bioengineering

MDPI

Article

Improving OCT Image Segmentation of Retinal Layers by Utilizing a Machine Learning Based Multistage System of Stacked Multiscale Encoders and Decoders

Arunodhayan Sampath Kumar [1], Tobias Schlosser [1], Holger Langner [2], Marc Ritter [2] and Danny Kowerko [1,*]

[1] Junior Professorship of Media Computing, Chemnitz University of Technology, 09107 Chemnitz, Germany; arunodhayan.sampath-kumar@cs.tu-chemnitz.de (A.S.K.); tobias.schlosser@cs.tu-chemnitz.de (T.S.)
[2] Professorship of Media Informatics, University of Applied Sciences Mittweida, 09648 Mittweida, Germany; langner@hs-mittweida.de (H.L.); ritter@hs-mittweida.de (M.R.)
* Correspondence: danny.kowerko@cs.tu-chemnitz.de

Abstract: Optical coherence tomography (OCT)-based retinal imagery is often utilized to determine influential factors in patient progression and treatment, for which the retinal layers of the human eye are investigated to assess a patient's health status and eyesight. In this contribution, we propose a machine learning (ML)-based multistage system of stacked multiscale encoders and decoders for the image segmentation of OCT imagery of the retinal layers to enable the following evaluation regarding the physiological and pathological states. Our proposed system's results highlight its benefits compared to currently investigated approaches by combining commonly deployed methods from deep learning (DL) while utilizing deep neural networks (DNN). We conclude that by stacking multiple multiscale encoders and decoders, improved scores for the image segmentation task can be achieved. Our retinal-layer-based segmentation results in a final segmentation performance of up to $82.25 \pm 0.74\%$ for the Sørensen–Dice coefficient, outperforming the current best single-stage model by 1.55% with a score of $80.70 \pm 0.20\%$, given the evaluated peripapillary OCT data set. Additionally, we provide results on the data sets Duke SD-OCT, Heidelberg, and UMN to illustrate our model's performance on especially noisy data sets.

Keywords: ophthalmology; ophthalmology diseases; OCT biomarkers; OCT segmentation; computer vision and pattern recognition; machine learning; deep learning

Citation: Sampath Kumar, A.; Schlosser, T.; Langner, H.; Ritter, M.; Kowerko, D. Improving OCT Image Segmentation of Retinal Layers by Utilizing a Machine Learning Based Multistage System of Stacked Multiscale Encoders and Decoders. *Bioengineering* **2023**, *10*, 1177. https://doi.org/10.3390/bioengineering10101177

Academic Editors: Alan Wang, Sibusiso Mdletshe, Brady Williamson and Guizhi Xu

Received: 31 August 2023
Revised: 2 October 2023
Accepted: 5 October 2023
Published: 10 October 2023

1. Introduction and Motivation

Humans are highly dependent on their vision for social interactions. Globally, 39 million people are visually impaired and are aged above 50 years [1]. The primary causes of blindness are cataracts (51%), glaucoma (8%), age-related macula degeneration (AMD) (5%), corneal opacities (4%), uncorrected refractive errors and trachoma (each 3%), and diabetic retinopathy (1%) [2,3]. In the KORA-Age study conducted in the southern part of Germany, 822 participants (49.6% women, 50.4% men, aged 68–96 years) were asked standard questions related to eye diseases. The most common eye diseases were cataracts (36%), dry eyes (15%), glaucoma (9%), and AMD (8%) [4]. Most of the participants were suffering from glaucoma or AMD while having cataracts. Another study estimated the incidences of severe visual impairment and blindness by using Germany's largest state blind registry and their projection rates for Germany in 2010 and 2030. The major causes of blindness and visual impairment were AMD (50%), glaucoma (15%), and diabetic retinopathy (10%) [5].

A non-invasive three-dimensional imaging modality used in eye clinics for the diagnosis of pathologies is optical coherence tomography (OCT) [6], a non-invasive imaging technique that uses light waves to capture biological tissue in high resolution. It is an important technique in organs where the traditional microscopic tissue diagnosis method employing a biopsy is unavailable [7]. For OCT, near-infrared light with wavelengths

ranging from about 800 to 1300 nm is typically employed, enabling deeper penetration into tissues such as the retina of the human eye compared to visible light [8]. Here, the OCT B-scan provides a two-dimensional, cross-sectional view to allow the visualization of the retina and cornea, which aids in the diagnosis of diseases such as AMD, glaucoma, and diabetic retinopathy [9].

Medical image semantic segmentations are a wide area of research that involves the extraction of regions of interest (ROIs) from medical images, such as in the form of computed tomography (CT) scans, magnetic resonance imaging (MRI) scans, X-ray scans, and OCT scans [10]. However, manual segmentation is an incredibly time-consuming task [11]. Therefore, current approaches also encompass the development of algorithms incorporating machine learning (ML) and deep learning (DL) in the form of deep neural networks (DNN), whereas models trained with OCT imagery enable the (semi-)automated classification of OCT biomarkers [12] and the segmentation of OCT imagery into the different layers of the human eye (retinal layers) [13,14]. Finally, (semi-)automated recommender systems can assist doctors with the identification and location of abnormalities such as AMD [15], diabetic retinopathy [16], and glaucoma [17], ultimately aiding in the reduction of visual impairments and blindness among patients.

1.1. Related Work

The success of DNNs such as convolution neural networks (CNN) in object detection has helped researchers to explore the feature learning capabilities of prediction problems such as image segmentation [18–20]. Research on semantic pixel-wise segmentation is an ongoing topic that is driven by complex data sets. Prior to the development of DNNs, the most effective techniques mainly used manually created characteristics to categorize individual pixels [21]. To estimate the class probabilities of a ROI, a patch is typically fed through a machine learning classifier, such as a random forest classifier [22], a voting classifier [23], or boosting [24]. Recent medical image segmentation methods primarily utilized autoencoder-based DNNs for end-to-end segmentation [25,26]. The most commonly used OCT-based segmentation models are the fully convolutional networks (FCNs) [27] and U-Net [28]. Other segmentation methods applied to medical image segmentation are the region-based CNNs (R-CNN) [29] and instance segmentation [30].

In terms of the state-of-the-art models for macular-based OCT segmentation, ReLayNet was proposed for the semantic segmentation of macular OCT B-scans into their retinal layers and fluid masses [25]. This method paved the way for a baseline for the automatic segmentation of retinal OCT layers. An attention-guided channel-to-pixel CNN for retinal layer segmentation with choroidal neovascularization was designed by [17], where a channel-to-pixel block is utilized along with an edge loss function to segment the retinal layers with blurry boundaries. The attention mechanism has been employed to address the sizeable morphological variation among retinal layers [31]. The work of [32] utilized a multiscale and dual attention (MDAN) U-Net with multiscale features and attention mechanisms to further improve the segmentation performance of U-Net. Additionally, commonly used OCT segmentation methods are the feature pyramid networks (FPNs) [33] for global feature extraction, followed by a Gaussian process and feature alignment with epistemic uncertainty [34,35]. Subsequently, [36] developed a so-called fully convolutional instance segmentation (FCIS) network with a segment proposal network and an object detection system.

Recent natural language processing (NLP) developments using recurrent neural networks (RNNs) have been explored for OCT segmentation. These models consider sequences between different scans for processing pixel sequences [37,38]. The work of [39] developed a polyp detection system using an autoencoder network with an encoder (pre-trained VGG19) and a decoder (U-Net) [28] branch to segment colon cancer cells. Y-Net [40], a model inspired by U-Net, was developed to segment cancer cells irrespective of shape, size, texture, and orientation. With state-of-the-art autoencoders, the encoder often produces low-resolution representations, while the decoder is responsible for producing multidi-

mensional features for each pixel for classification [41]. The authors utilized a single-stage training process where the encoder was the VGG16 with weights pre-trained on ImageNet. The decoder was an FCN architecture that learns to upsample its input feature maps along the encoder's feature map to produce the input to its respective decoder [40].

Often, decoder networks progressively add existing networks as single-stage approaches until no further improvement is achieved [42]. Further, prediction capabilities can be improved by appending an RNN to the existing FCN network as RNN layers mimic conditional random fields' (CRF) sharp boundary delineation capabilities while exploring the feature representation of the FCN [30]. Recently developed segmentation architectures using DNNs are not fed forward during the inference time [43,44]. They require aids such as region proposal networks for inference. Multiscale deep architectures have become a common approach for feature extraction. They use input images at multiple scales with their corresponding feature maps from different layers [45]. Others employ a combination of feature maps from different layers in a single deep architecture [46]. However, the common idea of incorporating global and local contexts is to extract features at multiple scales [47].

For peripapillary retinal images, ref. [48] developed an automatic segmentation method. With their approach, the boundaries of the optic discs (left and right) were determined based on the estimation of the position of Bruch's membrane openings in radially OCT B-scans by combining a CNN with a multigraph search algorithm that supports the segmentation of retinal boundaries. Finally, a Dilated-Residual U-Net (DRUNET) was proposed by [49] to segment five retinal layers and optic discs that are not thoroughly segmented from their connecting tissues.

In comparison to approaches that are commonly deployed for image segmentation, generative approaches such as variational autoencoders (VAE) [50] and generative adversarial networks (GAN) [51] are currently being investigated regarding their diagnostic abilities within ophthalmology [52]. In [52], the authors report on the capabilities of GANs, ranging from the segmentation of images to their augmentation, denoising, domain transfer, super resolution, post-intervention prediction, and feature extraction. However, given their limitations, such as the risk of generating additional noises and artifacts, generative approaches to OCT image segmentation are still being developed, whereas AUC scores within the range of 92 to 97% are being reported [52,53].

1.2. Contribution of This Work and Future Prospects

In this contribution, we propose a machine-learning-based multistage system of stacked multiscale encoders and decoders to perform the retinal layer segmentation task. This approach is motivated by current developments by [17] on the segmentation of peripapillary OCT imagery, where the segmentation task is realized using a multiscale graph convolutional network (GCN)-assisted two-stage network for the OCT image segmentation task. Here, we extend this approach by utilizing an approach where multiple encoders and decoders are stacked to obtain improved segmentation scores. As the results of this segmentation task can be potentially leveraged for the following classification stage, the importance of the segmentation task itself is emphasized, given the retinal layers' physiological and pathological states.

In regards to future prospects, retinal layer segmentation could serve as a visual aid during the education of (medical) students as well as medical assistants working towards becoming professionals in ophthalmology. With the provided retinal layer segmentations, they are further enabled to learn, explore, and understand the exact locations as well as the extent of the retinal conditions, which is crucial for effective treatment planning. In particular, retinal disorders and pathological conditions emerging from fluids are accessible in a quantitative manner. Combining multiple OCT B-scan segmentations could even enable future three-dimensional visualizations, therefore providing a better understanding of the anatomical structures and pathological conditions. Additionally, the obtained segmentations can be further utilized for subsequent processing steps such as OCT biomarker

classification, visual acuity predictions, as well as treatment predictions and adjustments. By leveraging community data sets with different physiological and pathological states, the diagnostic skills of healthcare professionals can be supported, and even patients can be educated about their eye conditions.

1.3. Section Overview

In the following text, our contribution is structured as follows: Firstly, our materials, methods, implementation of our data sets, model architecture, training setup, loss functions, and evaluation principles are introduced in Section 2. Our test results, evaluation, and discussion include a quantitative evaluation of the OCT image segmentation for retinal layers by utilizing our proposed system in comparison to currently developed and investigated baseline approaches in Section 3. In addition to our quantitative results, a qualitative evaluation using segmentation visualizations is provided to illustrate our results. Finally, we provide conclusions on the benefits of our system while giving additional insights into future developments and further possibilities for improvement.

2. Materials, Methods, and Implementation

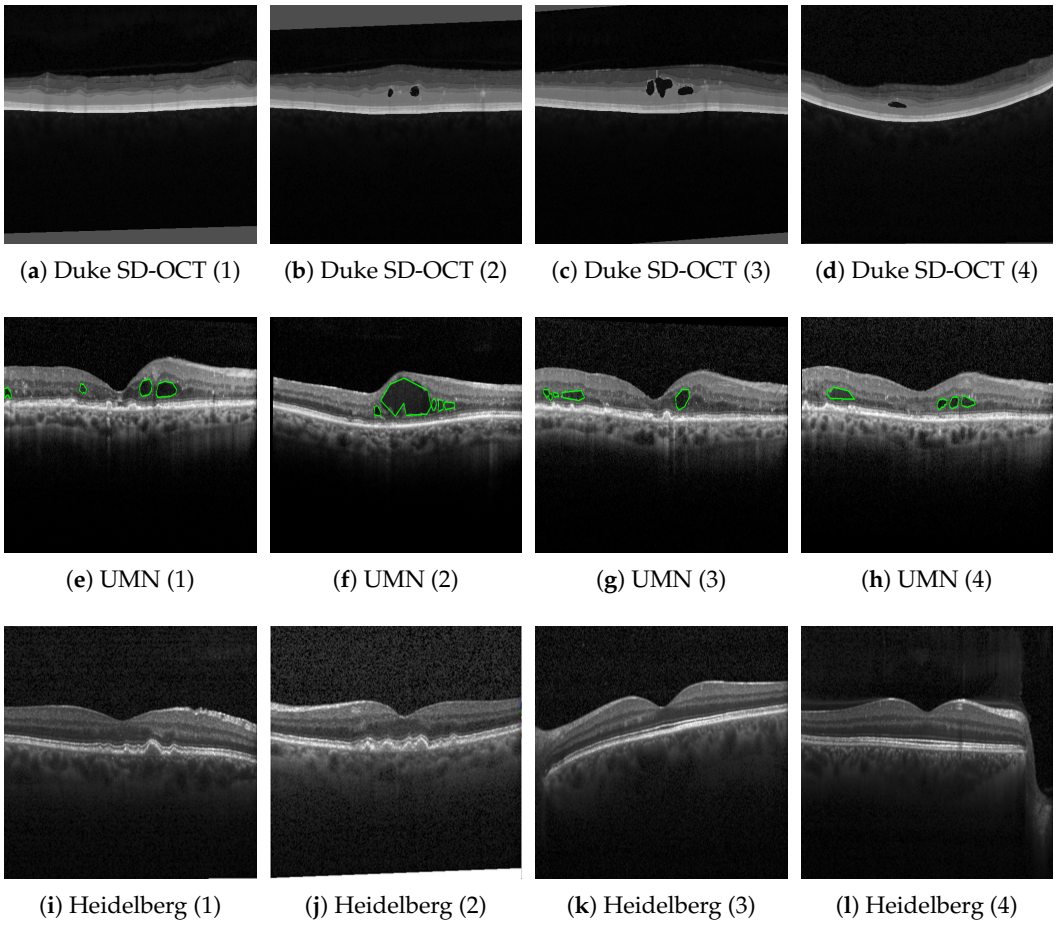
First, a suitable data corpus has to be defined to realize the OCT image segmentation of retinal layers. For this purpose, Section 2.1 introduces our data sets with their OCT imagery and segmentations. The model architecture used for our machine-learning-based multistage system of stacked multiscale encoders and decoders is introduced in Section 2.2 with its training setup and utilized loss functions presented in Sections 2.3 and 2.4. Finally, our evaluation principles and used metrics are highlighted in Section 2.5.

2.1. Data Sets

In the following text, the data sets used for the quantitative as well as the qualitative evaluation are introduced. For this purpose, Figure 1 gives an overview of our public data sets that were used for the qualitative assessment, Duke SD-OCT [14] (top row), UMN [54] (middle row), and Heidelberg [55] (bottom row), with and their annotations, if available. For the Duke SD-OCT data set, the retinal layers are highlighted in terms of their different levels of brightness. For the UMN data set, areas with fluids are additionally available. For the Heidelberg data set, no annotations are available. Additionally, Table 1 gives a tabular overview of the public data sets used in our further experiments.

2.1.1. Peripapillary Data Set

The peripapillary OCT data set [17] used for the evaluation of our system comprises 122 randomly selected radial OCT B-scans from 61 patients with 10 manually annotated labels (also see Figure 2). For the original data set, two graders annotated the OCT B-scans via ITK-SNAP [56] with the help of glaucoma specialists. The obtained annotations reflect the consensus of all graders [17]. With the provided radial OCT B-scans, 12 image slices are available per OCT B-Scan, where for each patient, one eye was imaged. Each patient contributed two randomly selected image slices, resulting in 122 image slices overall. A horizontal flip augmentation, additive Gaussian noises, and contrast adjustment were performed on the training data set to augment the number of samples by a factor of two, resulting in 244 samples in total. Here, we follow this principle to allow a comparison of the results previously obtained by [17]. Table 2 shows the utilized OCT image segmentation data set with its retinal layers, abbreviations, and color scheme for visualization [17]. Our training, validation, and testing sets were split with a data set split ratio of 60/20/20, resulting in 148/48/48 images, respectively. The image resolution of 1024×992 pixels is constant over all images and data subsets. All images contain annotations of layer segmentations, as given in Table 2.



(**a**) Duke SD-OCT (1) (**b**) Duke SD-OCT (2) (**c**) Duke SD-OCT (3) (**d**) Duke SD-OCT (4)

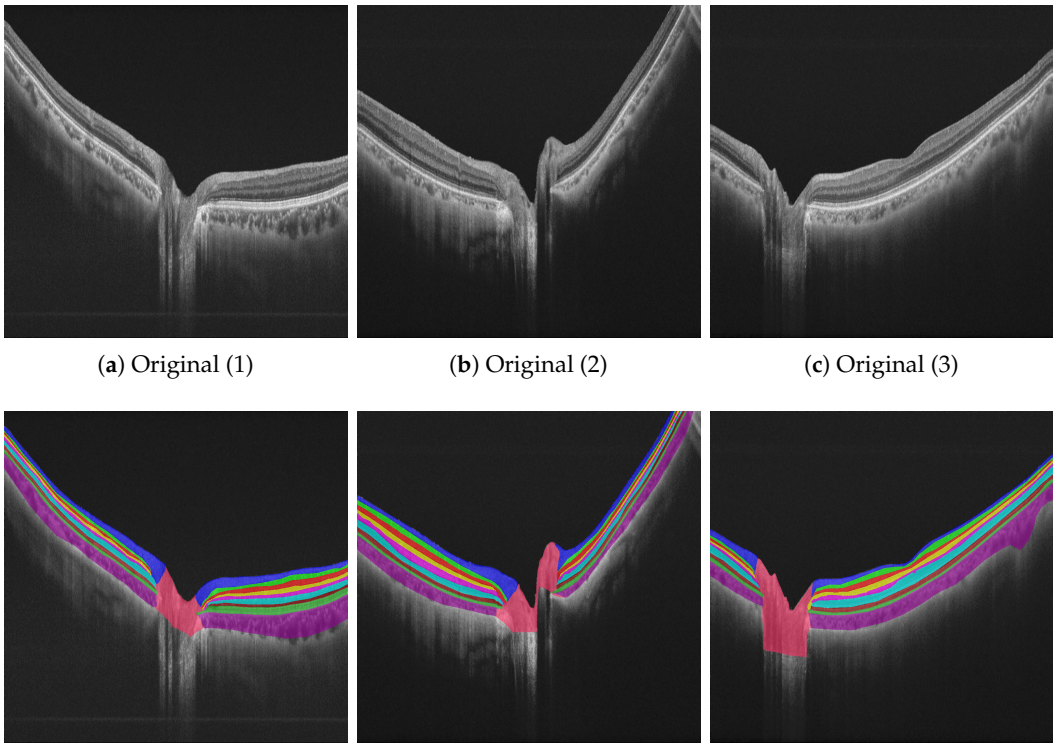
(**e**) UMN (1) (**f**) UMN (2) (**g**) UMN (3) (**h**) UMN (4)

(**i**) Heidelberg (1) (**j**) Heidelberg (2) (**k**) Heidelberg (3) (**l**) Heidelberg (4)

Figure 1. Overview of our public data sets for qualitative assessment, Duke SD-OCT [14] (**top row**), UMN [54] (**middle row**), and Heidelberg [55] (**bottom row**), with and their annotations, if available. For the Duke SD-OCT data set, the retinal layers are highlighted in terms of different levels of brightness. For the UMN data set, areas with fluids are additionally available (highlighted in green). For the Heidelberg data set, no annotations are available.

Table 1. Overview of our public data sets used for the qualitative assessment with their number of images, image resolution, and availability of annotations for segmentations (✓). There are no annotations available for the Heidelberg data set (-). For the annotation details see Sections 2.1.2 and 2.1.3.

Data Set Name	Number of Images	Image Resolution	Annotations
Duke SD-OCT [14]	88	536 × 496	✓
UMN [54]	125	1024 × 496	✓
Heidelberg [55]	125	1024 × 992	-



(**a**) Original (1) (**b**) Original (2) (**c**) Original (3)

(**d**) Annotated (1) (**e**) Annotated (2) (**f**) Annotated (3)

Figure 2. Overview of our data set with two selected samples and their visualized retinal layers following the color scheme presented in Table 2 as a legend for the present visualizations. Table 2 gives an overview of our data set with its samples and the proposed data set split ratio. In the top row, exemplary peripapillary OCT images are shown with their annotations in the bottom row.

Table 2. The data set color scheme for our data set with its retinal layer based segmentations, ranging from the retinal nerve fiber layer (RNFL) to the optic disc. Figure 2 gives an overview of our data set with a few selected samples and visualized retinal layers.

Layer	Color Scheme
Retinal nerve fiber layer (RNFL)	
Ganglion cell layer (GCL)	
Inner plexiform layer (IPL)	
Inner nuclear layer (INL)	
Outer plexiform layer (OPL)	
Outer nuclear layer (ONL)	
Inner/outer photoreceptor segment (IS/OS)	
Retinal pigment epithelium (RPE)	
Choroid	
Optic disc	

2.1.2. Duke SD-OCT Data Set

The Duke University provides the publicly available Duke SD-OCT data set [14], which comprises 110 OCT B-scans recorded from 10 diabetic macular edema (DME) patients. The data set includes fluid and non-fluid manual annotations of eight boundaries by two ophthalmologists, encompassing RNFL, GCL, INL, OPL, ONL, the inner segment ellipsoid (ISE), and the outer segment retinal pigment epithelium (OS-RPE).

2.1.3. UMN Data Set

The Minnesota Ophthalmology University Clinic collected the UWN data set [54], which comprises 600 OCT B-scans from 24 AMD patients. Within the UMN data set, retinal fluid regions—intraretinal fluid (IRF), subretinal fluid (SRF), and pigment epithelial detachment (PED)—were manually annotated and assessed by two ophthalmologists.

2.1.4. Heidelberg Data Set

The Heidelberg data set [55] comprises 108,312 OCT B-scans recorded from 4686 patients with retinal fluid annotations, including 37,206 images with choroidal neovascularization (CNV), 11,349 images with DME, 8617 images with drusen, and 51,140 healthy images. The retinal fluid annotations were manually annotated with a tiered grading system. The employed undergraduate and medical students were first-tier graders who reviewed the diagnostic information and discarded OCTs contaminated by severe artifacts. The following four ophthalmologists were second-tier graders who independently graded the images as CNV, DME, and drusen.

2.2. Model Architecture

The model architecture of our proposed system where multiple encoders and decoders are combined as stacks is shown in Figure 3. In Figure 3a, the general model is illustrated. Figure 3b shows our proposed and implemented architecture, consisting of two main parts: stack 1 with encoder 1 and decoder 1 (Section 2.2.1), where the first decoder, our Modified Attention U-Net (Section 2.2.3), uses a local context, as well as stack 2 with encoder 2 and decoder 2 (Section 2.2.2), where the second decoder, again our Modified Attention U-Net, uses a larger context as the denoiser. The proposed segmentation network predicts the segmentation map given an input OCT scan and its corresponding labels. A detailed description of the realized model and its first stack, its second stack, and our Modified Attention U-Net as decoders 1 and 2 is provided in the following text.

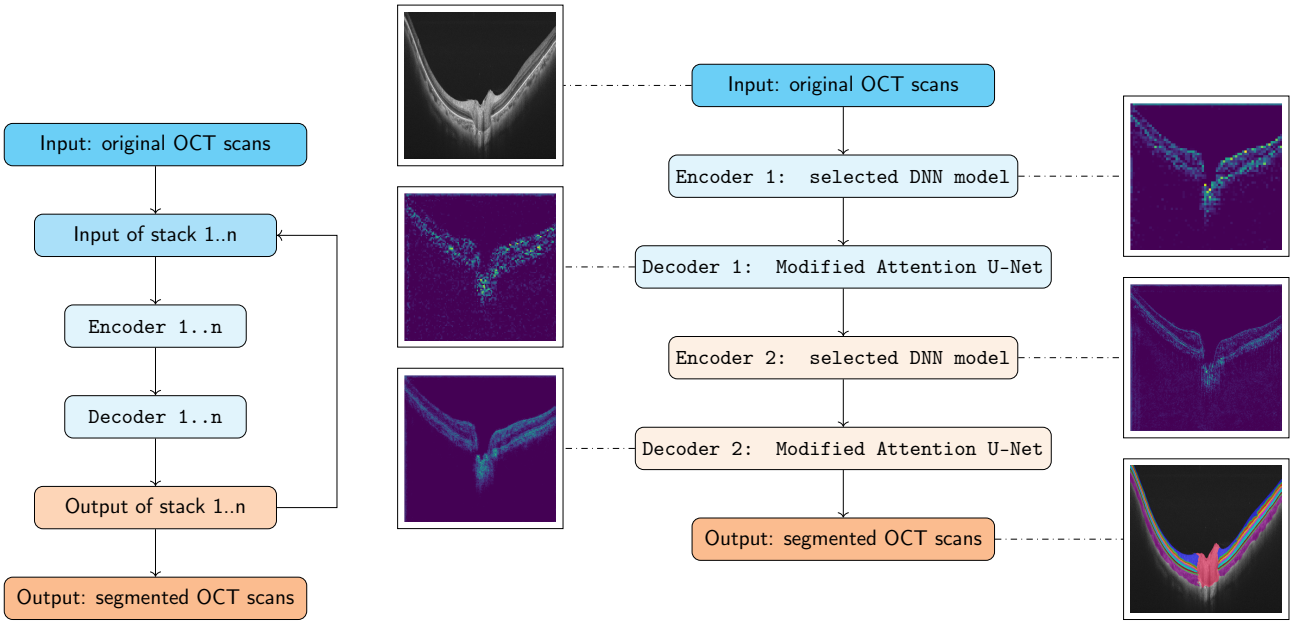
2.2.1. Stack 1: Encoder 1 + Decoder 1 (Local Context)

Within the first stack, the encoder captures contextual pieces of information of the input data in high resolution to show lower-dimensional representations. Each encoder layer possesses spatial attention, whereby the network can focus on certain areas of a feature map. The decoder operates at a higher resolution with a focus on fine-grained details. The term "local context" implies that this part of the architecture looks at smaller, more localized patterns or features in the input, for which the local context is essential for capturing details and subtle nuances in the data.

2.2.2. Stack 2: Encoder 2 + Decoder 2 (Denoiser)

Within the second stack, the architecture can naturally introduce multiscale processing by utilizing a second encoder. The second encoder allows the network to refine and further process the features from the output of the first decoder. By doing so, the model can capture higher-order interactions and details that might not be captured in the first stack. The features or outputs of decoder 1 possibly show inconsistencies, inaccuracies, or noise, and encoder 2 can be trained to correct or filter these artifacts by focusing on more robust features. The second decoder's primary purpose is to refine or "clean" the output. The decoder looks at broader patterns and features within the data. Subsequently, a larger context allows the model to understand the overall structure and global patterns, which

can help to refine or denoise the output. Denoising models remove noise or unwanted artifacts from the data. In images, this could lead to the reduction of pixel-level artifacts or the smoothing out of regions.



(**a**) General model (**b**) Model with two stacks

Figure 3. The model architecture and framework of our proposed model. (**a**) The general model used to stack multiple encoders and decoders is illustrated. (**b**) The model with two stacks of encoders and decoders is shown. For visualization purposes, cyan and orange represent the input and the output, light cyan represents the first stack with encoder 1 and decoder 1, and light orange represents the second stack with encoder 2 and decoder 2.

2.2.3. Modified Attention U-Net

There are two significant advantages to the addition of an attention mechanism before the upsampling step. Firstly, there is focused information upsampling. The attention mechanism enhances certain features and suppresses less important ones. Hence, during upsampling, only the most important information gets propagated to a higher resolution, leading to potentially more accurate reconstructions. Secondly, there is contextual awareness. The attention mechanism inherently takes a larger contextual view of the input feature map. This broad contextual awareness can guide the upsampling process by applying it before upsampling. This means that the resultant higher-resolution feature maps can be more contextually aligned with the broader features of the image data, potentially leading to better global coherence in the output.

The general architecture of U-Net [28] is symmetric and comprises two major parts. The left part is called the contracting part, which consists of the convolution process. The right part is the expansive part with transposed two-dimensional convolutional layers (upsampling). Motivated by vision transformers [57], we introduce an attention mechanism with residual learning [58,59] to facilitate advanced information fusion between feature maps in our network architecture. This module is crafted to integrate two feature maps: the primary feature map x and a secondary or skip feature map *skip*. We specifically employ concatenation as our fusion method to optimize the information flow. Given a primary feature map x with a channel size of C_x as well as a skip feature map *skip* with a channel size of C_{skip}, integration using concatenation is formalized in Equation (1). The primary and skip connection feature maps are concatenated along the channel dimension, resulting in a channel size of $C_x + C_{skip}$. Subsequently, a 1×1 convolution is employed to transform this

merged feature map back to a channel size of C_x. Concatenation-based fusion is designed to seamlessly integrate contextual information from the skip connection, enhancing the expressive power of our network.

$$x' = Convolution_{1\times1}(C_x + C_{skip} \rightarrow C_x)([x, skip]) \quad (1)$$

For implementation purposes, we used the Segmentation Models PyTorch library (Segmentation Models PyTorch library, https://smp.readthedocs.io/en/latest/index.html, accessed on 4 October 2023) [60] for U-Net while modifying its functionality by adapting it with the aforementioned attention mechanism.

2.3. Training Setup

Our training setup included the Adam optimizer [61] with a concatenating cosine annealing linear scheduler with an initial learning rate of 0.001, decaying by a factor of $0.01 \times$ *learning rate*, and a batch size of 32. For validation, we used four-fold cross-validation for our main experiments. Our models were trained for 75 epochs, whereas early stopping was introduced to prevent overfitting when no further training or validation progress could be observed within the earlier stages of the training process. As a hyperparameter tuning strategy, we utilized a random search for the selection of random combinations of hyperparameter values from pre-defined sets to evaluate our models' performance levels. This included our optimizer, its learning rate, and its batch size in order to finally obtain the selected model training parameterization. For example, for our batch size, we evaluated batch sizes of 16, 32, and 64. To fine-tune our encoder models, we additionally deployed ImageNet-based weights for the pre-training of the first encoder of the first stack.

To enable future on-site deployment of our realized system as a (semi-)automated recommender system for the OCT image segmentation of retinal layers, we evaluated our setup using general-purpose graphics processing units (GPGPU) within all of the following experiments. Our test environment was solely composed of current consumer-grade hardware. This test environment encompassed (i) our CPU, »Intel(R) Core(TM) i9-9900K CPU @ 3.60 GHz« with 7200 BogoMips and a maximum CPU load of 99%, (ii) our GPU, »TITAN RTX« with a maximum GPU load of 99%, (iii) our working memory with 128 GB of RAM, as well as (iv) our hard drive (SSD), »Samsung 970 EVO Plus SSD« with 500 GB [62].

2.4. Loss Functions

Combining different loss functions helps to capture the characteristics of the data better than using a single loss function. For all of our experiments, we computed a combined loss function composed of three popular loss functions, the Sørensen–Dice loss [63], also known as the Dice score, for the evaluation of image segmentations, the Lovász loss [64], and the Tversky loss [65]. For their use, we utilized the Segmentation Models PyTorch library [60], whereas the optimal weights for each loss function were determined via hyperparameter tuning.

2.4.1. Dice Loss

The binary Dice loss function $\mathcal{D}_{binary\ loss}$, which is often used for binary segmentation tasks, is based on the binary Dice score $\mathcal{D}_{binary\ score}$ [63]. It is defined given the ground truth and the predicted segmentations X and Y.

$$\mathcal{D}_{binary\ score} = \frac{2|X \cap Y|}{|X| + |Y|} \quad (2)$$

$$\mathcal{D}_{binary\ loss} = 1 - \mathcal{D}_{binary\ score}(X, Y) \quad (3)$$

For multiclass segmentations with C classes [66], the Dice loss $Dice_{multiclass\ loss}$ is based on the multiclass Dice score $\mathcal{D}_{multiclass\ score}$ with X_c and Y_c being the ground truth and the predicted segmentations for class c.

$$\mathcal{D}_{multiclass\ score} = \frac{1}{C}\sum_{c=1}^{C}\mathcal{D}_{binary\ score}(X_c, Y_c) \tag{4}$$

$$\mathcal{D}_{multiclass\ loss} = \frac{1}{C}\sum_{c=1}^{C}[1 - \mathcal{D}_{binary\ score}(X_c, Y_c)] \tag{5}$$

2.4.2. Lovász Loss

The Lovász–Softmax loss $\mathcal{L}$ [64] is a differentiable surrogate of the intersection over union (IoU) measure. It is defined as follows:

$$\mathcal{L} = \sum_{i=1}^{n}\alpha(i)[m(i) - \phi(i)]_+ \tag{6}$$

where:

$\phi(i)$ = predicted probability sorted in decreasing order
$m(i)$ = the corresponding ground truth label
p_i = the cumulative sum of ground truth labels up to index i
$\alpha(i) = p_i - p_{i-1}$
$[x]_+$ = denotes the positive part of x, defined as $\max(x, 0)$.

2.4.3. Tversky Loss

The Tversky loss $\mathcal{T}$ [65], a generalization of the Dice coefficient, is controlled via α and β influencing the magnitude of penalties for false positives and false negatives, respectively:

$$\mathcal{T} = 1 - \frac{|X \cap Y|}{|X \cap Y| + \alpha|X - Y| + \beta|Y - X|} \tag{7}$$

2.4.4. Combined Loss

By combining the Dice loss (Equation (5)), the Lovász loss (Equation (6)), and the Tversky loss functions (Equation (7)), we obtained our combined loss function, which was weighted as follows:

$$Combined\ loss = 0.5 \times \mathcal{D}_{multiclass\ loss} + 0.3 \times \mathcal{L} + 0.2 \times \mathcal{T} \tag{8}$$

2.5. Evaluation Principles

For the evaluation, we utilized the Dice score given our obtained OCT image segmentations [67,68] using Equation (4), which, in image segmentation, corresponds to the F1-score used for classification tasks [69,70]. Therefore, the Dice score was used to assess the alignment of our predicted segmentations with their labeled ground truth segmentations.

3. Test Results, Evaluation, and Discussion

For the test results, evaluation, and discussion of our machine-learning-based multi-stage system of stacked multiscale encoders and decoders, the following sections give an overview of our segmentation results using the introduced OCT image segmentation data sets presented in Section 2.1. For this purpose, a differentiation between the quantitative (Section 3.1) and qualitative results (Section 3.2) is made, for which we provide a discussion in the context of currently deployed comparable deep learning models.

3.1. Quantitative Results

For the evaluation, we utilized the EfficientNet (B0–B5) [71], ResNet34D and ResNet50D [58,72] models, as well as SEResNeXt50-32x4D [73,74] as encoders, whereas our Modified Attention U-Net was deployed as the decoder (decoders 1 and 2). Our models were pre-trained on ImageNet while utilizing the parameterizations given by their original authors. Table 3 gives an overview of our obtained results with one stack of our proposed system, while Table 4 gives an overview with two stacks using our deployed models. Additionally, all shown results are visualized in Figure 4 with their respective segmentation performances in terms of the Dice score. These models were utilized as they are considered to be well-established baselines for deep neural network performance as they deploy principles such as uniformly scaled network features [71], residual learning [58,59], and attention mechanisms such as squeeze-and-excitation blocks [73] while showing a wide range of applications in different medical image segmentation and classification tasks [75–78].

In addition, the following baseline deep neural network results are reported. Given one model without the explicit differentiation between encoders and decoders as well as their stacks, models such as U-Net, DRUNET, and ReLayNet allow for out-of-the-box segmentation. The previously reported scores were 80.5 ± 0.4, 80.6 ± 0.4, and $80.4 \pm 0.4\%$ for U-Net [28], DRUNET [49], and ReLayNet [25], respectively [17].

Table 3. Quantitative results for biomarker segmentation using nine different encoder–decoder stack combinations with one stack, evaluated on the peripapillary data set presented in Section 2.1. The Dice score was calculated using Equation (4). The best model is highlighted in bold. Table 4 shows our results using two stacks of our model.

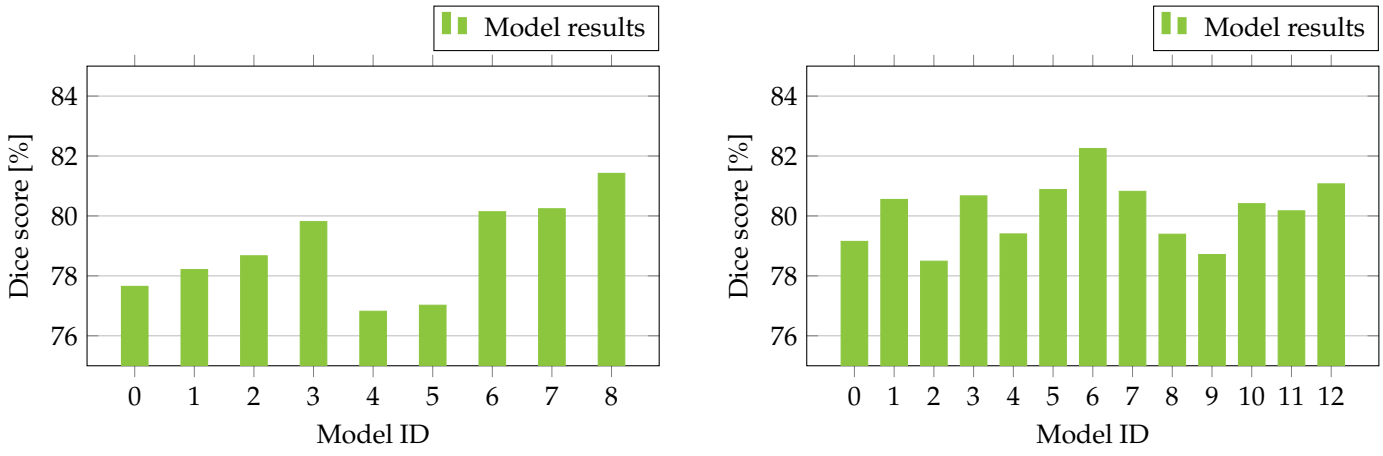
Model ID	Encoder	Decoder	Dice Score [%]
0	EfficientNet B0	Modified Attention U-Net	77.65
1	EfficientNet B1	Modified Attention U-Net	78.21
2	EfficientNet B2	Modified Attention U-Net	78.67
3	EfficientNet B3	Modified Attention U-Net	79.81
4	EfficientNet B4	Modified Attention U-Net	76.82
5	EfficientNet B5	Modified Attention U-Net	77.02
6	ResNet34D	Modified Attention U-Net	80.14
7	ResNet50D	Modified Attention U-Net	80.24
8	SEResNeXt50-32x4D	Modified Attention U-Net	**81.42**

Given our results in Table 3, using SEResNeXt50-32x4D and our Modified Attention U-Net as the encoder and decoder led to the best segmentation score when only utilizing one stack. Following this observation, the best segmentation scores were obtained from deploying SEResNeXt50-32x4D and ResNet34D as encoder 1 and encoder 2 with Modified Attention U-Net as decoders 1 and 2. With the worst-performing combination, EfficientNet B4 and EfficientNet B5 as encoder 1 and encoder 2, a range or model-based improvement of about 3.76% was obtained. Given the different obtained results, Table 5 shows a summarized comparison with the current state-of-the-art method given the data set presented by [17] with a single-stage approach. Furthermore, our results with a single encoder and decoder with attention pooling (one stack with SEResNeXt50-32x4D) and our best result with stacked encoders and decoders with attention pooling (two stacks with SEResNeXt50-32x4D and ResNet34D) are depicted. In comparison, our system resulted in an improvement in the segmentation score by 1.55% compared with the single-stage base-

line. From the single-stage baseline to one stack, an improvement of 0.72% was obtained, whereas the final improvement by adding stage 2 resulted in an additional improvement of 0.83%.

Table 4. The quantitative results of biomarker segmentation using 10 different encoder–decoder stack combinations with two stacks, evaluated on the peripapillary data set presented in Section 2.1. For the last three models, the three best models were evaluated by deploying their encoder as encoder 1 and encoder 2. The Dice score was calculated using Equation (4). The best model is highlighted in bold.

Model ID	Encoder 1	Encoder 2	Decoders 1 and 2	Dice Score [%]
0	EfficientNet B0	EfficientNet B1	Modified Attention U-Net	79.15 ± 0.48
1	EfficientNet B2	EfficientNet B3	Modified Attention U-Net	80.55 ± 0.28
2	EfficientNet B4	EfficientNet B5	Modified Attention U-Net	78.49 ± 0.31
3	EfficientNet B0	ResNet34D	Modified Attention U-Net	80.67 ± 0.26
4	EfficientNet B2	ResNet34D	Modified Attention U-Net	79.40 ± 0.19
5	ResNet50D	ResNet34D	Modified Attention U-Net	80.88 ± 0.28
6	SEResNeXt50-32x4D	ResNet34D	Modified Attention U-Net	$\mathbf{82.25 \pm 0.74}$
7	ResNet34D	SEResNeXt50-32x4D	Modified Attention U-Net	80.82 ± 0.41
8	SEResNeXt50-32x4D	EfficientNet B2	Modified Attention U-Net	79.39 ± 0.27
9	SEResNeXt50-32x4D	EfficientNet B3	Modified Attention U-Net	78.71 ± 0.34
10	ResNet34D	ResNet34D	Modified Attention U-Net	80.41 ± 0.48
11	ResNet50D	ResNet50D	Modified Attention U-Net	80.17 ± 0.19
12	SEResNeXt50-32x4D	SEResNeXt50-32x4D	Modified Attention U-Net	81.07 ± 0.42



(**a**) With one stack (**b**) With two stacks (also see Table 4)

Figure 4. Visualization of the quantitative results with (**a**) one stack and (**b**) two stacks from Tables 3 and 4.

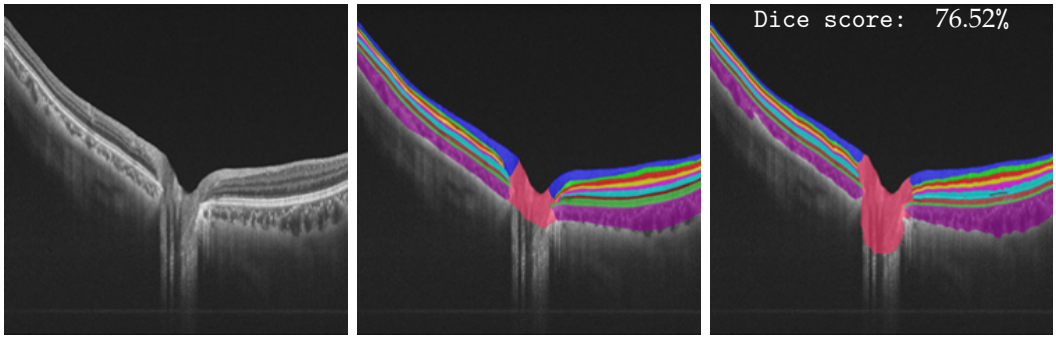
Table 5. Comparison of the current state-of-the-art method using a single-stage approach (first row) as well as our approaches with one stack with one encoder and one decoder (second row) and with two stacks of encoders and decoders (last row). The best model is highlighted in bold.

Model	Dice Score [%]
State-of-the-art model with a single stage [17]	80.70 ± 0.20
One stack with one encoder and one decoder with attention pooling (SEResNeXt50-32x4D)	81.42 ± 0.54
Two stacks of encoders and decoders with attention pooling (best model from Table 4, model ID 6: SEResNeXt50-32x4D and ResNet34D)	$\mathbf{82.25 \pm 0.74}$

From our experience, using the same encoder for different stacks leads to increased computational costs while facing problems such as vanishing gradients. Additionally, overfitting may appear in deeper models. To further investigate this observation, we added the three best-performing encoders from Table 4, deploying them as encoder 1 and encoder 2, respectively. These were ResNet34D, ResNet50D, and SEResNeXt50-32x4D. For ResNet34D and ResNet50D, no improvement could be obtained compared to our results with one stack, while SEResNeXt50-32x4D even resulted in a reduced performance in comparison, despite being the best-performing encoder in the three additional test runs.

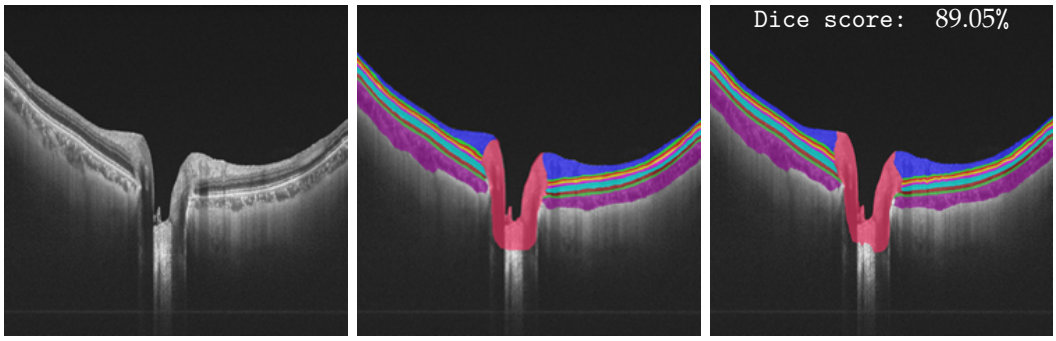
3.2. Qualitative Results

Following our quantitative results, Figures 5–8 give an overview of our qualitative results in the form of the obtained segmentation visualizations, where the different detected retinal layers are mapped on top of the original imagery. Figure 5 shows our results for the peripapillary data set, given two different segmentations with the original imagery (a,d), their ground truth annotations (b,e), and their predicted annotations (c,f). Additionally, Table 6 shows selected segmentation results in terms of the Dice score given in Figure 5c,f, whereas the class-wise score distributions are illustrated. Overall, it is observable that the ground truth annotations align closely in terms of the different layers' positions and their spatial separations. Only minor deviations, e.g., in the center of the annotations for the optic disc (c and f in comparison to b and e), are observable. This observation is also evident within the related results table, showing a segmentation score of 63.65% for Figure 5c. Besides the optic disc, GCL, IS/OS, and RPE are the more difficult retinal layers to segment.



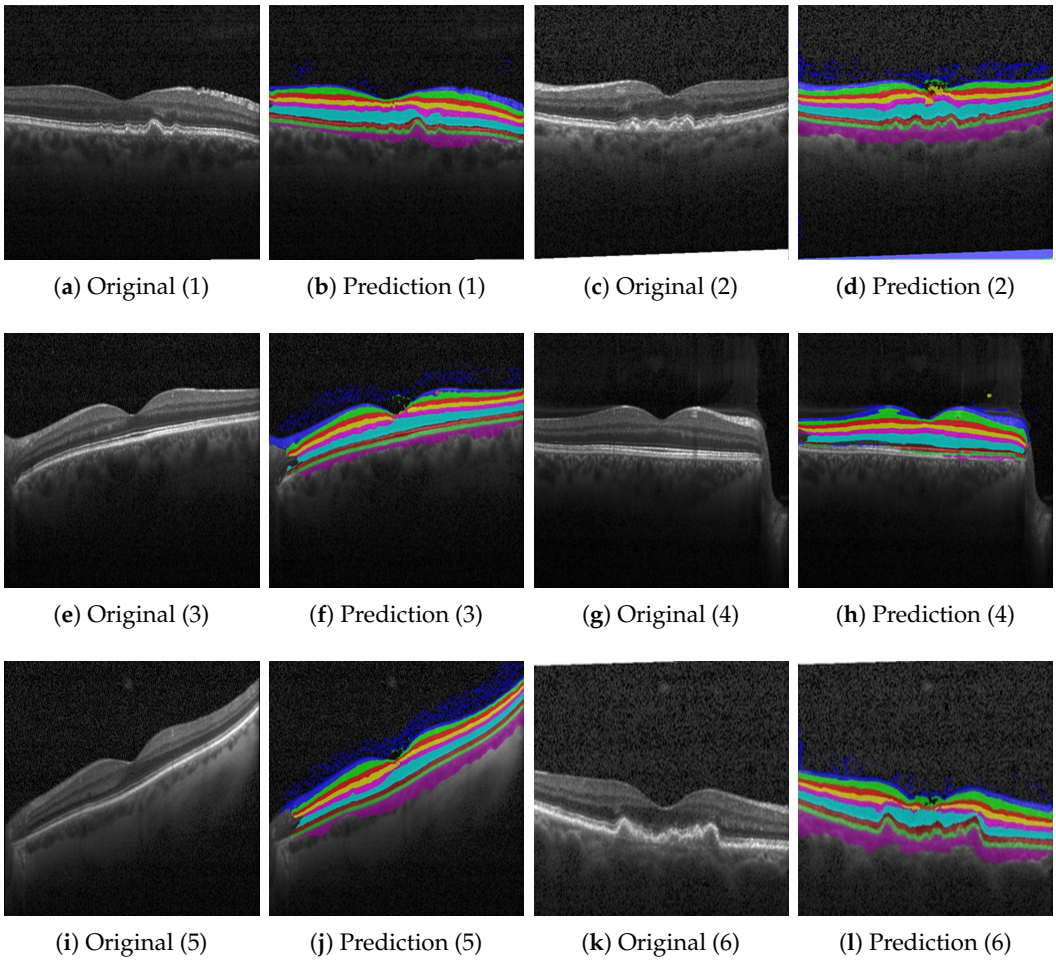
(**a**) Original (1) (**b**) Ground truth (1) (**c**) Prediction (1)

Figure 5. *Cont.*



(**d**) Original (2) (**e**) Ground truth (2) (**f**) Prediction (2)

Figure 5. Qualitative peripapillary data set with the original, ground truth segmentation, and segmentation model results as a color-coded overlay. The individual, class-wise Dice scores are shown in Table 6.

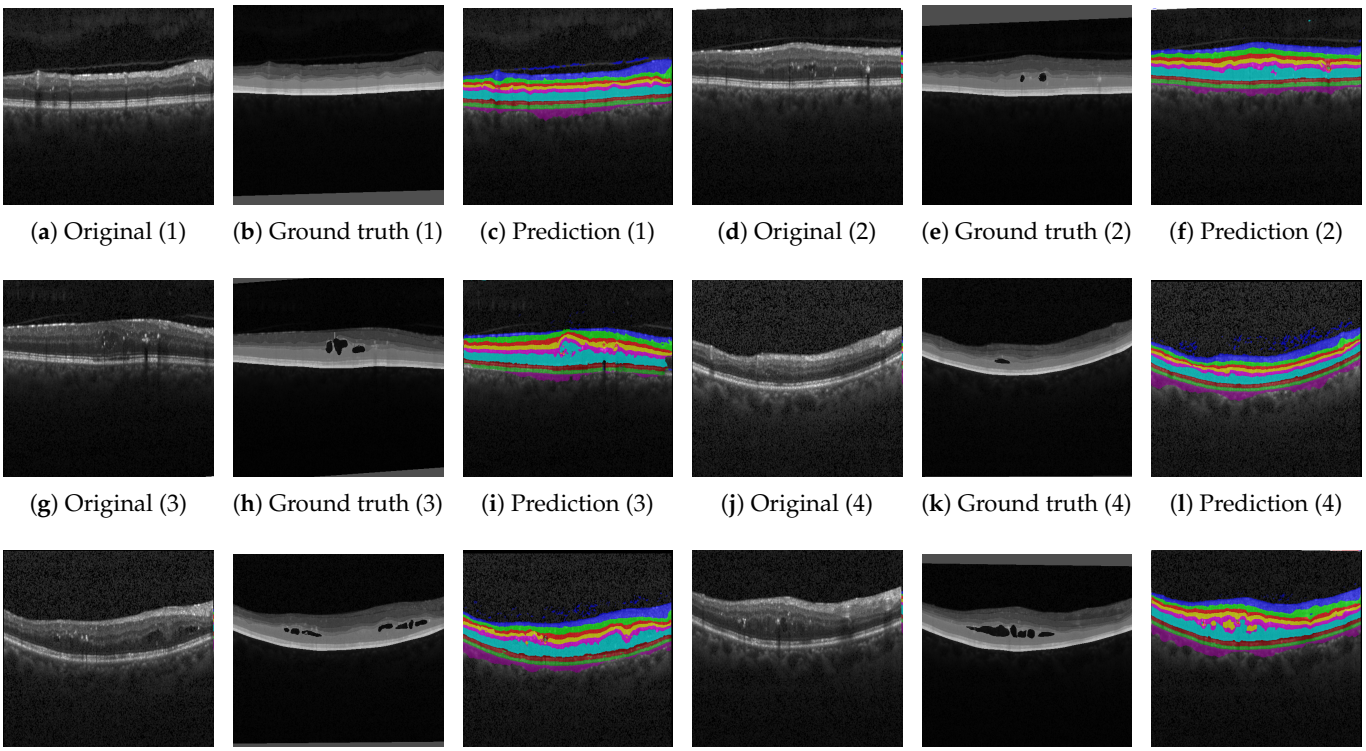


(**a**) Original (1) (**b**) Prediction (1) (**c**) Original (2) (**d**) Prediction (2)

(**e**) Original (3) (**f**) Prediction (3) (**g**) Original (4) (**h**) Prediction (4)

(**i**) Original (5) (**j**) Prediction (5) (**k**) Original (6) (**l**) Prediction (6)

Figure 6. Qualitative results based on the Heidelberg OCT data set described in Section 2.1.4. For the Heidelberg OCT data set, no annotations are available. We consider the Heidelberg data set to be a particularly noisy data set. This is apparent when comparing Figure 5 with (**c**) and (**k**): General image noise and artifacts are observed.

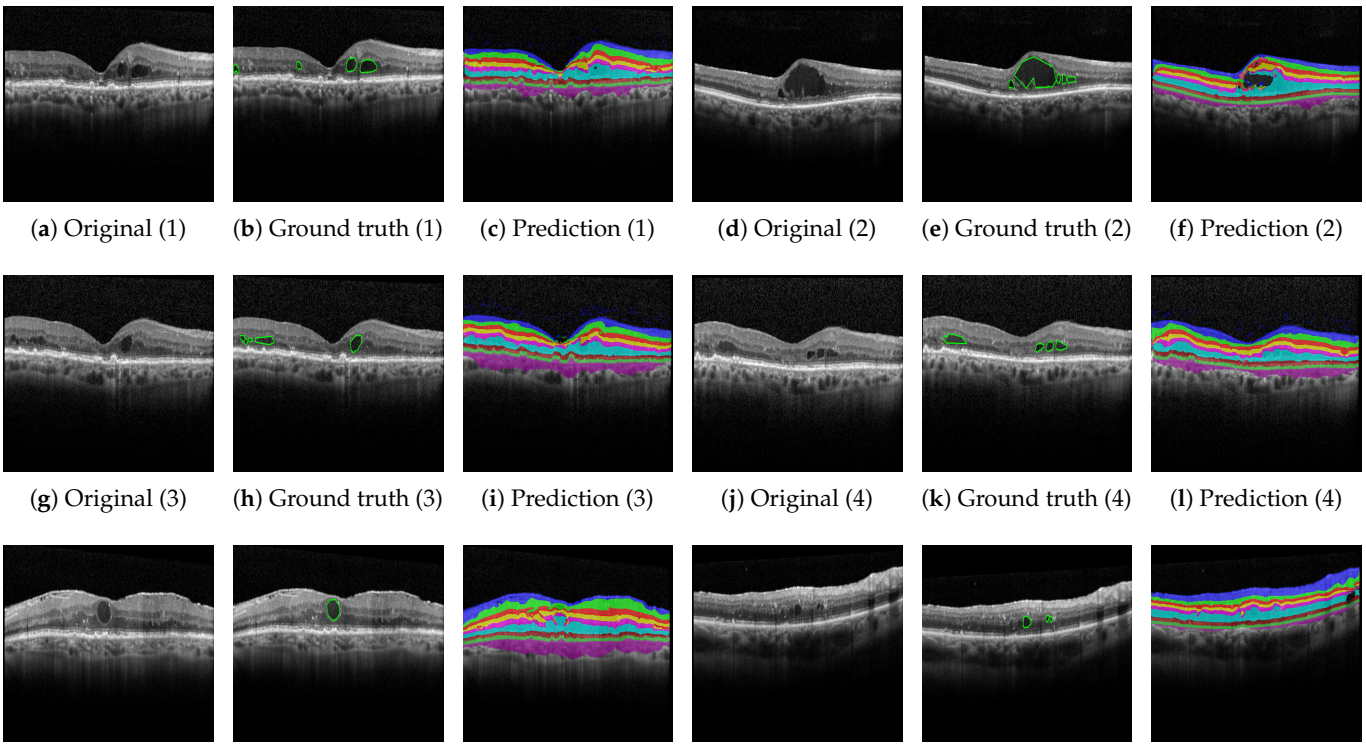


(**a**) Original (1) (**b**) Ground truth (1) (**c**) Prediction (1) (**d**) Original (2) (**e**) Ground truth (2) (**f**) Prediction (2)

(**g**) Original (3) (**h**) Ground truth (3) (**i**) Prediction (3) (**j**) Original (4) (**k**) Ground truth (4) (**l**) Prediction (4)

(**m**) Original (5) (**n**) Ground truth (5) (**o**) Prediction (5) (**p**) Original (6) (**q**) Ground truth (6) (**r**) Prediction (6)

Figure 7. Qualitative results based on the Duke SD-OCT data set described in Section 2.1.2. The ground truth comprises fluid and non-fluid manual annotations of eight boundaries.



(**a**) Original (1) (**b**) Ground truth (1) (**c**) Prediction (1) (**d**) Original (2) (**e**) Ground truth (2) (**f**) Prediction (2)

(**g**) Original (3) (**h**) Ground truth (3) (**i**) Prediction (3) (**j**) Original (4) (**k**) Ground truth (4) (**l**) Prediction (4)

(**m**) Original (5) (**n**) Ground truth (5) (**o**) Prediction (5) (**p**) Original (6) (**q**) Ground truth (6) (**r**) Prediction (6)

Figure 8. Qualitative results based on the UMN data set described in Section 2.1.3. The ground truth comprises the retinal fluid regions.

On the contrary, Figures 6–8 show results with the Heidelberg, Duke SD-OCT, and UMN data sets for six different segmentations with the original imagery, their ground truth annotations when available, and their predicted annotations. We consider the Duke SD-OCT, Heidelberg, and UMN data sets to be noisy data sets with different types and classes of retinal layers. This is especially apparent when comparing Figure 5 with Figure 6c,k: General image noise and artifacts can be observed. For our public data sets, Figure 6 (Heidelberg data set) depicts comparable visualizations of the obtained segmentations. Here, no noteworthy deviations from the observable retinal layers can be perceived. Finally, in Figure 7 (Duke SD-OCT data set) and Figure 8 (UMN data set), a comparison with the ground truth annotations of the evaluated data sets is made. While the Duke SD-OCT and UMN data sets do not provide any comprehensive annotations regarding the present retinal layers, we conclude that our trained model can be deployed instead to generate the said annotations with an improved quality for these data sets. Future contributions will have to investigate these automatically generated annotations further, for which subsequent annotations by medical doctors are planned.

In conclusion, we quantitatively evaluated our model with previously unseen data sets with different characteristics regarding the image quality as well as annotations. While we consider the Heidelberg data set to be a particularly noisy data set, it also shows different annotations with, in turn, different classes for retinal layers. Yet, qualitatively convincing results could be achieved over all additional data sets, despite fewer layers explicitly being annotated in their labeling processes. We therefore conclude that our model is able to adapt to previously unseen data, as evaluated on these data sets.

Table 6. Overview of individual, class-wise Dice score results following Figure 5. Overall there are 48 images.

Test Sample(s)	RNFL	GCL	IPL	INL	OPL	ONL	IS/OS	RPE	Choroid	Optic Disc	Overall Dice Score [%]
Figure 5c	85.05	75.21	79.12	81.73	84.13	83.63	51.89	67.84	92.91	63.65	76.52
Figure 5f	92.72	79.64	80.78	84.17	88.73	95.03	92.50	90.00	96.67	90.23	89.05
Overall	83.25	68.49	73.93	80.65	81.39	91.73	85.84	84.55	86.92	85.56	82.25

4. Conclusions and Outlook

To determine the progression of visual pathologies, optical-coherence-tomography-based retinal image is often utilized to investigate the influence of OCT biomarkers and their progression when visual pathologies such as age-related macula degeneration, cataracts, and glaucoma must be evaluated. While segmentations with improved quality might improve the treatment of patients with eye diseases, the realization of a (semi-)automated system could also provide the benefits of reduced treatment times and treatment costs, especially in areas with otherwise less or no access to medical aid in general.

As a couple of approaches to the task of OCT image segmentation of retinal layers already exist, fundamental approaches from machine learning and deep learning utilizing deep neural networks have shown the first promising results. In our contribution, we propose the application of a machine-learning-based multistage system of stacked multiscale encoders and decoders for the task of image segmentation. While this approach is able to leverage already-existing state-of-the-art deep neural networks in terms of their image segmentation performance, it also furthers their motivation by only combining models for the encoding and decoding stages that elevate the resulting image segmentation scores.

Given the evaluated peripapillary OCT data set, we obtained a final segmentation performance of up to $82.25 \pm 0.74\%$ in terms of the Sørensen–Dice coefficient, outperforming the current best single-stage model by 1.55% with a score of $80.70 \pm 0.20\%$. We conclude that, by adding additional stages to the image segmentation process, improved segmentation scores can be reached. With our trained model, the Duke SD-OCT, Heidelberg, and

UMN data sets were further investigated, showing that even noisy and previously unseen data sets' retinal layers can be (semi-)automatically segmented to complement available annotations with as yet non-contained OCT biomarker segmentations.

Future, as well as more application-focused contributions, could be used to extend our approach by assessing not only additional segmentation imagery but also the following OCT image classification stage. This includes further OCT biomarkers such as disorganizations of the retinal inner layers (DRIL) between GCL, IPL, INL, and OPL [79] as well as hyper-reflective foci (HRF). To explore novel approaches to image segmentation and classification, different transformer-based models [57] with improved prediction scores, encompassing, i.a., data-efficient image transformers (DeiTs) as well as hybrid transformers [80–82], will be further assessed. Additionally, further strategies regarding data augmentation will have to be investigated. Finally, a (semi-)automated recommender system can be realized to enable the future on-site deployment of our realized system.

Author Contributions: A.S.K., T.S. and D.K. conducted this contribution's writing process and the related research project's implementation and evaluation with the help of H.L. and M.R., who revised this manuscript. A.S.K., T.S. and D.K. designed this contribution, whereas M.R. and D.K. designed and supervised the related research project. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the European Social Fund for Germany as well as the Federal Ministry of Education and Research, namely the Medical Informatics Hub in Saxony (MiHUBx) under grant number 01ZZ2101C.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: We gratefully thank the authors of Li et al. [17] for providing us with their peripapillary OCT data set. The Duke SD-OCT [14], UMN [54], and Heidelberg [55] data sets are publically available.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AMD	Age-related macula degeneration
DNN	Deep neural network
GCL	Ganglion cell layer
INL	Inner nuclear layer
IPL	Inner plexiform layer
IRF	Intraretinal fluid
IS/OS	Inner/outer photoreceptor segment
ONL	Outer nuclear layer
OPL	Outer plexiform layer
PED	Pigment epithelial detachment
RNFL	Retinal nerve fiber layer
RNN	Recurrent neural network
RPE	Retinal pigment epithelium
SRF	Subretinal fluid

References

1. Anders, J.; Dapp, U.; Laub, S.; Renteln-Kruse, W. Impact of fall risk and fear of falling on mobility of independently living senior citizens transitioning to frailty: Screening results concerning fall prevention in the community. *Z. Gerontol. Geriatr.* **2007**, *40*, 255–267. [CrossRef]
2. E, J.Y.; Li, T.; Mclnally, L.; Thomson, K.; Shahani, U.; Gray, L.; Howe, T.; Skelton, D. Environmental and behavioural interventions for reducing physical activity limitation and preventing falls in older people with visual impairment. *Cochrane Database Syst. Rev.* **2020**, *9*, CD009233. [CrossRef]
3. Pascolini, D.; Mariotti, S. Global Estimates of Visual Impairment: 2010. *Br. J. Ophthalmol.* **2011**, *96*, 614–618. [CrossRef]
4. Reitmeir, P.; Linkohr, B.; Heier, M.; Molnos, S.; Strobl, R.; Schulz, H.; Breier, M.; Faus, T.; Küster, D.M.; Wulff, A.; et al. Common eye diseases in older adults of southern Germany: Results from the KORA-Age study. *Age Ageing* **2017**, *46*, 481–486. [CrossRef] [PubMed]
5. Finger, R.P.; Fimmers, R.; Holz, F.G.; Scholl, H.P. Incidence of blindness and severe visual impairment in Germany: Projections for 2030. *Investig. Ophthalmol. Vis. Sci.* **2011**, *52*, 4381–4389. [CrossRef] [PubMed]
6. Kansal, V.; Armstrong, J.J.; Pintwala, R.; Hutnik, C. Optical coherence tomography for glaucoma diagnosis: An evidence based meta-analysis. *PLoS ONE* **2018**, *13*, e0190621. [CrossRef]
7. Aumann, S.; Donner, S.; Fischer, J.; Müller, F. Optical coherence tomography (OCT): Principle and technical realization. In *High Resolution Imaging in Microscopy and Ophthalmology: New Frontiers in Biomedical Optics*; Springer: Berlin, Germany, 2019; pp. 59–85.
8. Shu, X.; Beckmann, L.; Zhang, H.F. Visible-light optical coherence tomography: A review. *J. Biomed. Opt.* **2017**, *22*, 121707. [CrossRef]
9. Maldonado, R.S.; Toth, C.A. Optical coherence tomography in retinopathy of prematurity: Looking beyond the vessels. *Clin. Perinatol.* **2013**, *40*, 271–296. [CrossRef]
10. Iqbal, S.; Qureshi, A.; Li, J.; Mahmood, T. On the Analyses of Medical Images Using Traditional Machine Learning Techniques and Convolutional Neural Networks. *Arch. Comput. Methods Eng.* **2023**, *30*, 3173–3233. [CrossRef] [PubMed]
11. Cardenas, C.E.; Yang, J.; Anderson, B.M.; Court, L.E.; Brock, K.B. Advances in Auto-Segmentation. *Semin. Radiat. Oncol.* **2019**, *29*, 185–197. [CrossRef]
12. Schlosser, T.; Beuth, F.; Meyer, T.; Kumar, A.S.; Stolze, G.; Furashova, O.; Engelmann, K.; Kowerko, D. Visual Acuity Prediction on Real-Life Patient Data Using a Machine Learning Based Multistage System. *arXiv* **2022**, arXiv:2204.11970.
13. Garvin, M.K.; Abramoff, M.D.; Wu, X.; Russell, S.R.; Burns, T.L.; Sonka, M. Automated 3-D Intraretinal Layer Segmentation of Macular Spectral-Domain Optical Coherence Tomography Images. *IEEE Trans. Med. Imaging* **2009**, *28*, 1436–1447. [CrossRef]
14. Chiu, S.J.; Allingham, M.J.; Mettu, P.S.; Cousins, S.W.; Izatt, J.A.; Farsiu, S. Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema. *Biomed. Opt. Express* **2015**, *6*, 1172–1194. [CrossRef]
15. Fang, L.; Cunefare, D.; Wang, C.; Guymer, R.H.; Li, S.; Farsiu, S. Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search. *Biomed. Opt. Express* **2017**, *8*, 2732–2744. [CrossRef] [PubMed]
16. Elgafi, M.; Sharafeldeen, A.; Elnakib, A.; Elgarayhi, A.; Alghamdi, N.S.; Sallah, M.; El-Baz, A. Detection of Diabetic Retinopathy Using Extracted 3D Features from OCT Images. *Sensors* **2022**, *22*, 7833. [CrossRef]
17. Li, J.; Jin, P.; Zhu, J.; Zou, H.; Xu, X.; Tang, M.; Zhou, M.; Gan, Y.; He, J.; Ling, Y.; et al. Multi-scale GCN-assisted two-stage network for joint segmentation of retinal layers and discs in peripapillary OCT images. *Biomed. Opt. Express* **2021**, *12*, 2204–2220. [CrossRef]
18. Everingham, M.; Eslami, S.A.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136. [CrossRef]
19. Brostow, G.J.; Fauqueur, J.; Cipolla, R. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognit. Lett.* **2009**, *30*, 88–97. [CrossRef]
20. Song, S.; Lichtenberg, S.P.; Xiao, J. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015. [CrossRef]
21. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495.
22. Shotton, J.; Johnson, M.; Cipolla, R. Semantic texton forests for image categorization and segmentation. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8. [CrossRef]
23. Leon, F.; Floria, S.A.; Badica, C. Evaluating the effect of voting methods on ensemble-based classification. In Proceedings of the 2017 IEEE International Conference on INnovations in Intelligent Systems and Applications (INISTA), Gdynia, Poland, 3–5 July 2017; pp. 1–6. [CrossRef]
24. Sturgess, P.; Alahari, K.; Ladicky, L.; Torr, P. Combining Appearance and Structure from Motion Features for Road Scene Understanding. In Proceedings of the BMVC-British Machine Vision Conference, London, UK, 7–10 September 2009. [CrossRef]
25. Roy, A.G.; Conjeti, S.; Karri, S.P.K.; Sheet, D.; Katouzian, A.; Wachinger, C.; Navab, N. ReLayNet: Retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks. *Biomed. Opt. Express* **2017**, *8*, 3627–3642. [CrossRef] [PubMed]

26. Kiaee, F.; Fahimi, H.; Rabbani, H. Intra-Retinal Layer Segmentation of Optical Coherence Tomography Using 3D Fully Convolutional Networks. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 2795–2799. [CrossRef]
27. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
28. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015.
29. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R.B. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
30. Lin, T.; Maire, M.; Belongie, S.J.; Bourdev, L.D.; Girshick, R.B.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014.
31. Pérez-Nicolás, M.; Colinas-León, T.; Alia-Tejacal, I.; Peña-Ortega, G.; González-Andrés, F.; Beltrán-Rodríguez, L. Morphological Variation in Scarlet Plume (*Euphorbia fulgens* Karw ex Klotzsch, Euphorbiaceae), an Underutilized Ornamental Resource of Mexico with Global Importance. *Plants* **2021**, *10*, 2020. [CrossRef]
32. Liu, W.; Sun, Y.; Ji, Q. MDAN-UNet: Multi-Scale and Dual Attention Enhanced Nested U-Net Architecture for Segmentation of Optical Coherence Tomography Images. *Algorithms* **2020**, *13*, 60. [CrossRef]
33. Lin, T.; Dollár, P.; Girshick, R.B.; He, K.; Hariharan, B.; Belongie, S.J. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
34. Orlando, J.I.; Seeböck, P.; Bogunović, H.; Klimscha, S.; Grechenig, C.; Waldstein, S.; Gerendas, B.S.; Schmidt-Erfurth, U. U2-net: A bayesian u-net model with epistemic uncertainty feedback for photoreceptor layer segmentation in pathological oct scans. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019; IEEE: Piscatway, NJ, USA, 2019; pp. 1441–1445.
35. Farshad, A.; Yeganeh, Y.; Gehlbach, P.; Navab, N. Y-Net: A Spatiospectral Dual-Encoder Networkfor Medical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Singapore, 18–22 September 2022. [CrossRef]
36. Li, Y.; Qi, H.; Dai, J.; Ji, X.; Wei, Y. Fully Convolutional Instance-aware Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
37. Tran, A.; Weiss, J.; Albarqouni, S.; Faghi Roohi, S.; Navab, N. Retinal Layer Segmentation Reformulated as OCT Language Processing. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, 4–8 October 2020; Proceedings, Part V; Springer: Berlin/Heidelberg, Germany, 2020; pp. 694–703. [CrossRef]
38. Kugelman, J.; Alonso-Caneiro, D.; Read, S.; Vincent, S.; Collins, M. Automatic segmentation of OCT retinal boundaries using recurrent neural networks and graph search. *Biomed. Opt. Express* **2018**, *9*, 5759. [CrossRef]
39. Meester, R.; Doubeni, C.; Zauber, A.; Goede, S.; Levin, T.; Corley, D.; Jemal, A.; Lansdorp-Vogelaar, I. 969 Public Health Impact of Achieving 80% Colorectal Cancer Screening RATES in the United States by 2018. *Cancer* **2015**, *81*, AB181–AB182. . [CrossRef]
40. Mohammed, A.; Yildirim, S.; Farup, I.; Pedersen, M.; Hovde, Ø. Y-net: A deep convolutional neural network for polyp detection. arXiv **2018**, arXiv:1806.01907.
41. Chen, S.; Guo, W. Auto-Encoders in Deep Learning—A Review with New Perspectives. *Mathematics* **2023**, *11*, 1777. [CrossRef]
42. Lin, G.; Shen, C.; Reid, I.D.; van den Hengel, A. Efficient piecewise training of deep structured models for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
43. Noh, H.; Hong, S.; Han, B. Learning Deconvolution Network for Semantic Segmentation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
44. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, 40, 834–848. [CrossRef]
45. Martí, M.; Maki, A. A multitask deep learning model for real-time deployment in embedded systems. *arXiv* **2017**, arXiv:1711.00146.
46. Hariharan, B.; Arbeláez, P.A.; Girshick, R.B.; Malik, J. Hypercolumns for Object Segmentation and Fine-grained Localization. *arXiv* **2014**, arXiv:1411.5752.
47. Mostajabi, M.; Yadollahpour, P.; Shakhnarovich, G. Feedforward semantic segmentation with zoom-out features. *arXiv* **2014**, arXiv:1412.0774.
48. Zang, P.; Wang, J.; Hormel, T.T.; Liu, L.; Huang, D.; Jia, Y. Automated segmentation of peripapillary retinal boundaries in OCT combining a convolutional neural network and a multi-weights graph search. *Biomed. Opt. Express* **2019**, *10*, 4340–4352. [CrossRef]
49. Devalla, S.K.; Renukanand, P.K.; Sreedhar, B.K.; Subramanian, G.; Zhang, L.; Perera, S.; Mari, J.M.; Chin, K.S.; Tun, T.A.; Strouthidis, N.G.; et al. DRUNET: A dilated-residual U-Net deep learning network to segment optic nerve head tissues in optical coherence tomography images. *Biomed. Opt. Express* **2018**, *9*, 3244–3265. [CrossRef]
50. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.

51. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*; MIT Press: Boston, MA, USA, 2014; Volume 27.
52. You, A.; Kim, J.K.; Ryu, I.H.; Yoo, T.K. Application of generative adversarial networks (GAN) for ophthalmology image domains: A survey. *Eye Vis.* **2022**, *9*, 6. [CrossRef] [PubMed]
53. Burlina, P.M.; Joshi, N.; Pacheco, K.D.; Liu, T.A.; Bressler, N.M. Assessment of deep generative models for high-resolution synthetic retinal image generation of age-related macular degeneration. *JAMA Ophthalmol.* **2019**, *137*, 258–264. [CrossRef] [PubMed]
54. Rashno, A.; Nazari, B.; Koozekanani, D.D.; Drayna, P.M.; Sadri, S.; Rabbani, H.; Parhi, K.K. Fully-automated segmentation of fluid regions in exudative age-related macular degeneration subjects: Kernel graph cut in neutrosophic domain. *PLoS ONE* **2017**, *12*, e0186949. [CrossRef]
55. Kermany, D.S.; Goldbaum, M.; Cai, W.; Valentim, C.C.; Liang, H.; Baxter, S.L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell* **2018**, *172*, 1122–1131.e9. [CrossRef]
56. Yushkevich, P.A.; Piven, J.; Hazlett, H.C.; Smith, R.G.; Ho, S.; Gee, J.C.; Gerig, G. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage* **2006**, *31*, 1116–1128. [CrossRef]
57. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 $\times$ 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
58. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
59. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity Mappings in Deep Residual Networks. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 630–645. [CrossRef]
60. Iakubovskii, P. Segmentation Models Pytorch. 2019. Available online: https://github.com/qubvel/segmentation_models.pytorch (accessed on 30 August 2023).
61. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
62. Schlosser, T.; Friedrich, M.; Beuth, F.; Kowerko, D. Improving automated visual fault inspection for semiconductor manufacturing using a hybrid multistage system of deep neural networks. *J. Intell. Manuf.* **2022**, *33*, 1099–1123. [CrossRef]
63. Sudre, C.H.; Li, W.; Vercauteren, T.; Ourselin, S.; Cardoso, M.J. Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. *arXiv* **2017**, arXiv:1707.03237.
64. Berman, M.; Blaschko, M.B. Optimization of the Jaccard index for image segmentation with the Lovász hinge. *arXiv* **2017**, arXiv:1705.08790.
65. Salehi, S.S.M.; Erdogmus, D.; Gholipour, A. Tversky loss function for image segmentation using 3D fully convolutional deep networks. *arXiv* **2017**, arXiv:1706.05721.
66. Liu, Q.; Tang, X.; Guo, D.; Qin, Y.; Jia, P.; Zhan, Y.; Zhou, X.; Wu, D. Multi-class gradient harmonized dice loss with application to knee MR image segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2019: 22nd International Conference, Shenzhen, China, 13–17 October 2019; Proceedings, Part VI 22; Springer: Berlin, Germany, 2019; pp. 86–94.
67. Dice, L.R. Measures of the Amount of Ecologic Association Between Species. *Ecology* **1945**, *26*, 297–302. [CrossRef]
68. Sorensen, T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biol. Skr.* **1948**, *5*, 1–34.
69. Van Rijsbergen, C.J. *The Geometry of Information Retrieval*; Cambridge University Press: Cambridge, UK, 2004.
70. Taha, A.A.; Hanbury, A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Med. Imaging* **2015**, *15*, 29. [CrossRef]
71. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
72. He, T.; Zhang, Z.; Zhang, H.; Zhang, Z.; Xie, J.; Li, M. Bag of tricks for image classification with convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 558–567.
73. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
74. Wightman, R. PyTorch Image Models. 2019. Available online: https://github.com/rwightman/pytorch-image-models (accessed on 30 August 2023).
75. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, 20 September 2018; Proceedings 4; Springer: Berlin, Germany, 2018; pp. 3–11.
76. Xia, K.j.; Yin, H.s.; Zhang, Y.d. Deep semantic segmentation of kidney and space-occupying lesion area based on SCNN and ResNet models combined with SIFT-flow algorithm. *J. Med. Syst.* **2019**, *43*, 2. [CrossRef] [PubMed]

77. Zhu, Z.; Wang, H.; Zhao, T.; Guo, Y.; Xu, Z.; Liu, Z.; Liu, S.; Lan, X.; Sun, X.; Feng, M. Classification of cardiac abnormalities from ECG signals using SE-ResNet. In Proceedings of the 2020 Computing in Cardiology, Rimini, Italy, 13–16 September 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–4. [CrossRef]
78. Abedalla, A.; Abdullah, M.; Al-Ayyoub, M.; Benkhelifa, E. Chest X-ray pneumothorax segmentation using U-Net with EfficientNet and ResNet architectures. *PeerJ Comput. Sci.* **2021**, *7*, e607. [CrossRef] [PubMed]
79. Midena, E.; Torresin, T.; Schiavon, S.; Danieli, L.; Polo, C.; Pilotto, E.; Midena, G.; Frizziero, L. The Disorganization of Retinal Inner Layers Is Correlated to Müller Cells Impairment in Diabetic Macular Edema: An Imaging and Omics Study. *Int. J. Mol. Sci.* **2023**, *24*, 9607. [CrossRef]
80. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 10347–10357.
81. Touvron, H.; Cord, M.; Sablayrolles, A.; Synnaeve, G.; Jégou, H. Going deeper with image transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 32–42.
82. Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; Wang, Y. Transformer in transformer. In *Advances in Neural Information Processing Systems*; MIT Press: Boston, MA, USA, 2021; Volume 34, pp. 15908–15919.