In Domain Transfer Learning for Prostate MRI Segmentation

Larbi Touijer
DIBRIS-MaLGa
University of Genova
Genoa, Italy
larbi.touijer@edu.unige.it

Matteo Santacesaria
DIMA-MaLGa
University of Genoa
Genoa, Italy
matteo.santacesaria@unige.it

Francesca Odone
DIBRIS-MaLGa
University of Genoa
Genoa, Italy
francesca.odone@unige.it

Vito Paolo Pastore
DIBRIS-MaLGa
University of Genoa
Genoa, Italy
vito.paolo.pastore@unige.it

***Abstract*—In recent years, deep learning has been widely applied to different medical image analysis tasks. However, large-scale annotated datasets are typically unavailable in such a domain, potentially hindering deep learning applications. Transfer learning with a fine-tuning framework is a commonly adopted solution to this issue, exploiting large-scale natural image datasets (e.g., ImageNet) to pre-train a deep neural network, and fine-tuning the resulting model on the target dataset. A potential alternative could be gathering data coming from different specialized centers to increase the number of available training data. However, privacy issues as well as diverse acquisition modalities are important challenges to such a solution. In this paper, we investigate if small-scale datasets for in-domain fine-tuning can be beneficial over natural image datasets pre-training only. Using popular small-scale benchmark datasets of prostate MRI volumes and ImageNet pre-trained models, we show that there is always a benefit when using in-domain data to fine-tune the ImageNet pre-trained model, before fine-tuning it on the target dataset. Our results provide insights for a potential improvement of deep-learning-based prostate segmentation in MRI images, showing benefits when using data acquired in different specialized centers within a transfer learning framework.**

***Index Terms*—MRI Prostate, Segmentation, Deep learning, Transfer Learning**

I. Introduction

The medical domain has considerably benefited from deep learning in recent years. Automating the analysis of clinical images may result in a faster diagnosis for the patient, possibly reducing the intervention time in crucial situations [1]. Such a need for automation has paved the way for a massive application of deep learning in medical image analysis. In this domain, deep learning covers multiple use cases such as image classification [2], landmark detection [3], and semantic segmentation [4]. However, the need for large quantities of high-quality annotated data may hinder deep-learning applications in the medical domain. A commonly exploited solution is the adoption of a transfer learning framework, starting from a model typically pre-trained on ImageNet, and transferring the acquired knowledge by fine-tuning the resulting model on the target medical images. An alternative to only using ImageNet as a source dataset lies in exploiting in-domain images. Recently, different works have investigated the effect of in-domain transfer learning with discording results depending on the specific application and downstream task [5], [6]. Thus, the problem of evaluating the benefit of in-domain transfer learning for medical image analysis remains an open question. One could expect better performances for the downstream task, as ImageNet images and semantics are quite far from the medical domain. Moreover, using in-domain datasets for fine-tuning in a transfer learning framework would solve the privacy issue, that prevents direct data sharing between different diagnostic centers. Within such a framework, a deep learning model could be trained locally in each center, and the resulting weights could be exchanged instead of the original images.

The described scenario is reminiscent of a federated learning framework [7], where generally an entity called *aggregator* collects weights of the same model trained locally in different centers, called *parties*, combines them (e.g., with an average), and send back the final weights to the single parties, to improve the performance of the downstream task for all involved parties.

Inspired by this approach, in this paper, we investigate the effect of in-domain transfer learning for Magnetic Resonance Imaging (MRI) analysis. We focus on the segmentation task, as it plays a crucial role in identifying specific anatomical regions in MRI images, highlighting organs or tumors, and providing size estimates. Our main contributions can be regarded as follows: we perform a systematic study on the benefit of in-domain transfer learning for the Prostate MRI segmentation task, involving publicly available small-scale datasets, differing in acquisition modalities (e.g., the presence of specific coils and the strength of the magnetic field). The choice of this setup simulates the real-life scenario where diverse medical centers acquire a limited number of images of the same anatomical district, and data privacy issue prevents the direct exchange of data between them. We perform multiple stages of fine-tuning exploiting all the possible combinations of the explored datasets. Our results show that there is always a benefit, in terms of an increase in performance on the test set, when the ImageNet pre-trained models are fine-tuned with an in-domain dataset, before being fine-tuned on the target one. Even if further tests are necessary, we believe that this work may provide insights for tackling both data scarcity and privacy issues in the task of prostate MRI image segmentation using in-domain transfer learning.

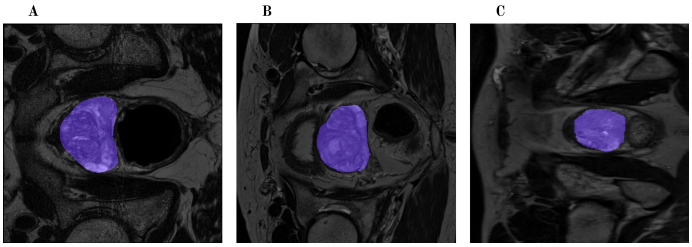

Fig. 1. Example slices of the Prostate MRI, from each one of the considered datasets, with an overlay of the corresponding segmentation mask. (A) is a sample slice from Site A; (B) is a sample slice from Site B; (C) is a sample slice from Site C.

II. Related Works

Over the years, there have been numerous contributions addressing the challenge of prostate MRI segmentation. The typical approach consists of the usage of various architectures based on Fully Convolutional neural Networks (FCN). For instance, Tian et al [8] propose PSnet (Prostate Segmentation network), and Cheng et al [9] opt for holistically nested networks, merging holistically nested edge detection with FCN. Zhang et al [10] modify the UNet architecture to design the ZNet model, specifically conceived for prostate segmentation. Milletari et al [11] propose another modified UNet-type architecture to address the volumetric prostate segmentation. All of the aforementioned works use single or multiple prostate datasets, without delving deep into the heterogeneity between the datasets coming from multiple medical centers [12].
Rundo et al [13] propose the USE-Net to generalize on heterogenous datasets coming from different hospitals. Their approach consists of multiple training/testing combinations merging the different datasets involved in their work. Liu et al [12] propose MS-Net, training their model on all the datasets merged and testing on each dataset individually.
Differently from the first cited works, our contribution regards the strategy of training in the scenario of multi-site datasets. Moreover, differently from [12] and [13], we keep all the datasets separated throughout the experiments, imitating the usual scenario of the data not being shared between medical centers and implementing in-domain fine-tuning in an attempt to improve the performance on the prostate segmentation task. Our results suggest that there is a benefit of using in-domain fine-tuning in the investigated framework.

III. Method

A. Dataset

In this section, we describe the datasets investigated in our experiments along with the adopted preprocessing procedures.

1) Dataset Details: In our work, we adopt the same two popular benchmark datasets for MRI prostate segmentation used in [12] and [13] : (i) *NCI-ISBI 2013 challenge* [14], (ii) *Initiative for Collaborative Computer Vision Benchmarking (Site C)* [15].
The first dataset includes data acquired in two different sites: Site A contains 30 MRI volumes collected at the Radboud University Nijmegen Medical Centre. Site B includes 30 MRI volumes acquired at the Boston Medical Center. The second dataset is referred to as Site C and presents 19 MRI volumes provided by the Hospital Center Regional University of Dijon-Bourgogne. These datasets perfectly illustrate the realistic scenario where multiple hospitals acquire images of the same anatomical district, with peculiar acquisition protocols and field strengths.

TABLE I
Summary of the three datasets details, listing the number of cases (patients), the MRI field strength in Tesla, the type of coil used in the machine to acquire the prostate MRI, and the number of 2D image slices extracted from the 3D volume.

Dataset	#Cases	Field Strength	Coil	#Slices
Site A	30	3T	Surface	396
Site B	30	1.5T	Endorectal	384
Site C	19	3T	No	468

2) Data pre-processing: The first transformation to the images is the extraction of the 2D image slices from the 3D volumes. The number of slices per volume differs from one site to another, but the selection strategy is the same: we select the slices where the segmentation mask is bigger than a fixed threshold (that we set at 200 pixels). The rationale is to evaluate the benefit of transfer learning when the prostate is evident in the MRI image. Sites A and B divide the prostate masks into the Central Gland (CG) and the Peripheral Zone (PZ). For our experiments and to be consistent with site C, we combine them into one region. Finally, we perform a Bias Field Correction [16], which is a common preprocessing step when dealing with MRI images. Table I summarizes the dataset details, highlighting the number of patients, the field strength, and the final number of image slices. Fig.1 showcases a sample slice from each of the three datasets involved in our experiments, with the segmentation mask overlayed and represented in purple.

B. Proposed Approach

In this section, we describe our proposed approach, schematically represented in Fig.2. We exploit a UNet++ FCN, with an EfficientNetB3 encoder backbone pre-trained on ImageNet. At this stage, the model is fine-tuned on a prostate MRI image dataset, exploiting the provided segmentation masks. In our work, we investigate three different small-scale

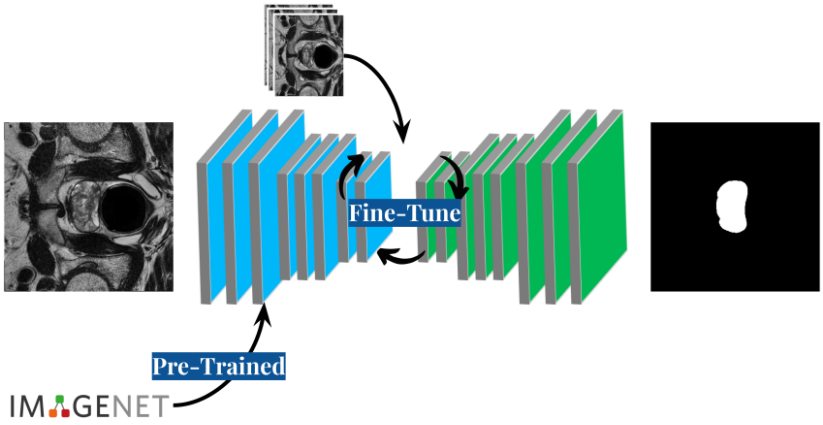

Fig. 2. Schematic representation of the proposed pipeline. The architecture is a UNet++ FCN with an EfficientNetB3 backbone pre-trained on ImageNet, which we train on a prostate MRI source dataset, and then fine-tune on the target dataset.

prostate datasets acquired in different centers and with diverse modalities. Considering each of them as a target dataset, our idea is to exploit the remaining two for in-domain fine-tuning, in an attempt to improve the performances of the deep learning model on the target one. To provide an example, considering the Site A dataset, we first fine-tune our ImageNet pre-trained model on either the Site B dataset or the Site C dataset, and then we further fine-tune the resulting model on the target Site A dataset. The same procedure applies when the other two datasets are considered as targets. To further investigate the benefit of in-domain small-scale datasets fine-tuning, we explore a double fine-tuning procedure, where for each one of the three datasets investigated in our work, the ImageNet pre-trained model is fine-tuned in sequence on the remaining two, and the resulting model is further fine-tuned on the actual target dataset.

IV. Experiments

In this section, we provide details on the implementation and training protocols for our experiments. Finally, we present and discuss the obtained results.

A. Experiment details

The three datasets used in this work do not come with a native test set. Thus, in our training protocol, we hold out 20% of the dataset for testing. To evaluate the robustness of the proposed approach and for model selection, we perform a 5-fold cross-validation procedure. We split the data per patient, to avoid the risk of ending up with 2D slices of the same patients spread among the training, validation, and test sets. For the bias field correction, we used the *N4BiasFieldCorrectionImageFilter* implementation from the SimpleITK library [17]. The FCN architectures and backbones are taken from the python library *segmentation_models_pytorch* [18] implementation. The models are trained for 200 epochs with a batch size of 16, with the Adam optimizer and a cosine-annealing scheduler with a learning rate of 10^{-3}. Images are standardized prior to being fed to the ImageNet pre-trained model, using the training dataset mean and standard deviation.

B. Results

1) Evaluation metrics: In our experiments, we evaluate the results of the segmentation model with two widely used evaluation metrics: Dice score and Intersection over Union (IoU) [19]. Both metrics are similarity measures between two images containing an area of interest, where one is the ground truth (GT) and the other is the model's prediction output.

The IoU (Eq.(1)) is the ratio between the area obtained by intersecting the prediction and ground truth and the area obtained by uniting the prediction and the ground truth. The IoU is comprised between 0 and 1. The higher the value the better the prediction and the closer to the ground truth.

$$IoU(Pred, GT) = \frac{Pred \cap GT}{Pred \cup GT}. \quad (1)$$

The Dice Score (Eq.(2)) is the ratio between two times the area obtained by intersecting the prediction and ground truth, and the sum of the two areas. The resulting score ranges between 0 and 1. The higher the value, the better the segmentation performance.

$$Dice(Pred, GT) = \frac{2\,|Pred \cap GT|}{|Pred| + |GT|}. \quad (2)$$

The adopted evaluation metrics are computed with pixel-wise comparisons. In our results, we report both metrics in percentage.

2) Comparative study on architecture and backbone: First, we perform a comparative study to select the best-performing architectures and backbones, to be used for the transfer learning experiment. Exploiting the Site A dataset, we consider four different FCN architectures (UNet, UNet++, MANet, and DeepLabV3), and seven different encoders including four EfficientNets and three ResNets models. Table II reports the validation Dice score averaged among the 5 folds (see Section IV-A) obtained for each possible combination of architectures and backbones. The color

TABLE II

THE VALIDATION DICE SCORE (IN PERCENTAGE), REPRESENTED AS A MEAN±STD OVER A 5-FOLD CV, OF TRAINING A MIXTURE OF BACKBONES AND FCN ARCHITECTURES ON THE SITE A DATASET. THE COLOR SCHEME RANGES FROM LIGHT YELLOW TO BLUE, WHERE THE HIGHER THE VALIDATION SCORE, THE MORE BLUE THE CELL IS. THE BEST COMBINATION IS EFFICIENTNETB3 WITH UNET++.

	UNet	UNet++	MANet	DeepLabV3
EfficientNetB0	87.79 ± 0.37	88.31 ± 1.37	87.38 ± 2.14	88.07 ± 0.42
EfficientNetB1	88.57 ± 2.33	88.18 ± 0.76	88.06 ± 1.1	87.85 ± 1.5
EfficientNetB2	89.09 ± 0.46	88.79 ± 0.05	88.25 ± 0.35	87.59 ± 0.71
EfficientNetB3	88.84 ± 1.76	**89.27** ± 1.62	88.71 ± 0.74	87.95 ± 1.21
ResNet18	86.72 ± 0.99	88.61 ± 0.64	86.16 ± 1.5	87.5 ± 0.21
ResNet34	87.29 ± 1.03	87.81 ± 0.87	88.17 ± 0.79	89.04 ± 0.13
ResNet50	87.14 ± 0.97	86.22 ± 2.99	85.26 ± 0.43	88.93 ± 1.24

TABLE III

TEST DICE SCORE OF FINE-TUNING EFFICIENTNETB3 WITH UNET++ ON THE TARGET DATASET, WITH THE MODEL TRAINED ON THE SOURCE DATASET. ALTERNATIVE: TEST DICE SCORE OF FINE-TUNING EFFICIENTNETB3 WITH UNET++ ON THE TARGET DATASETS, WITH THE MODEL TRAINED ON THE DIFFERENT COMBINATIONS OF SOURCE DATASETS.

Target dataset / *Source weights*	→Site A		→Site B		→Site C	
	Dice%	*IoU%*	*Dice%*	*IoU%*	*Dice%*	*IoU%*
Site A	89.09	82.52	91.42	84.89	**90.48**	**83.34**
Site B	90.90	84.34	88.68	81.41	88.27	80.41
Site C	**92.34**	**86.75**	91.22	84.74	87.11	80.26
Site B⟶Site C	91.65	85.53	-	-	-	-
Site C⟶Site B	91.48	85.07	-	-	-	-
Site A⟶Site C	-	-	**91.62**	**85.36**	-	-
Site C⟶Site A	-	-	90.89	84.13	-	-
Site A⟶Site B	-	-	-	-	89.36	81.92
Site B⟶Site A	-	-	-	-	89.94	82.44
All_Data	91.13	84.83	90.83	84.07	87.68	80.48

scheme used in the table reflects the value of the validation score. As the score increases from low to high values, the color scheme ranges from light yellow to blue. Based on the validation Dice score, the best-performing model is UNet++ with an EfficientNetB3 encoder, reaching an average Dice score (in percentage) of 89.27 ± 1.62. For this reason, we carry on the rest of the experiments using the EfficientNetB3 with the UNet++ combination.

3) Experiments on in-domain fine-tuning: In the next experiments, we exploit in-domain fine-tuning to improve the performances of our model on the prostate segmentation task. For these experiments, we report the results corresponding to the best validation fold. First, we derive a baseline, that is the test Dice score and IoU score corresponding to directly fine-tuning the UNet++ with ImageNet pre-trained EfficientNetB3 encoder on each of the target datasets. The obtained results are reported in Table III. The test Dice scores (in percentage) are 89.09 on Site A, 88.68 on Site B, and 87.11 on Site C.

Then, we apply the fine-tuning protocols described in Section III-B. In Table III, the operator $\longrightarrow$ refers to fine-tuning. To provide an example, the entry Site B $\longrightarrow$ Site C in the *Source weights* column, means that the model has been fine-tuned on Site B and further fine-tuned on Site C, before being fine-tuned again on the remaining target dataset (Site A). When only a dataset is reported in this Table, it means that the ImageNet pre-trained model has been fine-tuned on that dataset prior to being fine-tuned again on the remaining target datasets. Overall, there is an increase in test Dice scores and IoU scores with the application of in-domain fine-tuning. More specifically: Site A gains an average of 2.5% in Dice score and 3.1% in IoU score when the model is fine-tuned on Site B or Site C; Site B gains an average of 2.6% in Dice score and 3.4% in IoU score when fine-tuning the model on Site A or Site C; Site C gains an average of 2.4% in Dice score and 1.6% in IoU score when fine-tuning the model on Site A or Site C.

Regarding the double in-domain fine-tuning, it has a marginal or no benefit in comparison to the previous experiment.

As a benchmark, we fine-tune our EfficientNetB3-UNet++ model on a dataset obtained by merging the three datasets used in our work, computing the Dice score and IoU score for the test set of each dataset separately. Interestingly, as we can see in the last row of Table III (All_Data), this setup provides a higher Dice score and IoU score with respect to training each dataset separately, but lower performances when compared to the in-domain fine-tuning experiment. Finally, there seems to be a correlation between the dataset strength field and how effective the fine-tuning is. As highlighted in Table I, Site B has a field strength of 1.5T, whereas Site A and Site C have a field strength of 3T. In both cases where Site B is used for in-domain fine-tuning, the corresponding Dice score is inferior in comparison when the usage of the other two datasets as sources. This difference might be explained by the higher-quality images produced by 3T MRI machines.

V. CONCLUSION

In this paper, we evaluate the effect of in-domain fine-tuning for the segmentation task of Prostate MRI images. We employ a deep learning pipeline where a UNet++ with an ImageNet pre-trained EfficientNetB3 backbone is fine-tuned

on in-domain target MRI datasets. Using three benchmark prostate MRI datasets, we perform an empirical study considering all the possible combinations of in-domain fine-tuning and comparing the results in terms of test Dice score and IoU score, with the original model being only pre-trained on ImageNet. Our results show that there is always a benefit when using in-domain fine-tuning, with performance slightly superior to fine-tuning the same model on a dataset obtained by merging all three datasets. Even if further work is necessary, we believe that these results provide insights for improving prostate segmentation in MRI images, by presenting a possible way to exploit datasets coming from different specialized centers and acquired with different modalities within a transfer learning framework, thus providing a potential way around privacy issues that prevent sharing the data directly.

FUNDING

This work was funded by the European Union - NextGenerationEU. However, the views and opinions expressed are those of the authors alone and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

ACKNOWLEDGMENT

This work was carried out within the framework of the project "RAISE - Robotics and AI for Socio-economic Empowerment" and has been supported by European Union - NextGenerationEU.

REFERENCES

[1] Deeksha Kaul, Harika Raju, and BK Tripathy, "Deep learning in healthcare," Deep Learning in Data Analytics: Recent Techniques, Practices and Applications, pp. 97–115, 2022.

[2] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez, "A survey on deep learning in medical image analysis," Medical image analysis, vol. 42, pp. 60–88, 2017.

[3] Erik Meijering, "Deep learning in bioimaging," Computational and Structural Biotechnology Journal, vol. 18, pp. 2301–2313, 2020.

[4] Risheng Wang, Tao Lei, Ruixia Cui, Bingtao Zhang, Hongying Meng, and Asoke K Nandi, "Medical image segmentation using deep learning: A survey," IET Image Processing, vol. 16, no. 5, pp. 1243–1267, 2022.

[5] Andrea Maracani, Vito Paolo Pastore, Lorenzo Natale, Lorenzo Rosasco, and Francesca Odone, "In-domain versus out-of-domain transfer learning in plankton image classification," Scientific Reports, vol. 13, no. 1, pp. 10443, 2023.

[6] Larbi Touijer, Vito Paolo Pastore, and Francesca Odone, "Food image classification: The benefit of in-domain transfer learning," in International Conference on Image Analysis and Processing. Springer, 2023, pp. 259–269.

[7] Mashal Khan, Frank G Glavin, and Matthias Nickles, "Federated learning as a privacy solution-an overview," Procedia Computer Science, vol. 217, pp. 316–325, 2023.

[8] Zhiqiang Tian, Lizhi Liu, Zhenfeng Zhang, and Baowei Fei, "Psnet: prostate segmentation on mri based on a convolutional neural network," Journal of medical imaging, vol. 5, no. 2, pp. 021208–021208, 2018.

[9] Ruida Cheng, Holger R Roth, Nathan Lay, Le Lu, Baris Turkbey, William Gandler, Evan S McCreedy, Tom Pohida, Peter A Pinto, Peter Choyke, et al., "Automatic magnetic resonance prostate segmentation by deep learning with holistically nested networks," Journal of medical imaging, vol. 4, no. 4, pp. 041302–041302, 2017.

[10] Yue Zhang, Jiong Wu, Wanli Chen, Yifan Chen, and Xiaoying Tang, "Prostate segmentation using z-net," in 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE, 2019, pp. 11–14.

[11] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in 2016 fourth international conference on 3D vision (3DV). Ieee, 2016, pp. 565–571.

[12] Quande Liu, Qi Dou, Lequan Yu, and Pheng Ann Heng, "Ms-net: multi-site network for improving prostate segmentation with heterogeneous mri data," IEEE transactions on medical imaging, vol. 39, no. 9, pp. 2713–2724, 2020.

[13] Leonardo Rundo, Changhee Han, Yudai Nagano, Jin Zhang, Ryuichiro Hataya, Carmelo Militello, Andrea Tangherloni, Marco S Nobile, Claudio Ferretti, Daniela Besozzi, et al., "Use-net: Incorporating squeeze-and-excitation blocks into u-net for prostate zonal segmentation of multi-institutional mri datasets," Neurocomputing, vol. 365, pp. 31–43, 2019.

[14] Bloch Nicholas, Madabhushi Anant, Huisman Henkjan, Freymann John, Kirby Justin, et al., "Nci-proc. ieee-isbi conf. 2013 challenge: Automated segmentation of prostate structures," The Cancer Imaging Archive, vol. 5, 2015.

[15] Guillaume Lemaître, Robert Martí, Jordi Freixenet, Joan C Vilanova, Paul M Walker, and Fabrice Meriaudeau, "Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric mri: a review," Computers in biology and medicine, vol. 60, pp. 8–31, 2015.

[16] Jaber Juntu, Jan Sijbers, Dirk Van Dyck, and Jan Gielen, "Bias field correction for mri images," in Computer Recognition Systems: Proceedings of the 4th International Conference on Computer Recognition Systems CORES'05. Springer, 2005, pp. 543–551.

[17] Ziv Yaniv, Bradley C Lowekamp, Hans J Johnson, and Richard Beare, "Simpleitk image-analysis notebooks: a collaborative environment for education and reproducible research," Journal of digital imaging, vol. 31, no. 3, pp. 290–303, 2018.

[18] Pavel Iakubovskii, "Segmentation models pytorch," https://github.com/qubvel/segmentation_models.pytorch, 2019.

[19] Jeroen Bertels, Tom Eelbode, Maxim Berman, Dirk Vandermeulen, Frederik Maes, Raf Bisschops, and Matthew B Blaschko, "Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice," in Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22. Springer, 2019, pp. 92–100.