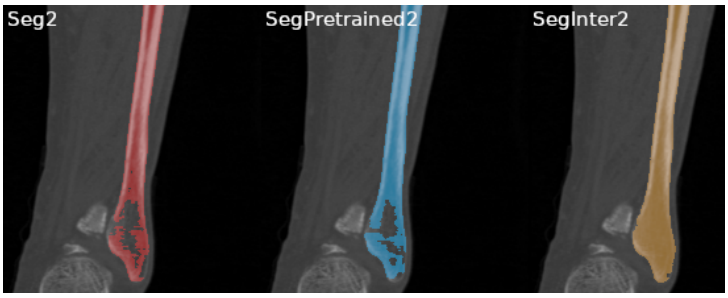
BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2020
http://excel.fit.vutbr.cz

# Benchmarking medical segmentation models with limited training sets

Kateřina Trávníčková*, Oldřich Kodym



**Abstract**
Deep learning based medical data segmentation methods can provide excellent results already. However, these results are obtained mostly thanks to the large training data sets. Obtaining the sufficient amount of correct annotations might be problematic in the medical field. This paper describes the problem of training medical segmentation models with limited annotations and proposes solutions to address the issue.

We compare the baseline segmentation model group with two other model groups. These groups use different means to battle the lack of data problem. First group is pretrained in unsupervised manner and the second one uses human interaction in form of guidance clicks. We train 14 models for each group on subsets with varying number of patients.

Segmentation model trained on small number of patients has better results when pretrained in unsupervised manner on the whole trainig set with 70 patients. Better results are obtained with the interactive method, where training on only two patients reaches Dice score 0.929 whereas the preitrained model reaches 0.830 and the baseline model only 0.749.

**Keywords:** Segmentation — Deep learning — Medical data

**Supplementary Material:** *N/A*

*xtravn24@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

## 1. Introduction

Medical data analysis with use of deep learning is an important field since the successful application of its methods can lessen the workload for medical professionals and can also generally lead to healthcare quality improvement. Although the methods of medical data segmentation have highly improved in the past few years, the shortage of professionally created annotations is still an issue.

This work focuses on the task of semantic segmentation of the long bones in human computed tomography (CT) scans, specifically on the possibilities of using small amounts of annotated data while maintaining satisfactory segmentation quality. Cases where no annotated data are present for the particular segmentation class are also considered and transfer learning view is used as a part of the selected solution. The main goal of this work is to examine ways of solving these challenging use-cases, and to observe the behavior of the selected solutions on varying number of training data rather than to compete with the quality of the state-of-the-art methods by tweaking the network architecture.

Segmentation of medical data requires high level of precision. This is why semi-automatic and more traditional methods, such as thresholding or graph cut [1], are still being used even with recent development of deep learning methods. This can be partly because of the lack of the training data for specific cases, partly because the low error tolerance is a necessity while dealing with real patients and high level of control over the segmentation method is a must. Alternatively, the classical methods can be used as an integrated part of the deep learning method, such as described in paper by Kodym et al. [2].

Segmentation models, even from different domains, usually share the encoder-decoder architecture type, as used in several other works [3, 4, 5]. In the medical domain, the most widespread architecture is the U-net [6] architecture. Authors of this model followed the encoder-decoder trend in segmentation models and developed an architecture that became a standard of medical data segmentation. Which also helps to improve the segmentation results is converting the whole model to 3D with use of 3D convolutions as Çiçek et al. have done in [7].

Apart from architecture development itself, some authors focus on other means of raising the quality of the segmentation results, for example transfer learning. Authors of the Models Genesis [8] experimented with transferring the knowledge obtained from image restauration task to the task of segmentation (and other tasks as well). The general model is trained to deal with several different types of distortion. Authors claim that this practise helps the model to learn general data character and structure from unlabeled data, speeding up the convergence of the segmentation training and providing better segmentation results than randomly initialized models.

As the automatic segmentation models can be sensitive to domain changes and their accuracy can drop in specific cases, several attempts to benefit from user interaction have also been made [9, 10, 11]. In the paper [11] for example Sakinis et al. use foreground and background click maps as a part of the model input to help the model with the right object selection. This can help to battle small domain shifts and even provide a mean to generalize the segmentation to previously unseen types of data, such as irregularly shaped tumors.

We are using a variation of the earlier mentioned U-net architecture as a baseline method for the task of longitudinal bone segmentation in axial CT slices. Apart from that, two extensions are selected and tested. The first extension benefits from transfer learning principles described by authors of Models Genesis [8]. The model is pretrained on image restauration task without needing any labels and then finetuned to the segmentation task. In the second extension, user interactions are used to improve the segmentation quality similarly to the work of Sakinis et al. [11]. This interactive model uses user clicks for object or background specification.

All three models have been trained on several different subsets of data with varying number of patients in each set. The smallest set only contains one patient. Three groups of models have been created, one for each method. This benchmarking helps to show the benefit of each method for different data amount scenarios.

The main contribution of this paper lies in the comparison of the selected models and examination of their behavior when being trained with variably sized training subsets. Our experiments suggest that using unlabeled data for pretraining can be beneficial to some extent. Best improvement was reached in the scenario with one labeled patient. This scenario best shows the benefit of using unsupervised pretraining, however the final reached Dice score is quite low. More research could be done with bigger amount of unlabeled pretraining data to prove that this method can be used in practice. Alternatively, this method can be used only as a preprocessing for semi-automatic methods, serving as a means of time reduction for human experts. Experiments with interactive segmentation yield promising results even in the one patient scenario, with minimal number of interaction. In this use-case, segmentation with only one click in each axial slice yields much better results than the baseline method. Interactive training with two patients reaches Dice score 0.929, which is considerably high value for such a small training set.

## 2. Selected solutions to CT segmentation

As already mentioned in the previous section, lack of data and, particularly, the expert-annotated data is one of the problems in the field of medical image segmentation. In this work, we set one baseline method inspired by the current medical segmentation models standards and then compare it to another two experimental approaches. These two approaches are based on unsupervised transfer learning and user interaction. The proposed model scheme and its inputs are shown in Figure 1.

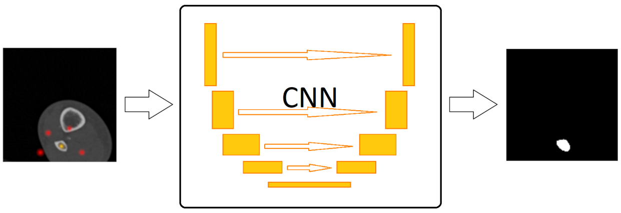


**Figure 1.** The U-net type model architecture with input and output patch. Each yellow block contains several layers consisting of convolution and ReLU. The orange arrows represent concatenation of the left block output with the right block input. Model input consists of only 1 channel of data for the baseline and the pretrained solutions. For the interactive solution, two other input channels are added (yellow color and red color in the input patch).

## 2.1 The baseline and the training process

The network architecture itself remains unchanged from 2D U-net [6] in the terms of level count and number of convolution filters in each convolution layer. Each covolution layer is followed by ReLU or Sigmoid (last layer) nonlinearity. In the training process, the network input is a $96 \times 96$ 2D patch randomly cut out of the CT data slice and corresponding groundtruth segmentation slice is used as a loss function input. Slices in each minibatch are chosen by random generator without repetition and the minibatch size is set to 32 slices. The generator is reset only after all the data has been used. We use cross entropy loss function and the Adam optimizer with learning step set to 1e-5.

We do not use the concept of epochal training because different models have been trained on different number of patients which would lead to unequally long epochs. Rather than that, we use a number of iterations for comparing the training processes of different models.

We are aware of the possibility of using regularization techniques to avoid problems with overfitting, but in this work we chose not to use them just yet, because we wanted to see the unaltered impact of different selected approaches.

The baseline model was proposed as a representation of a basic version of modern deep learning segmentational standard and serves as a reference point for comparing the efficiency of the two solutions for battling the lack of training data problem. The training process remains the same for the other two model versions if not said otherwise in the next two subsections.

## 2.2 Unsupervised pretraining

Inspired by the authors of Models Genesis [8], we use transfer learning as a means to improve the segmentation quality of a model trained on small amount of labeled data. At first, the model is trained on a secondary task of image restauration. The distortion methods are the same as in the original paper [8], specifically in- and out-painting, nonlinear intensity changes and local pixel shuffling. We use pregenerated pixel-shuffled data instead of using online generation, which we chose to do to reduce time consumption of the training process. The input of the model is a distorted patch and original patch is used as a label for simple L1 loss function. Please note that in this way the training process does not need any expert created annotations and is completely unsupervised. After the pretraining process, model can be trained on the segmentation task as described in the baseline subsection. Illustration of the restauration model data is shown in the Figure 2.

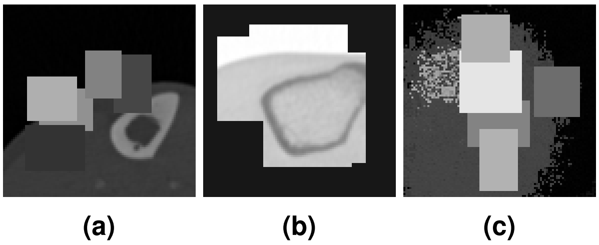


**Figure 2.** Three different training patches for the image restauration task. Inpainting (2a), Outpainting + intensity transformation (2b) and Inpainting and pixel shuffling (2c).

The benefit of this solution lies in the possibility of training on a big amount of unlabeled data, which should help the model learn the basic data structure resulting in increasing the segmentation quality even if trained only on a smaller amount of labeled data.

## 2.3 Adding user interaction

Allowing the possibility of user interaction in the final segmentation application, we can consider this approach to be another way to battle the lack of data issue. Same as Sakinis et al. [11], we use background and foreground click to specify the desired object within the input slice. We use the same method for creating the click maps as in the original paper which is as follows. Two click maps are created, one for object and one for background and both are then concatenated with the input data slice, creating a model input with 3 channels. When creating the clicks maps, pixels of a value 1 are placed to the desired click positions and whole map is then smoothed with Gaussian filter and normalized to range $\langle 0, 1 \rangle$.

In each training iteration, the model works in prediction mode at first and several user clicks are generated from the groundtruth data as well as from the previous interaction output. First click is placed in the innermost position of the groundtruth object area and following clicks are then placed in the biggest erroneous areas. Final number of interactions is randomly set to $t \in \langle 1, 5 \rangle$ for each slice in the minibatch. Unlike the original paper, we did not use multilabel data for the training. Because of that, we randomly convert the groundtruth to all-background and in such a case we only provide the background clicks. We hope that this will force the model to pay more attention to user clicks.

This solution was originally proposed to create more general model which should be able to segment previously unseen data, as suggested by the authors. Yielding good results, we investigate to what degree can this solution help to improve the results of a segmentation trained on only small amount of training data.

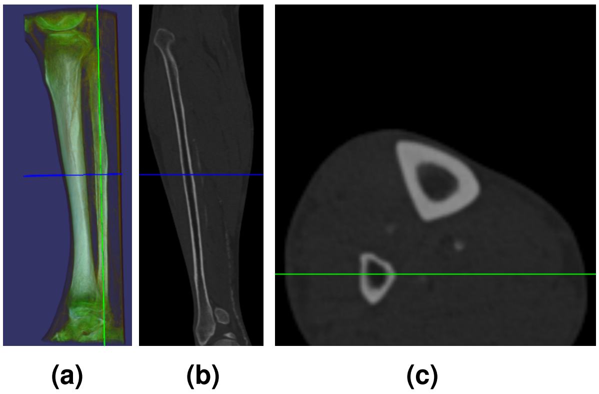


**Figure 3.** Example of the fibula bone and the surrounding tissue. 3D visualisation from the lateral view (3a), sagittal slice (3b) and axial slice (3c).

### 2.4 Dataset

All the models have been trained on a fibula subset of a fairly large dataset of human body CT scans provided by the TESCAN 3Dim company. The expert grountruth segmentation was provided as well. This subset consists of CT scans of fibula bone and its surroundings (Figure 3) from 95 patients. Up to 70 patients data was used for the training, other 9 patients data was used for validation during the training and another 10 patients data was used for evaluation of the final models. Note that the evaluation data was not used during the training process nor as the checkpoint selection criteria.

The data in Hounsfield units was clipped to range $\langle -1024, 3000 \rangle$ and then scaled within the range of $\langle 0, 1 \rangle$. Otherwise no changes or augmentations were done to this data.

## 3. Experiments with varying dataset sizes

For the best demonstration of the impact of each of the selected solutions, we created three groups of models, one for each method and one for the baseline. Each group consists of 14 models which uses the following numbers of patients for the training process: 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 40, 50, 60, 70. The naming convention for the models in this work is as follows. Three different names are used (Seg, SegPretrained and SegInter) for defining the models group. The following number identifies the number of patients used for the model training. Model named Seg1, for example, is a model that was trained on one patient and belongs to the baseline group.

The Dice score on validation data is being recorded during the training process. This metric is defined by Equation 1, where $Gt_i$ is the ground truth segmentation and $P_i$ is the prediction for the $i$-th pixel. The model checkpoint with the best validation Dice score is always saved and is considered to be the final model after the last iteration ends. All models groups were trained for 65 000 iterations of the segmentation training process.

$$Dice = \frac{2 \sum_i^N Gt_i \cdot P_i}{\sum_i^N Gt_i^2 + \sum_i^N P_i^2} \qquad (1)$$

### 3.1 The baseline models

One group of the models has been trained with the baseline method for 65 000 iterations. The graph of convergence (Fig. 4a) suggests that further Dice score improvement is possible even after the 65 000 iterations. The hard limit of 65 000 iterations has been set due to the computing grid time limits. Even though there is a possibility of a further improvement, we believe that models trained this way are sufficient to illustrate the results of our experiments. As expected, the models with fewer training patients reach much lower Dice score than the model trained on the maximum of 70 patients. It is also possible to observe the trend of decreasing impact of adding more data with higher patients counts.

### 3.2 The secondary task of image restauration

The image restauration task itself is an interesting experiment. The restauration model was trained for 43 000 (checkpoint 1) and 104 000 (checkpoint 2) iterations on the 70 patients data. Both checkpoints were compared and it is noticeable that the second checkpoint model is surprisingly good in restoring the

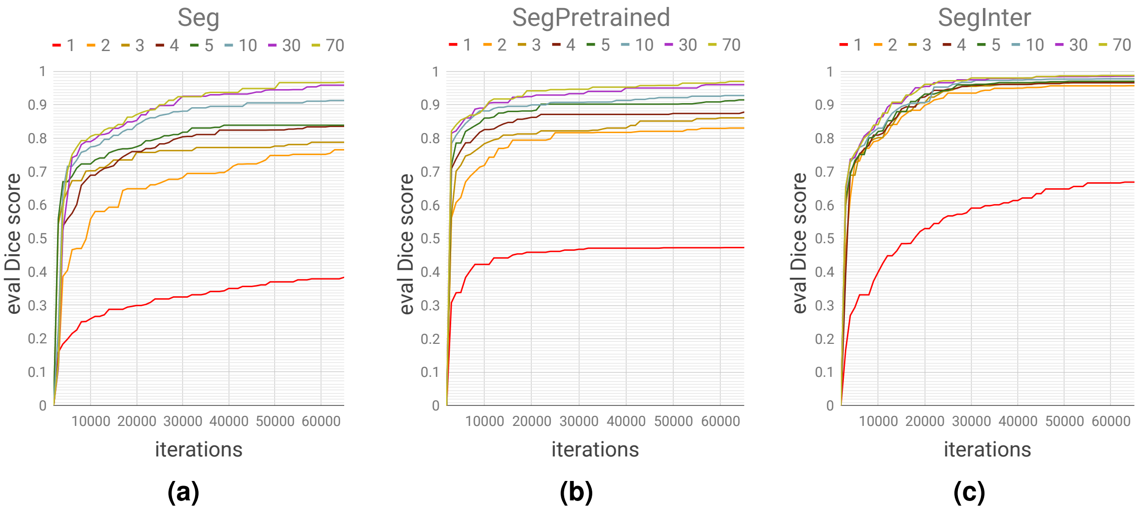


**Figure 4.** The comparison of the training process of the different models in the three groups. The graphs shows the development of the best Dice score reached on the validation data during the training process.

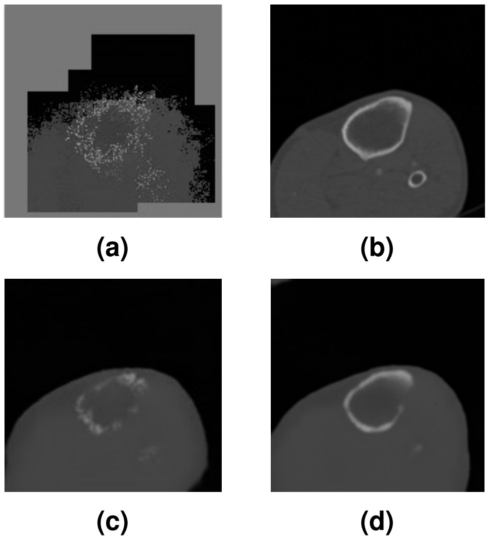


**Figure 5.** The secondary task of image restauration. Distorted input (5a), original lateral slice (5b), restauration model output after 43 000 iterations (5c) and after 104 000 iterations (5d).

bone details (Figure 5d). After the pretraining, SegPretrained models are trained the same way as the baseline group.

### 3.3 Interactive training process

As mentioned above, the interactive training process was implemented as in the work of Sakinis et al. [11]. However, there might be some changes needed in the future. The simulated interaction training process should mimic the way an actual user would click to improve the segmentation results, but in cases where the model reaches almost perfect results (mostly in later iterations), the simulated clicks degrade to clicking on very small areas of only few pixels. It is most likely that a real user would not place any guidance clicks on such areas and training the model to expect such a behavior might be inefficient. During the training process, there is maximum of 5 interactions for the training data, but only one click is used in the case of validation data and also during the evaluation of the final model.

## 4. Evaluation

For each model, the checkpoint with the best evaluation Dice score is considered to be the final one. Each of the final models in all the three groups has been tested on the evaluation data. The average Dice score for chosen models is shown in the Table 1.

Please note that there is always at least one user click present when evaluation the interactive model, which gives it a major advantage in comparison to the baseline. This is not considered a problem for the model comparison as we are researching the benefit of adding the user interaction.

### 4.1 Testing the final models

The numerical results on the testing data suggest that there is certain benefit in using either the unsupervised pretraining or the interactive approach. The most visible difference is, as expected, between the models with fewer number of training patients. In the case of the pretrained models, this is also caused by the fact that the difference between the pretraining data and the training data amount is increasing with the decreasing number of training patients, as there is only 70 patients data used for the pretraining iteself. The first noticeable thing about the training process is the

**Table 1.** Table of average dice reached by some of the models in the three groups (Seg, SegPretrained, SegInter) on the testing data. Patients column determines the number of training patients for particular model in the group.

| Patients | Seg | SegPretrained | SegInter |
|---|---|---|---|
| 1 | 0.348 | 0.392 | **0.588** |
| 2 | 0.749 | 0.830 | **0.929** |
| 3 | 0.795 | 0.855 | **0.949** |
| 5 | 0.819 | 0.858 | **0.943** |
| 10 | 0.831 | 0.904 | **0.948** |
| 30 | 0.974 | 0.977 | 0.977 |
| 70 | 0.981 | **0.984** | 0.983 |

speed up of the models convergence (Figure 4c). It seems that pretrained models reach higher Dice score on the validation data than the baseline models, especially in the scenarios with smaller amount of labeled data. The convergence speed up might bias the evaluation of the final model but although the baseline method might still improve a little in case of longer optimization, the trend is clear. Pretraining the model on larger unannotated data, possibly with adding whole human body data, would be an interesting extension to this experiment and might be examined in the future. The interactive approach yields slightly better results, which is most likely thanks to better localization of the segmented object, as described in the next subsection.

### 4.2 The benefits and the limitations of interactive training with one class

The visual examination of the outputs of different models reveals that interactive training helps the model with locating the correct object, which is shown in Figure 6. Even though it improves results of the model trained on the one patient data, it is apparent that model is still relying on the known bone shape rather than on the user clicks, at least to some extent. In some cases, clicking on the false positive areas does not help the model ignore them. It is most likely that the model just learned to either use or not use the filters trained to do the particular bone segmentation and it would probably require training on a mixed bone dataset and/or dataset with multiple classes to improve the utilization of the filters.

## 5. Conclusions

We compared the three possible solutions to bode segmentation in CT data with respect to the lack of data problem. The baseline solution was a rather traditional approach to the medical data segmentation, using the unpretrained U-net type architecture with training pairs consisting of 2D input patch and the corresponding ground truth segmentation patch. The two other solutions were chosen to battle the problem of the lack of expert-created annotation in the medical field. The first solution uses the unsupervised pretraining on a secondary task of image restauration. The second solution adds the possibility of a user interaction in form of click maps. We trained three groups of models, one for each solution. Each group consisted of 14 models with varying number of training patients, to illustrate the efficiency of each solution. The results on the testing data suggests that both of the solutions provide certain benefit.

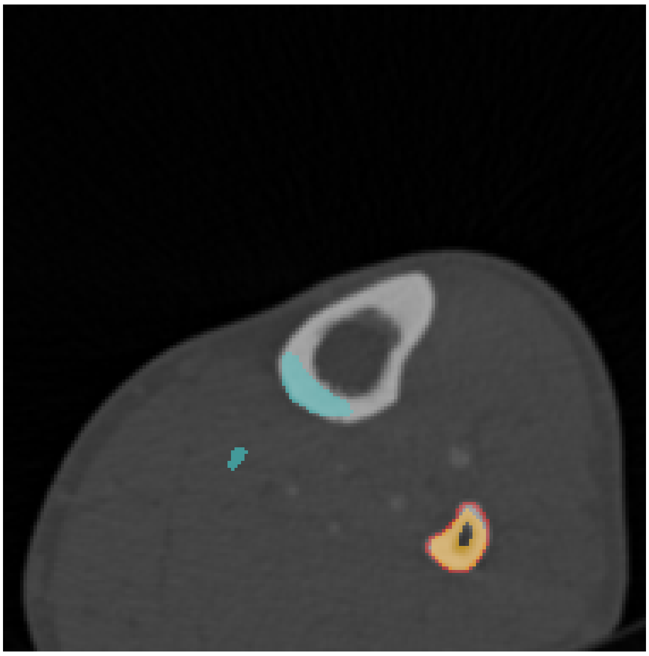


**Figure 6.** Comparison of SegInter1 and Seg1 models output. In this case, the interactive model (orange) trained on just one patient has much better result than the baseline (blue) model. Seg1 completely fails to identify the correct bone. (The ground truth segmentation is delineated with red color.)

The best improvement was between the model Seg1 and SegInter1. Both models were trained only on one training patient, but the interactive model test Dice score was 0.24 better than the baseline model. The interactive model trained on two patients already yields promising results with test Dice score 0.929. This was achieved while using only one user interaction per slice, which could be later replaced by drawing only one line in the lateral view.

Both of the selected solutions provide some degree of improvement to the lack of data problem. The use of the interactive approach could, once implemented, lessen the workload for medical experts by being the mean to reducing the necessary amount of training data. For data with a complexity of long bones such as fibula, which has both left and right variant, we only need to create manual segmentation for two bones (left and right) to obtain relatively good results. This can provide a significant boost to the creation of segmentation dataset for new types of data.

Even though we already have some promising results, the final method should include some improvements, such as using regularization methods, replacing the loss function with Diceloss or converting the whole model to 3D. The immediate future work will, however, lie mostly in the improving the interactive solution or combining both of the solutions to see if this brings some improvement. One of the possible improvements of the interactive model is adding the model output from the previous iteration/interaction as another channel to the next model input.

## Acknowledgements

I would like to thank Oldřich Kodym for contributing to this work with his knowledge from the deep learning field as well as for his patience as my supervisor. I would also like to thank the TESCAN 3Dim company for providing the CT datasets.

Computational resources were supplied by the project ”e-Infrastruktura CZ” (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

## References

[1] Y. Y. Boykov and M. . Jolly. Interactive graph cuts for optimal boundary region segmentation of objects in n-d images. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 1, pages 105–112 vol.1, July 2001.

[2] Oldřich Kodym, Michal Španěl, and Adam Herout. Segmentation of defective skulls from ct data for tissue modelling, 2019.

[3] Evan Shelhamer, Jonathon Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1–1, 05 2016.

[4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation, 2015.

[5] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, 2015.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. 2015.

[7] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation, 2016.

[8] Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael Gotway, and Jianming Liang. *Models Genesis: Generic Autodidactic Models for 3D Medical Image Analysis*, pages 384–393. 10 2019.

[9] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas Huang. Deep interactive object selection, 2016.

[10] G. Wang, W. Li, M. A. Zuluaga, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin, and T. Vercauteren. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE Transactions on Medical Imaging*, 37(7):1562–1573, July 2018.

[11] Tomas Sakinis, Fausto Milletari, Holger Roth, Panagiotis Korfiatis, Petro Kostandy, Kenneth Philbrick, Zeynettin Akkus, Ziyue Xu, Daguang Xu, and Bradley J. Erickson. Interactive segmentation of medical images through fully convolutional neural networks, 2019.