Nuclei Segmentation: Hyperparameter-Tuned U-Net DLA Analysis using Modified RandAugment

Joshua Bercich[1]

[1] *University of Sydney, Camperdown, Australia*

November 6, 2023

1 Introduction

Innovation in biomedical deep-learning research is challenging due to sparse data availability, ethical access, and wide contextual variety. Computer vision has the capacity to greatly impact prominent obstacles for biomedicine and support major medical developments. Leveraging image data from brain MRI-scans or cross-sectional pathological microscopies creates a range of challenges such as instance segmentation of malignant tumours or isolating certain nuclei categories and high-threshold detection counts [1]. The latter is pertinent for constructing high-throughput monitoring technologies to detect unexpected blood diseases or support biomedical research.

Major limitations in biomedical deep-learning include the availability of low-scale ethically sourced datasets, and flexibly efficient models with high transferability across different learning tasks [1], [2]. By focussing on instance segmentation for various phenotype nuclei provided by the BBBC039 dataset, this study seeks to address two aims: (1) demonstrate efficient tuning and rapid iteration of the U-Net Deep-Layer Aggregation (DLA) model architecture [3], [4], and (2) evaluate effective segmentation augmentation policies for small datasets. Utilising a small effective segmentation network opens rapid deployment and refinement of deep-learning solutions and suitably benchmarks further research within the field. Tuned augmentation, such as RandAugment, mitigates the issue of low data accessibility and enriches model training effectiveness through regularisation.

The combination of a segmentation-modified RandAugment policy strategy with the U-Net DLA model demonstrated impressive results by outperforming known research for binary nuclei semantic segmentation and presents a new benchmark for 3-class boundary instance segmentation. Efficient loss convergence highlighted the capacity of rapid development, and outperforming existing research indicates the importance of the results from this research [2]. The final selected model achieved 3-/2-class test metrics of 90.37/65.39% accuracy, 97.71/98.77 Dice, 84.37/97.96 AJI, and 90.60/65.98 F_1. This improvement indicates a new architecture for multi-task challenges, with proven success for instance segmentation of nuclei.

2 Methods

2.1 Previous Work

Research of nuclei segmentation models is sparse and deterministically tuned according to the specific dataset. Stacked U-Nets (SU-Nets) have been used for histopathological ensemble architecture producing binary inferences from aggregate overlap masks [5]. Another study of different models showed the success of U-Net variants, such as ASW-Net and U-Net DLA, for nuclei segmentation on the same BBBC039 dataset, but required post-processing methods for comparative results (79.806 versus 90.200 AJI) [2]. Additionally, U-Net DLA achieved $+0.023$ F_1 and second best -0.01 AJI to the SU-Net ensemble highlighting practical relevance [2].

Model Selection Segmentation has seen increased attention recently through newer architectures such as latest YOLO versions [6], or the multi-task capable Segment-Anything-Model (SAM) which can perform object detection and segmentation [7]. Due to the use of highly dense ResNet and GPT backbones, networks are significantly larger as SAM requires 100ms for inferences making it infeasible for rapid iterations [7]. Aforementioned U-Net architectures are efficient with proven success as the Fast-SAM model backbone [8]. With fewer classes, larger networks are not necessary particularly with low-scale datasets [2], [8]. Research of segmentation architectures reinforced U-Net applicability as top-1 performance was attained by Sharp U-Net and a default U-Net model achieved second-best for nuclei segmentation (-0.03 DICE; $+1.09\%$ accuracy) [9].

Data Augmentation Policies To enforce regularisation and extend the utility of low-scale datasets, image augmentation is required. Ideal for multiclass learning, policies are difficult to select and have expensive costs when balancing bias and variance. AutoAugment uses an RL learning method that rewards performance [10]. It reduced CIFAR100 classification error by $3.3 \pm 0.2\%$, and determines optimal policies for datasets [10]. RandAugment substitutes RL controllers for hyperparameters n, number of augment operations, and m, scaling magnitude of the operations [11]. The reduced Bayesian search space achieves near-equivalent results decreasing 10^{34} search iterations to 10^2 making it more effective for rapid development [11].

2.2 Experimental Methodology

The BBBC039 dataset was created using a Hoechst stain and fluorescence microscopy on high-throughput U2OS histopathological nuclear phenotypes from 200 different fields of view [1]. Each image contains labelled nuceli with annotated overlapping regions for distinguishing entire nuclei boundaries within a 16-bit 520×696 frame. Statically defined metadata defined a 50%/25%/25% training/validation/testing split which was maintained [1].

2.2.1 Dataset Preprocessing

Images were normalised from 16-bit grayscale RGBA channels to 8-bit pixels. Label masks were spliced to the single colour channel and dropped the image alpha. Micronuclei were removed using a threshold pixel size of $\phi = 50$. Marked nuclei boundaries were extracted and a 3-channel output mask was constructed corresponding to different classes: $p_{bkgrnd} = 0, p_{nuclei} = 1$, and $p_{bndry} = 2$ (see Figure 1). Raw 520×696 frames were too large for rapid iterations so an $n_{split} = 232$ square pixel side length split frames into 6 separate instances which were retained within each subset. The final dataset paritions held 600/300/300 observations.

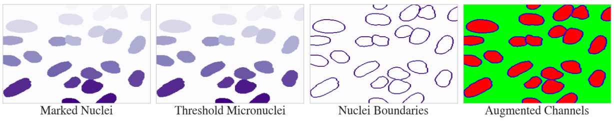


Figure 1: *Preprocessing stages of BBBC039 dataset labels; individual label masking, thresholding, boundary extraction and channel annotation.*

2.2.2 Experimental Method

The chosen U-Net DLA model was built for brain MRI abnormality detection [12]. Four paired convolutional (3×3) and max-pooling (2×2) layers comprise the encoder and decoder with channel dimensions determined by an initial feature hyperparameter [12]. DLA residuals connect respective encoder-decoder blocks without any combined recurrent residual blocks [12].

Model training used Adam optimisation with tuned learning rate $\gamma \in \{1 \times 10^{-3}, 1 \times 10^{-4}\}$ and Cross-Entropy loss. Evaluative metrics apply any softmax/sigmoid operation post-inference. For each model ablations, a batch size of $B = 8$ and epochs $E = 10$ were used. Ablations for the learning rate γ and initial feature channels $c_{1;in} \in \{16, 32, 64\}$ were performed.

For each iteration, a segmentation modified RandAugment, suitable for C-channelled image and label pairs (I_i, L_i) was implemented with hyperparameters n, m [11]. Operations included rotation, shear, translate, rescaling, horizontal/vertical flip, and centre crop; each with manually tuned parameters determined by $m \in (0, 1)$. Augmentations use nearest-pixel interpolation and do not impact colourisation. This introduced the additional search spaces $n \in \{2, 3, 4, 5\}$ and $m \in \{0.1, 0.25, 0.5\}$. Early stopping was set for validation F_1 for $E = 5$ with variance $\delta = 0.01$.

Models were trained using a NVIDIA T4 GPU[1] with single-core 13.5GB memory CPUs; hyperparameter tuning required approximately 9.5 hours of end-to-end compute time.

2.2.3 Model Evaluation

Evaluation of different models used several metrics to understand the performance of and general utility of each ablation. The multi-class average F_1 describes predictive power respective to minimising false positives.

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{1}$$

Another common segmentation metric is the DICE coefficient, similar to F_1 but designed for pixel-set overlap comparison rather than exact pixel-wise performance that the F_1 metric provides.

$$\text{Dice} = 2 \times \frac{TP}{TP + FP + FN} \tag{2}$$

[1] Provided by the Google Cloud Platform with accessible CUDA cores for faster training. Funded by Google Education Grant credits.

The Aggregate Jaccard Index (AJI) is another set-based metric exploiting pixel label overlap to describe similarity between inference I and target labels T for N different labels.

$$AJI(I,T) = \frac{1}{N}\sum_{i=1}^{N}\frac{|I_i \cap T_i|}{|I_i \cup T_i|} \tag{3}$$

Additional metrics were considered such as pixel accuracy to assess the inference exactness, and the coverage error to determine class bias where one may dominate other classes. Combinations of these evaluation metrics is necessary in determining an optimal model.

3 Results

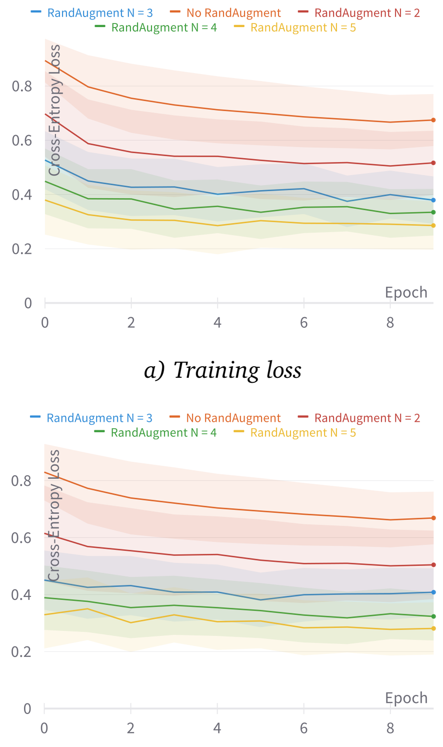
For the given experimental setting, impressive results were obtained for all aforementioned ablations. Hyperparameter tuning of the baseline model for input channels $c_{1;in}$ and learning rate γ achieved improved performance compared to leading contemporary research in 3-class Dice measures with top-97.71 and $+1.277$ gain [2]. See Table 1 for a detailed list of results across the 50 separate model variations.

3.1 Loss Convergence

Each model converged similarly for the Cross-Entropy loss as shown in Figure 2. There was no major variation between training and validation loss other than the immediate decrease indicating an efficient convergence to a localised minima. The RandAugment hyperparameter n mainly impacted the magnitude of loss due to the introduced noise from augmentation operations such as the fill space in *translate* and *rotate* creating easily inferable null classes. Loss converged slowly indicating the ease of learning nuclei features, but was constrained by the third class boundaries reinforced by the large difference between 3-/2-channel AJI results from Table 1. High variation in loss convergence is shared across all augmentation policies but there is clearly perturbations as n increases alongside less efficient decline in loss.

3.2 Hyperparameter Selection

Foundational hyperparameters for this experiment include the initial features $c_{1;in}$, learning rate γ, and modified RandAugment n, m parameters. Figure 3 shows the impact of different hyperparameters on the test Aggregate Jaccard Index (AJI). There is a clear correlation for γ preferencing higher learning rates, although Table 1 shows that top-performing models vary consistently between both values highlighting correlative dependency with other parameters. Additionally, there is a greater impact from the number of operations n over the scalar magnitude m where reduced operations, although showing variable results, lead to greater performing models. There is increased variance introduced with expressive augmentation policies which is expected by impacting convolutional bottlenecks under different $c_{1;in}$ and γ parameters.



a) Training loss

b) Validation loss

Figure 2: *Cross-entropy loss convergence for U-Net DLA grouped by RandAugment n per epoch for training and validation subsets.*

The optimal configuration for the RandAugment policy must balance the bias-variance trade-off and leverage a stable combination; $n = 3$ and $m = 0.25$ yields the optimal results. This is reinforced by Figure 4 where final test top-F_1 scores were attained by this combination. Consequently, increase variation was observed for n having skewed performance but the decrease deviation σ for m indicates greater stability over $m = 0.5$ and improved performance over $m = 0.1$, a similar result being observed for n.

Finally, selection of $c_{1;in}$ was more difficult due to the inter-correlation between other hyperparameters. Increase network parameters allows greater propagation of noise from harsher augmentation policies and so smaller networks are preferenced, although mid-range policies can allow for larger models to have stable convergence. Therefore, a bigger network with $c_{1;in} = 64$ can be matched with the chosen RandAugment configuration.

Conferring with Table 1 as a consistently high performing model across the listed evaluation metrics, the optimal hyperparameter configuration chosen for the U-Net DLA model uses $c_{1;in} = 64$, $\gamma = 0.0001$, $n = 3$,

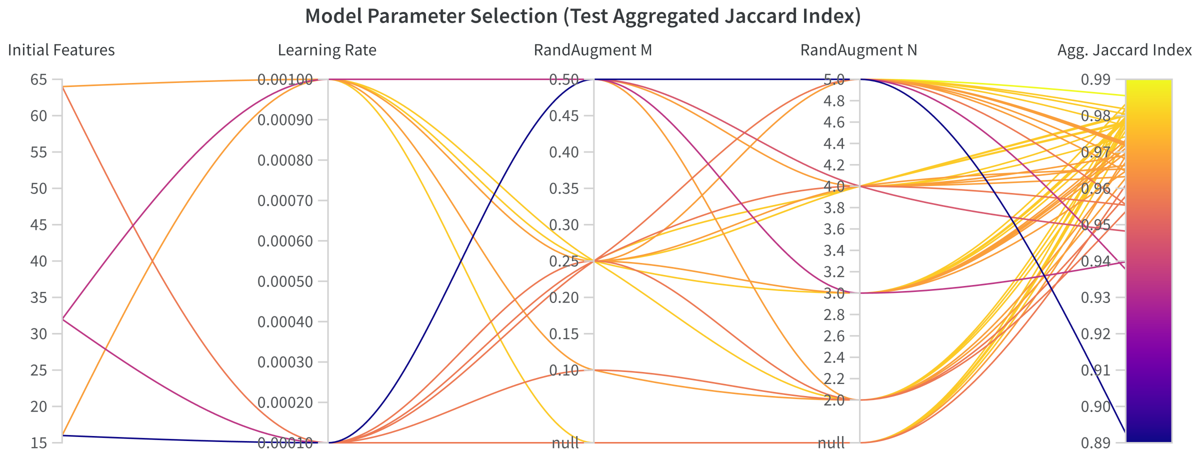


Figure 3: *U-Net DLA AJI parallel coordinates plot for hyperparameter ablation comparisons.*

and $m = 0.25$. Following a final testing loop for $E = 30$ epochs, the model attained $90.37/65.39\%$ accuracy, $97.71/98.77$ Dice, $84.37/97.96$ AJI, and $90.60/65.98$ F_1.

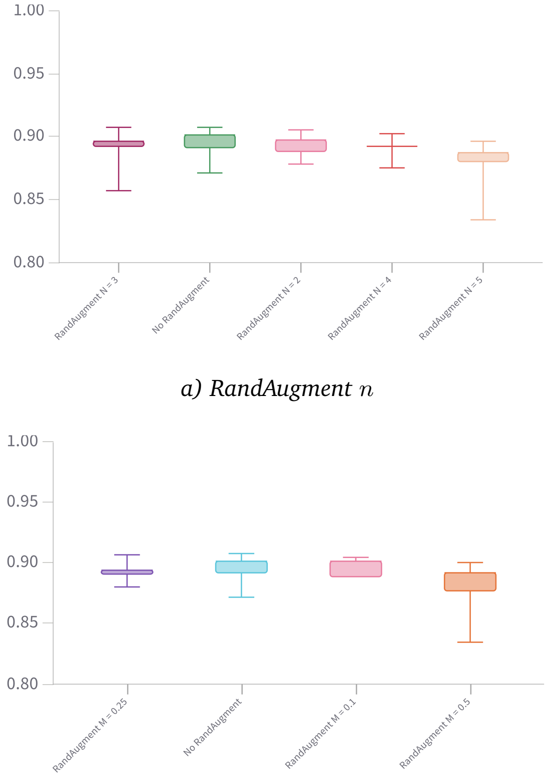
4 Discussion

Experimental results yielded key insights into important considerations targeted to designing a rapid iterable deep-learning architecture for effective instance segmentation learning tasks with high performance and address low-scale data availability.

Impressive results from the hyperparameter tuned U-Net DLA model already demonstrates comparative success to other contemporary research implementations for nuclei segmentation with very few ablations performed for model tuning. Although the selected model is quite large at approximately 31M training parameters, this can be quite restrictive across all research settings, but smaller $c_{1;in} = 32, 16$ networks did not demonstrate excessive loss in performance. This indicated that a low parameterised network with a simplified U-Net DLA design can be effective with minimal model iterations and be highly applicable for instance segmentation challenges.

Analysis of the modified RandAugment policy with specialised augmentation operations for segmentation image data was performed with a grid-search of 44 hyperparameter configurations. More intelligent search strategies can efficiently isolate the selected combination of n, m and reduce this search space further. But with the small number of trained instances, there was a clear increase in relative performance from benchmarked baselines despite the only having 3 target classes with contextually "simple" training instances. Translated to other more complex tasks, this result is expected to show more impressive outcomes.

Interestingly, the trade-off between network size and the augmentation policies was less expected which



a) RandAugment n

b) RandAugment m

Figure 4: *U-Net DLA final test F_1 performance across RandAugment hyperparameters n, m.*

Table 1: *Summary of model evaluation metrics for all tuned hyperparameters; metrics marked in bold are top 3- or 2-channel results for the baseline/modified RandAugment group. Metrics show 3-/2-class results (with/without the boundary class).*

	N-Ftrs	LR	RA-n	RA-m	Accuracy	Dice	AJI	F_1
Baseline	16	0.0001	-	-	**92.45/64.66**	96.96/98.37	82.36/96.89	89.14/65.61
	16	0.001	-	-	90.73/65.24	**97.73/98.78**	84.60/97.76	**90.78/65.91**
	32	0.0001	-	-	90.15/64.99	97.59/98.69	83.80/97.35	90.22/65.76
	32	0.001	-	-	87.08/65.66	96.93/98.34	**79.81/98.22**	87.12/66.07
	64	0.0001	-	-	91.50/65.22	97.58/98.70	84.22/97.72	90.51/65.89
	64	0.001	-	-	88.96/65.31	97.67/98.74	83.83/97.86	90.19/65.94
Mod-RA	16	0.0001	2	0.10	92.40/63.61	96.88/98.33	81.80/95.36	88.84/65.07
	16	0.001	2	0.10	90.77/64.55	97.50/98.65	83.65/96.72	90.14/65.54
	32	0.0001	2	0.10	90.05/65.35	**97.68/98.75**	84.12/97.89	90.43/65.95
	32	0.001	2	0.10	91.85/64.11	97.16/98.48	82.66/96.09	89.44/65.32
	64	0.001	2	0.25	**92.94/63.90**	97.01/98.39	82.52/95.78	89.35/65.22
	64	0.0001	2	0.25	90.67/65.13	97.66/98.74	84.31/97.58	**90.57/65.85**
	64	0.001	2	0.50	90.54/64.25	97.39/98.60	83.11/96.29	89.76/65.39
	64	0.0001	2	0.50	87.82/65.57	97.07/98.42	**80.64/98.19**	87.79/66.06
Mod-RA	16	0.0001	3	0.50	90.17/62.69	96.68/98.21	80.16/93.98	87.70/64.55
	16	0.001	3	0.25	92.55/64.36	96.99/98.39	82.45/96.47	89.24/65.46
	32	0.0001	3	0.25	91.52/64.80	97.41/98.61	83.53/97.10	90.03/65.68
	32	0.001	3	0.25	88.84/65.64	97.50/98.65	**82.78/98.32**	89.42/66.10
	64	0.0001	3	0.25	**91.73/64.95**	**97.62/98.72**	84.46/97.32	**90.70/65.76**
	64	0.001	3	0.25	89.74/65.40	97.55/98.68	83.43/97.96	89.92/65.98
	16	0.0001	3	0.50	91.08/64.72	97.14/98.46	82.46/96.97	89.22/65.64
	16	0.001	3	0.50	89.65/65.34	97.54/98.67	83.40/97.89	89.89/65.95
	32	0.0001	3	0.50	87.28/65.40	96.39/98.06	78.18/97.93	85.76/65.97
	32	0.001	3	0.50	89.94/65.18	97.43/98.62	83.05/97.66	89.64/65.88
	64	0.001	3	0.50	89.88/64.87	97.53/98.67	83.49/97.20	89.99/65.71
	64	0.0001	3	0.50	88.64/65.34	97.51/98.65	82.95/97.87	89.54/65.95
Mod-RA	16	0.001	4	0.25	**92.22/63.73**	96.70/98.23	81.28/95.54	88.42/65.13
	16	0.0001	4	0.25	89.58/64.42	97.30/98.54	82.45/96.53	89.25/65.48
	32	0.001	4	0.25	90.40/65.27	97.25/98.52	82.46/97.79	89.20/65.92
	32	0.0001	4	0.25	89.33/65.43	97.51/98.66	**83.04/98.03**	89.63/66.00
	64	0.001	4	0.25	90.79/65.29	**97.56**/98.69	83.84/97.83	**90.22/65.93**
	64	0.0001	4	0.25	89.70/65.42	97.49/98.65	83.18/97.99	89.73/65.99
	16	0.001	4	0.50	89.16/63.30	96.62/98.13	80.06/94.82	87.53/64.87
	16	0.0001	4	0.50	89.44/64.49	97.29/98.53	82.39/96.63	89.19/65.51
	32	0.001	4	0.50	89.63/65.29	97.32/98.56	82.46/97.79	89.20/65.92
	32	0.0001	4	0.50	90.42/65.02	97.30/98.54	82.70/97.42	89.41/65.79
	64	0.001	4	0.50	90.64/64.67	97.41/98.61	83.22/96.91	89.81/65.61
	64	0.0001	4	0.50	88.92/64.28	97.25/98.50	82.09/96.32	88.99/65.40
Mod-RA	16	0.001	5	0.25	**92.04/63.75**	96.55/98.15	80.80/95.56	88.04/65.14
	16	0.0001	5	0.25	88.57/65.55	97.32/98.56	81.91/98.18	88.77/66.05
	32	0.001	5	0.25	90.50/65.23	**97.38/98.59**	83.03/97.72	**89.63/65.90**
	32	0.0001	5	0.25	89.43/65.45	97.37/98.58	82.51/98.04	89.24/66.01
	64	0.001	5	0.25	91.13/64.87	97.09/98.44	82.23/97.21	89.05/65.72
	64	0.0001	5	0.25	87.70/64.62	97.28/98.53	81.84/96.81	88.73/65.57
	16	0.001	5	0.50	85.93/59.80	94.98/96.97	74.26/89.27	83.41/62.73
	16	0.0001	5	0.50	87.79/64.85	97.27/98.53	81.65/97.18	88.59/65.70
	32	0.001	5	0.50	89.61/62.56	96.51/98.12	79.45/93.77	87.15/64.47
	32	0.0001	5	0.50	85.11/65.82	97.12/98.44	**79.57/98.55**	86.86/66.18
	64	0.001	5	0.50	88.25/64.93	97.12/98.45	81.17/97.26	88.23/65.73
	64	0.0001	5	0.50	89.41/64.37	97.28/98.53	82.32/96.47	89.15/65.46

is likely a result of the simple dataset target masks which differs significantly from CIFAR-100 or ImageNet datasets that RandAugment was initially designed for. This is important to consider in future work when implemented augmentation policies for low-scale and contextually different datasets, further emphasising the importance of reduced search spaces for policy optimisation.

The difference between the 3-channel instance segmentation and binary semantic segmentation tasks is also quite interest considering the large discrepancies between difference evaluative metrics. This research revealed the importance of set-weighted metrics such as Dice and AJI for image segmentation evaluation over pixel-wise measures of accuracy and F_1 scores. These measures better capture generalised trends for classifying different regions and can be paired with other measures such as the coverage error to gain further insight into certain patterns of inference.

5 Conclusion

Histopathological nuclei instance segmentation is one of many biomedical research areas. Developing rapidly iterative and effective deep-learning models for image segmentation supports development of multi-task infrastructures and with additional augmentation strategies, mitigates issues of data accessibility across spare research challenges. Deep-Layer Aggregation (DLA) U-Net presents a low-scale model addressing said obstacles, demonstrating success in nuclei segmentation for the BBBC039 dataset. Leveraging a modified RandAugment policy strategy specifically for segmentation highlighted improved results with final binary masking metrics of 98.77 Dice and 97.96 AJI, $+2.337$ and $+7.76$ differences from second-best research results [2]. Future work should seek demonstrative multi-task success and improve modified RandAugment search strategies to minimise training iterations and increase the rapid deployment of research in new areas of biomedicine.

References

[1] V. Ljosa, K. L. Sokolnicki, and A. E. Carpenter, "Annotated high-throughput microscopy image sets for validation," *Nature Methods*, vol. 9, no. 7, pp. 637–637, 2012. DOI: 10.1038/nmeth.2083. [Online]. Available: https://doi.org/10.1038/nmeth.2083.

[2] W. Pan, Z. Liu, W. Song, X. Zhen, K. Yuan, F. Xu, and G. N. Lin, "An integrative segmentation framework for cell nucleus of fluorescence microscopy," *Genes*, vol. 13, no. 3, 2022, ISSN: 2073-4425. DOI: 10.3390/genes13030431. [Online]. Available: https://www.mdpi.com/2073-4425/13/3/431.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. arXiv: 1505.04597. [Online]. Available: http://arxiv.org/abs/1505.04597.

[4] F. Yu, D. Wang, and T. Darrell, "Deep layer aggregation," *CoRR*, vol. abs/1707.06484, 2017. arXiv: 1707.06484. [Online]. Available: http://arxiv.org/abs/1707.06484.

[5] Y. Kong, G. Z. Genchev, X. Wang, H. Zhao, and H. Lu, "Nuclear segmentation in histopathological images using two-stage stacked u-nets with attention mechanism," *Frontiers in Bioengineering and Biotechnology*, vol. 8, 2020, ISSN: 2296-4185. DOI: 10.3389/fbioe.2020.573866. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fbioe.2020.573866.

[6] D. Reis, J. Kupec, J. Hong, and A. Daoudi, *Real-time flying object detection with yolov8*, 2023. arXiv: 2305.09972 [cs.CV].

[7] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, *Segment anything*, 2023. arXiv: 2304.02643 [cs.CV].

[8] X. Zhao, W. Ding, Y. An, Y. Du, T. Yu, M. Li, M. Tang, and J. Wang, *Fast segment anything*, 2023. arXiv: 2306.12156 [cs.CV].

[9] A. Basu, P. Senapati, M. Deb, R. Rai, and K. G. Dhal, "A survey on recent trends in deep learning for nucleus segmentation from histopathology images," *Evolving Systems*, 2023. DOI: 10.1007/s12530-023-09491-3. [Online]. Available: https://doi.org/10.1007/s12530-023-09491-3.

[10] E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," *CoRR*, vol. abs/1805.09501, 2018. arXiv: 1805.09501. [Online]. Available: http://arxiv.org/abs/1805.09501.

[11] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical data augmentation with no separate search," *CoRR*, vol. abs/1909.13719, 2019. arXiv: 1909.13719. [Online]. Available: http://arxiv.org/abs/1909.13719.

[12] M. Buda, A. Saha, and M. A. Mazurowski, "Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm," *Computers in Biology and Medicine*, vol. 109, pp. 218–225, 2019. DOI: 10.1016/j.compbiomed.2019.05.002. [Online]. Available: https://doi.org/10.1016%2Fj.compbiomed.2019.05.002.