DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Biomedical Computing

Deep Learning Based Analysis of Tumor-infiltrating Lymphocytes in H&E Stained Histological Sections for Survival Prediction of Breast Cancer patients

Margaryta Olenchuk

DEPARTMENT OF INFORMATICS

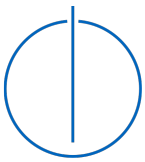
TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Biomedical Computing

Deep Learning Based Analysis of Tumor-infiltrating Lymphocytes in H&E Stained Histological Sections for Survival Prediction of Breast Cancer patients

Deep Learning basierte Analyse von tumorinfiltrierenden Lymphozyten in H&E gefärbten histologischen Schnitten zur Überlebensvorhersage von Brustkrebspatienten

Author:	Margaryta Olenchuk
Supervisor:	Prof. Dr. Peter Schüffler
Advisor:	Dr. Philipp Wortmann, Ansh Kapil
Submission Date:	15.12.2022



I confirm that this master's thesis in biomedical computing is my own work and I have documented all sources and material used.

Munich, 15.12.2022 Margaryta Olenchuk

Acknowledgments

Abstract

Kurzfassung

Contents

1. Related work

The focus of the following chapter is the most critical areas of computer vision: deep learning-based approaches for semantic image segmentation, particularly for medical images analysis. As Shephard, Adam et al. discuss [1] segmentation of tumor/stroma as well as the detection of TILs can be viewed as semantic segmentation problem.

1.1. Deep learning-based semantic segmentation

Semantic segmentation is a computer vision task that aims to differentiate regions by assigning a class label to each pixel. Due to the success of deep learning models in a wide range of vision applications, various deep learning-based algorithms for image segmentation have been developed and published in the literature [2]. One of the most prominent deep learning architectures used by the computer vision community include fully convolutional networks (FCNs) [3], encoder-decoders [4], generative adversarial networks (GANs) [5] and recurrent neural networks (RNNs) [6].

FCNs [3] are among the most widely used architectures for computer vision tasks and their general architecture consists of several learnable convolutions, pooling layers, and a final 1×1 convolution. While models based on this architecture perform well on challenging segmentation benchmarks, e.g. applied on scene segmentation [7] and instance aware semantic segmentation [8], they are also used on segmentation problems in histology domain such as colon glands segmentation [9], identification of muscle and messy regions in contexts of inflammatory bowel disease [10] as well as nuclei [11] and TILs [12] segmentation for breast cancer all performed on the Hematoxylin and Eosin (H&E) stained histopathology images. Moreover, the FCN method was applied for semantic segmentation of TCGA [13] breast data set [14], which is also used in this master thesis. However, despite its popularity, the conventional FCN model has limitations such as loss of localization and the inability to process potentially useful global context information due to a series of down-sampling and a high sampling rate.

A popular group of deep learning models for semantic image segmentation that aims to solve the aforementioned issues of FCNs is based on the convolutional encoder-decoder architecture [4]. Their model consists of two parts, an encoder consisting of convolutional layers and a deconvolution network that consists of deconvolution and unpooling layers that take the feature vector as input and generate a map of pixel-wise class probabilities. Example for such a convolutional encoder-decoder architecture for image segmentation is SegNet [15]. The SegNet's encoder network has 13 convolutional layers with corresponding layers in the decoder. The final decoder output is fed to a multi-class soft-max classifier to produce class

probabilities for each pixel independently. The main feature of SegNet is that the decoder uses pooling indices computed in the max-pooling step of the corresponding encoder to perform non-linear upsampling. This allows it to achieve high scores for road scene understanding problems [15], COVID-19 lung computed tomography image segmentation [16], liver tumor segmentation in computed tomography scans [17] and colon cancer histopathological images analysis [18]. There are several encoder-decoder models initially developed for biomedical image segmentation. Ronneberger et al. [19] proposed the U-Net model for segmenting biological microscopy images that can train with few annotated images effectively. U-Net has an FCN-like down-sampling part that extracts features with 3×3 convolutions and an up-sampling part. Feature maps from the encoder are copied to the corresponding decoder part of the network to avoid losing pattern information. Besides the segmentation of neuronal structures in electron microscopic recordings demonstrated in the original paper [19], U-Net was applied for numerous further tasks such as nuclei segmentation in histology images [20, 21], segmenting individual colon glands in histopathology images [22], epidermal tissue segmentation in histopathological images of skin biolsies [23] and cell segmentation on histopathology triple-negative breast cancer patients dataset [24]. A further example of an encoder-decoder model for semantic segmentation of histopathology images is HookNet [25]. The architecture consists of two encoder-decoder branches to extract contextual and fine-grained detailed information and combine it (hook up) for the target segmentation. The model showed improvement compared with single-resolution models and was applied to segment different histopathologies like breast cancer tissue sections [25], lung squamous cell carcinoma [25], invasive melanoma tumor [26] and cervical cancer [27] slides.

Another widely used group of deep learning models for semantic segmentation are the atrous (or dilated) convolutional models that include the DeepLab family [28, 29]. The use of atrous convolutions addresses the decreasing resolution caused by max-pooling and striding and Atrous Spatial Pyramid Pooling (ASPP) analyzes an incoming convolutional feature layer with filters at multiple sampling rates allowing to capture objects and image context at multiple scales to robustly segment objects at multiple scales. DeepLabv3+ [30] uses encoder-decoder architecture including atrous separable convolution, composed of a depthwise convolution (spatial convolution for each channel of the input) and pointwise convolution (1×1 convolution with the depthwise convolution as input). Authors [30] demonstrated the effectiveness of DeppLabv3+ model with modified Xception backbone at recognition of visual object classes in realistic scenes, but it also found multiple applications such as skin lesion segmentation [31], segmentation of H&E stained breast cancer [32] and colorectal carcinoma [33] histopathology images. Despite all the efforts, even this popular architecture has constraints in learning long-range dependency and spatial correlations due to the inductive bias of locality and weight sharing [34] that may result in sub-optimal segmentation of complex structures.

GANs [5] have been applied to a wide range of computer vision tasks, and have been adopted for image segmentation too. The general architecture of GANs consists of the discriminator and the generator. The generator learns the training data distribution and produces similar data, while the discriminator discriminates between real data and simulated

data. Hence the task of the generator is to learn to generate the best images to fool the discriminator. There are many extended models such as conditional GAN (cGAN) [35] where the additional information is added to both the generator and the discriminator as a condition. This architecture was used for semantic segmentation of brain tumor in magnetic resonance imaging [36] and nuclei segmentation in histopathology images [37]. Further extended version of cGAN, pix2pix [38] was developed for conversion between different types of images but also found use cases in medical setting such as cell image segmentation on the fluorescence liver images [39] and retinal blood vessel segmentation [40]. A further GAN extension originally developed for image transformation between two domains but also applicable for segmentation is CycleGAN [41]. The architecture has two mirror-symmetric GANs to form a ring network to find the mapping between domains. For instance, CycleGAN was applied to kidney tissue [42] segmentation. Some GAN-based models were specifically developed for semantic segmentation in the medical domain, such as Domain Adaptation and Segmentation GAN (DASGAN) [43] that performs image-to-image translation and semantic tumor epithelium segmentation. It has an extended CycleGAN architecture with discriminator networks adjusted to predict pixel-wise class probability maps on top of predicting the correct source of an image. As a further example the proposed architecture consisting of pyramid of GAN structures [44], each responsible for generating and segmenting images at a different scale, was applied to segment prostate histopathology images.

RNNs [6] have also proven to be useful in modeling the short/long-term dependencies among pixels to generate segmentation maps. Pixels can be linked together and processed sequentially to model global contexts and improve semantic segmentation. ReSeg [45] is an RNN-based model for semantic segmentation. Each layer is composed of four RNNs that go through the image horizontally and vertically in both directions to provide relevant global information, while convolutional layers extract local features that are then followed by up-sampling layers to recover the predictions at original image resolution. Another important development is a pixel-level segmentation of scene images using a long-short-term-memory (LSTM) network [46]. Segmentation is then carried out by 2D LSTM networks, allowing texture and spatial model parameters to be learned within a single model. But despite all further developments that showcase the potential even for histopathology image segmentation: RACE-net [47] applied for segmentation of the cell nuclei in H&E stained breast cancer slides, Her2Net [48] segmenting cell membranes and nuclei from human epidermal growth factor receptor-2 (HER2)-stained breast cancer images, etc., an important limitation of RNNs is that, due to their sequential nature, they are comparably slower, since this sequential calculation cannot be easily parallelized.

The Transformer in Natural Language Processing is an architecture that aims to solve sequence-to-sequence problems based on encoder-decoder architecture. These models rely on self-attention mechanisms and capture long-range dependencies among tokens (words) in a sentence without using RNNs or convolution. Transformers have also emerged into image semantic segmentation. Recent studies have shown that the Transformers can achieve superior performance than CNN-based approaches in various semantic segmentation applications [49]. The state-of-the-art Transformer-based semantic segmentation methods can be

often applied either as convolution-free models or/and as CNN-Transformer hybrid models. Swin-Transformer [50] for instance is a pure hierarchical Transformer that can serve as a backbone for various computer vision tasks including semantic segmentation. To tokenize the image, it brakes the image into windows that further consist of patches. It constructs a hierarchical representation of an image by starting from small-sized patches and gradually merging neighboring patches into deeper Transformer layers. Swin-Transformer or its slightly modified successors found its application in the medical domain as well, often as a backbone, for example for colon cancer segmentation in H&E stained histopathology images [51] or gland segmentation [52]. A further popular fully transformer-based model for semantic segmentation is Segmenter [53]. The encoder consists of Multi-head Self Attention and Multi-Layer Perceptron (MLP) blocks, as well as two-layer norms and residual connections after each block and a linear decoder that bilinearly up-samples the sequence into a 2D segmentation mask. While performing well on scene segmentation [53], is not particularly used in the medical domain. In the field of medical image segmentation, TransUNet [54] was the first attempt to establish self-attention mechanisms by combining transformer with U-Net and proved that transformers can be used as powerful encoders for medical image segmentation. A novel positional-encoding-free Transformer SegFormer [55] set new state-of-the-art in terms of efficiency and accuracy in publicly available semantic segmentation datasets and applied for instance in gland and nuclei segmentation [52]. This architecture remains promising also for semantic segmentation in medical applications due to positional-encoding-free encoder and lightweight MLP decoder.

2. Methods

2.1. Semantic segmentation

2.1.1. DeepLab

One of the challenges in semantic segmentation using standard CNNs is that as the input feature map goes through the network it gets smaller and the information about objects of a smaller scale can be lost. DeepLab family introduces atrous convolutions that extract more dense features which help to preserve the object's information. Compared to standard convolutions, atrous convolutions have an additional parameter, atrous rate, which is the stride at which the input is sampled (Figure 2.1 a). The atrous convolution is used in the last few blocks on features that were extracted from the backbone network (e.g. ResNet [56]).

One of the latest models in this family, DeepLabv3 [29], applies several parallel atrous convolutions with different atrous rates (Atrous Spatial Pyramid Pooling, or ASPP, Figure 2.1 b) to effectively capture multi-scale information. Image-level features, or image pooling, are also applied to incorporate global context information. Those are calculated by applying global average pooling on the last feature map of the backbone. After applying all the operations in parallel, the results of each operation along the channel is concatenated and 1×1 convolution is applied to get the output. The addition of atruos convolutions allows the enlargement of the field of view without increasing the size of the filtering kernel, therefore the computation time.

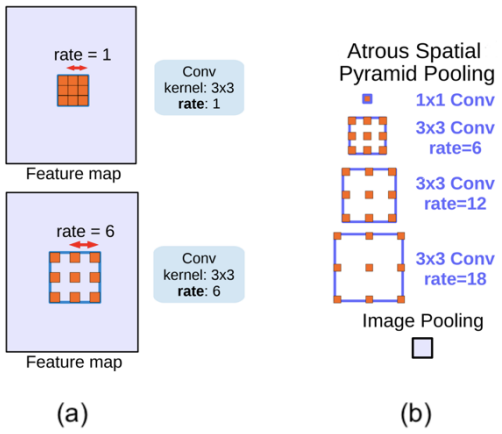
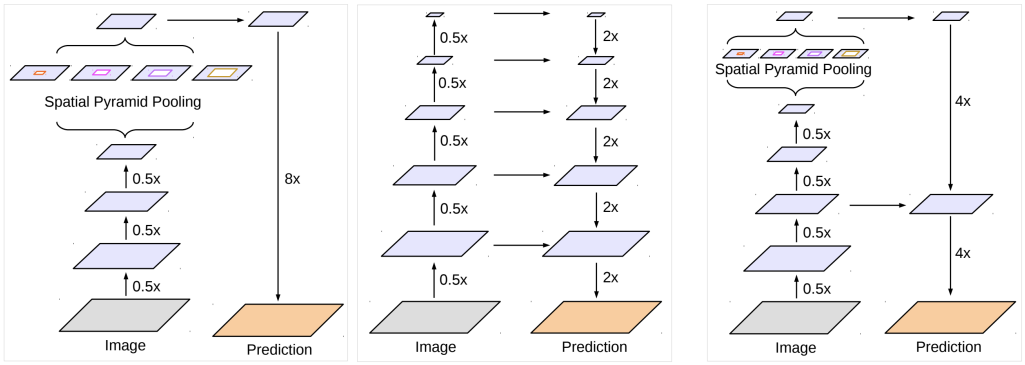


Figure 2.1.: (a) Atrous convolution, (b) ASPP augmented with Image Pooling (or Image-level features) [29]

DeepLabv3+

The reproduction of shape contours during semantic image segmentation remained difficult with DeepLabv3 [30]. DeepLabv3 bilinearly upsamples the logits both during training and evaluation (Fig. 2.2 a), hence the improvements were made to employ the encoder-decoder structure (Figure 2.2) to avoid using a naive decoder. DeepLabv3+ [30] adds the decoder



(a) Spatial Pyramid Pooling (b) Encoder-Decoder (c) Encoder-Decoder with Atrous Conv

Figure 2.2.: The spatial pyramid pooling module of DeepLabv3 (a), the encoder-decoder structure (b) and DeepLabv3+ adaptation (c) [30]

module on top of the encoder output, as shown in Fig. 2.3. In the decoder module, the 1×1 convolution reduces the channels of the low-level feature map from the encoder module which is then concatenated with the DeepLabv3 feature map and the 3×3 convolution obtains sharper segmentation results. As a result, DeepLabv3+ holds rich semantic information from the encoder module, while the detailed object boundaries are recovered by the decoder module and the spatial information is retrieved.

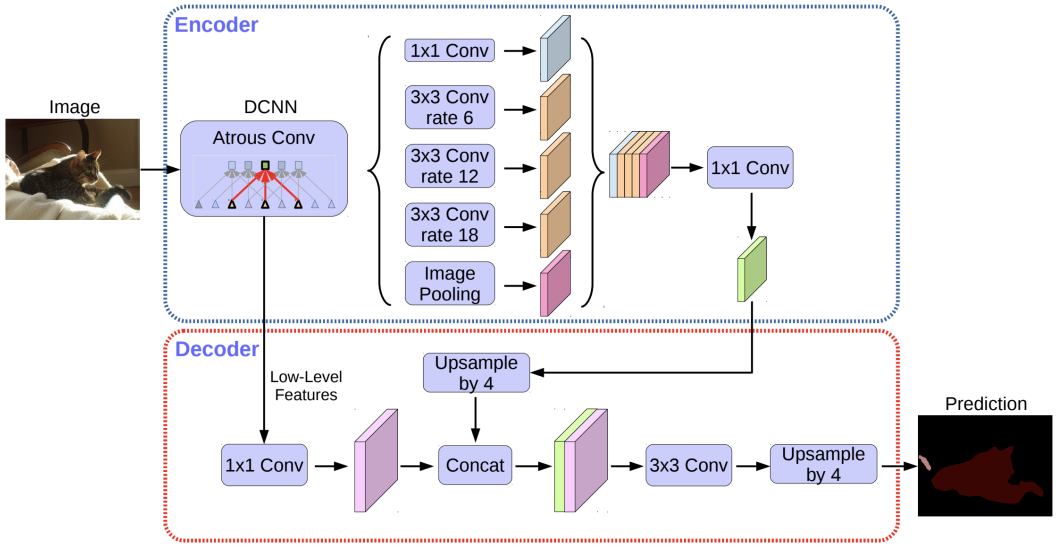


Figure 2.3.: DeepLabv3+ architecture. DeepLabv3 as encoder and proposed decoder structure for semantic image segmentation. [30]

2.1.2. Transformers

Transformers [57] were originally designed for the neural machine translation problem in NLP to capture long-range dependencies among words in a sentence. Their architecture converts 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.

The input and output are first embedded into an n-dimensional space. Since the network and the self-attention are permutation invariant, the positional encoding is added to create a representation of the position of the word in the sentence. The following modules consist mainly of Multi-Head Attention and Feed Forward layers. Encoder (Figure 2.4, left) and decoder (Figure 2.4, right) are composed of those modules that can be stacked on top of each other $N\times$ times.

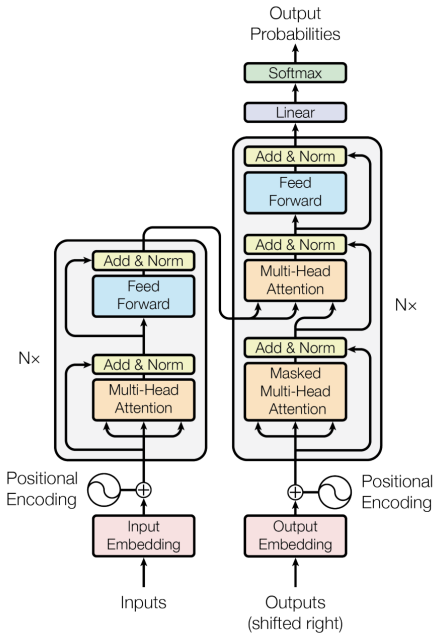


Figure 2.4.: Transformer model architecture. [57]

Self-attention is a sequence-to-sequence operation. It takes a weighted average over all the input vectors using dot product. Scaled Dot-Product Attention (Figure 2.5, left) can be described by the following equation:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (2.1)$$

where, in the context ot translation problem, 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. Q, K and V matrices are used to calculate the attention scores. These scores measure how much attention needs to be placed on words of the input sequence with respect to a word at a certain position. The scaling factor $\sqrt{d_k}$ is applied to avoid large values that after applying softmax would lead to vanishing gradients.

While Scaled Dot-Product Attention focuses on the whole sentence, Multi-Head Attention approaches different segments of the words. The word vectors are divided into a fixed number (number of heads) of parts, and then within Multi-Head Attention (Figure 2.5, right) the attention mechanism is repeated multiple times on those separate parts with linear projections of Q, K, and V. Since the Feed-Forward layer is expecting just one matrix, a vector for each word, the outputs are linearly concatenated. This allows the system to learn from different representations of Q, K, and V.

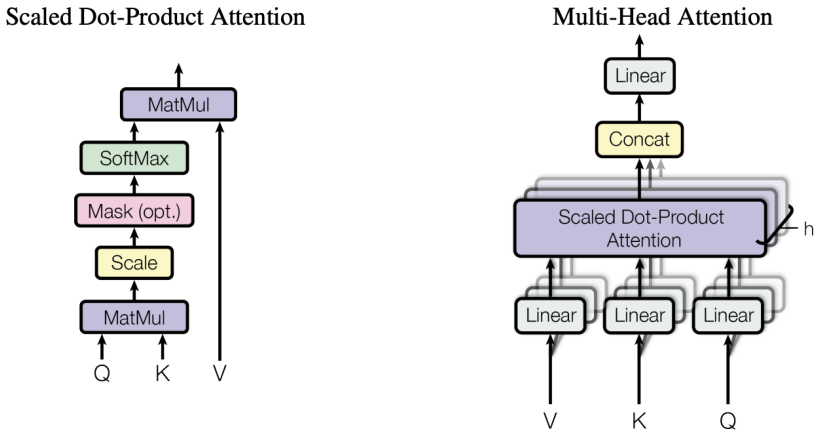


Figure 2.5.: Scaled Dot-Product Attention (left). Multi-Head Attention consists of several attention layers running in parallel (right). [57]

To add element-wise non-linearity transformation of incoming vectors, the transformer includes feed-forward networks. It processes the output from one attention layer so that it fits better for the next attention layer. Each of the layers in the encoder and decoder contains a fully connected feed-forward network, which is applied to each position separately and identically. These feed-forward layers can be described as a separate, identical linear transformation of each element from the given sequence.

Naive application of transformers 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) [58] is the first work to prove that a pure Transformer can achieve state-of-the-art performance in image classification. ViT converts the input image into a 1D series by cutting it into patches 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.

SegFormer

SegFormer [59] is a positional-encoding-free transformer based semantic segmentation method. As depicted in Figure 2.6, it consists of two main modules: a hierarchical Transformer encoder to generate high-resolution coarse features and low-resolution fine features; and a lightweight All-MLP decoder to fuse these multi-level features to produce the final semantic segmentation mask.

The $H \times W \times 3$ input image is divided into patches of size 4×4. Those patches are forwarded to the hierarchical Transformer encoder to obtain multi-level features at $\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}$

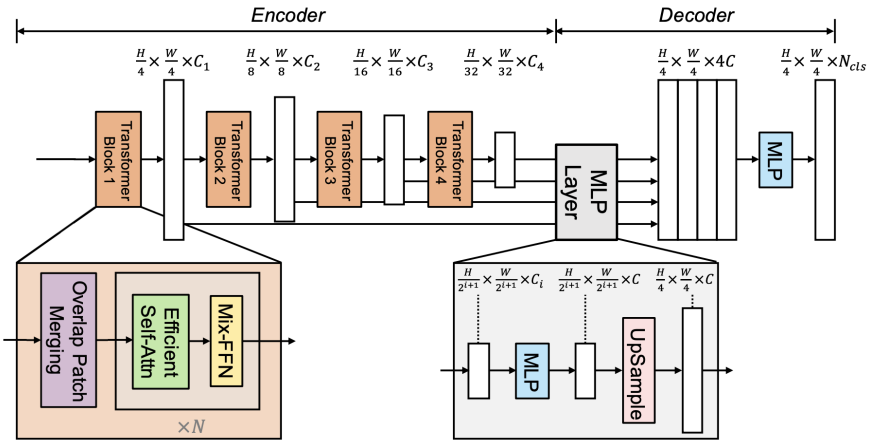


Figure 2.6.: SegFormer consists of two main modules: A hierarchical Transformer encoder to extract coarse and fine features; and a lightweight All-MLP decoder to directly fuse these multi-level features and predict the semantic segmentation mask. "FFN" indicates feed-forward network. (modified image [59] according to the official implementation)

resolution.

Overlapped Patch Merging produces features given an image patch and parameters: patch size K, stride between two adjacent patches S, and padding size P.

The main computation bottleneck of each transformer block in encoder is the self-attention layer. In SegFormer, before applying the self-attention according to the formula 2.1, the sequence K is reduced by ratio R:

$$\hat{K} = Reshape(\frac{N}{R}, C \cdot R)(K)$$

$$K = Linear(C \cdot R, C)(\hat{K})$$

where $N = H \times W$, $Reshape(\frac{N}{R}, C \cdot R)(K)$ refers to reshaping K to the the shape of $\frac{N}{R} \times (C \cdot R)$, and $Linear(C \cdot R, C)(\hat{K})$ refers to a linear layer taking a $(C \cdot R)$-dimensional tensor as input and generating a C-dimensional tensor as output. Therefore, the new K has dimensions $\frac{N}{R} \times C$.

Mix-FFN (feed-forward network) can be formulated as:

$$x_{out} = MLP(GELU(Conv3 \times 3(MLP(x_{in})))) + x_{in}$$

where x_{in} is the feature from the self-attention module.

The multi-level features are then passed to All-MLP decoder to predict the segmentation mask at $\frac{H}{4} \times \frac{W}{4} \times N_{cls}$ resolution, where N_{cls} is the number of classes.

The proposed All-MLP decoder consists of four main steps. First, multi-level features from the encoder go through an MLP layer to unify the channel dimension (2.2). Then, features are up-sampled to $\frac{1}{4}$th of the original image (2.3). Third, a MLP layer is adopted to fuse the

concatenated features (2.4). Finally, another MLP layer takes the fused feature to predict the segmentation mask (2.5).

$$\hat{F}_i = Linear(C_i, C)(F_i), \forall i \tag{2.2}$$

$$\hat{F}_i = Upsample(\frac{H}{4} \times \frac{W}{4})(\hat{F}_i), \forall i \tag{2.3}$$

$$F = Linear(4C, C)(Concat(\hat{F}_i)), \forall i \tag{2.4}$$

$$M = Linear(C, N_{cls})(F) \tag{2.5}$$

where F_i is the the feature and M is the final mask.

3. Data

3.1. Segmentation

The data comes from publicly released Tumor InfiltratinG lymphocytes in breast cancER (TiGER) challenge dataset containing digital pathology images of Her2 positive (Her2+) and Triple Negative (TNBC) breast cancer whole-slide images (WSI), regions of interest (ROIs) and manual annotations. More specifically, WSIROIS dataset was used for model training, validation, and testing (see Table 3.1). TiGER data, both at WSI and ROI level, was released at a spacing (pixel size) of approximately 0.5 µm/px, for more information please refer to the original challenge website[1].

Source	Tissue			TILs		
	#slides	#ROIs	median ROI size #pixels [k]	#slides	#ROIs	median ROI size #pixels [k]
TCGA-BRCA	151	151	4 983	124	1744	20
RUMC	26	81	1 312	26	81	1 312
JB	18	54	1 465	18	54	1 465
	195	286		168	1879	

Table 3.1.: TiGER data overview. Sources: Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA), Radboud University Medical Center (RUMC) and Jules Bordet Institute (JB). Tissue slides and ROIs refer to the segmentation images and annotations whereas TILs prefix specifies the data for TILs detection provided by the challenge.

The TiGER tissue annotations include eight labels that were reduced to three (see Table 3.2). The training masks were generated using available XML-files. In the provided mask images, in certain cases, regions not included in ROIs and non-annotated regions in ROIs where marked with the same label, which could not be directly used for training.

While for tissue segmentation the images and their masks could be used as directly extracted from the dataset, the data for TILs segmentation required some preprocessing. The TiGER fixed-size bounding box annotation for lymphocytes and plasma cells (see Table 3.3) was adapted for segmentation by transforming each bounding box into an annotation of the center pixel with a dilatation of three.

[1] `https://tiger.grand-challenge.org/Data/`

TiGER Tissue Label	Share	ID	new ID
Invasive tumor	0.283	1	1
In-situ tumor	0.029	3	1
Tumor-associated stroma	0.286	2	2
Inflamed stroma	0.096	6	2
Necrosis not in-situ	0.048	5	0
Healthy glands	0.0008	4	0
Background	0.231	0	0
Rest	0.026	7	0

Table 3.2.: Reduction of labels provided in TiGER challenge dataset. Resulting labels include three classes: Tumor (1), Stroma (2) and Rest (0) with shares of 0.312, 0.382 and 0.306. Shares were calculated by dividing the number of pixels belonging to some label by the number of the pixel in the current image and averaged over all images.

Source				Number of cells per ROI		
	#slides	#ROIs	#cells	min	max	median
TCGA-BRCA	124	1 744	19 115	0 (44.3%)	206	1
RUMC	26	81	4 728	0 (7.4%)	657	19
JB	18	54	5 523	0 (7.4%)	608	51.5
	168	1 879	29 366			

Table 3.3.: Data overview for TILs detection. Sources: Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA), Radboud University Medical Center (RUMC) and Jules Bordet Institute (JB). Number of cells here refers to the number of bounding boxes that were assigned for lymphocytes and plasma cells, further named TILs.

3.2. Survival Analysis

TiGER challenge aims to assess the prognostic significance of computer-generated TILs scores for predicting survival applying Cox proportional hazards model. The survival analysis is done internally, hence no corresponding data was released. The survival analysis within this thesis is done exclusively on publicly available TCGA-BRCA data. Where death (`vital_status` = 1) is considered as an event, and the time until the event or censoring is taken either from `days_to_death` (number of days to death from first diagnosis) or `days_to_followup` (number of days to last follow-up from first diagnosis).

`vital_status`	#cases	median age at diagnosis [years]	median time to event [months]
Dead	146	62	37.8
Alive	919	58	26.3
	1065	58	28.7

Table 3.4.: Survival data overview.

A. General Addenda

If there are several additions you want to add, but they do not fit into the thesis itself, they belong here.

A.1. Detailed Addition

Even sections are possible, but usually only used for several elements in, e.g. tables, images, etc.

B. Figures

B.1. Example 1

✓

B.2. Example 2

✗

List of Figures

List of Tables

Bibliography

[1] A. Shephard, M. Jahanifar, R. Wang, M. Dawood, S. Graham, K. Sidlauskas, S. A. Khurram, N. Rajpoot, and S. E. A. Raza. "TIAger: Tumor-Infiltrating Lymphocyte Scoring in Breast Cancer for the TiGER Challenge". In: *arXiv preprint arXiv:2206.11943* (2022).

[2] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos. "Image segmentation using deep learning: A survey". In: *IEEE transactions on pattern analysis and machine intelligence* (2021).

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

[4] H. Noh, S. Hong, and B. Han. "Learning deconvolution network for semantic segmentation". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1520–1528.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (2014).

[6] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. "Learning representations by back-propagating errors". In: *nature* 323.6088 (1986), pp. 533–536.

[7] C. Yu, J. Wang, C. Gao, G. Yu, C. Shen, and N. Sang. "Context prior for scene segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 12416–12425.

[8] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei. "Fully convolutional instance-aware semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2359–2367.

[9] A. BenTaieb and G. Hamarneh. "Topology aware fully convolutional networks for histology gland segmentation". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2016, pp. 460–468.

[10] J. Wang, J. D. MacKenzie, R. Ramachandran, and D. Z. Chen. "A deep learning approach for semantic segmentation in histology tissue images". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2016, pp. 176–184.

[11] V. A. Natarajan, M. S. Kumar, R. Patan, S. Kallam, and M. Y. N. Mohamed. "Segmentation of nuclei in histopathology images using fully convolutional deep neural architecture". In: *2020 International Conference on computing and information technology (ICCIT-1441)*. IEEE. 2020, pp. 1–7.

[12] M. Amgad, A. Sarkar, C. Srinivas, R. Redman, S. Ratra, C. J. Bechert, B. C. Calhoun, K. Mrazeck, U. Kurkure, L. A. Cooper, et al. "Joint region and nucleus segmentation for characterization of tumor infiltrating lymphocytes in breast cancer". In: *Medical Imaging 2019: Digital Pathology*. Vol. 10956. SPIE. 2019, pp. 129–136.

[13] D. A. Gutman, J. Cobb, D. Somanna, Y. Park, F. Wang, T. Kurc, J. H. Saltz, D. J. Brat, L. A. Cooper, and J. Kong. "Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data". In: *Journal of the American Medical Informatics Association* 20.6 (2013), pp. 1091–1098.

[14] M. Amgad, H. Elfandy, H. Hussein, L. A. Atteya, M. A. Elsebaie, L. S. Abo Elnasr, R. A. Sakr, H. S. Salem, A. F. Ismail, A. M. Saad, et al. "Structured crowdsourcing enables convolutional segmentation of histology images". In: *Bioinformatics* 35.18 (2019), pp. 3461–3467.

[15] V. Badrinarayanan, A. Kendall, and R. 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.

[16] A. Saood and I. Hatem. "COVID-19 lung CT image segmentation using deep learning methods: U-Net versus SegNet". In: *BMC Medical Imaging* 21.1 (2021), pp. 1–10.

[17] S. Almotairi, G. Kareem, M. Aouf, B. Almutairi, and M. A.-M. Salem. "Liver tumor segmentation in CT scans using modified SegNet". In: *Sensors* 20.5 (2020), p. 1516.

[18] A. B. Hamida, M. Devanne, J. Weber, C. Truntzer, V. Derangère, F. Ghiringhelli, G. Forestier, and C. Wemmert. "Deep learning for colon cancer histopathological images analysis". In: *Computers in Biology and Medicine* 136 (2021), p. 104730.

[19] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi. Cham: Springer International Publishing, 2015, pp. 234–241.

[20] A. Lagree, M. Mohebpour, N. Meti, K. Saednia, F.-I. Lu, E. Slodkowska, S. Gandhi, E. Rakovitch, A. Shenfield, A. Sadeghi-Naini, et al. "A review and comparison of breast tumor cell nuclei segmentation performances using deep convolutional neural networks". In: *Scientific Reports* 11.1 (2021), pp. 1–11.

[21] Z. Zeng, W. Xie, Y. Zhang, and Y. Lu. "RIC-Unet: An improved neural network based on Unet for nuclei segmentation in histology images". In: *Ieee Access* 7 (2019), pp. 21420–21428.

[22] H. Pinckaers and G. Litjens. "Neural ordinary differential equations for semantic segmentation of individual colon glands". In: *arXiv preprint arXiv:1910.10470* (2019).

[23] K. R. Oskal, M. Risdal, E. A. Janssen, E. S. Undersrud, and T. O. Gulsrud. "A U-net based approach to epidermal tissue segmentation in whole slide histopathological images". In: *SN Applied Sciences* 1.7 (2019), pp. 1–12.

[24] M. E. Bagdigen and G. Bilgin. "Cell segmentation in triple-negative breast cancer histopathological images using U-Net architecture". In: *2020 28th Signal Processing and Communications Applications Conference (SIU)*. IEEE. 2020, pp. 1–4.

[25] M. Van Rijthoven, M. Balkenhol, K. Siliņa, J. Van Der Laak, and F. Ciompi. "HookNet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images". In: *Medical Image Analysis* 68 (2021), p. 101890.

[26] A. Shah, A. Mehta, M. Wang, N. Neumann, T. McCalmont, and A. Zakhor. "Deep Learning Segmentation of Invasive Melanoma". In: *International Conference on Image Processing, Bordeaux, France*. 2022.

[27] Z. Meng, Z. Zhao, B. Li, F. Su, and L. Guo. "A cervical histopathology dataset for computer aided diagnosis of precancerous lesions". In: *IEEE Transactions on Medical Imaging* 40.6 (2021), pp. 1531–1541.

[28] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. "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.

[29] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. "Rethinking atrous convolution for semantic image segmentation". In: *arXiv preprint arXiv:1706.05587* (2017).

[30] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. "Encoder-decoder with atrous separable convolution for semantic image segmentation". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 801–818.

[31] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera. "Attention deeplabv3+: Multi-level context attention mechanism for skin lesion segmentation". In: *European conference on computer vision*. Springer. 2020, pp. 251–266.

[32] B. M. Priego-Torres, D. Sanchez-Morillo, M. A. Fernandez-Granero, and M. Garcia-Rojo. "Automatic segmentation of whole-slide H&E stained breast histopathology images using a deep convolutional neural network architecture". In: *Expert Systems With Applications* 151 (2020), p. 113387.

[33] H. Xu, Y. J. Cha, J. R. Clemenceau, J. Choi, S. H. Lee, J. Kang, and T. H. Hwang. "Spatial analysis of tumor-infiltrating lymphocytes in histological sections using deep learning techniques predicts survival in colorectal carcinoma". In: *The Journal of Pathology: Clinical Research* (2022).

[34] Y. Xie, J. Zhang, C. Shen, and Y. Xia. "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2021, pp. 171–180.

[35] M. Mirza and S. Osindero. "Conditional generative adversarial nets". In: *arXiv preprint arXiv:1411.1784* (2014).

[36] M. Rezaei, K. Harmuth, W. Gierke, T. Kellermeier, M. Fischer, H. Yang, and C. Meinel. "A conditional adversarial network for semantic segmentation of brain tumor". In: *International MICCAI Brainlesion Workshop*. Springer. 2017, pp. 241–252.

[37] F. Mahmood, D. Borders, R. J. Chen, G. N. McKay, K. J. Salimian, A. Baras, and N. J. Durr. "Deep adversarial training for multi-organ nuclei segmentation in histopathology images". In: *IEEE transactions on medical imaging* 39.11 (2019), pp. 3257–3267.

[38] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. "Image-to-image translation with conditional adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134.

[39] H. Tsuda and K. Hotta. "Cell Image Segmentation by Integrating Pix2pixs for Each Class". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2019.

[40] D. Popescu, M. Deaconu, L. Ichim, and G. Stamatescu. "Retinal blood vessel segmentation using pix2pix gan". In: *2021 29th Mediterranean Conference on Control and Automation (MED)*. IEEE. 2021, pp. 1173–1178.

[41] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232.

[42] M. Gadermayr, L. Gupta, V. Appel, P. Boor, B. M. Klinkhammer, and D. Merhof. "Generative adversarial networks for facilitating stain-independent supervised and unsupervised segmentation: a study on kidney histology". In: *IEEE transactions on medical imaging* 38.10 (2019), pp. 2293–2302.

[43] A. Kapil, T. Wiestler, S. Lanzmich, A. Silva, K. Steele, M. Rebelatto, G. Schmidt, and N. Brieu. "DASGAN–Joint Domain Adaptation and Segmentation for the Analysis of Epithelial Regions in Histopathology PD-L1 Images". In: *arXiv preprint arXiv:1906.11118* (2019).

[44] W. Li, J. Li, J. Polson, Z. Wang, W. Speier, and C. Arnold. "High resolution histopathology image generation and segmentation through adversarial training". In: *Medical Image Analysis* 75 (2022), p. 102251.

[45] F. Visin, M. Ciccone, A. Romero, K. Kastner, K. Cho, Y. Bengio, M. Matteucci, and A. Courville. "Reseg: A recurrent neural network-based model for semantic segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2016, pp. 41–48.

[46] W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki. "Scene labeling with lstm recurrent neural networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3547–3555.

[47] A. Chakravarty and J. Sivaswamy. "RACE-net: a recurrent neural network for biomedical image segmentation". In: *IEEE journal of biomedical and health informatics* 23.3 (2018), pp. 1151–1162.

[48] M. Saha and C. Chakraborty. "Her2Net: A deep framework for semantic segmentation and classification of cell membranes and nuclei in breast cancer evaluation". In: *IEEE Transactions on Image Processing* 27.5 (2018), pp. 2189–2200.

[49] C. Nguyen, Z. Asad, R. Deng, and Y. Huo. "Evaluating transformer-based semantic segmentation networks for pathological image segmentation". In: *Medical Imaging 2022: Image Processing*. Vol. 12032. SPIE. 2022, pp. 942–947.

[50] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. "Swin transformer: Hierarchical vision transformer using shifted windows". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 10012–10022.

[51] Z. Qian, K. Li, M. Lai, E. I. Chang, B. Wei, Y. Fan, Y. Xu, et al. "Transformer based multiple instance learning for weakly supervised histopathology image segmentation". In: *arXiv preprint arXiv:2205.08878* (2022).

[52] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang. "Ds-transunet: Dual swin transformer u-net for medical image segmentation". In: *IEEE Transactions on Instrumentation and Measurement* (2022).

[53] R. Strudel, R. Garcia, I. Laptev, and C. Schmid. "Segmenter: Transformer for semantic segmentation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 7262–7272.

[54] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou. "Transunet: Transformers make strong encoders for medical image segmentation". In: *arXiv preprint arXiv:2102.04306* (2021).

[55] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. *SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers*. 2021. DOI: 10.48550/ARXIV.2105.15203. URL: https://arxiv.org/abs/2105.15203.

[56] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[57] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[58] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[59] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. "SegFormer: Simple and efficient design for semantic segmentation with transformers". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 12077–12090.