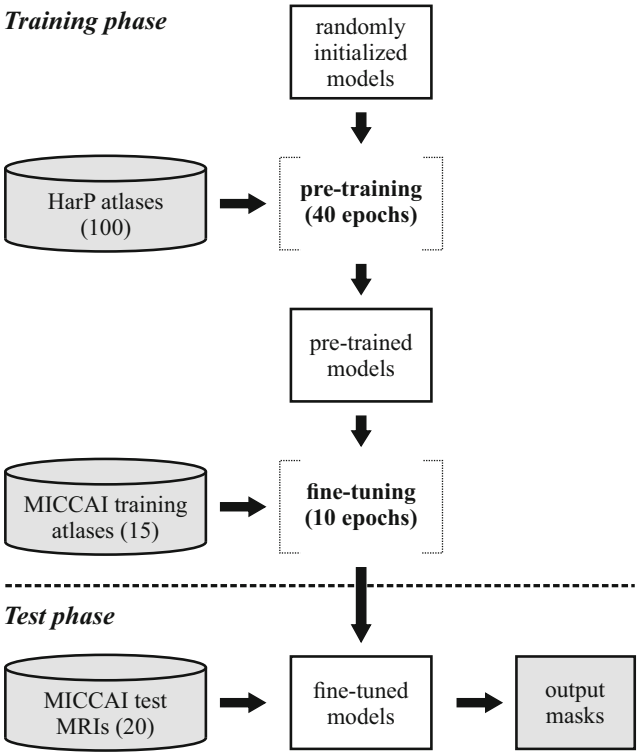


**Fig. 7** Transfer learning from HarP to the MICCAI dateset

Compared to the training procedure of the respective CNNs with the HarP dataset, two hyperparameters were altered during fine-tuning. The initial learning rate was set to $2 \times 10^{-5}$, as bigger values led to zero outputs after the first fine-tuning epoch, indicating that useful features for the detection of hippocampal regions were quickly forgotten, due to the dominant background class. Combined with the initial learning rate, which was set as high as possible, a fine-tuning duration of 10 epochs was enough to reach maximum performance.

Due to the decreased number of epochs and atlases in the MICCAI training set, fine-tuning was completed much faster. On average, 23 and 7.5 minutes were required for the fine-tuning of each segmentation and error correction CNN respectively.

### Thresholding

To obtain a binary segmentation mask during evaluation, we applied a threshold to the value of each voxel of the error corrected mask. Since they express the probability of a single voxel belonging to the hippocampus, the default threshold value was set to 0.5.

Segmentation masks were normally forwarded to the error correction module in continuous form, without any thresholding applied. However, to study the effect of different components in the overall performance, thresholds were also applied to the segmentation masks produced by the segmentation module (without error correction) and to the outputs of each individual CNN.

## Results

### Evaluation Metrics

The level of agreement between the outputs and the corresponding manual segmentations was quantified using the following metrics:

$$Dice = \frac{2 \cdot |A \cap B|}{|A| + |B|} \tag{2}$$

$$Jaccard = \frac{|A \cap B|}{|A \cup B|} \tag{3}$$

$$precision = \frac{|A \cap B|}{|B|} \tag{4}$$

$$recall = \frac{|A \cap B|}{|A|} \tag{5}$$

where $A$ is the set of voxels classified as part of the hippocampus by the proposed method and $B$ the corresponding set of the ground truth mask. The Wilcoxon signed-rank test was utilized to assess the statistical significance between the different outcomes.

## Results in the HarP Dataset

The performance of the proposed method in the HarP dataset was evaluated through a 5-fold cross-validation process. The 100 atlases of the HarP dataset were equally and proportionally divided into five folds, according to gender, age, medical diagnosis, MR scanner field strength and bilateral hippocampal volume. Training and testing of the entire method were repeated five times in the HarP dataset. In each round, the method was trained from scratch with a different set of 80 atlases (4 training folds) and subsequently tested with the remaining 20 atlases (test fold). First, segmentation CNNs were trained and tested for all cross-validation rounds, in order to provide intermediate segmentation masks for the entire dataset. Then, error correction CNNs were trained and tested in a similar manner. It is important to point out that no data utilized during training were also used to test the performance in any case.

Table 2 shows the mean Dice and standard deviation in the HarP dataset, averaged over all five test folds. Results are reported for the outputs of each segmentation CNN and after average fusion for both the segmentation and error correction steps.

We notice that individual segmentation CNNs exhibit different levels of accuracy, with coronal slices appearing to be best suited for hippocampus segmentation. This is clearly visible in Fig. 8, where red bars corresponds to the median value, blue boxes to the 25th-75th percentile range and red crosses to outlier values (more than 1.5× the interquartile range away from the box). Also, individual segmentation CNNs produce segmentations with various types of errors. For example, they may not recognize parts of the hippocampus in some slices or classify as foreground voxels far away from the actual position of the structure, resulting in spatial incoherence. These problems are eliminated after fusion (Fig. 9), while both the mean Dice value and standard deviation improve over the best

**Table 2** Mean Dice and standard deviation for the outputs of individual CNNs and after average fusion using the HarP dataset. "EC fusion" refers to the output of the error correction module

| | Sagittal | Coronal | Axial | Fusion | EC fusion |
|---|---|---|---|---|---|
| Dice | 0.8834 | 0.8898 | 0.8665 | 0.8965 | **0.9010** |
| std | 0.0278 | 0.0266 | 0.0348 | 0.0224 | **0.0182** |

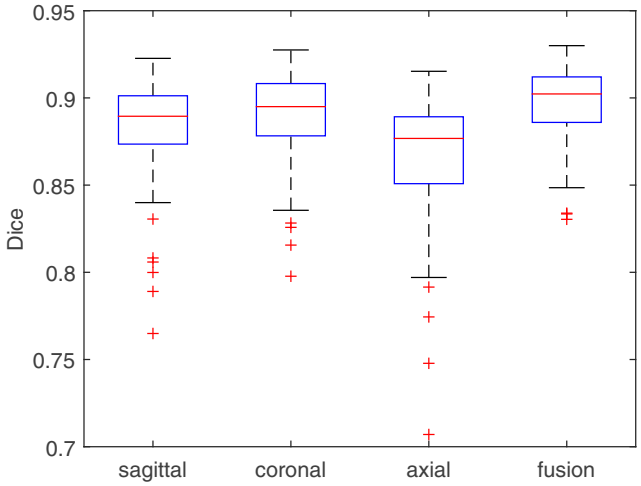


**Fig. 8** Dice distribution for the segmentation step using the HarP dataset

performing coronal segmentation CNN. The improvements of fusion over each individual CNN are statistically significant ($p < 8.5 \times 10^{-12}$).

The error correction step improves the quality of the received segmentation mask, both in terms of mean Dice and standard deviation. After fusion, the mean Dice value increases to 0.9010, which is 0.0045 higher than the best result obtained without error correction. The improvement offered by error correction is also statistically significant ($p < 1.4 \times 10^{-12}$). A qualitative example comparing the outputs before and after the error correction module on a difficult case of the HarP dataset is presented in Fig. 10. Iterative refinement using multiple stacked error correction modules was also explored, using either the same or different CNNs in successive modules and additional error correction modules consisting only of Refine CNNs. In terms of mean Dice, using CNN duplicates and two successive error correction modules led to a small performance improvement. Aiming to achieve the best trade-off between segmentation accuracy and total processing time, we chose to include a single error correction module for hippocampus segmentation.

The performance of the proposed method in different medical diagnoses is presented in Table 3. We notice that error correction offers consistent improvement regardless of the diagnosis, which highlights the robustness of the proposed method to different medical cases. As expected, Dice is lower for patients with AD, due to the deformation of the hippocampus structure, but also due to the reduced average hippocampal volume. Smaller structures have higher percentage of voxels near their surface. This affects the segmentation quality (Fig. 11), as it is more challenging for the automatic method to segment areas near the outer surface, than internal ones.

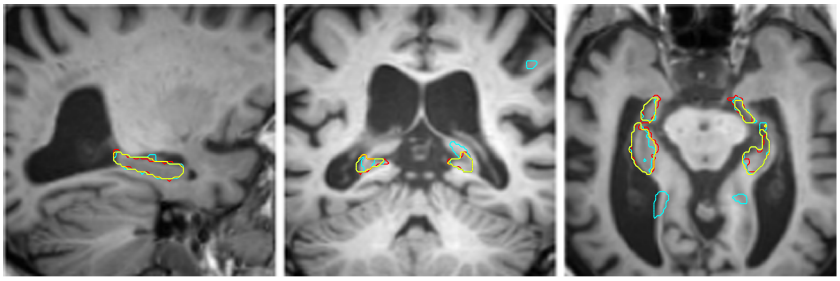

**Fig. 9** Effect of average fusion during segmentation. The output of the coronal segmentation CNN and the segmentation mask after average fusion are depicted in cyan and yellow respectively. The outline of the ground truth is depicted in red. Best viewed in color

Table 4 compares the proposed method to other published methods that use the same dataset. Values are shown as they appear in the respective publications, for each hippocampus separately or combined. The proposed method is also compared to two widely used segmentation tools, namely FreeSurfer (Fischl et al. 2002) and FIRST (Patenaude et al. 2011). Segmentation masks were obtained by executing the default test scripts (*recon-all -all* and *run_first_all*) provided with the respective packages. The performance of the tools was assessed for the structure of the hippocampus, treating all other output labels as background.

We observe that the proposed method compares favorably to previous methods in every evaluation metric. It is also the only one based on CNNs, which demonstrates their potential in brain MRI segmentation. The next two methods, exhibiting a mean Dice value over 0.88, are based on multi-atlas registration and fusion techniques. Trained and validated using a different set of ADNI MRIs, a patch-based label fusion method with structured discriminant embedding (Wang et al. 2018) acheived a mean Dice value of 0.879 and 0.889 for the left and right hippocampus respectively. A CNN-based method, also validated using MRIs from the ADNI database, was proposed in Chen et al. (2017). Despite the fact that only normal cases were included and the hippocampal region was manually cropped before segmentation, their final result (mean Dice value of 0.8929) is also inferior to that of the proposed method.

Furthermore, we compared each of the estimated hippocampal volumes with the corresponding ground truth volume. The proposed method produces segmentations that are on average $70mm^3$ smaller, taking into account both hippocampi. Although relatively small ($-1.3\%$ compared to the ground truth), the volume difference can be attributed to the many fine details in the manual tracings, which are more difficult to be captured. The correlation coefficient between the two volumes is 0.97. Based on these results, we conclude that there is high level of agreement, which is also evident in Fig. 12.

Qualitative results are presented in Fig. 13, where an indicative sagittal, coronal and axial slice is shown for the best, median and worst case. The ADNI image ID, medical diagnosis, bilateral ground truth hippocampal volume and Dice value are listed for each case. We observe that most details of the manual tracings are well preserved by the proposed method, while the outline from the automatic segmentation method is smoother. Output volumes are spatially consistent and close to the corresponding manual tracings, even when the boundary of the hippocampus is not visible, as in the sagittal slice of the best and worst cases and the axial slice of the median case. Automatic segmentation quality seems satisfactory even in the worst case, where the inferior Dice value can be attributed to the very small bilateral hippocampal volume.

## Results in the MICCAI Dataset

The MICCAI dataset was already divided into two sets. In our study, we used the same training and test sets. Since a single split of the dataset was used, a total of six CNNs needed to be trained. Fine-tuning of the error correction CNNs required segmentation masks for the 15 atlases of the MICCAI training set, which were not directly available, as testing of the segmentation module was conducted using the separate test set. To that end, a 5-fold inner cross-validation

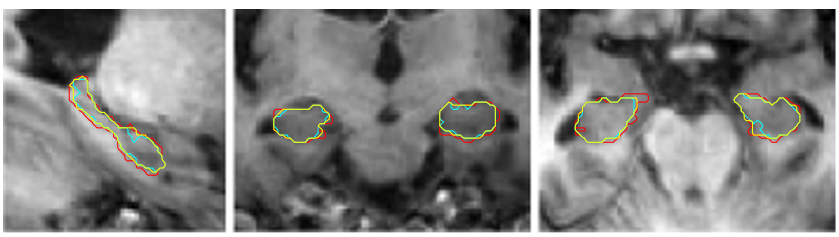

**Fig. 10** Effect of error correction on a difficult case of the HarP dataset (case #254893, diagnosed with AD). The outputs before and after the error correction are depicted in cyan and yellow respectively. The outline of the ground truth is depicted in red. Best viewed in color

**Table 3** Mean Dice and standard deviation per diagnosis in the HarP dataset. Results were obtained after average fusion for both the segmentation and error correction steps

| Diagnosis | Segmentation | Error correction |
|---|---|---|
| Normal (29) | 0.9071 (± 0.0201) | 0.9115 (± 0.0145) |
| MCI (21) | 0.8970 (± 0.0227) | 0.9020 (± 0.0176) |
| LMCI (13) | 0.9053 (± 0.0117) | 0.9075 (± 0.0115) |
| AD (37) | 0.8849 (± 0.0220) | 0.8898 (± 0.0173) |
| All (100) | 0.8965 (± 0.0224) | 0.9010 (± 0.0182) |

was performed with the MICCAI training set, using a different set of 12 atlases to fine-tune the segmentation CNNs at each round and the remaining 3 atlases to produce segmentation masks for later usage, during the fine-tuning of the error correction module.

Table 5 summarizes the results for the segmentation step using different training approaches. When training from scratch, with random parameter initialization, the mean Dice value in the MICCAI test set was 0.8182 after fusion. Evaluating the pre-trained with the HarP dataset CNNs without any additional fine-tuning resulted in inferior performance. Transfer learning led to far superior segmentation quality. Mean Dice value was significantly higher for all three individual segmentation CNNs and reached the value of 0.8711 after fusion ($p = 4.8 \times 10^{-5}$ compared to training from scratch). Furthermore, Dice improved for all 20 test atlases in relation to both training from scratch and evaluating the pre-trained CNNs. These results suggest that while CNN-based methods can benefit from larger datasets, a transfer learning procedure is essential to surpass the performance of training from

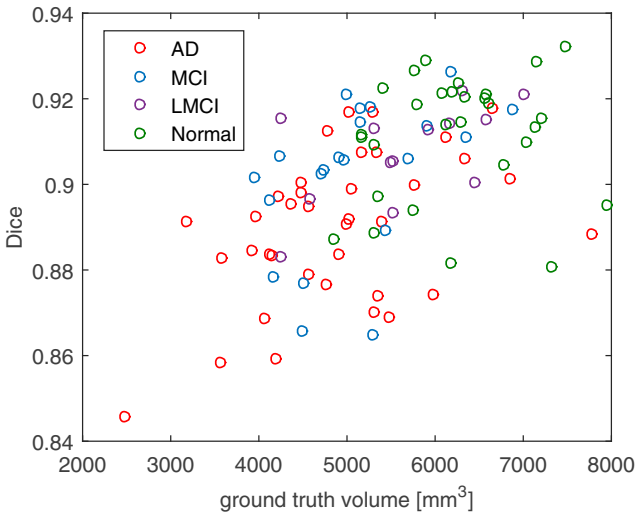


**Fig. 11** Dice values for the error corrected segmentation masks against bilateral ground truth hippocampal volumes in the HarP dataset. Best viewed in color

**Table 4** Performance comparison between the proposed method and other published methods and tools using the HarP dataset

| | | Giraud et al. (2016) | Zhu et al. (2017) | Ahdidan et al. (2016) | Maglietta et al. (2016) | Inglese et al. (2015) | Platero and Tobar (2017) | Chincarini et al. (2016) | FreeSurfer (v6.0) | FIRST (FSL v6.0) | Proposed method |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Dice | L | | 0.881 | 0.8739 | 0.8670 | 0.86 | | | 0.7039 | 0.8051 | **0.9000** |
| | R | | 0.885 | 0.8749 | 0.8594 | 0.86 | | | 0.6918 | 0.8033 | **0.9015** |
| | both | 0.898 | | | | | 0.850 | 0.85 | 0.6980 | 0.8044 | **0.9010** |
| Jaccard | L | | 0.788 | | 0.7664 | | | | 0.5450 | 0.6748 | **0.8187** |
| | R | | 0.795 | | 0.7591 | | | | 0.5306 | 0.6728 | **0.8213** |
| | both | | | | | | | | 0.5377 | 0.6738 | **0.8203** |
| precision | L | | 0.880 | | 0.8872 | 0.89 | | | 0.6199 | 0.7440 | **0.9051** |
| | R | | 0.884 | | 0.8772 | 0.88 | | | 0.6115 | 0.7406 | **0.9091** |
| | both | | | | | | | | 0.6156 | 0.7418 | **0.9073** |
| recall | L | | 0.884 | | 0.8598 | 0.85 | | | 0.8196 | 0.8819 | **0.8960** |
| | R | | 0.889 | | 0.8501 | 0.84 | | | 0.7993 | 0.8850 | **0.8952** |
| | both | | | | | | | | 0.8094 | 0.8833 | **0.8958** |

Metrics were calculated separately for each hippocampus and jointly for both hippocampi. Results for other methods are displayed as reported in the respective publications. Results for Freesurfer and FIRST were obtained by executing the default test scripts provided with the respective packages. Best performing values appear in bold

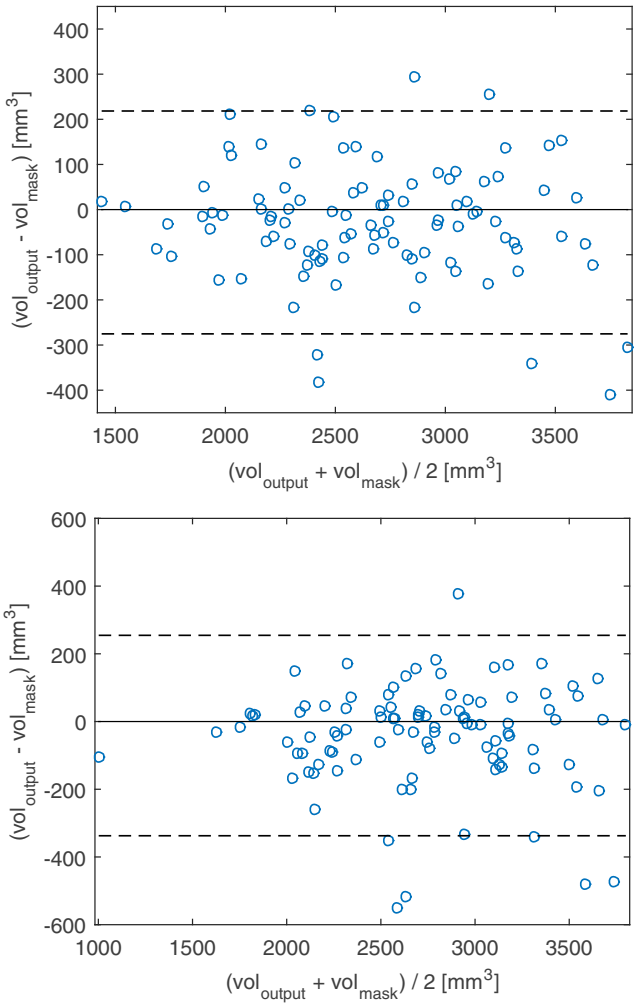


**Fig. 12** Bland-Altman plots showing the hippocampal volume agreement between the error corrected and ground truth masks for the left (top plot) and right (bottom plot) hippocampus in the HarP dataset. Dashed lines indicate the 95% confidence level interval

scratch, especially when different segmentation protocols and MR scanners are involved. In the remainder of this section, all results refer to the transfer learning training approach.

Table 6 presents the mean Dice and standard deviation values in the MICCAI test set for the outputs of each segmentation CNN and after average fusion for both the segmentation and error correction steps. Dice distributions for the segmentation step are presented in Fig. 14. Compared to the HarP dataset, the improvement of fusion over the individual CNNs is less obvious, but still statistically significant ($p < 0.0024$). However, one notable difference is the larger effect of error correction in the overall performance, which improves the mean Dice value by 0.0105 ($p = 2.4 \times 10^{-4}$), more than twice the amount compared to the improvement in the HarP dataset.

For our final results, which were obtained with error correction and after average fusion, we searched for the optimal threshold value, in order to further improve the adaptation of the proposed method to the new dataset, after the process of transfer learning. To that end, we utilized only the 15 atlases of the MICCAI training set and performed another 5-fold inner cross-validation, this time with respect to the error correction CNNs. Maximum Dice in the training set was achieved with $T = 0.42$. This value was then used when evaluating the proposed method with the MICCAI test set, further increasing the mean Dice by 0.0019 to the value of 0.8835 ($p = 0.0478$). In order to obtain the best possible result, the threshold search should be repeated when transfer learning to a new domain. However, it should be noted that this is not necessary, as the proposed method performs almost equally well and still surpasses the competition (refer to Table 7) using a wide range of threshold values, including the default value of 0.5, as presented in Fig. 15.

Table 7 compares the proposed method with the three top performing entries of the MICCAI challenge (PICSL BC, NonLocalSTAPLE and PICSL Joint), other published methods that use the same test set, as well as the FreeSurfer and FIRST segmentation tools. Segmentation masks for all entries of the MICCAI challenge are publicly available, which enabled us to calculate all evaluation metrics. Results for de Brébisson and Montana (2015) were produced using the provided official code. For the rest of the methods, we report only values included in the respective publications. As can be seen, the proposed method compares favorably to all other methods. Kushibar et al. (2017) and de Brébisson and Montana (2015) are both CNN-based methods that utilize orthogonal slices as inputs and combine them within a single CNN. The proposed method achieves superior performance, which strengthens our choice to form model ensembles, comprised of independent CNNs for each slicing operation and combine their outputs with late fusion.

Taking into account both hippocampi, the proposed method produces segmentations that are on average $126mm^3$ larger than the respective ground truth volumes. The larger output volume (1.7% compared to the ground truth) can be attributed to the selection of a lower threshold value ($T = 0.42$), which was optimized for maximum Dice. The correlation coefficient between them is 0.90. Volume agreement is graphically presented in Fig. 16.

Figure 17 shows qualitative results for the best, median and worst case. A representative sagittal, coronal and axial slice is presented, along with the MRI ID, bilateral hippocampal volume and Dice value.

## Processing Time

Processing time was measured with the whole MRI volume as input to the proposed method. The input dimensions

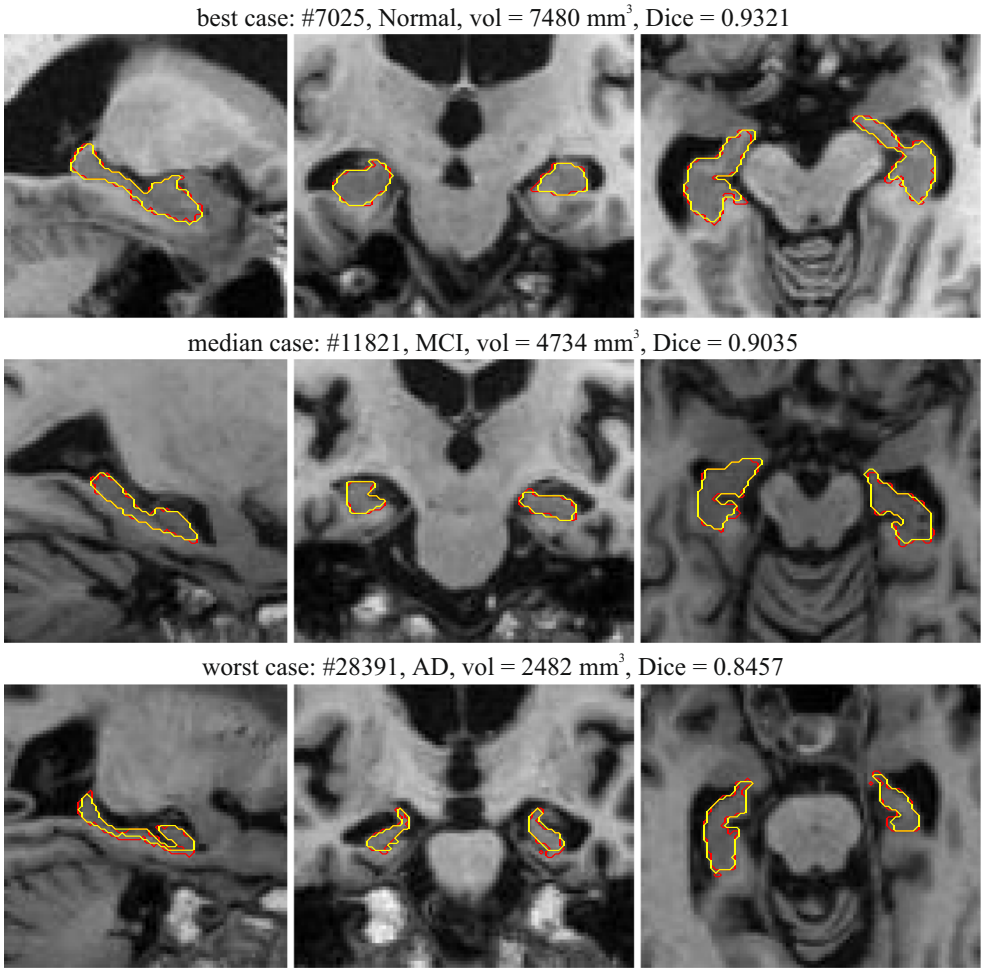


**Fig. 13** Qualitative results for the error corrected masks in the HarP dataset. The outlines of the automatic segmentations and the ground truth masks are depicted in yellow and red respectively. Best viewed in color

**Table 5** Mean segmentation Dice in the MICCAI test set using different training approaches

| | Sagittal | Coronal | Axial | Fusion |
|---|---|---|---|---|
| Train from scratch | 0.8143 | 0.8074 | 0.7351 | 0.8182 |
| Pre-trained HarP CNNs | 0.7723 | 0.7737 | 0.6712 | 0.7690 |
| Transfer learning | **0.8577** | **0.8655** | **0.8265** | **0.8711** |

**Table 6** Mean Dice and standard deviation for the outputs of individual CNNs and after average fusion in the MICCAI test set. "EC fusion" refers to the output of the error correction module

| | Sagittal | Coronal | Axial | Fusion | EC fusion |
|---|---|---|---|---|---|
| Dice | 0.8577 | 0.8655 | 0.8265 | 0.8711 | **0.8816** |
| std | 0.0338 | 0.0256 | 0.0429 | 0.0247 | **0.0150** |

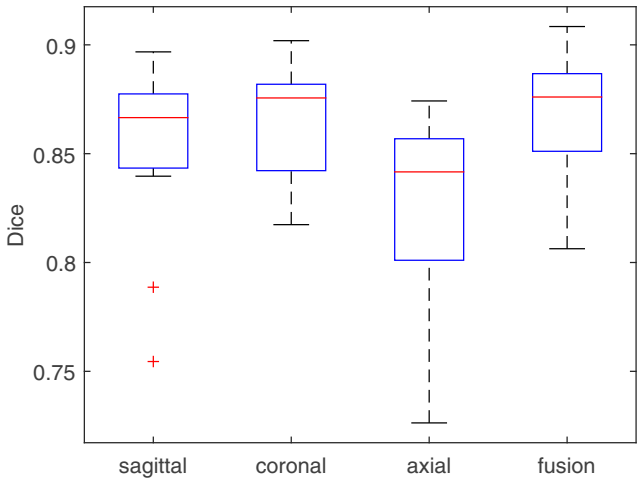


**Fig. 14** Dice distribution for the segmentation step using the MICCAI test set

were 197 × 233 × 189 and up to 256 × 334 × 256 voxels for the HarP and the MICCAI datasets respectively. For computational efficiency, when a mini-batch consists only of non-brain slices (the sum of all pixels is zero), we explicitly set the output to zero, without passing these slices through the CNNs.

Using a single NVIDIA GTX 1080, segmenting one MRI of the HarP dataset requires 14.8 seconds. In the MICCAI dataset, where the dimensions of the input MRI volume are larger, the equivalent required time is 21.8 seconds. In more detail, each segmentation CNN requires on average 3.7 or 6.0 seconds when segmenting a MRI from the Harp or the MICCAI dataset respectively. After cropping, error correction CNNs always receive fixed sized volumes in evaluation mode and require on average 1.2 seconds each, including both the Replace and Refine parts. Time is almost

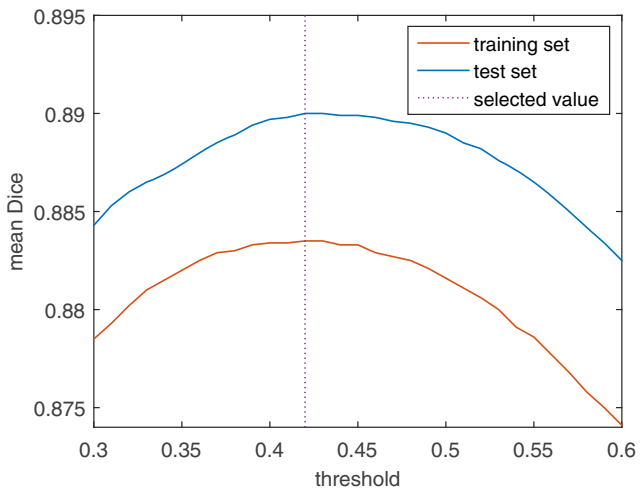


**Fig. 15** Effect of threshold value in the overall performance after transfer learning to the MICCAI dataset

**Table 7** Performance comparison between the proposed method and other published methods and tools using the MICCAI test set

| | | MICCAI Challenge (Landman and Warfield 2012) | | | Wang and Yushkevich 2013 | Zarpalas et al. 2014b | Ledig et al. 2015 | de Brébisson and Montana 2015 | Kushibar et al. 2017 | FreeSurfer (v6.0) | FIRST (FSL v6.0) | Proposed method |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1st | 2nd | 3rd | | | | | | | | |
| Dice | L | 0.8705 | 0.8648 | 0.8629 | 0.871 | | 0.869 | 0.8213 | 0.876 | 0.7897 | 0.8085 | **0.8825** |
| | R | 0.8692 | 0.8672 | 0.8617 | 0.872 | | 0.868 | 0.8197 | 0.879 | 0.8013 | 0.8094 | **0.8843** |
| | both | 0.8698 | 0.8660 | 0.8622 | | 0.870 | | 0.8205 | | 0.7956 | 0.8090 | **0.8835** |
| Jaccard | L | 0.7715 | 0.7627 | 0.7600 | | | | 0.6981 | | 0.6541 | 0.6787 | **0.7901** |
| | R | 0.7693 | 0.7665 | 0.7578 | | | | 0.6959 | | 0.6693 | 0.6803 | **0.7930** |
| | both | 0.7702 | 0.7645 | 0.7587 | | | | 0.6969 | | 0.6617 | 0.6794 | **0.7915** |
| precision | L | 0.8611 | 0.8492 | 0.8503 | | | | 0.8029 | | 0.7343 | 0.7721 | **0.8681** |
| | R | 0.8764 | 0.8673 | 0.8631 | | | | 0.8454 | | 0.7656 | 0.8067 | **0.8828** |
| | both | 0.8688 | 0.8584 | 0.8568 | | | | 0.8236 | | 0.7499 | 0.7892 | **0.8755** |
| recall | L | 0.8819 | 0.8829 | 0.8779 | | | | 0.8447 | | 0.8574 | 0.8516 | **0.8995** |
| | R | 0.8640 | 0.8692 | 0.8621 | | | | 0.7982 | | 0.8426 | 0.8154 | **0.8877** |
| | both | 0.8725 | 0.8756 | 0.8696 | | | | 0.8205 | | 0.8496 | 0.8326 | **0.8932** |

Metrics were calculated separately for each hippocampus and jointly for both hippocampi. Results for the MICCAI Challenge entries were calculated from the output segmentation masks, which were made available after the challenge. Results for de Brébisson and Montana 2015 were produced using the provided official code. Results for Freesurfer and FIRST were obtained by executing the default test scripts provided with the respective packages. Other results are displayed as reported in the respective publications. Best performing values appear in bold

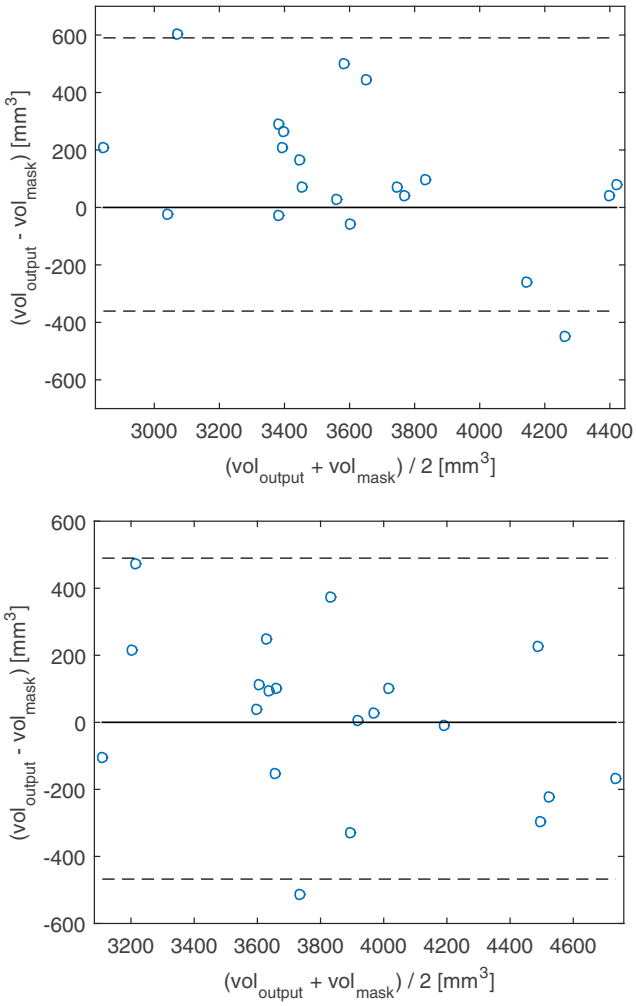


**Fig. 16** Bland-Altman plots showing the hippocampal volume agreement between the error corrected and ground truth masks for the left (top plot) and right (bottom plot) hippocampus in the MICCAI test set. Dashed lines indicate the 95% confidence level interval

entirely consumed by the CNNs, as all other processes, such as average fusion or thresholding, add a negligible amount. Thus, the total required time can be reduced by a factor of 3 through the parallel execution of sagittal, coronal and axial CNNs in different graphics cards.

To make fair comparison, we must also consider the preprocessing time required for a new MRI. In the literature, additional registrations that must be performed for each new MRI are often considered to be part of the preprocessing procedure and time spent is not counted or reported separately. For example, in Giraud et al. (2016), which has a competing performance in terms of Dice, authors report 5 minutes for preprocessing and only a few seconds for segmentation. In comparison, the preprocessing pipeline of the proposed method is completed within 30 seconds, with single-core CPU execution. In the MICCAI dataset, the second best performing method (Kushibar et al. 2017), which is CNN-based and utilizes the GPU, also requires a total of 5 minutes for every test MRI. Thus, the proposed method is overall more than 5× faster compared to those methods. Compared to other methods of Table 4, where Platero and Tobar (2017) is reported to require 17 minutes per MRI and Zhu et al. (2017) requires 20 minutes just for the label fusion step, the proposed method can be considered at least 23× faster, which is a dramatic improvement, given that it also offers superior segmentation accuracy.

## Discussion

In this work, we developed an automatic segmentation method of the hippocampus from magnetic resonance imaging, incorporating a number of different CNNs and exploiting their distinctive capabilities. The proposed architecture is composed of three modules. In the first one, we calculate a segmentation mask for the whole MRI volume. Then, based on that segmentation mask, we crop a wider area around both hippocampi. Finally, the error correction module, which uses a combination of Replace and Refine CNNs, corrects erroneous labels within the cropped region and further improves the performance of the entire method. The proposed architecture can be easily extended to consider multiple structures or even perform full brain segmentation by adjusting or totally removing the region cropping module, should sufficient and annotated datasets become available for all structures of interest.

Inside the segmentation and error correction modules, 3D inputs are decomposed into 2D orthogonal slices. We trained separate CNNs for each slicing operation and performed a late voxel-wise average fusion of their outputs. In practice, we designed an ensemble of three models, with common internal structure, but different training data. All CNNs are fully convolutional and operate with full slices, allowing efficient inference regardless of the input size.

We followed two different approaches while training the proposed method. Starting with random initialization, we managed to obtain state-of-the-art results in the HarP dataset, which included a sufficiently large number of atlases. This was not possible for the MICCAI dataset, due to the much smaller training set, contradicting the philosophy of CNN training. To overcome this inefficiency, we used the already trained networks with the HarP dataset as an initialization point and consequently fine-tuned them with the MICCAI training set. Overall, transfer learning showed spectacular relative improvement, proving that in CNN-based segmentation methods the combination of multiple datasets is beneficiary, even with different manual segmentation protocols. This is a unique advantage of CNNs compared to other segmentation methods, which can be exploited for applications in the medical domain, where the

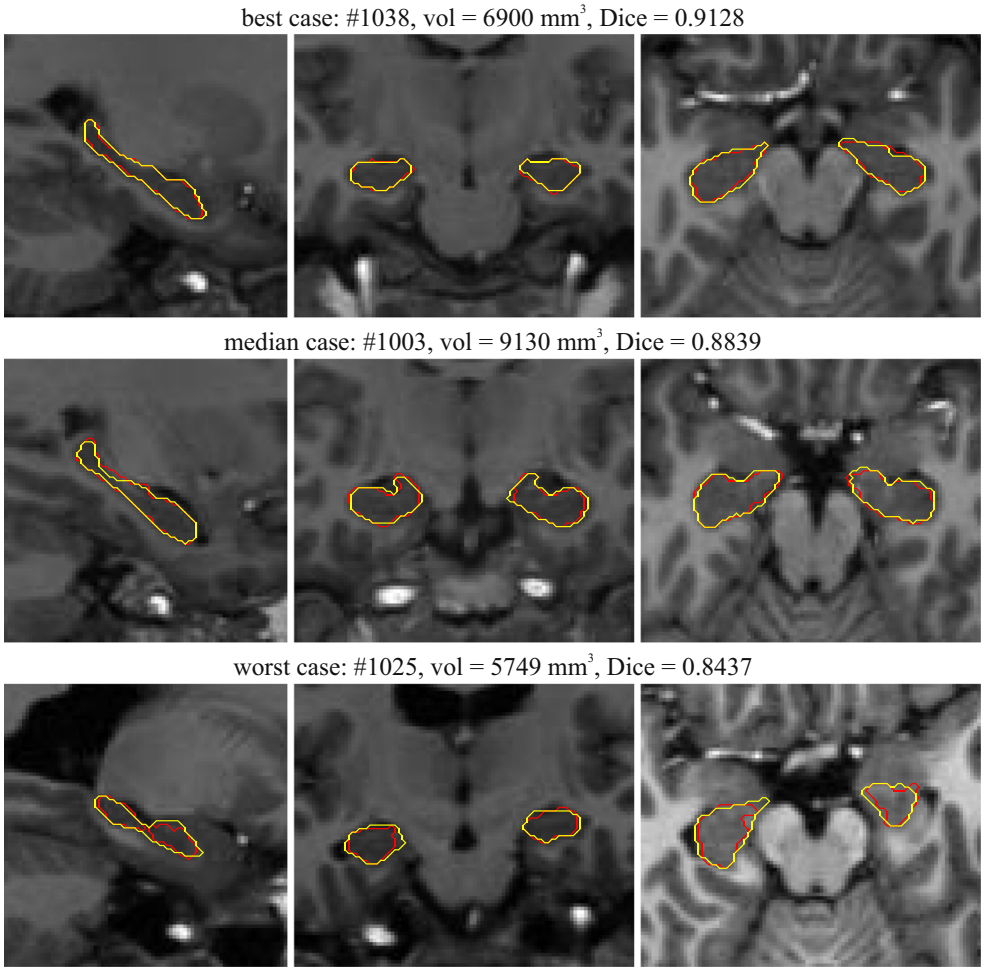


**Fig. 17** Qualitative results for the error corrected masks in the MICCAI test set. The outline of the automatic segmentation and the ground truth are depicted in yellow and red respectively. Best viewed in color

creation of large and manually annotated datasets is costly and time consuming.

The proposed method was validated using two different public datasets, which include cases with various demographic and clinical characteristics, demonstrating high robustness and surpassing in performance previously published methods. As shown by the results, the addition of the error correction mechanism led to a systematic improvement and helped our implementation take precedence over other competing methods. Overall, for a test MRI, the proposed method is dramatically faster when compared to multi-atlas registration and fusion methods.

**Acknowledgments** The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPU used for this research.

**Information Sharing Statement** Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (RRID:SCR_003007, adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

**Compliance with Ethical Standards**

**Conflict of interests** The authors declare no conflicts of interest.

## References

Ahdidan, J., Raji, C.A., DeYoe, E.A., Mathis, J., Noe, K., Rimestad, J., Kjeldsen, T.K., Mosegaard, J., Becker, J.T., Lopez, O. (2016). Quantitative neuroimaging software for clinical assessment of hippocampal volumes on MR imaging. *Journal of Alzheimer's Disease*, *49*(3), 723–732.

Aljabar, P., Heckemann, R.A., Hammers, A., Hajnal, J.V., Rueckert, D. (2009). Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. *NeuroImage*, *46*(3), 726–738.

Bateman, R.J., Xiong, C., Benzinger, T.L., Fagan, A.M., Goate, A., Fox, N.C., Marcus, D.S., Cairns, N.J., Xie, X., Blazey, T.M., et al. (2012). Clinical and biomarker changes in dominantly inherited Alzheimer's disease. *New England Journal of Medicine*, *367*(9), 795–804.

Bernasconi, N., Bernasconi, A., Caramanos, Z., Antel, S., Andermann, F., Arnold, D. (2003). Mesial temporal damage in temporal lobe epilepsy: a volumetric MRI study of the hippocampus, amygdala and parahippocampal region. *Brain*, *126*(2), 462–469.

Boccardi, M., Bocchetta, M., Morency, F.C., Collins, D.L., Nishikawa, M., Ganzola, R., Grothe, M.J., Wolf, D., Redolfi, A., Pievani,

M., et al. (2015). Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol. *Alzheimer's & Dementia*, *11*(2), 175–183.

de Brébisson, A., & Montana, G. (2015). Deep neural networks for anatomical brain segmentation. arXiv:150202445.

Bremner, J.D., Narayan, M., Anderson, E.R., Staib, L.H., Miller, H.L., Charney, D.S. (2000). Hippocampal volume reduction in major depression. *American Journal of Psychiatry*, *157*(1), 115–118.

Brosch, T., Tang, L.Y., Yoo, Y., Li, D.K., Traboulsee, A., Tam, R. (2016). Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE Transactions on Medical Imaging*, *35*(5), 1229–1239.

Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J. (2016). Human pose estimation with iterative error feedback. In: Proceedings of the IEEE conference on computer vision and pattern recognition, IEEE, pp. 4733–4742.

Chen, H., Dou, Q., Yu, L., Heng, P.A. (2016). Voxresnet: Deep voxelwise residual networks for volumetric brain segmentation. arXiv:160805895.

Chen, Y., Shi, B., Wang, Z., Sun, T., Smith, C.D., Liu, J. (2017). Accurate and consistent hippocampus segmentation through convolutional LSTM and view ensemble. In: International workshop on machine learning in medical imaging, Springer, pp. 88–96.

Chetlur, S., Woolley, C., Vandermersch, P., Cohen, J., Tran, J., Catanzaro, B., Shelhamer, E. (2014). cuDNN: Efficient primitives for deep learning. arXiv:14100759.

Chincarini, A., Sensi, F., Rei, L., Gemme, G., Squarcia, S., Longo, R., Brun, F., Tangaro, S., Bellotti, R., Amoroso, N., et al. (2016). Integrating longitudinal information in hippocampal volume measurements for the early detection of Alzheimer's disease. *NeuroImage*, *125*, 834–847.

Choi, H., & Jin, K.H. (2016). Fast and robust segmentation of the striatum using deep convolutional neural networks. *Journal of Neuroscience Methods*, *274*, 146–153.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O. (2016). 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention, Springer, pp. 424–432.

Collins, D.L., & Pruessner, J.C. (2010). Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting ANIMAL with a template library and label fusion. *NeuroImage*, *52*(4), 1355–1366.

Collobert, R., Kavukcuoglu, K., Farabet, C. (2011). Torch7: A matlab-like environment for machine learning. In: BigLearn, NIPS workshop, EPFL-CONF-192376.

Coupé, P., Manjón, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L. (2010). Nonlocal patch-based label fusion for hippocampus segmentation. In: International conference on medical image computing and computer assisted intervention, Springer, pp. 129–136.

Dolz, J., Desrosiers, C., Ayed, I.B. (2017). 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study. NeuroImage.

Du, A., Schuff, N., Amend, D., Laakso, M., Hsu, Y., Jagust, W., Yaffe, K., Kramer, J., Reed, B., Norman, D., et al. (2001). Magnetic resonance imaging of the entorhinal cortex and hippocampus in mild cognitive impairment and Alzheimer's disease. Journal of Neurology. *Neurosurgery & Psychiatry*, *71*(4), 441–447.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., et al. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, *33*(3), 341–355.

Gidaris, S., & Komodakis, N. (2017). Detect, replace, refine: Deep structured prediction for pixel wise labeling. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5248–5257.

Giraud, R., Ta, V.T., Papadakis, N., Manjón, J. V., Collins, D.L., Coupé, P., Alzheimer's Disease Neuroimaging Initiative, et al. (2016). An optimized patchmatch for multi-scale and multi-feature label fusion. *NeuroImage*, *124*, 770–782.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the 13th international conference on artificial intelligence and statistics, pp. 249–256.

Harrison, P.J. (2004). The hippocampus in schizophrenia: a review of the neuropathological evidence and its pathophysiological implications. *Psychopharmacology*, *174*(1), 151–162.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.M., Larochelle, H. (2017). Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, *35*, 18–31.

He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Inglese, P., Amoroso, N., Boccardi, M., Bocchetta, M., Bruno, S., Chincarini, A., Errico, R., Frisoni, G., Maglietta, R., Redolfi, A., et al. (2015). Multiple RF classifier for the hippocampus segmentation: Method and validation on EADC-ADNI harmonized hippocampal protocol. *Physica Medica*, *31*(8), 1085–1091.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv:150203167.

Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L Whitwell, J., Ward, C., et al. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, *27*(4), 685–691.

Jack, C.R., Barkhof, F., Bernstein, M.A., Cantillon, M., Cole, P.E., DeCarli, C., Dubois, B., Duchesne, S., Fox, N.C., Frisoni, G.B., et al. (2011). Steps to standardization and validation of hippocampal volumetry as a biomarker in clinical trials and diagnostic criterion for Alzheimer's disease. *Alzheimer's & Dementia*, *7*(4), 474–485.

Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, *36*, 61–78.

Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv:14126980.

Kleesiek, J., Urban, G., Hubert, A., Schwarz, D., Maier-Hein, K., Bendszus, M., Biller, A. (2016). Deep MRI brain extraction: a 3D convolutional neural network for skull stripping. *NeuroImage*, *129*, 460–469.

Krizhevsky, A., Sutskever, I., Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097–1105.

Kushibar, K., Valverde, S., Gonzalez-Villa, S., Bernal, J., Cabezas, M., Oliver, A., Llado, X. (2017). Automated sub-cortical brain structure segmentation combining spatial and deep convolutional features. arXiv:170909075.

Landman, B., & Warfield, S. (2012). MICCAI 2012 Workshop on Multi-Atlas Labeling. CreateSpace Independent Publishing Platform, ISBN: 1479126187.

Langerak, T.R., van der Heide, U.A., Kotte, A.N., Viergever, M.A., Van Vulpen, M., Pluim, J.P. (2010). Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). *IEEE Transactions on Medical Imaging*, *29*(12), 2000–2008.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Ledig, C., Heckemann, R.A., Hammers, A., Lopez, J.C., Newcombe, V.F., Makropoulos, A., Lötjönen, J., Menon, D.K., Rueckert, D. (2015). Robust whole-brain segmentation: application to traumatic brain injury. *Medical Image Analysis*, *21*(1), 40–58.

Leung, K.K., Barnes, J., Ridgway, G.R., Bartlett, J.W., Clarkson, M.J., Macdonald, K., Schuff, N., Fox, N.C., Ourselin, S., Alzheimer's Disease Neuroimaging Initiative, et al. (2010). Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. *NeuroImage*, *51*(4), 1345–1359.

Li, K., Hariharan, B., Malik, J. (2016). Iterative instance segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3659–3667.

Long, J., Shelhamer, E., Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440.

Maglietta, R., Amoroso, N., Boccardi, M., Bruno, S., Chincarini, A., Frisoni, G.B., Inglese, P., Redolfi, A., Tangaro, S., Tateo, A., et al. (2016). Automated hippocampal segmentation in 3D MRI using random undersampling with boosting algorithm. *Pattern Analysis and Applications*, *19*(2), 579–591.

Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L. (2007). Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, *19*(9), 1498–1507.

Mehta, R., Majumdar, A., Sivaswamy, J. (2017). BrainSegNet: a convolutional neural network architecture for automated segmentation of human brain structures. *Journal of Medical Imaging*, *4*(2), 024003.

Milletari, F., Ahmadi, S.A., Kroll, C., Plate, A., Rozanski, V., Maiostre, J., Levin, J., Dietrich, O., Ertl-Wagner, B., Bötzel, K., et al. (2017). Hough-CNN: deep learning for segmentation of deep brain regions in MRI and ultrasound. *Computer Vision and Image Understanding*, *164*, 92–102.

Moeskops, P., Viergever, M.A., Mendrik, A.M. (2016). de Vries LS, Benders MJ, Išgum I. *Automatic segmentation of MR brain images with a convolutional neural network. IEEE Transactions on Medical Imaging*, *35*(5), 1252–1261.

Morra, J.H., Tu, Z., Apostolova, L.G., Green, A.E., Toga, A.W., Thompson, P.M. (2010). Comparison of AdaBoost and support vector machines for detecting Alzheimer's disease through automated hippocampal segmentation. *IEEE Transactions on Medical Imaging*, *29*(1), 30.

Nair, V., & Hinton, G.E. (2010). Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10), pp. 807–814.

Patenaude, B., Smith, S.M., Kennedy, D.N., Jenkinson, M. (2011). A bayesian model of shape and appearance for subcortical brain segmentation. *Neuroimage*, *56*(3), 907–922.

Pereira, S., Pinto, A., Alves, V., Silva, C.A. (2016). Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Transactions on Medical Imaging*, *35*(5), 1240–1251.

Platero, C., & Tobar, M.C. (2017). Combining a patch-based approach with a non-rigid registration-based label fusion method for the hippocampal segmentation in Alzheimer's disease. *Neuroinformatics*, *15*(2), 165–183.

Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S. (2014). *CNN features off-the-shelf: an astounding baseline for recognition* (Vol. 2014, pp. 512–519). IEEE Conference on: IEEE.

Ronneberger, O., Fischer, P., Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention, Springer, pp. 234–241.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Scoville, W.B., & Milner, B. (2000). Loss of recent memory after bilateral hippocampal lesions. *The Journal of Neuropsychiatry and Clinical Neurosciences*, *12*(1), 103–a.

Sdika, M. (2010). Combining atlas based segmentation and intensity classification with nearest neighbor transform and accuracy weighted vote. *Medical Image Analysis*, *14*(2), 219–226.

Shakeri, M., Tsogkas, S., Ferrante, E., Lippe, S., Kadoury, S., Paragios, N., Kokkinos, I. (2016). Sub-cortical brain structure segmentation using F-CNN's. In: 2016 IEEE 13th international symposium on biomedical imaging (ISBI), IEEE, pp. 269–272.

Shen, D., Moffat, S., Resnick, S.M., Davatzikos, C. (2002). Measuring size and shape of the hippocampus in MR images using a deformable shape model. *NeuroImage*, *15*(2), 422–434.

Sled, J.G., Zijdenbos, A.P., Evans, A.C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging*, *17*(1), 87–97.

Smith, S.M. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, *17*(3), 143–155.

Sun, C., Shrivastava, A., Singh, S., Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In: 2017 IEEE international conference on computer vision (ICCV). IEEE, pp. 843–852.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A. (2017). Inception-v4, Inception-ResNet and the impact of residual connections on learning. In: AAAI conference on artificial intelligence.

Tong, T., Wolz, R., Coupé, P., Hajnal, J.V., Rueckert, D., Initiative, A.D.N., et al. (2013). Segmentation of MR images via discriminative dictionary learning and sparse coding: application to hippocampus labeling. *NeuroImage*, *76*, 11–23.

Wachinger, C., Reuter, M., Klein T (2017). DeepNAT, Deep convolutional neural network for segmenting neuroanatomy. NeuroImage.

Wang, H., & Yushkevich, P.A. (2013). Multi-atlas segmentation with joint label fusion and corrective learning an open source implementation. Frontiers in Neuroinformatics 7.

Wang, H., Das, S.R., Suh, J.W., Altinay, M., Pluta, J., Craige, C., Avants, B., Yushkevich, P.A., Alzheimer's Disease Neuroimaging Initiative, et al. (2011). A learning-based wrapper method to correct systematic errors in automatic image segmentation: consistently improved performance in hippocampus, cortex and brain segmentation. *NeuroImage*, *55*(3), 968–985.

Wang, H., Suh, J.W., Das, S.R., Pluta, J.B., Craige, C., Yushkevich, P.A. (2013). Multi-atlas segmentation with joint label fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(3), 611–623.

Wang, L., Wang, L., Lu, H., Zhang, P., Ruan, X. (2016). Saliency detection with recurrent fully convolutional networks. In: European conference on computer vision, Springer, pp. 825–841.

Wang, Y., Ma, G., Wu, X., Zhou, J. (2018). Patch-based label fusion with structured discriminant embedding for hippocampus segmentation. Neuroinformatics, 1–13.

Yang, J., Staib, L.H., Duncan, J.S. (2004). Neighbor-constrained segmentation with level set based 3-D deformable models. *IEEE Transactions on Medical Imaging*, *23*(8), 940–948.

Yosinski, J., Clune, J., Bengio, Y., Lipson, H. (2014). How transferable are features in deep neural networks? In: Advances in neural information processing systems, pp. 3320–3328.

Zarpalas, D., Gkontra, P., Daras, P., Maglaveras, N. (2014a). Accurate and fully automatic hippocampus segmentation using subject-specific 3D optimal local maps into a hybrid active contour model. *IEEE Journal of Translational Engineering in Health and Medicine*, *2*, 1–16.

Zarpalas, D., Gkontra, P., Daras, P., Maglaveras, N. (2014b). Gradient-based reliability maps for ACM-based segmentation of hippocampus. *IEEE Transactions on Biomedical Engineering*, *61*(4), 1015–1026.

Zhang, W., Li, R., Deng, H., Wang, L., Lin, W., Ji, S., Shen, D. (2015). Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *NeuroImage*, *108*, 214–224.

Zhu, H., Cheng, H., Yang, X., Fan, Y., Initiative, A.D.N., et al. (2017). Metric learning for multi-atlas based segmentation of hippocampus. *Neuroinformatics*, *15*(1), 41–50.