

Foveation for Segmentation of Mega-pixel Histology Images

Chen Jin¹, Ryutaro Tanno^{1,2}
Moucheng Xu¹, Thomy Mertzani¹, Daniel C. Alexander¹

¹ Centre for Medical Image Computing, Department of Computer Science,
University College London, UK

² Healthcare Intelligence, Microsoft Research Cambridge, UK

Abstract. Segmenting histology images is challenging because of the sheer size of the images with millions or even billions of pixels. Typical solutions pre-process each histology image by dividing it into patches of fixed size and/or down-sampling to meet memory constraints. Such operations incur information loss in the field-of-view (FoV) (i.e., spatial coverage) and the image resolution. The impact on segmentation performance is, however, as yet understudied. In this work, we first show under typical memory constraints (e.g., 10G GPU memory) that the trade-off between FoV and resolution considerably affects segmentation performance on histology images, and its influence also varies spatially according to local patterns in different areas (see Fig. 1). Based on this insight, we then introduce *foveation module*, a learnable “dataloader” which, for a given histology image, adaptively chooses the appropriate configuration (FoV/resolution trade-off) of the input patch to feed to the downstream segmentation model at each spatial location (Fig. 1). The foveation module is jointly trained with the segmentation network to maximise the task performance. We demonstrate, on the Gleason2019 challenge dataset for histopathology segmentation, that the foveation module improves segmentation performance over the cases trained with patches of fixed FoV/resolution trade-off. Moreover, our model achieves better segmentation accuracy for the two most clinically important and ambiguous classes (Gleason Grade 3 and 4) than the top performers in the challenge by 13.1% and 7.5%, and improves on the average performance of 6 human experts by 6.5% and 7.5%.

1 Introduction

The histology images are ultra-high resolution microscope images from Hematoxylin and Eosin-stained biopsy, which form the primary source of information for cancer detection, grading and treatment planning. However manual analysis of histology images is expensive and prone to false negative detection due to their enormous size (up to 100,000² pixels), motivating the development of accurate automated methods. Deep learning (DL) based approaches have been recently adopted to improve the segmentation of high-resolution images in recent years. However, modern DL methods cannot operate on mega(giga)-pixel histology images, given the limited GPU memory constraints.

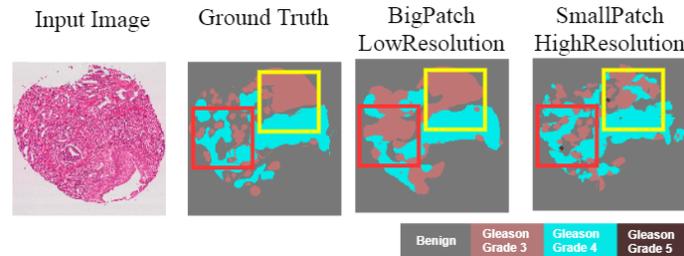


Fig. 1: A 4-class Gleason grade segmentation example under two trade-off configurations: “BigPatch-LowResolution” with patch size of 5000^2 pixels and downsampling rate of 0.2 and “SmallPatch-HighResolution” with patch size of 2000^2 pixels and downsampling rate of 0.5.

To mitigate this issue, histology images are typically dissected into smaller patches and/or down-sampled to fit into the available GPU memory [1]. To exploit all the available GPU memory thus requires one to trade off field of view (FoV), i.e. spatial extent of context, against resolution, i.e. level of image detail. Tuning this trade-off exhaustively is expensive [2], and as a result, it is commonly set by crude developer intuition.

A considerable amount of work has attempted to alleviate the issue of subjectivity in tuning this trade-off by learning to merge multi-scale information, in both medical imaging [3] and computer vision [4,5]. These works optimise model performance by exploiting the information from multi-scale sources. Specifically, they learn feature representations from multiple parallel networks, then aggregate learnt multiple scale representations before making the final prediction. DeepMedic [3] is one pioneering example in this category, having two parallel networks and applied to brain lesion segmentation. Another is [6], which is designed to work specifically with histopathology images. In general computer vision, the authors of [5] boost the learning efficiency of multi-scale information by enforcing global and local streams to interactively exchange information with each other. On the other hand, Chen *et al.*[4] perform multi-scale feature aggregation via an attention mechanism, which weights the prediction score from multi-scale parallel networks. However, these approaches: 1) construct multiple parallel networks, which can be computationally expensive; 2) typically use limited (2 or 3) manually selected scales; 3) need to rely on specific choices of neural network architecture.

Our contributions: In this work, we first demonstrate empirically on a public histology segmentation dataset that the choice of the input patch configuration (i.e., FoV/resolution trade-off) considerably influences the segmentation performance on different classes. Secondly, motivated by this finding, we then propose *foveation module*, a data-driven “data loader” that learns to provide the segmentation network with the most informative patch configuration for each location in a ultra-high-resolution image. The foveation module can be trained jointly in an end-to-end fashion with the segmentation network to optimise the task performance. The inspiration for our method roots from the ways in which pathologist segment high-resolution images —

starting from a low-resolution bird’s-eye view of the whole image ³, the annotators navigate their gaze through different locations and zoom in to the right extent to collect both local and contextual information. The magnification scale is controlled by, what is called, *foveation* (i.e., the process of adjusting the focal length of the eye, the distance between the lens and fovea). We also note that our work bears some similarity with the recent approach proposed by Katharopoulos *et al.*[7] in which attention weights are learned over the mega-pixel histology image to sample patches from a small informative sub-locations for the downstream classification task. However, our work differs from theirs in that we aim to select the best patch configuration at every spatial location for the downstream segmentation task, while their method tries to select the most informative subset of all spatial locations for the classification task. We evaluate the benefits of our foveation module on the Gleason2019 challenge dataset, where we show it boosts segmentation performance with little extra computational cost.

2 Methods

In this section, we first perform empirical analysis to illustrate the impact of the patch FoV/resolution trade-off on the histology segmentation performance and its spatial variation across the image. Motivated by this finding, we then propose *foveation module*, a module that learns to provide the segmentation network with the most informative patch configuration for each location in an ultra-high-resolution histology image.

2.1 Effects of Patch Configuration on Histology Segmentation

The first part of our work is a comprehensive empirical analysis, investigating a key question: “How does the FoV/resolution of training input patches affect the final segmentation performance?”. To this end, we validate our method on a multi-class segmentation dataset: the Gleason2019 Challenge ⁴ which contains 322 histology images of average size $\approx 5000 \times 5000$ pixels. Each image is labelled by a subset of 6 annotators (all experts). Each pixel is labelled into one from four classes (Benign, Gleason Grade 3,4,5). Preprocessing for empirical analysis: A subset of 298 training examples have been used in the first part of our work. We fuse 6 annotations into 1 using pixel-level probabilistic analysis by STAPLE [8]. Each image is paired with 1 STAPLE fused annotation as gold standard. Pre-processing for foveation experiments: we extract central part of input images \mathbf{x} of size 4400^2 from the original histology images to ensure the constant size of input image. We randomly split 298 images into a training subset (268) and a testing subset (30).

Then, we perform three sets of experiments: 1) the first investigates the impact of FoV only; 2) the second investigates the impact of resolution; and 3) the last studies the combination of the FoV and resolution (ie the FoV/resolution trade-off). To study the impact of the FoV, we divide the original 5000^2 images into 256^2 , 512^2 , 768^2 patches, respectively. It is noted all sampled patches share the same original resolution. In the

³ Screen display or human vision typically have lower resolutions than that of the ultra-high resolution images of interest in this work.

⁴ <https://gleason2019.grand-challenge.org>

second set of experiments, we downsample the original 5000^2 images using four different downsampling rates of 0.03, 0.06, 0.12 and 0.22, respectively. To examine the trade-off between FoV and resolution, we first fix our memory limit at 10G. Then we divide the original images into 1100^2 , 2000^2 , 3000^2 , 4000^2 , 5000^2 , respectively, to create five sets of patches, which share the same resolution. Afterwards, we further perform downsampling on each set of the five sets of patches, with different downsampling rates all to an identical size of 1100^2 . We illustrate the results of the total three sets of experiments in blue curves in Fig. 2(a). In Fig. 2(a), the red star represents the trade-off which has the best mean performance of the 4 classes across all of the experiments. Meanwhile, the trade-offs of best performances for each class across all of the experiments, are also highlighted in different shaped marks (see Legends in Fig. 2(a)). It is clear that there is no single “best fit” sampling size for best mean performance and best performances for each class, simultaneously. We also show visual results of the best segmentation for each class in Fig. 2(b). As visually illustrated in Fig. 2(b), each trade-off is only optimal for each pattern at each spatial location. This shows that the standard sampling strategy is not optimal overall. This motivates our novel Foveation module to learn to sample the most optimal training patches, which will be introduced in the following section.

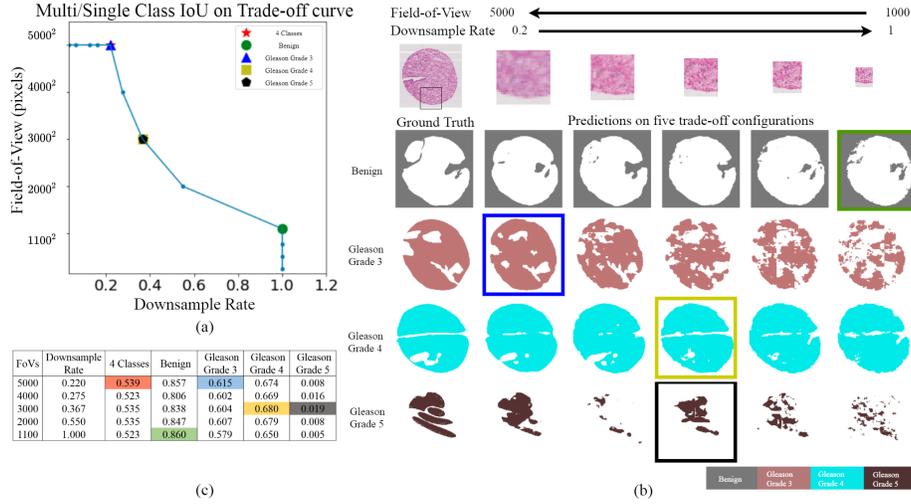


Fig. 2: Quantitative (left) and qualitative (right) evidence of different optimal FoV/resolution trade-offs for different classes. (a): The best patch configuration (i.e., FoV/resolution trade-off) for segmentation performance overall and for individual classes are highlighted. (b): A visual illustration of the class-wise variation of the best patch configuration. Here the size of FoV decreases from left to right and Downsampling rate increases from left to right. In each row, the best segmentation result among the five trade-off configurations is highlighted, which corresponds to the trade-off configuration in Left. (c): The corresponding segmentation performance measured in *intersection over union* (IoU) are reported.

2.2 Foveation Module

The inspiration for our method is to combine information from multiple scales and locations, as the fovea of the eye does that automatically adjusting the focal length according to different regions of interests via simply a glimpse over the scene. Our method consists of two components (see Fig.3 for a schematic); (1) *Foveation module* that takes a low-resolution version of a mega-pixel input image and generates importance weights over a set of patches with varying spatial FoV/resolution at different pixel locations, (2) *Segmentation network* that processes the input patches based on the outputs of the foveation module, and estimates the corresponding segmentation probabilities. This segmentation network can be any existing models.

For each mega-pixel image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ where H, W, C denote the height, width and channels respectively, we compute its lower resolution version $\mathbf{x}_r \in \mathbb{R}^{h \times w \times C}$. The resolution of \mathbf{x}_r is empirically determined based on performance. We also define a ‘‘patch-extractor’’ function $PE(\mathbf{x}, i) = \{\mathbf{p}_1^{(i)}(\mathbf{x}), \dots, \mathbf{p}_D^{(i)}(\mathbf{x})\}$ that extracts a set of D patches of varying field-of-view/resolution (but the same number of pixels to tackle the FoV-resolution trade-off when the input patch size is fixed due to limited memory) from the full resolution image \mathbf{x} centered at the corresponding i^{th} pixel in \mathbf{x}_r (see Fig.3 for a set of examples). *Foveation module*, F_θ , parametrised by θ , takes the low-resolution image \mathbf{x}_r as the input and generates the probability distributions $F_\theta(\mathbf{x}_r) \in [0, 1]^{h \times w \times D}$ over patches $PE(\mathbf{x}, i)$ at respective spatial locations $i \in \{1, \dots, wh\}$ in \mathbf{x}_r .

Based on the outputs of the foveation module, at each location i , we compute the input patch by taking the weighted average of the extracted patches of varying resolutions/sizes:

$$\mathbf{p}^{(i)}(\mathbf{x}) := \sum_{d=1}^D f_d^{(i)}(\mathbf{x}_r) \cdot \mathbf{p}_d^{(i)}(\mathbf{x}) \quad (1)$$

where $f_d^{(i)}(\mathbf{x}_r)$ denotes the value of $F_\theta(\mathbf{x}_r)$ at i^{th} pixel, and quantifies the ‘‘importance’’ of the d^{th} patch at that location. The importance weights $[f_1^{(i)}(\mathbf{x}_r), \dots, f_D^{(i)}(\mathbf{x}_r)]$ sum up to one. The weighted average of the multiple patches based on the estimated probabilities ensures the full differentiability of the objective function with respect to θ . This approach can be viewed as the mean approximation of the ‘‘stochastic hard’’ attention employed in [7], similar to the approach in [9] and the ‘‘deterministic soft’’ attention in [10]. We then subsequently feed this input patch to *segmentation network*, $S_\phi(\mathbf{p}^{(i)}(\mathbf{x}))$, parametrised by ϕ to estimate the segmentation probabilities within the spatial extent covered by the patch with the smallest field of view in $PE(\mathbf{x}, i)$. During training, the parameters $\{\theta, \phi\}$ of both the foveation module and the segmentation network are jointly learned to minimise the segmentation specific loss function (e.g., cross entropy) by performing stochastic gradient descent. We extract patches at different locations and process them separately, which can be computationally expensive. For computational efficiency, for each mega-pixel image \mathbf{x} , we randomly select a subset of pixels (locations) from its low resolution counterpart \mathbf{x}_r , compute the corresponding input patches according to eq. (1), feed them to the segmentation network and compute the losses. At inference time, we segment the whole mega-pixel image \mathbf{x} by aggregating predictions $S_\phi(\mathbf{p}^{(i)}(\mathbf{x}))$ at different locations i . We open-source the code at <https://github.com/lxasqj/Foveation-for-Segmentation-of-Mega-pixel-Histology-Images>.

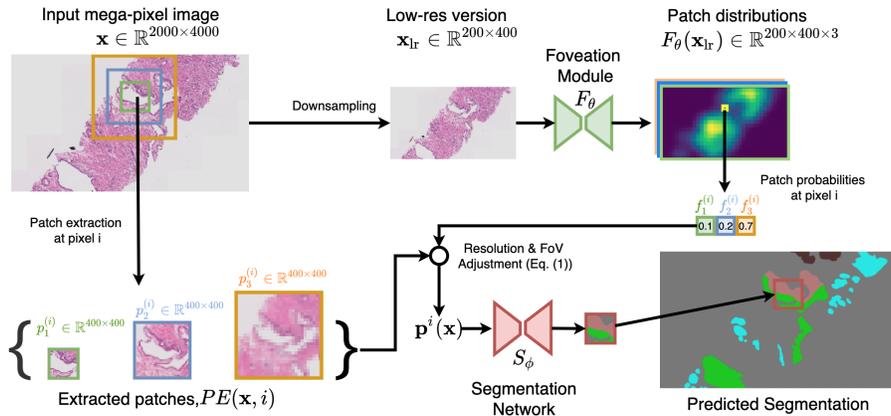


Fig. 3: Architecture schematic.

3 Network Architectures and Implementation Details

Training: For all experiments, we employ the same training scheme unless otherwise stated. We optimize parameters using Adam [11] with initial learning rate of $2e10^{-5}$ and $\beta=0.9$ and train for 50 epochs. We apply batches size of 2. For patch extractor we extract a set of 5 patches at field-of-view $\{1100^2, 2000^2, 2900^2, 3800^2, 4400^2\}$ pixels and apply a down-sampling factor of $\{1, 0.55, 0.38, 0.29, 0.25\}$ to have identical 1100^2 pixels in extracted patches.

Architectures: The Foveation module was defined as a light weighted CNN architecture (0.1M parameter) comprised of 3 convolution layers, each with 33 kernels followed by BatchNorm and Relu. The number of kernels in each respective layer is $\{40, 40, 5\}$. A softmax layer is added at the end. All convolution layers are initialised following He initialization [12]. The Segmentation module was defined as a deep CNN architecture (66M parameter) referring to HRNetV2-W48 in [13] (details provided in the original literature). The segmentation network is initialized with HRNetV2-W48 pre-trained on Imagenet dataset as provided by the author[13].

4 Results

In this section, we 1) qualitatively inspect the learnt spatial distribution of the FoV/Resolution Trade-off; 2) quantitatively compare our method with seven baselines, average expert and top Gleason2019 challenge performers; 3) provide visual results of our segmentation performance against the set of six baselines.

For one input image \mathbf{x} , based on the output of foveation module $F_\theta(\mathbf{x}_r)$, we calculate the weighted average patch size over all localtions, and refer to as Average Patch Size Map (APSM). In Fig.4, we pick one validation image as input and plot its APSM at three different epochs. As shown in Fig.4, as training progresses, the

average patch sizes at regions with small patterns (e.g. nuclei) shrink, while the average patch sizes at regions with large patterns (e.g. glands or background) expand. The observation evidenced that our method can learn the spatial distribution the FoV/Resolution Trade-off.

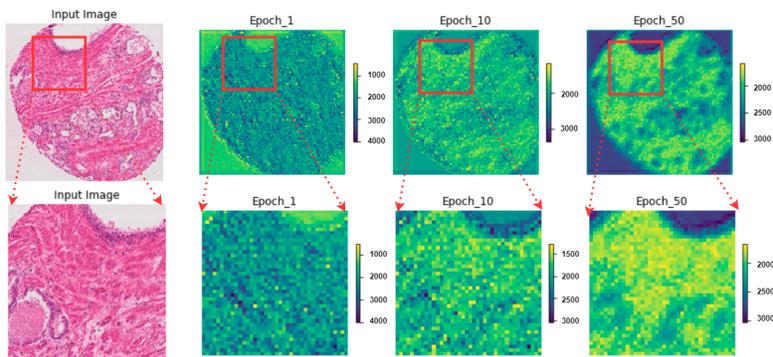


Fig. 4: The evolution of the learned weighted average patch size during training for one validation image. The brighter the colour, the higher the resolution and smaller patch.

To quantitatively evaluate our method we perform comparisons with three groups of baselines: Group 1) five one-hot baselines which force the probability distributions from foveation module $F_\theta(\mathbf{x}_{lr})$ to be fixed one-hot vectors, thus selecting only one scale per baseline from the given set of 5 patches with varying FoV/resolution. A uniform random one-hot baseline that at each location randomly selects one from the given set of 5 patches with varying FoV/resolution, referred to as Baseline-Random. And an average baseline that assigns equal probability $f_d^{(i)}$ of $1/5$ over the set of 5 patches with varying FoV/resolution, referred to as Baseline-Average; Group 2) average expert baseline measure the 6 annotators’ performance taking STAPLE fused golden standard as ground truth; Group 3) top 2 results of the Gleason2019 challenge leaderboard <https://gleason2019.grand-challenge.org/Results/> ranked by overall all classes average segmentation accuracy. We also collect highest segmentation accuracy of each classes as a third Single-Class-Best case for comparison. We quantify segmentation performance of group 1 and group 2 by Intersection over Union (mIoU). We quantify segmentation performance of group 3 via pixel accuracy for each class, to be consistent and comparable against results released on the leaderboard. It is worth noting that for all results we remove the “Gleason Grade 5” class in evaluation, as it is under represented - only 2% pixels in the given dataset.

We first compare our method against the baseline approaches in Group 1. The results are shown in Table 1. Our method achieved better segmentation performance (mIoU/IoU) over the 7 one-hot baselines described above in i) overall four class average, ii) benign class and iii) class Grade 4. The performance is also comparable with the best of class Grade 3. The results show that our method combines the advantages of dif-

ferent FoV/Resolution trade-off over the five baselines. This point is also qualitatively illustrated in Fig 5. The results against the baseline of Group 2 are shown in last row of Table 1, where our model improves on the average performance of 6 human experts by 6.5% and 7.5% for the two most clinically important and ambiguous classes (Gleason Grade 3 and 4), and gives comparable performance for overall average and the Benign class. Group 3 results are shown in Table 2, where our model achieves better segmentation accuracy against the top performers in the challenge for the two most clinically important and ambiguous classes (Gleason Grade 3 and 4) by 13.1% and 7.5%.

Table 1: Mean IoU (column 2) and IoU (column 3-5) on Gleason2019 Histology dataset. row 2-6: 5 one-hot baselines; row 7: uniform random one-hot baselines; row 8: average baseline; row 9: our result with foveation approach; row 10: average expert baseline.

Baselines	Overall	Benign	Grade 3	Grade 4
Baseline-1100 ²	0.520	0.810	0.618	0.650
Baseline-2000 ²	0.525	0.805	0.644	0.653
Baseline-2900 ²	0.530	0.810	0.649	0.661
Baseline-3800 ²	0.505	0.764	0.614	0.640
Baseline-4400 ²	0.460	0.675	0.556	0.610
Baseline-Random	0.487	0.722	0.586	0.638
Baseline-Average	0.493	0.788	0.522	0.660
Ours	0.533	0.824	0.630	0.678
Average Expert	0.569	0.839	0.564	0.603

Table 2: Quantitative comparison to Gleason2019 challenge leaderboard results measured by single class pixel accuracy

Experiment	Benign	Grade 3	Grade 4
Ours	88.3	83.8	78.1
Overall Top1	95.9	2.24	16.5
Overall Top2	83.0	52.7	54.0
Single-Class-Best	95.9	70.7	70.6

5 Discussion

Motivated by our observation that the FoV/resolution trade-off varies widely, we introduced a new theoretically grounded algorithm for simultaneously learning the spatial distribution of the trade-off and training a segmentation network that exploits it. Our method is simple to implement, requiring simply the addition of a "patch-extractor" and a light weighted foveation module (CNN). Our approach is complementary to a

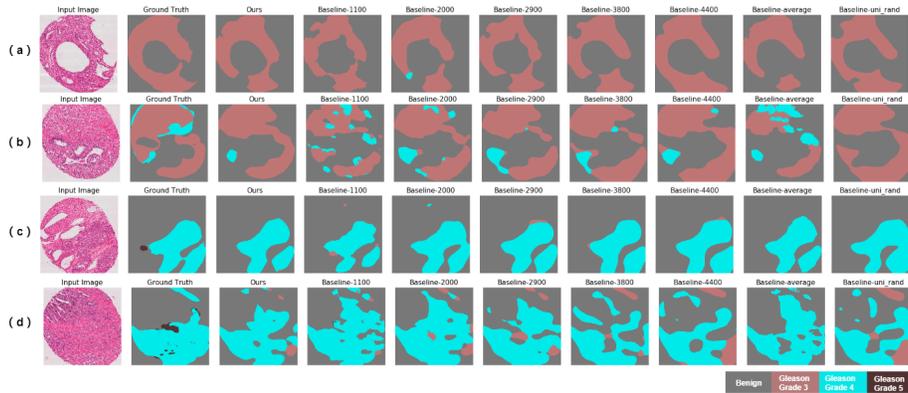


Fig. 5: Qualitative comparison of our method versus the seven baselines on four example validation images

wide range of existing “multi-resolution” segmentation architectures. In this work we used HRNetV2, a SoTA very deep multi-scale architecture and enhanced it with the proposed foveation module. Experiments on the Gleason2019 challenge segmentation data set show superior performance over single “best fit” trade-offs, average expert performance, and challenge leaderboard top results, especially for the two most clinically important and ambiguous classes (Gleason Grade 3 and 4). For the typically easy Benign class, the performance of our approach is slightly lower but still competitive to the challenge leaderboard top results, and the variation compared to other classes is anticipated: a) Table 1 shows our method outperforms all seven baselines on the Benign class and is comparable (1% difference) to average expert performance; b) Table 2 shows that our Benign-class performance is similar to the leading published.

We acknowledge the noisy nature of the dataset used to demonstrate our approach, however we believe that it is a good example to illustrate its robustness. The noise in the annotations is due to minor pattern variations between adjacent Gleason classes that leads to high variation in experts’ performance in Table 1 and Fig 1 (in supplementary material). Our method shown promising results on Gleason2019 Tissue microarrays (also TMAs) images, we would expect it to perform well on WSI and more generic non-medical ultra-high resolution image dataset too and would test in future work.

In the current implementation, all scales are searched equally, which means that all local patterns/classes are treated equally. This does not account well for class imbalance. In the Gleason challenge, as in many medical image segmentation tasks, rare classes can be overlooked unless explicitly emphasised during training. Therefore a key focus for future is to add a weighting mechanism taking care imbalanced classes.

Acknowledgements. We sincerely acknowledge: Marnix Jansen for inspirational pathological advice. Hongxiang Lin for the insightful discussions. C.J., T.M. and D.A. acknowledge funding by the EPSRC grants EP/R006032/1, EP/M020533/1, the CRUK/EPSRC grant NS/A000069/1, and the NIHR UCLH Biomedical Research Centre.

References

1. Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep neural network models for computational histopathology: A survey. *arXiv preprint arXiv:1912.12378*, 2019.
2. Nikhil Seth, Shazia Akbar, Sharon Nofech-Mozes, Sherine Salama, and Anne L Martel. Automated segmentation of dcis in whole slide images. In *European Congress on Digital Pathology*, pages 67–74. Springer, 2019.
3. Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.
4. Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3640–3649, 2016.
5. Wuyang Chen, Ziyu Jiang, Zhangyang Wang, Kexin Cui, and Xiaoning Qian. Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8924–8933, 2019.
6. Yuqian Li, Junmin Wu, and Qisong Wu. Classification of breast cancer histology images using multi-size and discriminative patches based on deep learning. *IEEE Access*, 7:21400–21408, 2019.
7. Angelos Katharopoulos and François Fleuret. Processing megapixel images with deep attention-sampling models. *arXiv preprint arXiv:1905.03711*, 2019.
8. Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, 23(7):903–921, 2004.
9. Maximilian Ilse, Jakub M Tomczak, and Max Welling. Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712*, 2018.
10. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.
11. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
12. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.
13. Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019.