Multi-scale fully convolutional neural networks for histopathology image segmentation: from nuclear aberrations to the global tissue architecture

Rüdiger Schmitz[1,2,3], Frederic Madesta[3,4], Maximilian Nielsen[3,4], René Werner[3,4⋆], and Thomas Rösch[1⋆]

[1] Department for Interdisciplinary Endoscopy, University Medical Center Hamburg-Eppendorf, Germany
[2] Institute of Anatomy and Experimental Morphology, University Medical Center Hamburg-Eppendorf, Germany
[3] DAISYlabs, Forschungszentrum Medizintechnik Hamburg, Germany
[4] Department of Computational Neuroscience, University Medical Center Hamburg-Eppendorf, Germany
r.schmitz@uke.de

Abstract. Histopathologic diagnosis is dependent on simultaneous information from a broad range of scales, ranging from nuclear aberrations ($\approx \mathcal{O}(0.1\,\mu m)$) over cellular structures ($\approx \mathcal{O}(10\,\mu m)$) to the global tissue architecture ($\gtrapprox \mathcal{O}(1\,mm)$). Bearing in mind which information is employed by human pathologists, we introduce and examine different strategies for the integration of multiple and widely separate spatial scales into common U-Net-based architectures. Based on this, we present a family of new, end-to-end trainable, multi-scale multi-encoder fully-convolutional neural networks for human modus operandi-inspired computer vision in histopathology.

This work is based upon and extends our PAIP 2019 challenge submission which has achieved top-10 results for both tasks.

Keywords: Deep learning · Computer Vision · Segmentation · Histopathology · Histology · Multi-scale · Fully-convolutional neural nets

1 Introduction

If the rumour is tumour, the issue is tissue. Histopathology is the gold standard and backbone of cancer diagnosis, providing important information for various stages of the treatment process.

This includes, amongst others, the assessment of resection specimens with respect to whether the resection margins are free of tumour cells and how close to the resection margins they reach, which can be viewed as a segmentation

⋆ Equal contribution.

task. In addition, individualised treatment planning strongly relies on a fine-grained grading of precursor and differently malignant cancer lesions, which can be understood as multi-class segmentation and/or classification problems.

Human pathologists meet these challenges with the help of highly specialised grading systems for all kinds of cancer and cancer precursors. Even though these grading systems vary for different kinds of cancer, most of them rely on a combination of features such as

- Local alterations of the nuclear inner structure
- Deformed and varying nuclear shapes or global alterations of the nuclei
- Altered nucleus to stroma ratio for individual tumour cells and
- Altered positions of the nuclei within their parent cells (e.g. not at the bottom anymore, as observed in many glandular tumours)
- Altered cellular shapes and
- Altered relative positions of neighbouring cells (e.g. not all on the same, single level but some stacked over others)
- Disorganised long-range order (e.g. atypical or deformed glandular shapes)
- Invasion, i.e. disrespecting global tissue order and borders between different layers

As can be seen from this (in-extensive) list, diagnosis and grading of malignancy inherently involves a range of different scales. These scales span a factor of about or more than a thousand, ranging from sub-nuclear features (which lie on a spatial scale of $\approx \mathcal{O}(0.1\,\mu\text{m})$) over nuclear, cellular ($\approx \mathcal{O}(10\,\mu\text{m})$), intercellular ($\approx \mathcal{O}(100\,\mu\text{m})$) to glandular and other higher organisational features ($\gtrapprox \mathcal{O}(1\,\text{mm})$).

The importance of the integration of information from different scales is reflected in how human pathologists approach these tasks: regions of interest are typically viewed at several different scales by turning the objective revolver of the microscopy back and forth again and again.

With its success in various computer vision tasks, deep learning methods have opened up a myriad of perspectives for computer-aided diagnoses (CADx) in histopathology. For the segmentation tasks, fully convolutional neural networks (FCNs, [9]) and, most prominently, U-Net-based architectures [12], have successfully been adopted [3, 4, 8]. Whilst in many of these cases, generic computer vision networks have been applied to histopathology images, there are some approaches for specialised histopathology architectures, [2, 7, 14], and training techniques [4, 15]. Some of these works address the question on how additional context can be provided to the network, however, they are confined to local, same-scale context, [2, 7], and/or sliding-window CNN techniques [2].

In this paper, we introduce a family of U-Net-based deep neural nets that are specifically designed for the extensive integration of largely different spatial scales.

2 Materials & Methods

2.1 Dataset

Throughout this work, we use the dataset provided for the PAIP 2019 challenge (part of the MICCAI 2019 Grand Challenge for Pathology) for training and evaluation. It is comprised of 50 de-identified whole-slide histopathology images from 50 patients that underwent resection for hepatocellular carcinoma (HCC) in three Korean hospitals (SNUH, SNUBH, SMG-SNU BMC) from 2005 to June 2018. The slides have been stained by hematoxylin and eosin and digitalised using an Aperio AT2 whole-slide scanner at $\times 20$ power.

Figure 1 shows an exemplary whole-slide image from the PAIP dataset. It is worth noting that all individual cases of the given dataset do indeed include cancerous regions.

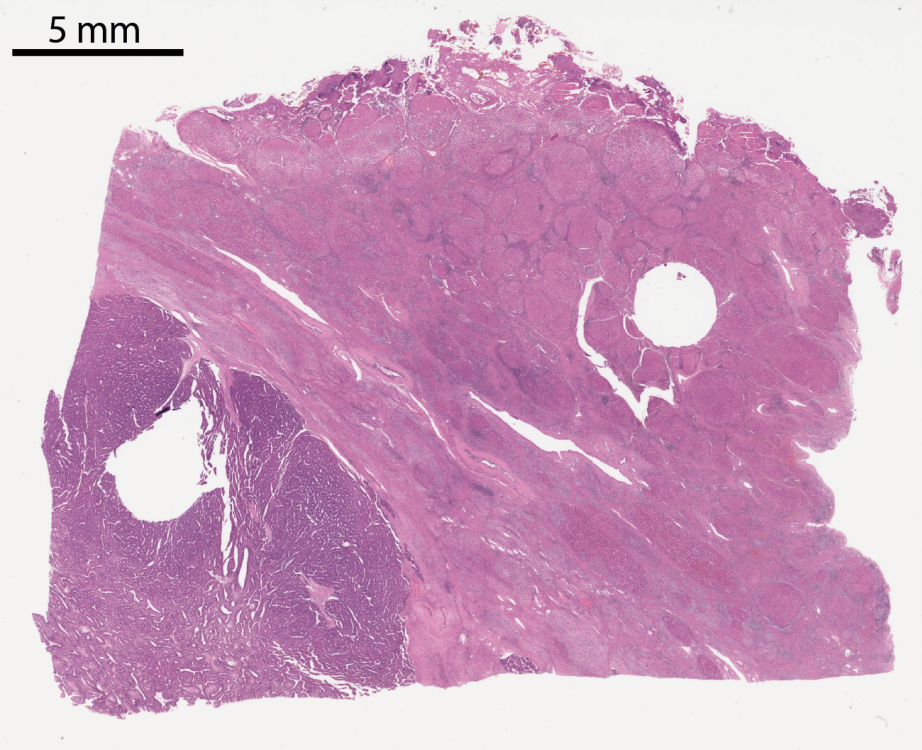

Fig. 1. An exemplary case from the PAIP dataset. The tissue sample is stained by hematoxylin and eosin. A hepatocalluar carcinoma (HCC) can easily be identified in the bottom left corner of the image. The so-called "pseudocapsule" around it is a typical, though not obligatory, feature of HCC.

All de-identified pathology images and annotations used in this research were prepared and provided by the Seoul National University Hospital under a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI18C0316).

2.2 Data Processing & Training

As compared to the annotations provided for the PAIP challenge, which comprise the classes "viable tumour" and "whole tumour" (i.e. including viable tumour cells, stroma, necrosis, ...), we added "overall tissue" as an another class for improved patch balancing and enhanced training stability. The "annotations" for the "overall tissue" class are created fully automatically before training, namely by thresholding of the original images with [R, G, B] $\leq$ [235, 210, 235] and subsequent application of morphological opening and closing operations. This threshold is the same as the one that has been used by the challenge organisers for cleaning of the whole and viable tumour annotations.

Images are standardised by Reinhard colour normalisation [11] with respect to the tissue regions and normalised to channel-wise zero average and unit variance.

The training process is split into (short pseudo-)epochs of 1920 patches each, with all patches randomly redrawn after each individual "epoch". (For this reason, the number of patches per epoch is arbitrary.) This allows for a close inspection of the training process, first, and, as compared to the use of a fixed number of pre-selected patches, it effectively reduces overfitting and/or resource requirements when facing enormously large images as in histopathology. For each individual "epoch", the patches are balanced with respect to both the individual cases in the training set and the available classes (background, overall tissue, whole tumour, viable tumour).

Our models are implemented using PyTorch [10]. Training is conducted by use of an in-house developed training pipeline for generic multi-modal medical image deep learning.

Optimisation is performed using Adam [6] employing a learning rate of 10^{-3} and a learning rate decay with $\gamma = 0.5$ every 57,600 iterations (30 epochs). During training, we employ online data augmentation including the following operations: rotation, flip along the horizontal axis, saturation and brightness transformation. All models are trained for 228,480 iterations (120 epochs), which is found to be well above what is needed for convergence for all of our models.

We choose the binary cross entropy (BCE) loss as a common and widely used loss function for this study. In order to balance for class inhomogeneities and draw attention to the tumour regions, we use class-weights of [0, 1, 2, 6] for background, overall tissue, whole tumour and viable tumour, respectively.

2.3 Evaluation & Metrics

We have chosen the task of viable tumour cell segmentation, such as presented in task 1 of the PAIP 2019 challenge, as an example task that may profit from extensive multi-scale integration. In addition to task 1 of that challenge, we also examine the predictions for the "whole tumour" class. The segmentation of the viable tumour cells, however, remains our primary interest, also due to the fact that the respective annotations appear to be very well consistent.

To this end, we train and evaluate all of the models described in the forthcoming on the PAIP 2019 challenge dataset using a 5-fold cross-validation (CV) split. We examine both the BCE loss on the validation dataset, as well as the Jaccard index which has been used for assessment of the PAIP 2019 task 1 and provides an intuitive parameter for segmentation quality.

Validation is conducted at the following steps (number of iterations): 0, 3,840, 7,680, 13,440, 19,200, 28,800, 38,400, 57,600, 76,800, 96,000, 115,200, 134,400, 153,600, 172,800, 192,000, 211,200, 228,400. At each such step, four 3072×3072 px-sized sub-images per case are evaluated (totalling 40 validation sub-images from 10 cases per split). The positions of these sub-images are randomly chosen once, such that for any validation case, all four classes are represented in at least one of the sub-images. The sub-image positions are kept fixed for each individual split, such that all models for a given split are evaluated with respect to the same sub-images.

We evaluate all models with respect to their underlying models, i.e. all models are compared with the baseline U-Net. In addition, msYI-Net and msY^2-Net are compared with the msY-Net they are built upon. As in all of these comparisons, one model can reproduce the other, a one-sided paired-sample t-test is used. In order to correct for multiple testing, we use the BenjaminiHochberg procedure.

2.4 Baseline method

Beyond the still popular sliding-window CNN-based techniques, U-Net-based FCN architectures form the de facto standard in histopathology image segmentation. Therefore, we have chosen a ResNet18-based U-Net [5] as baseline for our studies. An implementation of this architecture can be found in [1]. The ResNet18, which forms the encoder of this architecture (cf. figure 2), has been pre-trained on the ImageNet dataset [13]. Four our study, the baseline model is trained at full-resolution patches of the size 512×512.

2.5 msY-Net - Integrating local tissue architecture

Whole-slide images (WSI) are commonly saved as pyramid representations that underlie the common WSI formats such as `*.mrxs`, `*.czi`, `*.svs` and others. Therefore, when working with WSI images, retrieval of multiple scales and of very broad context of a given patch comes with a moderate overhead only.

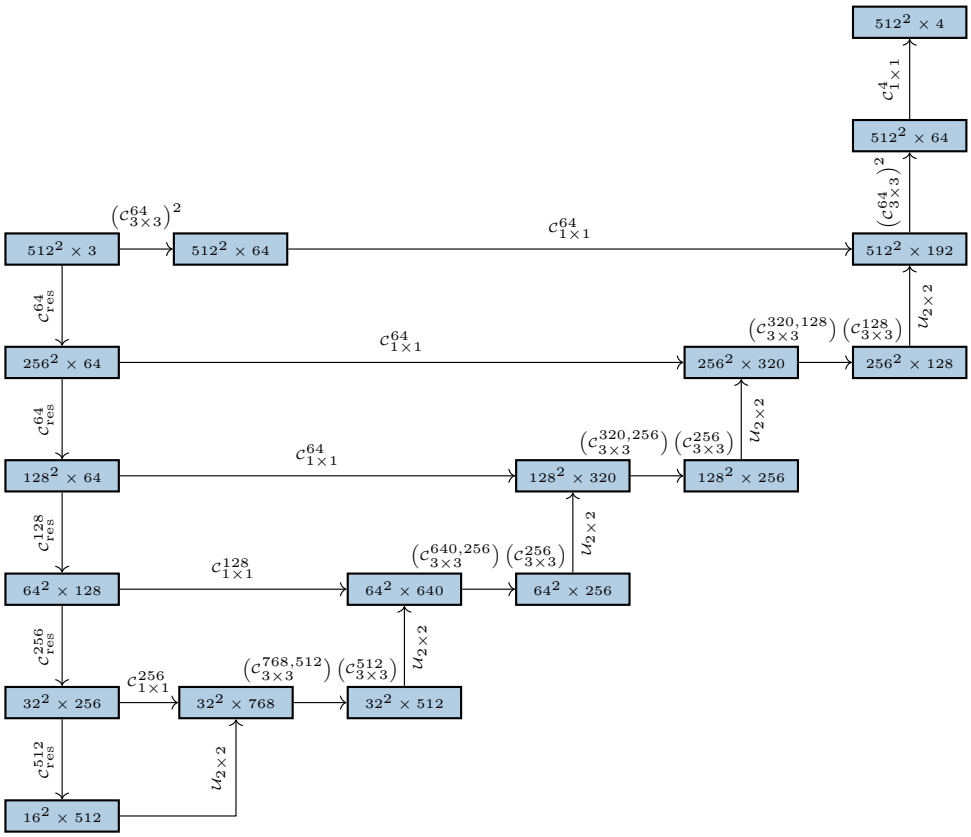

Fig. 2. ResNet18-based U-Net architecture (baseline model). For each block the spacial image shape as well as the number of channels are given. Here, $\mathcal{C}^{f_i,f_o}_{m\times n}$ denotes a single $m \times n$ convolution with f_i input and f_o output feature maps, followed by ReLU activation. $\left(\mathcal{C}^{f}_{m\times n}\right)^l$ shall represent l consecutive $m \times n$ convolutions with f feature input and output maps, each followed by a ReLU activation function. For the encoding part, the blocks ($\mathcal{C}^{f}_{\text{res}}$) of a ResNet18 are used where f denotes number of the respective output feature maps. Each individual block introduces a spatial downscaling by a factor of 2, either through max pooling or strided convolutions. The decoder uses $m \times n$ bilinear upsampling ($\mathcal{U}_{m\times n}$) to enlarge the spacial dimensions.

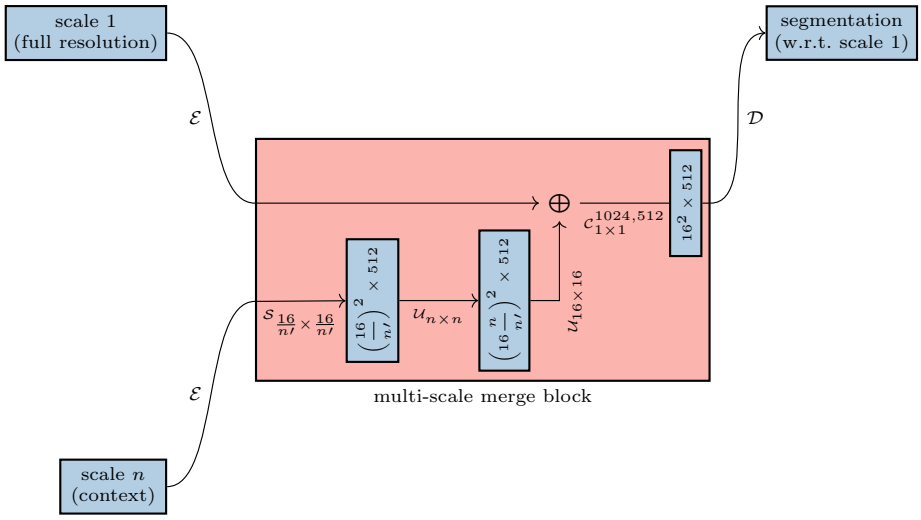

Fig. 3. Schematic illustration of the msY-Net with the multi-scale merge operation as its main building block. Here, $\mathcal{E}$, $\mathcal{D}$ are the encoder and decoder path of the baseline network (cf. figure 2), respectively. In order to spatially match the output of the full resolution encoder, a $n\prime \times n\prime$ centre cropping ($\mathcal{S}_{\frac{16}{n\prime}\times\frac{16}{n\prime}}$) of the n-times down-scaled context path is performed followed by $n \times n$ bilinear upsampling ($\mathcal{U}_{n\times n}$), where $n\prime = 4$ if $n = 4$ and $n\prime = 8$ if $n = 16$. For the case $n\prime \neq n$, another centre cropping with 16×16 is conducted. Both, now spatially consistent, paths are then merged by concatenation ($\oplus$). Finally, the number of feature maps is reduced to the original number by a 1×1 convolution. Cf. section 2.5 for a detailed description of the rationale.

In order to provide the network with local architectural information (cf. figure 4 and the considerations in section 1), we construct a multi-scale Y-Net (msY-Net) as a dual-scale dual-encoder network building upon the baseline Res-U-Net architecture, cf. figure 3.

The msY-Net is provided with two input patches of the scales 1 and 4 (which correspond to the inner two rectangles in figure 4). The former, the full-resolution patch, is fed into the standard U-Net architecture. The latter is passed through a separate but analogous encoder architecture ("context encoder") built from another ResNet18. As the skip connections in the U-Net are for helping the decoder re-localise the high-level features and only the full-resolution patch is the one that needs to be segmented, the context encoder does *not* have any skip connections to the decoder.

The information from the context encoder is provided to the underlying U-Net at its bottleneck by concatenation. At this point, however, it is crucial to concatenate both paths in a way such that their spatial scales agree. This means that just before concatenation, the context-encoder high-level feature maps are cropped to the region that spatially corresponds to the bottleneck feature maps from the full-resolution encoder. The cropped context-encoder high-level feature maps are then bilinearly upsampled by a factor of 4 and concatenated to the bottleneck feature maps from the full-resolution encoder. Before being fed to the decoder, the resulting $2 \times 512 = 1024$ feature maps are reduced to 512 feature maps by 1×1 convolution. This operation shall learn which of the feature maps from the two paths are relevant and how they need to be related to each other. These steps form the "multi-scale merge block" that is at the heart of the msY-Net and figure 3.

2.6 msY*I*-Net and msY^2-Net - Integration of larger context

In order to provide the model with large scale context information, cf. figure 4, we construct two models that have another large-context encoder for patches of the scale 16.

We examine two different ways of how to add this information to the underlying msY-Net: First, we add the large-context encoder in the same way the first context-encoder of the msY-Net has been added before, i.e. through another multi-scale merge block following the first multi-scale merge block. We refer to this model as msY^2-Net.

For the second model, we keep the large-context encoder separate, *parallel*, so to say, to the underlying msY-Net. The I in msYI-Net shall refer to the large-context encoder that parallels the underlying msY-Net. The large-context encoder consists of another ResNet18 that is used for classification of the overall content of the full-resolution patch. Hence, the network has two outputs, the segmentation of the full-resolution patch from its msY-Net part and the classification of the full-resolution patch contents from the large-context encoder, its I-part. In addition, the classification output from the I-part is concatenated to the msY-Net logits just before the final output of the network and merged by a

1×1 convolution, so that the msY-Net-part can use the large-context classification information for adjusting its final predictions. For training of this model, however, both the segmentation and the classification output are treated separately and fed to a segmentation and a classification loss function, respectively. The model is then optimised with respect to the arithmetic mean of the two losses. For both loss functions, binary cross entropy losses are used.

In the msY-Net and the msY^2-Net architectures, spatial correspondence between the full resolution encoder and the context encoder(s) is enforced in the multi-scale merge block (cf. figure 3). It should be noted that for the large-context encoder in the msYI-Net, there is no such requirement. Therefore, this model can, in principle, be fed with large-context patches of entirely arbitrary scales.

3 Results

3.1 Validation loss

All of the models that we have introduced above are trained and evaluated on fixed 5-fold splits of the dataset. The best validation loss that is individually reached is shown in table 3.1. For the msYI-Net, only the segmentation part of the loss is given.

	CV folds					**Mean**	
	0	**1**	**2**	**3**	**4**		
U-Net	0.111	0.104	0.186	0.136	0.151	0.138 ± 0.015	⋆
msY-Net	0.080	0.107	0.194	0.101	0.120	0.120 ± 0.020	
msYI-Net	0.099	0.080	0.232	0.079	0.088	0.116 ± 0.029	
msY^2-Net	0.077	0.088	0.166	0.126	0.106	**0.113 ± 0.016**	

Table 1. Best validation BCE loss per split and model. A ⋆ denotes statistical significance at the level of 0.05 (cf. section 2.3). Results significantly above those of the baseline model are marked bold.

3.2 Jaccard metrics

Table 3.2 shows the best Jaccard metric values that are reached for each individual model and CV split. As it can be seen, all the context-enhanced models perform significantly better than the baseline U-Net.

This also holds when the models that are best converged with respect to the BCE loss are evaluated for their mean Jaccard index on the viable and the whole

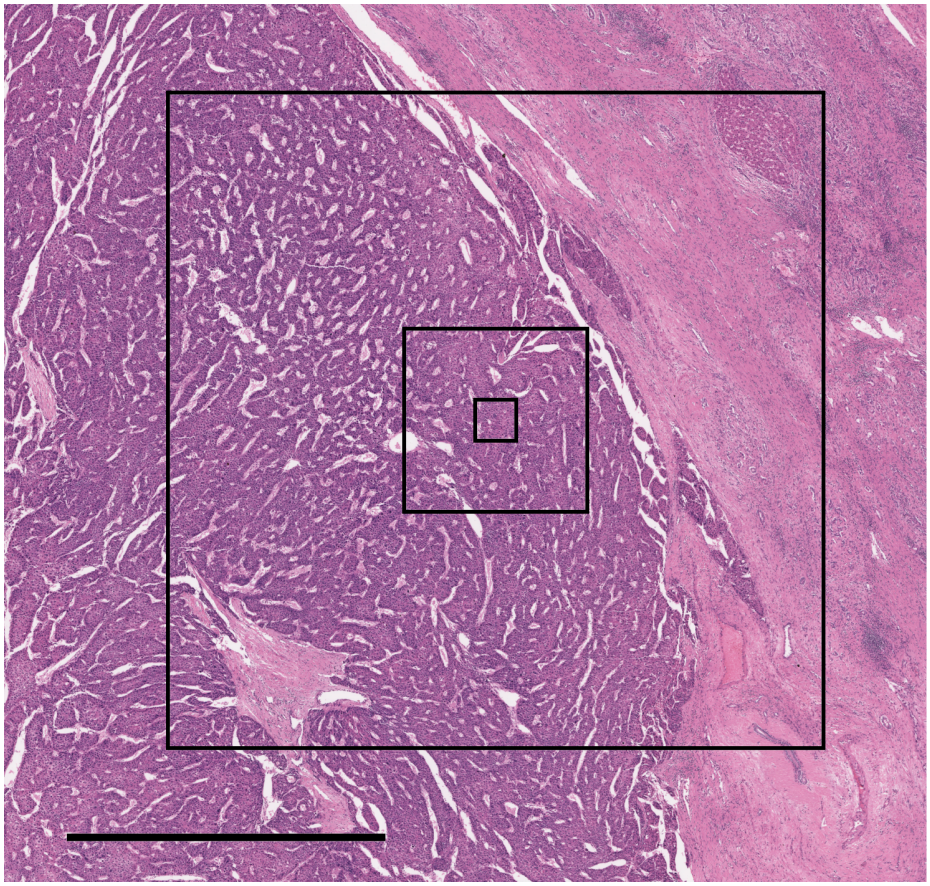

Fig. 4. Input patches to the different models under consideration, shown in an illustrative region of the whole-slide image depicted in figure 1. The innermost rectangular corresponds to the input with which a U-Net is provided that is trained on full-resolution patches of the size 512×512. The innermost plus the next larger rectangular is the input to the msY-Net model described in section 2.5. As can be seen, this scale contains information on how the cells are organised in strands and "trabeculae" (or whether the cells violate these patterns) that are impossible to deduce using the innermost patch alone. In this sense, it adds "architectural" information. The msYI-Net and the msY^2-Net can make use of the inputs represented by all the three rectancles. The outermost rectangle contains information on the large-scale organisation of the tissue, such as the pseudocapsule, for instance, that is typical for hepatocellular carcinoma. The scale bar is 2 mm.

tumour classes, cf. table 3.2. In addition, these results suggest that there may be an additional benefit from the large-context encoder as compared to the msY-Net context encoder alone. Though, this difference is statistically significant only for one of the two three-arm models, namely the msY^2-Net.

Further studies and results in favour of the understanding that there is a benefit from the context-encoder over the baseline U-Net and that there might be an additional benefit from another large-context encoder can be found in the Supplementary Materials.

	CV folds					**Mean**
	0	**1**	**2**	**3**	**4**	
U-Net	0.852	0.815	0.714	0.767	0.749	0.776 ± 0.024
msY-Net	0.924	0.849	0.737	0.840	0.854	**0.841 ± 0.030**
msY*I*-Net	0.880	0.864	0.815	0.889	0.908	**0.871 ± 0.016**
msY^2-Net	0.927	0.880	0.751	0.845	0.808	**0.842 ± 0.030**

Table 2. Best Jaccard index for the viable tumour area. A $\star$ ($\star\star$) denotes statistical significance at the level of 0.05 (0.005). Results significantly above those of the baseline model are marked bold.

	CV folds					**Mean**
	0	**1**	**2**	**3**	**4**	
U-Net	0.760	0.690	0.607	0.700	0.689	0.689 ± 0.024
msY-Net	0.883	0.752	0.655	0.755	0.709	**0.751 ± 0.038**
msY*I*-Net	0.819	0.807	0.674	0.902	0.780	**0.797 ± 0.037**
msY^2-Net	0.878	0.819	0.724	0.804	0.751	**0.795 ± 0.027**

Table 3. Class-average of the Jaccard index for whole and viable tumour, at best validation loss. A $\star$ ($\star\star$) denotes statistical significance at the level of 0.05 (0.005). Results significantly above those of the baseline model are marked bold.

4 Conclusion

Using the segmentation of hepatocellular carcinoma in hematoxylin-eosin stained whole-slide images of the liver as an example task, our results show that the

extensive integration of widely different spatial scales, as a "mimicry" of how humans approach analogous tasks, can benefit U-Net-based architectures. As the detailed structure of the encoder and the decoder are left entirely untouched, the approach presented herein can be seamlessly integrated into various encoder-decoder models.

This study has important limitations: First of all, it is restricted to one particular task and organ and disease entity. It remains to future work to examine how these findings generalise to other tasks, organs and types of cancer or even other diseases. In addition, further work is needed to identify the optimal approach for connecting (particularly) the large-context encoder to the base network.

This proof-of-principle study advocates the benefit of extensive multi-scaling in histopathology deep learning. It goes without saying that future research is likely to optimise the models presented herein much further.

Acknowledgments

RS gratefully acknowledges funding by the Studienstiftung des deutschen Volkes and the Günther Elin Krempel foundation. The authors would like to thank NVIDIA for the donation of a graphics card under the GPU Grant Program. In addition, the authors are grateful towards Hinnerk Stüben for his excellent technical support. This study was partially supported by an unrestricted grant from Olympus Co Hamburg, Germany, and by the Forschungszentrum Medizintechnik Hamburg (02fmthh2017).

Conflict of interest

RS, FM and MN declare that there are no conflicts to disclose. TR receives study support for various projects from Olympus Co Hamburg, Germany, but declares that there is no conflict to disclose with regards to this project. RW received funding from Siemens Healthcare, Erlangen, Germany, but declares that there is no conflict to disclose regarding this project.

References

1. Simple pytorch implementations of u-net/fullyconvnet (fcn) for image segmentation. `https://github.com/usuyama/pytorch-unet`, accessed: 2019-09-19
2. Bejnordi, B.E., Zuidhof, G., Balkenhol, M., Hermsen, M., Bult, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J.: Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images. Journal of Medical Imaging **4**(04), 1 (Dec 2017). https://doi.org/10.1117/1.JMI.4.4.044504, 00021
3. Bulten, W., Bndi, P., Hoven, J., Loo, R.v.d., Lotz, J., Weiss, N., Laak, J.v.d., Ginneken, B.v., Hulsbergen-van de Kaa, C., Litjens, G.: Epithelium segmentation using deep learning in H&E-stained prostate specimens with immunohistochemistry as reference standard. Scientific Reports **9**(1) (Dec 2019). https://doi.org/10.1038/s41598-018-37257-4, 00000

4. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine (Jul 2019). https://doi.org/10.1038/s41591-019-0508-1, 00000
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs] (Dec 2015)
6. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
7. Li, J., Sarma, K.V., Ho, K.C., Gertych, A., Knudsen, B.S., Arnold, C.W.: A Multi-scale U-Net for Semantic Segmentation of Histological Images from Radical Prostatectomies. AMIA Annu Symp Proc. pp. 1140–1148 (2017), 00002
8. Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G.E., Kohlberger, T., Boyko, A., Venugopalan, S., Timofeev, A., Nelson, P.Q., Corrado, G.S., Hipp, J.D., Peng, L., Stumpe, M.C.: Detecting Cancer Metastases on Gigapixel Pathology Images. arXiv:1703.02442 p. 13
9. Long, J., Shelhamer, E., Darrell, T.: Fully Convolutional Networks for Semantic Segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 3431–3440 (2015), 04357
10. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
11. Reinhard, E., Ashikhmin, M., Gooch, B., Shirley, P.: Color Transfer between Images. IEEE Computer Graphics and Applications p. 8 (2001), 02092
12. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention MICCAI 2015 **9351**, 234–241 (2015)
13. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision **115**(3), 211–252 (Dec 2015). https://doi.org/10.1007/s11263-015-0816-y, 06647
14. Vu, Q.D., Kwak, J.T.: A Dense Multi-Path Decoder for Tissue Segmentation in Histopathology Images. Computer Methods and Programs in Biomedicine (Mar 2019). https://doi.org/10.1016/j.cmpb.2019.03.007, 00000
15. Wang, S., Zhu, Y., Yu, L., Chen, H., Lin, H., Wan, X., Fan, X., Heng, P.A.: RMDL: Recalibrated Multi-instance Deep Learning for Whole Slide Gastric Image Classification. Medical Image Analysis p. 101549 (Aug 2019). https://doi.org/10.1016/j.media.2019.101549, 00000

Supplementary Materials

	CV folds					Mean	
	0	**1**	**2**	**3**	**4**		
U-Net	0.852	0.693	0.680	0.707	0.719	0.730 ± 0.031	⋆
msY-Net	0.920	0.849	0.704	0.772	0.808	**0.811 ± 0.036**	
msY*I*-Net	0.847	0.850	0.671	0.889	0.848	0.821 ± 0.038	
msY^2-Net	0.927	0.864	0.751	0.822	0.808	**0.834 ± 0.029**	⋆

Table 4. Viable tumour area Jaccard index at best validation BCE loss. A ⋆ denotes statistical significance at the level of 0.05. Results significantly above those of the baseline model are marked bold.

	CV folds					Mean		
	0	**1**	**2**	**3**	**4**			
U-Net	0.700	0.757	0.612	0.694	0.675	0.688 ± 0.023	⋆	
msY-Net	0.845	0.692	0.687	0.754	0.701	0.736 ± 0.030		
msY*I*-Net	0.831	0.840	0.811	0.916	0.819	**0.843 ± 0.019**	⋆	⋆
msY^2-Net	0.823	0.778	0.696	0.830	0.695	**0.766 ± 0.030**		

Table 5. Best Jaccard index for the whole tumour area (i.e. including viable tumour cells, stroma, necrosis, ...). A ⋆ denotes statistical significance at the level of 0.05. Results significantly above those of the baseline model are marked bold.

	CV folds					Mean	
	0	**1**	**2**	**3**	**4**		
U-Net	0.667	0.687	0.535	0.694	0.659	0.648 ± 0.029	⋆
msY-Net	0.845	0.655	0.605	0.739	0.611	0.691 ± 0.030	⋆
msY*I*-Net	0.791	0.763	0.678	0.916	0.711	**0.771 ± 0.041**	⋆
msY^2-Net	0.830	0.775	0.696	0.787	0.693	**0.756 ± 0.027**	

Table 6. Whole tumour area Jaccard index at best validation BCE loss.A ⋆ denotes statistical significance at the level of 0.05. Results significantly above those of the baseline model are marked bold.