3D U-Net For Segmentation of Covid-19-Associated Pulmonary Infiltrates Using Transfer Learning: State-Of-The-Art Results on Affordable Hardware

Keno K. Bressem
Department of Radiology Charité Universitätsmedizin Berlin
Stefan M. Niehues
Department of Radiology Charité Universitätsmedizin Berlin
Günther Engel
Department of Radiology Charité Universitätsmedizin Berlin
Bernd Hamm
Department of Radiology Charité Universitätsmedizin Berlin
Marcus R. Makowski
Technical University of Munich
Janis L. Vahldiek
Department of Radiology Charité Universitätsmedizin Berlin
Lisa C. Adams (lisa.adams@charite.de)
Department of Radiology Charité Universitätsmedizin Berlin

Research Article

Keywords: COVID-19, U-Net, Segmentation, Computed Tomography

Posted Date: March 11th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-259319/v1

Abstract

Segmentation of pulmonary infiltrates can help assess severity of COVID-19, but manual segmentation is labor and time-intensive. Using neural networks to segment pulmonary infiltrates would enable automation of this task. However, training a 3D U-Net from computed tomography (CT) data is time- and resource-intensive. In this work, we therefore developed and tested a solution on how transfer learning can be used to train state-of-the-art segmentation models on limited hardware and in shorter time. We use the recently published RSNA International COVID-19 Open Radiology Database (RICORD) to train a fully three-dimensional U-Net architecture using an 18-layer 3D ResNet, pretrained on the Kinetics-400 dataset as encoder. The generalization of the model was then tested on two openly available datasets of patients with COVID-19, who received chest CTs (Corona Cases and MosMed datasets). Our model performed comparable to previously published 3D U-Net architectures, achieving a mean Dice score of 0.679 on the tuning dataset, 0.648 on the Coronacases dataset and 0.405 on the MosMed dataset. Notably, these results were achieved with shorter training time on a single GPU with less memory available than the GPUs used in previous studies.

1 Introduction

The Coronavirus Disease-2019 (COVID-19) is an infectious disease of the respiratory tract and lungs, with more than 110 million confirmed cases worldwide and nearly 2.5 million deaths as of February 2021.[1,2] For the management of COVID-19, rapid diagnosis is critical to quickly isolate affected patients and prevent further spread of the disease.[3] Presently, the diagnostic standard for COVID-19 is real-time reverse transcription polymerase chain reaction (RT-PCR) from pharyngeal or deep nasal swaps.[4] However, in the clinical setting, computed tomography (CT) is increasingly used in patients with suspected COVID-19. The role of CT to diagnose COVID-19 has been critically debated, and currently there is consensus that CT should not be used in place of RT-PCR.[5] Nevertheless, CT remains an important tool to assess pulmonary infiltrates associated with COVID-19 and to estimate the severity of the disease.[6] On CT imaging, COVID-19 typically shows multifocal ground glass opacities as well as consolidations in predominantly peripheral and basal distribution.[7] Although the relationship is not strictly linear, a larger affected lung area is associated with more severe disease. Therefore, knowing how much of the lung is affected by COVID-19 may allow for a more accurate assessment of disease severity.

Manual segmentation of the affected lung area is a tedious task. In their recent work, *Ma et al.* manually segmented 20 openly available CT scans of patients affected by COVID-19 an reported a mean duration of 400 minutes per CT volume.[8] Clearly, this amount of time is too high to be implemented in routine clinical practice, and research is being conducted on methods to automate these tasks. One of the most promising techniques for automatic segmentation is deep neural networks, in particular the U-Net architecture.[9]

U-Nets consist of a down-sampling block that extracts features from input images and an up-sampling part that generates segmentation masks from the previously extracted features. Spatial information decreases in the deeper layers of a convolutional neural network; therefore, the U-Net has skip connections that allow the up-sampling block to use both the feature information of the deeper layers as well as the spatial information from earlier layers to generate high-resolution segmentation masks.[9] An advantage of the U-Net architecture is the relatively small amount of data required to obtain accurate results, which is especially important in medical imaging where data are usually sparse.[9,10] However, a drawback is the higher memory requirements of the U-Net, since multiple copies of feature maps must be kept in memory to enable the skip connections, so that training a U-Net either requires access to multiple graphics processing units (GPUs) to perform distributed training with a larger batch size, or the batch size must be greatly reduced. This is even more important when U-Nets are extended to three-dimensional space, since each item in a batch of 3D data is even larger. Another method to increase the accuracy of a model on limited data is to use transfer learning, where a model architecture is first trained on another task, and then fine-tuned on a novel task. [11]

In this work, we developed and evaluated an approach to effectively train a fully three-dimensional U-Net in a single GPU achieving state-of-the-art accuracy by using transfer learning.

2 Results

The 3D U-Net was trained on the RICORD data (n = 117 CT volumes) which was randomly split into a training dataset consisting of 100 volumes (85%) and a tuning dataset including 17 volumes (15%). The total training duration was 10 hours and 49 minutes with an average duration of 45 seconds per epoch for the lower input resolution and 2:30 minutes for the higher image resolution. While at the beginning of each training session the loss on the training data was higher than on the tuning data, the overall training loss showed a faster decline so that after 200 epochs it was slightly lower than the loss on the tuning data. After 200 epochs, however, we found no obvious signs of overfitting, as the average valid loss was still slowly decreasing.

The library we developed to achieve these results can be accessed at: *github.com/kbressem/faimed3d* with example code for the segmentation task at: https://github.com/kbressem/faimed3d/blob/main/examples/3d_segmentation.md.

2.1. Dice score

The Dice score was used to compare the original segmentation mask with the predicted mask. There are several implementations of the Dice score available that may affect the calculated score and thus limit comparability. We used the implementation by *Ma et al.*, for which the code is freely available.[8]

Because the lung areas affected by COVID-19 can differ substantially from case to case, we calculated the Dice score for each patient and then macro-averaged the scores. This resulted in slightly poorer

scores compared with micro-averaging across the entire data set but is more similar to clinical feasibility.

We obtained the highest scores on the tuning dataset with a mean Dice score of 0.679 and a standard deviation of 0.13. When applied to new datasets, the performance of the segmentation model decreased with a mean Dice score of 0.648 ±

0.132 for the Coronacases from the COVID-19 CT Lung and Infection Segmentation Dataset, and 0.405 ± 0.213 for the MosMed dataset. A summary of the Dice scores obtained on the datasets is shown in Table 1. Please refer to Fig. 1 for an overview of example images taken from the three datasets used in this study with segmentation masks from a human annotator (red) and the corresponding predicted masks from our model (green).

Table 1
Volumetric Dice scores

Dataset	CT scans (n)	Dice score Mean and std.	Dice score Lowest	Dice score Highest
RICORD	17	0.679 ± 0.130	0.398	0.846
Coronacases	10	0.648 ± 0.132	0.362	0.783
MosMedData	50	0.405 ± 0.213	0.008	0.675

Overview of the Dice scores obtained for the task of segmenting lung tissue affected by COVID-19 from healthy lung tissue. Abbreviation: Std = standard deviation.

2.2. Shape similarity

Because the normal Dice score is insensitive to shape, we also used the normalized surface Dice (NSD) to assess model performance based on shape similarity.[12] To ensure comparability of our results, we again used the implementation of the metric of Ma et al.[8] Again, the highest scores were achieved on the tuning dataset with a mean NSD of 0.781

± 0.124. On MosMed, the NSD was lowest with a score of 0.597 ± 0.270. On the ten images of the Coronacases dataset, the model achieved an NSD of 0.716 ± 0.135. A summary of the NSD can be found in Table 2.

Example images of the segmentation maps generated by the model compared to the ground truth are shown in Fig. 1. Table 3. provides an overview of the results we obtained and those reported in the published literature.

Table 2
Normalized surface Dice scores

Dataset	CT scans (n)	NSD Mean and std.	NSD Lowest	NSD Highest
RICORD	17	0.781 ± 0.124	0.480	0.911
Coronacases	10	0.716 ± 0.135	0.457	0.862
MosMedData	50	0.597 ± 0.270	0.060	0.926

Overview of the achieved normalized surface Dice scores (NSD) as a measure- ment of shape similarity between two regions. Abbreviation: Std = standard deviation.

Table 3
Overview of the results from previous studies.

Publication	Dataset	Dice score Tuning data	Dice score Hold-out data	Training time	Hardware
Our approach	RICORD	0.69810h, 49min1	GeForce RTX 2080ti		
	Coronacases		0.623		(11GB VRAM)
	MosMedData		0.403		
Müller et al.[10]*	RICORD	0.761	-	130h	1 Nvidia Quadro P6000
					(24 GB VRAM)
Yan et al.[13]	proprietary	-	0.726	-6 Nvidia TITAN RTX	
				(24 GB VRAM)	
Ma et al.[8]**	Coronacases	0.642	-	-	
	MosMedData		0.443		
Pu et al.[14]***	proprietary	-	0.81	-	
*Müller et al. report the accuracy for 5-fold cross-validation; we report the mean of 5 folds.					
**Ma et al. defined different tasks for segmentation, of which we report the accuracy of subtasks 3, as it is the most similar to our methods and thus most comparable.					
***Pu et al. report the Dice score only for lung areas > 200mm^3 and rated each infiltration separately.					

3 Discussion

In the present study, we propose a transfer learning approach using a 3D U-Net for segmenting pulmonary infiltrates associated with COVID-19 implemented on a single GPU with 11 GB VRAM. We used a transfer learning approach with an 18-layer 3D ResNet pretrained on a video classification dataset serving as encoder for the 3D U-Net, and obtained state-of-the-art results within comparably short training times using an in-house-developed library (*github.com/kbressem/faimed3d*).

There have been previous efforts to automatically segment pulmonary infiltrates using U-Nets, but few used fully three-dimensional models, while most studies applied a layer-by-layer approach. In our opinion, the metrics obtained from these two approaches are not comparable because the slice-wise approach may introduce selection bias into the data by excluding slices that do not show lung or infiltrates. For 3D models, the input volume shows the entire lung, including healthy and diseased lung tissue as well as portions of the neck and abdomen that do not contain lung tissue. *Müller et al.* proposed a 3D U-Net with an architecture similar to our model.[10] Because of limited training data, they used 5-fold cross-validation during training and reported a mean Dice score of 0.761 on the 5 validation folds. The model of *Müller et al.* was trained for 130h (more than 10 times longer than the model presented in this work) on a GPU with twice as much VRAM (Nvidia Quadro P6000). However, since the models were evaluated on a proprietary dataset, the obtained Dice scores cannot be compared without reservations, as differences in segmentation ground-truth may exist.

Lessmann et al. developed CORADS-AI, a deep learning algorithm for predicting the CO-RADS grade on non- contrast CT images.[15] CO-RADS (COVID-19 Reporting and Data System) is a categorical score between 1–5 that indicates the likelihood of pulmonary involvement, with a CO-RADS score of 1 corresponding to a very low probability of pulmonary involvement and a score of 5 representing a very high probability.[16] Interestingly, the interrater agreement on CO-RADS is only moderate, with a Fleiss kappa value of 0.47. CO-RADS grading differs from manual segmentation of pulmonary infiltrates in patients with proven COVID-19 and the kappa values are therefore not transferable. Nevertheless, the question is whether there is also a significant interrater difference in segmentation and how this would affect model performance and comparability between studies. For the RICORD dataset and the dataset provided my *Ma et al.*, each CT volume was annotated by multiple experts, including at least one board-certified radiologist, to reduce bias coming from poor interrater agreement. However, for the MosMed dataset the number of annotators per CT volume is not available.

Ma et al. also developed a data-efficient 3D U-Net model that achieved a mean Dice score of 0.642 in the 5-fold cross validation and a Dice score of 0.443 during interference on the MosMed dataset.[8]

The highest Dice score achieved with a 3D U-Net architecture was published by *Pu et al.* with a value of 0.81 for infiltration greater than 200 mm^3 on a proprietary dataset [21]. It is important to note, however, that the measurement of *Pu et al.* differs from other published results as well as from ours because the Dice score is calculated at a per-lesion level and then averaged, rather than at a per-patient level.

Yan et al. proposed a novel adaption of the U-Net architecture to increase segmentation performance for COVID-19 [20]. Their COVID-SegNet achieved a Dice score of 0.726 on the independent hold-out dataset. To achieve this, they used a proprietary dataset of 861 patients (8 times larger than the RICORD dataset and 40 times larger than the data from *Ma et al.*) and trained their model on six Nvidia Titan RTXs with 24 GB VRAM each.

By comparison, the model developed in this study achieved a higher Dice score than *Ma et al.* and had substantially shorter training times and lower hardware requirements than previously published studies. However, this comparison should be taken with caution because the datasets, training methods and calculation of metrics differed. Nonetheless, this study demonstrates the added benefit of using a pre-trained encoder for 3D U-Nets, as one can quickly achieve state-of-the-art results with lower hardware requirements and shorter training times. Transfer learning may help to provide better access and use of 3D segmentation models for the diagnostic community and for researches without access to high performance computing clusters.

4 Materials And Methods

4.1. Datasets and Annotations

Three openly available datasets of CT scans from patients affected by COVID-19 are used in this work. These include the following:

- RSNA International COVID-19 Open Radiology Database (RICORD)[17]
- MosMedData[18]
- COVID-19 CT Lung and Infection Segmentation Dataset[8]

RICORD is a multi-institutional and multi-national, expert annotated dataset of chest CT and radiographs. It consists of three different collections:

- Collection 1a includes 120 CT studies from 110 patients with COVID-19, in which the affected lung areas were segmented pixel by pixel.
- Collection 1b contains 120 studies of 117 patients without evidence of COVID-19
- Collection 1c contains 1,000 radiographs from 361 patients with COVID-19

Only collection 1a was included in the present work.

The MosMedData contains data from a single institution. Overall, 1,110 studies are included in the dataset. Pixel-wise segmentation of COVID-19-associated pulmonary infiltrates is available for 50 studies in the MosMedData, which we used for our work.

The COVID-19 CT Lung and Infection Segmentation Dataset consists of ten CT volumes from the Coronacases Initiative and ten CT volumes extracted from Radiopaedia, for which the authors have

added a pixel-wise segmentation of infiltrates. Because the ten CT volumes extracted from Radiopaedia have already been windowed and converted to PNG (Portable Network Graphics) format, we included only the ten Coronacases Initiative volumes in this study.

4.2. Data Preparation

The RICORD data are provided as DICOM (Digital Imaging and Communications in Medicine) slices for the different CT images, and the annotations are available in JSON format. We used SimpleITK to read the DICOM slices, scale the

images according to the rescale intercept and rescale slope, and clip the pixel-values to the range of -2000 and + 500.[19] The annotations were converted from JSON (JavaScript Object Notation) to a pixel array and matched to the respective DICOM slice using the study- and SOP instance UID. Both the original volume and annotations were then stored in NIfTI (Neuroimaging Informatics Technology Initiative) format.

The MosMedData and COVID-19 CT Lung and Infection Segmentation Dataset were already available in NIfTI format, so no further preprocessing was performed.

4.3. Model Architecture

The 3D U-Net architecture was implemented using PyTorch (version 1.7.0)[20] and fastai (version 2.1.10).[21] We used a fully three-dimensional U-Net architecture for CT volume segmentation. The encoder part consisted of an 18-layer 3D ResNet, as described by *Tran et al.*, pretrained on the Kinetics-400 dataset.[22] We removed the fully connected layers from the 3D ResNet and added an additional 3D convolutional layer and four upscaling blocks. Each upscaling block consisted of one transposed convolutional layer and two normal convolutional layers. Each convolutional layer was followed by a rectified linear unit (ReLU) as activation function. Instance normalization was applied to the lower layer features before the double convolution was performed. The final block of the U-Net consisted of a single residual block without dilation and a single convolutional layer with a kernel size and stride of one for pooling of the feature maps. The model architecture is visualized in Fig. 2.

4.4. Model Training

We randomly split the RICORD dataset into a training (85%) and a tuning (15%) dataset and used both the MosMedData and COVID-19 CT lung and infection segmentation datasets as hold-out datasets to only evaluate the trained model. A progressive resizing approach was used in which we first trained the U-Net on volumes consisting of 18 slices with a resolution of 112 x 112 px per slice, allowing to use a batch size of 6. In a second training session, we increased the resolution to 256 x 256 px for 20 slices and used a batch-size of 1. During training, we used various augmentations, including perspective distortion, rotation, mirroring, adjusting contrast and brightness, and adding random Gaussian noise to the volumes. For the loss function, we used a combination of the dice loss (as described by *Milletari et al.*[23]) and pixel-wise cross-entropy loss.

Regarding the learning rate, we used the cyclic learning rate approach described by Leslie Smith, as implemented in fastai.[24] Here, one specifies a base learning rate at the beginning of the training, which is then varied cyclically during each epoch. In addition, the first epochs of the training were warm-up epochs, where only a fraction of the final learning rate was used.

For the first training session, the weights of the pretrained encoder were not allowed to change for the first 10 epochs, and only the randomly initialized weights of the decoder part of the U-Net were trained. To do this, we used a base learning rate of 0.01. We then trained the model for 200 more epochs with a base learning rate of 0.001 and a weight decay of 1e-5. During training, the Dice score on the tuning data was monitored and the checkpoint of the model that achieved the highest dice score was reloaded after training.

For the second training session on the higher resolution input data, we set the learning rate to 1e-4 and the weight decay to 1e-5, training for 200 epochs and saving the checkpoint with the highest Dice score.

All training was performed on a single GPU (NVIDIA GeForce RTX 2080ti) with 11 GB of available VRAM.

Declarations

Data availability

All relevant data are available within the manuscript.

Acknowledgements

MRM is grateful for support from the German Research Foundation (DFG, SFB 1340/1 2018, 5943/31/41/91). LCA is grateful for her participation in the BIH Charité—Junior Clinician and Clinician Scientist Program and KBB is grateful for his participation in the Digital Clinician Scientist Program funded by the Charité—Universitaetsmedizin Berlin and the Berlin Institute of Health.

Author contributions

K.K.B and J.L.V. had the idea for the present idea. S.M.N. and B.H. supervised the project. K.K.B. conducted the analysis. L.C.A. was a major contributor in writing the manuscript. All authors (K.K.B., J.L.V., L.C.A, G.E., M.R.M., B.H., S.M.N.) read and revised the manuscript critically, approving the final version.

Competing interests

The authors declare no competing interests.

References

1. Ali, F., Kasry, A. & Amin, M. The New SARS-CoV-2 Strain Shows a Stronger Binding Affinity to ACE2 Due to N501Y Mutation. arXiv preprint arXiv:2101.01791(2021).
2. Johns-Hopkins-University. Johns Hopkins Coronavirus Resource Center. (2021).
3. Peck, K. R. Early diagnosis and rapid isolation: response to COVID-19 outbreak in Korea.*Clinical Microbiology and Infection*(2020).
4. Corman, V. M. *et al.* Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Eurosurveillance.* **25**, 2000045 (2020).
5. Rubin, G. D. *et al.* The role of chest imaging in patient management during the COVID-19 pandemic: a multinational consensus statement from the Fleischner Society. *Chest.* **158**, 106–116 (2020).
6. Liu, K. C. *et al.* CT manifestations of coronavirus disease-2019: a retrospective analysis of 73 cases by disease severity. *European journal of radiology.* **126**, 108941 (2020).
7. Bai, H. X. *et al.* Performance of radiologists in differentiating COVID-19 from non-COVID-19 viral pneumonia at chest CT. *Radiology.* **296**, E46–E54 (2020).
8. Ma, J. *et al.* Towards Data-Efficient Learning: A Benchmark for COVID-19 CT Lung and Infection Segmentation. *Medical physics* (2020).
9. Ronneberger, O., Fischer, P. & Brox, T. in *International Conference on Medical image computing and computer-assisted intervention.* 234–241 (Springer).
10. Müller, D., Rey, I. S. & Kramer, F. Automated Chest CT Image Segmentation of COVID-19 Lung Infection based on 3D U-Net. arXiv preprint arXiv: 2007.*04774* (2020).
11. Hussain, M., Bird, J. J. & Faria, D. R. in *UK Workshop on Computational Intelligence*.191–202(Springer).
12. Nikolov, S. *et al.* Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. arXiv preprint arXiv:1809.04430(2018).
13. Yan, Q. *et al.* COVID-19 Chest CT Image Segmentation–A Deep Convolutional Neural Network Solution. *arXiv preprint arXiv:2004.10987* (2020).
14. Pu, J. *et al.* Automated quantification of COVID-19 severity and progression using chest CT images. *Eur. Radiol.* **31**, 436–446 (2021).
15. Lessmann, N. *et al.* Automated assessment of CO-RADS and chest CT severity scores in patients with suspected COVID-19 using artificial intelligence.*Radiology*(2020).
16. Prokop, M. *et al.* CO-RADS–A categorical CT assessment scheme for patients with suspected COVID-19: definition and evaluation.*Radiology*,201473(2020).
17. Tsai, E. B. *et al.* The RSNA International COVID-19 Open Annotated Radiology Database (RICORD).*Radiology*,203957(2021).
18. Morozov, S. *et al.* MosMedData: Chest CT Scans With COVID-19 Related Findings Dataset. *arXiv preprint arXiv:2005.06465* (2020).
19. Lowekamp, B. C., Chen, D. T., Ibáñez, L. & Blezek, D. The design of SimpleITK. *Frontiers in neuroinformatics.* **7**, 45 (2013).

20. Paszke, A. *et al.* in *Advances in neural information processing systems*.8026–8037.
21. Howard, J., Gugger, S. & Fastai A layered API for deep learning. *Information.* **11**, 108 (2020).
22. Tran, D. *et al.* in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.* 6450–6459.
23. Milletari, F., Navab, N. & Ahmadi, S. A. in 2016 *fourth international conference on 3D vision (3DV).* 565–571 (IEEE).
24. Smith, L. N. in 2017 *IEEE Winter Conference on Applications of Computer Vision (WACV).* 464–472 (IEEE).

Figures

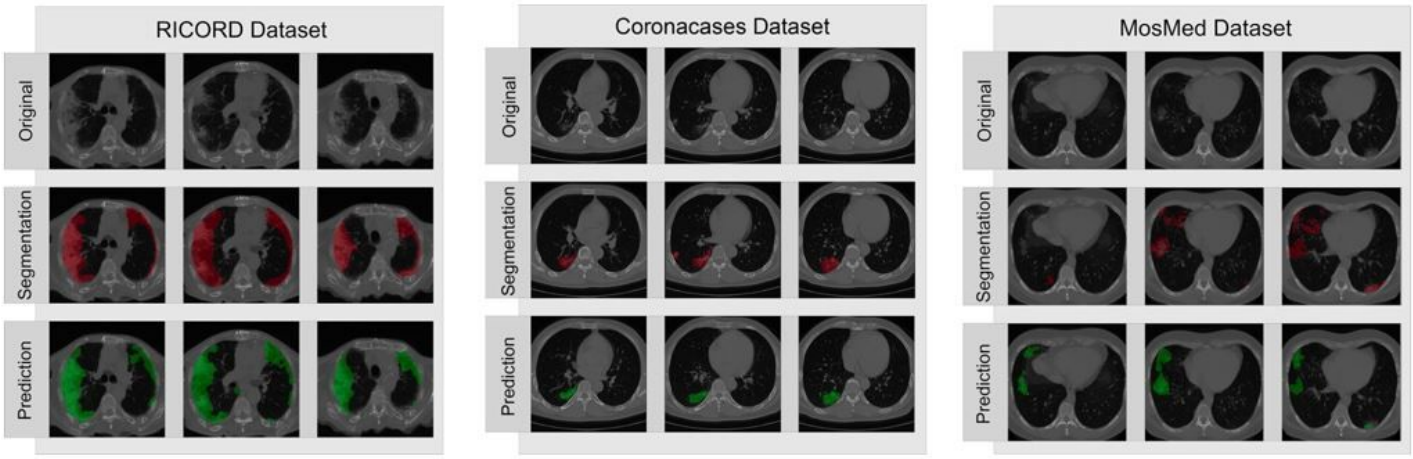

Figure 1

Example images taken from the three datasets used in this study with segmentation masks from a human annotator (red) and the corresponding predicted masks from our model (green). The CT from the MosMed dataset was originally acquired in prone position but images were flipped for this figure.

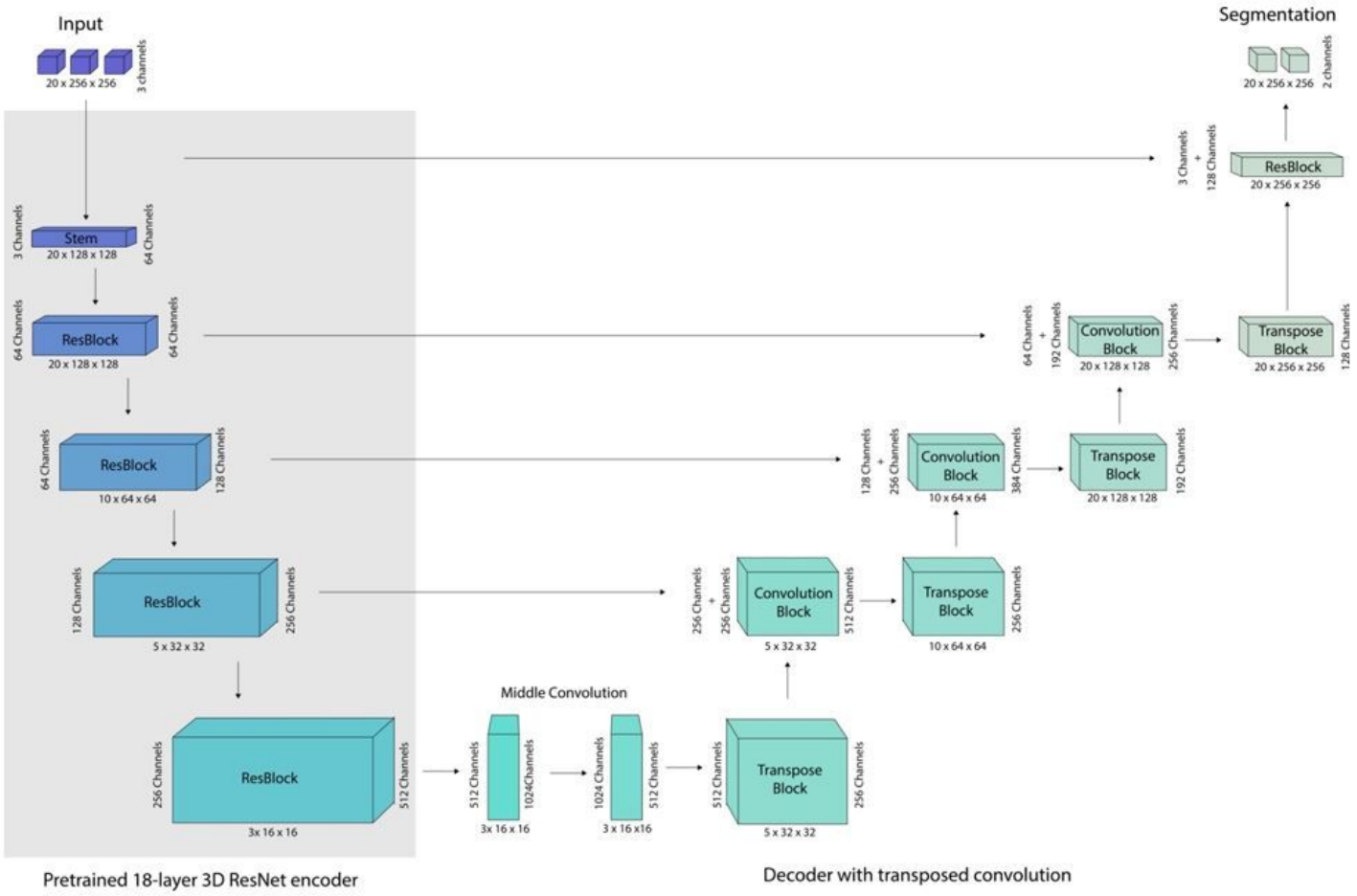

Figure 2

A schematic overview of the network architecture. As the encoder was pre-trained on color images, the expected input size was B x 3 x D x H x W, where B is the batch dimension, D the number of slices and H and W the height and width of each slice. To meet this requirement, the input images were tripled and stacked on the color channel. The encoder consisted out of a basic stem with single convolution, batch normalization and a rectified linear unit. Then, four 3D Residual Block (ResBlock) were sequentially connected to extract the image features. After each ResBlock, a skip connection to the upscaling blocks was implemented. The lower-level features were passed from the last encoder block to a double convolutional layer and then to four sequentially connected upscaling blocks. Each upscaling block consisted of a transposed convolution, which increased the spatial resolution of the feature maps and a double convolutional layer which received the output from the transposed convolution along with the feature maps from the skip connection. The final block of the decoder was again a ResBlock, which reduced the number of feature maps to the specified number of output classes.