



# Shortcomings of Ventricle Segmentation Using Deep Convolutional Networks

Muhan Shao<sup>1(✉)</sup>, Shuo Han<sup>2,3</sup>, Aaron Carass<sup>1,4</sup>, Xiang Li<sup>1</sup>, Ari M. Blitz<sup>5</sup>,  
Jerry L. Prince<sup>1,4</sup>, and Lotta M. Ellingsen<sup>1,6</sup>

<sup>1</sup> Department of Electrical and Computer Engineering,  
The Johns Hopkins University, Baltimore, MD 21218, USA  
muhan@jhu.edu

<sup>2</sup> Department of Biomedical Engineering,  
The Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

<sup>3</sup> Laboratory of Behavioral Neuroscience, National Institute on Aging,  
National Institutes of Health, Baltimore, MD 20892, USA

<sup>4</sup> Department of Computer Science, The Johns Hopkins University,  
Baltimore, MD 21218, USA

<sup>5</sup> Department of Radiology and Radiological Science, The Johns Hopkins University,  
Baltimore, MD 21287, USA

<sup>6</sup> Department of Electrical and Computer Engineering, University of Iceland,  
Reykjavik, Iceland

**Abstract.** Normal Pressure Hydrocephalus (NPH) is a brain disorder that can present with ventriculomegaly and dementia-like symptoms, which often can be reversed through surgery. Having accurate segmentation of the ventricular system into its sub-compartments from magnetic resonance images (MRI) would be beneficial to better characterize the condition of NPH patients. Previous segmentation algorithms need long processing time and often fail to accurately segment severely enlarged ventricles in NPH patients. Recently, deep convolutional neural network (CNN) methods have been reported to have fast and accurate performance on medical image segmentation tasks. In this paper, we present a 3D U-net CNN-based network to segment the ventricular system in MRI. We trained three networks on different data sets and compared their performances. The networks trained on healthy controls (HC) failed in patients with NPH pathology, even in patients with normal appearing ventricles. The network trained on images from HC and NPH patients provided superior performance against state-of-the-art methods when evaluated on images from both data sets.

**Keywords:** MRI · Hydrocephalus · Segmentation · CNN

## 1 Introduction

The ventricular system of the human brain is composed of four interconnected cavities: the left and right lateral, the third and the fourth ventricles. Each ventricle contains choroid plexus, a network of ependymal cells producing cerebrospinal

fluid (CSF). Normal pressure hydrocephalus (NPH) is a brain disorder usually caused by disruption of CSF flow but with normal CSF pressure. The ventricles expand and press against the brain tissue nearby, which can lead to the distortion of the brain shape and eventually cause brain damage. NPH is characterized by gait unsteadiness, urinary incontinence, and dementia [1]. However, unlike most forms of dementia, the symptoms in NPH are potentially reversible to a certain extent on properly selected patients. Diversion of CSF through shunt surgery has been reported to improve the symptoms of NPH [10]. However, it remains a challenge to identify NPH patients who respond to treatment, and differentiate NPH from other neurodegenerative disorders, such as Alzheimer’s disease [11].

Currently, NPH is diagnosed based on characteristic clinical symptoms and brain imaging [11]. The ventricular dilation in NPH can be observed through magnetic resonance (MR) images. Examples of T1-weighted (T1w) Magnetically Prepared Rapidly Acquired Gradient Echo (MPRAGE) images of NPH patients are shown in Fig. 3(a). Disproportionate dilation of components of the ventricular system in NPH is relative to the specific point of CSF disruption, which could have an impact on the diagnosis [11]. Therefore, accurate segmentation of the ventricular system into its four cavities could help characterize the pathophysiology and potentially lead to better surgical planning of NPH patients.

Previously published segmentation methods include the popular FreeSurfer [6] method and many multi-atlas segmentation methods [15, 20]. However, these methods require long processing times (several hours) and often fail to capture the boundary of the greatly enlarged ventricles in NPH patients. A recently developed segmentation algorithm, RUDOLPH [3, 5], is a combined patch-based and multi-atlas segmentation method designed for subjects with ventriculomegaly. Although this method is robust in ventricular parcellation, it also has a long runtime. In recent years, various methods based on deep convolutional neural networks (CNN) have been proposed to tackle neuroimage segmentation [2, 12]. The U-Net [16] is one of the most well-known CNN architectures in medical image analysis. The skip connections between contracting and expanding paths in the U-Net improve the network performance.

In this paper, we present a 3D U-Net method for segmenting the ventricular system. We trained three networks on images from two data sets, two comprising healthy controls (HC) and the other a mix of HC and NPH patients, and show the difference of their performances. The first network was trained on 13 HC and performed well when evaluated on subjects from the same data set. However, it performed poorly on the NPH data set, even on images with normal sized ventricles. The second network was trained on 38 HC, including elderly subjects with enlarged ventricles, and performed even worse than the first network when evaluated on NPH data set. The third network was trained on a mixture of 13 HC and 25 NPH images and provided dramatically improved results on both data sets, demonstrating the importance of training data selection.

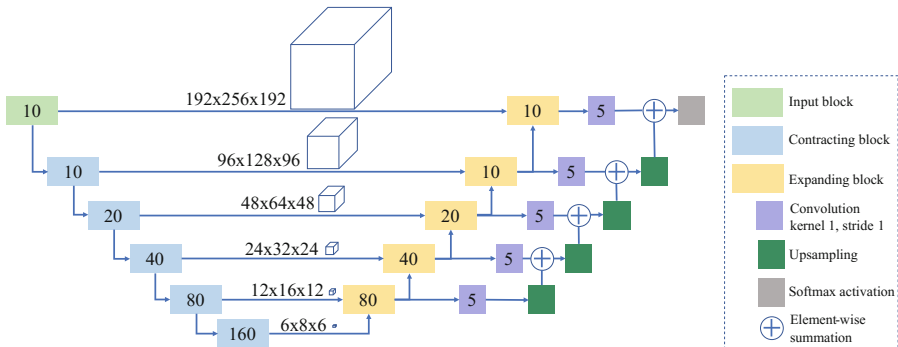
## 2 Methods

### 2.1 Data and Preprocessing

We evaluated our segmentation network using 3D brain MR images from two data sets. The first one comprised 38 T1w MR images from Neuromorphometrics Inc (NMM)<sup>1</sup>. Each image was manually delineated by experts into 138 brain structures. For our purposes, we converted the 139 labels (138 brain structure labels and 1 background label) into five: left and right lateral ventricles, third ventricle, fourth ventricle, and a catch-all background label. The inferior lateral ventricle label was included with the corresponding lateral ventricle label. The T1w MR images were sorted by the volume of the ventricular system and 13 images were used as training data for the first and third network, covering the entire spectrum of ventricle sizes in the data set. All 38 images were used as training data for the second network.

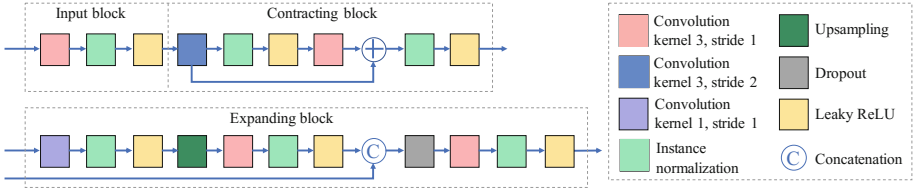
The second data set was from our NPH database comprising 95 NPH patients with a wide range of ventriculomegaly. They were acquired on a 3T (Siemens Corporation, Germany) scanner with T1w MPRAGE with TR = 10.3 ms, TE = 6 ms, and  $0.82 \times 0.82 \times 1.17 \text{ mm}^3$  voxel size. We manually delineated the ventricular system in all 95 NPH patients from our database into our five labels. A total of 25 NPH images, ranging from mild to severe cases, were chosen as our training data for the second network.

The images from the two data sets were run through a preprocessing pipeline, including N4 bias correction [18], rigid registration to MNI 152 atlas space [7], and skull stripping [17].



**Fig. 1.** Architecture of the ventricle-segmentation network. The numbers in the contracting and expanding blocks indicate the output number of features. The shape of the tensor is denoted next to the box in each resolution scale.

<sup>1</sup> <http://www.neuromorphometrics.com/>.



**Fig. 2.** Architecture of the input, contracting, and expanding blocks used in the segmentation network.

## 2.2 Ventricle Segmentation Network

A 3D U-Net [13] was modified to segment the left and right lateral ventricles, and the third and fourth ventricles. In this network (Fig. 1), a series of contracting blocks extract image features from local to global context and a series of expanding blocks, with shortcut to contracting blocks, act as “learnable” upsampling interpolation to restore the feature map resolution (Fig. 2). Using learned features, the projection convolution connected to each expanding block (Fig. 1) along with the softmax operation further classify the voxels into five labels including the four ventricles and the background.

The contracting block is similar to the building block for increasing dimensions of the pre-activation ResNet [9], since the shortcut within a block can make the optimization easier and increase accuracy [8]. In contrast to ResNet, however, the identity mapping and the residue encoding paths share the first convolution in this design to reduce overfitting. Instance normalization [19] was used since it is invariant to mean and covariant shift of image intensities. The negative slope of Leaky ReLUs [22] was 0.1 and the dropout rate was 0.2.

## 2.3 Training Procedure

We used data augmentation by applying right-left flipping, elastic deformation, and rotation to the training images. The images were cropped to  $192 \times 256 \times 192$  and sent to the input block. The loss function was one minus the mean Dice coefficient [4] of each label. The network was trained for 50 epochs using the Adam optimizer [14].

## 3 Experiments and Results

We trained three networks, VenSeg1 using 13 T1w MR images from NMM, VenSeg2 using 38 T1w MR images from NMM, and VenSeg3 using 38 T1w MR images including the same 13 in VenSeg1 and 25 from our NPH cohort. The 95 MR images (25 from NMM and 70 from NPH) formed the testing data set. We only evaluated the performance of VenSeg2 on the 70 NPH testing images.

The 25 testing images from NMM data set were processed by VenSeg1, VenSeg3, and three state-of-the-art brain segmentation methods: FreeSurfer 6.0,

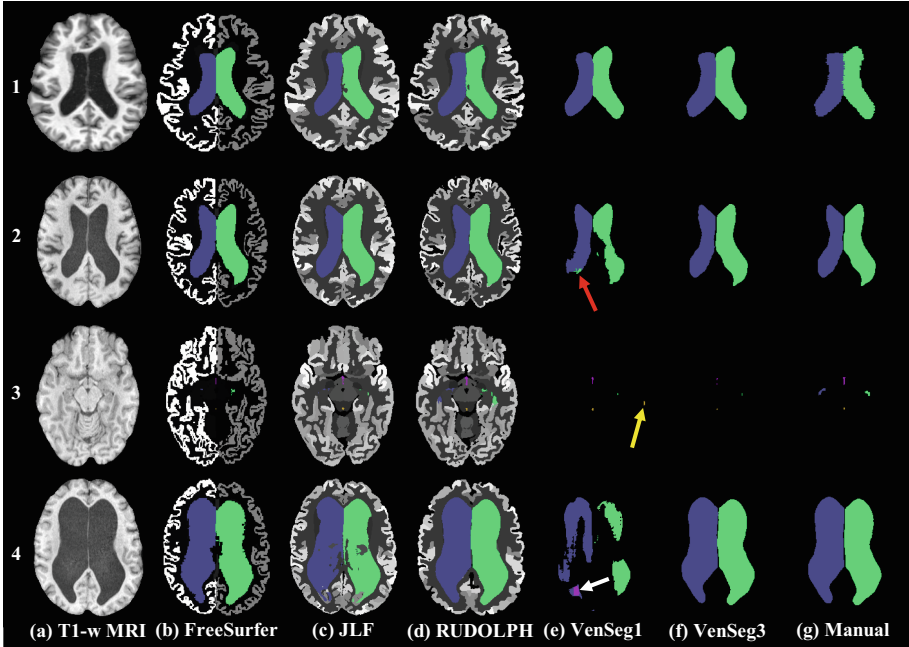
Joint label fusion (JLF) [20] and RUDOLPH [3]. The 70 testing images from the NPH cohort were processed by all the six segmentation methods. We provided FreeSurfer with skull-stripped data to speed up the process and turned on the `-bigventricles` switch for NPH subjects to handle the enlarged ventricles.

Visual comparisons of the five methods (excluding VenSeg2) on one NMM image and three NPH images are shown in Fig. 3. The VenSeg1 network provided accurate segmentation on the NMM image (Fig. 3(a), subject #1). However, it yielded erroneous segmentations on MR images of NPH patients. A truly surprising failure of VenSeg1 is subject #2; Subject #2 has a similar shape and volume to subject #1 from the NMM cohort (129 ml for subject #2 and 132 ml for subject #1) and yet VenSeg1 failed to capture the boundary of the lateral ventricles and mislabeled portions of the right ventricle as left. Subject #3 in Fig. 3 shows an NPH patient with mild pathology, however VenSeg1 incorrectly labeled some cortex as the 4th ventricle (yellow arrow in Fig. 3(e), subject #3).

We computed the Dice coefficient on a cohort of subjects only from NMM and a cohort of subjects only from NPH for the methods and report the results in Tables 1 and 2, respectively. We note that VenSeg2 performed worse than VenSeg1 on NPH data set despite having more training data (see Table 2). We used a paired Wilcoxon signed-rank test [21] to compare the methods. For the results on the NMM testing images, we found no significant differences between VenSeg1 and VenSeg3 in terms of Dice coefficients. Both networks performed significantly better ( $p < 0.001$ ) than FreeSurfer and RUDOLPH on the lateral ventricles and the 3rd ventricle, and better than FreeSurfer on the 4th ventricle. For the results on the NPH image testing set, VenSeg3 performed significantly better ( $p < 0.001$ ) than all the other methods on all the ventricle labels.

**Table 1.** The mean Dice coefficient (and standard deviation) over 25 T1w images from Neuromorphometrics. Ventricular system key: Merged four ventricle labels (Whole), right lateral ventricle (RLV), left lateral ventricle (LLV), third ventricle (3rd), and fourth ventricle (4th). The asterisks mean significantly different ( $p$ -value  $< 0.001$ ) to VenSeg1 and VenSeg3.

	Whole	RLV	LLV	3rd	4th
FreeSurfer	0.843*( $\pm 0.04$ )	0.848*( $\pm 0.04$ )	0.848*( $\pm 0.04$ )	0.700*( $\pm 0.12$ )	0.760*( $\pm 0.04$ )
JLF	0.881( $\pm 0.03$ )	0.879( $\pm 0.03$ )	0.888( $\pm 0.03$ )	0.796( $\pm 0.04$ )	0.844( $\pm 0.03$ )
RUDOLPH	0.883*( $\pm 0.03$ )	0.883*( $\pm 0.03$ )	0.888*( $\pm 0.03$ )	0.777*( $\pm 0.08$ )	0.839( $\pm 0.04$ )
VenSeg1	0.902( $\pm 0.03$ )	0.903( $\pm 0.03$ )	0.907( $\pm 0.03$ )	0.821( $\pm 0.07$ )	0.844( $\pm 0.04$ )
VenSeg3	0.902( $\pm 0.03$ )	0.904( $\pm 0.03$ )	0.907( $\pm 0.03$ )	0.817( $\pm 0.07$ )	0.842( $\pm 0.04$ )



**Fig. 3.** Segmentation results from three state-of-the-art methods (FreeSurfer, JLF, and RUDOLPH) and two proposed deep networks (VenSeg1 and VenSeg3) compared with a manual rater (column g). Subject #1: T1w image and segmentation results from NMM data set. Subjects #2–4: T1w images and segmentation results from NPH data set, showing moderate, mild, and severe cases. The red arrow in (e2) shows the right lateral ventricle inaccurately labeled as the left lateral ventricle. The yellow arrow in (e3) points to cortex mislabeled as the 4th ventricle. The white arrow in (e4) points to the right ventricle mislabeled as the 3rd ventricle. (Color figure online)

**Table 2.** The mean Dice coefficient (and standard deviation) over the 70 testing images from the NPH data set. Bold: VenSeg3 is significantly better ( $p$ -value  $< 0.001$ ) than the other five methods on all the labels.

	Whole	RLV	LLV	3rd	4th
FreeSurfer	0.937( $\pm 0.03$ )	0.942( $\pm 0.03$ )	0.939( $\pm 0.03$ )	0.840( $\pm 0.06$ )	0.730( $\pm 0.08$ )
JLF	0.930( $\pm 0.04$ )	0.931( $\pm 0.05$ )	0.933( $\pm 0.04$ )	0.865( $\pm 0.06$ )	0.862( $\pm 0.04$ )
RUDOLPH	0.942( $\pm 0.05$ )	0.943( $\pm 0.05$ )	0.944( $\pm 0.05$ )	0.875( $\pm 0.07$ )	0.838( $\pm 0.06$ )
VenSeg1	0.833( $\pm 0.14$ )	0.839( $\pm 0.15$ )	0.832( $\pm 0.15$ )	0.727( $\pm 0.21$ )	0.787( $\pm 0.11$ )
VenSeg2	0.482( $\pm 0.24$ )	0.484( $\pm 0.25$ )	0.480( $\pm 0.25$ )	0.275( $\pm 0.28$ )	0.684( $\pm 0.18$ )
VenSeg3	<b>0.971</b> ( $\pm 0.02$ )	<b>0.971</b> ( $\pm 0.02$ )	<b>0.974</b> ( $\pm 0.02$ )	<b>0.915</b> ( $\pm 0.06$ )	<b>0.903</b> ( $\pm 0.04$ )

## 4 Discussion and Conclusions

We present a 3D U-Net architecture to segment and label the ventricular system in patients with enlarged ventricles. We trained three models on two different data sets using manual delineations as training data. The models were evaluated on 25 NMM subjects and 70 NPH patients and compared to FreeSurfer, JLF, and RUDOLPH.

The model trained on 13 NMM data showed improvement over the state-of-the-art segmentation methods in terms of overlap with expert delineations on the same data set. However, it showed poor performance on the NPH data set, even on images with ventricle size similar to the training data. The segmentation results from this model on subjects #1 and #2 were inconsistent. The model failed to identify the boundary of the lateral ventricles and mislabeled portions of the right ventricle as left on subject #2 (see the red arrow in Fig. 3(e2)). This failure occurred despite the fact that the size of the ventricles in subject #2 is very similar to the ventricle size of subject #1 from NMM. In some cases with small ventricular volume, the model mislabeled the cortex as ventricle (see the yellow arrow in Fig. 3(e3)). In severe cases of NPH, this model cannot handle the pathology as its training data set does not include similar examples; Furthermore it labeled posterior portions of the right ventricle as the 3rd ventricle (see the white arrow in Fig. 3(e4)).

The second network was trained on 38 NMM images, including elderly subjects with enlarged ventricles, since more training data could potentially improve the performance. However, this network provided worse segmentation results than the first one when evaluated on NPH patients. One possible explanation is that adding more training data made the network overfitted on the NMM data set.

The failure of these two networks on NPH patients indicates that the network did not learn only the intensity and spatial information from the training data, since the first network successfully segmented a subject from NMM but failed on a subject with similar ventricle size from the NPH data set. The dominant features learned by the network—that are driving the segmentation—remain a mystery.

The third network was trained on 38 images from both data sets. It performed significantly better than all of the other methods on the entire testing data set, demonstrating both the robustness of the network to high variations of ventricle sizes, but also the importance of careful training data selection for deep learning methods.

## References

1. Adams, R., Fisher, C., Hakim, S., Ojemann, R., Sweet, W.: Symptomatic occult hydrocephalus with normal cerebrospinal-fluid pressure: a treatable syndrome. *N. Engl. J. Med.* **273**(3), 117–126 (1965)
2. de Brebisson, A., Montana, G.: Deep neural networks for anatomical brain segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 20–28 (2015)

3. Carass, A., et al.: Whole brain parcellation with pathology: validation on ventriculomegaly patients. In: Wu, G. (ed.) *Patch-MI 2017*. LNCS, vol. 10530, pp. 20–28. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-67434-6\\_3](https://doi.org/10.1007/978-3-319-67434-6_3)
4. Dice, L.R.: Measures of the amount of ecologic association between species. *Ecology* **26**(3), 297–302 (1945)
5. Ellingsen, L.M., Roy, S., Carass, A., Blitz, A.M., Pham, D.L., Prince, J.L.: Segmentation and labeling of the ventricular system in normal pressure hydrocephalus using patch-based tissue classification and multi-atlas labeling. In: *Proceedings of SPIE—the International Society for Optical Engineering*, vol. 9784 (2016)
6. Fischl, B.: Freesurfer. *NeuroImage* **62**(2), 774–781 (2012)
7. Fonov, V.S., Evans, A.C., McKinstry, R.C., Almlri, C., Collins, D.: Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage* **47**, S102 (2009)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint [arXiv:1512.03385](https://arxiv.org/abs/1512.03385) (2015)
9. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. arXiv preprint [arXiv:1603.05027](https://arxiv.org/abs/1603.05027) (2016)
10. Hebb, A.O., Cusimano, M.D.: Idiopathic normal pressure hydrocephalus: a systematic review of diagnosis and outcome. *Neurosurgery* **49**(5), 1166–1186 (2001)
11. Ishikawa, M., et al.: Guidelines for management of idiopathic normal pressure hydrocephalus. *Neurol. Med.-Chir.* **48**(Suppl.), S1–S23 (2008)
12. Kamnitsas, K., et al.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* **36**, 61–78 (2017)
13. Kayalibay, B., Jensen, G., van der Smagt, P.: CNN-based segmentation of medical imaging data. arXiv preprint [arXiv:1701.03056](https://arxiv.org/abs/1701.03056) (2017)
14. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
15. Ledig, C., et al.: Robust whole-brain segmentation: application to traumatic brain injury. *Med. Image Anal.* **21**(1), 40–58 (2015)
16. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) *MICCAI 2015*. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-24574-4\\_28](https://doi.org/10.1007/978-3-319-24574-4_28)
17. Roy, S., Butman, J.A., Pham, D.L.: Alzheimers disease neuroimaging initiative, others: robust skull stripping using multiple MR image contrasts insensitive to pathology. *NeuroImage* **146**, 132–147 (2017)
18. Tustison, N.J., et al.: N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imag.* **29**(6), 1310–1320 (2010)
19. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: the missing ingredient for fast stylization. arXiv preprint [arXiv:1607.08022](https://arxiv.org/abs/1607.08022) (2016)
20. Wang, H., Suh, J.W., Das, S.R., Pluta, J.B., Craige, C., Yushkevich, P.A.: Multi-atlas segmentation with joint label fusion. *IEEE Trans. Patt. Anal. Mach. Intell.* **35**(3), 611–623 (2013)
21. Wilcoxon, F.: Individual comparisons by ranking methods. *Biom. Bull.* **1**(6), 80–83 (1945)
22. Xu, B., Wang, N., Chen, T., Li, M.: Empirical evaluation of rectified activations in convolutional network. arXiv preprint [arXiv:1505.00853](https://arxiv.org/abs/1505.00853) (2015)