Context Aware 3D CNNs for Brain Tumor Segmentation

Siddhartha Chandra, Maria Vakalopoulou, Lucas Fidon, Enzo Battistella, Théo Estienne, Roger Sun, Charlotte Robert, Eric Deutsch, Nikos Paragios

Context Aware 3D CNNs for Brain Tumor Segmentation

Siddhartha Chandra[1], Maria Vakalopoulou[1,2], Lucas Fidon[1,3], Enzo Battistella[1,2], Théo Estienne[1,2], Roger Sun[1,2], Charlotte Robert[2], Eric Deutsch[2], and Nikos Paragios[1,3]

[1] CVN, CentraleSupélec, Université Paris-Saclay, France
[2] Gustave Roussy Institute, Paris, France
[3] TheraPanacea, Paris, France

Abstract. In this work we propose a novel deep learning based pipeline for the task of brain tumor segmentation. Our pipeline consists of three primary components: (i) a preprocessing stage that exploits histogram standardization to mitigate inaccuracies in measured brain modalities, (ii) a first prediction stage that uses the V-Net deep learning architecture to output dense, per voxel class probabilities, and (iii) a prediction refinement stage that uses a Conditional Random Field (CRF) with a bilateral filtering objective for better context awareness. Additionally, we compare the V-Net architecture with a custom 3D Residual Network architecture, trained on a multi-view strategy, and our ablation experiments indicate that V-Net outperforms the 3D ResNet-18 with all bells and whistles, while fully connected CRFs as post processing, boost the performance of both networks. We report competitive results on the BraTS 2018 validation and test set.

Keywords: Brain Tumor Segmentation · 3-D Fully Convolutional CNNs · Fully-Connected CRFs.

1 Introduction

Cancer is currently the second leading cause of death worldwide with overall 14.1 million new cases and 8.2 million deaths in 2012 [12]. Brain tumors, with gliomas being one of the most frequent malignant types, are among the most aggressive and dangerous types of cancer [5]. According to recent classifications malignant gliomas are classified into four WHO grades. From these low grade gliomas (LGG), including grade I and II are considered as relatively slow-glowing while high grade gliomas (HGG), including grade III and grade IV glioblastoma are more aggressive with the average survival time of approximately 1 year for patients with glioblastoma (GBM) [13, 21]. Besides being very aggressive, gliomas are very costly to treat, so accurately diagnosing of them at early stages is very important.

Multimodality magnetic resonance imaging (MRI) is the primary method of screening and diagnosis for gliomas. However, due to inconsistency and diversity of MRI acquisition parameters and sequences, there are large differences

in appearance, shape and intensity ranges, adding variability to the one that gliomas can have between different patients. Currently, tumor regions are segmented manually by radiologists, but this process is very time consuming while the inter-observer agreement between them is considerably low. In order to address all these challenges, the multimodal brain tumor segmentation challenge (BraTS) [22, 1–3] is organized annually, in order to highlight efficient approaches and the way forward for the accurate segmentation of brain tumors.

Currently, the emergence of deep learning as disruptive innovation method in the field of computer vision has impacted significantly the medical imaging community, with numerous architectures being proposed addressing task-specific problems. Fully Convolutional Networks (FCN) [20] and their extension to 3D [23, 14] are among the most commonly used architectures, boosting considerably the accuracies of the semantic segmentation. Inspired by these recent advances of deep learning, in this paper we exploit 3D CNNs coupled with fully-connected Conditional Random Fields for segmentation of brain tumor. More specifically, we compare two popular network architectures: V-Net [23] and 3D Residual-Nets [16] (ResNet), trained using a multi-view strategy, and provide preliminary results which indicate that V-Net architecture is better suited for dense-per-voxel brain tumor segmentation.

In the next sections, we discuss our contributions in detail, and we report our performance on the Training, Validation and Test Dataset of BraTS 2018.

2 Context-Aware 3D Networks

In this section, we give an overview of the different methods and strategies (Fig. 1,Fig. 2) we follow and we discuss in detail the different components of our pipeline.

2.1 Preprocessing using Histogram Standardization

MRI is the most popular medical imaging tool to capture the images of the brain and other internal organs. It is preferred due to its non-invasive nature and its ability to capture diverse types of tissues and physiological processes. It measures the response of body tissues to high-frequency radio waves when placed in a strong magnetic field, to produce images of the internal organs. MRI scans typically suffer with a bias due to artefacts produced by inhomogeneity in the magnetic field or small movements made by the patient during acquisition. Since MRI intensities are expressed in arbitrary units and may vary across acquisitions, this bias can adversely impact segmentation algorithms. State-of-the-art approaches typically employ bias correction strategies to pre-process the data corresponding to different modalities in order to mitigate this bias.

After careful comparison of existing bias-correction literature, we decide to use the recently proposed histogram standardizing approach [24] for bias correction. The authors in [24] propose a two phase algorithm that exploits the statistics of the different modalities in a dataset to transform the dataset in a

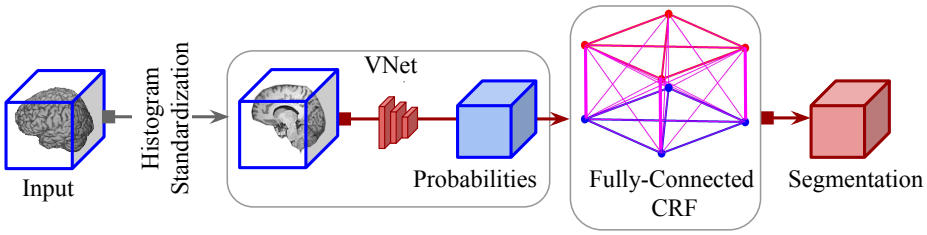

Fig. 1. A schematic overview of our approach. We first perform bias correction in the input brain volume using histogram standardization. A V-Net architecture is then trained on these data to deliver first phase of segmentation prediction. Further, we use a bilateral filtering performing fully-connected CRF to post-process our network predictions.

manner where similar intensities correspond to similarity in the tissue semantics. We pre-process all our data in this work using this strategy.

2.2 V-Net for 3D Semantic Segmentation

The first prediction stage in this work uses the V-Net architecture introduced by Milletari *et al.* in [23]. The V-Net architecture is a 3D fully convolutional neural network which can be trained end-to-end to deliver dense, per voxel class probabilities. The V-Net architecture has been exploited in literature for a variety of 3D segmentation tasks. Further, we use the generalized dice overlap loss as presented in [28] to optimise V-Net, which is a surrogate for the Dice coefficient used for evaluation. Using this loss function for training alleviates the need to compensate for the imbalance between the number of training samples for the different classes. We encourage the readers to refer to the original paper [23] for details on the network architecture.

2.3 Custom 3D ResNets for Semantic Segmentation

In this section, we discuss the 3D ResNets, as an alternative to the V-Net architecture described above. Residual networks were introduced by He *et al.* in [16]. ResNets ease the training of networks by adding 'residual' connections to the network architecture. These residual connections induce a short-cut connection of identity mapping without adding any extra parameters or computational complexity, thereby recasting the original mapping $F(x)$ as $F(x) = F'(x) + x$. We encourage the readers to refer to the original paper [16] for details.

ResNets are the building blocks of the majority of approaches on a variety of computer vision image segmentation benchmarks [11, 32, 31, 14], and thus were a natural starting point in this work. However, the vanilla ResNets lack certain desirable characteristics which make their application to the task of brain tumor segmentation challenging. For the BraTS 2018 benchmark, we addressed these challenges by extending the 3D ResNet architecture from [14]. We briefly discuss these challenges one by one and describe our strategies to overcome them.

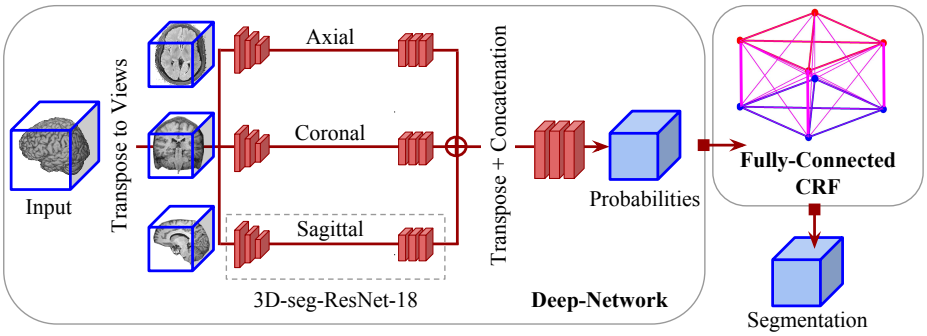

Fig. 2. Overview of our pipeline with 3D ResNets. Our network consists of three parallel ResNet-18 branches, each computing mid-level features on one of the axial, sagittal and coronal views of the input. These mid-level features are fused by transposing to a common view and concatenation. Linear classifiers on top are trained to produce probabilities for each of whole tumor, tumor core and enhancing tumor categories. Further, we use a bilateral filtering performing fully-connected CRF to post-process our network predictions.

Network Stride. Approaches to semantic segmentation use 'fully-convolutional networks' (FCNs) [20, 19] which are networks composed entirely of stacks of convolution operations, thereby producing per-patch outputs which spatially correspond to patches in the input image. A major challenge that presents itself in the use of FCNs is the *network stride*, also referred to as the *downsampling factor*. The output activations delivered by FCNs are smaller in spatial size than the input image due to the repeated max-pooling and convolutional strides. Thus, obtaining a labeling that is the same size as the input image requires upsampling of the output scores via interpolation, resulting in quantization and approximation errors and over-smooth predictions which do not capture the finer details in the input.

The downsampling factor of ResNets, like other popular network architectures such as [18, 27] is 32. This means that each output unit corresponds to a 32×32 patch in the input image. For the BraTS 2018 data where the size of the input volume is $240 \times 240 \times 155$, the vanilla ResNet delivers outputs of the size $8 \times 8 \times 5$. A popular approach to reduce the downsampling factor is using a deconvolution filter which is a backwards convolution operation to upsample the output, as proposed by Long and Shelhamer in [20]. However, this results in an increased number of parameters, which will lead to overfitting for smaller datasets like the BraTS 2018 dataset where obtaining pixel-accurate ground truth is tedious.

In this work, we use atrous convolutions proposed by Chen *et al.* [7]. The atrous algorithm introduces holes in the convolution kernel, thereby allowing us to reduce the loss in spatial resolution without any increase in the number of parameters. Authors in [30] use the same operation, rebranding it as 'dilated convolutions'. With a strategic use of atrous convolutions, we reduce the down-

sampling factor of ResNets to 4. This amounts to an output of size $60 \times 60 \times 39$ for the BraTS 2018 data.

Context Awareness. Standard deep networks do not have a built-in capacity to estimate the scale of the input [8]. This limitation becomes especially crippling for brain tumor segmentation where the scales of the whole tumor, tumor core and enhancing tumor categories depend on a variety of factors, therefore estimating the correct scale of tumors is a challenging task. Approaches typically address this shortcoming by feeding the input to the network at different scales and averaging the network responses across scales [8, 9]. A number of recent methods have proposed using feature pyramids [32, 10, 7] which instead capture features at multiple scales. The feature pyramids are finally fused into a single feature map via element-wise maximization, averaging or concatenation. In this work, we use the atrous spatial pyramid pooling (ASPP) approach proposed in [10]. ASPP uses a stack of convolutional filters with increasing degrees of dilation, thereby simulating filtering at multiple sampling rates and receptive fields. This captures visual context at multiple scales and leads to performance boosts for a variety of segmentation benchmarks [7]. The features at different scales are fused via averaging. This strategy enhances the context-awareness of the network.

Richness of Features (Network depth) vs Training/Inference speed. Deeper networks typically learn richer, more meaningful features as indicated by performance boosts over shallower networks [16, 15]. However, an increase in depth also increases training / inference time because the network represents a sequential directed acyclic graph and prior activations need to be computed before subsequent ones.

3D FCNs are much slower than their 2D counterparts. Unlike 2D convolutions which have benefitted from both software and hardware level optimizations, 3D convolutions still involve slow computations as the research into their optimization is in its infancy. To allow fast experimentation and validation, the network architecture design needs careful consideration.

The authors in [31] demonstrate that decreasing the depth and increasing the width of ResNets leads to both better accuracy and reduced training / testing time. Inspired by them, rather than using very deep ResNets, we use the smallest residual network ResNet-18 in our experiments. To increase the width of the network, we use a multi-view fusion architecture where our network has three branches, each computing features on one of axial, sagittal and coronal views. The features from the three views are transposed to a common view and concatenated, and linear classifiers for the three categories whole tumor, tumor core and enhancing tumor are trained on the fused features. This increases the speed at which the network operates as the activations of the three branches of the network can be computed in parallel. Here, we want to emphasize that each of these three branches is using the 3D input, in contrast to the 2.5D methods [26]. Further our preliminary experiments indicate that this multi-view fusion

leads to better performance on a validation set, compared to deeper variants: ResNet-34 and ResNet-50. Our approach is described in Fig. 2.

2.4 Fully-Connected Conditional Random Fields.

Fully convolutional deep networks such as V-Net and ResNets that produce per-voxel predictions consist of several downsampling phases followed by several upsampling phases. These phases involve quantization and approximations due to which these pipelines typically produce oversmooth predictions which do not capture the finer details in the input data. To address this limitation, we follow up the first pass of prediction using the network with a post-processing refinement pass. The refinement of the network prediction is done using a fully-connected Conditional Random Field (CRF). The fully-connected CRF performs bilateral filtering to refine the predictions made by our network, and uses the objective function proposed in [17]. Precisely, the CRF expresses the energy of a fully-connected CRF model as the sum of unary and pairwise potentials given by

$$E_I(\mathbf{l}) = \sum_i \psi_u(l_i) + \sum_i \sum_{j<i} \psi_p(l_i, l_j), \tag{1}$$

where

$$\psi_p(l_i, l_j) = \mu(l_i, l_j) \sum_{m=1}^{K} w_m^1 \exp(\underbrace{-\frac{|s_i - s_j|^2}{2\theta_\alpha^2} - \frac{|p_i - p_j|^2}{2\theta_\beta^2}}_{\text{appearance}}) + w_m^2 \exp(\underbrace{-\frac{|s_i - s_j|^2}{2\theta_\gamma^2}}_{\text{smoothness}}). \tag{2}$$

Here $\mathbf{l} = \{l_i\}$ denotes the labels for all the pixels indexed by i coming from a set of candidate labels $l_i \in \{1, 2, \ldots, L\}$. ψ_u denotes the image dependent unary potentials, and the image dependent pairwise potentials $\psi_p(l_i, l_j)$ are expressed by the product of a label compatibility function μ and a weighted sum over Gaussian kernels. The pixel intensities are expressed using the 4 modalities in the input data $p_i = (\textit{flair}, \textit{t1}, \textit{t2}, \textit{t1ce})$ and spatial positions are simply the coordinates in $3D$ space $s_i = (\mathtt{x}, \mathtt{y}, \mathtt{z})$. These are used together to define the appearance kernel, and the spatial positions alone are used to define the smoothness kernel. The appearance kernel tries to assign the same class labels to nearby pixels with similar intensity, and the hyperparameters θ_α and θ_β control the degrees of nearness and similarity. The smoothness kernel aims to remove small isolated regions. The model parameters $(\theta_\alpha, \theta_\beta, \theta_\gamma, w_m^1, w_m^2)$ are set by doing parameter sweeps using a validation set.

Having discussed our method in detail, we now delve into the experimental details and results in the next section.

3 Experiments and Results

3.1 Training Protocol

As described in Sec. 2, we use histogram standardization [24] for data preprocessing.

V-Net.
We train the V-Net from scratch on randomly cropped 3D patches, as presented in Fig. 1, of size $128 \times 128 \times 128$ voxels. We do not employ any other form of data augmentation. Our network takes as input all 4 input modalities ($\textit{flair}, \textit{t1}, \textit{t2}, \textit{t1ce}$) and is trained using the generalized Dice loss [28] to output class probabilities for the 3 classes in the dataset alongside 2 additional classes (void and background / healthy tissue). We use the standard stochastic gradient descent algorithm for training, with a weight-decay of $1 \times e^{-5}$ and momentum of 0.9. We use a polynomially decaying learning rate policy, with a starting learning-rate of $1 \times e^{-4}$ and we train for $10K$ iterations. Our implementation uses the pytorch [25] library.

ResNet-18.
As described in Sec. 2 and in Fig. 2, the three branches of our 3D ResNet-18 are initialized from the 3D ResNet-18 network from [14] which was trained for action recognition in videos. We augment the first convolutional layer (conv1) of the network from [14] with an additional input channel since we have 4 modalities (flair,t1,t2,t1ce), as opposed to 3 channels in natural images (r, g, b). We train our networks with randomly sampled input patches of size $97 \times 97 \times 97$, and our network outputs predictions of size $25 \times 25 \times 25$. The input brain volume is preprosessed by subtracting the per-image mean for each modality independently. We use the weighted Softmax Cross-Entropy loss to train our network for three classes: whole tumor, tumor core, and enhancing tumor. The weights for these three classes are 5,10,10 respectively. We use random flipping across the axial axis, and random scaling of the input between scales $0.25 - 2.5$ for data augmentation. We use the standard stochastic gradient descent algorithm for training, with a weight-decay of $1 \times e^{-5}$ and momentum of 0.9. We use a polynomially decaying learning rate policy, with a starting learning-rate of $1 \times e^{-4}$ and we train for $100K$ iterations. Our implementation is based on the Caffe2 library.

Fully connected CRF.
Our CRF parameters (Sec. 2.4) are estimated using the validation set of 85 patients. We use these probabilities as unary terms along hand-crafted pairwise terms (Sec. 2.4) for the CRF post-processing.

3.2 Testing Protocol

For results on the training set, we use a random train-test split of $200 - 85$ patients respectively. For results on the validation and test sets, we use all the

285 training patients to train, and evaluate on the 66 validation, and 191 test patients.

V-Net. Our testing is done in a sliding window fashion on 3D patches of size $128 \times 128 \times 128$ voxels and predictions of ovelapping voxels are obtained via averaging.
ResNet-18. Our testing is done in a sliding window fashion on 3D patches of size $97 \times 97 \times 97$ voxels and predictions of ovelapping voxels are obtained via averaging. We use multi-scale testing alongside flipping along the axial plane and average the probabilities delivered by the network.

3.3 Results

Our results on the BraTS 2018 Validation and Test datasets are tabulated in Tab. 1 and Tab. 2 respectively. On the Validation set, we compare the two network architectures we considered, ResNet-18 and V-Net, with and without using CRF post-processing. Our best results on the validation set are achieved when we use V-Net followed by CRF post-processing. For this reason our final submission Tab. 2 for the test dataset of BraTS 2018 have been performed using the V-Net architecture. These results were generated by the evaluation server on the official BraTS 2018 website and have been also summarized in [4]. Qualitative results are shown in Fig. 3 and Fig. 4.

method	**Dice**			**Sensitivity**			**Specificity**			**Hausdorff95**		
	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC
ResNet	0.740	0.868	0.801	0.771	0.811	0.769	0.991	0.992	0.997	5.312	4.971	9.891
ResNet+CRF	0.741	0.872	0.799	0.795	0.829	0.789	0.997	0.994	0.997	5.575	5.038	9.588
V-Net	0.766	0.896	0.810	0.821	0.909	0.815	0.992	0.989	0.952	7.211	6.541	7.821
V-Net+CRF	**0.767**	**0.901**	**0.813**	**0.839**	**0.916**	**0.819**	**0.998**	**0.994**	**0.997**	**7.569**	**6.68**	**7.630**

Table 1. Results on BraTS 2018 Validation dataset.

Label	**Dice**			**Hausdorff95**		
	ET	WT	TC	ET	WT	TC
Mean	0.61824	0.82991	0.73334	24.93432	20.45375	26.48868
StdDev	0.3083	0.16348	0.27445	33.86977	26.42336	31.0645
Median	0.75368	0.88719	0.85481	4.12311	6.16441	8.66025
25 quantile	0.48567	0.82071	0.65831	2.20361	3.60555	3.0
75 quantile	0.84363	0.92246	0.91996	49.78338	28.60328	47.69619

Table 2. Results on BraTS 2018 Test dataset.

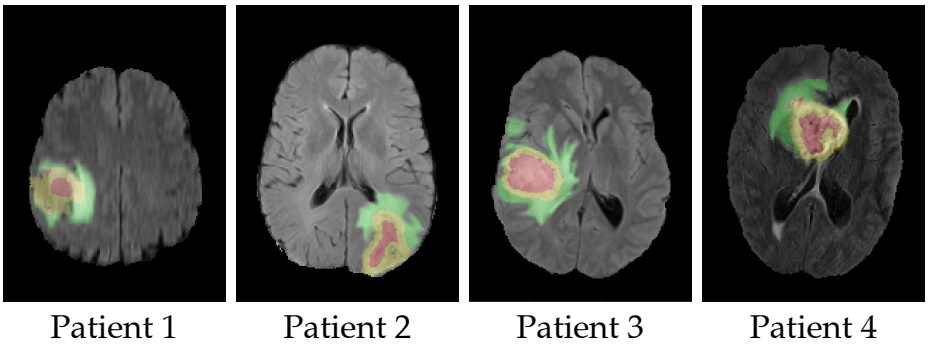

Fig. 3. Example segmentations on the Brats 2018 Validation set delivered by our approach for four patients. Green: edema, Red: non-enhancing tumor core; Yellow: enhancing tumor core.

4 Conclusions and Future Work

In this work, we have described a novel deep-learning architecture for automatic brain tumor segmentation. More specifically, our pipeline uses histogram standardization for input bias correction and uses the V-Net architecture for the first phase of segmentation prediction. We also describe a fully-connected CRF to refine the network outputs in a post-processing step, while we also investigate the use of the multiview approach to fuse 3D features. Our approach delivers competitive results on the BraTS 2018 dataset. In the future, we would like to incorporate spatial pyramids for richer feature representation, and adapt techniques that perform data augmentation in a natural way as persented in [29]. Finally, we will try to investigate techniques that integrate CRFs into the network training [6].

References

1. S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, J. Freymann, K. Farahani, and C. Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4, 9 2017.
2. S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, J. Freymann, K. Farahani, and C. Davatzikos. Segmentation labels and radiomic features for the pre-operative scans of the tcga-gbm collection, 07 2017.
3. S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, J. Freymann, K. Farahani, and C. Davatzikos. Segmentation labels and radiomic features for the pre-operative scans of the tcga-lgg collection, 07 2017.
4. S. Bakas, M. Reyes, et Int, and B. Menze. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *CoRR*, abs/1811.02629, 2018.
5. E. C. Holland. Progenitor cells and glioma formation. 14:683–8, 01 2002.

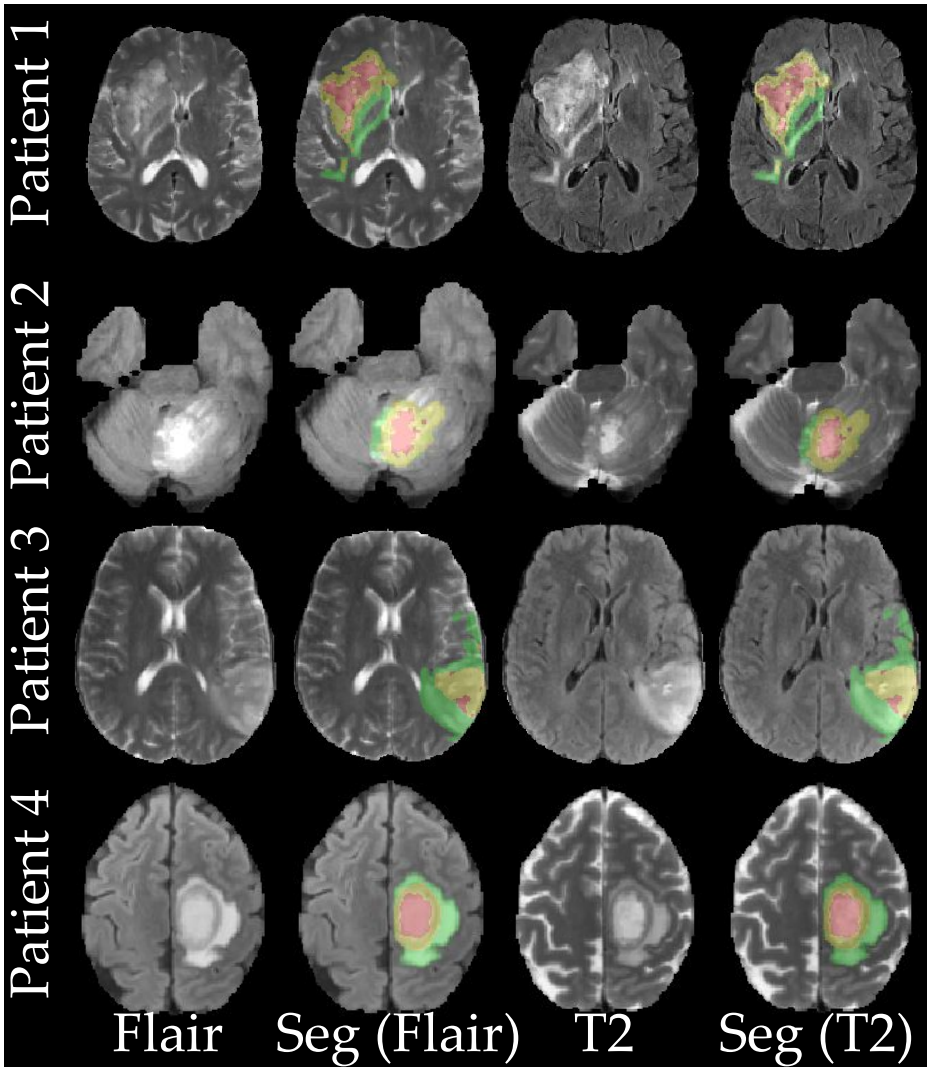

Fig. 4. Example of segmentations corresponding to four different patients, superimposed on Flair and T2 modalities on the Brats 2018 Test set delivered by our approach. Green: edema, Red: non-enhancing tumor core; Yellow: enhancing tumor core.

6. S. Chandra and I. Kokkinos. Fast, exact and multi-scale inference for semantic image segmentation with deep gaussian crfs. In *ECCV*, 2016.
7. L. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017.
8. L. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. *CVPR*, 2016.
9. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv*

preprint arXiv:1412.7062, 2014.
10. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv:1606.00915*, 2016.
11. L.-C. Chen, G. Papandreou, K. Murphy, and A. L. Yuille. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. *ICCV*, 2015.
12. J. Ferlay, I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. Parkin, D. Forman, and F. Bray. Cancer incidence and mortality worldwide, 2013.
13. M. Hadziahmetovic, K. Shirai, and A. Chakravarti. Recent advancements in multimodality treatment of gliomas. *Future oncology*, 7 10:1169–83, 2011.
14. K. Hara, H. Kataoka, and Y. Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, 2018.
15. K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. *ICCV*, 2017.
16. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
17. P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011.
18. A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.
19. Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, 1998.
20. J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.
21. D. N. Louis, A. Perry, G. Reifenberger, A. von Deimling, D. Figarella-Branger, W. K. Cavenee, H. Ohgaki, O. D. Wiestler, P. Kleihues, and D. W. Ellison. The 2016 world health organization classification of tumors of the central nervous system: a summary. *Acta Neuropathologica*, 131(6):803–820, Jun 2016.
22. B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M. A. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, G. Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz, R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H. C. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. V. Leemput. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10), 2015.
23. F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 565–571. IEEE, 2016.
24. L. G. Nyúl, J. K. Udupa, and X. Zhang. New variants of a method of mri scale standardization. *IEEE transactions on medical imaging*, 19(2):143–150, 2000.
25. A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.
26. H. R. Roth, L. Lu, A. Seff, K. M. Cherry, J. Hoffman, S. Wang, J. Liu, E. Turkbey, and R. M. Summers. A new 2.5d representation for lymph node detection using

random sets of deep convolutional neural network observations. In P. Golland, N. Hata, C. Barillot, J. Hornegger, and R. Howe, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014*, Cham, 2014. Springer International Publishing.
27. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.
28. C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 240–248. Springer, 2017.
29. M. Vakalopoulou, G. Chassagnon, N. Bus, R. Marini, E. I. Zacharaki, M.-P. Revel, and N. Paragios. Atlasnet: Multi-atlas non-linear deep networks for medical image segmentation. In A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, Cham, 2018. Springer International Publishing.
30. F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *ICLR*, 2016.
31. S. Zagoruyko and N. Komodakis. Wide residual networks. In *BMVC*, 2016.
32. H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. *CoRR*, abs/1612.01105, 2016.