U-Net#: Nuclei Segmentation in Histopathology Images

Mindaugas Kazimieras Stagys
Faculty of Mathematics and Informatics
Vilnius University
Vilnius, Lithuania
mindaugas.stagys@mif.stud.vu.lt

***Abstract*—Digital pathology is one of the most significant developments in modern medicine. Pathological examinations are the gold standard of medical protocols and play a fundamental role in diagnosis. Recently, with the advent of digital scanners, tissue histopathology slides can now be digitized and stored as digital images. As a result, digitized histopathological tissues can be used in computer-aided image analysis programs and machine learning techniques. Detection and segmentation of nuclei are some of the essential steps in the diagnosis of cancers. In this paper, we propose U-Net#, an improved architecture for medical image segmentation. The proposed architecture uses SWSL ResNext as an encoder and utilizes full-scale skip connections to combine low-level features with high level semantic information. Also, the attention modules are added to suppress irrelevant regions. Finally, atrous spatial pyramid pooling is used to account for different object scales. We demonstrate the effectiveness of the proposed model on PanNuke dataset, achieving 55.93% average IoU and 71.74% average Dice coefficient.**

***Index Terms*—Computational pathology, Histopathology, UNet, Nuclei segmantation, PanNuke**

I. Introduction

Microscopy is a central technology of biomedical research. Various microscopy techniques allow capturing structural and functional properties of biological model systems, including cultured cells, tissues and organoids. As microscopy makes progress to capture such systems in greater detail and throughput and as the development of novel assays reveals more complex properties of living organisms, the need for robust and easy to use microscopy image analysis methods becomes critical to answer a wider variety of biological questions.

Many image analysis workflows involve the segmentation of cell nuclei as a first step to extract meaningful biological signals. Research studies may involve counting cells, tracking moving populations, localizing proteins and classifying phenotypes or profiling treatments; in all of these and more, the nucleus is a reliable compartment of reference for identifying single cells in microscopy images.

However, selecting strategies to segment nuclei is not an easy task for nonexpert users in regular biology labs. Most existing user-friendly bioimage analysis tools identify nuclei using classical segmentation algorithms such as thresholding, watershed or active contours. These need to be configured for each study to account for different microscopy modalities, scales and experimental conditions, often requiring great expertise to select the algorithm that suits the problem and to adjust its parameters. For advanced users, the choice can also be daunting, considering that hundreds of papers are published every year presenting new methods for cell and nucleus segmentation. And even under controlled experimental conditions, no single parameter choice can segment all images correctly, because classical algorithms can fail to adapt to the heterogeneity of biological samples or can be sensitive to technical artifacts. Altogether, this situation slows down the pace of research and hinders biological laboratories from adopting imaging technologies owing to the time and expertise required.

II. Related Work

Deep learning, particularly convolutional neural networks, shows promising results in automatic segmentation for a variety of medical applications. A popular deep learning architecture in the field of semantic segmentation for biomedical applications is U-Net [1]. UNet's symmetric, encoder–decoder architecture allows automatic learning of features at different levels. Nevertheless, low-level features learned in the encoder are rich with feature space information but lack semantic information, and high-level features learned from the decoder are the opposite. Thus, the direct concatenation of these features may not produce the most optimal results. Researchers have already offered plenty of improvement schemes based on U-Net. For example, UNet++ [2] further strengthens skip connections by introducing nested and dense skip connections, aiming at reducing the semantic gap between the encoder and decoder. UNet 3+ [3] use the multi-scale features by introducing full-scale skip connections, which incorporates low-level details with high-level semantics from feature maps in full scales, but with fewer parameters. An addition of residual connections in ResUNet [4] architecture propagates information over layers, allowing building of deeper neural networks and reducing the impact of exploding or vanishing gradients, which ultimately alleviates training performance. Moreover, contrary to UNet, architectures such as RefineNet [5] or SegNet [6] only use the highest layer features while they lack in low-level representation. This was further improved in DeepLabV3 [7] and ResUNet++ [8], where extracted features are passed through the Atrous Spatial Pyramid Pooling (ASPP) module to obtain multi-scale information [9]. It was also realized that not all features are equally important. The attention mechanism dynamically allocates the input weights of neurons to selectively focus on the most critical part of the information [10] and is often introduced into U-Net type architectures to improve their performance [11], [12].

III. Methods

A. Stain Normalization

H&E images are the result of applying two stains to a tissue sample: hematoxylin and eosin. The hematoxylin binds to the cell nuclei and colors them purple, while the eosin binds to the cytoplasm and extracellular matrix, coloring them pink. Stain deconvolution is the process of untangling these two superimposed stains from an H&E image.

Digital pathology images can vary for many reasons, including:

- variation in stain intensity due to inconsistencies of technicians while applying stains to specimens;
- variation in image qualities due to differences in slide scanners;
- variation due to differences in lighting conditions when slide is scanned.

For these reasons, the proposed framework includes color normalization as the first data preprocessing step.

There are several traditional approaches which try to normalize the color space by estimating a color deconvolution matrix that allows identifying the underlying stain [13], [14], [15]. In this work we decided to use stain normalization technique proposed by Macenko et al. [13].

B. SWSL ResNeXt as Backbone

ResNeXt [16] is a homogeneous neural network which reduces the number of hyperparameters required by conventional ResNet. ResNeXt repeats a building block that aggregates a set of transformations with the same topology. Compared to a ResNet, it exposes a new dimension, cardinality (the size of the set of transformations), as an essential factor in addition to the dimensions of depth and width.

The semi-weakly supervised (SWSL) [17] framework suggests that a highly-accurate teacher model is first pre-trained with weakly supervised data sets, and then subsequently fine-tuned with all available labeled data. This teacher network is then used to assign “more accurate” scores or class predictions for the weakly supervised data that was previously pre-trained with. Next, the target student model is pre-trained with the weakly supervised data with improved labels from the teacher. Finally, the labeled data are used to fine-tune the student model.

SWSL ResNext101 32×4d was selected to be used as an encoder part of our proposed architecture. Initial weights were pre-trained on IG-1B-Targeted and ImageNet datasets. In the first step we train the decoder part for 5 epochs with encoder layers being frozen. Then all layers except batch normalization are unfreezed to fine-tune the encoder for our specific task.

C. Attention Gates

Attention mechanism firstly emerged in the natural language processing tasks and quickly gained dominance. Attention gates were successfully integrated into U-Net type architectures [11], [12] and improved segmentation performance while preserving computational efficiency.

The inputs of the attention gate are the upsampling features in the expansion path and the corresponding features from the encoder. The former one is used as a gating signal to enhance the learning of the target area related to the

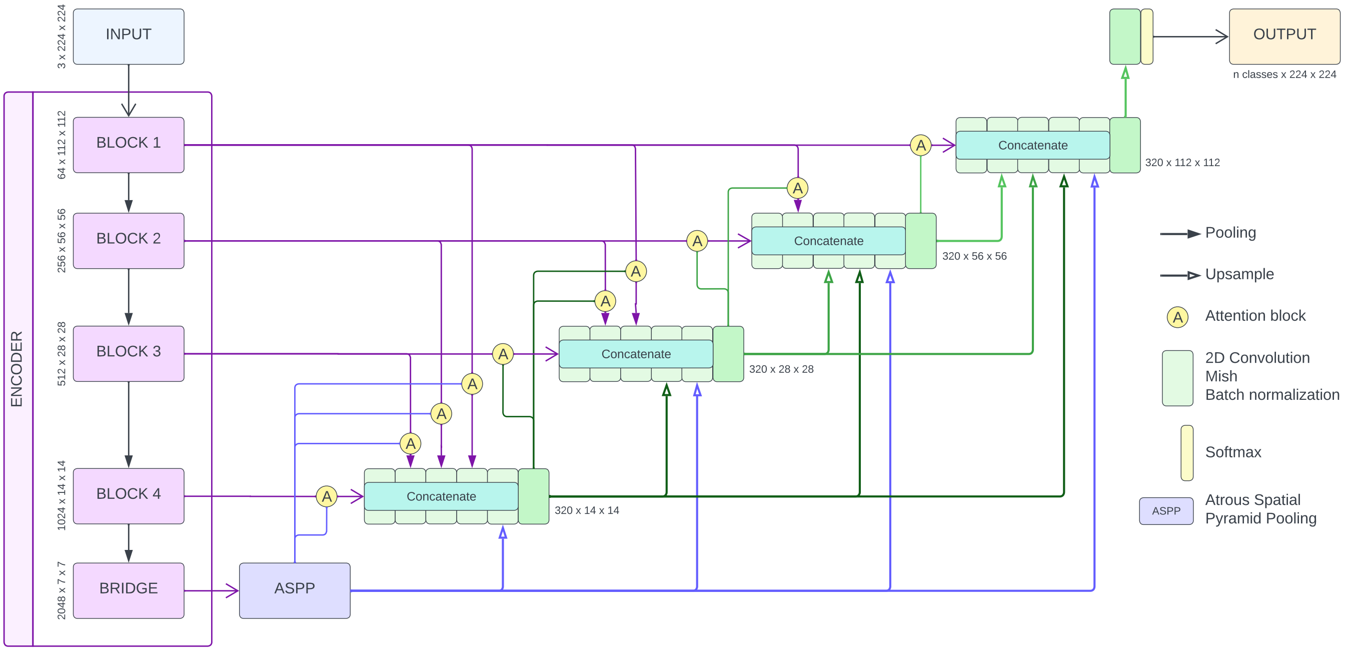

Figure 1 – U-Net# Architecture

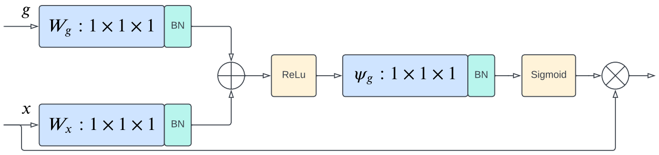
g
$W_g : 1 \times 1 \times 1$
BN
x
$W_x : 1 \times 1 \times 1$
BN
ReLu
$\psi_g : 1 \times 1 \times 1$
BN
Sigmoid

Figure 2 – Attention mechanism

segmentation task while suppressing the area irrelevant in the task. Therefore, attention gate can improve the efficiency of propagating semantic information through skip connections. Next, the sigmoid activation function is selected to train the convergence of the parameters in the gate and to get the attention coefficient α. Finally, the output can be obtained by multiplying the encoder feature by coefficient α pixel by pixel.

D. Atrous Spatial Pyramid Pooling (ASPP)

The idea of ASPP comes from spatial pyramid pooling, which is successful at re-sampling features at multiple scales. In ASPP, the contextual information is captured at various scales and many parallel atrous convolutions with different rates in the input feature map are fused. Atrous convolution allows controlling the field-of-view for capturing multi-scale information precisely.

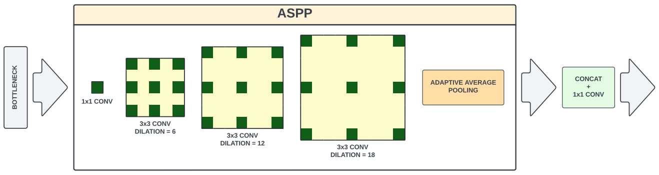

Figure 3 – Atrous Spatial Pyramid Pooling

In the proposed architecture, ASPP acts as a bridge between encoder and decoder. As in deepLabV3 [7], the ASPP consists of one 1×1 convolution and three 3×3 convolutions with dilation rates 6, 12 and 18.

E. Activation Function

In this work we use Mish [18] which is a novel smooth and non-monotonic neural activation function which can be defined as:

$$f(x) = x \cdot \tanh\left(\ln\left(1 + \mathrm{e}^{x}\right)\right) \qquad (1)$$

Advantages of Mish:

- Being unbounded above avoids saturation which generally causes training to drastically slow down due to near-zero gradients [19].
- Being bounded below results in strong regularization effects.
- The non-monotonic property of Mish causes small negative inputs to be preserved as negative outputs, which improves expressivity and gradient flow.
- The order of continuity being infinite for Mish is also a benefit over ReLU since ReLU has an order of continuity as 0 which means it is not continuously differentiable causing some undesired problems in gradient-based optimization.

IV. Data and Preprocessing

PanNuke [20], [21] is the largest and the most diverse to date dataset for nucleus segmentation and classification, that has been annotated in a semi-automated manner and quality-controlled by clinical professionals. The dataset consists of nuclei labels across 19 different tissue types, 481 visual fields, of which 312 are randomly sampled from more than 20K whole slide images at different magnifications, from multiple data sources. In total the dataset contains 205,343 labeled nuclei, categorized into 5 clinically important classes each with an instance segmentation mask. The authors of PanNuke dataset splitted the data into training, validation and test folds which contains 2656, 2523 and 2722 pairs of images and masks respectively. For every fold every tissue was splitted into three sections by ensuring that each contains an equal portion of the smallest class within it.

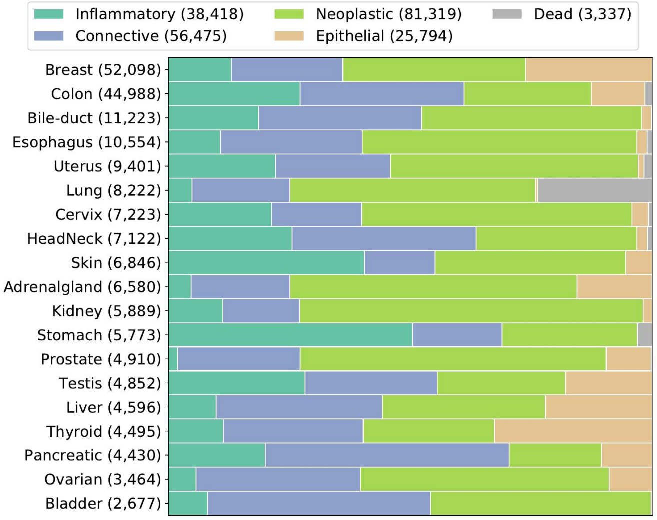

Figure 4 – A comparative plot of class distributions per tissue. Numbers in parenthesis represent the total number of nuclei within that category or tissue type.
Source: Adapted from [21].

Heavy data augmentations have been randomly applied to prevent overfitting. Augmentations include cropping, rotation, flipping, transposition, Gaussian noise and additive Gaussian noise, motion blur and median blur, shifting and scaling, various distoritions and affine transformations, change of sharpness, contrast and brightness. All images were resized to size 224×224.

V. Training

A. Loss Function

Based on performance comparison of various loss functions [22], we decided to use Focal Tversky loss.

Tversky loss [23] was designed to optimise segmentation on imbalanced medical datasets by utilising constant β that adjusts how harshly different types of error are penalised in the loss function. Tversky loss is defined as:

$$\mathrm{TL}(p, \hat{p}) = \frac{p\hat{p}}{p\hat{p} + \beta\left(1-p\right)\hat{p} + \left(1-\beta\right)p\left(1-\hat{p}\right)}, \qquad (2)$$

where $p \in [0, 1]$ and $p \in \hat{[}0, 1]$ represent the ground truth label and the prediction probability, respectively. With $\beta = 0.5$, this loss becomes equivalent to Dice Loss.

Focal Tversky loss [24] is an extension of Focal loss and Tversky loss functions. Similar to Focal loss, it focuses on hard examples, such as with small region of interest, by down-weighting easy/common ones with a help of γ coefficient:

$$\mathrm{FTL} = \sum_{c} \left(1 - FL_c\right)^{\gamma}, \qquad (3)$$

where c represents classes. When $\gamma = 1$, the FTL simplifies to the TL. Based on previous experiments [24], we set coeffiecients $\beta = 0.7$ and $\gamma = \frac{4}{3}$.

B. Optimizer

Adabound [25] optimizer was chosen to train the model for 50 epochs. AdaBound can be regarded as an adaptive method at the beginning of training, and thereafter it gradually and smoothly transforms to SGD as the time step increases. It have been shown that Adabound can eliminate the generalization gap between adaptive methods and SGD and maintain higher learning speed early in training at the same time.

Table I – Adabound hyperparameters

lr	betas	final lr	gamma	eps	weight decay
1e-4	[0.9, 0.999]	0.1	1e-3	1e-8	0

UNet# has 57.2M trainable parameters.

VI. Results

For the evaluation we used Intersection over Union (IoU), also known as the Jaccard index, and the Dice similarity coeficient metrics.

Table II – Results

	IoU	Dice
Neoplastic	0.6349	0.7767
Inflammatory	0.4644	0.6343
Connective	0.4112	0.5828
Dead	0	0
Epithelial	0.5500	0.7096
Background	0.9132	0.9546
Average w/o background	**0.5593**	**0.7174**

It seems that most segmentation errors are caused by either misclassification of nuclei type or unclear boundaries between nuclei that are close to each other. Due to the data imbalance, there were no predictions for the dead nuclei.

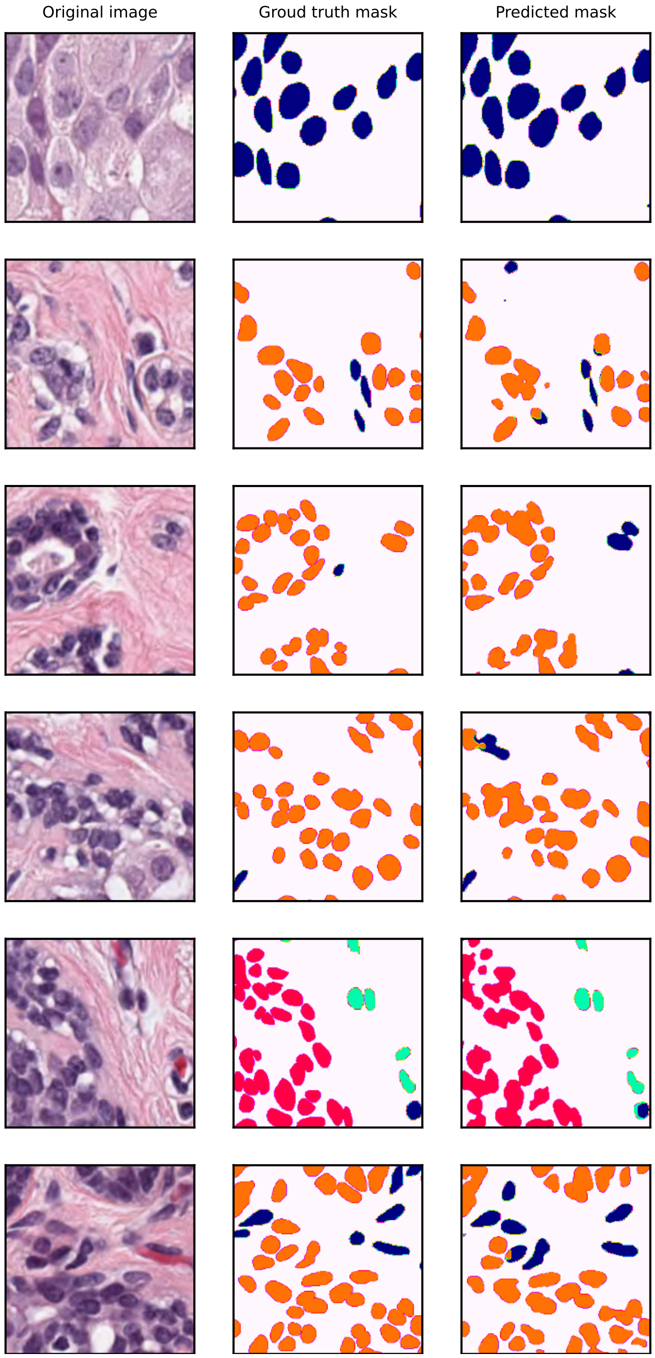

Figure 5 – Test fold images, masks and predictions

VII. Conclusions

In this paper, we presented an end-to-end workflow for the multi-class semantic nuclei segmentation in histopathology images. The proposed architecture successfully combines pre-trained SWSL ResNext encoder and methods such as Macenko stain normalization, full-scale skip connections, attention mechanism, ASPP, Focal Tversky loss and Mish activations. Further investigations could be done by hyperparameter tuning, comparing the proposed framework against various state-of-art architectures, training the model on different datasets or fusing nuclear morphology information with other features for networks thatcan improve detection, classification, grading and prognosis from histopathology images.

VIII. Acknowledgment

We would like to acknowledge IT Research center of Vilnius University, Faculty of Mathematics and Informatics for the provided HPC resources.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[2] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," 2018.

[3] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," 2020.

[4] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, apr 2020.

[5] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," 2016.

[6] V. Badrinarayanan, A. Handa, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," 2015.

[7] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017.

[8] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. de Lange, P. Halvorsen, and H. D. Johansen, "Resunet++: An advanced architecture for medical image segmentation," 2019.

[9] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[11] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," 2018.

[12] C. Li, Y. Tan, W. Chen, X. Luo, Y. Gao, X. Jia, and Z. Wang, "Attention unet++: A nested attention-aware u-net for liver ct image segmentation," in *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 345–349, 2020.

[13] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas, "A method for normalizing histology slides for quantitative analysis," in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 1107–1110, 2009.

[14] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Computer Graphics and Applications*, vol. 21, no. 5, pp. 34–41, 2001.

[15] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter, I. Esposito, and N. Navab, "Structure-preserving color normalization and sparse stain separation for histological images," *IEEE transactions on medical imaging*, vol. 35, no. 8, pp. 1962–1971, 2016.

[16] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," 2016.

[17] I. Z. Yalniz, H. Jégou, K. Chen, M. Paluri, and D. Mahajan, "Billion-scale semi-supervised learning for image classification," 2019.

[18] D. Misra, "Mish: A self regularized non-monotonic activation function," 2019.

[19] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (Y. W. Teh and M. Titterington, eds.), vol. 9 of *Proceedings of Machine Learning Research*, (Chia Laguna Resort, Sardinia, Italy), pp. 249–256, PMLR, 13–15 May 2010.

[20] J. Gamper, N. A. Koohbanani, K. Benet, A. Khuram, and N. Rajpoot, "Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification," in *European Congress on Digital Pathology*, pp. 11–19, Springer, 2019.

[21] J. Gamper, N. A. Koohbanani, S. Graham, M. Jahanifar, S. A. Khurram, A. Azam, K. Hewitt, and N. Rajpoot, "Pannuke dataset extension, insights and baselines," *arXiv preprint arXiv:2003.10778*, 2020.

[22] S. Jadon, "A survey of loss functions for semantic segmentation," in *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, IEEE, oct 2020.

[23] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3d fully convolutional deep networks," 2017.

[24] N. Abraham and N. M. Khan, "A novel focal tversky loss function with improved attention u-net for lesion segmentation," 2018.

[25] L. Luo, Y. Xiong, Y. Liu, and X. Sun, "Adaptive gradient methods with dynamic bound of learning rate," 2019.