# Multi-Attention Y-Net: A Spatiospectral Dual-Encoder Network for Medical Image Segmentation

M.A. Nusrath Amana ( Index no. 200022X)
*Department of Electronic and Telecommunication Engineering*
*University of Moratuwa, Sri Lanka*

***Abstract*—Automated segmentation of retinal optical coherence tomography (OCT) images has become a critical focus in medical image processing. The anatomic structure of retinal layers and their high-frequency variations make OCT imaging ideal for leveraging both spectral and spatial domain features. The proposed extension, MA-Ynet build on Y-Net, which integrates these features for improved segmentation, by enhancing its spatial encoder and decoder. This extension introduces a residual encoder with a simple attention module that strengthens the spatial encoder's ability to extract fine-grained features. By applying multi-head self-attention at the lowest-level features, I reconstruct the semantic representation of the feature map to improve feature extraction and leave the spectral encoder unchanged. Proposed approach achieves a mean Dice score of 0.86, advancing automated OCT image analysis.**

***Index Terms*—OCT segmentation; Y-Net; multi-head self-attention; channel attention; spatial attention**

## I. Introduction

Ocular Optical Coherence Tomography (OCT) is a widely used imaging modality in ophthalmology, aiding in the diagnosis and treatment of eye diseases such as diabetic macular edema (DME) and age-related macular degeneration (AMD). Accurate segmentation of intraretinal fluid pockets in OCT images is crucial, as it helps determine the presence, extent, and treatment response of retinal abnormalities. However, existing methods often fail to achieve efficient segmentation of these fluid regions due to limited feature extraction capabilities.

Recent advancements have demonstrated the potential of combining spectral and spatial domain features for improved segmentation performance. Previous work, such as Y-Net [1], introduced a dual-branch architecture that integrates spectral and spatial information, outperforming the widely used U-Net architecture in fluid segmentation by a minimum of 13% in dice score. Despite these improvements, challenges remain in extracting fine-grained features and addressing multiscale variations in complex scenes.

To address these limitations, I propose Multi-Attention Y-Net (MA-YNet), an enhanced version of Y-Net [1] that integrates multiple attention mechanisms to improve segmentation accuracy. My contributions include:

- Residual Structure in the Spatial Encoder: incorporate residual units with simple attention mechanisms to alleviate gradient vanishing issues and enhance the generalization capability of the backbone.
- Multi-Head Self-Attention: By applying multi-head self-attention to the lowest-level features, proposed method reconstruct feature maps, enabling better refinement of pixel-level segmentation, especially for adjacent regions.
- Channel and Spatial Attention Modules: To address the challenge of multiscale variations, proposed method introduce channel and spatial attention modules in the feature fusion stage. This ensures effective integration of features across scales, improving segmentation for targets of varying sizes.

## II. Related Work

Early methods for retinal OCT image segmentation primarily relied on graph-based approaches, such as graph cuts and shortest-path algorithms [2]. Following these initial efforts, hybrid methods combining neural networks with graph-based techniques were proposed to estimate retinal layer boundaries or to integrate graph convolutional networks with other architectures [3].

He et al. investigated OCT segmentation through a series of works [4], leveraging OCT scan topology for more robust segmentation. Fully convolutional networks (FCNs) were also explored for predicting segmentation maps, with additional corrections applied to ensure adherence to specific topology constraints [5].

Recent developments in medical image segmentation have focused on autoencoder-based deep neural networks [6] for end-to-end segmentation tasks. U-Net [7], one of the most widely adopted autoencoder architectures for 2D medical image segmentation, has inspired numerous advancements. For example, MDAN-U-Net [8] leverages multiscale features and attention mechanisms to enhance segmentation performance. Feature Pyramid Networks (FPNs) [9], which extract global features, have also been adapted for medical image segmentation.

In the context of OCT segmentation, specialized approaches have utilized Gaussian processes, feature alignment techniques, and epistemic uncertainty modeling [10]. Recurrent Neural Networks (RNNs) have also been applied, modeling inter-scan sequences or pixel sequences within scans [11].

Y-Net [1] introduced an autoencoder architecture with dual encoder branches that combine spatial and spectral features. The spectral branch employed fast Fourier transforms (FFTs) to capture high-frequency speckles, which are critical for OCT segmentation due to their tissue- and layer-specific properties. This approach addressed the limitations of spatial-only models, which struggle with high-frequency variations, by enabling feature disentanglement across frequency distributions.

Inspired by Y-Net [1] and advancements in attention mechanisms, I propose a novel Multi-Attention Y-Net (MA-YNet), incorporating features from MA-UNet [12] to address the challenges of fine-grained segmentation and multiscale target representation. Specifically, prposed MA-YNet employs a residual encoder with a simple attention module to enhance the extraction of fine-grained features. To refine semantic representations, proposed method utilize multi-head self-attention at the lowest feature level, reconstructing feature maps for improved segmentation of complex pixel categories. Additionally, proposed method integrate channel attention and spatial attention at various stages of feature fusion, effectively merging information from targets at diverse scales. These enhancements build upon the foundation of Y-Net [1] and further advance the state of the art in OCT image segmentation, enabling better performance across complex and multiscale regions.

## III. Method

In this section, I describe the foundational components of the proposed approach.

### A. Segmentation Framework

The proposed Y-Net segmentation framework predicts the segmentation map $y$ for a given input image $x \in \mathbb{R}^{H\times W}$, where $H$ and $W$ denote the image height and width. As shown in Fig.1, Y-Net consists of two encoder branches:

- **Spatial Encoder ($E_c$)**: This branch utilizes convolutional blocks with residual structure based on simple attention module (simAM) to capture local spatial features.
- **Spectral Encoder ($E_f$)**: This branch is designed to extract global frequency-domain features using fast Fourier convolutional (FFC) blocks.

The decoder network $G$, which receives fused spatial and spectral features, generates the segmentation map $y$ as:

$$y = G(E_c(x), E_f(x)).$$

Similar to U-Net, Y-Net employs skip connections between the spatial encoder and decoder blocks for enhanced feature propagation.

*1) Spatial Encoder:* The spatial encoder $E_c$ adopts a U-Net-like structure, comprising four encoder blocks. I use a residual structure and a simAM to construct the encoder and build an attention-based residual encoder, to improve the fine extraction ability of the backbone for target features and max pooling for down-sampling. The input image is processed sequentially through these blocks to extract spatial features.

*2) Spectral Encoder:* The spectral encoder $E_f$, introduced to augment frequency-domain feature extraction, receives the same input image as the spatial encoder. It consists of four FFC blocks, each designed to process local and global spectral information. The spectral features are processed through Fourier transformations, convolutional operations, and inverse transformations, ensuring seamless integration into the spatial domain.

*3) Spatial Decoder:* The decoder $G$ concatenates spectral and spatial features from the respective encoders and uses multi-head self attention(MSA) for the lowest level feature to rebuild the feature map and enhance the semantic representation of each feature point on the feature map. then uses up-convolutional blocks with channel attention and spatial attention mechanisms to generate the final segmentation map.

### B. Spatial Encoder Components

*1) Residual Block:* The residual structure, originally introduced in ResNet, addresses gradient vanishing and explosion issues by enabling shortcut connections. proposed method adopt a conventional residual structure, where each block includes two 3x3 convolutions and a parameter-free attention module (simAM) for neuron importance evaluation.

*2) simAM:* The simAM module is a parameter-free attention mechanism designed to enhance feature extraction in convolutional neural networks with minimal computational overhead. It evaluates the importance of each neuron in a feature map using an energy-based attention mechanism.

*a) Feature Space Mean::*

$$d = (X - X.\text{mean}(\text{dim} = [2, 3]))^2$$

*b) Variance::*

$$v = \frac{d.\text{sum}(\text{dim} = [2, 3])}{H + W - 1}$$

*c) Energy Distribution::*

$$E = \frac{d}{4(v + \rho)} + 0.5$$

*d) Enhanced Feature Map::*

$$X' = \text{sigmoid}(E) \odot X$$

*3) Channel and Spatial Attention Modules::* the shallow features, because of their relatively large feature map resolutions and spatial feature distribution, have a greater impact on feature fusion, so Spatial Attention Module (SAM) is utilized to integrate the fused features at the first two scales. High dimensional features tend to be compressed in channels, so fused features are integrated by using Channel Attention Module (CAM) at the last two scales.The structures of the SAM and CAM are shown in Fig. 2.

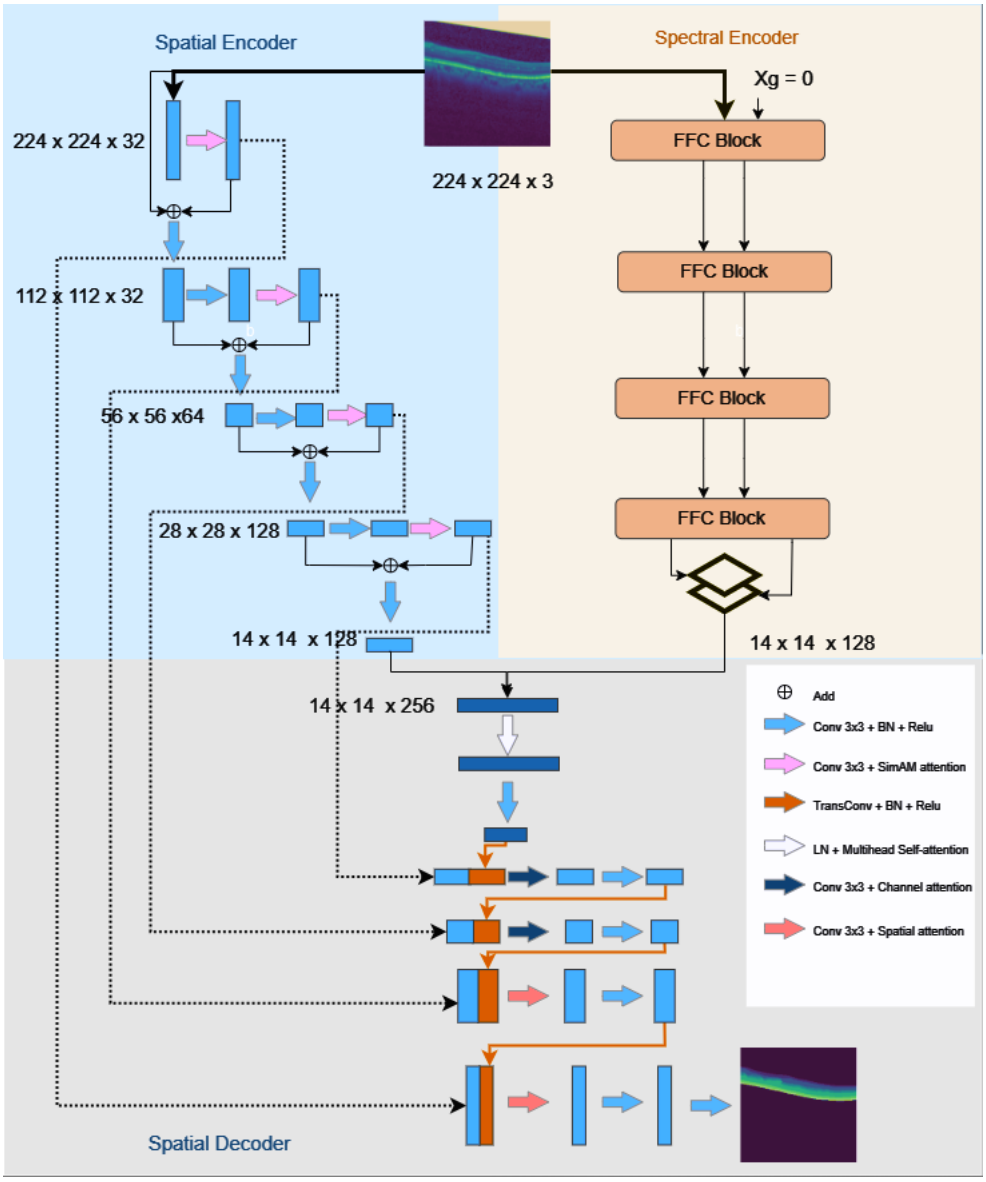


Fig. 1. The structure of MA-YNet; LN in the figure represents layer normalization and BN represents batch normalization.

*4) Multi-Head Self-Attention (MSA):* After down-sampling, MSA enhances global correlations by generating query, key, and value vectors to capture relationships between spatial features. The feature map ($14 \times 14 \times 256$) is flattened into a $256 \times 196$ matrix, processed through layer normalization and MSA, and then reshaped back to its original size. This approach preserves spatial context while leveraging global attention.

*5) Feature Fusion Based on Attention Enhancement:* The Fourier unit applies the Fast Fourier Transform (FFT) to input features, obtaining real and imaginary components ($a + bi$). These components are processed via convolutional layers, followed by activation and normalization. Finally, the features are converted back to the spatial domain using the inverse FFT.

### C. Spectral Encoder Components

*1) Fast Fourier Convolutional (FFC) Block:* Each FFC block, as shown in Fig.4 takes local and global feature maps ($x_l, x_g$) as input. These are processed through convolutional layers and a spectral normalization step to extract enriched spectral features. The resulting features are normalized, activated, and pooled for subsequent blocks.

*2) Fourier Unit:* The Fourier unit applies the Fast Fourier Transform (FFT) to input features, obtaining real and imaginary components ($a+bi$). These components are processed via convolutional layers, followed by activation and normalization. Finally, the features are converted back to the spatial domain using the inverse FFT.

### D. Loss Functions

The training process uses a combined loss function to optimize segmentation performance. Specifically, I use a com-

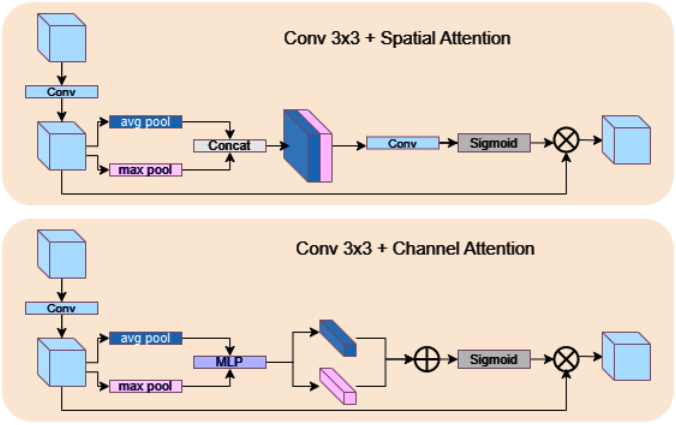


Fig. 2. Feature enhancement based on SAM and CAM.

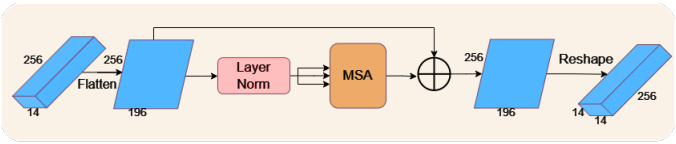


Fig. 3. The structure of MSA applying to image features rebuilding.

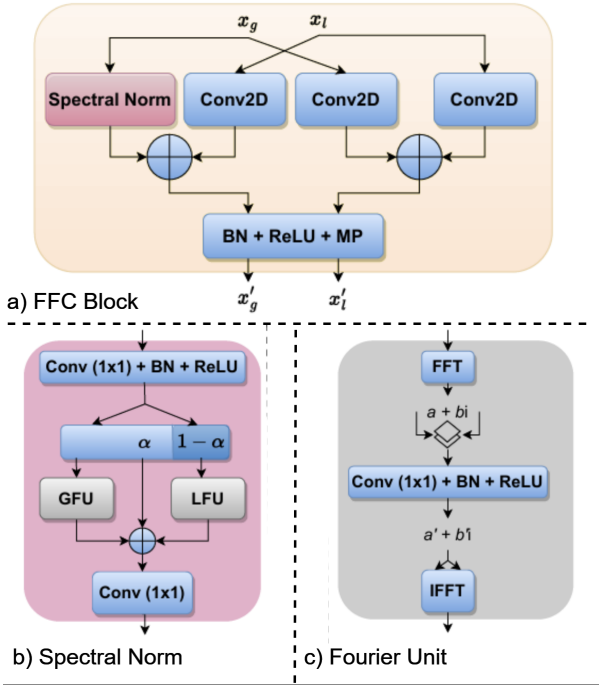


Fig. 4. FFc block architecture

bination of Dice loss and cross-entropy loss as Y-Net [1]:

*a) Dice Loss::* Measures the overlap between predicted and ground truth segmentation maps:

$$L_{\text{Dice}}(y, \hat{y}) = 1 - \frac{2 \cdot \sum y \cdot \hat{y} + \epsilon}{\sum y + \sum \hat{y} + \epsilon}$$

where $\epsilon$ ensures numerical stability.

*b) Cross-Entropy Loss::* Maximizes information gain between true and predicted labels:

$$L_{\text{CE}}(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

*c) Combined Loss::* The total loss is a weighted sum of Dice and cross-entropy losses:

$$L_{\text{total}} = \lambda_{\text{Dice}} \cdot L_{\text{Dice}} + \lambda_{\text{CE}} \cdot L_{\text{CE}}$$

## IV. Experiments

In this section, I evaluate proposed method, MA-YNet, and compares it with Y-Net[1] and baseline approaches using the Duke OCT dataset [13].

### A. Experimental Setup

The dataset comprises OCT scans from 10 patients, annotated by two experts. The experimental protocol aligns with prior works, where scans from the first six subjects are designated for training, those from subjects seven and eight for validation, and the remaining two subjects for testing. Training was conducted with a batch size of 10, a learning rate of $5 \times 10^{-4}$, weight decay of $1 \times 10^{-4}$, and a maximum of 100 epochs using the Adam optimizer. The number of training epochs was determined based on the best validation accuracy. The images were resized to $224 \times 224$, and the combined loss function utilized equal weights for Dice loss and cross-entropy loss. Additional experiments with focal loss were conducted but did not yield significant performance improvements over the baseline loss function.

### B. Results

The performance of MA-YNet was compared against Y-Net [1], U-Net [7] and other prior works for retinal layer and fluid segmentation tasks. The Dice scores, as summarized in Table I, demonstrate that MA-YNet outperformed Y-Net [1] in segmenting all retinal layers. However, Y-Net[1] exhibited superior performance in fluid segmentation.

### C. Discussions and Conclusion

This paper introduce MA-YNet, an enhanced autoencoder-based architecture for retinal layer and fluid segmentation in OCT images. By integrating multi-attention mechanisms with spectral and spatial domain features, MA-YNet effectively captures fine-grained details and frequency-dependent patterns. While MA-YNet outperforms Y-Net [1] and other baselines in retinal layer segmentation, it shows slightly lower performance in fluid segmentation compared to Y-Net [1]. This highlights the potential for further optimization in the spectral encoding branch.

TABLE I
MEAN AND PER LAYER DICE SCORE COMPARED TO RELATED WORKS ON THE PUBLICLY AVAILABLE DUKE OCT DATASET [13]

| Method | ILM | NFL-IPL | INL | OPL | ONL-ISM | ISE | OS-RPE | Fluid | Mean |
|---|---|---|---|---|---|---|---|---|---|
| RelayNet [14] | 0.84 | 0.85 | 0.70 | 0.71 | 0.87 | 0.88 | 0.84 | 0.30 | 0.75 |
| Language [15] | 0.85 | 0.89 | 0.75 | 0.75 | 0.89 | 0.90 | **0.87** | 0.39 | 0.78 |
| Alignment [16] | 0.85 | 0.89 | 0.75 | 0.74 | 0.90 | 0.90 | **0.87** | 0.56 | 0.81 |
| U-Net [7] | 0.84 | 0.89 | 0.77 | 0.76 | 0.89 | 0.89 | 0.85 | 0.80 | 0.836 |
| Y-Net [1] | 0.86 | 0.89 | **0.78** | 0.75 | 0.90 | 0.88 | 0.85 | **0.93** | 0.855 |
| MA-YNet(Proposed) | **0.87** | **0.91** | **0.78** | **0.78** | **0.91** | **0.91** | 0.86 | 0.85 | **0.86** |

## REFERENCES


[1] A. Farshad, Y. Yeganeh, P. Gehlbach, and N. Navab, "Y-net: A spatiospectral dual-encoder network for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds. Cham: Springer Nature Switzerland, 2022, pp. 582–592.

[2] K. Hu, B. Shen, Y. Zhang, C. Cao, F. Xiao, and X. Gao, "Automatic segmentation of retinal layer boundaries in oct images using multiscale convolutional neural network and graph search," *Neurocomputing*, vol. 365, pp. 302–313, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231219310860

[3] J. Li, P. Jin, J. Zhu, H. Zou, X. Xu, M. Tang, M. Zhou, Y. Gan, J. He, Y. Ling, and Y. Su, "Multi-scale gcn-assisted two-stage network for joint segmentation of retinal layers and disc in peripapillary oct images," 02 2021.

[4] Y. He, A. Carass, B. Jedynak, S. Solomon, S. Saidha, P. Calabresi, and J. Prince, "Topology guaranteed segmentation of the human retina from oct using convolutional neural networks," 03 2018.

[5] Y. He, A. Carass, Y. Yun, C. Zhao, B. Jedynak, S. Solomon, S. Saidha, P. Calabresi, and J. Prince, "Towards topological correct segmentation of macular oct from cascaded fcns," vol. 10554, 09 2017, pp. 202–209.

[6] F. Kiaee, H. Fahimi, and H. Rabbani, "Intra-retinal layer segmentation of optical coherence tomography using 3d fully convolutional networks," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 2795–2799.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[8] W. Liu, Y. Sun, and Q. Ji, "Mdan-unet: Multi-scale and dual attention enhanced nested u-net architecture for segmentation of optical coherence tomography images," *Algorithms*, vol. 13, p. 60, 03 2020.

[9] S. Feng, H. Zhao, F. Shi, X. Cheng, M. Wang, Y. Ma, D. Xiang, W. Zhu, and X. Chen, "Cpfnet: Context pyramid fusion network for medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. PP, pp. 1–1, 03 2020.

[10] J. Orlando, P. Seeböck, H. Bogunović, S. Riedl, C. Grechenig, S. Waldstein, B. Gerendas, and U. Schmidt-Erfurth, "U2-net: A bayesian u-net model with epistemic uncertainty feedback for photoreceptor layer segmentation in pathological oct scans," 01 2019.

[11] A. Tran, J. Weiss, S. Albarqouni, S. Roohi, and N. Navab, *Retinal Layer Segmentation Reformulated as OCT Language Processing*, 09 2020, pp. 694–703.

[12] Y. Sun, F. Bi, Y. Gao, L. Chen, and S. Feng, "A Multi-Attention UNet for Semantic Segmentation in Remote Sensing Images," *Symmetry*, vol. 14, no. 5, 2022. [Online]. Available: https://www.mdpi.com/2073-8994/14/5/906

[13] S. J. Chiu, M. J. Allingham, P. S. Mettu, S. W. Cousins, J. A. Izatt, and S. Farsiu, "Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema." *Biomedical optics express*, vol. 6 4, pp. 1172–94, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:12597457

[14] A. Guha Roy, S. Conjeti, S. P. Karri, D. Sheet, A. Katouzian, C. Wachinger, and N. Navab, "Relaynet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks," *Biomedical Optics Express*, vol. 8, 07 2017.

[15] A. Tran, J. Weiss, S. Albarqouni, S. Faghi Roohi, and N. Navab, "Retinal layer segmentation reformulated as oct language processing," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, Eds. Cham: Springer International Publishing, 2020, pp. 694–703.

[16] H. Maier, S. Faghihroohi, and N. Navab, "A line to align: Deep dynamic time warping for retinal oct segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, Eds. Cham: Springer International Publishing, 2021, pp. 709–719.