

Automated Retinal Segmentation of Hyperspectral Images Using a Modified U-Net with Spectral Multi-head Attention

Arrsh Ali ^{a,b}, Minh Tran ^{a,b}, Isioma Emordi ^{a,b}, Michelle Bryarly ^{a,b},
Muhammad Saad Yousuf ^c, Brodie Jean Woodall ^c, Baowei Fei ^{a,b,d,*}

^a Center for Imaging and Surgical Innovation, University of Texas at Dallas, Richardson, TX 75080

^b Department of Bioengineering, University of Texas at Dallas, Richardson, TX 75080

^c Department of Neuroscience, University of Texas at Dallas, Richardson, TX 75080

^d Department of Radiology, University of Texas Southwestern Medical Center, Dallas, TX 75235

* Corresponding author: bfei@utdallas.edu, Website: <https://fei-lab.org>

ABSTRACT

Hyperspectral imaging (HSI) can capture spatial and spectral data of the retina. Retinal vessel segmentation enables quantitative analysis and assessment of retinal health. In this work, we developed an HSI system for mouse retina and a deep learning architecture to segment retinal vessels. Retinal images were manually segmented to establish ground truth. We proposed the use of multi-head attention block for the spectral data, as well as Tversky loss for segmentation refinement. The model's performance was assessed using Dice similarity coefficient (DSC) and the Jaccard (IoU) metrics. A comparative analysis was conducted against the manual segmentation, and five other segmentation methods: UNet, UNet++, UNet3+, segment anything model (SAM), and SAM2. We demonstrated that our spectral network achieved an IoU and DSC scores of 0.80 ± 0.04 and 0.89 ± 0.02 , respectively. Our network outperforms other networks. The automatic segmentation method provides a tool for retinal analysis and quantification.

Keywords: Hyperspectral imaging (HSI), retina, fundus imaging, deep learning, vessel segmentation, U-Net, spectral multi-head attention

1. Introduction

The retinal fundus is the only region in the human body that allows for non-invasive visualization of the central nervous system (CNS), offering crucial insight into overall health [1]. Beyond ocular diseases, the retina provides a biomarker-rich gateway for neurodegenerative disease, such as Alzheimer's disease (AD) and cardiovascular dysfunction [1]. Thus, the need for the analysis of retinal structure can have many clinical applications. Fundus imaging captures the complex structure of retinal vasculature including the retinal blood vessels, optic disk (OD), macula, fovea, and retinal abnormalities. Such imaging when paired with segmentation algorithms are key for automatic retinal disease screening [2]. Numerous algorithms have been implemented for automating retinal vessel segmentation, from classical methods such as Otsu algorithm [3] and morphological thresholding [4] to advanced deep learning-based methods such as U-Net [5], U-Net++ [6], UNet3+ [7], and segment anything model (SAM) [8]. However, most of these models have been trained on segmenting vessels of human retina and rely on fundus RGB imaging modality. This presents a notable gap as mouse models, which closely resemble human retina vasculature and disease progression, remain prevalent in pre-clinical research. Deep convolutional neural networks (CNNs) were used for mouse retinal segmentation of multiphoton images [9]. However, current retinal segmentation methods are limited to RGB images, which primarily provides spatial data.

Hyperspectral imaging (HSI), however, provides both spatial and spectral data and can be powerful for retinal analysis [10]. To the best of our knowledge, no existing methods address the automated segmentation of HSI retinal fundus images. This study has the following contributions: (1) The development of a novel U-Net model tailored to HSI retinal fundus images. (2) The application of our U-Net model for the novel application of mouse retina vessel segmentation. This study offers a promising and scalable solution for accurate retinal vessel segmentation of HSI images across species, providing an advanced tool for preclinical and clinical translation while opening new avenues for diagnostics and research applications.

2. Materials and Methods

Hyperspectral Imaging System

The system used to acquire images of the retina has been described in a previous proceeding [11]. Briefly, the system consists of two cameras: An RGB camera and a hyperspectral snapshot camera. The RGB camera (ToupTek Photonics, China) captures high-resolution RGB images (2048×3072 pixels). The snapshot camera captures hyperspectral images ($270 \times 512 \times 16$ pixels) with 16 channels in the range 460 – 600 nm. The two cameras share the same field of view, achieved by connecting a beam splitter toward both cameras. From previous optical tests, the system achieved an average spectral root-mean squared error (RMSE) of 3.87 ± 1.89 % and an optical magnification factor of $2.15 \mu\text{m}/\text{pixel}$.

Image Acquisition

We acquired high-resolution HSI of the retina in C57BL/6 black mice and wild type mice (Jackson National Labs, USA). Our methods was based on the procedure described by More *et al.* [12]. Fifteen minutes prior to the anesthesia, we applied 0.5% phenylephrine and 1% tropicamide per eye for pupil dilation. Mice were anesthetized by inhalation of 5% isoflurane and medical air mixture until immobile, then remained sedated by inhalation of 1.5% isoflurane. Mice were then placed in a customized imaging platform, that allowed restraints and rotations of the animal while keeping the mouse sedated. We applied local anesthetic (Proparacaine, Fisher Scientific, USA), followed by a gel eye drop (Gentel Tears, Alcon, USA) to the cornea. To keep eyes hydrated and prevent cataracts, a plano-concave lens (LC2969, Thorlabs, USA) was then placed on each eye during image acquisition. We placed mice on top of the vertical translation platform, then rotated the imaging platform toward the surface of the cornea. Then, we translated the entire imaging system forward until the scope contacted the cornea. We then altered the focuses on the FFL lens until the retina was in focus. The exposure time for the RGB camera and HSI camera was 100ms and 200ms respectively. The entire imaging process lasted less than 15 minutes per mouse, which allowed the anesthetized mouse to return to their cage and recover.

Image Preprocessing

Following image acquisition, the 3D hyperspectral hypercube images underwent quality filtration and preprocessing to ensure high quality data. The images were evaluated based on image clarity, glare presence, and vessel visibility. Any images deemed low quality and artifact-heavy were excluded from the dataset, resulting in 53 remaining images initially, which later expanded to 273 images after a second round of retinal imaging. These images then underwent a preprocessing pipeline including brightness and contrast adjustments to enhance vessel visibility. Manual segmentation of each image was performed in Photoshop, during which the background of the image was masked out and the retinal vasculature was outlined in white (Figure 1). Segmentation was independently conducted by two trained annotators, each blinded to each other's annotations to eliminate bias.

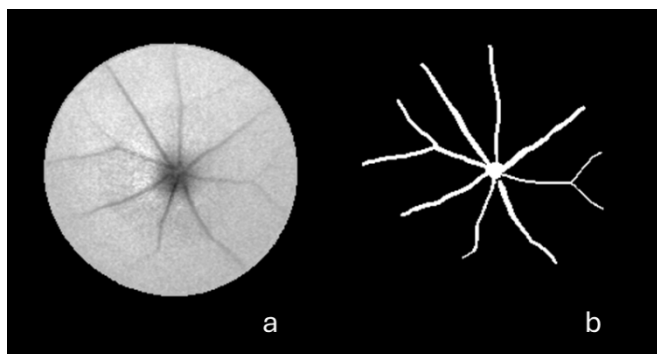


Fig. 1 Example segmentation maps. (a) Processed hyperspectral image of mouse retina, with the blood vessels already enhanced. (b) Manual segmentation map.

Image Segmentation Network

In this paper, we improved upon the UNet3+CBAM network proposed by Xu *et al.* [13]. The backbone of the hyperspectral segmentation network is a UNet3+ architecture [14], with an added convolution block attention module (CBAM) layer after each encoder layer. In this paper, we proposed the following modifications for our hyperspectral retina vessel segmentation task: (1) We kept the original UNet3+ architecture to leverage the full skip connection ability; (2) we used a modified CBAM using multi-head spectral transformer modules to learn useful wavelengths, which we subsequently named as the self-attention block module (SABM); (3) we moved SABM toward the encoder to better extract the useful wavelengths; (4) we used Tversky loss, which was demonstrated by Salehi *et al.* [15] to show improvements over Dice score in class imbalance. Figure 2 shows the UNet3+Transformer architecture that was used in our network. Figure 3(a) shows the structure of SABM. Figure 3(b) shows the implementations of the spectral attention module, and Figure 3(c) shows the implementation of the spatial attention module.

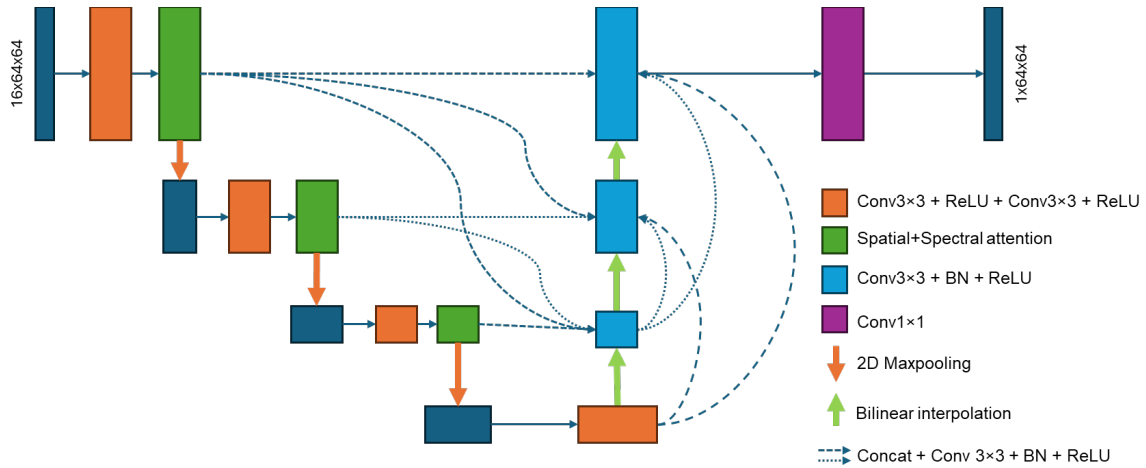


Fig 2. The structure of our UNet3+ network with modified convolution block attention module. This figure shows only a simplified version; the actual network consists of 4 layers of down-sampling.

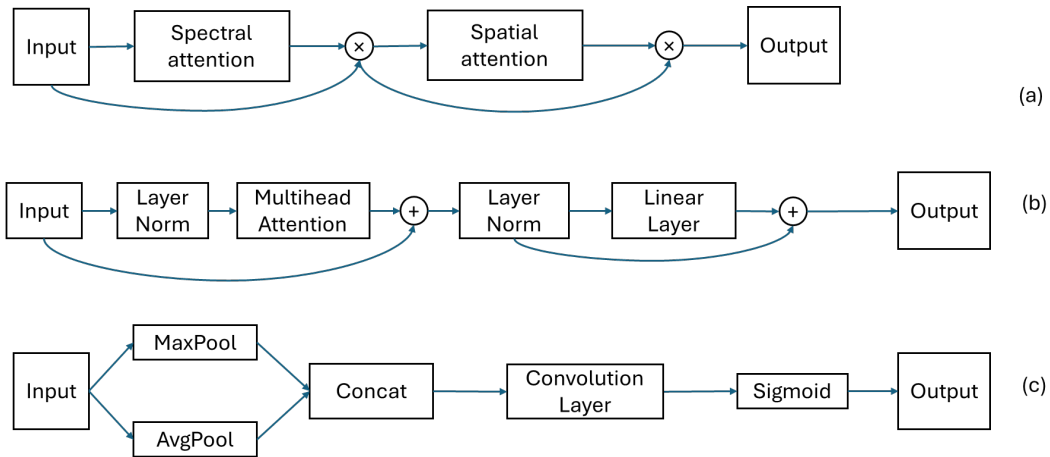


Fig. 3 (a) The structure of SABM. (b) The structure of self-attention spectral module. (c) The structure of spatial attention module.

In our proposed design, we moved SABM toward the output of the encoder, so that SABM can better refine the immediate output. We further replaced the spectral attention module in CBAM with a multi-head self-attention (MHSA) transformer block to better model spectral dependencies in hyperspectral images. Given an input feature map $X \in \mathbb{R}^{B \times C \times H \times W}$, spectral features are reshaped to $\mathbb{R}^{B \times HW \times C}$ and passed through learnable projections W_Q , W_K and W_V to produce the matrices for queries, keys, and values Q , K , V respectively:

$$Q = XW_Q, K = XW_K, V = XW_V$$

The attention weights are computed as $Attention(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$. In this set of equations, X denotes the input feature map. B is the batch size, C is the number of channels, H and W are the spatial height and width of the feature map, Q (query), K (key), and V (value) are the transformed representations of X obtained through learned weight matrices W_Q , W_K and W_V , respectively, d_k is the dimensionality of the key vectors, and the SoftMax operator is defined as $\text{Softmax}(z_i) = \frac{\exp(z_i)}{\sum_{j=1}^n \exp(z_j)}$ where z_i is the i -th element of the input vector z and n is the vector length.

We further used the Tversky loss, a generalization of the Dice loss that allows asymmetric weighting of false positive (FP) and false negative (FN). The Tversky loss is defined as:

$$\mathcal{L}_T = 1 - \frac{TP}{TP + \alpha FP + \beta FN}$$

Here, TP, FP, and FN denote the number of true positive, false positive, and false negative pixels, respectively. The parameters α , β control the penalty for FP and FN. In our case, we set $\alpha = 0.3$ and $\beta = 0.7$ to bias the losses toward higher recall, which is beneficial for detecting small vessel segments. The loss function is a combination of Tversky loss and binary cross entropy loss with equal 0.5 weighing on each. The network was trained using deep supervision, which meant that all decoder outputs were weighed equally when calculating the loss. During inference, only the output of the final decoder was considered.

Model Training

We trained the network using two separate rounds using different datasets. First, we trained the network using 53 images using five-fold cross-validation. In each fold, data from four mice were used for training and hyperparameter tuning, while data from the remaining mouse served as the test set. The test dataset was withheld entirely during training and tuning. This process was repeated five times, with each mouse serving as the test subject once. A second round of imaging led to acquisition of 220 images from 16 mice. Five-fold cross-validation was similarly used during training. Here, four of the folds consisted of 3 mice, and one fold consisted of 4 mice. Training was conducted on a high-performance GPU cluster equipped with four NVIDIA 64 GB GPUs. For augmentation of the training data, we used random rotation and flip of the hyperspectral images. The network was trained with the following hyperparameters: input size of 64×64 pixels with 16 spectral channels, batch size of 64, Adam optimizer with an initial learning rate of 1×10^{-4} and a 50% decay in learning rate every 5000 epochs, and 20,000 training epochs.

For comparison, we trained several deep learning segmentation models on the same hyperspectral retinal vessel dataset, using identical data augmentation strategies and loss functions. The tested models included a vanilla U-Net[16], a U-Net++[17], a UNet3+, a UNet3+ with regular CBAM architecture, a UNet3+ with SABM architecture, Segment Anything Model (SAM), and Segment Anything Model 2 (SAM2) [8]. All comparison models were trained using the same set of hyperparameters, with the exception of the loss function (Dice instead of Tversky + Dice) and the learning rate (1×10^{-3}).

Segmentation Evaluation

The metrics used to evaluate our network were Dice similarity coefficient (DSC) and intersection over union (IoU). DSC measures the degree of spatial overlap between the predicted segmentations of each model and the ground truth. It is defined as twice the area of overlap divided by the number of pixels in both the predicted and ground truth:

$$DSC = \frac{2 \times TP}{2 \times TP + FP + FN}$$

In this equation, TP, FP, and FN correspond to the number of pixels that are true positive (agreement between the predictions and ground truth), false positive (predicted segmentation but do not contain blood vessels), and false negative

(contain blood vessels but is not segmented), respectively. DSC has a range of [0,1], where a value closer to 1 indicates high agreement between the predicted and ground truth segmentations. IoU quantifies the overlap between the predicted region and the ground-truth region by calculating the ratio of the intersection to the union of the two masks. IoU penalizes false positives and negatives, making it a robust metric of segmentation quality. The formula for IoU is:

$$IoU = \frac{TP}{TP + FP + FN}$$

Similar to DSC, the range of IoU is 0 to 1, where one indicates a perfect overlap between predicted and ground truth regions.

3. Results

Table 1 shows the Jaccard score (IoU) of the models on the test dataset. Table 2 shows the Dice score (F1 score) of the models on the test dataset. Our network reported an IoU score of 0.80±0.04 and a Dice score of 0.89±0.02. As shown in both tables, our model reported the highest average IoU and DSC scores, indicating that it was the better performing model. Aside from our model, the UNet3+SABM appear to have the most promising segmentation capabilities. The reported Dice score is larger compared to the IoU for our model. This is expected, because our loss functions (Tversky and Dice) were optimized for Dice score. We found that the vanilla UNet3+ did not perform as well as UNet++ on both the training and the test dataset. This goes against our expectations since UNet3+ was developed to address many shortcomings of UNet++. We found that the addition of CBAM or SABM improved the performance of the network. We believe that addition of spatial attention modules forces the U-Net to focus on the vessel-like features. The addition of Tversky loss and lowered learning rate also improved the test result.

Figure 4 shows the segmentation masks of each network on a variety of hyperspectral images of the retina. Looking at the qualitative results, we found that our network shows promise in generating results. In Figure 4(a), our network trained on Tversky loss identified a missing blood vessel, which was omitted by every other network and by the annotator. In Figure 4(c), all networks identified a gap due to glare artifact, however our network bridges that gap.

Table 1. Jaccard score (intersection over union) of the segmentation methods.

	Fold					Total
	1	2	3	4	5	
UNet	0.73±0.04	0.72±0.05	0.73±0.05	0.71±0.04	0.74±0.04	0.73±0.04
UNet++	0.77±0.04	0.76±0.05	0.75±0.05	0.74±0.04	0.76±0.05	0.76±0.05
UNet3+	0.76±0.05	0.74±0.06	0.73±0.05	0.74±0.05	0.73±0.05	0.74±0.05
SAM	0.79±0.04	0.80±0.03	0.79±0.04	0.80±0.03	0.78±0.04	0.79±0.04
SAM 2	0.78±0.03	0.80±0.03	0.79±0.04	0.79±0.03	0.78±0.04	0.79±0.03
UNet3+CBAM	0.78±0.04	0.79±0.03	0.78±0.04	0.77±0.04	0.77±0.03	0.78±0.04
UNet3+SABM	0.80±0.03	0.78±0.03	0.79±0.04	0.78±0.02	0.78±0.04	0.79±0.03
<i>Ours</i>	0.82±0.04	0.81±0.03	0.80±0.04	0.79±0.04	0.78±0.05	0.80±0.04

Table 2. Dice similarity coefficient of the segmentation methods.

	Fold					Total
	1	2	3	4	5	
UNet	0.86±0.01	0.86±0.01	0.85±0.02	0.85±0.03	0.83±0.03	0.85±0.02
UNet++	0.88±0.01	0.88±0.01	0.87±0.01	0.87±0.01	0.87±0.02	0.87±0.01
UNet3+	0.88±0.01	0.88±0.01	0.87±0.01	0.88±0.01	0.87±0.02	0.88±0.01
SAM	0.85±0.03	0.86±0.02	0.87±0.02	0.86±0.02	0.84±0.02	0.86±0.02
SAM 2	0.85±0.03	0.87±0.02	0.87±0.02	0.87±0.02	0.84±0.02	0.86±0.02
UNet3+CBAM	0.88±0.02	0.88±0.01	0.88±0.01	0.88±0.01	0.86±0.02	0.88±0.01
UNet3+SABM	0.88±0.02	0.87±0.01	0.88±0.01	0.88±0.01	0.87±0.02	0.88±0.01
<i>Ours</i>	0.89±0.01	0.88±0.01	0.90±0.02	0.88±0.02	0.88±0.02	0.89±0.02

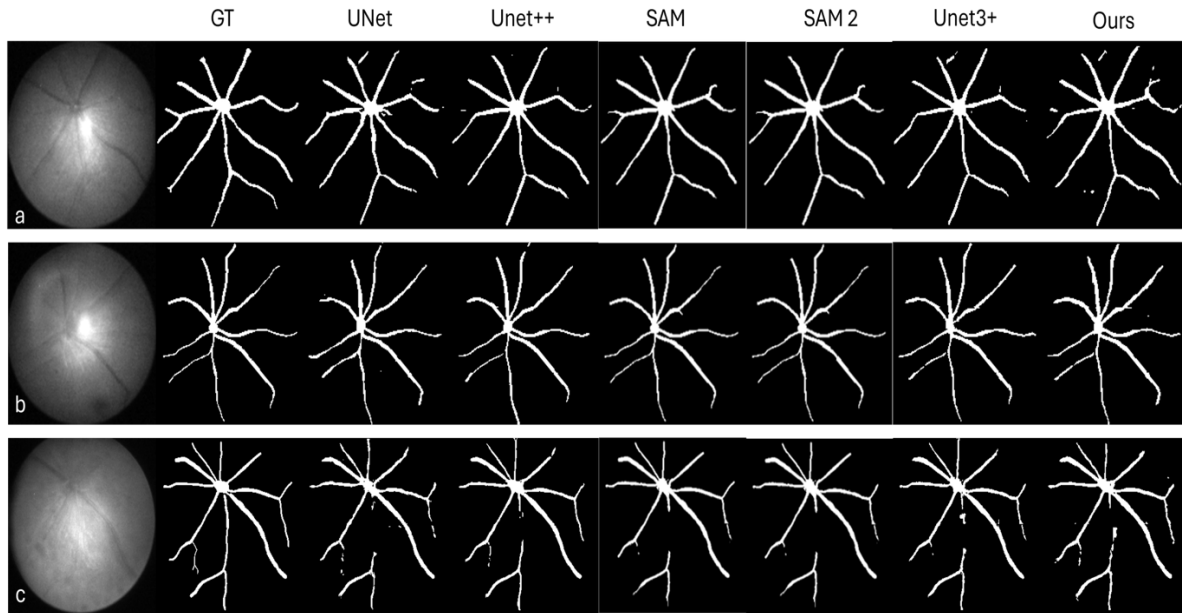


Fig. 4 The segmentation output produced by different networks. From left to right: one frame of the hyperspectral image, showing the image at wavelength 460 nm, ground truth used for training, output made by UNet, UNet++, SAM, SAM2, UNet3+, and our network, respectively.

4. Discussion and conclusion

We developed a deep learning architecture based on the UNet3+ algorithm with the novel addition of SABM and Tversky loss to segment retinal vessels on mouse HSI data. After acquiring 273 high quality mouse retina HSI and conducting manual retinal segmentation, we trained and tested our novel U-Net model, and other standard U-Net models, to evaluate the accuracy of their segmentation performances using DSC and IoU. Across five folds, our model achieved the highest mean performance among all tested models, with superior IoU and DSC scores. This supports that the incorporation of spectral multi-head attention with Tversky loss to a standard UNet3+ architecture helps improve the segmentation performance. The promising results of this segmentation model may provide a tool for retinal vessel analysis which may be applied to pre-diagnostic tests for various diseases affecting the retina and CNS. In the future, we will investigate the SABM architecture further to improve segmentation results. We will also compare different types of loss functions to determine their effects on segmentation accuracy. Future work will also include the evaluation of the segmentation model with human retinal datasets to make the model cross-species generalizable.

ACKNOWLEDGEMENTS

Research reported in this publication was supported in part by the National Cancer Institute of the National Institutes of Health under Award Number R01CA288379 and by the Cancer Prevention and Research Institute of Texas (CPRIT) under Award Number RP240289 and RP240542. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

- [1] A. London, I. Benhar, and M. Schwartz, "The retina as a window to the brain—from eye research to CNS disorders," *Nature Reviews Neurology*, vol. 9, no. 1, pp. 44-53, 2013/01/01 2013, doi: 10.1038/nrneurol.2012.227.
- [2] C. Chen, J. H. Chuah, R. Ali, and Y. Wang, "Retinal Vessel Segmentation Using Deep Learning: A Review," *IEEE Access*, vol. 9, pp. 111985-112004, 2021, doi: 10.1109/ACCESS.2021.3102176.

- [3] J. Dash and N. Bhoi, "Retinal blood vessel segmentation using Otsu thresholding with principal component analysis," in *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, 19-20 Jan. 2018 2018, pp. 933-937, doi: 10.1109/ICISC.2018.8398938.
- [4] K. BahadarKhan, A. A Khaliq, and M. Shahid, "A Morphological Hessian Based Approach for Retinal Blood Vessels Segmentation and Denoising Using Region Based Otsu Thresholding," *PLOS ONE*, vol. 11, no. 7, p. e0158996, 2016, doi: 10.1371/journal.pone.0158996.
- [5] C. Guo, M. Szemenyei, Y. Yi, W. Wang, B. Chen, and C. Fan, "SA-UNet: Spatial Attention U-Net for Retinal Vessel Segmentation," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 10-15 Jan. 2021 2021, pp. 1236-1242, doi: 10.1109/ICPR48806.2021.9413346.
- [6] Z. Wang, Z. Xie, and Y. Xu, "Automatic Segmentation for Retinal Vessel Using Concatenate UNet++ ", Cham, 2022: Springer International Publishing, in *The 2021 International Conference on Machine Learning and Big Data Analytics for IoT Security and Privacy*, pp. 10-18.
- [7] K.-W. Huang, Y.-R. Yang, Z.-H. Huang, Y.-Y. Liu, and S.-H. Lee, "Retinal Vascular Image Segmentation Using Improved UNet Based on Residual Module," *Bioengineering*, vol. 10, no. 6, p. 722, 2023. [Online]. Available: <https://www.mdpi.com/2306-5354/10/6/722>.
- [8] Z. Wu and X. Xiong, "Vessel-SAM2: Adapting Segment Anything 2 for Patch-Free Retinal Vessel Segmentation in Ultra-High Resolution Fundus Images," *IEEE Sensors Letters*, pp. 1-4, 2025, doi: 10.1109/LSENS.2025.3595139.
- [9] M. Haft-Javaherian, L. Fang, V. Muse, C. B. Schaffer, N. Nishimura, and M. R. Sabuncu, "Deep convolutional neural networks for segmenting 3D in vivo multiphoton images of vasculature in Alzheimer disease mouse models," *PLOS ONE*, vol. 14, no. 3, p. e0213539, 2019, doi: 10.1371/journal.pone.0213539.
- [10] S. Lemmens *et al.*, "Hyperspectral Imaging and the Retina: Worth the Wave?," (in eng), *Transl Vis Sci Technol*, vol. 9, no. 9, p. 9, Aug 2020, doi: 10.1167/tvst.9.9.9.
- [11] M. H. Tran *et al.*, "A dual-camera high-resolution hyperspectral imaging system for the retina," in *SPIE Medical Imaging*, 2025, vol. 13410: SPIE. [Online]. Available: <https://doi.org/10.1117/12.3047906>. [Online]. Available: <https://doi.org/10.1117/12.3047906>
- [12] S. S. More, J. M. Beach, and R. Vince, "Early Detection of Amyloidopathy in Alzheimer's Mice by Hyperspectral Endoscopy," *Investigative Ophthalmology & Visual Science*, vol. 57, no. 7, pp. 3231-3238, 2016, doi: 10.1167/iovs.15-17406.
- [13] Y. Xu, S. Hou, X. Wang, D. Li, and L. Lu, "A Medical Image Segmentation Method Based on Improved UNet 3+ Network," *Diagnostics*, vol. 13, no. 3, p. 576, 2023. [Online]. Available: <https://www.mdpi.com/2075-4418/13/3/576>.
- [14] H. Huang *et al.*, "UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4-8 May 2020 2020, pp. 1055-1059, doi: 10.1109/ICASSP40776.2020.9053405.
- [15] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks," Cham, 2017: Springer International Publishing, in *Machine Learning in Medical Imaging*, pp. 379-387.
- [16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," Cham, 2015: Springer International Publishing, in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234-241.
- [17] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," Cham, 2018: Springer International Publishing, in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 3-11.