Brain extraction network trained with "silver standard" data and fine-tuned with manual annotation for improved segmentation

Roberto Souza[1,2,*], Oeslle Lucena[3], Mariana Bento[1,2,4], Julia Garrafa[5], Letícia Rittner[6], Simone Appenzeller[5], Roberto Lotufo[6] and Richard Frayne[1,2]

[1]Radiology and Clinical Neurosciences, Hotchkiss Brain Institute, University of Calgary, Canada
[2]Seaman Family MR Research Centre, Foothills Medical Centre, Canada
[3]School of Biomedical Engineering and Imaging Sciences, Faculty of Life Sciences and Medicine, King's College London, UK
[4]Calgary Image Processing and Analysis Centre, Foothills Medical Centre, Canada
[5]Division of Rheumatology, Faculty of Medical Science, University of Campinas, Brazil
[6]Medical Imaging and Computing Laboratory, University of Campinas, Brazil
[*]roberto.medeirosdeso@ucalgary.ca

***Abstract*—Training convolutional neural networks (CNNs) for medical image segmentation often requires large and representative sets of images and their corresponding annotations. Obtaining annotated images usually requires manual intervention, which is expensive and time consuming, as it typically requires a specialist. An alternative approach is to leverage existing automatic segmentation tools and combine them to create consensus-based "silver-standards" annotations. A drawback to this approach is that silver-standards are usually smooth and this smoothness is transmitted to the output segmentation of the network. Our proposal is to use a two-staged approach. First, silver-standard datasets are used to generate a large set of annotated images in order to train the brain extraction network from scratch. Second, fine-tuning is performed using much smaller amounts of manually annotated data so that the network can learn the finer details that are not preserved in the silver-standard data. As an example, our two-staged brain extraction approach has been shown to outperform seven state-of-the-art techniques across three different public datasets. Our results also suggest that CNNs can potentially capture inter-rater annotation variability between experts who annotate the same set of images following the same guidelines, and also adapt to different annotation guidelines.**

I. Introduction

U-nets are fully convolutional neural networks (CNNs) commonly used for biomedical image segmentation [1]. Training these networks requires considerable amounts of annotated data, which is usually generated manually by experts. Manual annotation often is performed following strict guidelines. These sets of annotation guidelines define what structures or sub-structures to include in the final segmentation. Ideally, raters who annotate images by following the same guidelines, should get the same annotation results, regardless of the software used and/or the starting point for the annotation (*i.e.*, annotating from scratch or fixing a mask generated by automatic method A or B). However, in practice inter-rater annotation agreement is seldom perfect. Brain extraction in magnetic resonance (MR) imaging [2]–[9], for example, is a very common brain image processing task, usually used as an initial processing step for more complex analysis, that is built around image segmentation guidelines that include only brain structures (*e.g.*, white and grey matter) and exclude the skull and other brain surrounding tissues.

Lucena *et al.* [10] leveraged existing automatic segmentation techniques and combined them to generate "silver standard" annotations. The drawback to their method was that silver standard annotations tend to be smooth and, thus, omitted finer segmentation details. We investigate a two-staged approach for brain extraction that combines 1) training of a U-net from scratch with silver-standard annotated data, followed by 2) fine-tuning of the network using smaller amounts of manually annotated data.

Our goal in this work is not to find the best architecture for skull-stripping, but rather to assess the impact of our two-staged approach when building data-driven models, such as CNNs. For a more comprehensive analysis of different CNNs architectures for skull-stripping, see [11]–[13] and [14]. Our method compared favorably to seven public, commonly used, skull-stripping techniques [2]–[8]. The two-staged approach also potentially reduced the number of manually annotated datasets necessary to train a network capable of outputting sharp segmentations (*i.e.*, optimal delineation of gyri and sulci). Through this study, we enhanced the previously published *Calgary-Campinas* dataset [9] (https://sites.google.com/view/calgary-campinas-dataset/home) by adding additional manual annotations.

II. Materials

A. Brain MR datasets

Three publicly available datasets were used in this study. Sample renderings of manually segmented data for each

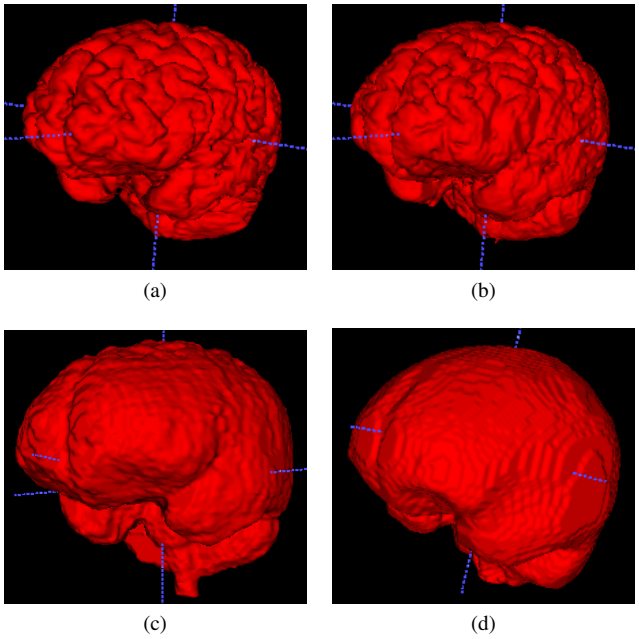


Fig. 1. Sample 3D reconstruction of manual annotations for (a) *CC-12* Manual 1, (b) *CC-12* Manual 2, (c) LPBA40, and (d) OASIS datasets. *CC-12* Manual 1 and Manual 2 annotations were generated using the same guidelines, but with different initialization masks as starting point. Note that *CC-12*, LPBA40, and OASIS were annotated following different guidelines.

dataset are shown in Figure 1.

1) LONI Probabilistic Brain Atlas (LPBA40) [15]: This dataset consists of 40 coronal 3D T1-weighted spoiled gradient echo MR image volumes. Manual annotation (*i.e.*, brain masks) were obtained by manually correcting masks generated with the Brain Extraction Tool (BET) [16]. All image volumes were obtained on a General Electric (GE) 1.5-T scanner.

2) Open Access Series of Imaging Studies (OASIS) [17]: This dataset included 77 subjects, of which 20 were classified as having some degree of cognitive impairment. For each subject, three to four T1-weighted MP-RAGE scans were acquired and co-registered. Manual brain mask annotations were obtained using an in-house method based on atlas registration and specialist review of the masks. The images were collected on a Siemens 1.5-T scanner.

3) Calgary-Campinas-359 (CC-359) [9]: The *CC-359* is a dataset of brain T1 volumetric images acquired in 359 presumed normal subjects. The image volumes were acquired on scanners from three vendors (GE, Philips, Siemens) and at two magnetic field strengths (1.5 T, 3 T). The dataset is balanced - with approximately 60 subjects in each of the six vendor-field strength subgroups. Age and gender balance in each subgroup were comparable. The *CC-359* dataset was divided into 1) a group of 347 subjects who had silver-standard brain masks (*CC-347*) and 2) a group of 12 subjects who had both silver standard and manually annotated brain masks (*CC-12*).

a) CC-347: These image volumes were accompanied by only their silver standard brain masks. These masks were generated by consensus using the simultaneous truth and performance level estimation (STAPLE) algorithm [18]. The input to STAPLE was the output of eight publicly available skull-stripping methods: Advanced Normalization Tools (ANTs) [2], Brain Extraction based on non-local Segmentation Technique (BEaST) [4], BET, Brain Surface Extractor (BSE) [8], Hybrid Watershed Approach (HWA) [7], Marker Based Watershed Scalper (MBWSS) [3], Optimized Brain Extraction (OPTIBET) [6], and Robust Brain Extraction (ROBEX) [5]. The complete consensus building process is described in [9]. We refer to the silver standard brain masks as *STAPLE-automatic*.

b) CC-12: Data from twelve subjects (consisting of two image volumes randomly selected for each of the six vendor-field strength subgroups) were manually and independently annotated by two experts. Both experts annotated using the same guidelines. They were instructed to segment the same brain structures with as much detail (*i.e.*, follow the brain contours) as they could. We refer to the annotated data generated by the two experts as *Manual 1* and *Manual 2*. The experts used different masks generated with automatic skull-stripping methods as a starting point for the annotation procedure. Manual 1 masks were obtained by manually correcting masks generated with BEaST [4]. Manual 2 masks were obtained by correcting the output of a 3D watershed segmentation approach [19]. Both experts used the software ITK-snap [20] to complete their annotation. The STAPLE-derived consensus masks of the two sets of twelve manual annotations generated by each expert was also computed (*STAPLE-manual*). As well, the silver standard brain masks derived from consensus among the eight publicly available skull-stripping methods (*STAPLE-automatic*) were also formed.

B. U-net segmentation

The U-net-like architecture used in the experiments is depicted in Figure 2. It is a 2D architecture that receives a three-channel image as input to incorporate 3D context. The channels correspond to the superior, inferior as well as the slice currently being segmented. An area-open filter [21] implemented in [22] is applied to the resulting 3D output of the network in order to filter out small connected components, leaving only the largest component.

III. Experimental methodology

A two-staged approach was implemented. For all networks, 70% of the data was used for training and 30% was left for validation (*i.e.*, model selection). The networks were trained using coronal images with 64×64 patches. These parameters (image orientation and patch size) and the number of patches were set experimentally. Training methodology (number of patches, and number of training epochs) is summarized Table I. The Dice coefficient overlap metric of the silver standard-trained networks (*Stage 1*) and the fine-tuned networks (*Stage 2*), were computed versus the manually annotated masks and also *STAPLE-manual* and *STAPLE-automatic* on the *CC-12* data. The Dice coefficient served as the training loss function of the network [23].

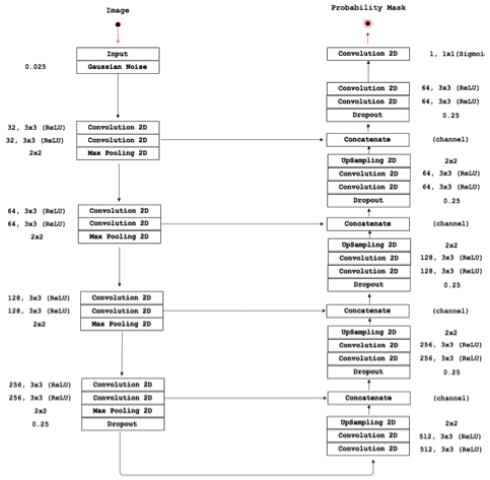


Fig. 2. U-net architecture used in the experiments.

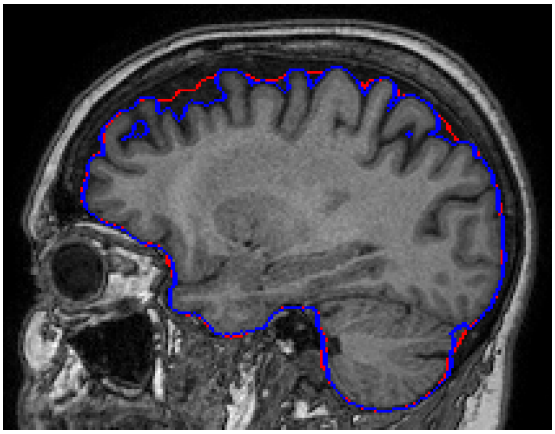


Fig. 3. Representative segmentation contours output of Stage 1 (red) and Stage 2 fine-tuned with Manual 2 (blue). The network after fine-tuning is capable of better following the brain surface curvature.

TABLE I
SUMMARY OF THE TRAINING (SILVER STANDARD, STAGE 1; OR FINE-TUNED, STAGE 2), NUMBER OF PATCHES AND NUMBER OF EPOCHS FOR THE DIFFERENT NETWORKS.

Method	Training	Patches	Epochs
CC-347	scratch	124,920	50
*CC-12**	fine-tune	11,520	25
LPBA40	fine-tune	10,000	25
OASIS	fine-tune	19,000 / 19,500**	25

*The parameters for the four versions of the *CC-12* fine-tuning (Manual 1, Manual 2, *STAPLE-manual*, *STAPLE-automatic* were the same.
**The OASIS dataset has an odd number of images, therefore the number of patches between folds one and two differ slightly.

a) Stage 1: The networks were initially trained from scratch using the silver standard masks available in the *CC-347* dataset. The Stage 1 network was trained using the Adam optimizer [24] with a learning rate of 10^{-3}.

b) Stage 2: Different networks were created by fine-tuning the silver standard-derived network (from Stage 1) using a two-fold cross validation procedure. The fine-tuned networks correspond to fine-tuning with the manual annotations from LPBA40, OASIS, and *CC-12* datasets. In the case of *CC-12*, fine tuning was performed separately with each of the four different annotation datasets (Manual 1, Manual 2, *STAPLE-manual*, and *STAPLE-automatic*, see previous section). The Stage 2 networks were trained using the Adam optimizer with a learning rate of 10^{-4}.

A Wilcoxon signed-rank test [25] was used to assess statistical significant differences between Stage 1 and Stage 2 results. A p-value < 0.05 was deemed statistically significant. We also compared our approach against seven publicly available skull-stripping techniques (ANTs, BEaST, BET, HWA, MBWSS, OPTIBET, and ROBEX). We left BSE out of the comparison, because it has been previously reported as a poorer performing technique [9]. Where appropriate, mean $\pm$ standard deviation are reported.

IV. RESULTS

The computed Dice coefficient results are summarized in Table II. Our fine-tuned techniques (Stage 2) achieved the best results in 4 of 6 (66.7%) comparisons (in the two cases, it ranked second). It was outperformed only by BEaST in the *CC-12* dataset when using the annotations from Manual 1 as the reference mask, though the difference was not significant ($p = 0.16$). Stage 2 performance was better than Stage 1 performance ($p < 0.05$) in all datasets, except *CC-12* when *STAPLE-automatic* was used as the reference annotation mask (difference not significant, $p = 0.34$).

Sample output segmentation results for Stage 1 and Stage 2 are depicted in Figure 3. Sample 3D reconstructions are also shown for Stage 2 fine-tuned with OASIS and fine-tuned with *CC-12 STAPLE-manual* results (Figure 4). The Dice coefficient agreement between Manual 1 and Manual 2 annotations was $96.4\% \pm 0.7\%$.

Dice coefficient curves during Stage 1 and Stage 2 training using the LPBA40 and OASIS dataset are presented in Figure 5. The Dice coefficients training, validation and test curves for Stage 1 using the four annotations of *CC-12* as reference and Stage 2 Dice coefficient curves for the *CC-12* dataset are depicted in Figures 6 and 7, respectively.

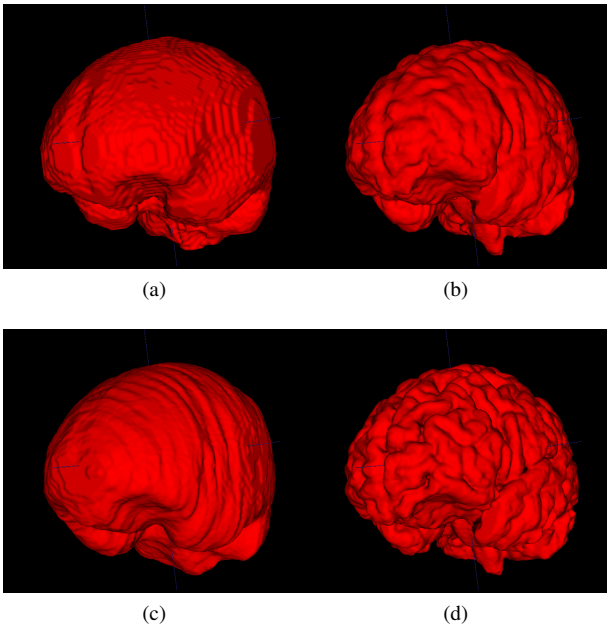
V. DISCUSSION

Our proposed methodology that combines training from scratch using silver standard masks annotations (Stage 1) and then fine-tuning with manual annotations (Stage 2) had the overall best result. The only two scenarios where it was outperformed (ranking second in both cases) were with 1) the *CC-12* (Manual 1) dataset, where it performed worse than BEaST, and 2) with the *CC-12 STAPLE-automatic* dataset, where it was worse than the Stage 1 results. Neither difference, however, was found to be statistically significant ($p > 0.05$). These findings can be explained by the fact that Manual 1 masks were generated by manually adjusting an initial segmentation generated by BEaST, resulting a bias in favor of

TABLE II

PERCENTAGE DICE COEFFICIENT RESULTS (MEAN ± STANDARD DEVIATION). TRUTH WAS MANUAL SEGMENTATION AND *STAPLE-automatic*, IN ONE CC-12 CASE, PROVIDED WITH EACH OF THE THREE DATASETS. THE BEST RESULT FOR EACH DATASET IS EMBOLDENED.

Dataset	***CC-12***				**LPBA40**	**OASIS**
Method	**Man-1**	**Man-2**	**STAPLE-man**	**STAPLE-auto**		
ANTS	96.2 ± 0.9	96.0 ± 0.8	96.1 ± 0.8	96.2 ± 0.9	97.3 ± 0.6	95.3 ± 1.9
BEaST	$\mathbf{97.9 \pm 1.0}$	95.7 ± 1.2	96.8 ± 1.1	95.4 ± 1.7	96.3 ± 0.5	92.5 ± 1.3
BET	94.0 ± 1.6	95.4 ± 1.0	94.7 ± 1.3	97.4 ± 0.5	96.6 ± 0.7	93.5 ± 2.7
HWA	89.7 ± 1.8	91.8 ± 1.3	90.7 ± 1.4	93.8 ± 1.5	92.5 ± 1.2	94.0 ± 1.4
MBWSS	96.2 ± 1.4	95.7 ± 1.4	90.7 ± 1.4	93.8 ± 1.5	92.2 ± 0.8	90.2 ± 4.4
OPTIBET	94.9 ± 1.2	95.6 ± 0.7	95.2 ± 0.9	97.1 ± 0.5	95.9 ± 0.6	94.5 ± 1.1
ROBEX	94.3 ± 1.0	95.9 ± 0.6	95.1 ± 0.7	97.7 ± 0.6	96.8 ± 0.2	95.6 ± 0.8
Stage 1	95.6 ± 1.4	97.0 ± 0.6	96.3 ± 0.9	$\mathbf{98.6 \pm 0.2}$	97.1 ± 0.9	95.4 ± 0.9
Stage 2	97.6 ± 0.5	$\mathbf{97.5 \pm 0.3}$	$\mathbf{97.4 \pm 0.4}$	98.5 ± 0.2	$\mathbf{98.2 \pm 0.2}$	$\mathbf{96.6 \pm 0.5}$



(a) (b) (c) (d)

Fig. 4. Sample 3D reconstruction of the manual annotation and the U-net results for the OASIS dataset. (a) Manual annotation reference, (b) Stage 1, (c) Stage 2 fine-tuned with OASIS and (d) fine-tuned with *CC-12 STAPLE-manual* results. Notice that the network is capable of adapting to two different segmentation guidelines: the one used in OASIS, which gives a rough segmentation, and the one used in *STAPLE-manual*, which tries to follow the brain surface curvature.

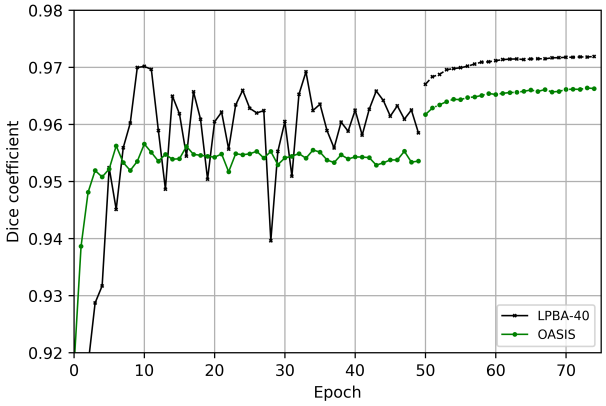


Fig. 5. Dice coefficient test set curves for the LPBA40 and OASIS datasets. The first 50 epochs correspond to training from scratch with the silver standard masks (Stage 1) and the last 25 epochs correspond to the fine-tuning (Stage 2). There is a noticeable effect on the curves when transitioning from Stage 1 to Stage 2, probably due the fact that the number of subjects available for fine-tuning LPBA40 and OASIS are larger compared to *CC-12*.

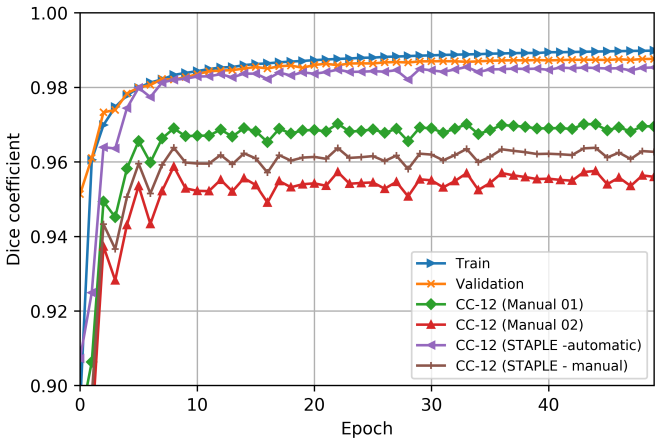
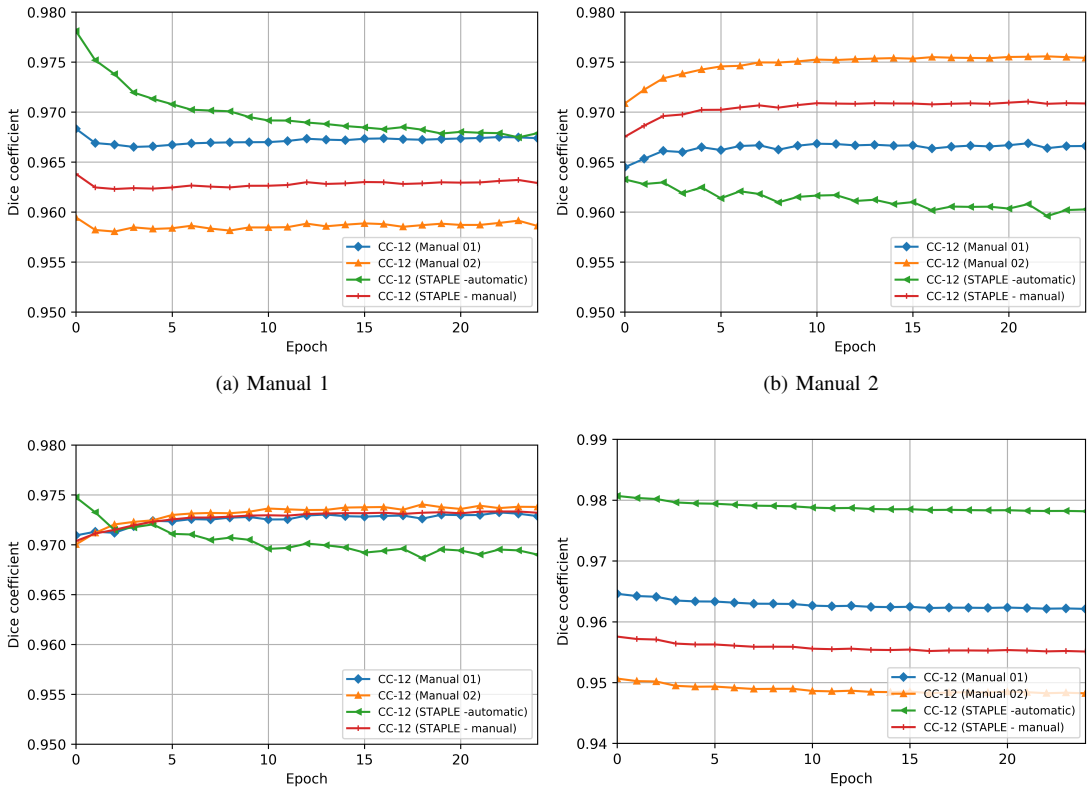


Fig. 6. Training, validation and test curves for the *CC-12* dataset: Dice coefficient curves are shown for the different epochs during the Stage 1 training. The test curve results is higher for *STAPLE-automatic* due to its similar nature to the silver standards used to train the network.

BEaST. In the *CC-12 STAPLE-automatic* case, this result is justified due to the masks used for fine-tuning were generated using the same methodology as the masks used for Stage 1. Also, Stage 1 was trained on 347 images as opposed to the fine-tuning stage (Stage 2) that was trained on only six image volumes in each of the two folds.

Note that the relative ranking of the best skull-stripping techniques changes according to the reference used to compute the Dice coefficient [26] (*e.g.*, Manual 1 or Manual 2), although the experts were performing the task under the same annotation guidelines, the only difference being the mask they used as a starting point for their annotation.

By observing the training, validation and testing Dice coefficient curves for Stage 1 using the different annotations of *CC-12* as reference (Figure 6), we can see that the test curves had a similar shape. As expected in this case, the



(a) Manual 1

(b) Manual 2

(c) *STAPLE-manual*

(d) *STAPLE-automatic*

Fig. 7. Dice coefficient test set curves for the *CC-12* dataset on different epochs of the fine-tuning (Stage 2). Fine-tuning with (a) Manual 1, (b) Manual 2, (c) *STAPLE-manual*, (d) *STAPLE-automatic*. The curves illustrate that the network is able to adapt to the annotation quality used in the fine-tuning step. Manual 1 and Manual 2 curve improves while the *STAPLE-automatic* curve decreases, when fine-tuning with the respective manual annotation. When using *STAPLE-manual*, we can see that Manual 1, Manual 2 and *STAPLE-manual* curves improve, because *STAPLE-manual* is created based on Manual 1 and Manual 2 information. Fine-tuning curves for *STAPLE-automatic* are nearly flat, because theses annotations were similar to the ones used when training from scratch on Stage 1.

curve of *CC-12 STAPLE-automatic* data was higher than the other references. This observation was because its reference annotation was similar to the the silver standard annotations used for its training. The fine-tuning curves of the *CC-12* dataset (Figure 7) illustrate that the network was able to adapt to the details of the specific annotation being used to fine-tune the network. When fine-tuning with Manual 1, the Manual 1 curve improves while the *STAPLE-automatic* curve decreases. The same trend was observed with Manual 2. When using *CC-12 STAPLE-manual*, we can see that Manual 1, Manual 2 and *STAPLE-manual* curves improve, which is expected, because *STAPLE-manual* incorporates Manual 1 and Manual 2 information. *CC-12 STAPLE-automatic* decreased. Finally, when fine-tuning with *CC-12 STAPLE-automatic* the Dice coefficients were nearly flat, because theses annotations were similar to the ones used when training from scratch. This indicates that the network can potentially capture inter-rater annotation variability and adapt to different annotation guidelines. This is supported quantitatively (see Table II) and also by the results presented in Figures 3 and 4; where the amount of detail in the segmentation varies according to the reference (different guideline or different rater) used to fine-tune the network. The results of Stage 1 and Stage 2 when fine-tuned with *CC-12 STAPLE-automatic* had less brain-surface contour detail due the nature of the annotation masks used for training and fine-tuning. The results of Manual 1 and Manual 2 provided more detail. Not surprisingly, *CC-12 STAPLE-manual* captures an amount of detail in between Manual 1 and Manual 2.

It is important to clarify that in practice we do not have access to the test Dice coefficient curves when selecting our model. In this paper, we have only used them to depict network behaviors. The model was always chosen based on the validation set error, and the metrics reported in Table II are computed on the test set. Often the model selected based on the validation set is not the model that will give the best results on the test set, but currently there are no alternatives to tackle this challenging problem.

The same behavior can be seen for LPBA40 and OASIS (*cf.*, Figure 5). In this case, the network starts to fine-tune to the guidelines followed when manually annotating both of these datasets. These results were interesting, and opens up some interesting possibilities. For instance, many brain image analysis software packages, such as FreeSurfer [27] and FSL

[28], can take several hours to process a single subject. A suitably trained network could potentially model the output of these packages in order to yield similar results. The advantage of the network is that it would be much faster. A similar idea has been recently applied to hippocampus segmentation [29].

Another important advantage of our technique is that generating the silver standard masks is relatively inexpensive compared to manual annotation. Therefore, allowing the creation of large silver standard annotated datasets that can be used to train a network from scratch potentially exposing it to a larger variability of data (multi-centre, healthy/diseased), making it more robust and generalizable. Then, smaller amount of manually annotated data can be used to capture finer segmentation details. In our skull-stripping case study, as few as six image volumes (*i.e.*, the number in each fold of *CC-12* processing) was sufficient to fine-tune the network by learning the finer segmentation details.

VI. Conclusions

We presented a two-staged methodology that combines, in Stage 1, silver standards with, in Stage 2, manual annotation to train CNNs for segmentation tasks. While the case study demonstrates advantages for the specific case of skull-stripping, we believe that the two-staged methodology can be generalized to most segmentation tasks. Further, the methodology can potentially capture inter-rater and annotation guideline variability to yield state-of-the-art results with reduced amounts of manually annotated data, as was shown in our skull-stripping case study. The data used in the experiments are publicly available. The annotations Manual 1, Manual 2 and *STAPLE-manual* were incorporated to the *Calgary-Campinas* dataset. In future studies, we would like to evaluate this methodology in other brain structures, specially structures of smaller size to more fully validate our method.

Acknowledgments

The authors would like to thank NVidia for providing a Titan V GPU, Amazon Web Services for access to cloud-based GPU services, FAPESP CEPID-BRAINN (2013/07559-3) and CAPES PVE (88881.062158/2014-01). R.S. was supported by an NSERC CREATE I3T Award and currently holds the T. Chen Fong Fellowship in Medical Imaging from the University of Calgary. R.F. holds the Hopewell Professorship of Brain Imaging at the University of Calgary. L.R. thanks CNPq (308311/2016-7), and R.L. thanks CNPq(311228/2014-3).

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 2015, pp. 234–241.

[2] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, "A reproducible evaluation of ANTs similarity metric performance in brain image registration," *NeuroImage*, vol. 54, no. 3, pp. 2033 – 2044, 2011.

[3] R. Beare, J. Chen, C. Adamson, T. Silk, D. Thompson, J. Yang, V. Anderson, M. Seal, and A. Wood, "Brain extraction using the watershed transform from markers," *Frontiers in Neuroinformatics*, vol. 7, no. 32, December 2013.

[4] S. F. Eskildsen, P. Coupé, V. Fonov, J. V. Manjòn, K. K. Leung, N. Guizard, S. N. Wassef, L. R. Østergaard, and D. L. Collins, "BEaST: Brain extraction based on non-local segmentation technique," *NeuroImage*, vol. 59, no. 3, pp. 2362 – 2373, 2012.

[5] J. E. Iglesias, C. Y. Liu, P. M. Thompson, and Z. Tu, "Robust brain extraction across datasets and comparison with publicly available methods," *IEEE Transactions on Medical Imaging*, vol. 30, no. 9, pp. 1617–1634, Sept 2011.

[6] E. S. Lutkenhoff, M. Rosenberg, J. Chiang, K. Zhang, J. D. Pickard, A. M. Owen, and M. M. Monti, "Optimized brain extraction for pathological brains (OPTIBET)," *PLoS ONE*, vol. 9, no. 12, pp. 1–13, 12 2014.

[7] F. Ségonne, A. M. Dale, B. E. Busa, B. M. Glessner, B. D. Salat, B. H. K. Hahn, and B. F. A, "A hybrid approach to the skull stripping problem in MRI," *NeuroImage*, vol. 22, 2004.

[8] D. W. Shattuck, S. R. Sandor-Leahy, K. A. Schaper, D. A. Rottenberg, and R. M. Leahy, "Magnetic resonance image tissue classification using a partial volume model," *NeuroImage*, vol. 13, no. 5, pp. 856–876, may 2001.

[9] R. Souza, O. Lucena, J. Garrafa, D. Gobbi, M. Saluzzi, S. Appenzeller, L. Rittner, R. Frayne, and R. Lotufo, "An open, multi-vendor, multi-field-strength brain mr dataset and analysis of publicly available skull stripping methods agreement," *NeuroImage*, vol. 170, pp. 482 – 494, 2018.

[10] O. Lucena, R. Souza, L. Rittner, R. Frayne, and R. Lotufo, "Silver standard masks for data augmentation applied to deep-learning-based skull-stripping," in *International Symposium on Biomedical Imaging*. IEEE, 2018.

[11] J. Kleesiek, G. Urban, A. Hubert, D. Schwarz, K. Maier-Hein, M. Bendszus, and A. Biller, "Deep MRI brain extraction: A 3D convolutional neural network for skull stripping," *NeuroImage*, vol. 129, pp. 460–469, 2016.

[12] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Auto-context convolutional neural network (auto-net) for brain extraction in magnetic resonance imaging," *IEEE Transactions on Medical Imaging*, vol. 36, no. 11, pp. 2319–2330, 2017.

[13] G. Zhao, F. Liu, J. A. Oler, M. E. Meyerand, N. H. Kalin, and R. M. Birn, "Bayesian convolutional neural network based mri brain extraction on nonhuman primates," *NeuroImage*, vol. 175, pp. 32–44, 2018.

[14] O. Lucena, R. Souza, L. Rittner, R. Frayne, and R. Lotufo, "Convolutional neural networks for skull-stripping in brain mr imaging using silver standard masks," *Artificial Intelligence in Medicine*, vol. 98, pp. 48 – 58, 2019.

[15] D. W. Shattuck, M. Mirza, V. Adisetiyo, C. Hojatkashani, G. Salamon, K. L. Narr, R. A. Poldrack, R. M. Bilder, and A. W. Toga, "Construction of a 3D probabilistic atlas of human cortical structures," *NeuroImage*, vol. 39, no. 3, pp. 1064–1080, 2008.

[16] S. M. Smith, "Fast robust automated brain extraction," *Human Brain Mapping*, vol. 17, no. 3, pp. 143–155, Nov. 2002.

[17] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open access series of imaging studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults," *J. Cognitive Neuroscience*, vol. 19, no. 9, pp. 1498–1507, Sep. 2007.

[18] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation," *IEEE Transactions on Medical Imaging*, vol. 23, no. 7, pp. 903–921, July 2004.

[19] H. K. Hahn and H. Peitgen, "The skull stripping problem in MRI solved by a single 3D watershed transform," in *Proceedings of the Third International Conference on Medical Image Computing and Computer-Assisted Intervention*, ser. MICCAI '00. London, UK, UK: Springer-Verlag, 2000, pp. 134–143.

[20] P. A. Yushkevich, J. Piven, H. Cody Hazlett, R. Gimpel Smith, S. Ho, J. C. Gee, and G. Gerig, "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *NeuroImage*, vol. 31, no. 3, pp. 1116–1128, 2006.

[21] L. Vincent, "Morphological area openings and closings for grey-scale images," in *Shape in Picture*. Springer, 1994, pp. 197–208.

[22] R. Souza, L. Rittner, R. Machado, and R. Lotufo, "iamxt: Max-tree toolbox for image processing and analysis," *SoftwareX*, vol. 6, pp. 81–84, 2017.

[23] T. Sørensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to

analyses of the vegetation on danish commons," *Biol. Skr.*, vol. 5, pp. 1–34, 1948.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[26] R. Souza, O. Lucena, M. Bento, J. Garrafa, S. Appenzeller, L. Rittner, R. Lotufo, and R. Frayne, "Reliability of using single specialist annotation for designing and evaluating automatic segmentation methods: a skull stripping case study," in *International Symposium on Biomedical Imaging*. IEEE, 2018.

[27] A. M. Dale, B. Fischl, and M. I. Sereno, "Cortical surface-based analysis. I. Segmentation and surface reconstruction," *NeuroImage*, pp. 179–194, 1999.

[28] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, "FSL," *NeuroImage*, vol. 62, no. 2, pp. 782 – 790, 2012.

[29] B. Thyreau, K. Sato, H. Fukuda, and Y. Taki, "Segmentation of the hippocampus by transferring algorithmic knowledge for large cohort processing," *Medical Image Analysis*, vol. 43, pp. 214–228, 2018.