Inter-site variability in prostate segmentation accuracy using deep learning

Anonymous Author[1], Anonymous Author[2], Anonymous Author[1], Anonymous Author[2], and Anonymous Author[1]

[1] Anonymous institution
[2] Anonymous institution

Abstract. Deep-learning-based segmentation tools are yielding higher reported segmentation accuracies for many medical image segmentation problems. However, inter-site variability in medical image acquisition protocols and quality can challenge the translation of these tools to data from unseen sites. This study quantifies the impact of inter-site variability on the accuracy of deep-learning-based segmentation of the prostate from magnetic resonance (MR) images, and evaluates two strategies for mitigating the performance discrepancies for data from unseen sites: training on multi-site datasets and retraining with data from the unseen site. Using 424 T2-weighted prostate MR images from six sites, we compare the segmentation accuracy of three deep-learning-based networks trained on data from a single site and on various configurations of data from multiple sites. We found that the segmentation accuracy of a single-site network was substantially worse on data from unseen sites than on data from the training site, and that training on multi-site data gives only marginal improvement. However, including as few as 8 subjects from the unseen site, e.g. during commissioning each new clinical system, yields substantial improvement (regaining 75% of the difference).
Keywords: segmentation, neural network, deep learning, inter-site variability, prostate

1 Introduction

With the development of deep-learning-based segmentation methods for medical images, reported segmentation accuracies have improved substantially for many segmentation problems including prostate [7], brain tumors [1] and abdominal organs [6]. Applying these methods in practice, however, remains challenging, with few segmentation methods achieving previously reported accuracies on new data sets. This may be due, in part, to inter-site variability in medical image acquisition equipment, protocols and quality.

Inter-site variability has remained a challenge in medical image analysis for decades [12, 9]. Data sets used to design, train and validate segmentation algorithms are, for logistical and financial reasons, sampled in clusters from one or a small number of hospitals, rather than independently sampled from the population of all images. The distribution of images in collected data sets may not be representative of, and may have less variability than, the whole population. Furthermore, an algorithm optimized for data from one site may not be optimal for data from another site, and reported estimates of segmentation accuracy may overestimate the expected accuracy.

Deep-learning-based methods may be more susceptible to this problem than previous approaches. Segmentation methods with hand-crafted image and algorithmic features, such as bias field correction [12], can explicitly encode high-level priors that reflect knowledge of sources of inter-site variability. Deep-learning-based methods, in contrast, typically aim to use weaker priors and rely on learning corresponding patterns from the data. Accordingly, the performance of deep-learning-based methods may depend more heavily on having a large training data set that is representative of the images to which the method will be applied.

In this study, we aimed to quantify the impact of inter-site variability on the accuracy of deep-learning-based segmentation of the prostate from T2-weighted MRI of three deep-learning-based methods, and evaluated two strategies to mitigate the performance loss at an unseen site: training on multi-site data sets, and retraining with some data from the unseen site. To identify general trends, we conducted these experiments using three different deep-learning based methods. Specifically, this study addresses the following questions:

1) How accurate are prostate segmentations using networks trained on data from a single site when evaluated on data from the same site and from unseen sites? 2) How accurate are prostate segmentations using networks trained on data from multiple sites when evaluated on data from the same sites and from unseen sites? 3) Can the accuracy of these prostate segmentations be improved by including a small sample of data from the unseen site?

2 Methods

2.1 Imaging

This study used T2-weighted 3D prostate MRI from 6 sites (256 from SITE1[3], 48 from SITE2 and SITE5, and 24 from SITE3, SITE4 and SITE6), drawn from publicly available data sets and clinical trials requiring manual prostate delineation. Reference standard manual segmentations were performed at one of 3 sites: SITE1, SITE2 or SITE5. Images were acquired with anisotropic voxels, with in-plane voxel spacing between 0.5 and 1.0 mm, and out-of-plane slice spacing between 1.8 and 5.4 mm. All images and reference standard segmentations were resampled to 256 by 256 by 32 before automatic segmentation.

2.2 Experimental design

These data were randomized within each center and then used in various combinations to measure the performance under different choices of training data.

[3] Anonymized for review

Each experiment was conducted for three different neural networks architectures, described in Section 2.3. Segmentation performance was measured using the Dice coefficient and the symmetric boundary distance (BD).

Experiment 1: single-site networks To evaluate the within- and inter-site segmentation performance of networks trained on data from one site, we trained the network on 232 subjects from SITE1, and evaluated the data on the remaining 24 subjects from SITE1 and all subjects from each of the other sites.

Experiment 2: multi-site networks To evaluate the segmentation performance of networks trained on data from multiple sites, we used two experimental setups. First, we conducted a patient-level 6-fold cross-validation where, in each fold, 16 subjects from each site were used for training, and 8 subjects from each site were used for testing. This patient-level setup has been used in online challenges, such as the PROMISE12 segmentation challenge [7]. Because this may overestimate the performance at a site that has not been seen in training, we conducted a second site-level 6-fold cross-validation where, in each fold, 24 subjects from 5 sites were used for training, and all subject from the remaining site were used for testing.

Experiment 3: mitigating inter-site variation In practice, when introducing new imaging technology, hospitals typically undergo a commissioning process to calibrate and validate the technology. In principle, such a process could include re-training or fine-tuning a neural network on data from that site. To evaluate the utility of this approach, we conducted a hierarchical cross-validation where in each fold the network was trained using 24 subjects from 5 sites and a subset of subjects from the remaining site (3 subsets with 8 subjects and 3 subsets with 16 subjects).

2.3 Neural networks

To distinguish general trends from network-specific properties, three different neural network architectures, illustrated in Fig. 1 were used in this study: DenseVNet [4], ResUNet [3], and VoxResNet [2]. Like most medical image segmentation networks, these networks are all variants of U/V-Net architectures[11, 8] comprising a downsampling subnetwork, an upsampling subnetwork and skip connections. ResUNet segments 2D axial slices using a 5-resolution U-Net with residual units, max-pooling, and additive skip connections. DenseVNet segments 3D volumes using a 4-resolution V-Net with memory-efficient batch-wise spatial dropout and dense feature stacks to preserve high-resolution 3D information through the network; skip connections comprise a single convolution and are concatenated before a final segmentation convolution. VoxResNet segments 3D volumes using a 4-resolution V-Net with residual units [5], transpose-convolution upsampling, and deep supervision to improve gradient propagation. It is important to note that this study is not designed to compare the absolute performance

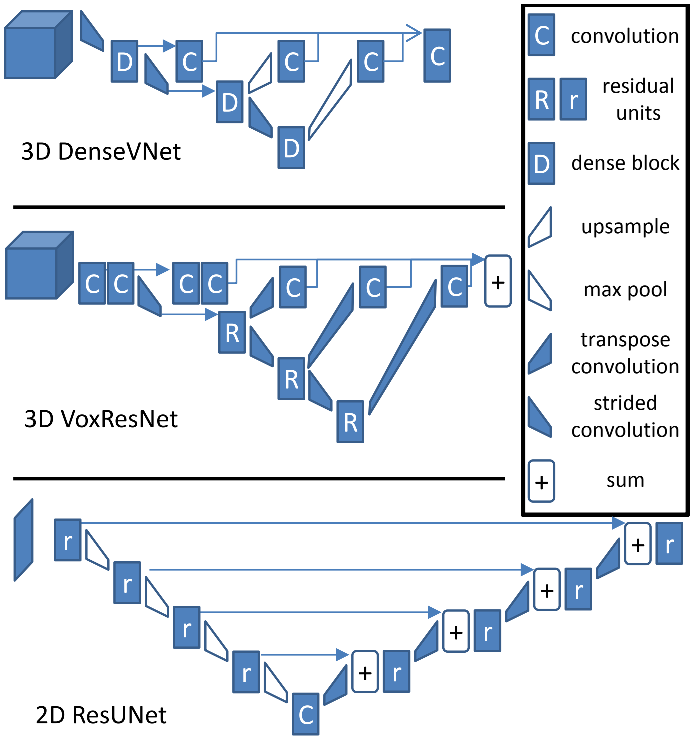

Fig. 1. Architectures of the neural networks.

of these networks; accordingly, the network dimensionality and features, hyperparameter choices, and training regimen were not made equivalent.

2.4 Network training and inference

For each fold of each experiment, the network was trained by minimizing the Dice loss using the Adam optimizer for 10000 iterations. The training data set was augmented online using affine perturbations. Final segmentations were post-processed to eliminate spurious segmentation by taking the largest connected component.

3 Results

The described experiments entailed the training of 174 networks and nearly 2500 segmentations. The aggregated segmentation accuracies for DenseVNet, VoxResNet and ResUNet are reported in Tables 3, 3, 3, respectively, and summarized below.

Table 1. DenseVNet

	SITE1	SITE2	SITE3	SITE4	SITE5	SITE6	pooled
			Dice coefficient				
Single-site	0.88	0.88	0.84	0.83	0.73	0.86	0.84
Patient-level	0.87	0.90	0.87	0.88	0.86	0.88	0.88
Site-level	0.87	0.88	0.85	0.74	0.78	0.85	0.83
Site-level + 8	0.87	0.89	0.87	0.86	0.86	0.88	0.87
Site-level + 16	0.87	0.89	0.88	0.86	0.86	0.88	0.87
			Boundary distance				
Single-site	1.58	1.69	1.95	2.17	4.95	1.91	2.37
Patient-level	1.74	1.52	1.54	1.49	2.02	1.56	1.64
Site-level	1.76	1.67	1.80	4.19	3.18	1.99	2.43
Site-level + 8	1.80	1.64	1.60	1.69	2.03	1.65	1.73
Site-level + 16	1.82	1.62	1.50	1.68	2.08	1.58	1.71

For networks trained on data from a single site, the mean accuracy on data from other sites was generally lower and varied substantially between sites, with the Dice score dropping by a mean (SD;range) of 0.13 (0.14;0.00–0.45).

For networks trained on multi-site data in a patient-level cross-validation, the mean accuracies (Dice: 0.88, 0.84, 0.84; BD: 1.7 mm, 2.2 mm, 3.0 mm) were nearly identical to those of the single-site networks on test data from the single site (Dice 0.88, 0.85, 0.86; BD: 1.6 mm, 2.1 mm, 2.3 mm), suggesting that it was not inherently more difficult to train the networks on multi-site data. However, for networks trained on multi-site data in a site-level cross-validation, mean accuracies for the unseen site (averaged Dice: 0.83, 0.74, 0.74; BD: 2.4 mm, 4.3 mm, 5.3 mm) were only marginally better than the accuracy of the single-site network for unseen sites (averaged Dice: 0.82, 0.70, 0,70; BD: 3.0 mm, 4.8 mm, 5.3 mm), suggesting training on data from 5 sites does not yield substantially better generalization.

For networks trained on data from five sites with some 'commissioning' data from the sixth site, segmentation accuracies on test data from the sixth site regained most of the difference between the patient- and site-level cross-validations. With only 8 subjects used as commissioning data, segmentation accuracies regained a mean (SD;range) 75% (19;31–96%) of the difference (averaged Dice: 0.87, 0.83, 0.82; BD: 1.8 mm, 2.5 mm, 3.3 mm) when the Dice score discrepancy was > 0.02. With 16 subjects used as commissioning data, segmentation accuracies regained a mean (SD;range) 95% (18;64–¿100%) of the difference (averaged Dice: 0.87, 0.85, 0.83; BD: 1.7 mm, 2.3 mm, 3.5 mm) when the Dice score discrepancy was > 0.02.

4 Discussion

In this work, we demonstrated that multiple deep-learning-based segmentation networks have poor accuracy when applied to data from unseen sites. This chal-

Table 2. VoxResNet

	SITE1	SITE2	SITE3	SITE4	SITE5	SITE6	pooled
			Dice coefficient				
Single-site	0.85	0.83	0.83	0.59	0.40	0.80	0.72
Patient-level	0.84	0.87	0.86	0.84	0.80	0.86	0.84
Site-level	0.83	0.83	0.85	0.66	0.50	0.83	0.75
Site-level + 8	0.85	0.86	0.85	0.83	0.79	0.84	0.84
Site-level + 16	0.85	0.88	0.86	0.85	0.82	0.85	0.85
			Boundary distance				
Single-site	1.95	2.30	2.06	7.92	10.58	2.59	4.57
Patient-level	2.08	1.85	1.71	1.91	2.72	1.86	2.02
Site-level	2.20	2.24	1.83	5.84	6.62	2.26	3.50
Site-level + 8	2.00	1.94	1.78	2.06	2.87	2.02	2.11
Site-level + 16	2.01	1.68	1.70	1.81	2.53	1.88	1.94

Table 3. ResUNet

	SITE1	SITE2	SITE3	SITE4	SITE5	SITE6	pooled
			Dice coefficient				
Single-site	0.87	0.85	0.77	0.47	0.57	0.83	0.73
Patient-level	0.85	0.88	0.87	0.87	0.81	0.84	0.85
Site-level	0.83	0.84	0.83	0.71	0.51	0.80	0.75
Site-level + 8	0.84	0.85	0.86	0.84	0.74	0.82	0.83
Site-level + 16	0.84	0.86	0.85	0.86	0.78	0.85	0.84
			Boundary distance				
Single-site	1.74	1.90	2.39	8.16	7.00	2.15	3.89
Patient-level	1.99	1.69	1.56	1.61	2.39	2.07	1.89
Site-level	2.14	2.02	1.95	3.92	8.38	2.52	3.49
Site-level + 8	2.09	2.05	1.64	1.96	3.73	2.30	2.30
Site-level + 16	2.10	1.79	1.69	1.65	2.83	2.05	2.02

lenges the translation of segmentation tools based on these networks to other research sites and to clinical environments.

In our experiments, and more broadly in medical image analysis, different methods have different capacity to generalize to new sites. Since this is important for the clinical and research impact of these methods, generalization ability should become a metric evaluated by our community. This will require the creation of multi-site datasets, such as PROMISE12 [7] and ADNI [10], to design and evaluate methods. Standardized evaluation protocols, in independent studies and in MICCAI challenges, should include unseen sites in the test set to evaluate generalizability.

For both single- and multi-site training data set, some sites consistently yielded poorer accuracy when no data from that site was included in training. For example, SITE5 yielded low accuracies in many analyses, likely due sitespecific differences in prostate MRI protocol: for example, the median inter-slice spacing at SITE5 was 4.7 mm compared to 2.8 mm across the other sites. One solution to this problem would be to adjust clinical imaging at this site to be more consistent with other sites; however, such a solution could be very disruptive. Note that this effect almost disappears in the patient-level cross-validation suggesting that these cases are probably not substantially harder to segment, as long as they are represented in the training data to some extent. This suggests that the more practical solution of retraining the segmentation network with some data from each site during the commissioning process may be effective.

The conclusions of this study should be considered in the context of its limitations. Our study focused exclusively on prostate segmentation, where deep-learning-based segmentation methods have become dominant and multi-site data sets are available. Reproducing our finding on other segmentation problems, once appropriate data is available, will be valuable. We observed variability between networks in their generalization to new sites; while we evaluated three different networks, we cannot conclude that all networks will need commissioning with data from each new site. Evaluating each network required training 58 networks, so a more exhaustive evaluation was not feasible for this work.

Our analysis of the accuracy of deep-learning-based segmentation methods demonstrated the performance on networks trained and tested on data from one or more sites can overestimate the performance at an unseen site. This suggests that segmentation evaluation and especially segmentation challenges should include data from one or more completely unseen sites in the testing data to estimate how well methods generalize. They also suggest that commissioning segmentation methods at a new site by retraining networks with a limited number of additional samples from that site could be an effective way to mitigate this problem.

References

1. Bakas, S., Menze, B., Davatzikos, C., Reyes, M., Farahani, K. (eds.): International MICCAI BraTS Challenge (2017)
2. Chen, H., Dou, Q., Yu, L., Qin, J., Heng, P.A.: VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images. NeuroImage (2017)
3. Ghavami, N., Hu, Y., Bonmati, E., Rodell, R., Gibson, E., Moore, C., Barratt, D.: Automatic slice segmentation of intraoperative transrectal ultrasound images using convolutional neural networks. In: SPIE Medical Imaging (Feb 2018)
4. Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., Gurusamy, K., Davidson, B., Pereira, S.P., Clarkson, M.J., Barratt, D.C.: Automatic multi-organ segmentation on abdominal ct with dense v-networks. IEEE Transactions on Medical Imaging (2018)
5. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. arXiv:1603.05027 (2016)
6. Landman, B., Xu, Z., Igelsias, J.E., Styner, M., Langerak, T.R., Klein, A.: MICCAI multi-atlas labeling beyond the cranial vault - workshop and challenge (2015), accessed July 2017

7. Litjens, G., Toth, R., van de Ven, W., Hoeks, C., Kerkstra, S., van Ginneken, B., Vincent, G., Guillard, G., et al.: Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. Med Image Anal 18(2), 359–373 (2014)
8. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: IEEE 3D Vis. pp. 565–571 (2016)
9. Mirzaalian, H., de Pierrefeu, A., Savadjiev, P., Pasternak, O., Bouix, S., Kubicki, M., Westin, C.F., Shenton, M.E., Rathi, Y.: Harmonizing diffusion mri data across multiple sites and scanners. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 12–19. Springer (2015)
10. Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C., Jagust, W., Trojanowski, J.Q., Toga, A.W., Beckett, L.: The alzheimer's disease neuroimaging initiative. Neuroimaging Clinics 15(4), 869–877 (2005)
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)
12. Styner, M.A., Charles, H.C., Park, J., Gerig, G.: Multisite validation of image analysis methods: assessing intra-and intersite variability. In: Medical Imaging 2002: Image Processing. vol. 4684, pp. 278–287. International Society for Optics and Photonics (2002)