Automatic Segmentation of Multiple Sclerosis Lesions with Swin Transformer

Automatische Segmentierung von Multiple-Sklerose-Läsionen mit Swin Transformer

Guided research project

Study Program:	M. Sc., Biomedical Computing
Submission Date:	11.05.2022

Abstract

Multiple sclerosis (MS) is an inflammatory-demyelinating and degenerative disease of the central nervous system. Magnetic Resonance Imaging (MRI) can assess the inflammatory activity by detecting MS lesions. Automated segmentation of MS lesions in brain MRI scans is actively researched to help medical doctors diagnose and monitor the disease. White matter hyperintensities (WMH) that can be seen on MRI scans are associated with multiple sclerosis as it indicates inflammation associated with the disease. Manual delineation of WMHs is a tedious, and time-consuming task and since real-world clinical MRI scans are often 3D and heterogeneous in terms of image resolutions, contrasts, and noise levels. It requires a high level of expertise to annotate the desired regions correctly. Automated segmentation techniques were shown to be more objective and free from user bias compared to manual annotation. This guided research project explores Transformer-based deep learning method for automatic semantic segmentation of WMHs, as the Transformer based methods for semantic segmentation reach the state-of-the-art performances or even outperform it while also shown to produce more reliable and less texture biased predictions [1].

Contents

List of Figures

List of Tables

1. Introduction

According to World Health Organization, Multiple sclerosis (MS) is the most common primary neurological disorder of young adults, especially in Europe and North America [2]. The disease may affect various parts of the central nervous system (CNS), including the spinal cord, brainstem, cerebellum, cerebrum, and optic nerves, but the peripheral nerves are not affected. Pathologically, MS is characterized by numerous, discrete lesions scattered throughout the CNS white matter. The presence of these lesions in patients' brains causes multiple, varied symptoms and signs of neurological dysfunction. Magnetic resonance imaging (MRI) became an important imaging technique for identifying demyelinating lesions which can be used to support a clinical diagnosis and generally monitoring of patients with MS. White matter hyperintensities (WMH) that can be seen on MRI scans are associated with multiple sclerosis as it indicates inflammation associated with the disease. Manual delineation of WMHs is possible and can provide ground truth for the volumetric quantification of WMHs. However, it is a laborious, tedious, and time-consuming task and requires a high level of expertise to avoid unacceptable levels of intra- and inter-variability. Besides, this becomes more problematic with the growing size of the datasets motivated by advances in medical techniques, encouraging automated segmentation. The automated segmentation techniques were shown to be more objective and free from user bias compared to the visual WMH ratings [3]. Automatic detection of the WMH is still an ongoing challenge and would be a valuable addition to diagnostics and treatment effectiveness monitoring routines. The WMH Segmentation Challenge 2017 [4] was held to compare the state-of-the-art algorithms in conjunction with the 20th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2017). A convolutional neural network has proven to be an effective computational model for automatically extracting image features. The winning method is a variant of a fully convolutional network architecture based on U-Net [5], which takes as an input the axial slices of two modalities FLAIR and T1 from the brain MR scans during both training and testing. This guided research project examines whether the results of the proposed model can be improved by applying a transformer-based method, which became the state-of-the-art technique in both natural language processing and computer vision tasks.

2. Related work

This guided research project focuses on applying existing deep learning semantic segmentation methods in the context of medical imaging. State-of-the-art methods for this task rely on Fully Convolutional Networks (FCN) [6] and perform well on challenging segmentation benchmarks [7, 8, 9]. These methods are usually based on learnable stacked convolutions that capture semantically rich information. Methods based on FCN that are integrated into encoder-decoder architectures have become popular methods for semantic segmentation. Some approaches are solely based on a stack of consecutive convolutions followed by spatial pooling [10, 11], further approaches combine upsampled high-level and low-level feature maps during decoding to capture global information [12, 13]. Some papers also propose spatial pyramid pooling to capture multi-scale contextual information and capture global information [7, 14]. One such model would be Deeplabv3+ [7].

For semantic segmentation, labeling local patches often depends on the global image context and the local nature of convolutional filters limits the access to the global image information. To avoid this, DeepLab methods [15, 16] introduce feature aggregation with dilated convolutions and spatial pyramid pooling. It enlarges the receptive fields of convolutional networks and provides multi-scale features.

The progress in Natural Language Processing (NLP) provided several alternative aggregation schemes based on various attention strategies [17, 18] that can be used in the context of segmentation to better capture contextual information. Transformers [19] are state-of-the art for many NLP tasks. They rely on self-attention mechanisms and capture long-range dependencies among tokens (words) in a sentence. Some works employ self-attention layers to replace some or all of the spatial convolution layers in the popular ResNet [20]. Another usage in the context of segmentation would be the application of the self-attention layers as an addition to backbones [21, 22, 8] or head networks [23, 24] by providing the capability to encode distant dependencies or heterogeneous interactions. Transformer-based models have also gained a lot of attraction in medical image analysis. Transformer bottleneck was added to a 2D U-Net architecture for Multi-Atlas Abdomen Labeling Challenge [25], Transformer model was also fed with feature maps to extend the approach of 3D brain tumor segmentation [26]. During this guided research project we will focus on applying Swin Transformer for WMH segmentation of MRI scans which was already applied for Synapse multi-organ segmentation [27].

3. Methods

Semantic segmentation is a challenging computer vision problem. The goal of semantic segmentation is to assign each image pixel to a category label corresponding to the underlying object. Recent approaches to semantic segmentation typically rely on convolutional encoder-decoder architectures where the encoder generates low-resolution image features and the decoder upsamples features to segmentation maps with per pixel class scores.

3.1 Transformer

Transformer [19] is an architecture for transforming one sequence into another one based on encoder-decoder architecture, but it differs from the previously existing sequence-to-sequence models because it does not imply any Recurrent Networks. Encoder (Figure 1,

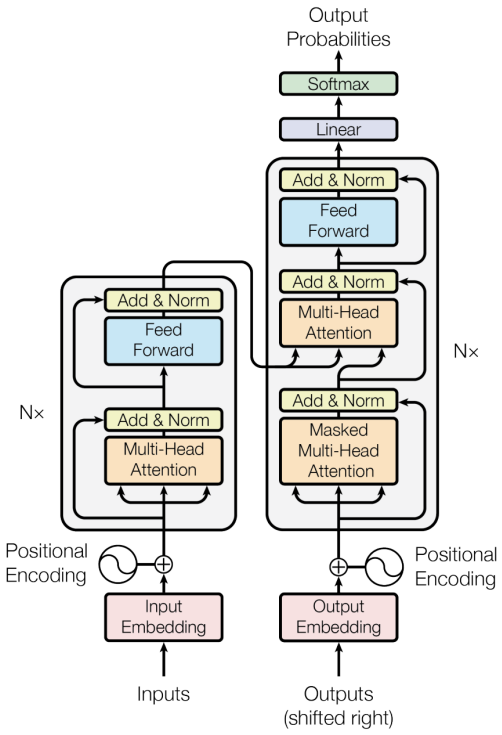


Figure 1: The Transformer - model architecture. [19]

left) and decoder (Figure 1, right) are composed of modules that can be stacked on top

of each other $N\times$ times. The modules consist mainly of Multi-Head Attention and Feed Forward layers. The input and output strings are first embedded into an n-dimensional space.

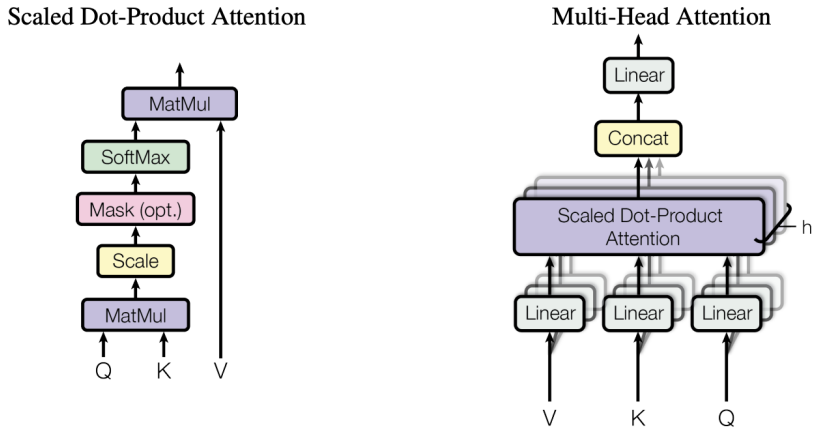


Figure 2: Scaled Dot-Product Attention (left). Multi-Head Attention consists of several attention layers running in parallel (right). [19]

Scaled Dot-Product Attention (Figure 2, left) can be described by the following equation:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

where Q is a matrix of vector representation of one word in the sequence, K contains vector representations of all the words in the sequence and V contains again the vector representations of all the words in the sequence. For the multi-head attention modules in the encoder and decoder, V consists of the same word sequence as Q. However, for the attention module that is taken into account, the encoder and the decoder sequences, V, and Q are different.

The Multi-Head Attention (Figure 2, right) concatenates multiple attention outputs linearly to expected dimensions. It can be parallelized into multiple mechanisms. The attention mechanism is repeated multiple times with linear projections of Q, K, and V. This allows the system to learn from different representations of Q, K, and V. These linear representations are achieved by multiplying Q, K, and V by weight matrices that are learned during the training.

This Feed-Forward layers can be described as a separate, identical linear transformation of each element from the given sequence. They have identical parameters for each position.

3.1.1 Visual Transformer

Transformers were originally designed for the neural machine translation problem in NLP to capture long-range dependencies among words in a sentence. Naive application of this

approach into the image domain would require evaluation of relations between each pixel and every other pixel, which is obviously not scalable. The Visual transformer (ViT) [28] converts the input image into a 1D series by cutting it into patches (patches of 16 by 16 pixels, see Figure 3) and feeding it to a linear layer. It yields a patch embedding. Position embeddings are added to the image patch embeddings. Adding the learnable position embeddings to each patch allows the model to learn the structure of the image. The rest of the pipeline is a standard encoder and decoder blocks of the transformer. The decoder learns to map patch-level encodings coming from the encoder to patch-level class scores. Next, these patch-level class scores are upsampled by bilinear interpolation to pixel-level scores.

3.1.2 Swin Transformer

Shifted Window (SWIN) Transformer [29] is a hierarchical transformer whose representation is computed with Shifted WINdows. It constructs a hierarchical representation of an image by starting from small-sized patches and gradually merging neighboring patches into deeper Transformer layers. Each patch is also named a *token*, and its feature is a concatenation of raw pixel values (e.g. RGB pixel values). The image is broken into windows that consist of patches (see Figure 3). An overview of the Swin Transformer architecture

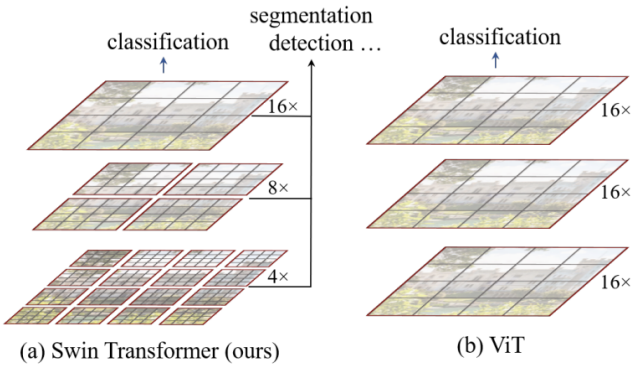


Figure 3: (a) Swin Transformer builds hierarchical feature maps by merging image patches (shown in gray) in deeper layers and has linear computation complexity to input image size due to computation of self-attention only within each local window (shown in red). (b) In contrast, previous vision Transformers [28] produce feature maps of a single low resolution and have quadratic computation complexity to input image size due to computation of selfattention globally. [29]

is presented in Figure 4. After the already mentioned Patch Partition (similar to ViT, see Figure 3), a linear embedding layer is applied to the raw-valued feature to project it to an arbitrary dimension C. To produce a hierarchical representation, the number of *tokens* is reduced by patch merging layers as the network gets deeper. Swin Transformer blocks are applied afterward for feature transformation, which also maintains the number of *tokens*. Swin Transformer is built by replacing the standard multi-head self-attention module in

a Transformer block with a module based on shifted windows. For efficient modeling, self-attention is then applied between all patches within each window but not outside it. The shifted window partitioning approach introduces connections between neighboring non-overlapping windows in the previous layer and is found to be effective in semantic segmentation. As illustrated in Figure 5, the first module uses a regular window partitioning strategy which starts from the top-left pixel, and for instance, the 8×8 feature map is evenly partitioned into 2×2 windows of size 4×4. Then, the next module adopts a windowing configuration that is shifted from that of the preceding layer, by displacing the windows by $(\lfloor \frac{4}{2} \rfloor, \lfloor \frac{4}{2} \rfloor)$ pixels from the regularly partitioned windows. The first Swin Transformer block module (Figure 4 (b) left) uses a regular window partitioning strategy (as in Layer 1 in Figure 5), whereas the next block (Figure 4 (b) right) adopts a shifted window partitioning strategy (as in Layer 1+1 in Figure 5). This adds connections across windows. Thereby Swin Transformer can act as a general-purpose backbone for semantic segmentation and computer vision in general.

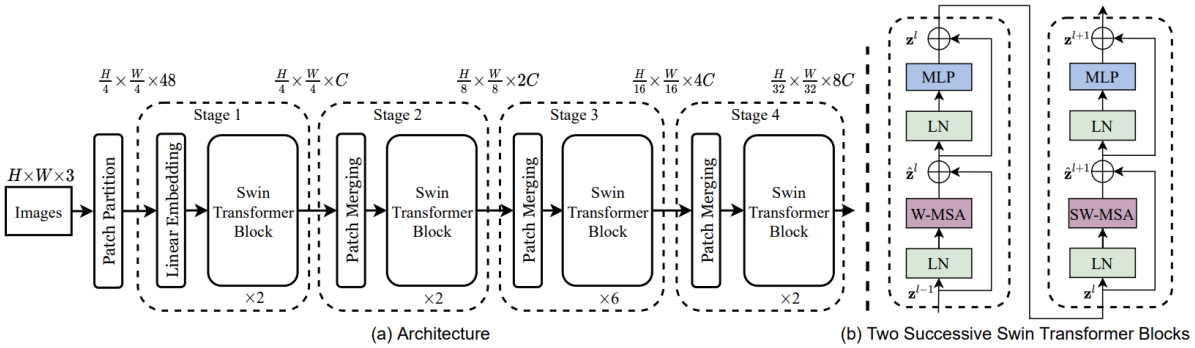


Figure 4: (a) The architecture of a Swin Transformer (Swin-T); (b) two successive Swin Transformer Blocks. W-MSA and SW-MSA are multi-head self attention modules with regular and shifted windowing configurations, respectively. [29]

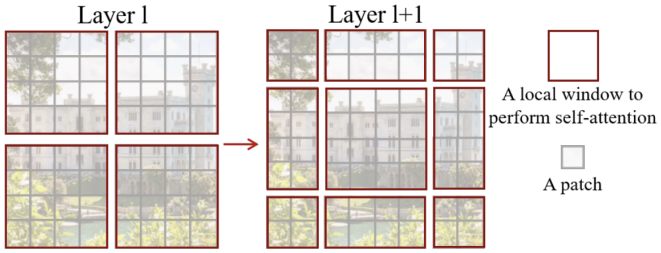


Figure 5: An illustration of the shifted window approach for computing self-attention in the proposed Swin Transformer architecture. In layer l (left), a regular window partitioning scheme is adopted, and self-attention is computed within each window. In the next layer l + 1 (right), the window partitioning is shifted, resulting in new windows. The self-attention computation in the new windows crosses the boundaries of the previous windows in layer l, providing connections among them. [29]

4. Experiments

4.1 Dataset and Evaluation Protocol

Images. This guided research project relies on the publicly available dataset from the MICCAI WMH Challenge [4]. Properties of the data are summarised in Table 1. The publicly released images were acquired from three different scanners from two hospitals in the Netherlands and one from Singapore. As shown in Table 1, there exists large difference in acquisition settings, in particular voxel sizes of the captured images differ significantly among the three scanners.

Datasets	Scanners Name	Voxel Size (m^3)	Size of FLAIR Scans	#
Utrecht	3T Philips Achieva	$0.96 \times 0.95 \times 3.00$	$240 \times 240 \times 48$	20
Singapore	3T Siemens TrioTim	$1.00 \times 1.00 \times 3.00$	$252 \times 232 \times 48$	20
GE3T	3T GE Signa HDxt	$0.98 \times 0.98 \times 1.20$	$132 \times 256 \times 83$	20

Table 1: Characteristics of MICCAI WMH Challenge dataset. The training set consists 60 subjects' data from three scanners [4].

For each subject, a 3D T1-weighted image, and a 2D multi-slice FLAIR image were provided. In our training only the FLAIR images were used for training due to the specific structure of the used segmentation framework MMSegmentation [30] and since the manual reference standard was defined on the FLAIR images. The 2D images for the training of deep learning models were generated using only FLAIR scans, excluding 16% of slices from top and bottom to avoid including images with artifacts and slices with extremely low information levels.

Evaluation protocol. Two different partitioning strategies were used for training and evaluation. First, at the patient partitioning level, the patients from each medical center were divided into 80%/20% cohorts. Hence, 16 patients from each center were used for training and four for validation. The final 1952 training and 488 validation images were converted to .png files for training. Secondly, to evaluate the model's generalizability to potential domain shift, data from one specific medical center was entirely used for validation and the rest for the training. Specifically, data yielded from Utrecht and Singapore centers were used for training and GE3T for validation. This formed the next separation

level - by a medical center to test model generalizability. The division of training set and validation set is visualised in Table 2.

	Patient level				Medical center level			
Datasets	Train		Validation		Train		Validation	
	Patients	Slides	Patients	Slides	Patients	Slides	Patients	Slides
Utrecht	16	528	4	132	20	660	0	0
Singapore	16	528	4	132	20	660	0	0
GE3T	16	896	4	224	0	0	20	1120
Total	48	1952	12	488	40	1320	20	1120

Table 2: Two data partition strategies of the MICCAI WMH Challenge dataset for training and evaluation.

Evaluation metrics. For evaluation, six different metrics proposed by the MICCAI WMH Challenge organizers were used to compare and rank the methods. Those metrics evaluate the segmentation performance in different aspects: Dice similarity coefficient (DSC), Hausdorff distance (95th percentile), Average volume difference (in percentage), Sensitivity for individual lesions (recall), precision and F1-score for individual lesions. For detailed information please refer to the original paper of the challenge [4].

4.2 Models and Training Details

Model for comparison. During this guided research project, the models were trained using the Github[1] repository called MMSegmentation [30] by open-mmlab. It can be used as a segmentation framework where different components and modules can be constructed into a customized semantic segmentation framework. For our project, the Swin Transformer backbone was selected. The overview of the models that were included in the Results sections is available in Table 3 below.

Backbone	Decode head	Auxiliary head	Loss Functions
Swin Transformer	UPerHead	FCNHead	Dice Loss Cross Entropy Loss Dice + Cross Entropy Loss Focal Loss
UNet	ASPPHead	FCNHead	Dice Loss

Table 3: Training parameters overview. FCNHead - Depthwise-Separable Fully Convolutional Network for Semantic Segmentation, head is implemented according to Fast Semantic Segmentation Network [31]; UPerHead - head implementation of UPerNet [32]; UNet - UNet backbone [5]; ASPPHead - head implementation of DeepLabV3 [16]

[1] https://github.com/open-mmlab/mmsegmentation

Training setting. All models were trained using *AdamW* optimizer for 80K iterations. A batch size of 16 was used for Cross Entropy Loss and a batch size of 8 for other models. The learning rate was set to an initial value of 1e-06 and then used a poly learning rate schedule with factor 1.0 by default. During training, data augmentation was used through random horizontal flipping with probability 0.5, gamma correction with gamma equals to 1 and application of photometric distortion to image sequentially with a probability of every transformation of 0.5.

5. Results

Class imbalance is a common problem in medical images. For WMH segmentation, the numbers of positives and negatives are highly imbalanced. Hence, Dice loss [33] and Focal loss [34] are applicable here as the loss functions for training the model. Moreover, the Focal Loss was specifically designed to address the one-stage object detection scenario in which there is an extreme imbalance between foreground and background classes during training, just like WMH regions and background. The model with Swin Transformer backbone was also trained with classical Cross Entropy Loss. For overview of the trained models please refer to the Table 4.

Train/test split	Model	Parameters	Training time
Patient-wise	SwinT + Dice Loss	59.94 M	1d 22h 42m 29s
	SwinT + Cross Entropy Loss	59.94 M	10h 25m 06s
	SwinT + Dice and Cross Entropy Loss	59.94 M	2d 0h 32m 18s
	SwinT + Focal Loss	59.94 M	12h 08m 21s
	UNet + Dice Loss	29M	9h 13m 49s
Center-wise	SwinT + Dice Loss	59.94 M	1d 22h 54m 44s
	UNet + Cross Entropy Loss	29M	17h 35m 52s

Table 4: Training time and number of parameters overview together with specific data split runs.

5.1 Evaluation on patient-wise split

The results of predictions done with Focal Loss (Figure 6) were significantly worse in comparison to Cross Entropy Loss (Figure 8) and Dice Loss (Figure 9) overall. While the results of training with combination of Cross Entropy and Dice Loss (Figure 7) show better results only according to Average volume difference compared to Cross Entropy Loss. Dice Loss results (Figure 9) report better average Dice similarity coefficient and Precision, lower Hausdorff distance and Average volume difference compared to Cross Entropy Loss results, but it is important to state that the Cross Entropy model was able to be trained on the doubled batch size, which could be an advantage in this comparison.

As a baseline, the model with a UNet backbone was also trained with Dice Loss (Figure 10). Here, Transformer with Dice Loss model achieves higher Precision, but

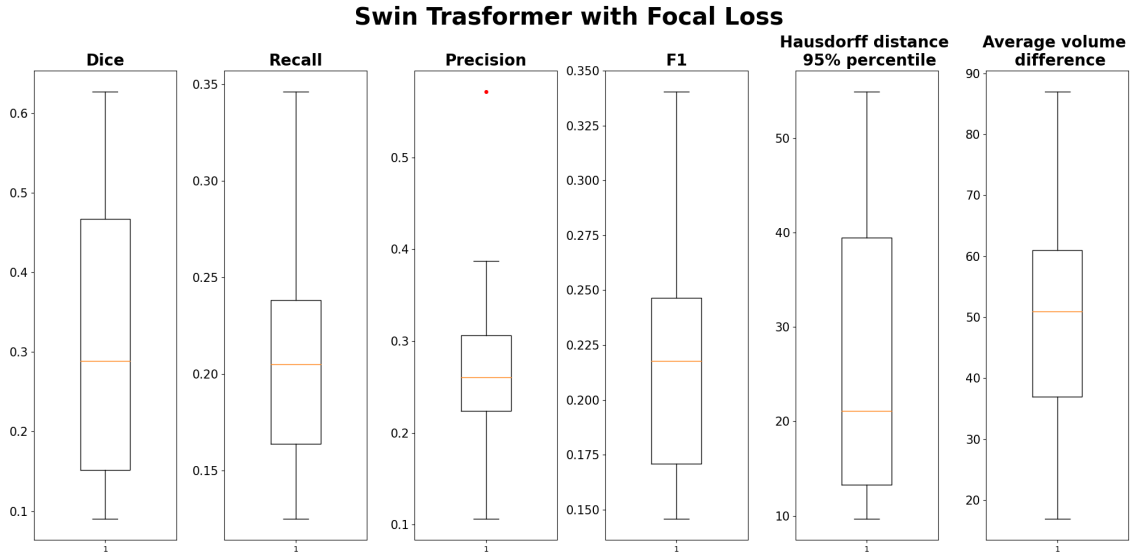


Figure 6: Dice similarity coefficient, Sensitivity for individual lesions (Recall), Precision, F1-score for individual lesions, Hausdorff distance (95th percentile) and Average volume difference computed on validation data for **Swin Transformer backbone trained with Focal Loss**.

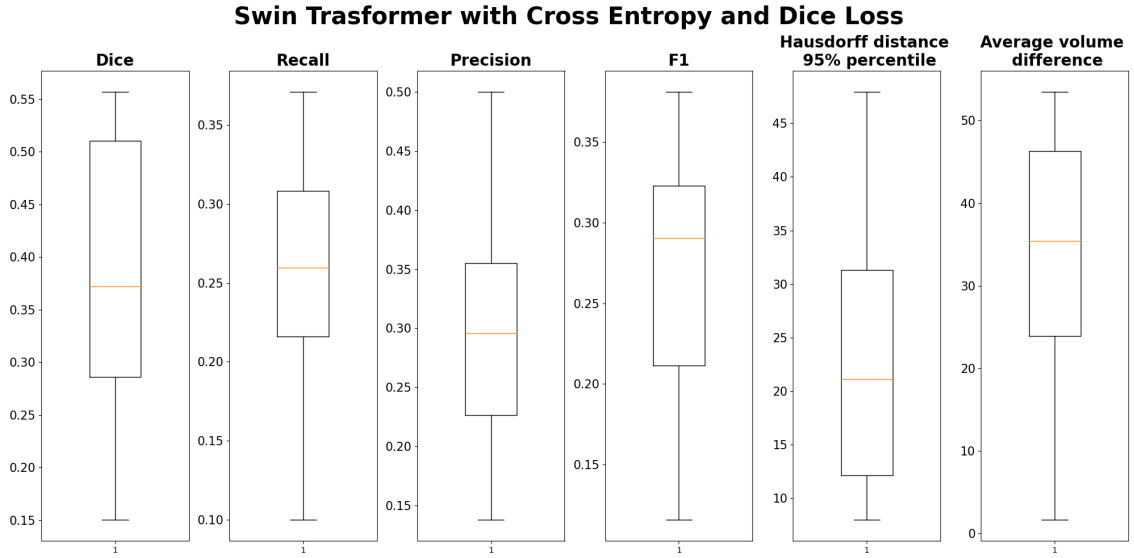


Figure 7: Dice similarity coefficient, Sensitivity for individual lesions (Recall), Precision, F1-score for individual lesions, Hausdorff distance (95th percentile) and Average volume difference computed on validation data for **Swin Transformer backbone trained with Cross Entropy and Dice Loss**.

lower Dice similarity coefficient, Recall, and F1. Hausdorff distance and Average volume difference are also elevated in comparison to UNet results.

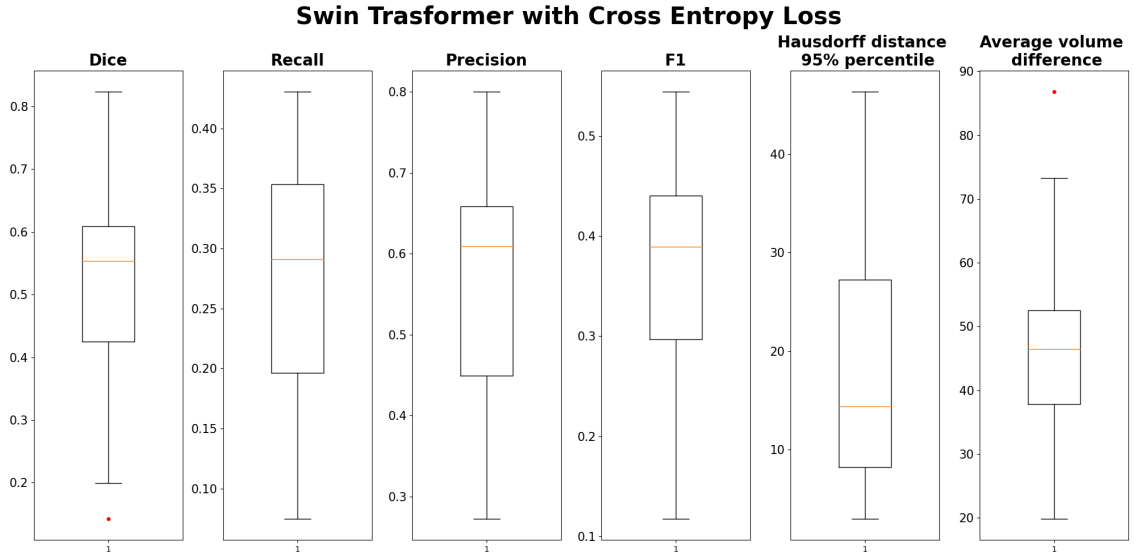


Figure 8: Dice similarity coefficient, Sensitivity for individual lesions (Recall), Precision, F1-score for individual lesions, Hausdorff distance (95th percentile) and Average volume difference computed on validation data for **Swin Transformer backbone trained with Cross Entropy Loss**.

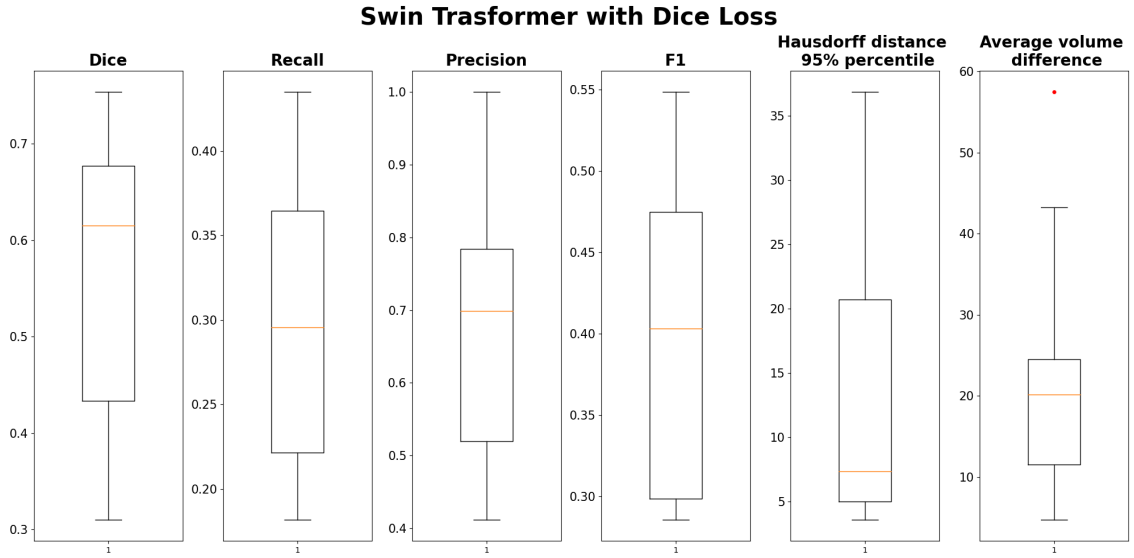


Figure 9: Dice similarity coefficient, Sensitivity for individual lesions (Recall), Precision, F1-score for individual lesions, Hausdorff distance (95th percentile) and Average volume difference computed on validation data for **Swin Transformer backbone trained with Dice Loss**.

5.2 Evaluation on center-wise split

Training on data from Singapore and Utrecht medical centers and validation on GE3T data center degraded the performance of UNet (Figure 11) compared to the patient level (Figure 10) in Dice similarity, Recall, and F1-score for individual lesions, but UNet on medical center separation managed to achieve better Precision. The Swin Transformer

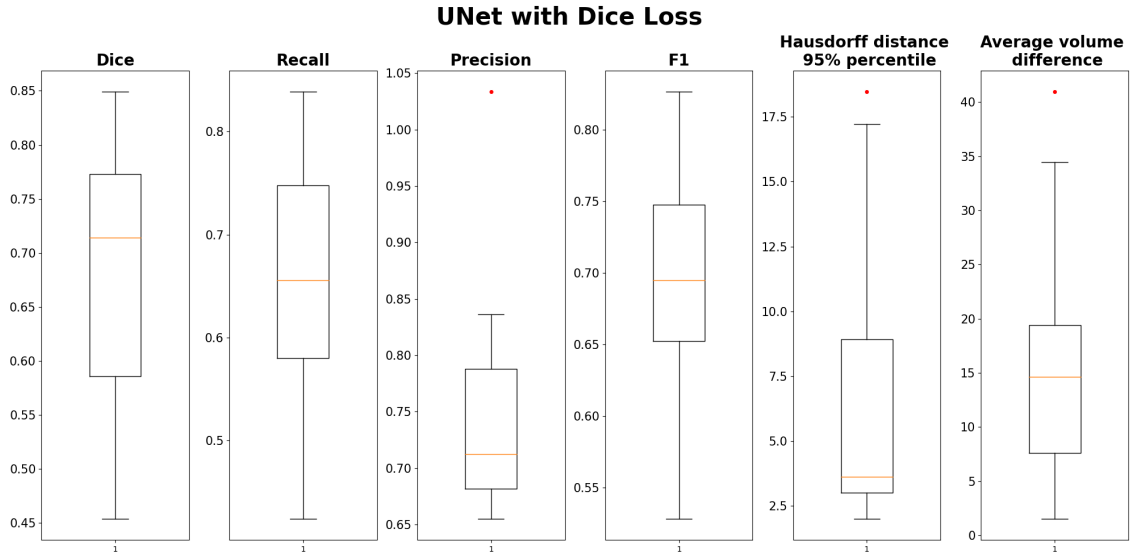


Figure 10: Dice similarity coefficient, Sensitivity for individual lesions (Recall), Precision, F1-score for individual lesions, Hausdorff distance (95th percentile) and Average volume difference computed on validation data for **UNet backbone trained with Dice Loss**.

results (Figure 11) compared to the patient level training (Figure 9) shows slightly worse results in Dice similarity, but better Sensitivity, Precision, and F1-score for individual lesions. Compared to each other (Figure 11) the results of Dice score and Sensitivity are slightly worse by Swin Transformer but Precision is narrowly better. By viewing the segmentation results of some selected slides (Figure 13) it might be assumed that the UNet-based method might experience compilations while predicting within the regions of WMHs. Particularly by patient 132 slide 54 and patient 132 slide 51 (Figure 13), that UNet predicts a considerate amount of false negatives within the WMH masses. It is when the masses are smaller, and there is not as much heterogeneity within, that the UNet-based method performs better than the Swin Transformer-based approach. Swin Transformer seems to handle wider regions better by sometimes achieving quite good results as is shown by patient 132 slide 54 (Figure 13) while still struggling with regions that are located at the very border of WMH and non-WMH regions or even missing the small regions all together (Figure 13 patient 105 slide 45).

The comparison of Swin Transformer and UNet metrics while validating on Singapore medical data scans and training with GE3T and Utrecht (Figure 12) also indicates poor results of Transformer based network: worse Dice score and Sensitivity but also higher Hausdorff distance and Average volume difference.

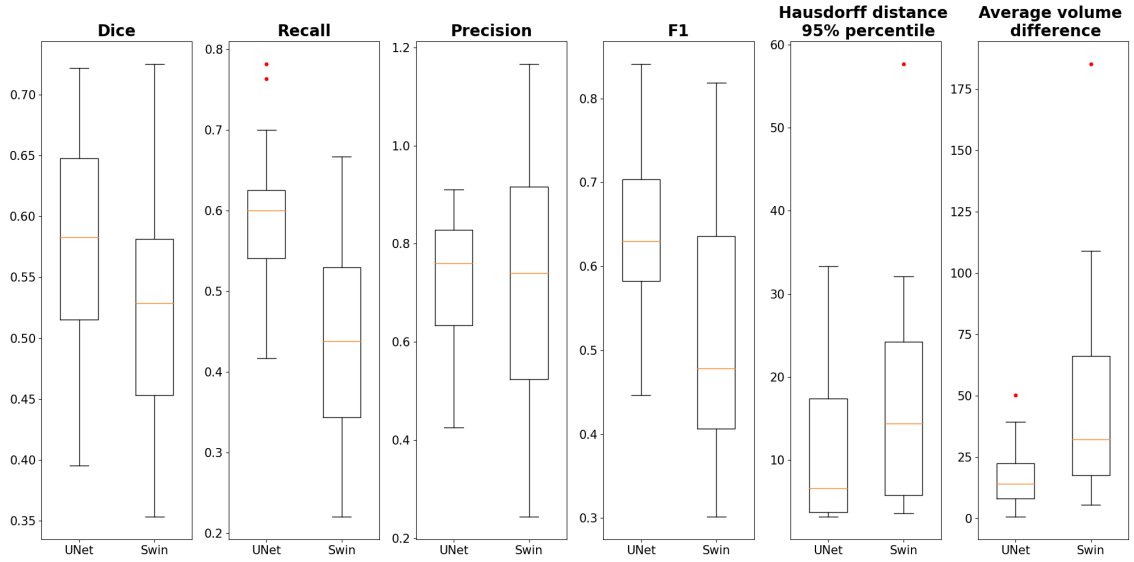


Figure 11: Dice similarity coefficient, Sensitivity for individual lesions (Recall), Precision, F1-score for individual lesions, Hausdorff distance (95th percentile) and Average volume difference computed on validation data for **UNet and Swin Transformer backbone trained with Dice Loss validated on GE3T medical center scans**.

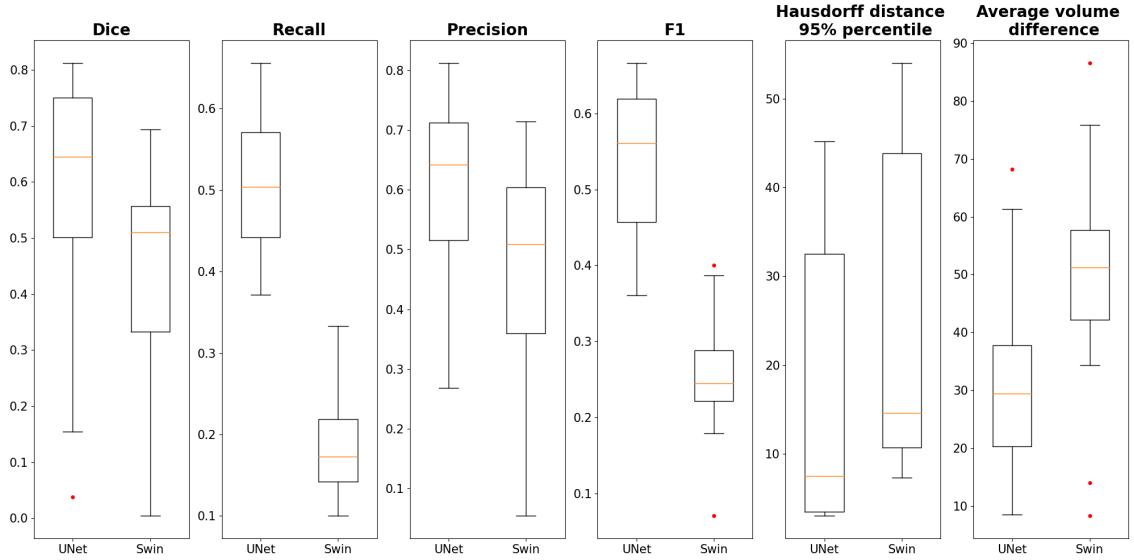
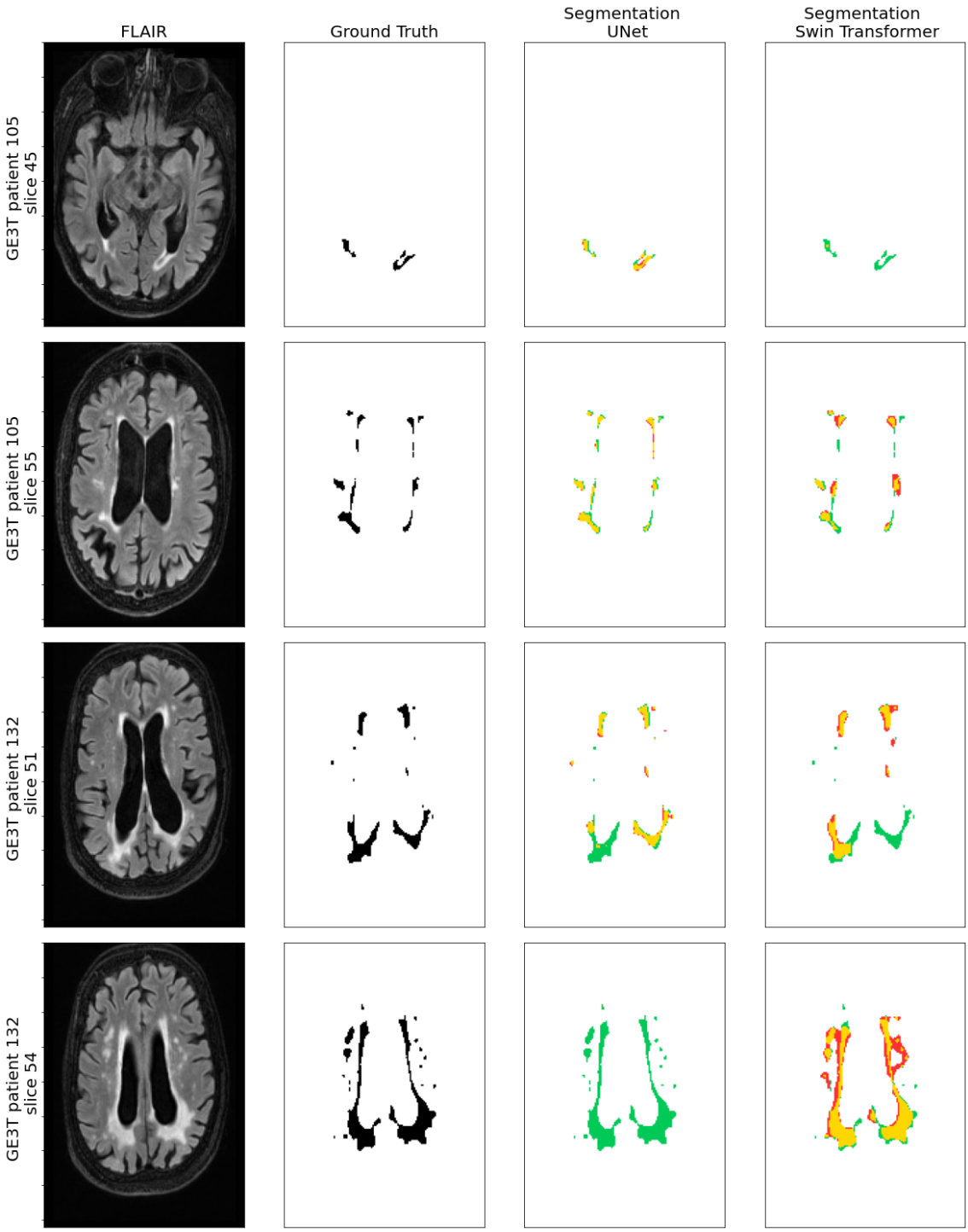


Figure 12: Dice similarity coefficient, Sensitivity for individual lesions (Recall), Precision, F1-score for individual lesions, Hausdorff distance (95th percentile) and Average volume difference computed on validation data for **UNet and Swin Transformer backbone trained with Dice Loss validated on Singapore medical center scans**.



Patient ID	Model	Dice	Recall	Precision	F1	H95	AVD
105	UNet	**0.665**	0.622	0.636	0.629	3.741	5.654
	Swin	0.624	0.511	0.851	0.638	5.000	29.815
132	UNet	0.546	0.469	0.425	0.446	7.280	50.068
	Swin	**0.708**	0.428	0.538	0.477	5.000	10.209

Figure 13: Two cases of segmentation results with two models validated GE3T data. From left to right: FLAIR MR image, the associated ground truth, segmentation result using UNet and Swin Transformer backbones. For the segmentations the yellow area is the overlap between the segmentation maps and the ground truth (true positives), the green pixels are the false negatives and the red ones are the false positives. (AVD = Average volum difference, H95 = Hausdorff distance 95% percentile)

6. Discussion

This guided research project attempted to apply and evaluate Transformer models in the case of semantic segmentation of WMH in MR scans. The achieved results indicate promising results for this task. There is a lot of room for improvement to have a more fair and clear comparison with current state-of-the-art models and architectures that have already gone through the phases of adjustment and tuning with regard to medical imaging applications. One of the severe disadvantages in the current work is the usage of only one modality (FLAIR) for training and validation. The available framework and toolboxes (in our case MMSegmentation [30]) for the Transformer networks are still orientated on the three-channel input, and after long and excruciating attempts to aggregate the inputs into a suitable data format, the decision was made to convert solely FLAIR images into RGB input in form of png files. That is of course not desired, as the ideal input would be a two-channel matrix with embedded T1 and FLAIR information. The fact that the presented results and baseline did not fully achieve the final challenge evaluations [35] proves that there are more pre and post-processing steps needed to be done.

Furthermore, the Swin Transformer method mainly considers the design of the Transformer encoder, neglecting the contribution of the decoder for further improvements. The possible improvement of the performed experiments would be applying the more advanced models like Segformer [36]. The prior interest of this guided research project was the application of SegFormer for WMH segmentation. The official implementation of this method is a part of the MMSegmentation [30] model zoo, but due to before mentioned channel complications and difficulties with the installation of the framework the project needed to be refocused on Swin Transformers for which the setup and application appeared more accessible. The further obstacle related to the chosen framework was the inability to train on multiple GPUs, which was caused by still unknown incompatibilities with the working environment.

Using transformers allows for making more detailed and globally consistent predictions compared to convolutional networks. In particular, performance is improved when a large amount of training data is available. Despite the broad applications of the transformer model, it struggles to perform well for some tasks with limited training data. If that is the case for us, there are theoretically justified optimization strategies to train deeper transformer models with improved generalization and faster convergence speed on small

datasets [37].

Specifically for Swin Transformer, one could further experiment with different loss functions, such as the Sensitivity-specificity loss function and since Dice Loss showed promising results, Soft Dice Loss might an option as well. Furthermore, all trainings in this guided research were performed with a tiny Swin Transformer. There are further, small and big Swin Transformers available, that are able to reach more increased depths that could be interesting to apply.

Overall, Transformer models are definitely worth looking into further to find the suitable models, parameters, and setting applicable for the medical image domain. For this, a suitable framework must be chosen very carefully. The potentially better option would HuggingFace's Transformers [38] which were not explored during this project due to the time constraints.

Bibliography

[1] Kishaan Jeeveswaran et al. *A Comprehensive Study of Vision Transformers on Dense Prediction Tasks*. 2022. DOI: 10.48550/ARXIV.2201.08683. URL: https://arxiv.org/abs/2201.08683.

[2] Sharon Warren, Kenneth G Warren, and World Health Organization. *Multiple sclerosis / Sharon Warren, Kenneth G. Warren*. 2001.

[3] Martha E Payne et al. "Development of a semi-automated method for quantification of MRI gray and white matter lesions in geriatric subjects". In: *Psychiatry Research: Neuroimaging* 115.1 (2002), pp. 63–77. ISSN: 0925-4927. DOI: https://doi.org/10.1016/S0925-4927(02)00009-4. URL: https://www.sciencedirect.com/science/article/pii/S0925492702000094.

[4] Hongwei Li et al. "Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images". In: *NeuroImage* 183 (2018), pp. 650–665. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2018.07.005. URL: https://www.sciencedirect.com/science/article/pii/S1053811918305974.

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by Nassir Navab et al. Cham: Springer International Publishing, 2015, pp. 234–241.

[6] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[7] Liang-Chieh Chen et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 801–818.

[8] Minghao Yin et al. "Disentangled non-local neural networks". In: *European Conference on Computer Vision*. Springer. 2020, pp. 191–207.

[9] Changqian Yu et al. “Context prior for scene segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 12416–12425.

[10] Clement Farabet et al. “Learning hierarchical features for scene labeling”. In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2012), pp. 1915–1929.

[11] Pedro Pinheiro and Ronan Collobert. “Recurrent convolutional neural networks for scene labeling”. In: *International conference on machine learning*. PMLR. 2014, pp. 82–90.

[12] Md Amirul Islam et al. “Gated feedback refinement network for dense image labeling”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 3751–3759.

[13] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. “Segnet: A deep convolutional encoder-decoder architecture for image segmentation”. In: *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017), pp. 2481–2495.

[14] Hengshuang Zhao et al. “Pyramid scene parsing network”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2881–2890.

[15] Liang-Chieh Chen et al. “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs”. In: *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2017), pp. 834–848.

[16] Liang-Chieh Chen et al. “Rethinking atrous convolution for semantic image segmentation”. In: *arXiv preprint arXiv:1706.05587* (2017).

[17] Jun Fu et al. “Scene segmentation with dual relation-aware attention network”. In: *IEEE Transactions on Neural Networks and Learning Systems* 32.6 (2020), pp. 2547–2560.

[18] Jun Fu et al. “Dual attention network for scene segmentation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 3146–3154.

[19] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[20] Han Hu et al. “Local relation networks for image recognition”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 3464–3473.

[21] Irwan Bello et al. “Attention augmented convolutional networks”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 3286–3295.

[22] Yue Cao et al. “Gcnet: Non-local networks meet squeeze-excitation networks and beyond”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2019, pp. 0–0.

[23] Jiayuan Gu et al. “Learning region features for object detection”. In: *Proceedings of the european conference on computer vision (ECCV)*. 2018, pp. 381–395.

[24] Han Hu et al. “Relation networks for object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 3588–3597.

[25] Jieneng Chen et al. “Transunet: Transformers make strong encoders for medical image segmentation”. In: *arXiv preprint arXiv:2102.04306* (2021).

[26] Wenxuan Wang et al. “Transbts: Multimodal brain tumor segmentation using transformer”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 109–119.

[27] Hu Cao et al. *Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation*. 2021. DOI: 10.48550/ARXIV.2105.05537. URL: https://arxiv.org/abs/2105.05537.

[28] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[29] Ze Liu et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 10012–10022.

[30] MMSegmentation Contributors. *MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark*. https://github.com/open-mmlab/mmsegmentation. 2020.

[31] Rudra P K Poudel, Stephan Liwicki, and Roberto Cipolla. *Fast-SCNN: Fast Semantic Segmentation Network*. 2019. DOI: 10.48550/ARXIV.1902.04502. URL: https://arxiv.org/abs/1902.04502.

[32] Tete Xiao et al. *Unified Perceptual Parsing for Scene Understanding*. 2018. DOI: 10.48550/ARXIV.1807.10221. URL: https://arxiv.org/abs/1807.10221.

[33] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. *V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation*. DOI: 10.48550/ARXIV.1606.04797. URL: https://arxiv.org/abs/1606.04797.

[34] Tsung-Yi Lin et al. *Focal Loss for Dense Object Detection*. DOI: 10.48550/ARXIV.1708.02002. URL: https://arxiv.org/abs/1708.02002.

[35] Hugo J. Kuijf et al. “Standardized Assessment of Automatic Segmentation of White Matter Hyperintensities and Results of the WMH Segmentation Challenge”. In: *IEEE Transactions on Medical Imaging* 38.11 (2019), pp. 2556–2568. DOI: 10.1109/TMI.2019.2905770.

[36] Enze Xie et al. *SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers*. 2021. DOI: 10.48550/ARXIV.2105.15203. URL: https://arxiv.org/abs/2105.15203.

[37] Peng Xu et al. DOI: 10.48550/ARXIV.2012.15355. URL: https://arxiv.org/abs/2012.15355.

[38] Thomas Wolf et al. “Transformers: State-of-the-Art Natural Language Processing”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.