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

Rüdiger Schmitz [a,b,c,∗], Frederic Madesta [b,c], Maximilian Nielsen [b,c], Jenny Krause [d], Stefan Steurer [e], René Werner [b,c,∗∗], Thomas Rösch [a,∗∗]

[a] Department for Interdisciplinary Endoscopy, Hamburg, Germany
[b] Center for Biomedical Artificial Intelligence (bAIome), Hamburg, Germany
[c] Department of Computational Neuroscience, Hamburg, Germany
[d] I. Department of Internal Medicine, Hamburg, Germany
[e] Department of Pathology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

ARTICLE INFO

ABSTRACT

Histopathologic diagnosis relies on simultaneous integration of information from a broad range of scales, ranging from nuclear aberrations ($\approx \mathcal{O}(0.1\mu\,m)$) through cellular structures ($\approx \mathcal{O}(10\mu\,m)$) to the global tissue architecture ($\gtrsim \mathcal{O}(1mm)$).

To explicitly mimic how human pathologists combine multi-scale information, we introduce a family of multi-encoder fully-convolutional neural networks with deep fusion. We present a simple block for merging model paths with differing spatial scales in a spatial relationship-preserving fashion, which can readily be included in standard encoder-decoder networks. Additionally, a context classification gate block is proposed as an alternative for the incorporation of global context.

Our experiments were performed on three publicly available whole-slide images of recent challenges (PAIP 2019: hepatocellular carcinoma segmentation; BACH 2020: breast cancer segmentation; CAMELYON 2016: metastasis detection in lymph nodes). The multi-scale architectures consistently outperformed the baseline single-scale U-Nets by a large margin. They benefit from local as well as global context and particularly a combination of both. If feature maps from different scales are fused, doing so in a manner preserving spatial relationships was found to be beneficial. Deep guidance by a context classification loss appeared to improve model training at low computational costs. All multi-scale models had a reduced GPU memory footprint compared to ensembles of individual U-Nets trained on different image scales. Additional path fusions were shown to be possible at low computational cost, opening up possibilities for further, systematic and task-specific architecture optimisation.

The findings demonstrate the potential of the presented family of human-inspired, end-to-end trainable, multi-scale multi-encoder fully-convolutional neural networks to improve deep histopathologic diagnosis by extensive integration of largely different spatial scales.

1. Introduction

1.1. Clinical relevance and motivation

If the rumor 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 (Brierley et al., 2016). For instance, a fine-grained grading and stag-

ing of dysplasia and malignancy in precursor and cancer lesions, respectively, underlies individualised treatment planning in many tumour entities. Moreover, in curative surgery, the assessment of whether the resection specimen margins are free of tumour cells is of vital importance and a core task of clinical pathology.

Human pathologists meet these challenges with the help of elaborated diagnostic criteria and grading systems for all kinds of cancer and cancer precursors. Even though their specific details vary for different kinds of cancer, many rely on a combination of features such as

- Nuclear inner density, i.e. colour, (Zink et al., 2004) and
- Deformed and varying nuclear shapes or global alterations of the nuclei (Zink et al., 2004).
- Increased nucleus to stroma ratio and
- Loss of nuclear polarity (e.g. nucleus not anymore at the bottom), as observed in many glandular tumours.
- Deformed cellular shapes and heterogeneous cell sizes,
- Loss of organ- and function-defining positions on small scales (e.g., neighboring cells not in a single layer, but some stacked over each other) and larger scales (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 (not-exhaustive) list, diagnosis and grading of malignancy inherently involve a range of different scales. These scales may span a factor of more than a thousand-fold, ranging from sub-nuclear features (which lie on a spatial scale of $\approx \mathcal{O}(0.1\mu m)$) via nuclear, cellular ($\approx \mathcal{O}(10\mu m)$), inter-cellular ($\approx \mathcal{O}(100\mu m)$) to glandular and other higher organisational features ($\gtrsim \mathcal{O}(1mm)$). The importance of the integration of information from different scales is reflected in how human pathologists approach these tasks: Regions of interest are repeatedly viewed at several different magnifications by turning the objective revolver of the microscopy back and forth.

In this work, we aim to develop a family of deep learning models that architecturally mimic this behaviour.

1.2. Related works

With their success in various computer vision tasks, deep learning methods have opened up a myriad of perspectives for computer vision and computer-aided diagnosis (CADx) in histopathology (Litjens et al., 2017). Image segmentation is a standard task in computer vision and machine learning and has a direct clinical use in the field of pathology, be it the analysis of the margin status (i.e., distance of tumour cells to resection margin), area-dependent grading systems (with the Gleason score in prostate cancer as a prominent example (Epstein et al., 2016; Karimi et al., 2020b; Nir et al., 2018), or specific research applications, such as the analysis of 3d-tumour morphology from a multiplicity of tissue sections (Schmitz et al., 2018; Segovia-Miranda et al., 2019).

In medical image segmentation, standard computer vision models, including fully convolutional neural networks (FCNs, Long et al., 2015) and, most prominently, U-Net-based architectures (Ronneberger et al., 2015) have successfully been applied to various scenarios and imaging modalities (Litjens et al., 2017; Isensee et al., 2018), including computational histopathology (Bulten et al., 2019; Liu et al., 2017; Campanella et al., 2019).

In addition, more specialised network architectures (Bejnordi et al., 2017b; Li et al., 2018a; Vu and Kwak, 2019) and training techniques (Campanella et al., 2019; Wang et al., 2019) have been proposed to address the challenges of computational histopathology. Some of these works touch upon the question on how additional context can be provided to the network, but are mainly confined to local, similar-scale context (Bejnordi et al., 2017b; Li

et al., 2018a) and/or sliding-window convolutional neural network (CNN) techniques (Bejnordi et al., 2017b; Wetteland et al. 2019) or classification tasks. Early on, Nir et al. (2018) described the potential of multi-scale features from more separate scales in classical, "hand-crafted" feature-driven machine learning. By integrating the features from different scales by use of a support vector machine (SVM), they paved the way for many works to follow. Recently, the same research group advocated the use of individually trained CNNs as feature extractors whose outputs were, in an ensembling-like fashion, combined by a logistic regression model into a final Gleason grade classification in prostate cancer (Karimi et al., 2020b). Similarly, Wetteland et al., (2019) suggested to train distinct CNNs as feature extractors and merge their information in a classification network in replacement of the original fully-connected layers, which can again be viewed as an ensembling approach. For application to breast cancer and its differential diagnoses, Ning et al. (2019) proposed a similar ensembling technique, but again based on classical, hand-crafted feature extractors and using an SVM for integration of multi-scale information.

In 3d imaging, multi-path end-to-end trainable models have incorporated similar-scale and local context to reduce the memory footprint and alleviate the problem of the otherwise extremely limited input size in memory-costly 3d-nets (Kamnitsas et al., 2017), with remarkable success in e.g. the sub-acute stroke lesion segmentation challenge ISLES 2015 (Maier et al., 2017). In histopathology, there also have been attempts toward end-to-end trainable multi-scale models (Gu et al., 2018; Li et al., 2018a), but which have so far shown only minor benefits as compared to the aforementioned ensembling variants. This observation indicates that multi-scale deep learning-based segmentation of histopathology data is still in its infancy and its actual potential remains to be unveiled.

Introduction of additional image context has been a topic of interest also in the natural image domain, resulting in prominent techniques like dilated or atrous convolutions and atrous spatial pyramid pooling (Chen et al., 2017a; 2017b). Further, Zhang et al. (2018) proposed an FCN architecture that explicitly predicts "scaling factors" for the possible classes from the bottleneck layer, which are then used to multiply and, thus, highlight the respective feature maps at the final layer. The scaling factors can capture the overall image content and can be trained by use of an additional classification loss. Similarly, Zhou et al. (2019) introduced a reinforcement-based strategy involving two sub-nets, one for encoding context and one for the actual segmentation task. By the properties of the natural image domain, namely the limited image size as compared to histopathologic whole-slide images (WSIs), the scales of detailed and contextual features in these works are, however, much more similar than in histopathology. Congruously, the primary aim of these approaches has been to "help clarify local confusion" (Liu et al., 2015) or to make better use of what is fed into to the net anyways, rather than to add large and otherwise unavailable context. For histopathology image segmentation as a specific task, however, we aim for the integration of otherwise unavailable information from much different scales into a single, end-to-end trainable model.

1.3. Contributions

There exist plenty of highly optimised, U-Net-derived architectures employing, for instance, elaborate skip connections (Badrinarayanan et al., 2017), dense connections (Li et al., 2018b), attention gating techniques (Oktay et al., 2018) and newer FCN architectures like DeepLab (Chen et al., 2017a). Nevertheless, standard U-Nets have turned out to be robust work horses for many medical computer vision tasks and are hard to beat by internal modifications of the base architecture (Isensee et al., 2018; 2019).

However, histopathology diagnosis is, by the nature of the large whole-slide images and with closely interwoven features from very different scales, a very specific and challenging task. Therefore, drawing on the U-Net as a standard base model, this work explores whether an architectural mimicry of how human experts approach this specific task can improve the performance of FCNs for histopathology image segmentation.

The main contributions of this paper are as follows: We introduce a family of U-Net-based fully convolutional deep neural nets that are specifically designed for the extensive integration of largely different spatial scales. First, we propose a simple building block that can fuse various encoders with different spatial scales in a manner that preserves relative spatial scales. As a lightweight alternative, we also propose the use of an independent context classification model for gating the segmentation model output. Second, we integrate these building blocks into different multi-scale FCNs and compare their segmentation performance to U-Net baseline architectures. To illustrate generalisability, the evaluation is based on three different publicly available image datasets provided by recent challenges. Third, by a systematic, stepwise analysis, we identify relevant aspects of the proposed multi-scale FCN family, including the necessity of preserving spatial relationships between different encoders, the benefits from deep guidance by an additional classification loss and possible generalisations through multiple path fusions. Based on these observations, we narrow down the possible multi-scale setups and comment on how to systematically adapt the presented multi-scale FCN family to specific deep learning tasks in histopathology.

To foster reproducibility and further research, our proposed models are publicly provided as open source[1] to the community.

2. Model architectures

2.1. Baseline architectures

2.1.1. U-Net architecture

Beyond the still popular sliding-window CNN-based techniques, U-Net-based FCN architectures form the de facto standard in the medical image domain (Isensee et al., 2018), including histopathology (Litjens et al., 2017). Beating the standard U-Net by internal modifications is evidently hard (Isensee et al., 2018; 2019) and beyond the scope of this work. Rather, as outlined in the introduction, this study examined whether by designing a model of standard components but with its larger architecture mimicking human expert diagnostic procedures, further improvement can be made.

Therefore, we chose a non-modified ResNet18-based U-Net (He et al., 2015) as a common, standard U-Net variant to form the baseline for this study[2] For a detailed description of the baseline model, the reader is referred to the Supplementary Materials. In brief, the ResNet18 forms the encoder of the otherwise standard U-Net architecture (cf. Fig. S1), where the encoding ResNet18 has been pre-trained on the ImageNet dataset (Russakovsky et al., 2015). For our study, the baseline model was trained at full-resolution patches of 512×512 pixels of the WSI images.

2.1.2. Multi-scale ensembles as an upper bound for state-of-the-art performance

We additionally compared our proposed models to ensembling techniques to see if we can reach or even excel their performance

with a single, and computationally less costly, model. Inspired by the approach of Karimi et al. (2020b) to Gleason grading, individual U-Nets were trained on the different image scales. The predicted probabilities were then re-sampled to the target resolution and merged by use of different ensembling techniques, namely hard majority votes, average ensembles (soft majority voting), and logistic regression ensembles. For each ensemble, the best individual model per scale was selected. The logistic regression model was trained and evaluated on the same train and test sets as the individual models. By this procedure, we aimed to define a thorough and systematic upper bound for state-of-the-art multi-scale ensembling performance.

2.2. The msY model family: Multi-scale multi-encoder architectures

To provide the network with context and architectural information (cf. Fig. 1 and the considerations in Section 1), we constructed a family of multi-scale multi-encoder networks building upon the baseline Res-U-Net architecture.

In the following, we first introduce the underlying blocks for integration of multi-scale context, namely the *multi-scale merge block* and the *context classification gate*. Afterwards, we describe the different variants of our setup that are examined in this study.

Technically, it is worth noting that common whole-slide image (WSI) formats use so-called pyramid representations, which contain the original image in multiple, downsampled versions. Therefore, multiple scales can directly be loaded from file, with no need for resampling and, hence, only a moderate overhead only.

2.2.1. Multi-scale merge block: Spatial relationship preserving path fusion in multi-scale multi-encoder models

Fig. 2a sketches the functioning of the multi-scale merge block. At the bottleneck level, the feature maps from both the main encoder and the side (context) encoder have sizes of $16 \times 16 \times 512$. In order to spatially match the output of the full resolution encoder, a $n' \times n'$ center cropping ($\mathcal{S}_{\frac{16}{n'} \times \frac{16}{n'}}$) of the n-times downscaled context path is performed, followed by $n \times n$ bilinear upsampling ($\mathcal{U}_{n \times n}$), where $n' = 4$ if $n = 4$ and $n' = 8$ if $n = 16$. For the case $n' \neq n$, another center cropping with 16×16 is conducted. Both, now spatially consistent paths are then merged by concatenation. Finally, the number of feature maps is reduced to the original number by a 1×1 convolution. This operation is meant to learn which of the feature maps from the two paths are relevant and how they need to be combined.

For application to multiple context encoders, spatial alignment is ensured for any individual context encoder in the same manner as described above. The spatially-aligned feature maps from all encoders are then concatenated. Afterwards, the feature map size is reduced by a 1×1 convolution with $512 \cdot (m + 1)$ input feature maps and 512 output feature maps, $c_{1 \times 1}^{512 \cdot (m+1), 512}$, where m denotes the number of side encoders.

2.2.2. Context classification gate

Alternatively to the use of multiple encoders that are merged by multi-scale merge blocks, we examined the potential of gating the class-wise predictions from the segmentation network by a context classifier. The context classifier tries to predict the content of the central detail patch from the low-resolution global context patch alone (cf. Fig. 1). As multiple, even mutually exclusive classes can be included in the detail patch, this states a multi-label classification problem.

The context classification gate is depicted in Fig. 2b. The classification net outputs a probability value for each individual class (illustrated by the coloured boxes in the bottom left corner).
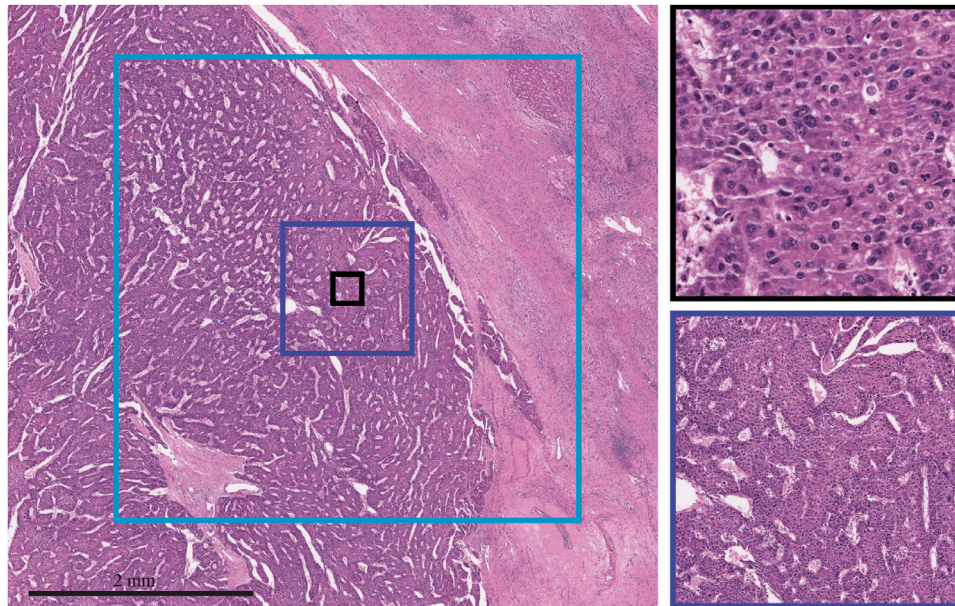

Fig. 1. Input patches at different spatial scales, shown in an illustrative region of the same whole-slide image as depicted in Fig. 4. The innermost, black rectangle corresponds to a 512×512 pixel patch of the scale 1, which we refer to as the "detail patch" (zoomed view in the top right inset). The next, dark blue rectangle corresponds to a 572×572 pixel patch of the scale 4. It contains information on how the cells are organised in strands and "trabeculae" – or whether the cells violate these patterns. These features are hard or impossible to deduce using the innermost patch alone. In this sense, the dark blue patch adds "architectural" information. We refer to it as "local context" or the "local context patch". A zoomed view of it is shown in the bottom right inset. The outermost, light blue rectangle, which we call a "global context patch", contains information on the large-scale organisation of the tissue, such as the pseudocapsule, which is typical for hepatocellular carcinoma. Whilst a standard U-Net is provided with the information from the detail patch solely, a msY-Net architecture (Section 2.2.3) can integrate information from the detail patch plus either the local or the global context patch. The msY²- and msY*I*-Net architectures are two options for integration of all three scales. The scale bar is 2 mm. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

These are multiplied to the probability maps of the segmentation network in a channel-wise manner (similar to how excitations in squeeze-and-excite blocks are handled, cf. (He et al., 2015)), thereby emphasising probable classes and suppressing unlikely diagnoses. To allow the segmentation network to either use or ignore this guidance, a "leak" is constructed by concatenation of the original, un-excited feature maps to the excited ones, followed by a 1×1 convolution that is to learn how to combine the excited and the leaked feature maps.

2.2.3. msY-Net: Integrating context and tissue architecture

The msY-Net is provided with two patches of the scales 1 and 4 (which correspond to the inner two rectangles in Fig. 1) or 1 and 16 (inner and outer rectangle) as input. The full-resolution patch (scale 1) is fed into the standard U-Net architecture. The other 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 two paths are merged at the bottleneck of the original U-Net by use of the multi-scale merge block (cf. Section 2.2.1). The resulting architecture is sketched in Fig. 3b.

In the following, we refer to a msY-Net that uses detail patches of the scale 1 and context patches of the scale n as msY$_{(n)}$-Net.

2.2.4. msY*I*-Net and msY²-Net: Integration of global and local context

In order to provide the model with large- and small-scale context information at the same time, we constructed two models that either use two context encoders or one context encoder plus one context classification gate. For the former variant, we added two context encoders using a single multi-scale merge block. We refer to this model as msY²-Net.

For the latter variant, a context classification gate that uses the large-context patch was added to an underlying msY-Net. This model is outlined in Fig. 3e. The *I* in the name msY*I*-Net refers to the large-context classification sub-net paralleling the underlying msY-Net *without* any fusion at the bottleneck or before. This network has two outputs: the segmentation of the full-resolution patch from its msY-Net part and the classification of the full-resolution patch content from the large-context encoder, its *I*-part. Finally, the final logits of the two paths are combined by a context classification gate that modifies the segmentation output by the classification of its context (cf. Section 2.2.2).

In the msY-Net and the msY²-Net architectures, spatial correspondence between the full resolution encoder and the context encoder(s) is enforced in the multi-scale merge block (cf. Fig. 2). It should be noted that for the large-context encoder in the msY*I*-Net that ends in a classification gate instead of a multi-scale merge block, there is no such requirement. Therefore, this model can, in principle, be fed with large-context patches of arbitrary scales.

2.2.5. Context classification loss

Additional loss functions can improve the training of specific parts of U-Net-based architectures and are used in various manners (Kickingereder et al., 2019; Li and Tso, 2019), including a classification loss on an additional output derived from the bottleneck feature maps (Mehta et al., 2018).

Analogously, we computed a classification output from the feature maps of global context encoders (i.e., those with input scale 16) and used it to compute an additional classification loss. The classification loss is computed with respect to the content of the detail patch. As described in Section 4, we used a binary cross entropy (BCE) loss for the classification problem and added it to the segmentation loss.
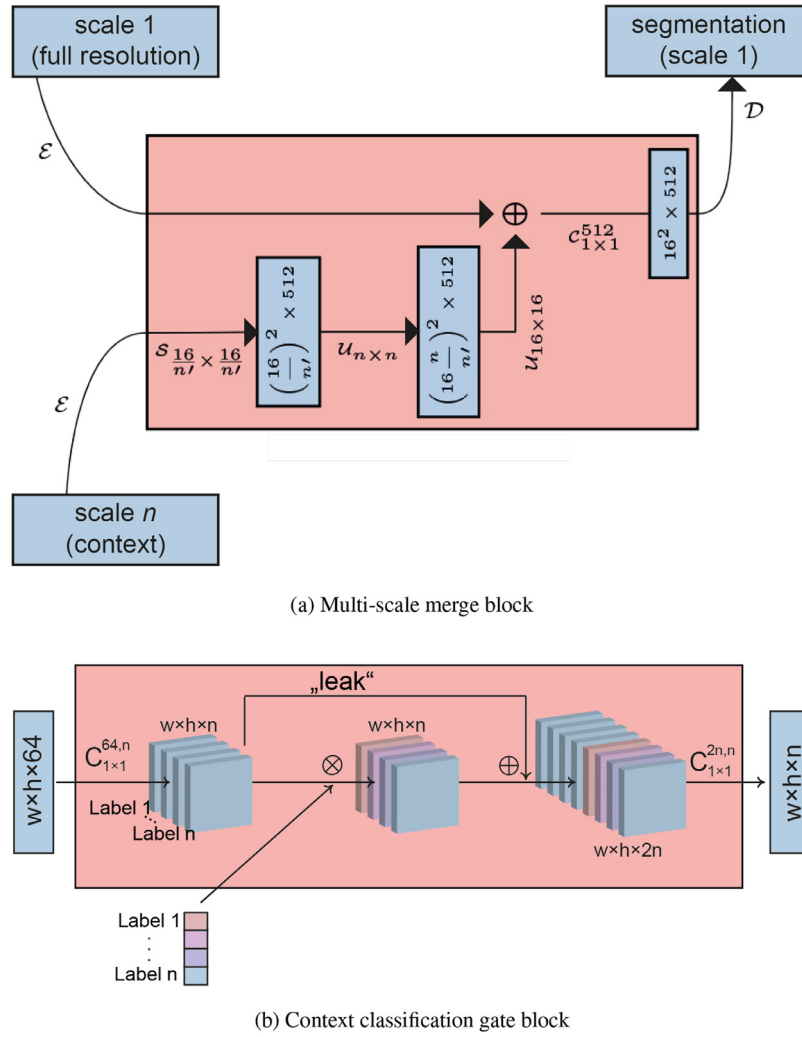
(a) Multi-scale merge block

(b) Context classification gate block

Fig. 2. Schematic illustration of the multi-scale merge block (a) and the context classification gate block (b). $\mathcal{E}$, $\mathcal{D}$ are the encoder and decoder path of the network (cf. Fig.), respectively. The blue upward boxes are the feature maps with their respective sizes printed inside. $\oplus$ denotes concatenation and $\otimes$ channel-wise multiplication. The "leak" connection is a copy, followed by concatenation. $\mathcal{S}_{x \times y}$ stands for the central cropping of the size $x \times y$ in the spatial dimensions and $\mathcal{U}_{a \times b}$ the bilinear upsampling by factors of a and b in the dimensions 1 and 2. w and h denote the spatial width and height, n the number of classes. $\mathcal{C}^{a,b}_{n_1,n_2}$ means $n_1 \times n_2$ convolution with a input feature maps and b output features maps (where a may be omitted if the size of its input is explicitly given), followed by ReLU activation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

By the additional classification loss, we wanted to ease gradient flow through the deeper layers of the context encoder and to explicitly force it to focus on the content of the detail patch.

2.2.6. Early and multiple fusions by multi-scale merge blocks

Positioning the multi-scale merge block at the bottleneck level is not obligatory. Technically, a multi-scale merge block can merge any two encoders at any level at which the one shall influence the other (the connection is directed, not mutual). Also, usage of multiple multi-scale merge blocks is possible.

Very early merge blocks will, however, not provide much context due to the small receptive fields of the center crop. On the other hand, earlier and in particular multiple fusions may allow for an additional processing of combined features and might facilitate modelling of complex combined features.

It should be noted that these multiple merges are computationally inexpensive, as they only introduce additional learnable parameters through the 1×1 convolutions inside the merge blocks.

All models are implemented using PyTorch v1.2.0 (Paszke et al., 2017).

3. Materials & methods

3.1. Datasets

To examine generalisability of our findings, experiments were conducted on three different datasets for different entities of cancer, collected by different centers and scanned by different scanners. To ensure reproduciblity, we employed the following three publicly available challenge datasets:

3.1.1. PAIP 2019

The PAIP (Pathology Artificial Intelligence Platform) 2019 challenge (part of the MICCAI 2019 Grand Challenge for Pathology) dataset comprises 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 (Kim et al., 2021). The slides have been stained by hematoxylin and eosin and digitalised using an Aperio AT2 whole-slide scanner at $\times 20$ power and 0.5021μ m/px resolution, resulting in image sizes between $35,855 \times 39,407$ and $64,768 \times 47,009$ pixels (1.399 to 3.044 gigapixels).
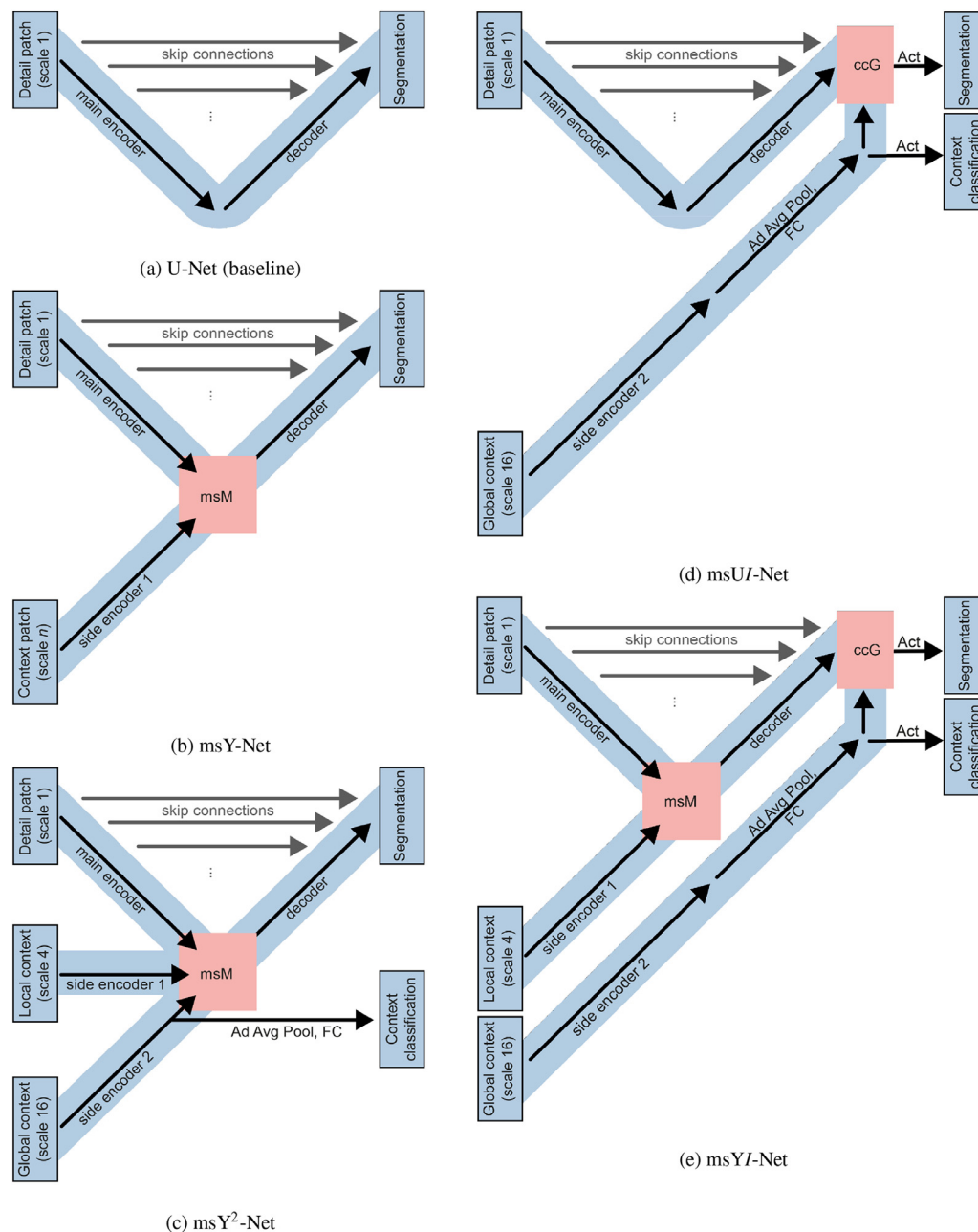
Fig. 3. Schematic illustration of the main architectures studied in this paper. All main and side encoders use an ImageNet-pretrained ResNet-18. The skip connections and the decoder are the same throughout all of our models (see supplemental materials for details). The multi-scale merge (msM) and context classification gate (ccG) blocks are described in sections 2.2.1 and 2.2.2 and sketched in Fig. 2. Ad Avg Pool denotes adaptive average pooling, FC a fully-connected layer and Act the activation function.

Regions of viable cancer cells as well as whole cancer regions (additionally including stroma cells and so forth) have been annotated manually. As described in (Kim et al., 2021), one pathologist with 11 years of experience in liver histopathology drew the initial annotations that were then reviewed by another expert pathologist. All cases of the given dataset include cancer regions. With respect to the Edmonson-Steiner grading system (Edmondson and Steiner, 1954), their distribution is as follows: $N = 7$ cases of grade 1, $N = 23$ grade 2 tumours, and $N = 20$ grade 3 samples.

All de-identified pathology images and annotations 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 HI18C0316).

Fig. 4(A) shows an exemplary whole-slide image from the dataset. It illustrates why hepatocellular carcinoma (HCC) is a prominent example of a cancer that is characterised not only by nuclear abberations, but also by local tissue abnormalities and large-scale features such as a so-called pseudocapsule (as illustrated in the figure). In fact, (low-grade) HCC is challenging to diagnose, as it is often identifiable *only* by abberations of the long-range tissue architecture with only minor nuclear abnormalities, if any (Bosman et al., 2010). This makes the dataset well-suited for demonstration of the importance of multi- and particularly large-scale context features.

3.1.2. BACH 2018

The BACH (BreAst Cancer Histology) 2018 challenge (Aresta et al., 2019) included the classification of small
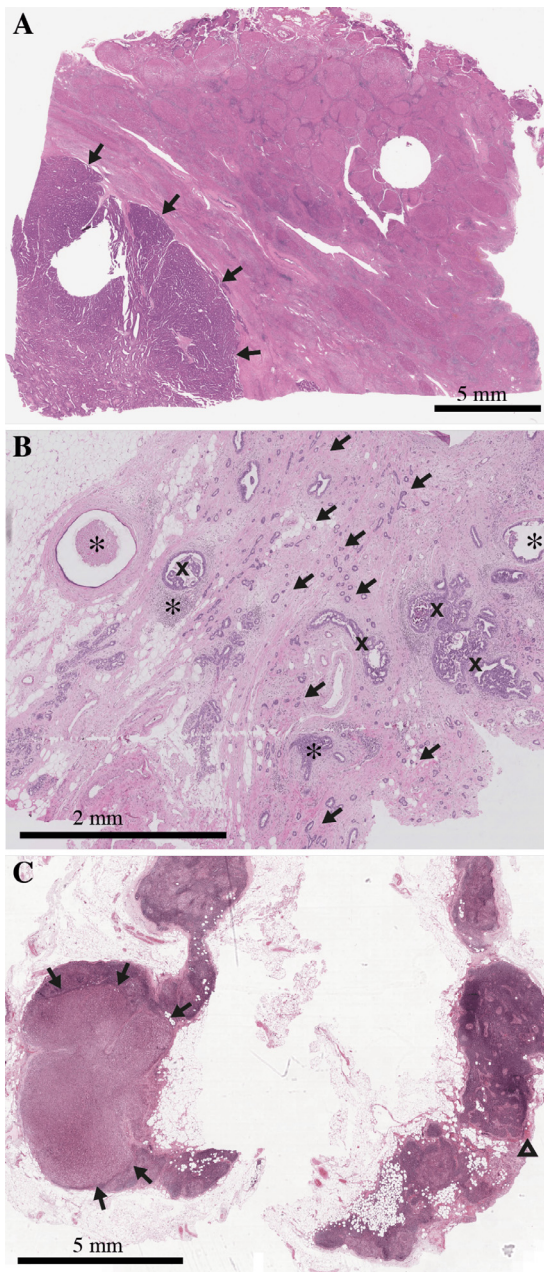
Fig. 4. Exemplary cases from the PAIP 2019 (A), BACH 2018 (B) and CAMELYON 2016 (C) datasets. The arrows mark regions of hepatocallular carcinoma, breast carcinoma and a lymph node macrometastasis in A, B and C, respectively. In B, benign and in situ lesions are indicated by stars and x's. Beside the macrometastasis in C, also isolated tumour cells are found (next to the triangle).

(2.048×1.536) image patches as one task, and the segmentation of WSIs of breast biopsies as another. For the latter task, 10 pixel-wise annotated WSIs were provided for training. Two medical experts performed the image segmentation using the following four labels: (1) normal, (2) benign, (3) in situ and (4) invasive breast carcinoma. WSIs were acquired by a Leica SCN400 ($\times$20 power, pixel scale: 0.467μ m/px) in the period from 2013 to 2015 at the Centro Hospitalar Cova da Beira. The image sizes range from $50,529 \times 36,833$ to $64,703 \times 45,808$ pixels (1.861 to 2.964 gigapixels).

Differential diagnosis in suspected breast cancer is known as a challenging task to human pathologists, and also to machine learning approaches. This is reflected in the results of the top-performing teams, reaching only moderate scores of 0.50 to 0.69 with respect to the custom challenge metric (Aresta et al., 2019). Fig. 4 (B) illustrates how close the different diagnoses can be interwoven, with invasive and in situ parts of the carcinoma directly neighboring each other.

3.1.3. CAMELYON 2016/MM subset

For the original CAMELYON (Cancer Metastases in Lymph Nodes) 2016 challenge (Bejnordi et al., 2017a), 399 WSIs of lymph nodes from women with confirmed breast cancer were gathered from two different centers (Radboud UMC and UMC Utrecht) during the first half of 2015. 240 of the 399 slides contained one or more nodal metastases. The images were scanned by a Panoramic 250 Flash II ($\times$20, pixel scale: 0.243μ m/px) and a NanoZoomer-XR Digital slide scanner C12000-01 ($\times$40, pixel scale: 0.226μ m/px), respectively. Image sizes are $61,440 \times 53,760$ to $217,088 \times 103,936$ pixels (3.303 to 22.563 gigapixels).

Initial annotations of lymph node metastases were drawn by medical students and then reviewed and corrected by two expert pathologists, as detailed in (Bejnordi et al., 2017a). This procedure fits the clinical observation that the detection of lymph node metastasis is a much easier task to human pathologists than diagnosis of HCC and pathologies of the breast. The original dataset contains macrometastases (tumour cell clusters with a diameter $\geq$ $2mm$), micrometastases ($0.2mm \leq$ diameter $< 2mm$) as well as yet smaller clusters down to isolated tumour cells. Whilst these are all clinically relevant, isolated tumour cells and very small clusters do not infer with the global lymph node architecture (Lakhani et al., 2012).

For our experiments, we therefore created a subset of the CAMELYON 2016 dataset, consisting of 20 WSIs with at least one macrometastasis. We refer to this subset as CAMELYON 2016/MM, with MM for macrometastasis. The procedure for the establishment of the CAMELYON 16/MM subset is detailed in Section. Fig. 4 (C) provides a typical example of the CAMELYON 2016/MM dataset. As illustrated in Fig. 4 (C), this dataset still contains isolated tumour cells (triangle) and micrometastases along with the macrometastases (arrows), but with a higher weight put to the macrometastases and the larger length scales as compared to the original CAMELYON 2016 dataset. However, it should be noted that this task still differs from the other two, namely diagnosis of HCC and pathologies of the breast. Whilst macrometastases infer with the global architecture of the lymph node, they can still be diagnosed at the single- or few-cell level, by virtue of the differences of individual tumour and autochthonous cells. Therefore, even though larger scales certainly help the human pathologist with the diagnosis of macrometastases, the consideration of larger scales and the integration of multi-scale information is of lesser importance.

3.2. Preprocessing

WSI data preprocessing was similarly performed for all three datasets. Compared to the annotations originally provided for the respective challenges, we automatically generated annotations for an "overall tissue" as an additional class. This was achieved by thresholding of the original images with [R, G, B] $\leq$ [235, 210, 235], followed by binary morphological opening and closing operations. The rationale for introducing the additional class was threefold: First, it makes the "background" class, that would otherwise include not only all-white background *but also* healthy tissue, much less heterogeneous. Second, it facilitates sampling of healthy tissue patches as meaningful negative examples (cf. Section 4). Third, the automatically generated "overall tissue" annotations let us generalise the Reinhard colour normalisation (Reinhard et al., 2001) to our application. Reinhard colour normalisation, as originally introduced, implicitly assumes the colour statistics to be computed

from meaningful image areas. In contrast, when directly applied to a WSI image, the fraction of all-white background would determine the colour statistics and hence a colour normalisation, which is undesired. To avoid this, images were standardised by Reinhard colour normalisation (Reinhard et al., 2001) with respect to the "overall tissue" regions only, and then normalised to channel-wise zero average and unit variance, again in the tissue regions only.

3.3. Evaluation metrics

The primary outcome parameter for all experiments was the Jaccard index with the classes weighted as detailed in Section 4. Secondary, for experiments on the BACH challenge data, we additionally report the custom metric used in that specific challenge (Aresta et al., 2019). Simply put, the metric is based on a pixel-wise accuracy measure that penalises predictions that are farther from the given ground truth segmentation class. This is possible since the segmentation classes can be ordered according to their malignancy. We refer to that metric as the "BACH metric".

For each dataset, we employed a 5-fold cross validation (CV) strategy with the split conducted at the level of the entire WSIs. Splitting is performed such that all classes were present in the validation set. The splits were computed once and then kept fixed throughout all experiments. The number of WSIs in the validation set was 10 for PAIP 2019, 2 for BACH 2018 and 4 for CAMELYON 2016/MM. Validation was performed after the following epochs (number of iterations): 1 (1,920), 3 (5,760), 5 (9,600), 8 (15,360), 11 (21,120), 16 (30,720), 21 (40,320), and then every 10 epochs (19,200 iterations). At each validation step, a fixed number of $3,072 \times 3,072$ pixel-sized sub-images per image was evaluated. We evaluated four such sub-images per validation WSI for PAIP, ten for BACH and six for CAMELYON 2016/MM, resulting in 40, 40 and 24 of these per split for PAIP, BACH and CAMELYON 2016/MM, respectively. The positions of the sub-images were randomly sampled with the condition that, for any WSI, all available classes shall be represented in at least N_I/N_c of the sub-images. Here, N_I denotes the number of validation sub-images from that WSI and N_c is the number of available classes. Sampling of the sub-image positions was done once, before the first experiment, and then kept fixed throughout all experiments. This means that all models for a given split are evaluated with respect to the exact same sub-images at all validation steps.

3.4. Statistical analysis

For a given model, performance scores were evaluated per split, from which the mean and 95%-confidence interval (CI) across the five splits were computed. To set a robust baseline for the experiments in Section 5.1, the experiments for the baseline, single-scale U-Net were repeated three times and averaged, corresponding to a conservative estimate (over clustered experiments). The results for the individual runs of the baseline U-Net are provided in the Section.

As detailed in Section, we applied a corrected *t*-test for use in cross-validation settings (Nadeau and Bengio, 2000; Bouckaert and Frank, 2004) to test for statistical significance of the differences between the multi-scale models and the baseline U-Net. As further described in the same section, our study is underpowered for examining differences between different multi-scale models for significance. Therefore, we restricted the tests to comparisons between multi-scale architectures and the baseline U-Net. All statistical analyses were performed using R v4.0.0 (R Core Team, 2013).

4. Experiments

Our experiments were designed as follows: First, we explored the behaviour of the msY-family on the PAIP 2019 dataset. Second, to examine generalisability of our findings, we evaluated the performance of the best performing architectures on the BACH 2018 and the CAMELYON 2016/MM dataset. Third, to aetiologically understand the behaviour of the msY-models and to narrow down the possible design variants, we conducted further experiments with selected variants of that models on the PAIP 2019 dataset.

The training strategy and all hyperparameters were manually optimised for training of the baseline Res18-Unet on PAIP 2019 before the experiments and then kept fixed for all models and throughout all experiments. We aimed to reproduce a standard and widely used setup as much as possible, including the use of a common model as the baseline. For the same reason, we chose the binary cross entropy (BCE) loss as a common and widely used loss function (see, e.g., Heller et al. (2019)). BCE was used as a loss function for both classification (if the model has an additional classification output, cf. Section 2.2.2) and segmentation. The training process was split into (short pseudo-)epochs of 1,920 patches each, with all patches re-sampled after each individual "epoch". Compared to the use of a fixed number of pre-selected patches, this reduces overfitting given the limited memory resources and the large whole-slide images. For each pseudo-epoch, the patches were balanced with respect to both the individual cases in the training set and the available classes. Optimisation was performed using Adam (Kingma and Ba, 2014) employing a learning rate of 10^{-3} and a learning rate decay with $\gamma = 0.5$ every 30 epochs (57,600 iterations). During training, we employed online data augmentation including the following standard operations: rotation, flip along the horizontal axis, saturation and brightness transformation.

4.1. Comparison of msY-family and baseline models

4.1.1. PAIP 2019

The clinical scenario in our PAIP 2019 experiments is the segmentation of hepatocellular carcinoma for evaluation of tumour extent and margin status after the resection of hepatocellular carcinoma.

The comparison of all multi-scale models (including the msY-family and the multi-scale ensembling variants) to the baseline U-Net resulted in a total of $N = 9$ pairwise model comparisons. The evaluation of the corresponding results were the primary focus of the PAIP 2019 experiments. In addition, we also compared the different multi-scale architectures amongst each other and to multi-scale ensembles. However, as our study is underpowered for finding statistically significant results in between different multi-scale models (cf. supplemental materials), we only descriptively present the corresponding results.

For the PAIP 2019 experiments, all models were trained for at least 120 pseudo-epochs (230,400 iterations) or until convergence of the validation loss was reached. According to the clinical scenario and due to the fact that the localisation of viable tumour cells with respect to the resection margins determines the resection status, we put additional weight to the viable tumour class as compared to the whole tumour class and weight the loss by [0, 1, 2, 6] for background, overall tissue, whole tumour and viable tumour classes, respectively.

Segmentation performance as the primary outcome parameter was measured by the weighted average Jaccard index in the classes of interest, i.e., "whole tumour" and "viable tumour", where the weights are the same as above.

As a secondary outcome, we furthermore examined the resource requirements of the models. We report the number of trainable parameters and measured the GPU memory footprint. For the

latter, we observed the GPU memory usage of the training process, including forward and backward pass, at different batch sizes using the NVIDIA System Management Interface. The memory footprint of the model without system overhead was then deduced by linear regression and reported as gigabytes per patch in batch. The detailed measurements are reported in the supplemental materials. For multi-scale models, the term "patch" is meant as a multi-scale patch, i.e., including the image patches for all individual scales. The input patch sizes in all models and for all scales are 512×512 pixels, except at scale 4 (576×576 for the U-Net in the ensembles, 572×572 for the msY-family models).

4.1.2. BACH 2018

As described in Section 3.1.2, the BACH 2018 dataset allows us to evaluate the models with respect to their capability in the differential diagnoses of breast lesions, which represents the clinical scenario for this experiment.

We started from the models from the PAIP 2019 experiments and, after replacement of the output layers, fine-tuned them on the BACH 2018 dataset. Without loss of generality, we fine-tuned the models from the split i ($i \in [0, 1, 2, 3, 4]$) of the PAIP dataset on the split with the same index i ($i \in [0, 1, 2, 3, 4]$) of the BACH 2018 dataset. Training was performed exactly as described in Section 4.1.1, with the following differences: First, for the first 10 pseudo-epochs (19,200 iterations), only the output layers were trained and the other weights were kept fixed. Fine-tuning of all weights was then continued for another 70 pseudo-epochs (134,400 iterations). Second, in line with the clinical scenario of breast lesion differential diagnosis, we put equal weights to all classes of interest, i.e., "normal", "benign", "in situ" and "invasive carcinoma", both for the computation of the loss function and the weighted-average Jaccard index. In order to examine whether our models are on par with state-of-the-art results on this task, we additionally evaluated the scores from the custom BACH 2018 metric as introduced in (Aresta et al., 2019).

4.1.3. CAMELYON 2016/MM

For our CAMELYON 2016/MM experiments, we again started using the models from the PAIP 2019 experiments and, after replacement of the output layers, fine-tuned them on the new dataset. Training was again performed as described in Section 4.1.1 except that for the 10 pseudo-epochs (19,200 iterations) only the output layers were trained and the other weights were kept fixed. For CAMELYON 2016/MM dataset, we fine-tuned for at least 60 pseudo-epochs (115,200 iterations) or until convergence. Again, without loss of generality, we fine-tuned the models from the split i ($i \in [0, 1, 2, 3, 4]$) of the PAIP 2019 dataset on the split with the same index i ($i \in [0, 1, 2, 3, 4]$) of the CAMELYON 2016/MM dataset.

4.2. Context classification loss

We hypothesised that when using a global-scale encoder, introduction of an additional context classification loss as a means of "guidance" may benefit the model (cf. Section 2.2.5). In order to examine whether this is consistently the case, we trained the following models with and without context classification loss on the PAIP 2019 dataset: $msY_{(16)}$, msY^2, $msY_{(16),MM}$.

4.3. Spatial alignment in multi-scale merge blocks

We hypothesised that spatial matching is an essential step in the multi-scale merge block. To test this hypothesis, we compared a msY^2-Net, as an example where three different scales are merged, to a variant of the same model but with the multi-scale merge block replaced by a pure concatenation followed by a 1×1 convolution. This corresponds to a multi-scale merge block but

without alignment of the spatial scales and without preserving spatial relationships. The rest of the spatially non-aligned msY^2-Net variant remained unchanged.

If, in line with the hypothesis, the spatially non-aligned msY^2-Net performs worse than its standard variant with the correct multi-scale merge block, it remains to be examined whether this is "only" due to the clumsy initialisation and can potentially be overcome without spatial merging. One might hypothesise that the deficit is not architecturally, but due to the introduction of new, untrained convolutions in the middle of the otherwise pretrained encoder. Therefore, we examined two different variants of the non-aligned model: one variant with randomly initialised weights and another variant where the 1×1 convolution was initialised with the unit matrix at the channels belonging to the main encoder and with 0′s everywhere else, both superimposed with random noise (normal distribution with standard deviation 10^{-4}). In the latter variant, the main encoder corresponded to the unperturbed, pretrained model at the beginning of the training process, up to noise.

4.4. Multiple merging

The multi-scale merge block as presented in Fig. 2a can be introduced at any level of the encoders. Therefore, it also allows for earlier or multiple fusions, which may allow for the additional processing of combined features. Very early merge blocks, however, do not provide much additional context, due to the cropping step inside the multi-scale merge block. Therefore, it is not a priori clear, at which level the merge connection shall be established or if multiple path merges can further benefit the model.

In order to study whether models with multiple merges can be trained robustly and whether an effect through multiple merges can be found, we compared the segmentation performance of a $msY_{(16)}$-, a msY^2- and a $msYI$-Net to the analogues of them with multi-scale merge blocks at all encoder levels. In this experiment, to examine a "maximum" msY-family variant trainable on our hardware, we used ResNet34 instead of a ResNet18 for the global context encoder of the msY^2-model. All other models and encoders were ResNet 18-based, as by our standard. Additionally, we examined to which extent the introduction of multiple merges increases the GPU memory footprint of the models.

5. Results

5.1. Multi-scale multi-encoder models improve histopathology image segmentation

Table 1 reports the weighted average Jaccard index for the two non-trivial classes of the PAIP 2019 dataset, viable tumour and whole tumour (including stroma etc.). According to the clinical scenario of evaluating tumour extent and resection margin status in hepatocelullar carcinoma, 3x more weight is put onto the viable tumour class than on the whole tumour class (cf. Section 4.1.1 for details).

The first experiment shows a number of aspects: First and most importantly, there is a considerable improvement over the baseline U-Net by adding multi-scale input, either through ensembling or by multiple encoders as in the msY-family. The effect is found to be statistically significant in our experiments, even when tested only for five CV folds and even with a conservative correction for both the multiplicity of the pairwise comparisons and the violation of the independent samples-assumption underlying standard t-statistics.

The proposed multi-scale multi-encoder models reach the same segmentation performance as our best multi-scale ensemble, but as individual end-to-end trainable models and with much lower

Table 1

Different models from the proposed multi-scale multi-encoder family and multi-scale ensembles versus the baseline U-Net. The figures in the table depict the class-weighted Jaccard index for whole and viable tumour classes on the PAIP 2019 dataset. A $\star$ denotes statistical significance at the level of 0.05 when compared to the scale 1-U-Net. For a quick overview, the best results per split and overall are marked in bold, ignoring differences < 0.005.

Arch.	Scales	# pms.[b]	Mem.[c]	Weighted Jaccard						
				Per CV fold					Mean (95% CI)	
U-Net	1	17.804	1.066	0.814	0.754	0.686	0.735	0.729	0.744 (0.707, 0.780)	
Avg. Ens.	1, 4, 16	53.412[a]	3.506[a]	0.890	0.853	**0.836**	0.900	0.835	**0.863 (0.839, 0.887)**	$\star$
Log. Ens.	1, 4, 16	53.412[a]	3.506[a]	0.859	0.814	0.802	0.915	0.818	0.842 (0.805, 0.878)	$\star$
Maj. Ens.	1, 4, 16	53.412[a]	3.506[a]	0.883	0.838	0.770	0.875	0.816	0.836 (0.800, 0.873)	$\star$
msY-Net	1, 4	29.282	1.422	**0.902**	0.800	0.712	0.796	0.794	0.801 (0.748, 0.854)	$\star$
	1, 16	29.284	1.175	0.861	0.872	0.789	0.873	0.889	0.857 (0.826, 0.887)	$\star$
msUI-Net	1, 4	28.983	1.389	0.833	0.811	0.779	0.858	0.898	0.836 (0.800, 0.871)	$\star$
	1, 16	28.983	1.355	0.864	0.841	0.785	0.896	0.889	0.855 (0.820, 0.890)	$\star$
msYI-Net	1, 4, 16	40.460	1.398	0.847	**0.871**	0.804	0.895	**0.910**	**0.865 (0.833, 0.898)**	$\star$
msY^2-Net	1, 4, 16	40.761	1.561	0.847	**0.873**	0.795	**0.934**	0.876	**0.865 (0.825, 0.905)**	$\star$

[a] sum of three individual U-Nets (scales: 1, 4, 16).
[b] in units of one million parameters.
[c] GPU memory footprint given in gigabytes per image patch in batch.

Table 2

The best performing multi-scale architectures versus the baseline U-Net on the BACH 2018 dataset. The figures in the table depict the class-average of the Jaccard index for normal tissue, benign lesions and in situ and invasive carcinoma. A $\star$ denotes statistical significance at the level of 0.05 when compared to the scale 1-U-Net. The best results per split and overall are marked in bold, ignoring differences < 0.005.

Arch.	Scales	# pms.[b]	Mem.[c]	Weighted Jaccard						
				Per CV fold					Mean (95% CI)	
U-Net	1	17.804	1.066	0.523	0.476	0.478	0.510	0.501	0.498 (0.482, 0.514)	
Avg. Ens.	1, 4, 16	53.412[a]	3.506[a]	0.593	0.540	0.518	0.581	**0.622**	0.571 (0.538, 0.603)	$\star$
msYI-Net	1, 4, 16	40.460	1.398	**0.629**	0.569	0.488	0.633	0.553	**0.574 (0.527, 0.621)**	$\star$
msY^2-Net	1, 4, 16	40.761	1.561	0.578	**0.594**	**0.536**	**0.662**	0.524	**0.579 (0.536, 0.622)**	$\star$

[a] sum of three individual U-Nets (scales: 1, 4, 16).
[b] in units of one million parameters.
[c] GPU memory footprint given in gigabytes per image patch in batch.

resource requirements. Taking *msY^2-Net* as an example, it has only 76.3% of the parameters of a the corresponding ensemble of three U-Nets and comes with a GPU footprint reduced by 54.5% (if trained in parallel). Concerning the pairwise comparisons between different msY family architectures, our data suggest that, for the PAIP 2019 dataset, the global context patch provides more valuable information than the local context patch. However, the combination of both local and global context appears to lead to a further improvement, as the msY^2- and the msYI-Net models are consistently found amongst the top performing approaches, but this effect seems marginal and cannot be reliably detected through this study.

We next examined whether the superior performance of the multi-scale architectures translates to other datasets, tumour entities and tasks. For the task of breast lesion segmentation and differential diagnosis, the class-wise mean Jaccard indices are reported in Table 2. The results are in line with our findings on the PAIP 2019 dataset. Notably, the Jaccard indices on the BACH 2018 dataset are globally lower than on the PAIP 2019 dataset for all architectures. This observation fits the considerations in Section 3.1.2 that this is a particularly challenging task, also for the human pathologist. In addition, Table reports the performances of the same architectures as evaluated by the custom Bach metric used in the original challenge (cf. (Aresta et al., 2019) for details). The metric values for all architectures discussed in this section are in the range of the challenge results, with the multi-scale architectures, including the msY-family models and the multi-scale ensembles, reaching top performance.

Finally, Table 3 shows the performance of corresponding models applied to lymph node metastases segmentation. The results are again in line with the findings on PAIP 2019 and BACH 2018 data. However, the performance increase through the use of multiple scales is smaller in CAMELYON 2016/MM (6.8% as compared

to 16.3% for both PAIP 2019 and BACH 2018), which fits our considerations on the different nature and difficulty of these tasks (cf. Section 3.1.3).

With respect to the generalisability of these results, we further note that, later, an independent group has been able to present additional evidence in favor of these findings on yet other data (van Rijthoven et al., 2020). The group used a related but different end-to-end-trainable multi-scale model that confirms the benefit from the introduction of a context-encoder, though they train their context-encoder using an additional decoder that makes the model much heavier in terms of trainable parameters and GPU usage and voids the resource improvements as compared to multi-scale ensembles.

5.2. Context classification loss-based deep guidance for context encoder training

In order to examine whether the additional classification loss is indeed helpful for "guidance" of the global context encoder during training, we re-examined the msY^2-, the msY$_{(16)}$- and the msYI-Net models on the PAIP 2019 dataset when trained with or without the additional classification loss.

From the results in Table 4, it appears that the additional use of the classification loss consistently improves the model training. Recalling that the relative sizes of the detail and the global context encoder input patches (cf. Fig. 1) translate to a corresponding center-crop operation on the context encoder feature maps (as part of the multi-scale merge block, cf. Fig. 2a), this behaviour might be understood intuitively as the classification loss helping the context encoder acquire additional gradients for training.

Importantly, this technique comes with a very moderate overhead both in terms of the number of additional parameters and the effective GPU memory footprint.

Table 3

The best performing multi-scale architectures versus the baseline U-Net on the CAMELYON 2016/MM dataset. A ($\star$) denotes statistical significance at the level of 0.05 when individually compared to the scale 1-U-Net (not correcting for multiple testing). For the comparisons of other models to the baseline U-Net, the null hypothesis cannot be rejected at 0.05. The best results per split and overall are marked in bold, ignoring differences < 0.005.

Arch.	Scales	# pms.[b]	Mem.[c]	Jaccard (lymph node metastases)						
				Per CV fold					Mean (95% CI)	
U-Net	1	17.804	1.066	0.696	0.770	0.805	0.903	0.803	0.796 (0.737, 0.854)	
Avg. Ens.	1, 4, 16	53.412[a]	3.506[a]	0.748	**0.944**	0.798	0.923	0.883	**0.859 (0.794, 0.924)**	
msYI-Net	1, 4, 16	40.460	1.398	0.710	0.756	0.822	0.874	0.859	0.804 (0.750, 0.859)	
msY^2-Net	1, 4, 16	40.761	1.561	**0.775**	0.804	**0.865**	0.918	**0.889**	0.850 (0.804, 0.897)	($\star$)

[a] sum of three individual U-Nets (scales: 1, 4, 16).
[b] in units of one million parameters.
[c] GPU memory footprint given in gigabytes per image patch in batch.

Table 4

Context classification loss guidance for training of the global context encoder. Numbers are the class-weighted Jaccard indices for the PAIP 2019 dataset. The best results per pairwise comparison and split are marked in bold, where differences < 0.005 are ignored.

Arch.	Clss. loss	# pms.[a]	Mem.[b]	Weighted Jaccard					
				Per CV fold					Mean (95% CI)
msY^2-Net	with	40.761	1.561	0.847	**0.873**	**0.795**	**0.934**	**0.876**	**0.865 (0.825, 0.905)**
	without	40.759	1.436	**0.903**	0.851	0.737	0.826	0.779	0.819 (0.769, 0.869)
msY$_{16}$-Net	with	29.284	1.281	0.861	**0.872**	**0.789**	**0.873**	**0.889**	**0.857 (0.826, 0.887)**
	without	29.282	1.175	**0.883**	0.847	0.713	0.715	0.764	0.785 (0.724, 0.845)

[a] in units of one million parameters.
[b] GPU memory footprint given in gigabytes per image patch in batch.

Table 5

Spatial alignment at multi-scale path fusion. Numbers are the class-weighted Jaccard indices for the PAIP 2019 dataset. All models are variants of the msY^2 architecture, either with spatial alignment at the multi-scale merge block (1, as per default) or without (2, 3). The best results per pairwise comparison and split are marked in bold, where differences < 0.005 are ignored.

msY^2-Net variant	Weighted Jaccard					
	Per CV fold					Mean (95% CI)
spatially aligned[1]	**0.870**	**0.835**	**0.796**	**0.902**	**0.865**	**0.854 (0.822, 0.885)**
non-aligned, init[2]	**0.872**	0.814	0.725	0.822	0.827	0.812 (0.770, 0.854)
non-aligned, random[3]	**0.875**	0.787	0.777	0.808	0.832	0.816 (0.785, 0.847)

5.3. Necessity for spatial alignment in multi-scale merge blocks

Drawing on theoretical considerations, we have constructed the multi-scale merge blocks such that spatial relationships between the different paths are preserved upon fusion. Table 5 repeats the results for the msY^2-Net and compares them to those of two variants of the same architecture which both do *not* adhere to this condition. These are constructed by concatenating the bottleneck feature maps from the different paths without any spatial alignment as part of the multi-scale merge block, where the two variants differ in that the merging 1×1 convolution is either initialised randomly or with pre-defined weights leaving the main path unperturbed at training startup (cf. Section 4.2 for details).

It can be seen that both these variants perform consistently worse than the proposed msY^2-Net with correct spatial alignment in the multi-scale merge block. This is irrespective of whether the weights of the 1×1 convolution are randomly initialised or whether they are initialised such that the ResNet18 encoder of the underlying U-Net is left untouched by the un-aligned merge connections. This suggests that the deficit through the missing alignment step may be architectural rather than only a disturbance of the pretrained encoder by the additional 1×1 convolution. We therefore conclude that, when merging paths at the bottleneck level in a manner as done by the multi-scale merge block, spatial matching is a necessary step.

5.4. Additional path fusions can be added by introduction of multiple multi-scale merge blocks

Table 6 compares the standard msY^2-, the msY$_{(16)}$- and the msYI-Net with multi-scale merge blocks only at the bottleneck to their respective analogues with multi-scale merge connections at every level of the encoder.

Our study cannot find relevant differences between single-merge and multiple-merge setups. However, the additional merge connections did at least not lead to any instability in model training. Moreover, they come with a moderate increase in the number of trainable parameters and GPU memory requirements only. Therefore, multiple path fusions can be readily included in msY family models and can be used for further optimisation, e.g. by systematic neural architecture searches.

6. Discussion and conclusions

Using the segmentation of carcinoma in hematoxylin-eosin (H&E) stained whole-slide images 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, offers significant and relevant improvements over baseline single-scale U-Nets as the de facto standard in histopathology image segmentation. The improvement has been consistently

Table 6

Multi-level multi-scale merging: For three different multi-scale architectures, the standard variant with multi-scale fusion by a single multi-scale merge block only at the bottleneck ("bottleneck") is compared to an analogous architecture, but with multi-scale merge blocks at each individual encoder level ("multiple"). Numbers are the class-weighted Jaccard indices at the PAIP 2019 dataset. The best results per pairwise comparison and split are marked in bold, where differences < 0.005 are ignored.

msY^2-Net variant	Merge block(s)	# pms.[a]	Mem.[b]	Weighted Jaccard					
				Per CV fold					Mean (95% CI)
$msY_{(16)}$-	bottleneck	29.284	1.175	**0.861**	**0.872**	**0.789**	0.873	0.889	**0.857 (0.826, 0.887)**
Net	multiple	29.464	1.281	0.854	0.848	0.769	**0.881**	**0.903**	0.851 (0.811, 0.891)
$msYl$-	bottleneck	40.460	1.398	0.847	**0.871**	**0.804**	0.895	**0.910**	**0.865 (0.833, 0.898)**
Net	multiple	41.032	1.556	**0.859**	**0.872**	**0.805**	**0.911**	0.897	**0.869 (0.836, 0.901)**
$msY^2_{(Res34)}$-	bottleneck	50.869	1.516	0.870	0.835	0.796	**0.902**	0.865	0.854 (0.822, 0.885)
Net	multiple	51.140	1.608	**0.876**	**0.879**	**0.802**	0.863	**0.915**	**0.867 (0.835, 0.899)**

[a] in units of one million parameters.

[b] GPU memory footprint given in gigabytes per image patch in batch.

shown for three different datasets, clinical scenarios and tumour entities.

From a methodical and architectural perspective, our study presents a family of models that can integrate context from multiple scales and at various levels. As an overarching effect, it shows that when fusing encoder paths from different scales, spatial alignment and the preservation of spatial relationships is necessary. The proposed multi-scale merge block fulfils this requirement. The performance of the proposed architectures in terms of segmentation accuracy was further shown to be (at least) on par with ensembles of U-Nets, where the proposed multi-scale models models are end-to-end trainable systems with a reduced number of parameters and smaller memory footprint. It should, however, be noted that only common ensembling techniques have been studied, and more sophisticated techniques could potentially lead to further improvement. Moreover, different to Karimi et al. (2020b), the logistic regression-based ensembling did not lead to better performance than standard averaging of the class probabilities of the U-Net trained for the different scales. The reason remains so far unclear; potential explanations could be the different application addressed (Gleason scoring vs. tumour segmentation) and differences in implementation.

It goes without saying that future research is also likely to optimise the multi-scale models presented herein much further. As the detailed structure of encoder, decoder and possible skip connections are left entirely untouched, the proposed multi-scale architectures can seamlessly be adopted to various encoder-decoder models. In particular, increased receptive fields in the individual encoders may benefit msY family models, as these would allow the center-crop of multi-scale merge block to acquire high-level features from a larger region. Therefore, the use of dilated or atrous convolutions and atrous spatial pyramid pooling (Chen et al., 2017a; 2017b) in the context encoders might considerably benefit the models of the proposed multi-scale architecture family. Continuing with potential for improvement, we note that – as a straightforward extension of standard single scale CNN-based segmentation – we trained our networks and encoders on the original "raw" image information. This approach neglects classical work on multi-resolution image representation, which could also be advantageous in the given application context. Moreover, adding multi-level wavelet transforms to the CNN architecture has recently be shown as a very promising means to enlarge perceptive fields (Liu et al., 2018; Savareh et al., 2019), which, as said, can be particularly beneficial when implemented in the context encoders.

We have shown how the proposed building blocks and the underlying intuition can be extended to any number of arbitrarily sized spatial scales, as relevant for the particular organ and disease of interest. Moreover, additional and early merge connections may offer the possibility to model more complex relations between the different scales. We have shown that the proposed multi-scale merge block can be flexibly used at many levels, and that even

msY family models with path fusions at all possible encoder levels can be trained robustly and with minor GPU memory overhead. As both the relevant spatial scales as well as the complexity of the tasks vary between applications, we finally envision that the msY family architectures with multiple merges open up a rich environment and search space also for neural architecture searches.

Apart from the pure network architecture perspective, we used binary cross entropy as a very common, widely used and accepted loss function for the current study. The results, however, already reveal that the integration of an additional context content loss (although again implemented as cross entropy) improves performance. This suggests that the application- and/or architecture-specific loss function design bears the potential for a further improvement of WSI segmentation performance. Furthermore, we directly made use of the labels and annotations as provided by the respective challenge organisers for our experiments. It is, however, well known that the annotation of WSI slides is prone to errors and inter-observer variability. Over the past years, handling of label noise and uncertain annotation has attracted increasing interest and attention in the medical image analysis domain and in computational pathology, where, again, design of specific loss functions provides a promising approach (see Karimi et al. (2020a) for a recent review).

Concerning the limitations of this study, the results are so far based on three publicly available H&E stained WSI image datasets; it remains to be seen in future work whether the results generalise to different image data, other tasks, including tumour grading and regression tasks (e.g. for survival prediction), and other diseases. Furthermore, our analysis was built on a cross validation strategy; therefore, confirmation of our findings on separate independent test databases would clearly be desirable. Finally, whilst superior performance was consistently found for the three independent datasets, there still is a dataset dependency. In line with the underlying motivation of a human pathologist's mimicry, the data suggest that the performance increase is larger on tasks that require the human pathologist to integrate information from different scales (PAIP 2019, BACH 2018) and smaller on simpler tasks with more focus on individual cells (CAMELYON 2016/MM).

Despite the remaining limitations and potential for future work, the presented study provides clear evidence that a mimicry of how human experts approach a specific task can be successfully used to develop specialised machine learning architectures. It advocates the integration of extensive multi-scale context into deep learning models for complex tasks in computational histopathology.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Rüdiger Schmitz: Conceptualisation, Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft, Visualisation. **Frederic Madesta:** Software, Investigation, Writing - review & editing, Visualisation. **Maximilian Nielsen:** Investigation. **Jenny Krause:** Validation. **Stefan Steurer:** Validation. **René Werner:** Writing - review & editing, Supervision. **Thomas Rösch:** Supervision.

Acknowledgments

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2021.101996.

References

Aresta, G., Araújo, T., Kwok, S., Chennamsetty, S.S., Safwan, M., Alex, V., Marami, B., Prastawa, M., Chan, M., Donovan, M., Fernandez, G., Zeineh, J., Kohl, M., Walz, C., Ludwig, F., Braunewell, S., Baust, M., Vu, Q.D., To, M.N.N., Kim, E., Kwak, J.T., Galal, S., Sanchez-Freire, V., Brancati, N., Frucci, M., Riccio, D., Wang, Y., Sun, L., Ma, K., Fang, J., Kone, I., Boulmane, L., Campilho, A., Eloy, C., Polónia, A., Aguiar, P., 2019. BACH: Grand challenge on breast cancer histology images. Med. Image Anal. 56, 122–139. doi:10.1016/j.media.2019.05.010.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495. doi:10.1109/TPAMI.2016.2644615.

Bejnordi, B.E., Veta, M., van Diest, P.J., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J.A.W.M., Hermsen, M., Manson, Q.F., Balkenhol, M., Geessink, O., Stathonikos, N., van Dijk, M.C., Bult, P., Beca, F., Beck, A.H., Wang, D., Khosla, A., Gargeya, R., Irshad, H., Zhong, A., Dou, Q., Li, Q., Chen, H., Lin, H.-J., Heng, P.-A., Haß, C., Bruni, E., Wong, Q., Halici, U., Ümit Öner, M., Cetin-Atalay, R., Berseth, M., Khvatkov, V., Vylegzhanin, A., Kraus, O., Shaban, M., Rajpoot, N., Awan, R., Sirinukunwattana, K., Qaiser, T., Tsang, Y.-W., Tellez, D., Annuscheit, J., Hufnagl, P., Valkonen, M., Kartasalo, K., Latonen, L., Ruusuvuori, P., Liimatainen, K., Albarqouni, S., Mungal, B., George, A., Demirci, S., Navab, N., Watanabe, S., Seno, S., Takenaka, Y., Matsuda, H., Phoulady, H.A., Kovalev, V., Kalinovsky, A., Liauchuk, V., Bueno, G., Fernandez-Carrobles, M.M., Serrano, I., Deniz, O., Racoceanu, D., and, R.V., 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 318 (22), 2199. doi:10.1001/jama.2017.14585.

Bejnordi, B.E., Zuidhof, G., Balkenhol, M., Hermsen, M., Bult, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J., 2017. Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images. J. Med. Imaging 4 (04), 1. doi:10.1117/1.JMI.4.4.044504.

, 2010. In: Bosman, F.T., Carneiro, F., Hruban, R.H., Theise, N.D. (Eds.), WHO Classification of Tumours of the Digestive System, 3, 4. p. 209

Bouckaert, R.R., Frank, E., 2004. Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. In: Kanade, T., Kittler, J., Kleinberg, J.M., Mattern, F., Mitchell, J.C., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M.Y., Weikum, G., Dai, H., Srikant, R., Zhang, C. (Eds.), Advances in Knowledge Discovery and Data Mining, 3056. Springer, Berlin, Heidelberg, pp. 3–12. doi:10.1007/978-3-540-24775-3_3.

, 2016. In: Brierley, J.D., Gospodarowicz, M.K., Wittekind, C. (Eds.), TNM Classification of malignant tumours, 8th edition. Wiley-Blackwell.

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., 2019. Epithelium segmentation using deep learning in h&e-stained prostate specimens with immunohistochemistry as reference standard. Sci. Rep. 9 (1), 864. doi:10.1038/s41598-018-37257-4.

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., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. 25 (8), 1301–1309. doi:10.1038/s41591-019-0508-1.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. arXiv:1606.00915 [cs].

Chen, L.-C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587 [cs].

Edmondson, H.A., Steiner, P.E., 1954. Primary carcinoma of the liver. a study of 100 cases among 48,900 necropsies. Cancer 7 (3), 462–503. doi:10.1002/1097-0142(195405)7:3<462::aid-cncr2820070308>3.0.co;2-e.

Epstein, J.I., Egevad, L., Srigley, J.R., Humphrey, P.A., 2016. The 2014 international society of urological pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma. Am J Surg Pathol 40 (2), 9.

Gu, F., Burlutskiy, N., Andersson, M., Wiln, L.K., 2018. Multi-resolution Networks for Semantic Segmentation in Whole Slide Images. In: Stoyanov, D., Taylor, Z., Ciompi, F., Xu, Y., Martel, A., Maier-Hein, L., Rajpoot, N., van der Laak, J., Veta, M., McKenna, S., Snead, D., Trucco, E., Garvin, M.K., Chen, X.J., Bogunovic, H. (Eds.), Computational Pathology and Ophthalmic Medical Image Analysis. Springer International Publishing, Cham, pp. 11–18. doi:10.1007/978-3-030-00949-6_2.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. arXiv:1512.03385 [cs].

Heller, N., Isensee, F., Maier-Hein, K.H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., Yao, G., Gao, Y., Zhang, Y., Wang, Y., Hou, F., Yang, J., Xiong, G., Tian, J., Zhong, C., Ma, J., Rickman, J., Dean, J., Stai, B., Tejpaul, R., Oestreich, M., Blake, P., Kaluzniak, H., Raza, S., Rosenberg, J., Moore, K., Walczak, E., Rengel, Z., Edgerton, Z., Vasdev, R., Peterson, M., McSweeney, S., Peterson, S., Kalapara, A., Sathianathen, N., Papanikolopoulos, N., Weight, C., 2019. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge. doi:10.1016/j.media.2020.101821.

Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., Maier-Hein, K.H., 2018. Nnu-net: self-adapting framework for u-net-based medical image segmentation. arXiv:1809.10486 [cs].

Isensee, F., Petersen, J., Kohl, S.A.A., Jger, P.F., Maier-Hein, K.H., 2019. Nnu-net: breaking the spell on successful medical image segmentation. arXiv:1904.08128 [cs].

Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3d CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. 36, 61–78. doi:10.1016/j.media.2016.10.004.

Karimi, D., Dou, H., Warfield, S.K., Gholipour, A., 2020. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. Med. Image Anal. 65, 101759. doi:10.1016/j.media.2020.101759.

Karimi, D., Nir, G., Fazli, L., Black, P.C., Goldenberg, L., Salcudean, S.E., 2020. Deep learning-based gleason grading of prostate cancer from histopathology images-role of multiscale decision aggregation and data augmentation. IEEE J. Biomed. Health Inform. 24 (5), 1413–1426. doi:10.1109/JBHI.2019.2944643.

Kickingereder, P., Isensee, F., Tursunova, I., Petersen, J., Neuberger, U., Bonekamp, D., Brugnara, G., Schell, M., Kessler, T., Foltyn, M., Harting, I., Sahm, F., Prager, M., Nowosielski, M., Wick, A., Nolden, M., Radbruch, A., Debus, J., Schlemmer, H.-P., Heiland, S., Platten, M., von Deimling, A., van den Bent, M.J., Gorlia, T., Wick, W., Bendszus, M., Maier-Hein, K.H., 2019. Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. The Lancet Oncology 20 (5), 728–740. doi:10.1016/S1470-2045(19)30098-1.

Kim, Y.J., Jang, H., Lee, K., Park, S., Min, S.-G., Hong, C., Park, J.H., Lee, K., Kim, J., Hong, W., Jung, H., Liu, Y., Rajkumar, H., Khened, M., Krishnamurthi, G., Yang, S., Wang, X., Han, C.H., Kwak, J.T., Ma, J., Tang, Z., Marami, B., Zeineh, J., Zhao, Z., Heng, P.-A., Schmitz, R., Madesta, F., Rsch, T., Werner, R., Tian, J., Puybareau, E., Bovio, M., Zhang, X., Zhu, Y., Chun, S.Y., Jeong, W.-K., Park, P., Choi, J., 2021. Paip 2019: liver cancer segmentation challenge. Med. Image Anal. 67, 101854. doi:10.1016/j.media.2020.101854.

Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv:1412.6980.

Lakhani, S., Ellis, I., Schnitt, S., Tan, P., van de Vijver, M. (Eds.), 2012. WHO Classification of Tumours of the Brest. Vol. 4. 4 edition.

Li, J., Sarma, K.V., Chung Ho, K., Gertych, A., Knudsen, B.S., Arnold, C.W., 2018. A multi-scale u-net for semantic segmentation of histological images from radical prostatectomies. AMIA Annual Symposium Proceedings 2017, 1140–1148.

Li, S., Tso, G.K.F., 2019. Bottleneck supervised u-net for pixel-wise liver and tumor segmentation. arXiv:1810.10331 [cs].

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.-W., Heng, P.A., 2018. H-Denseunet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. arXiv:1709.07330 [cs].

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., Snchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88. doi:10.1016/j.media.2017.07.005.

Liu, P., Zhang, H., Zhang, K., Lin, L., Zuo, W., 2018. Multi-level Wavelet-CNN for Image Restoration. In: 2018 IEEE/CVF Conference on Computer Vision and Pat-

tern Recognition Workshops (CVPRW). IEEE, Salt Lake City, UT, pp. 886–88609. doi:10.1109/CVPRW.2018.00121.

Liu, W., Rabinovich, A., Berg, A.C., 2015. Parsenet: looking wider to see better. arXiv:1506.04579 [cs].

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., 2017. Detecting cancer metastases on gigapixel pathology images. arXiv:1703.02442 13.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 3431–3440.

Maier, O., Menze, B.H., von der Gablentz, J., Hni, L., Heinrich, M.P., Liebrand, M., Winzeck, S., Basit, A., Bentley, P., Chen, L., Christiaens, D., Dutil, F., Egger, K., Feng, C., Glocker, B., Gtz, M., Haeck, T., Halme, H.-L., Havaei, M., Iftekharuddin, K.M., Jodoin, P.-M., Kamnitsas, K., Kellner, E., Korvenoja, A., Larochelle, H., Ledig, C., Lee, J.-H., Maes, F., Mahmood, Q., Maier-Hein, K.H., McKinley, R., Muschelli, J., Pal, C., Pei, L., Rangarajan, J.R., Reza, S.M., Robben, D., Rueckert, D., Salli, E., Suetens, P., Wang, C.-W., Wilms, M., Kirschke, J.S., Krmer, U.M., Mnte, T.F., Schramm, P., Wiest, R., Handels, H., Reyes, M., 2017. ISLES 2015 - A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. Med. Image Anal. 35, 250–269. doi:10.1016/j.media.2016.07.009.

Mehta, S., Mercan, E., Bartlett, J., Weaver, D., Elmore, J.G., Shapiro, L., 2018. Y-Net: Joint Segmentation and Classification for Diagnosis of Breast Biopsy Images. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-Lpez, C., Fichtinger, G. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2018, 11071. Springer International Publishing, Cham, pp. 893–901. doi:10.1007/978-3-030-00934-2_99.

Nadeau, C., Bengio, Y., 2000. Inference for the generalization error. Adv. Neural. Inf. Process. Syst. 12, 307–313.

Ning, Z., Zhang, X., Tu, C., Feng, Q., Zhang, Y., 2019. Multiscale context-cascaded ensemble framework (msc2ef): application to breast histopathological image. IEEE Access 7, 150910–150923.

Nir, G., Hor, S., Karimi, D., Fazli, L., Skinnider, B.F., Tavassoli, P., Turbin, D., Villamil, C.F., Wang, G., Wilson, R.S., Iczkowski, K.A., Lucia, M.S., Black, P.C., Abolmaesumi, P., Goldenberg, S.L., Salcudean, S.E., 2018. Automatic grading of prostate cancer in digitized histopathology images: learning from multiple experts. Med. Image Anal. 50, 167–180. doi:10.1016/j.media.2018.09.005.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention U-Net: Learning Where to Look for the Pancreas.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in pytorch. NIPS 2017 workshop.

R Core Team, 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P., 2001. Color transfer between images. IEEE Comput. Graph. Appl. 21 (5), 34–41. doi:10.1109/38.946629.

van Rijthoven, M., Balkenhol, M., Silia, K., van der Laak, J., Ciompi, F., 2020. HookNet multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images. doi:10.1016/j.media.2020.101890.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 9351, 234–241.

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., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 211–252. doi:10.1007/s11263-015-0816-y.

Savareh, B.A., Emami, H., Hajiabadi, M., Azimi, S.M., Ghafoori, M., 2019. Waveletenhanced convolutional neural network: a new idea in a deep learning paradigm. Biomedical Engineering / Biomedizinische Technik 64 (2), 195–205. doi:10.1515/bmt-2017-0178.

Schmitz, R., Krause, J., Krech, T., Rsch, T., 2018. Virtual endoscopy based on 3-dimensional reconstruction of histopathology features of endoscopic resection specimens. Gastroenterology 154 (5), 1234–1236.e4. doi:10.1053/j.gastro.2017.11.291.

Segovia-Miranda, F., Morales-Navarrete, H., Kcken, M., Moser, V., Seifert, S., Repnik, U., Rost, F., Brosch, M., Hendricks, A., Hinz, S., Rcken, C., Ltjohann, D., Kalaidzidis, Y., Schafmayer, C., Brusch, L., Hampe, J., Zerial, M., 2019. Three-dimensional spatially resolved geometrical and functional models of human liver tissue reveal new aspects of NAFLD progression. Nat. Med. 25 (12), 1885–1893. doi:10.1038/s41591-019-0660-7.

Vu, Q.D., Kwak, J.T., 2019. A dense multi-path decoder for tissue segmentation in histopathology images. Comput Methods Programs Biomed doi:10.1016/j.cmpb.2019.03.007.

Wang, S., Zhu, Y., Yu, L., Chen, H., Lin, H., Wan, X., Fan, X., Heng, P.-A., 2019. RMDL: Recalibrated multi-instance deep learning for whole slide gastric image classification. Med Image Anal 101549. doi:10.1016/j.media.2019.101549.

Wetteland, R., Engan, K., Eftestol, T., Kvidstad, V., M., J.E.A., 2019. Multiscale deep neural networks for multiclass tissue classification of histological whole-slide images. Medical Imaging with Deep Learning (MIDL) 2019.

Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., Agrawal, A., 2018. Context encoding for semantic segmentation. arXiv:1803.08904 [cs].

Zhou, Y., Sun, X., Zha, Z.J., Zeng, W., 2019. Context-reinforced semantic segmentation. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 4046–4055.

Zink, D., Fischer, A.H., Nickerson, J.A., 2004. Nuclear structure in cancer cells. Nat. Rev. Cancer 4 (9), 677–687. doi:10.1038/nrc1430.