DEEP NEURAL NETWORKS COMPARISON FOR MRI SEGMENTATION OF THE BRAINSTEM

Seoyoung Oh, Mélanie Pélégrini-Issac, Hélène Urien, Véronique Marchand-Pauvert, Jérémie Sublime

DEEP NEURAL NETWORKS COMPARISON FOR MRI SEGMENTATION OF THE BRAINSTEM

Seoyoung Oh[⋆†]*, Mélanie Pélégrini-Issac*[†]*, Hélène Urien*[⋆]*, Véronique Marchand-Pauvert*[†]*, Jérémie Sublime*[⋆]

[⋆] ISEP - School of Digital Engineers, Paris, France
[†] Sorbonne Université, Inserm, CNRS, Laboratoire d'Imagerie Biomédicale, LIB, Paris, France

ABSTRACT

Deep learning networks are the standard for medical image segmentation, yet the network architectures in medical applications are poorly understood. A precise segmentation of the brainstem is crucial in neurological conditions like Amyotrophic Lateral Sclerosis (ALS), which is a rare neurodegenerative disease affecting respiratory muscles by weakening motor neurons in the brain and spinal cord, but it is challenging due to the lack and low resolution of Magnetic Resonance Imaging (MRI) data. In this context, this paper explores neural network properties for brainstem segmentation and presents an efficient model with strong results. We find that minimal gains come from transfer learning in the encoder while optimizing the decoder and loss function improves performance. Our work also provides valuable insights into model components for MRI segmentation of the brainstem.

***Index Terms*—** MRI, Brainstem, Segmentation, Deep Learning, Neural Network Properties

1. INTRODUCTION

Deep learning has been used extensively in advancing medical research, yet selecting and designing appropriate models for specific challenges like brainstem nuclei segmentation from MRI data remains complex [1]. This segmentation task is pivotal for the diagnosis and the prognosis of the functional outcome, e.g. in neurodegenerative diseases.

ALS is a progressive neurodegenerative disease affecting nerve cells in the brain and spinal cord, causing loss of muscle control and respiratory system degeneration. MRI is crucial for diagnosing and predicting disease evolution [2, 3], particularly in the brainstem, which plays a major role in motor neuron pathways and breathing. However, studying brainstem structures *in vivo* in clinical MRI settings is challenging due to the limited acquisition time, resulting in low-resolution images (Figure 1). Most studies on segmentation of brainstem nuclei have used high-spatial-resolution *ex vivo* MRI with hours of acquisition to obtain data, making them challenging to apply to *in vivo* MRI [1, 4].

Code is available at: `https://doi.org/10.5281/zenodo.10657987`

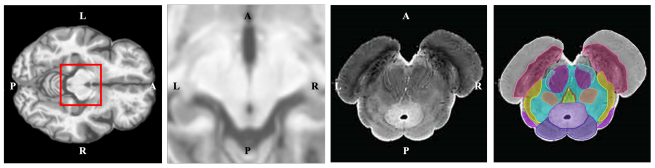

Fig. 1. (a) Sample axial slice of an *in vivo* MR T_1-weighted structural image, spatially registered to a standard space; (b) crop to a size of 50×50 1-mm isotropic voxels of the red pan in (a), corresponding to the brainstem; (c) high-resolution *ex vivo* T_2-weighted MR axial slice of the brainstem with 50-micron isotropic voxels; (d) detail of the brainstem nuclei, which are hard to visualize in (b). A: anterior; P: posterior; R: right; L: left; (c)-(d): adapted from [4].

In this paper, we propose a new segmentation model for *in vivo* MRI data, N-DecoNet, based on deep learning methods. We perform a fine-grained study of the components of deep learning models for brainstem MRI segmentation. Our main contributions are:

- We introduce a new simple and powerful segmentation model that shows competitive results while using fewer parameters than previously proposed models for the segmentation of brainstem nuclei from MRI data.
- We show that transfer learning does not improve segmentation performance on medical images, but properly configuring the decoder and loss function does.
- We propose guidelines for designing segmentation models for brainstem MRI through in-depth evaluation of multiple neural network architectures.

2. DEEP LEARNING NETWORKS

Semantic segmentation classifies image pixels into their respective classes. It typically involves encoder-decoder network models [5].

An encoder network performs convolution with a filter bank to produce a set of feature maps to capture higher semantic information. Frequently used classification networks include: ResNet [6], SENet [7] containing additional channel-wise attention mechanisms, EfficientNet [8] using the compound scaling method to build a more complicated network,

and transformer-based networks such as MiT [9].

Decoder networks are used to rebuild the original image based on the features passed by the encoder, and various models have been proposed depending on how this process is configured. They can be classified into **UNet-type** and **Pyramid-type** depending on the structure of the decoder.

UNet-type models refer to methodologies devised based on U-Net [10], and have two most prominent features. First, the feature map is gradually expanded in each stage of the decoder. Second, a skip connection method is used to combine the low-level detailed feature maps from the encoder with the high-level semantic feature maps of the decoder. This model has been improved by making various changes to the skip connections. To compensate for the weak connection of feature maps between the encoder and decoder, which causes signals to be gradually diluted due to multiple downsampling and upsampling operations, UNet++ [11] adds several sub-convolution blocks to the skip connections to capture fine-grained details. MAnet [12] applies an attention mechanism to skip connections.

Pyramid-type models stack and combine feature maps of different resolutions that can be obtained from the network. Networks of this type include PSPnet [13], which introduces the pyramid pooling method to extract feature maps of different sizes, FPN [14], which takes the last feature maps of each stage of the encoder to combine the pyramid module, DeepLabV3 [15] and DeepLabV3+ [16], which integrate separable convolutions with dilation between kernel elements, which helps expand the receptive field without increasing the number of parameters.

UNet-type models are commonly used for brainstem region segmentation in larger areas than the nuclei [17], while pyramid-type models are less frequently applied to medical image studies [18, 19]. Few studies have explored the use of encoder models for medical images [20], and research on decoder networks for medical image segmentation remains limited, hindering the development of new models for MRI segmentation of the brainstem.

3. METHOD

Our segmentation model, N-DecoNet, consists of two networks (Figure 2): an encoder network pre-trained with ImageNet [21] for classification tasks and a decoder network newly designed for the given segmentation task.

The decoder network takes feature maps from the last four pooling layers of the encoder network. It performs convolutions and upsamples the feature maps to make all feature maps the same size. The high-dimensional feature representation acquired by summing all feature maps is fed to a trainable softmax classifier [22]. This softmax classifies each pixel independently. The output of the softmax classifier is a K-channel image of probabilities, where K is the number of classes.

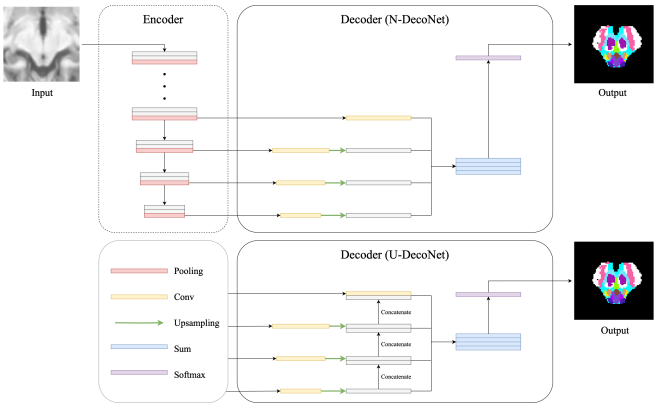

Fig. 2. Two types of decoder networks proposed for brainstem neuroanatomy segmentation: N-DecoNet (top), and U-DecoNet (bottom). Any pretrained classification network can be used as an encoder. N-DecoNet and U-DecoNet use the feature maps from the last four pooling steps. The colors represent the different compositions of the model.

N-DecoNet reduces the loss of low-level features trained from the encoder network by removing connections between each layer of the decoder. It can be trained efficiently with fewer parameters than UNet-type models that perform additional calculations before upsampling and add more complex structures to the skip connection. Also, unlike most pyramid-type models, it preserves more information to recover feature maps to the original image size by using multiple pooling stages.

In this paper, to check the impact of UNet-type operations on training, we also implemented a version of our model with an extra step (denoted as U-DecoNet): the upsampled feature maps of the previous levels are added together.

Model training minimizes the DiceCE loss, which is defined by linearly combining Dice loss [23] and cross entropy (CE) loss [24] with label smoothing (denoted as SoftCE). For experiments, we define DiceFocal loss, which is a compound loss function of Dice loss and Focal loss [25]:

$$\mathcal{L}_{\text{DiceCE}} = \alpha\mathcal{L}_{\text{SoftCE}} + (1-\alpha)\mathcal{L}_{\text{Dice}} \quad (1)$$

$$\mathcal{L}_{\text{DiceFocal}} = \alpha\mathcal{L}_{\text{Focal}} + (1-\alpha)\mathcal{L}_{\text{Dice}} \quad (2)$$

where $\alpha \in (0, 1)$.

4. EXPERIMENTS

We selected multiple neural network architectures and evaluated their performances on the given task. First, we trained different models with the same encoder with different loss criteria. Then we evaluated models with different pre-trained encoder networks with the same loss function. As far as we know, exploration of the impact of decoder architecture and loss function on medical image segmentation tasks is limited. This line of analysis is especially important for MRI segmentation of the brainstem.

Table 1. The number of training parameters in millions (M) and macro mIoU results of each model when it trains with ResNet-34 encoder and diverse loss functions. Macro mIoU was estimated with 95% confidence interval.

DECODER	PARAMS	DICE [23]	JACCARD [26]	FOCAL [25]	SOFTCE [24]	DICEFOCAL	DICECE
UNET [10]	24.4M	0.359 ± 3.4E-2	0.329 ± 2.0E-2	0.605 ± 1.3E-3	0.607 ± 1.2E-2	0.610 ± 1.7E-3	**0.684 ± 2.9E-2**
UNET++ [11]	26.1M	0.368 ± 4.3E-2	0.371 ± 2.1E-2	0.580 ± 5.6E-3	0.614 ± 9.3E-3	0.587 ± 8.4E-3	**0.646 ± 6.9E-3**
MANET [12]	31.8M	0.283 ± 3.6E-2	0.220 ± 7.8E-2	0.605 ± 6.2E-3	0.604 ± 4.6E-4	0.604 ± 1.1E-2	**0.610 ± 4.2E-2**
PSPNET [13]	21.6M	0.643 ± 1.0E-2	0.610 ± 3.3E-3	0.687 ± 3.9E-4	0.740 ± 6.5E-4	0.750 ± 2.0E-4	**0.756 ± 3.2E-3**
FPN [14]	23.2M	0.674 ± 1.5E-2	0.641 ± 1.3E-2	0.628 ± 2.5E-4	0.759 ± 7.0E-4	0.643 ± 6.8E-4	**0.777 ± 1.2E-3**
DEEPLABV3 [15]	26.0M	0.667 ± 5.1E-3	0.643 ± 1.4E-2	0.662 ± 2.5E-3	0.758 ± 1.5E-3	0.680 ± 2.8E-3	**0.789 ± 3.1E-3**
DEEPLABV3+ [16]	22.4M	0.616 ± 4.1E-2	0.553 ± 9.8E-2	0.644 ± 2.8E-3	0.719 ± 5.4E-3	0.655 ± 2.4E-3	**0.761 ± 2.0E-3**
N-DECONET	**21.5M**	0.688 ± 2.6E-2	0.656 ± 3.5E-2	0.630 ± 2.7E-3	0.731 ± 4.4E-3	0.655 ± 3.9E-3	**0.753 ± 1.3E-2**
U-DECONET	**21.5M**	0.530 ± 2.1E-2	0.549 ± 2.8E-2	0.615 ± 2.2E-3	0.703 ± 3.5E-3	0.621 ± 3.1E-3	**0.714 ± 1.0E-2**

4.1. Dataset

The MRI ALS dataset consists of T_1-weighted scans obtained on a 3T Verio Siemens MR scanner (CENIR, Brain Institute, Paris, France) from 26 patients with ALS (in the early stage of the disease) and 26 sex- and age-matched healthy controls (60.8 ± 11.0 yrs old).

Images were preprocessed using the SPM12 software[1], including skull stripping, bias field correction, intensity normalization, nonlinear warping to the Montreal Neurological Institute (MNI) standard space, and resampling to 1 mm^3 isotropic resolution. The size of a preprocessed image is $197 \times 233 \times 189$ voxels, of which the area of the brainstem is approximately $42 \times 42 \times 73$ voxels. Each image was then sliced along the axial z-axis and cropped to a size of 50×50 pixels centered on the brainstem area, then resized to a size of 224×224 pixels. In total, 3,796 2D images were created and 1,100 images were used for training, 798 for validation, and 1,898 images were used as test data.

We created a new atlas in MNI space for the ground truth by combining widely used atlases [27, 28] and masks from Nilearn project[2]. The atlas contains a total of $K = 34$ labels, including extra-brainstem (denoted as background) and brainstem regions: white matter, gray matter other than the nuclei classes (denoted as gray matter class), and 31 nuclei classes. The ground truth was generated by registering the atlas to the given images and transforming it to 2D images.

4.2. Training and Evaluation

To compare the quantitative performance of different models, we used macro mean Intersection over Union (mIoU) [29] for 34 labels, taking into account differences in the area occupied by each class. The training and evaluation process in the experiments was repeated 5 times, and Adam optimizer [30] was used. Each model was trained for 10 epochs with a batch size of 16, and a learning rate of 0.0001 was applied. It was confirmed that all models sufficiently converged under the same conditions.

[1]SPM12: https://www.fil.ion.ucl.ac.uk/spm/software/spm12/

[2]Nilearn: https://nilearn.github.io/dev/index.html

4.3. Results

N-DecoNet, U-DecoNet, and selected models were compared from three perspectives: structures of decoder networks, loss functions, and transfer learning of encoder networks. Table 1 presents the model performance with different loss functions using the same encoder network (ResNet-34) and through this, we can see how the configuration of each model affects solving the given problem. N-DecoNet demonstrated competitive results, despite its simple structure and the lowest parameter count among the models assessed. Both proposed networks performed more effectively compared to UNet-type models. Pyramid-type models were both lighter and more effective. The best results of the models are shown in Figure 3.

4.3.1. Impact of the Decoder Network

Through the results, we propose new considerations when designing a decoder network, due to the complexity of the dataset. First, in the process of upsampling, duplicating the previous layer like in UNet-type networks can reduce performance. Because the MRI used is less complex than the natural image dataset, using the previous layer together during upsampling resulted in the loss of low-level features. This can be confirmed by the fact that the Pyramid-type models and our model outperformed UNet-type models and that the performance deteriorated just by adding a combination step on the decoder (U-DecoNet).

Second, in the process of recovering the encoder's feature maps, combining feature maps of different sizes led to worse results. Likewise, with data of relatively low complexity, it may be difficult for the model to focus only on the details and learn the overall shape. Using the same size, U-DecoNet performed better than the UNet-type model. When using a loss function that specifically helps a model to learn the details such as Dice loss and Jaccard loss, the UNet-type model performed worse.

Lastly, changing the skip connections between the encoder network and the decoder network did not help improve model performance. Because the features that can be trained from the given data were limited, complex skip connections do not allow the model to discover more features.

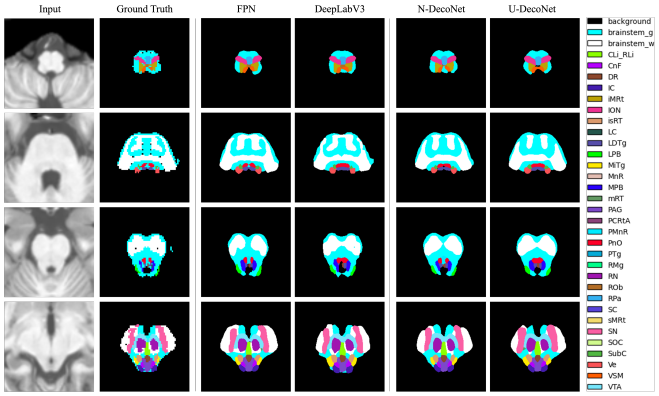

Fig. 3. Best results of each decoder. The FPN result used ResNet-50 as the encoder, and DeepLabV3 used SENet-101 as its encoder. N-DecoNet and U-DecoNet used ResNet-152 and ResNet-101 as encoders. All networks were trained with DiceCE loss ($\alpha = 0.1$). Each color represents a label class.

4.3.2. *Importance of the Loss Function*

As shown in Table 1, the results of each model differed depending on the loss function. All models performed best when trained in a way that minimizes the DiceCE loss function. In particular, N-DecoNet recorded 0.904 mIoU in the nuclei classes. However, the performance was greatly affected by the combining ratio (α) of the DiceCE loss (Figure 4).

As the proportion of SoftCE increased, performance deteriorated. Indeed, as the ratio of SoftCE grows, more emphasis is placed on training the overall shape rather than the details of the image. This is because the mIoU for nuclei classes drops more steeply than the mIoU for white matter and gray matter classes, which contain relatively more pixels. Interestingly, DeepLabV3 was not significantly affected and maintained similar performance regardless of the ratio being varied. The standard deviation of DeepLabV3 results was 0.006, while it was 0.040, 0.026, and 0.034 for FPN, N-DecoNet, and U-DecoNet results, respectively. Therefore, in DeepLabV3, we may find clues to create more efficient decoders, such as using atrous convolutions.

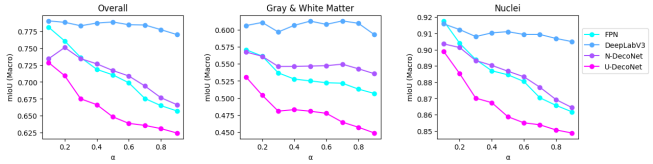

Fig. 4. Macro mIoU results based on different ratios of SoftCE in DiceCE loss. All networks used ResNet-34 as the encoder. (left): overall mIoU results; (middle): mean mIoU for gray and white matter classes; (right): mIoU for nuclei classes. α is the ratio factor in equations (1) and (2).

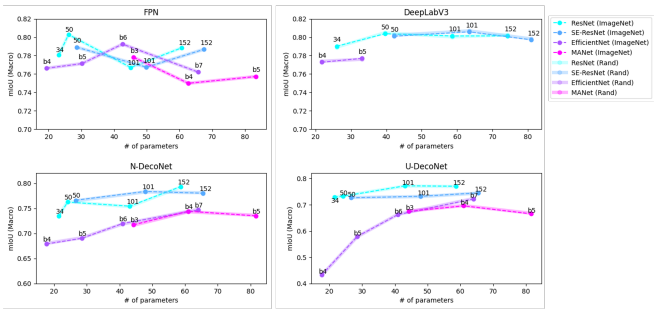

Fig. 5. Macro mIoU results of selected decoders with different encoder architectures for transfer learning and random initialization. All models were trained with DiceCE loss ($\alpha = 0.1$).

4.3.3. *Influence of the Encoder Network*

Figure 5 shows the results for each encoder and for each decoder. All models were trained with the DiceCE loss function with $\alpha = 0.1$. DeepLabV3, unlike other networks, has only results for three encoders. This is because some encoders could not be tested due to computational issues.

The results show that using deeper and more complex encoders did not improve the results. Additionally, there is no significant difference between the results of transfer learning with an encoder pre-trained on ImageNet and the results of training with random initialization. After checking the activation maps of each layer, we found that this was because the complexity of the given data set was lower than ImageNet. Perhaps because the encoder has already learned enough features, it could be more valuable to devise an appropriate decoder and loss function to restore the feature map to the size of the original image with less dilution process rather than applying state-of-the-art classification networks as encoders.

5. CONCLUSION

In this paper, we proposed a deep neural network to segment brainstem nuclei from *in vivo* structural MR images. By making a decoder with fewer dilution steps, we were able to design an efficient model that required less number of parameters and showed competitive results compared with state-of-the-art networks. We also investigated central questions on designing deep learning models for MRI segmentation of the brainstem. Evaluating different architectures with diverse loss functions, we found that constructing an appropriate decoder network and loss function improved the performance, even though transfer learning on an encoder network offered limited performance gains. In the future, this research can be used to implement a model with improved performance, which can expand the direction of research on diseases such as ALS, and the direction of research into multi-modal medical AI by improving the trustworthiness and interpretability of the model.

6. ACKNOWLEDGEMENTS

This study was funded by Sorbonne University, Inserm, and CNRS. S.O. was supported by a Ph.D. grant from the French Ministry of Higher Education and Research. The authors have no conflict of interest to disclose.

7. COMPLIANCE WITH ETHICAL STANDARDS

This study was performed in line with the Declaration of Helsinki. Approval was granted by the French Ethics Committee (IRCB 2018-A00789-52; Clinical Trials n° NCT 03694132).

8. REFERENCES

[1] R. Sclocco et al., "Challenges and opportunities for brainstem neuroimaging with ultrahigh field MRI," *NeuroImage*, vol. 168, pp. 412–26, 2018.

[2] G. Querin et al., "Multimodal spinal cord MRI offers accurate diagnostic classification in ALS," *J Neurol Neurosurg Psychiatry*, vol. 89, pp. 1220–1, 2018.

[3] M. Khamaysa et al., "Quantitative brainstem and spinal MRI in amyotrophic lateral sclerosis: implications for predicting noninvasive ventilation needs," *J Neurol*, vol. doi: 10.1007/s00415-023-12045-x, 2023, Epub ahead of print.

[4] R. J. Rushmore et al., "3D exploration of the brainstem in 50-micron resolution MRI," *Front Neuroanat*, vol. 14, pp. 40, 2020.

[5] V. Badrinarayanan et al., "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *CoRR*, 2015.

[6] K. He et al., "Deep residual learning for image recognition," in *2016 IEEE CVPR*, 2016, pp. 770–8.

[7] J. Hu et al., "Squeeze-and-excitation networks," in *2018 IEEE CVPR*, 2018, pp. 7132–41.

[8] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th ICML*, 2019, pp. 6105–14.

[9] E. Xie et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *35th Conference on NeurIPS*, 2021.

[10] O. Ronneberger et al., "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–41.

[11] Z. Zhou et al., "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans Med Imaging*, vol. 39, pp. 1856–67, 2020.

[12] R. Li et al., "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Trans Geosci Remote Sens*, vol. 60, pp. 1–13, 2022.

[13] H. Zhao et al., "Pyramid scene parsing network," in *2017 IEEE CVPR*, 2017, pp. 6230–9.

[14] T.-Y. Lin et al., "Feature pyramid networks for object detection," in *2017 IEEE CVPR*, 2017, pp. 936–44.

[15] L. C. Chen et al., "Rethinking atrous convolution for semantic image segmentation," *ArXiv*, vol. abs/1706.05587, 2017.

[16] L. C. Chen et al., "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Computer Vision – ECCV 2018*, 2018, pp. 833–51.

[17] A. V. Dalca et al., "Unsupervised deep learning for Bayesian brain MRI segmentation," in *MICCAI*, 2019, pp. 356–65.

[18] Ö Çiçek et al., "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *MICCAI*, 2016, pp. 424–32.

[19] Y. Tang et al., "Self-supervised pre-training of Swin Transformers for 3D medical image analysis," in *2022 IEEE CVPR*, 2022, pp. 20698–708.

[20] M. Raghu et al., "Transfusion: Understanding transfer learning for medical imaging," in *33rd Conference on NeurIPS*, 2019.

[21] J. Deng et al., "ImageNet: A large-scale hierarchical image database," in *2009 IEEE CVPR*, 2009, pp. 248–55.

[22] J. Bridle, "Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters," in *2nd Conference on NeurIPS*, 1989, vol. 2, pp. 211–217.

[23] C. H. Sudre et al., "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*, 2017, pp. 240–8.

[24] C. Szegedy et al., "Rethinking the inception architecture for computer vision," in *2016 IEEE CVPR*, 2016, pp. 2818–26.

[25] T. Y. Lin et al., "Focal loss for dense object detection," *IEEE Trans Pattern Anal Mach Intell*, vol. 42, pp. 318–27, 2018.

[26] J. Bertels et al., "Optimizing the Dice score and Jaccard index for medical image segmentation: Theory and practice," in *MICCAI*, 2019, pp. 92–100.

[27] J. L. Lancaster et al., "Automated Talairach atlas labels for functional brain mapping," *Hum Brain Mapp*, vol. 10, pp. 120–31, 2000.

[28] K. Singh et al., "Probabilistic template of the lateral parabrachial nucleus, medial parabrachial nucleus, vestibular nuclei complex, and medullary viscerosensory-motor nuclei complex in living humans from 7 Tesla MRI," *Front Neurosci*, vol. 13, pp. 1425, 2020.

[29] M. Everingham et al., "The pascal visual object classes challenge: A retrospective," *IJCV*, vol. 111 (1), pp. 98–136, 2014.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd ICLR*, 2015.