A COMPARATIVE STUDY OF DEEP ARCHITECTURES FOR VOXEL SEGMENTATION IN VOLUME IMAGES

F. Wagner[1,2], H.-G. Maas[1]

[1]Institute of Photogrammetry and Remote Sensing, TU Dresden, Germany
[2]franz.wagner@tu-dresden.de

KEY WORDS: Deep Learning, 3D Segmentation, 3D CNN, Tomography, Magnetic Resonance Imaging, CNN Comparison

ABSTRACT:

This study investigates the performance of eight different deep learning architectures for voxel segmentation in volume images. The motivation is to segment carbon in carbon reinforced concrete (CRC) in micro-tomography (µ-CT) data. Although there are many 3D convolutional neural networks (CNNs) available, it is not yet clear which one works best for these specific tasks. In this study, the authors compare the following networks: DenseVoxNet, HighResNet, Med3D, Residual 3D U-Net, 3D SkipDenseSeg, 3D U-Net, V-Net, and LV-Net. To provide a more general recommendation for selecting a neural network, three medical datasets were added in addition to the three CRC datasets to facilitate the selection of an appropriate network based on the dataset characteristics. The experiments emphasize the importance of the initial random state, used for example to initialize the network weights. On average, the 3D U-Net is the best generalizing CNN, followed by the Residual 3D U-Net and the 3D SkipDenseSeg. While the 3D U-Net is a good architecture to start with, the experiments show that it does not perform best on all domains. To achieve optimal results, the authors propose recommendations for selecting a 3D neural network based on the dataset attributes.

1. INTRODUCTION

To minimize the use of materials in buildings or other structures, carbon reinforced concrete can be used (Beckmann et al., 2021). As less concrete is used herein, the position of the carbon elements is of great importance. Therefore, we used a micro-tomography instrument (µ-CT), which consists of an X-ray source and a camera that takes a large number of projections as the object rotates. From these images, a volumetric reconstruction of the object is created. In these volumes, the task is to segment the carbon components, which can be done using convolutional neural networks (CNNs).

In 2012, (Ciresan et al., 2012) proposed a deep neural network approach to the task of pixel-wise classification, also known as semantic segmentation. Although the approach worked, it was relatively slow, so (Ronneberger et al., 2015) introduced the famous U-Net, which outperformed the previous approach on both speed and performance. Since then, many other segmentation-related neural networks have been published (e.g.: SegNet (Badrinarayanan et al., 2015), DeepLab (Chen et al., 2018), GCN (Peng et al., 2017) or UPerNet (Xiao et al., 2018)) to solve 2D segmentation problems. These types of CNNs have also been successfully applied to 3D data such as in computed tomography (CT), magnetic resonance imaging (MRI), or electron microscopy (EM). However, the performance of 3D convolutions, such as in the 3D U-Net (Çiçek et al., 2016), has further improved the segmentation accuracy on such datasets. (Mester et al., 2022) successfully applied a 3D U-Net to segment carbon rovings (bundles of single carbon fibers aligned in a grid) in concrete. However, they have not determined whether this is the optimal network for this task due to the lack of comprehensive reviews and comparisons of 3D CNNs. In medicine, 3D datasets are very common and thus 3D CNNs are well-known. However, most medical studies dealing with 3D data still use 2D CNNs for their analysis. According to (Singh et al., 2020) and (Niyas et al., 2022), only 8-11 % of published medical papers use 3D CNNs, although they would be suitable for this purpose. To fill this gap and to determine the best network for our research, this study compares 8 different 3D CNNs using the AiSeg project (https://gitlab.com/frawa/aiseg). The investigated networks are:

- DenseVoxNet (Yu et al., 2017)
- HighResNet (Li et al., 2017)
- Med3D (Chen et al., 2019)
- Residual 3D U-Net (Lee et al., 2017)
- 3D SkipDenseSeg (Bui et al., 2019)
- 3D U-Net (Çiçek et al., 2016)
- V-Net (Milletari et al., 2016)
- LV-Net (Lei et al., 2020)

To the best of our knowledge, there are currently no other comprehensive 3D CNN comparisons. With the exception of the Med3D publication, all networks listed were tested on a single domain only. Therefore, in this study, all networks were compared on 6 different datasets: Three public and three new datasets representing electron microscopy, CT, and MRI data. All datasets have their own challenges that networks must overcome.

2. DATASETS

A typical volumetric dataset consists of a large number of 2D images stacked on top of each other. What is represented by a pixel in a 2D image corresponds to a voxel in three-dimensional space. The datasets were acquired by different acquisition devices such as magnetic resonance imaging, electron microscopy and computed tomography. An overview of the devices used and the size of the datasets can be found in table 1. All datasets in use come with a dataset specific challenge that the networks need to handle. The new datasets for carbon rovings, concrete pores, and polyethylene fibers were created using Dragonfly (Object Research Systems (ORS), 2021).

This contribution has been peer-reviewed.

Method	Dataset	Classes	Voxels (million)
CT	Carbon Rovings	2	9326.6
CT	Concrete Pores	2	9932.1
CT	PE Fibers	2	1486.4
EM	Brain Mitochondria	2	129.8
MRI	BraTS	4	9222.6
MRI	Head and Neck Cancer	9	975.2

Table 1: Overview of the datasets including their data size represented in voxels

2.1 Carbon Rovings

At RWTH Aachen University, Germany, a laboratory mortar extruder is used to integrate soft impregnated carbon textiles, also called rovings, into structural components (figure 1). (Mester et al., 2022) were interested in the surface area of the rovings inside the concrete in order to use them in the context of a coupled multiscale method. This is a challenging task due to the fact that µ-CT basically determines the physical density of a voxel, and physical densities of carbon and some concrete constituents are rather similar. The dataset created for this purpose (Wagner, 2023a) is represented by 3 different CT scans. The first two scans were sliced into multiple training and validation sub-volumes of size 128 x 256 x 256 (Depth (d) x Height (h) x Width(w)) voxels, while the third scan is used for testing only. The unaugmented dataset consists of 129 training and 33 validation volumes plus the test volume. The training data was augmented using random rotations around the X, Y and Z axes, resulting in 1134 volumes with a voxel size of 9.4 µm. This dataset contains rather large structures that should be segmented by the models.

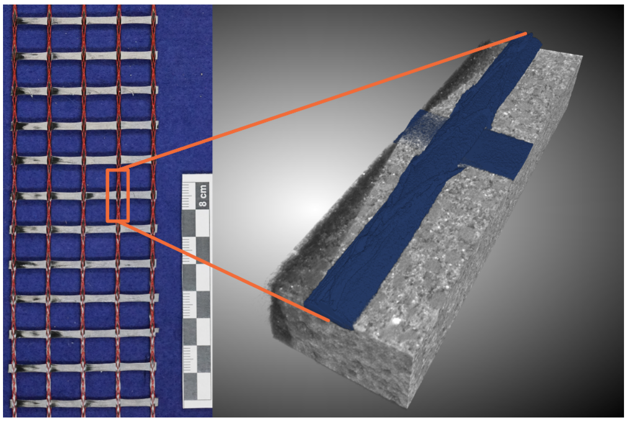

Figure 1: Carbon roving grid (left) and CT scan of a roving in concrete (blue, right) visualized with Dragonfly.

2.2 Concrete Pores

Segmenting pores in concrete is fairly straightforward task, for the obvious reason of different physical density. However, applying a simple thresholding may not be sufficient because the surrounding air is also segmented and some reconstructions may be extremely noisy depending on the power of the x-ray source used. Furthermore, depending on the reconstruction settings, the threshold has to be manually adjusted for each CT volume. Therefore, a new dataset for the segmentation of pores in concrete has been created (figure 2) (Wagner, 2023b). In its current state, it consists of 8 different CT scans, reduced to regions of interest. They were manually labeled and each scan was sliced into multiple sub-volumes of size 256 x 512 x 512 (d x h x w) voxels resulting in 148 training, 43 validation and 21 test volumes with different voxel sizes. This dataset contains very small to very large structures that the models should segment.

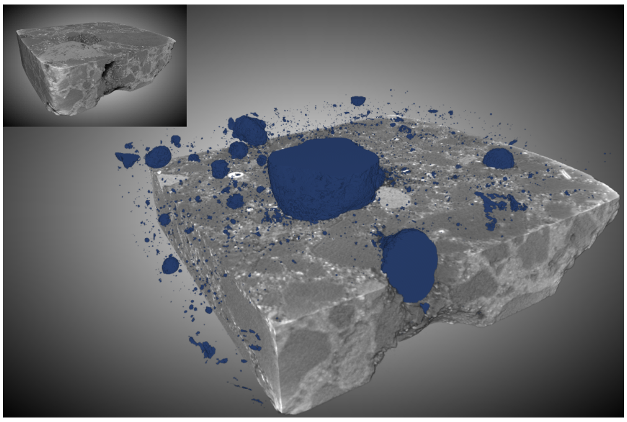

Figure 2: Miniature without labels and labeled pores (blue) visualized with Dragonfly.

2.3 Polyethylene Fibers

Segmentation of polyethylene fibers in strain-hardened cement-based composites is a very difficult task, since individual fibers of carbon or polyethylene bring the challenge that their density is very similar to one of quartz sand particles, resulting in almost identical gray values (Lorenzoni et al., 2020). The created PE fibers dataset (Wagner, 2023c) consists of only 3 spatially disjoint volumes of size 20 x 512 x 512 (d x h x w) voxels (figure 3) (voxel size: 4 µm). Since this is a rather small dataset, it was geometrically enlarged by combinations of rotation (using multiple angles), resizing, flipping, tilting, and squeezing using the AiSeg project. This is the only extensively augmented dataset, as training on so few images would lead to overfitting. A total of 397 training and 100 validation volumes were created. The original volumes were used as test volumes. The use of geometric augmentation results in different shapes for most of the new volumes. This dataset presents the challenge of using very little and only augmented data to predict real volumes. In addition, only very thin objects are included in this dataset.

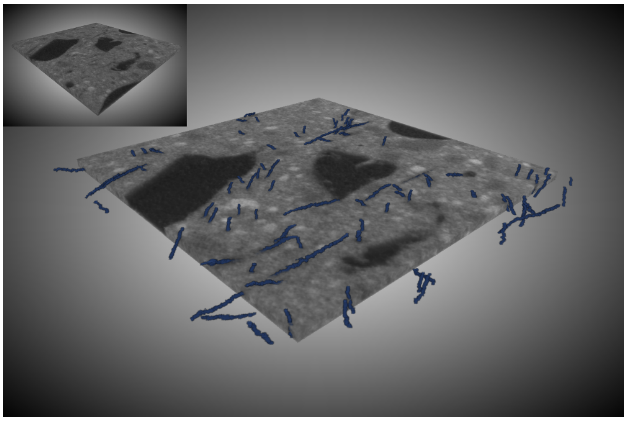

Figure 3: Miniature without labels and labeled volume containing fibers (blue) visualized with Dragonfly.

2.4 Brain Mitochondria

The Electron Microscopy dataset was created at EPFL in Lausanne to segment brain mitochondria in three-dimensional data

(Lucchi et al., 2013) (figure 4). It represents a small section of the hippocampal CA1 region of the brain and consists of two annotated volumes with a voxel size of 5 nm. The dimensions of the volumes are as follows: 165 x 768 x 1024 (d x h x w) voxels. The data represents rather large structures in a small dataset. However, the images are not perfectly aligned. The challenge for the models is to learn these features with limited data.

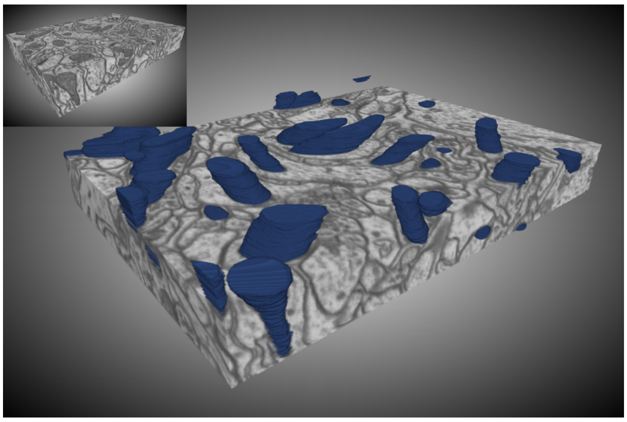

Figure 4: Miniature without labels and labeled brain mitochondria (blue) visualized with Dragonfly.

2.5 Brain Tumor Segmentation Challenge (BraTS)

The BraTS (2020) multimodal magnetic resonance imaging data were collected to develop and evaluate state-of-the-art methods for segmentation of brain tumors (namely gliomas). The published training data consists of 369 x 4 preoperative MRI scans of human brains, acquired at 19 different institutions. Each brain was imaged using native (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR) scans. The masks contain four labels: background, necrotic and non-enhancing tumor core (NCR/NET), peritumoral edema (ED) and GD-enhancing tumor (ET) (figure 5). For this study, all 1476 scans (369 x 4) were divided into 1033 training, 297 validation and 146 test volumes. Each volume is of size 155 x 240 x 240 (d x h x w) voxels with a voxel size of 1 mm. (Menze et al., 2015), (Bakas et al., 2017), (Bakas et al., 2018)
The challenges of this dataset are that it represents a multiclass problem and that the structures are intergrown.

2.6 Head and Neck Cancer

Radiation therapy is an important approach in the treatment of tumors. To prevent damage, the contours of tumors must be segmented with a high degree of confidence. The dataset of the Brain and neck cancer detection AAPM RT-MAC Grand Challenge 2019 (Cardenas et al., 2020), published in the Cancer Imaging Archive (Clark et al., 2013), aims to reduce common observer variability by making segmentation algorithms comparable. The MRI dataset consists of 55 scans, using T2-weighted images. 31 of them are used as training, 12 as validation and 12 as test data. Each volume has a size of 120 x 512 x 512 (d x h x w) voxels with a pitch of 0.5 mm per pixel and 2 mm per slice (2 x 0.5 x 0.5 mm³ (d x h x w)). The ground truth contains nine classes, resulting from a right and left subdivision and the background. The four main contours are: parotid glands, submandibular glands, level 2 and level 3 lymph nodes (figure 6). (Cardenas et al., 2019)

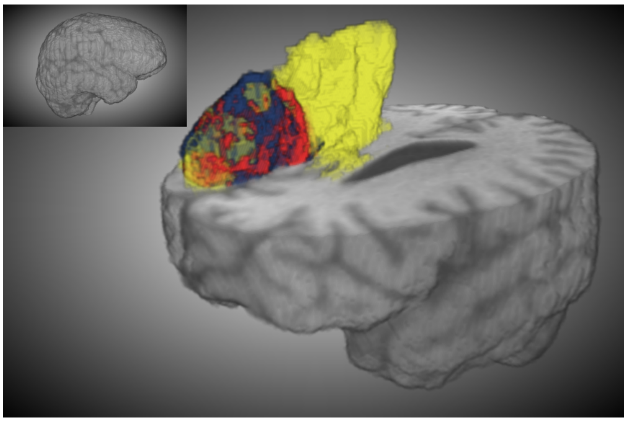

Figure 5: Miniature without labels and labeled BraTS data. Blue: Necrotic and non-enhancing tumor core; Yellow: Peritumoral edema; Red: Gadolinium-enhancing tumor visualized with Dragonfly.

The dataset consists of rather large structures. However, it has a different pitch in depth than in height and width. Also, some of the structures merge into each other.

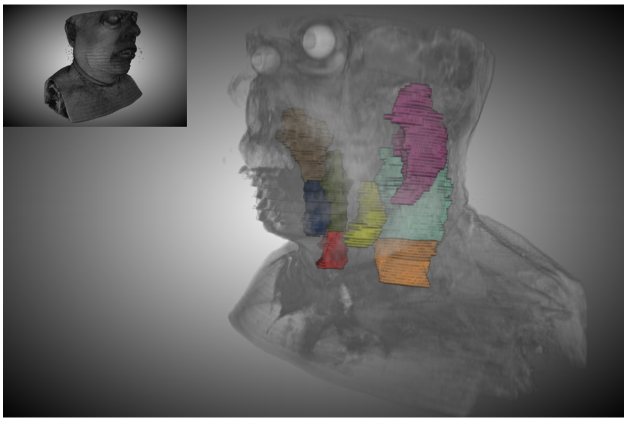

Figure 6: Miniature without labels and labeled organs at risk or tumors visualized with Dragonfly. Pink: submandibular gland (right); Teal: lymph node level 2 (right); Yellow: submandibular gland (right); Orange: lymph node level 3 (right)

3. METHODS

3.1 Data Preprocessing

When training 3D CNNs, most volumes do not fit into the video RAM (VRAM). Therefore, a training volume is divided into several equally overlapping sub-volumes. For example, in figure 7, the initial volume has the shape of 154 x 240 x 240 (depth (d) x height (h) x width (w)) voxels and is divided into 8 sub-volumes of 96 x 144 x 144 (d x h x w) voxels each. Each dataset presented is preprocessed according to this scheme. All datasets were, if they not already are, divided into training (needed to adjust the models weights), validation (needed to tune hyperparameters) and test (to evaluate the final model) data.

3.2 Hyperparameters

Hyperparameters are parameters that are set before a machine learning model is trained and affect how the model is trained and how it performs. Since we are comparing different architectures, the parameters for hidden layers, number of neurons

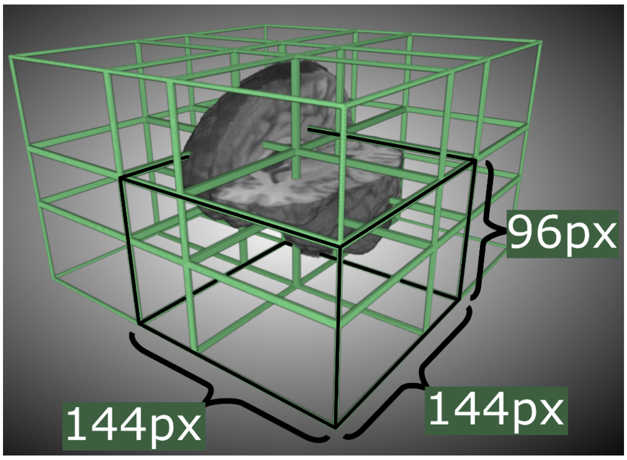

Figure 7: Example of a BraTS test volume with a size of 154 x 240 x 240 (d x h x w) voxels divided into 8 sub-volumes with a size of 96 x 144 x 144 (d x h x w) voxels. Black contours: one sub-volume.

in each layer and activation function are of course not identical. Due to the different dataset sizes, different input heights and widths (and depths for the fiber dataset) had to be used (table 2). Depending on the complexity and the number of parameters of the CNNs, the batch size is different for most networks. Since all batch sizes used are smaller than 16, group normalization was used as the normalization layer instead of batch normalization. The advantage of this technique is that it is more stable and produces better results than batch normalization on small batch sizes (Wu and He, 2018). The remaining hyperparameters related to training, such as total iterations, learning rate, optimizer, etc., are set the same to ensure a fair comparison. To evaluate the performance during training and validation, the cross-entropy loss (eq. 1) was used on a multi-class problem. For a binary segmentation, the binary cross entropy loss (eq. 2) was used. The optimizer was set to Adam (Kingma and Ba, 2015) with an initial learning rate of 0.001.

Dataset	Height and Width	Depth	GPUs
Rovings	128	64	8
Pores	128	64	8
Fibers	256	16	4
Mitochondria	128	64	4
BraTS	128	64	8
Head Neck Cancer	128	64	4

Table 2: Overview of the datasets, their input dimensions and used GPUs during training.

$$L_{multi\,class} = -\sum_{c=1}^{C} y_{o,c} \log(p_{o,c}) \tag{1}$$

$$L_{binary} = -(y \log(p) + (1-y) \log(1-p)) \tag{2}$$

where: L = Loss value
p = Probability: observation o is of class c
y = ∈[0, 1]; indicator if classification c is correct
C = Number of classes

3.3 Metrics

During the training of a neural network, its performance was monitored using the loss (section 3.2) and the accuracy, which refers to the ratio of correctly classified voxels to all voxels in a volume. During training, the weights were adjusted iteratively in such a way that the loss value gets minimized. Using the validation loss and accuracy, the following statements can be derived:

- Low loss, low accuracy: many small errors
- Low loss, high accuracy: very little errors - best case
- High loss, low accuracy: many big errors - worst case
- High loss, high accuracy: very little but big errors

However, accuracy is not a reliable measure: Consider a volume of size 10 x 10 x 10 voxels containing only 50 voxels that belong to the foreground. If the network predicts every voxel to be background, the accuracy is still 95%, even though it failed badly. Therefore, we used the common *DICE* coefficient, also called the overlap index or F1-score (Taha and Hanbury, 2015), to measure the equality between the ground truth and the segmentation. Using the True Positives (TP), False Positives (FP) and False Negatives (FN), it is calculated by:

$$DICE = \frac{2TP}{2TP + FP + FN} \tag{3}$$

In the case of a multi-class problem (*BraTS* and *Head and Neck Cancer*), the *DICE* per class, calculated using eq. 3, and the class-biased mean (*mDICE*, eq. 4) will be be provided. For a general comparison, we also present a weighted mean *DICE* (*wDICE*, eq. 6). The weighting ensures that the score represents a global performance rather than a class-biased result. The weight is calculated by dividing the voxels (N) representing a class (c) by all voxels in the current volume (v).

$$mDICE = \frac{\sum_{c=1}^{n} DICE_c}{n} \tag{4}$$

$$w_c = \frac{\sum_{v}^{t} N_{c_v}}{\sum_{v}^{t} N_v} \tag{5}$$

$$wDICE = \sum_{c=1}^{n} w_c \cdot DICE_c \tag{6}$$

where: w_c = weighting per class
t = total volumes
n = all classes

During testing, the test volumes were subdivided as explained in section 3.1. Therefore, a voxel is represented by 2 or more logits at the overlapping regions. A logit is the direct output of the CNN at a single voxel. Furthermore, due to the convolutions and missing information at the edges, the output of a network is less certain at the corners than at the center of the prediction. Therefore, a common practice is to apply a 3D Gaussian weighting. In MONAI (Medical Open Network for AI, a PyTorch-based, open-source framework for deep learning) (Cardoso et al., 2022), a default sigma of $\sigma = input\_size \cdot$

0.125 is used, which depends on the input dimension. This results in very different weightings, e.g. from 0.00038 (input dimension of 128) to 0.61 (input dimension of 512) at the corners. Instead, we decided to weight the logits in such a way that the values at the corners are always weighted $1/3$ to those in the center (eq. 7). After weighting and combining all the subvolumes, the final prediction was evaluated against the ground truth.

$$\sigma = \sqrt{\frac{-0.5 \cdot (d^2 \cdot h^2 \cdot w^2)}{\log \frac{1}{3}}} \tag{7}$$

3.4 Experimental Design

The experiments were performed on a High Performance Computing (HPC) system. Each training had 12 cores at 2.0 GHz and 60 GB RAM. Depending on the size of the dataset, 4 or 8 NVIDIA A100-SXM4 GPUs (40 GB VRAM) were used to train a network (table 2). To save computational time, reproducibility was disabled for all trainings since "deterministic operations are often slower than non-deterministic operations" (Paszke et al., 2019). This causes the values of the final weights to vary between two training runs due to different initial random states. The random state refers to the seed value used by random number generators within a machine learning algorithm. Therefore, the experiment is divided into two parts: In the first part, a rough estimation of the standard deviation (σ) of the *DICE* coefficient is performed using the head and neck cancer dataset (section 2.6). The performance of the networks is then further evaluated using the remaining five datasets.

Since the head and neck cancer dataset is quite small, the computation time is expected to be short, so this dataset was chosen to compute the standard deviation. However, since the training is still very computationally intensive, each network was trained only three times on the head and neck cancer dataset. For the other five datasets, the networks were trained only once.

With the exception of the PE fiber dataset (section 2.3), we did not perform strong augmentation, as this would change the properties of the datasets and thus make interpretation difficult. However, since this is a common practice, we refer to the AiSeg project for better and more robust results. The software is able to perform 3D offline and online augmentation as described in (Wagner et al., 2023).

4. RESULTS AND DISCUSSION

This section is divided into two parts. First, a rough estimate of the standard deviation regarding the *DICE* coefficient is made using the Head and Neck Cancer Dataset in dependence of the random initialization is conducted. Second, the performance of each network on the remaining five datasets is investigated. All neural networks were trained from scratch.

To make a rough estimate of the training duration, we approximated the duration as if we trained each dataset on a single A100-SXM4 GPU, resulting in over 101 days of training. If we had used a single RTX 3090, the training duration would have increased to almost a year (359 days).

4.1 Impact of Initial Random States (Head and Neck Cancer Dataset)

The purpose of this section is to make the reader aware that different initializations are likely to produce different results. To give a rough estimate, we trained all networks three times on the relatively small (table 1) Head and Neck Cancer dataset (section 2.6) to calculate an average and the standard deviation. Table 3 shows that the Med3D architecture with the ResNet10 backbone achieves the lowest σ with 0.08% and the V-Net the highest with 0.62%. Using this information, the results in the following sections should be treated with keeping these results in mind. Also, the standard deviation should decrease with the size of a dataset because there are many more volumes and therefore the variance of a dataset is likely to be higher.

Head and Neck Cancer	Parameters	***wDICE* (%)**		
Network (Backbone)	(million)	**Mean**	$\boldsymbol{\sigma}$	**Rank**
DenseVoxNet	1.7	96.54	0.32	14.
HighResNet	0.8	98.30	0.26	8.
Med3D (ResNet10)	17.3	98.61	0.08	1.
Med3D (ResNet18)	36.1	98.52	0.13	3.
Med3D (ResNet34)	66.5	98.34	0.34	7.
Med3D (ResNet50)	52.3	98.50	0.15	4.
Med3D (ResNet101)	91.3	98.12	0.27	9.
Med3D (ResNet152)	123.5	98.12	0.30	10.
Med3D (ResNet200)	132.7	98.10	0.10	11.
Residual 3D U-Net	141.2	98.43	0.34	6.
3D SkipDenseSeg	7.1	98.57	0.09	2.
3D U-Net	16.3	98.48	0.20	5.
V-Net	45.6	97.79	0.62	12.
LV-Net	12.2	97.56	0.60	13.

Table 3: Parameter count of all networks, mean testing *wDICE* coefficient of three training runs and standard deviation (σ) of the Head and Neck Cancer dataset.

In terms of the weighted performance, the Med 3D (ResNet 10) achieves the highest *wDICE*, closely followed by the SkipDenseSeg on this multiclass problem. Both networks have a small standard deviation of 0.08% and 0.09%, respectively, which covers the average difference of these two networks ($98.61\% - 98.57\% = 0.04\%$). Therefore, the assumption that the Med 3D (ResNet 10) performs best is not significant. For the Med3D backbones, it can be seen that fewer layers in the backbone are more powerful as fewer parameters need to be adjusted which is crucial for small datasets. The DenseVoxNet performs the worst of all the networks (96.54%).

Since this is a multiclass dataset, we also present the class-wise *DICE* in table 4 with the *mDICE* and their standard deviations in table 5 as the overall performance does not provide detailed information about the results. Compared to the other CNNs, the SkipDenseSeg was shown to perform best with an average of 1.39% (*mDICE*), although the achieved *DICE* scores are rather poor, which is reflected in figure 8. However, this is most likely due to the fact that the dataset is rather small, which is also supported by the comparison of the standard deviations per class. The more voxels a class contains, the lower the standard deviations become. In the following, the number of voxels associated with each class is shown, presenting the bias towards the background class which on its own consists of 98.71% of all voxels. Background: 962 594 506, C1: 590 569, C2: 589 474, C3: 1 121 325, C4: 1 140 637, C5: 2 207 852, C6: 2 379 147, C7: 2 226 010, C8: 2 326 160. The background class, has a low σ while all other classes have high standard deviations.

4.2 Carbon Rovings

The Carbon Rovings dataset is the second largest (table 1) and aims at segmenting quite large structures. The experiments have shown that 3D U-Net performs best on such elements (98.56%, table 6, figure 9). The gap to the second place (DenseVoxNet) is 0.39%, which is outside the standard deviation range of the two networks (DenseVoxNet: 0.32%; 3D U-Net:

Mean *DICE* (%) per class (Head and Neck Cancer Dataset)

Network (Backbone)	**B**	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	***mDICE***
DenseVoxNet	97.37	33.80	28.70	20.83	14.37	33.43	26.77	45.77	47.43	38.72
HighResNet	98.93	45.37	37.70	35.70	34.27	49.93	44.83	59.53	59.60	51.76
Med3D (ResNet10)	**99.23**	48.30	43.20	**43.63**	**37.97**	53.53	**48.67**	59.57	58.07	54.69
Med3D (ResNet18)	**99.23**	44.07	38.13	36.43	35.83	48.63	44.27	47.27	47.87	49.08
Med3D (ResNet34)	99.13	34.57	31.47	32.20	27.90	42.70	37.00	43.13	41.40	43.28
Med3D (ResNet50)	99.20	36.40	39.33	36.77	29.30	46.80	45.30	48.57	56.03	48.63
Med3D (ResNet101)	99.00	26.23	28.53	26.57	19.43	31.60	30.00	38.10	45.03	38.28
Med3D (ResNet152)	99.07	20.13	27.37	27.40	23.83	34.87	28.93	24.03	19.77	33.93
Med3D (ResNet200)	98.97	18.13	13.53	26.47	23.47	34.00	31.60	41.23	31.93	35.48
Residual 3D U-Net	99.13	48.23	**52.33**	33.13	27.97	44.33	38.87	58.63	59.47	51.34
3D SkipDenseSeg	99.17	**49.20**	49.37	41.17	35.63	**55.67**	46.97	63.93	63.57	**56.08**
3D U-Net	99.07	46.97	46.83	35.00	33.63	47.53	48.50	**65.97**	**69.83**	54.81
V-Net	98.67	25.33	16.73	19.37	25.47	32.40	33.53	37.00	42.33	36.76
LV-Net	98.40	12.77	7.30	32.33	23.53	31.77	32.80	44.90	41.77	36.17

Table 4: Mean testing *DICE* coefficient per class on the Head and Neck Cancer Dataset. Labels: background (B), submandibular glands (left (C1) and right (C2)), level 2 (left (C3) and right (C4)) and level 3 (left (C5) and right (C6)) lymph nodes and parotid glands (left (C7) and right (C8)).

σ (%) per class (Head and Neck Cancer Dataset)

Network (Backbone)	**B**	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**
DenseVoxNet	0.32	10.43	8.89	1.62	**0.81**	3.05	2.68	**0.51**	8.21
HighResNet	0.21	**2.12**	**2.11**	6.38	3.71	3.68	3.97	5.78	3.54
Med3D (ResNet10)	**0.06**	6.50	5.10	**0.42**	2.35	1.01	**1.59**	2.85	6.16
Med3D (ResNet18)	**0.06**	4.79	6.43	4.05	2.71	**0.61**	7.96	4.29	17.17
Med3D (ResNet34)	0.31	9.95	11.92	4.73	4.41	8.25	8.49	5.95	11.69
Med3D (ResNet50)	0.10	9.77	12.55	4.91	2.27	2.19	5.12	13.50	4.68
Med3D (ResNet101)	0.30	6.95	11.14	8.86	1.72	1.57	1.81	18.47	8.10
Med3D (ResNet152)	0.32	9.98	16.46	3.64	5.71	5.23	15.90	13.32	24.72
Med3D (ResNet200)	0.12	16.06	4.53	3.82	3.27	4.77	4.64	3.24	13.16
Residual 3D U-Net	0.29	10.37	6.76	5.28	6.25	9.02	6.04	8.60	7.78
3D SkipDenseSeg	**0.06**	5.65	2.70	3.14	3.22	2.36	3.20	4.25	2.28
3D U-Net	0.21	10.46	11.11	8.15	7.14	7.12	3.83	1.76	**1.77**
V-Net	0.47	13.95	21.46	7.62	8.06	17.08	13.79	31.60	9.58
LV-Net	0.46	6.50	5.84	14.12	11.94	15.63	19.21	16.30	3.47

Table 5: Standard deviation of the *DICE* coefficient per class on the Head and Neck Cancer Dataset. Labels: background (B), submandibular glands (left (C1) and right (C2)), level 2 (left (C3) and right (C4)) and level 3 (left (C5) and right (C6)) lymph nodes and parotid glands (left (C7) and right (C8)).

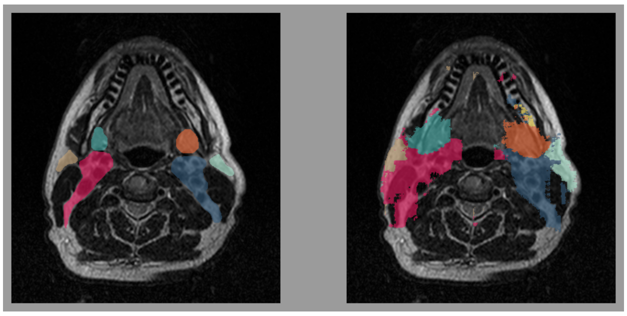

Figure 8: Visualization of ground truth (left) and segmentation (right) of the Head and Neck Cancer dataset using the 3D SkipDenseSeg. Orange: left submandibular gland; teal: right submandibular gland; yellow: left level 2 lymph node (not present in GT); gray: right level 2 lymph node (not present in GT and segmentation); blue: level 3 lymph node left; magenta: level 3 lymph node right; mint: left parotid gland; beige: right parotid gland.

0.20%). However, since this is a different dataset, the results of section 4.1 are not appropriate. Furthermore, the dataset is quite large, and therefore it is to be expected that the standard deviation will be lower. For this reason, we assume that the 3D U-Net performs best in binary segmentation with large structures and sufficient amount of training data. Although the 3D SkipDenseSeg performs best on the head and neck cancer dataset (*mDICE*) and achieves the lowest loss on this dataset, the gap in *DICE* with respect to the test data is 2.3% to the 3D U-Net, resulting in the second worst performance on this dataset.

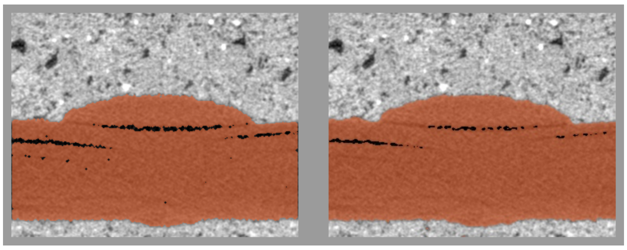

Figure 9: Visualization of ground truth (left) and segmentation (right) of the Carbon Rovings dataset using the 3D U-Net.

4.3 Concrete Pores

The Concrete Pores dataset consist of many tiny to large pores. Again, the 3D U-Net performs best, although not by a significant margin (gap to second: 0.11%, *DICE*: 87.66% (table 6, figure 10). On the second, the 3D SkipDenseSeg achieves a *DICE* of 87.55%, followed by the LV-Net, which achieves 86.95%. Although this dataset is similar in size to the Rovings dataset, all networks performed significantly worse. This is likely due to the fact that the pore structure is much more diverse in shape and size compared to a roving.

4.4 Polyethylene Fibers

For the Polyethylene Fibers dataset, the challenge is to predict tiny structures with limited data, which is a common problem for many researchers facing new domains. To overcome the problem of limited data, we used only geometrically augmented volumes to train and validate the networks. Testing was done

DICE (and *wDICE*) **(%)**							
Network (Backbone)	**Head and Neck Cancer**	**Carbon Rovings**	**Concrete Pores**	**PE Fibers**	**Brain Mitochondria**	**BraTS**	**avg Rank**
DenseVoxNet	96.54	98.17	85.77	39.70	67.45	97.56	7.
HighResNet	98.30	96.09	84.85	46.38	67.61	98.24	6.
Med3D (ResNet10)	**98.61**	97.87	79.64	36.26	63.65	96.49	9.
Med3D (ResNet18)	98.52	97.36	80.41	37.83	63.47	96.55	13.
Med3D (ResNet34)	98.34	97.52	83.73	36.93	61.03	97.80	12.
Med3D (ResNet50)	98.50	97.37	82.93	40.29	70.97	97.88	4.
Med3D (ResNet101)	98.12	96.36	85.71	37.24	71.60	97.99	5.
Med3D (ResNet152)	98.12	97.39	75.61	41.15	65.94	97.58	10.
Med3D (ResNet200)	98.10	97.41	80.24	38.56	63.48	97.73	11.
Residual 3D U-Net	98.43	97.57	83.44	**63.49**	73.66	**98.59**	2.
3D SkipDenseSeg	98.57	96.28	87.55	52.98	69.78	97.98	3.
3D U-Net	98.48	**98.56**	**87.66**	58.45	76.38	98.08	1.
V-Net	97.79	97.05	72.10	16.21	63.74	96.47	14.
LV-Net	97.56	97.40	86.95	16.16	**78.89**	97.05	8.

Table 6: Testing *DICE* (and *wDICE* for multi-class datasets) coefficient regarding all networks and datasets. On the right hand side, the average rank is given, representing a mean ranking of the networks regarding all datasets.

Validation Loss						
Network (Backbone)	**Head and Neck Cancer (avg)**	**Carbon Rovings**	**Concrete Pores**	**PE Fibers**	**Brain Mitochondria**	**BraTS**
DenseVoxNet	0.284	0.046	0.114	0.094	0.252	0.599
HighResNet	0.327	0.051	0.178	0.089	0.298	0.670
Med3D (ResNet10)	0.209	0.050	0.289	0.141	0.336	0.773
Med3D (ResNet18)	0.232	0.062	0.268	0.152	0.370	0.738
Med3D (ResNet34)	0.248	0.060	0.253	0.149	0.396	0.748
Med3D (ResNet50)	0.239	0.059	0.197	0.114	0.266	0.701
Med3D (ResNet101)	0.314	0.053	0.223	0.127	0.279	0.690
Med3D (ResNet152)	0.293	0.061	0.235	0.134	0.318	0.690
Med3D (ResNet200)	0.291	0.067	0.173	0.125	0.297	0.651
Residual 3D U-Net	0.315	0.048	0.139	0.083	0.211	0.576
3D SkipDenseSeg	0.228	**0.037**	0.158	0.082	0.281	0.571
3D U-Net	**0.202**	0.041	**0.093**	**0.067**	**0.207**	**0.560**
V-Net	0.638	0.052	0.465	0.298	0.730	0.804
LV-Net	0.736	0.102	0.307	0.305	0.525	0.757

Table 7: Validation Loss regarding all networks and datasets.

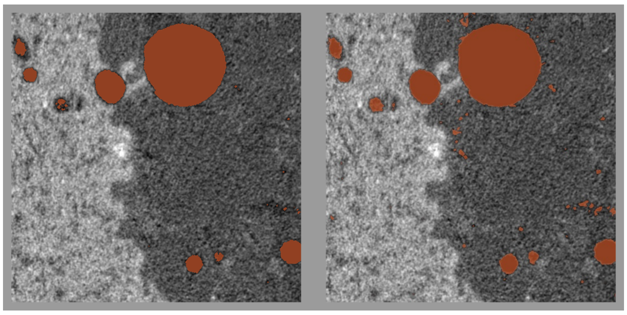

Figure 10: Visualization of ground truth (left) and segmentation (right) of the Concrete Pores dataset using the 3D U-Net.

on the unaugmented volumes, which explains the rather poor results of, for example, the V-Net and the LV-Net. We suspect a weakness in their architecture that makes them inefficient at learning generalized features on such data. The 3D-U-Net and its residual version are superior. In (Ronneberger et al., 2015), the U-Net was shown to perform well on small datasets, which has been proven in the third dimension as well. The Residual 3D U-Net significantly outperformed all other non U-Net architectures by a large margin (11.51%, table 6), and although it is the largest network in terms of parameters (141.2M), the inner architecture shows its efficiency. In total, only three networks were able to achieve a *DICE* of > 50% and can be considered useful for such problems: 1. Residual U-Net 3D, 2. 3D U-Net, 3.: 3D SkipDenseSeg. Even though the Residual 3D U-Net achieved a *DICE* of only 63.49% it reliably found all fibers in the test data (figure 11). The rather poor score results to the fact that fibers marked in the ground truth are very thin and maybe too conservative in size. The predicted structures are thicker and, on such a scale, this difference explains the achieved *DICE*.

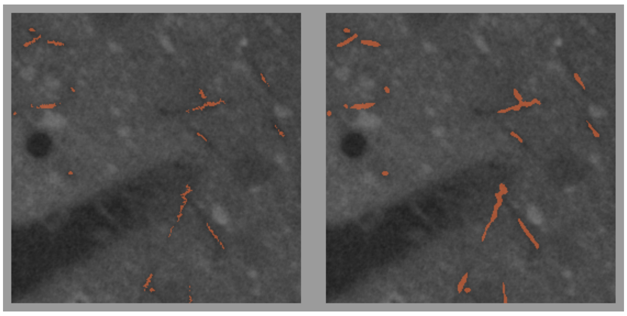

Figure 11: Visualization of ground truth (left) and segmentation (right) of the PE Fibers dataset using the Residual 3D U-Net.

4.5 Brain Mitochondria

The Brain Mitochondria dataset is quite small. Also, the training data is very similar to the test data, which is why the LV-Net, last place in the PE Fibers dataset test, performed best with a *DICE* of 78.89%. It seems that this architecture has a high potential to learn the features of the training dataset, but cannot adapt them to slightly different volumes. It requires a sufficient amount of training volumes to successfully learn and adapt features on unseen data. However, the loss is quite high with 0.525

(table 7). This happens when the network performs well in general, but produces a few large errors as shown in figure 12 (top right), where the network falsely predicts a large structure. This also supports the hypothesis that this architecture has a very narrow applicability. The second and third best models are the two U-Net variants, which again proves their usability on small datasets. However, the difference in *DICE* between the LV-Net and the 3D U-Net (second best) is still significant at 2.51% although the 3D U-Net was able to learn more features than the LV-Net when comparing the validation loss (table 7).

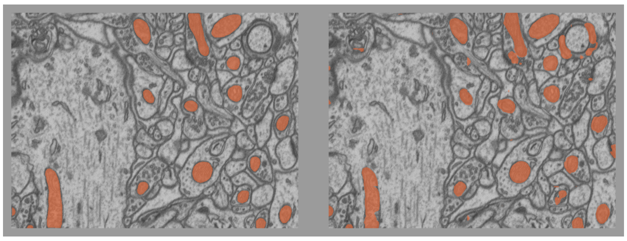

Figure 12: Visualization of ground truth (left) and segmentation (right) of the Brain Mitochondria dataset using the LV-Net.

4.6 BraTS

The large BraTS dataset contains complex structures to learn. However, all architectures perform quite well overall (*wDICE* from 96.47% to 98.59%, table 6), because the background class consists of the most voxels (98.7% background). Therefore, we also show the *DICE* per class as the overall performance is misleading, as already stated in section 4.1. The class distribution of voxels associated to each class is as follows: Background: 10 357 065 861, C1: 30 280 730, C2: 79 352 001, C3: 26 598 048. Table 8 indicates that the prediction of the single classes is a difficult task for all tested CNNs. The Residual 3D U-Net performs best (*mDICE*: 62.5%, visualization in figure 13), closely followed by the 3D SkipDenseSeg (*mDICE*: 59.5%). The worst network is the Med3D (ResNet10) with a *mDICE* of 42.6%.

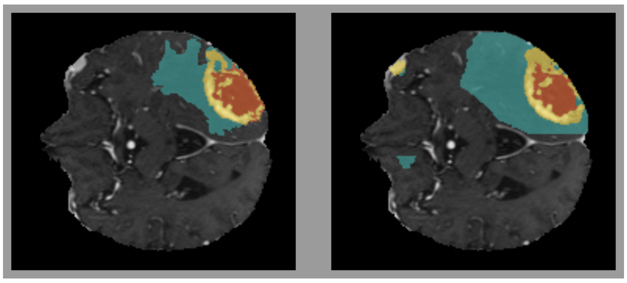

Figure 13: Visualization of ground truth (left) and segmentation (right) of the BraTS dataset using the Residual 3D U-Net. Orange: NCR/NET; teal: ED; yellow: ET

5. CONCLUSION

The experiments have emphasized the importance of the initial random state. The minimum and maximum standard deviations ranged from 0.08% (Med3D (ResNet10)) to 0.62% (V-Net) on the Head and Neck Cancer dataset. In addition to choosing the correct hyperparameters, this can have a significant impact on the performance, and a final network should be trained several times to achieve the best results. In table 6, the rankings of all networks have been averaged across all datasets. Overall, the standard 3D U-Net performed best. The next best CNNs are the Residual 3D U-Net and the 3D SkipDenseSeg, although it never performed best on any domain. The comparisons have shown that the domain has an impact, albeit less than expected, and that the U-Net variants are fairly general architectures applicable to any dataset. However, although the 3D U-Net or its residual version can be a good starting point, in some cases, other networks are superior. From our results, we propose the following recommendations for selecting a 3D neural network based on dataset attributes:

***DICE* (%) per class** (BraTS)					
Network (Backbone)	**B**	**NCR/NET**	**ED**	**ET**	***mDICE***
DenseVoxNet	98.2	45.4	33.2	43.2	55.0
HighResNet	98.8	48.2	38.9	**51.1**	59.2
Med3D (ResNet10)	97.3	24.6	21.4	27.0	42.6
Med3D (ResNet18)	97.3	36.9	21.0	26.8	45.5
Med3D (ResNet34)	98.6	35.0	27.7	30.9	48.0
Med3D (ResNet50)	98.6	46.4	30.4	33.7	52.3
Med3D (ResNet101)	98.7	36.2	32.0	37.1	51.0
Med3D (ResNet152)	98.3	39.1	26.1	38.1	50.4
Med3D (ResNet200)	98.5	38.6	28.4	28.2	48.4
Residual 3D U-Net	**99.2**	**60.2**	**43.4**	47.0	**62.5**
3D SkipDenseSeg	98.6	53.9	34.7	50.9	59.5
3D U-Net	98.8	39.8	33.0	45.3	54.2
V-Net	97.2	29.8	23.1	36.3	46.6
LV-Net	97.8	43.8	25.1	45.4	53.0

Table 8: Testing *DICE* coefficient per class on the BraTS dataset. Labels: background (B), necrotic and non-enhancing tumor core (NCR/NET), peritumoral edema (ED) and GD-enhancing tumor (ET).

- Larger datasets with large, coherent structures: 3D U-Net or DenseVoxNet
- Larger datasets with tiny to large structures: 3D U-Net or 3D SkipDenseSeg
- Larger datasets with complex structures: Residual 3D U-Net, HighResNet or 3D U-Net
- Medium datasets with large and very similar structures: 3D U-Net
- Small datasets with simpler, intergrown structures: a Med3D version or 3D SkipDenseSeg
- Very small datasets with very thin structures: Residual 3D U-Net or 3D U-Net

Although the LV-Net performs best on the Brain Mitochondria dataset, the experiments suggest that it has a weakness in its architecture that reduces its generalizability. Therefore, we cannot confidently recommend its use on such medium-sized datasets and have decided to recommend the 3D U-Net instead. The experiments also showed that neither the V-Net nor most of the Med3D versions (ResNet: 34, 50, 101, and 200) could achieve a top 3 result in any domain.

6. ACKNOWLEDGEMENTS

The research work presented in the publication has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB/TRR 280; Projekt-ID: 417002380
The authors are grateful to the Center for Information Services and High Performance Computing [Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH)] at TU Dresden for providing its facilities for high throughput calculations.

REFERENCES

Badrinarayanan, V., Handa, A., Cipolla, R., 2015. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Robust Semantic Pixel-Wise Labelling. *CoRR*, abs/1505.07293. https://doi.org/10.48550/arXiv.1505.07293.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., Freymann, J. B., Farahani, K., Davatzikos, C., 2017. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data*, 4, 170117. https://doi.org/10.1038/sdata.2017.117.

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R. T., Berger, C., Ha, S. M., Rozycki, M., Prastawa, M., Alberts, E., Lipková, J., Freymann, J. B., Kirby, J. S., Bilello, M., Fathallah-Shaykh, H. M., Wiest, R., Kirschke, J., Wiestler, B., Colen, R. R., Kotrotsou, A., LaMontagne, P., Marcus, D. S., Milchenko, M., Nazeri, A., Weber, M., Mahajan, A., Baid, U., Kwon, D., Agarwal, M., Alam, M., Albiol, A., Albiol, A., Varghese, A., Tuan, T. A., Arbel, T., Avery, A., B., P., Banerjee, S., Batchelder, T., Batmanghelich, K. N., Battistella, E., Bendszus, M., Benson, E., Bernal, J., Biros, G., Cabezas, M., Chandra, S., Chang, Y., et al., 2018. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. *CoRR*, abs/1811.02629. https://doi.org/10.48550/arXiv.1811.02629.

Beckmann, B., Bielak, J., Bosbach, S., Scheerer, S., Schmidt, C., Hegger, J., Curbach, M., 2021. Collaborative research on carbon reinforced concrete structures in the CRC/TRR 280 project. *Civil Engineering Design*, 3(3), 99-109. https://doi.org/10.1002/cend.202100017.

Bui, T. D., Shin, J., Moon, T., 2019. Skip-connected 3D DenseNet for volumetric infant brain MRI segmentation. *Biomedical Signal Processing and Control*, 54, 101613. https://doi.org/10.1016/j.bspc.2019.101613.

Cardenas, C. E., Mohamed, A. S. R., Yang, J., Gooding, M., Veeraraghavan, H., Kalpathy-Cramer, J., Ng, S. P., Ding, Y., Wang, J., Lai, S. Y., Fuller, C. D., Sharp, G., 2020. Head and neck cancer patient images for determining auto-segmentation accuracy in T2-weighted magnetic resonance imaging through expert manual segmentations. *Medical Physics*, 47(5), 2317-2322. https://doi.org/10.1002/mp.13942.

Cardenas, C., Mohamed, A., Sharp, G., Gooding, M.and Veeraraghavan, H., Yang, J., 2019. Data from AAPM RT-MAC Grand Challenge 2019. The Cancer Imaging Archive. https://doi.org/10.7937/tcia.2019.bcfjqfqb.

Cardoso, M. J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., Nath, V., He, Y., Xu, Z., Hatamizadeh, A., Myronenko, A., Zhu, W., Liu, Y., Zheng, M., Tang, Y., Yang, I., Zephyr, M., Hashemian, B., Alle, S., Darestani, M. Z., Budd, C., Modat, M., Vercauteren, T., Wang, G., Li, Y., Hu, Y., Fu, Y., Gorman, B., Johnson, H., Genereaux, B., Erdal, B. S., Gupta, V., Diaz-Pinto, A., Dourson, A., Maier-Hein, L., Jaeger, P. F., Baumgartner, M., Kalpathy-Cramer, J., Flores, M., Kirby, J., Cooper, L. A. D., Roth, H. R., Xu, D., Bericat, D., Floca, R., Zhou, S. K., Shuaib, H., Farahani, K., Maier-Hein, K. H., Aylward, S., Dogra, P., Ourselin, S., Feng, A., 2022. MONAI: An open-source framework for deep learning in healthcare. https://doi.org/10.48550/arXiv.2211.02701.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4), 834-848. https://doi.org/10.1109/TPAMI.2017.2699184.

Chen, S., Ma, K., Zheng, Y., 2019. Med3D: Transfer Learning for 3D Medical Image Analysis. *CoRR*, abs/1904.00625. https://doi.org/10.48550/arXiv.1904.00625.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., Ronneberger, O., 2016. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. *CoRR*, abs/1606.06650. https://doi.org/10.48550/arXiv.1606.06650.

Ciresan, D., Giusti, A., Gambardella, L., Schmidhuber, J., 2012. Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images. *Advances in Neural Information Processing Systems*, 25.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., Prior, F., 2013. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *Journal of Digital Imaging*, 26(6), 1045–1057. https://doi.org/10.1007/s10278-013-9622-7.

Kingma, D. P., Ba, J., 2015. Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. https://doi.org/10.48550/arXiv.1412.6980.

Lee, K., Zung, J., Li, P., Jain, V., Seung, H. S., 2017. Superhuman Accuracy on the SNEMI3D Connectomics Challenge. *CoRR*, abs/1706.00120. https://doi.org/10.48550/arXiv.1706.00120.

Lei, T., Zhou, W., Zhang, Y., Wang, R., Meng, H., Nandi, A. K., 2020. Lightweight V-Net for Liver Segmentation. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1379-1383. https://doi.org/10.1109/ICASSP40776.2020.9053454.

Li, W., Wang, G., Fidon, L., Ourselin, S., Cardoso, M. J., Vercauteren, T., 2017. On the Compactness, Efficiency, and Representation of 3D Convolutional Networks: Brain Parcellation as a Pretext Task. *Information Processing in Medical Imaging*, 348–360. https://doi.org/10.1007/978-3-319-59050-9$_2$8.

Lorenzoni, R., Curosu, I., Paciornik, S., Mechtcherine, V., Oppermann, M., Silva, F., 2020. Semantic segmentation of the micro-structure of strain-hardening cement-based composites (SHCC) by applying deep learning on micro-computed tomography scans. *Cement and Concrete Composites*, 108, 103551. https://doi.org/10.1016/j.cemconcomp.2020.103551.

Lucchi, A., Li, Y., Fua, P., 2013. Learning for Structured Prediction Using Approximate Subgradient Descent with Working Sets. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 1987-1994. https://doi.org/10.1109/CVPR.2013.259.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.-A., Arbel, T., Avants, B. B., Ayache, N., Buendia, P., Collins, D. L., Cordier, N., Corso,

J. J., Criminisi, A., Das, T., Delingette, H., Demiralp, Ç., Durst, C. R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K. M., Jena, R., John, N. M., Konukoglu, E., Lashkari, D., Mariz, J. A., Meier, R., Pereira, S., Precup, D., Price, S. J., Raviv, T. R., Reza, S. M. S., Ryan, M., Sarikaya, D., Schwartz, L., Shin, H.-C., Shotton, J., Silva, C. A., Sousa, N., Subbanna, N. K., Szekely, G., Taylor, T. J., Thomas, O. M., Tustison, N. J., Unal, G., Vasseur, F., Wintermark, M., Ye, D. H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., Van Leemput, K., 2015. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans. Med. Imaging*, 34(10), 1993–2024. https://doi.org/10.1109/TMI.2014.2377694.

Mester, L., Wagner, F., Liebold, F., Klarmann, S., Maas, H.-G., Klinkel, S., 2022. Image-based modelling and analysis of carbon-fibre reinforced concrete shell structures. S. Stokkeland (ed.), *Concrete innovation for sustainability: : held in Oslo, Norway, June 12-16, 2022*, fib proceedings, fib, 1631–1640.

Milletari, F., Navab, N., Ahmadi, S., 2016. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *CoRR*, abs/1606.04797. https://doi.org/10.48550/arXiv.1606.04797.

Niyas, S., Pawan, S., Anand Kumar, M., Rajan, J., 2022. Medical image segmentation with 3D convolutional neural networks: A survey. *Neurocomputing*, 493, 397-413. https://doi.org/10.1016/j.neucom.2022.04.065.

Object Research Systems (ORS), 2021. Dragonfly 2021.1. http://www.theobjects.com/dragonfly. Computer Software.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems 32*, 8024–8035. https://doi.org/10.48550/arXiv.1912.01703.

Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J., 2017. Large Kernel Matters - Improve Semantic Segmentation by Global Convolutional Network. *CoRR*, abs/1703.02719. https://doi.org/10.48550/arXiv.1703.02719.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.

Singh, S. P., Wang, L., Gupta, S., Goli, H., Padmanabhan, P., Gulyás, B., 2020. 3D Deep Learning on Medical Images: A Review. https://doi.org/10.48550/ARXIV.2004.00218.

Taha, A. A., Hanbury, A., 2015. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, 15(1), 29. https://doi.org/10.1186/s12880-015-0068-x.

Wagner, F., 2023a. Carbon Rovings Segmentation Dataset. https://doi.org/10.34740/KAGGLE/DS/2920892.

Wagner, F., 2023b. Concrete Pores Segmentation Dataset. https://doi.org/10.34740/KAGGLE/DS/2921245.

Wagner, F., 2023c. Fiber Segmentation Dataset. https://doi.org/10.34740/KAGGLE/DS/2894881.

Wagner, F., Eltner, A., Maas, H.-G., 2023. River water segmentation in surveillance camera images: A comparative study of offline and online augmentation using 32 CNNs. *International Journal of Applied Earth Observation and Geoinformation*, 119, 103305. https://doi.org/10.1016/j.jag.2023.103305.

Wu, Y., He, K., 2018. Group Normalization. *CoRR*, abs/1803.08494. https://doi.org/10.48550/arXiv.1803.08494.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J., 2018. Unified Perceptual Parsing for Scene Understanding. *CoRR*, abs/1807.10221. https://doi.org/10.48550/arXiv.1807.10221.

Yu, L., Cheng, J., Dou, Q., Yang, X., Chen, H., Qin, J., Heng, P., 2017. Automatic 3D Cardiovascular MR Segmentation with Densely-Connected Volumetric ConvNets. *CoRR*, abs/1708.00573. https://doi.org/10.48550/arXiv.1708.00573.