

FOCAL ATTENTION NETWORKS: OPTIMISING ATTENTION FOR BIOMEDICAL IMAGE SEGMENTATION

Michael Yeung^{1,2,3}, Leonardo Rundo^{2,4,5}, Evis Sala^{2,4}, Carola-Bibiane Schönlieb⁶, Guang Yang³

¹ School of Clinical Medicine, University of Cambridge, Cambridge, UK

² Department of Radiology, University of Cambridge, Cambridge, UK

³ National Heart & Lung Institute, Imperial College London, London, UK

⁴ Cancer Research UK Cambridge Centre, University of Cambridge, Cambridge, UK

⁵ Dept. of Information Eng., Electrical Eng. and Applied Mathematics, University of Salerno, Fisciano, Italy

⁶ Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK

ABSTRACT

In recent years, there has been a rising interest to incorporate attention into deep learning architectures for biomedical image segmentation. The modular design of attention mechanisms enable flexible integration into convolutional neural network architectures such as the U-Net. Whether attention is appropriate to use, what type of attention to use, and where in the network to incorporate attention modules, are all important considerations that are currently overlooked. In this paper, we investigate the role of the Focal parameter in modulating attention, revealing a link between attention in loss functions and networks. By incorporating a *Focal distance penalty term*, we extend the Unified Focal loss framework to include boundary-based losses. Furthermore, we develop a simple and interpretable, dataset and model-specific heuristic to integrate the Focal parameter into the Squeeze-and-Excitation block and Attention Gate, achieving the best results with fewer number of attention modules on three well-validated biomedical imaging datasets, suggesting judicious use of attention modules results in better performance and efficiency. The source code is available at: <https://github.com/mlyg/focal-attention-networks>.

Keywords—Biomedical Imaging, Image Segmentation, Machine Learning, Cost Function

I. INTRODUCTION

Attention provides neural networks with the capacity to selectively process salient inputs. In the context of image segmentation, attention mechanisms may be broadly divided into two complementary types: spatial attention, which enhances processing of salient locations within the image, and channel attention, which calibrates feature maps based on relative importance [1]–[3]. Both forms of attention may be encapsulated into modules, which enable flexible integration into existing convolutional neural network (CNN) architectures. The most widely used CNN for biomedical image segmentation is the U-Net, consisting of a symmetrical encoder-decoder structure with skip connections [4]. Attention-based

variants of the U-Net include the Attention U-Net, which incorporates spatial attention using the Attention Gate (AG), and the USE-Net, which uses the channel-based Squeeze-and-Excitation (SE) block [1], [2], [5].

Despite the widespread use of attention in CNNs for image segmentation, it remains unclear when attention is appropriate to use, which type of attention to use, and where the optimal location is for use within the network. Currently, performance benchmarking is the only method available for evaluating the usefulness of attention. However, the contribution of individual attention modules cannot be inferred, and even with ablation studies, only minor a subset of all positional combinations can be reasonably evaluated. Without understanding how individual attention modules affect performance, it is not possible to determine where to optimally place attention modules.

The main contributions of this work may be summarised as follows:

- 1) We leverage the Focal parameter to modulate attention across both loss function and network contexts, incorporating a Focal Distance Penalty Term to further generalise the Unified Focal loss framework, and integrating a Focal layer into network attention modules.
- 2) We demonstrate consistently improved performance using the Unified Focal loss and Focal Attention networks across three, well-validated open-source biomedical imaging datasets.
- 3) We develop a simple and interpretable, model and dataset-specific heuristic for deciding the optimal type, location and strength of attention.

II. MATERIALS AND METHODS

II-A. Focal attention in loss functions

The Focal loss was designed as a variant of the cross-entropy loss to address class imbalanced datasets for classification, by selectively downweighting well-classified examples [6]. Image segmentation often involves additional

class imbalance at the pixel level, and in biomedical image segmentation, class imbalance is frequently observed with objects such as cell nuclei or tumours which occupy a small area relative to the image. In our previous work, we generalised distribution-based and region-based loss functions into a single framework known as the Unified Focal loss, to handle class imbalanced datasets [7]. For a given softmax output for classes c , where r and t refer to the rare class and ground truth respectively, the Unified Focal loss (\mathcal{L}_{UF}) is defined as the weighted sum of the Asymmetric Focal loss (\mathcal{L}_{AF}) and Asymmetric Focal Tversky loss (\mathcal{L}_{AFT}):

$$\mathcal{L}_{\text{UF}} = \lambda \mathcal{L}_{\text{AF}} + (1 - \lambda) \mathcal{L}_{\text{AFT}}, \quad (1)$$

where:

$$\mathcal{L}_{\text{AF}} = -\frac{\delta}{N} y_{i:r} \log(p_{t,r}) - \frac{1-\delta}{N} \sum_{c \neq r} (1 - p_{t,c})^\gamma \log(p_{t,r}), \quad (2)$$

$$\mathcal{L}_{\text{AFT}} = \sum_{c \neq r} (1 - \text{TI}) + \sum_{c=r} (1 - \text{TI})^{1-\gamma}. \quad (3)$$

The Tversky Index (TI) is defined as:

$$\text{TI} = \frac{\sum_{i=1}^N p_{0i} g_{0i}}{\sum_{i=1}^N p_{0i} g_{0i} + \delta \sum_{i=1}^N p_{0i} g_{1i} + (1 - \delta) \sum_{i=1}^N p_{1i} g_{0i}}, \quad (4)$$

where p_{0i} is the probability of pixel i belonging to the foreground class and p_{1i} is the probability of pixel belonging to background class. g_{0i} is 1 for foreground i.e. segmentation target and 0 for background and conversely g_{1i} takes values of 1 for background and 0 for foreground.

The three hyperparameters in the Unified Focal loss are λ , which controls the relative weights of the two component losses, δ , which controls the relative weighting of positive and negative examples, and γ , which controls the relative weighting of easy and difficult examples.

Another class of loss functions are boundary-based loss functions, which compute Distance Transform Maps (DTM) based on Euclidean distances to penalise predictions relative to class boundaries [8]. The Distance Penalty Term (DPT) is defined as the inverse of DTM, penalising incorrect predictions close to boundaries [9]. We further generalise the DPT, defining the Focal Distance Penalty Term (FDPT) by applying a Focal parameter, ϵ :

$$\mathcal{W}_c^{\text{FDPT}} = (\mathcal{W}_c^{\text{DPT}})^\epsilon. \quad (5)$$

The FDPT establishes a continuity between no boundary awareness ($\epsilon = 0$), to varying degrees of boundary awareness, revealing that the DPT is a particular solution where $\epsilon = 1$ (Fig. 1).

We integrate the FDPT into the Unified Focal loss framework by replacing the ground truth with its respective FDPT,

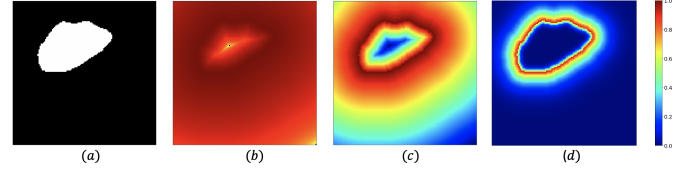


Fig. 1. Focal Distance Penalty Term distance maps visualised as heatmaps where (a) label image, (b) $\epsilon = 0.1$, (c) $\epsilon = 1$ (equivalent to DPT) and (d) $\epsilon = 10$

and together with the other Focal parameters, derive a loss function where optimisation involves selective attention to difficult to segment regions and boundaries (Fig. 2).

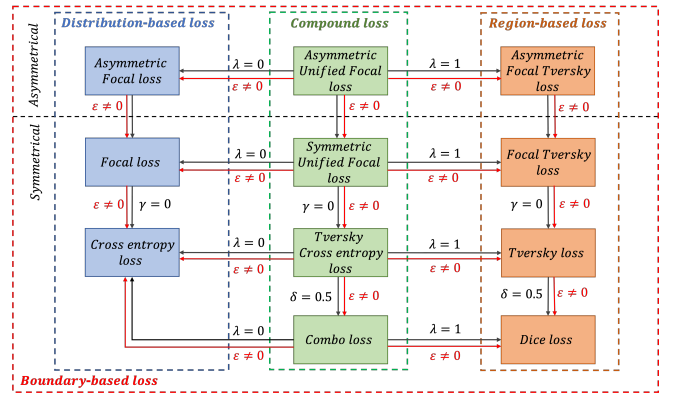


Fig. 2. The extended Unified Focal loss framework generalises numerous distribution-based, region-based and boundary-based loss functions. By fixing hyperparameter values, particular loss functions may be derived as specified by the arrows

II-B. Focal attention in networks

Analogous to its effect on loss functions, the Focal parameter generalises network attention mechanisms by modulating attention strength [3]. As examples, we select the SE block for channel attention and the AG for spatial attention, which has been integrated into the U-Net architecture in the USE-Net and Attention U-Net respectively (Fig. 3) [2], [5]. Briefly, SE blocks achieves channel attention by performing initial feature aggregation using global average pooling along the spatial axis ('squeeze'), followed by two fully connected (FC) layers with ReLU and sigmoid activations producing the 'excitation' operation. In contrast, the AG uses contextual information from the gating signal g_x to prune skip connections x_i . The Focal layer is inserted after the generation of attention weights and prior to recalibration, and involves a single, trainable parameter, that modulates attention strength. By initialising the Focal parameter to 1, the Focal attention module initially behaves identical to its respective attention module, but during training, the

attention strength is optimised through parameter updating using backpropagation.

We investigate a further use of the Focal layer to determine usefulness of individual attention modules. By initialising the Focal parameter to 0 and monitoring the Focal layer weights during training, we expect attention modules that are either not useful or harmful to performance to remain close to 0. Attention modules where the Focal parameter remains close to 0 at convergence are removed to complete the pruning procedure. Given the stochastic nature of initialisation and training, and the dependence on other factors such as learning rate and schedule, there are no set cut-off values to distinguish likely useful from neutral or harmful attention modules. In these experiments, for consistency we remove attention modules whose Focal parameter < 0.2 at convergence.

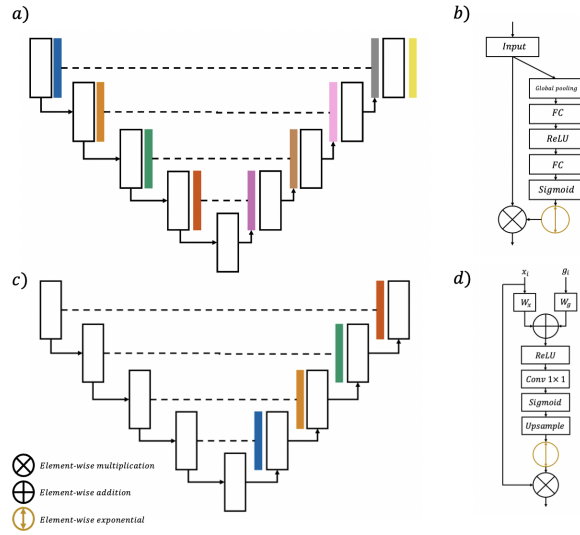


Fig. 3. Simplified diagrams of the (a) USE-Net, (b) SE block, (c) Attention U-Net and (d) AG. The coloured blocks indicate the position of attention modules

II-C. Dataset descriptions and evaluation metrics

To evaluate the extended Unified Focal loss and Focal Attention networks, we select three well-validated, open-source biomedical imaging datasets: Digital Retinal Images for Vessel Extraction (DRIVE), 2018 Data Science Bowl (2018DSB) and CVC-ClinicDB [10]–[12]. Briefly, the DRIVE dataset consists of 40 coloured fundus photographs for retinal vessel segmentation, 2018DSB comprises 670 light microscopy images for nuclei segmentation, and the CVC-ClinicDB dataset consists of 612 frames containing polyps obtained during optical colonoscopy. A summary of the datasets and training details are presented in Table I.

For evaluation, we calculate Dice Similarity Coefficient (DSC), precision and recall metrics per image and average over the independent test set.

Table I. Details of datasets and training setup used for our experiments

Dataset	Segmentation	#Images	Size	#Training	#Validation	#Test	%Foreground	UFL γ
DRIVE	Retinal vessel	40	$512 \times 512 \times 3$	16	4	20	8.70	0.1
2018DSB	Cell nuclei	670	$256 \times 256 \times 3$	428	108	134	14.5	0.2
CVC-ClinicDB	Colorectal polyp	612	$288 \times 384 \times 3$	392	98	122	9.30	0.3

II-D. Implementation details

For our experiments, we used the Medical Image Segmentation with Convolutional Neural Networks (MIScnn) open-source Python library [13]. Our implementations made use of Keras with Tensorflow backend and were trained using NVIDIA P100 GPUs. For all experiments, except for the DRIVE dataset that is already partitioned into 20 training images and 20 testing images, we randomly partitioned each dataset into 80% development and 20% test set, and further divided the development set into 80% training set and 20% validation set. All images were normalised to $[0, 1]$ using the z-score, and we used the following data augmentation: scaling, rotation, mirroring, elastic deformation and brightness.

Model parameters were initialised using Xavier initialisation. We trained each model using the Adam optimizer with an initial learning rate of 1×10^{-3} and used ReduceLROnPlateau to reduce the learning rate by 0.1 if the validation loss did not improve after 25 epochs, and the EarlyStopping callback to terminate training if the validation loss did not improve after 50 epochs.

We used the asymmetric variant of the Unified Focal loss with $\delta = 0.6$ and $\lambda = 0.5$, and empirically determined the optimal γ value for each dataset. The FDPT was empirically set with $\epsilon = 0.1$. For the SE block, we set the reduction ratio $r = 8$ [1], [5]. For all experiments, we used instance normalisation and trained with a batch size of 1 [4], [14].

III. RESULTS

III-A. Loss function experiments

We compare the performance of the U-Net when optimised with the DSC loss, DSC + cross entropy (CE) loss, and UFL, with the FDPT set either to $\epsilon = 0$ (equivalent to no penalty), $\epsilon = 1$ (equivalent to DPT) and $\epsilon = 0.1$. The results for the loss function comparisons on the three datasets are shown in Table II.

The UFL + FDPT is associated with the highest DSC score across all three datasets, with a DSC of 0.8155, 0.9165 and 0.8993 for the DRIVE, 2018DSB and CVC-ClinicDB respectively. The UFL achieves consistently higher DSC scores compared to the DSC and DSC + CE losses. Worse performance was often observed with the DPT compared to no penalty, suggesting the DPT may focus too strictly on boundaries, while better performance is observed when the attention is relaxed by setting a lower ϵ value.

III-B. Focal Attention network experiments

The results using Focal SE blocks are shown in Table III. Integrating Focal variants of the SE block achieved the

Table II. Performance comparisons using DSC loss, DSC + CE loss and UFL loss, with and without boundary attention, on DRIVE, 2018DSB and CVC-ClinicDB. The highest scores are denoted in bold

Dataset	Metrics	DSC	DSC + DPT	DSC + FDPT	DSC + CE	DSC + CE + DPT	DSC + CE + FDPT	UFL	UFL + DPT	UFL + FDPT
DRIVE	DSC	0.8082	0.8086	0.8105	0.8093	0.8070	0.8116	0.8142	0.8142	0.8155
	Precision	0.8473	0.8498	0.8440	0.8480	0.8596	0.8468	0.8199	0.8298	0.8075
	Recall	0.7766	0.7751	0.7836	0.7776	0.7640	0.7829	0.8127	0.8031	0.8276
2018DSB	DSC	0.9147	0.9016	0.9150	0.9148	0.9085	0.9159	0.9157	0.9129	0.9165
	Precision	0.9205	0.9230	0.9191	0.9140	0.9236	0.9163	0.9061	0.9008	0.9196
	Recall	0.9168	0.8891	0.9184	0.9233	0.9012	0.9231	0.9324	0.9327	0.9204
CVC-ClinicDB	DSC	0.8826	0.8174	0.8833	0.8917	0.8536	0.8993	0.8937	0.8622	0.8993
	Precision	0.9175	0.9058	0.9117	0.9166	0.9135	0.9155	0.8965	0.8913	0.9074
	Recall	0.8759	0.7614	0.8816	0.8874	0.8171	0.9024	0.9096	0.8563	0.9092

highest DSC score. Interestingly, the highest DSC score for the DRIVE and 2018DSB dataset was observed with the pruned Focal USE-Net, suggesting certain attention modules that were removed may have worsened performance.

Table III. Performance comparisons using the U-Net, USE-Net and its Focal variant, before and after attention module selection. The highest scores are denoted in bold

		U-Net	USE-Net	Focal USE-Net	Focal USE-Net
DRIVE	#SE	0	9	9	2
	DSC	0.8142	0.8145	0.8152	0.8159
	Precision	0.8199	0.8252	0.8056	0.8114
2018DSB	Recall	0.8127	0.8079	0.8289	0.8246
	#SE	0	9	9	4
	DSC	0.9157	0.9131	0.9159	0.9179
CVC-ClinicDB	Precision	0.9061	0.9001	0.9024	0.9085
	Recall	0.9324	0.9354	0.9374	0.9343
	#SE	0	9	9	5
CVC-ClinicDB	DSC	0.8937	0.8952	0.9005	0.8952
	Precision	0.8965	0.9084	0.9097	0.9111
	Recall	0.9096	0.9023	0.9032	0.9059

The results using Focal AG are shown in Table IV. The performance improvements were less consistent than with the USE-Net. For the DRIVE dataset, the best performance was observed with the U-Net, although better performance was observed with the Focal AG in the 2018DSB and CVC-ClinicDB dataset. This may be unsurprising, given that the retinal vessels in the DRIVE dataset extend to cover the entire field of view, making spatial attention effectively redundant for this task.

Table IV. Performance comparisons using the U-Net, Attention U-Net and its Focal variant, before and after attention module selection. The highest scores are denoted in bold

		U-Net	Attention U-Net	Focal Attention U-Net	Focal Attention U-Net
DRIVE	#AG	0	4	4	0
	DSC	0.8142	0.8138	0.8138	-
	Precision	0.8199	0.8049	0.8005	-
2018DSB	Recall	0.8127	0.8274	0.8325	-
	#AG	0	4	4	1
	DSC	0.9157	0.9127	0.9145	0.9158
CVC-ClinicDB	Precision	0.9061	0.8997	0.9018	0.9057
	Recall	0.9324	0.9345	0.9365	0.9341
	#AG	0	4	4	2
CVC-ClinicDB	DSC	0.8937	0.9051	0.9063	0.9118
	Precision	0.8965	0.9185	0.9134	0.9195
	Recall	0.9096	0.9078	0.9126	0.9220

The Focal layer weights monitored during training are shown in Fig. 4. The SE blocks in the decoder position converged towards consistently higher attention weights than in the encoder position, suggesting greater attention strength may be beneficial downstream compared to earlier in the network. The weights for both the SE and AG modules varied across datasets, with the least variation observed in the

DRIVE dataset and the most in the CVC-ClinicDB dataset. This matches the variation observed within each dataset, with retinal vessels displaying considerably homogeneity in the DRIVE dataset, in comparison to colorectal polyps in the CVC-ClinicDB dataset which vary considerably in location, shape, size, colour and texture.

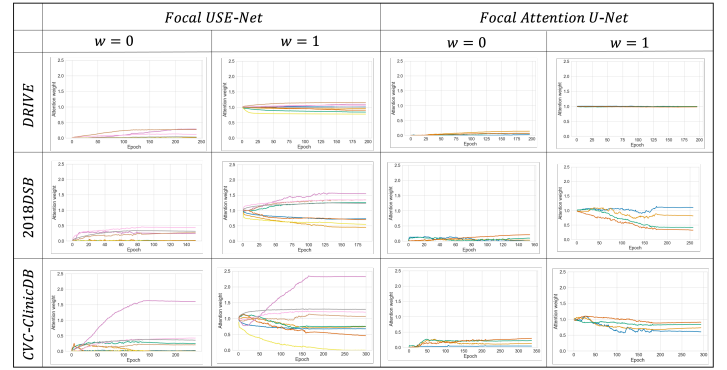


Fig. 4. Focal layer weights, w , monitored during training, when initialised to 0 and 1, for the three datasets. The colour encoding corresponds to individual attention modules in the positions illustrated in Figure 3

IV. CONCLUSION

In this work, we highlight the role of the Focal parameter in modulating attention for both loss function and network contexts. We derive an FDPT, and incorporate this into the UFL framework to generalise boundary-based losses. We demonstrate improved performance using the extended UFL over the DSC and DSC + CE losses. In the network context, we incorporate a Focal layer into the SE blocks and AG modules to optimise attention strength, and observe better performance using Focal variants of the USE-Net and Attention U-Net. Finally, we develop a simple and interpretable heuristic, by monitoring zero initialised Focal layer weights, to query the usefulness of a given attention module at a particular position. Interestingly, we often observed better performance after removing certain attention modules, suggesting that judicious use of attention modules is necessary, requiring consideration of the dataset, the model, the attention type and position, to optimise performance and efficiency.

V. ACKNOWLEDGMENT

This work was partially supported by The Mark Foundation for Cancer Research and Cancer Research UK Cambridge Centre [C9685/A25177], the CRUK National Cancer Imaging Translational Accelerator (NCITA) [C42780/A27066] and the Wellcome Trust Innovator Award, UK [215733/Z/19/Z]. Additional support was also provided by the National Institute of Health Research (NIHR) Cambridge Biomedical Research Centre [BRC-1215-20014] and the Cambridge Mathematics of Information in Healthcare (CMIH) [funded by the EPSRC grant EP/T017961/1], as well as in part by the UK Research and Innovation Future Leaders Fellowship [MR/V023799/1], in part by the Medical Research Council [MC/PC/21013], in part by the European Research Council Innovative Medicines Initiative [DRAGON, H2020-JTI-IMI2 101005122], and in part by the AI for Health Imaging Award [CHAI MELEON, H2020-SC1-FA-DTS2019-1 952172]. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care.

VI. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using anonymised human subject data made available in open access. Ethical approval was not required as confirmed by the license attached with the open access data.

VII. REFERENCES

- [1] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis.*, 2018, pp. 7132–7141.
- [2] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197–207, 2019.
- [3] M. Yeung, E. Sala, C.-B. Schönlieb, and L. Rundo, "Focus U-Net: A novel dual attention-gated CNN for polyp segmentation during colonoscopy," *Comput. Biol. Med.*, vol. 137, p. 104815, 2021.
- [4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015, pp. 234–241.
- [5] L. Rundo, C. Han, Y. Nagano, J. Zhang, R. Hataya, C. Militello, A. Tangherloni, M. S. Nobile, C. Ferretti, D. Besozzi *et al.*, "USE-Net: Incorporating squeeze-and-excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets," *Neur. Comp.*, vol. 365, pp. 31–43, 2019.
- [6] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. International Conference on Computer Vision (ICCV)*. IEEE, Oct 2017, pp. 2999–3007.
- [7] M. Yeung, E. Sala, C.-B. Schönlieb, and L. Rundo, "Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation," *Comput. Med. Imaging Graph.*, vol. 95, p. 102026, 2022.
- [8] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. B. Ayed, "Boundary loss for highly unbalanced segmentation," in *Proc. International Conference on Medical Imaging with Deep Learning (MIDL)*. PMLR, 2019, pp. 285–296.
- [9] T. Sugino, T. Kawase, S. Onogi, T. Kin, N. Saito, and Y. Nakajima, "Loss weightings for improving imbalanced brain structure segmentation using fully convolutional networks," in *Healthcare*, vol. 9, no. 8, 2021, p. 938.
- [10] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. Van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Trans Med Imaging*, vol. 23, no. 4, pp. 501–509, 2004.
- [11] J. C. Caicedo, A. Goodman, K. W. Karhohs, B. A. Cimini, J. Ackerman, M. Haghighi, C. Heng, T. Becker, M. Doan, C. McQuin *et al.*, "Nucleus segmentation across imaging experiments: the 2018 data science bowl," *Nat. Methods*, vol. 16, no. 12, pp. 1247–1253, 2019.
- [12] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilarino, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Comput. Med. Imaging Graph.*, vol. 43, pp. 99–111, 2015.
- [13] D. Müller and F. Kramer, "MIScnn: a framework for medical image segmentation with convolutional neural networks and deep learning," *BMC Med. Imaging*, vol. 21, no. 1, pp. 1–11, 2021.
- [14] X.-Y. Zhou and G.-Z. Yang, "Normalization in training U-Net for 2-D biomedical semantic segmentation," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1792–1799, 2019.