applied sciences

MDPI

Article

BEAC-Net: Boundary-Enhanced Adaptive Context Network for Optic Disk and Optic Cup Segmentation

Lincen Jiang [1,2], Xiaoyu Tang [1], Shuai You [1], Shangdong Liu [1,3] and Yimu Ji [1,*]

[1] School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210023, China; 2022010301@njupt.edu.cn (L.J.); 1022041009@njupt.edu.cn (X.T.); 2021070708@njupt.edu.cn (S.Y.); lsd@njupt.edu.cn (S.L.)

[2] School of Computer and Software, Nanjing Vocational University of Industry Technology, Nanjing 210023, China

[3] Nanjing Yunzhi Data Technology Co., Ltd., Nanjing 210012, China

[*] Correspondence: jiym@njupt.edu.cn

Abstract: Accurately segmenting the optic disk (OD) and optic cup (OC) on retinal fundus images is important for treating glaucoma. With the development of deep learning, some CNN-based methods have been implemented to segment OD and OC, but it is difficult to accurately segment OD and OC boundaries affected by blood vessels and the lesion area. To this end, we propose a novel boundary-enhanced adaptive context network (BEAC-Net) for OD and OC segmentation. Firstly, a newly designed efficient boundary pixel attention (EBPA) module enhances pixel-by-pixel feature capture to collect the boundary contextual information of OD and OC in the horizontal and vertical directions. In addition, background noise makes segmenting boundary pixels difficult. To this end, an adaptive context module (ACM) was designed, which simultaneously learns local-range and long-range information to capture richer context. Finally, BEAC-Net adaptively integrates the feature maps from different levels using the attentional feature fusion (AFF) module. In addition, we provide a high-quality retinal fundus image dataset named the 66 Vision-Tech dataset, which advances the field of diagnostic glaucoma. Our proposed BEAC-Net was used to perform extensive experiments on the RIM-ONE-v3, DRISHTI-GS, and 66 Vision-Tech datasets. In particular, BEAC-Net achieved a Dice coefficient of 0.8267 and an IoU of 0.8138 for OD segmentation and a Dice coefficient of 0.8057 and an IoU value of 0.7858 for OC segmentation on the 66 Vision-Tech dataset, achieving state-of-the-art segmentation results.

Keywords: glaucoma; boundary enhanced; adaptive context; OD and OC segmentation

check for updates

Citation: Jiang, L.; Tang, X.; You, S.; Liu, S.; Ji, Y. BEAC-Net: Boundary-Enhanced Adaptive Context Network for Optic Disk and Optic Cup Segmentation. *Appl. Sci.* **2023**, *13*, 10244. https://doi.org/10.3390/app131810244

Academic Editors: Min Xia and Kai Hu

Received: 26 July 2023
Revised: 24 August 2023
Accepted: 29 August 2023
Published: 12 September 2023

1. Introduction

Glaucoma is a blinding eye disease that is very dangerous and eventually leads to blindness. It can cause irreversible damage to vision if not diagnosed and treated in time. Ophthalmologists make a clinical diagnosis of glaucoma using the CDR (cup–disk ratio, OC/OD) indicator, which helps detect and diagnose glaucoma at an early stage. Existing research has shown that glaucoma can generally be considered when the CDR is larger than 0.65 [1,2]. As a result, the accurate segmentation of the optic disk (OD) and optic cup (OC) in fundus images can provide quantitative assessment and diagnostic assistance to physicians in screening and treating glaucoma. Figure 1 shows the average unaffected eyes and the eyes of glaucoma patients. The accurate segmentation of OD and OC is important to calculate the CDR indicator. Doctors face a large number of fundus images of patients every day. Manual segmentation takes a lot of time and effort, and it requires a high level of professionalism. Meanwhile, manual segmentation is inefficient and more subjective, which leads to inaccuracy in segmented boundaries. Therefore, for ophthalmologists, segmenting the OD and OC accurately from fundus images is an important task.

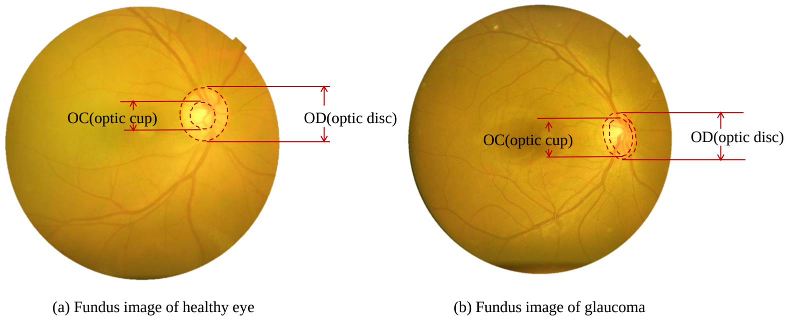


Figure 1. The comparison of healthy eye and glaucoma.

Deep learning has made great progress in the field of medical image segmentation. For instance, a series based on convolutional neural networks (CNN) [3–6] has been developed for automatic feature extraction from medical images, so applying it to OD and OC segmentation has become a major research direction and outperforms traditional methods in terms of segmentation effects. Deep-learning-based segmentation methods are based on the fully convolutional network (FCN) [7]. The widely used U-Net network [8] has become the main neural network architecture for biomedical image segmentation tasks due to its multi-scale skip connections and learnable deconvolution layers [9–11]. M-Net, based on U-Net, introduces the idea of deep supervision and adds a loss function in the middle layer to successfully achieve the joint segmentation of OD and OC [12]. CDED-Net uses a tightly connected network of OD and OC decoders to achieve better results in the joint segmentation of OD [13]. CCNet provides a more efficient way of capturing contextual information, using a self-attentive mechanism to make any location in the feature graph perceive feature information at all locations [14]. The advantages of the DeepLabv3+ [15] model are that it explores multi-scale contextual information and achieves accurate segmentation by applying multi-sample rate dilation convolution, multi-receiver field convolution, or pooling on the input feature. However, due to the complexity of fundus images, most existing methods used to segment medical images are CNN-based. However, the segment results are often unsatisfactory because object boundaries are inaccurate. These problems are caused by the insufficiently detailed features acquired after the deep convolution operation, especially the limited contextual information of OC.

To solve the above problem, the Swin Transformer uses a technique called "self-attentiveness" to automatically learn semantic features in fundus images and overcomes the shortcomings of the Vision Transformer (Vit) [16]. The Swin Transformer improves feature extraction to eliminate and suppress interference from parts such as blood vessels and bright lesions in the OD region of the original image.

In this paper, we proposed using a novel boundary-enhanced adaptive context network (BEAC-Net), a pure Transformer network, for optic disk and optic cup segmentation. BEAC-Net can automatically segment OD and OC in retinal fundus images and consists of the classical encoder–decoder structure. We verify the efficacy of BEAC-Net on two public fundus image datasets and introduce the 66 Vision-Tech dataset to test the generalization property of the model.

To this end, the contributions of this paper are four-fold:

1. We propose using a novel BEAC-Net network for OD and OC segmentation based on the ACM (adaptive context module), which eliminates noise in critical pixels when fully effective.
2. We design an EBPA (efficient boundary pixel attention) module that can impose boundary awareness on the proposed network via cumulative contextual information in the horizontal and vertical directions to enhance pixel-by-pixel feature capture.
3. We construct an AFF (attentional feature fusion) module to integrate the feature maps from high-level and low-level features adaptively. Multi-scale hierarchical feature extraction avoids an excessive loss of key information in the original images.

4. We provide a high-quality retinal fundus image dataset named the 66 Vision-Tech dataset. The fundus images are from 66 VISION TECH Co., Ltd., No. 9 Jinfeng Road, High-tech District, Suzhou 215163, China.
   The experiment results demonstrate the good generalization properties of the model.

2. Related Work

The traditional methods used to segment the optic disk and optic cup are mainly based on manual feature extraction. Aquino et al. and Lu et al. [17,18] used a circular transform-based method to segment OD. Sukanya et al. [19] used super pixel point classification to segment OD and OC, which transformed the problem of locating the OC boundary to a pixel-by-pixel classification problem. Cheng et al. [20] attempted to segment the optic disk using a weakly supervised approach. Compared with optic disk segmentation, the low contrast ratio causes optic cup segmentation to be more difficult. The segmentation accuracy of traditional segment methods is determined by the image quality and the depth of the manually extracted image features, and when effective features cannot be extracted, the segmentation accuracy will be significantly reduced.

With the continuous development and popularity of deep-learning technology, Feng et al. [21] discovered that the application of convolutional neural networks (CNNs) to extract image features achieved better results than traditional algorithms on various medical image segmentation tasks. FCN is the basis of many current semantic segmentation methods and has achieved better segmentation results than traditional methods on natural image segmentation tasks. Therefore, deep-learning techniques have been introduced into medical image processing, and more and more studies have been conducted using convolutional neural networks. OD and OC segmentation networks based on deep learning have been proposed one after another and achieved better segmentation results than traditional segmentation methods. M-Net (multi-label deep network) added multi-scales based on U-Net, introduced the idea of depth supervision, and added additional loss functions in the middle interlayer with an additional loss function. The introduction of the polar coordinate transformation operation successfully realizes the simultaneous segmentation of the OD and OC. Huang et al. [22] proposed DenseNet which applied the FCN to the OD and OC segmentation tasks. However, most existing CNN-based methods cannot obtain detailed boundary features due to multiple pooling and downsampling, resulting in the ambiguity of segmentation boundaries. Gu et al. [23] proposed CENet (context encoder network), which realized a context-encoding module consisting of a residual multi-path pooling module and a multi-scale dense dilated convolution module. CENet can capture features with high-level semantic information at multi-scales, but it is not used in the segmentation of OC. CDED-Net adopted a tightly connected optic cup–optic disk decoder network structure and achieves high performance in the joint segmentation of cups and disks, but extensive networks for segmentation tasks increase the number of required parameters. Wang et al. [24] proposed a feature-embedding framework that effectively improves the generalization ability of convolutional neural networks in completing the OD and OC segmentation tasks. To summarize the above, these manners enhance the information extraction ability of the network by improving the structure, but the utilization of multi-level features in the middle of the network is not sufficient.

DeepLabv3+ [15] used the convolution of upsampling filters to extract dense feature maps and capture a long-distance context, which is mainly divided into encoder and decoder parts. Xception is used as the backbone network in this paper. Then, the ASPP structure is used to solve the multi-scale problem. The decoder part is introduced to combine the underlying features with high-level feature fusion and obtain a notably high segmentation boundary accuracy. Despite these improvements in the segmentation network, DeepLabv3+ has more parameters, which may lead to long training times.

Transformers in the field of computer vision have taken shape, and the Transformer architecture model has achieved significant results in the field of CV for its self-attention feature. The Transformer can pay attention to information in different subspaces and capture richer feature information. However, the computation of the Transformer is too large. Therefore, Liu et al. [25] proposed using the Swin Transformer. Different from the previous ViT, the Swin Transformer made two improvements: (1) The Swin Transformer can obtain the global attention capability via the W-MSA and SW-MSA operations. (2) Reducing the computation from a squared relationship of image size to a linear relationship greatly reduces the number of operations and increases the speed of model inference. Determining how to fuse CNN and powerful ViT to achieve better segmentation effects has become a research hotspot. CNN can extract local detail information using hierarchical feature representation with a strong local contextual feature extraction ability. However, the local characteristics of the convolutional layer limit the network to capture global information. The Transformer network has a natural advantage for global information extraction but is not enough for local detailed information extraction because of its self-attention structure. It also has a better ability to deal with long-range dependencies, so integrating the two domains is considered to complement each other to improve the segmentation performance. Chen et al. [26] proposed Trans-Unet, which adopts U-Net as the overall network architecture and uses a Transformer in the encoder structure to extract more features. Lin et al. [27] proposed DS-TransUnet, which incorporated a deeper size hierarchical Swin transform in the encoder and decoder for feature extraction and enhanced modeling capabilities that preserve a wide range of contextual information. Cao et al. proposed Swin-Unet, incorporating the Swin Transformer module to effectively solve the problems of uneven illumination and noise interference in underwater image segmentation. Reza Azad et al. [28] proposed TransDeepLab, based on Deeplabv3 incorporated into the Swin Transformer module, which has excellent performance on the Synapse Multi-Organ Segmentation and Skin Lesion Segmentation tasks.

3. Methods

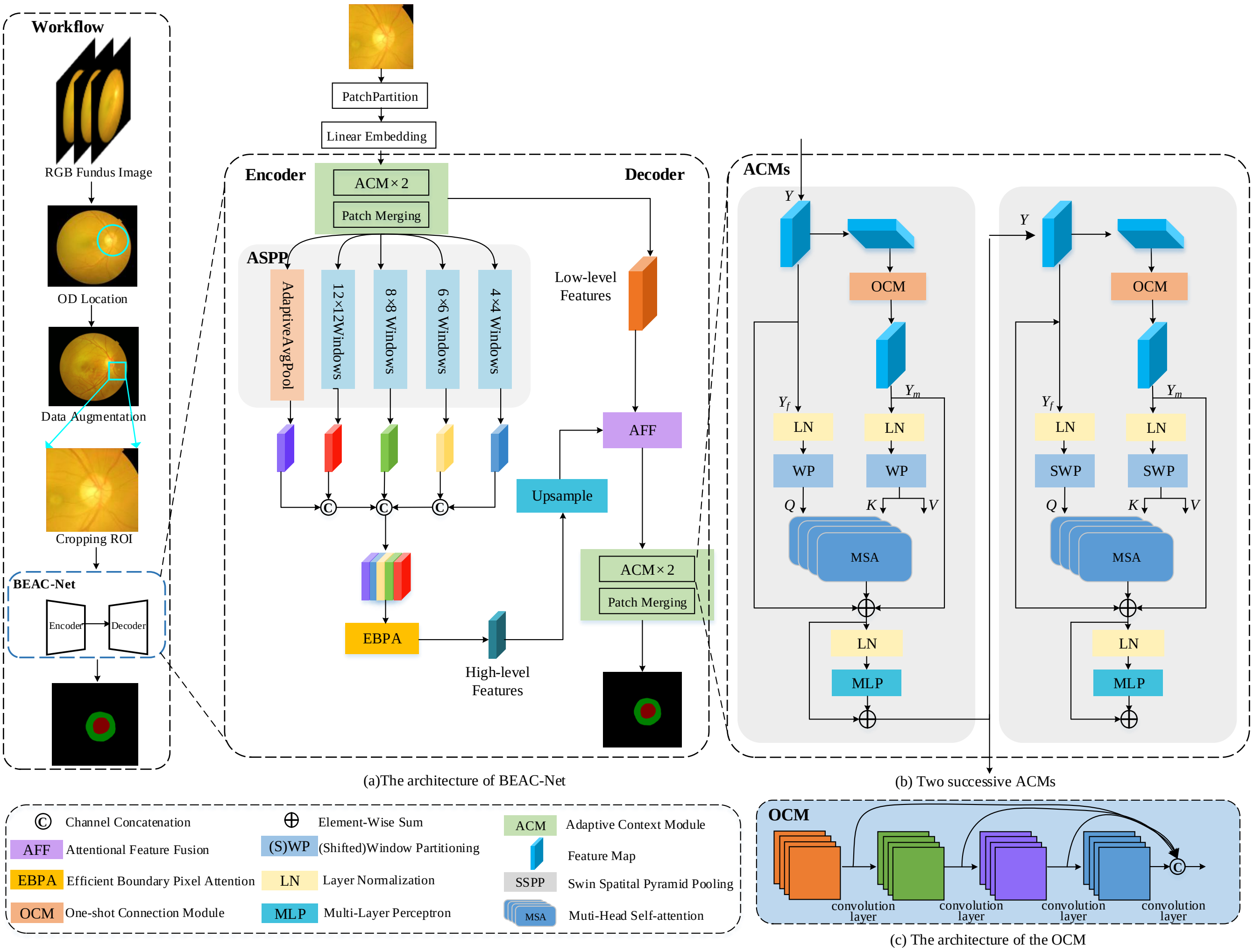
Our proposed method follows the workflow of the automatic segmentation of OD and OC, as shown in Figure 2. Firstly, fundus images are localized with OD via a morphological image processing method to locate the center of the OD. Secondly, image enhancement operations on the original image due to the contrast of the photographs taken on a 66 Vision-Tech fundus camera are not enough. Figure 3 shows that image enhancement is achieved using the SRGAN image generation network by alternately training the generator and discriminator to convert low-resolution images to high-resolution images to improve the quality and clarity of the images. In addition, the boundary parts of the OD are blurred, and the data enhancement can clarify the boundary parts of the OD to improve the contrast of the image. Thirdly, the original image is cropped to a region of interest (ROI). Finally, the cropped optic disk images and labels are trained on the BEAC-Net segmentation network to complete the segmentation.

3.1. Overview

Cropped and data-augmented fundus images are the input sent to the BEAC-Net network. In Figure 2a, we propose a BEAC-Net that consists of encoders and decoders. The encoder captures the multi-scale contextual information to progressively reduce the feature map while capturing more advanced semantic information, and the encoder implements the gradual recovery of spatial information to capture clearer OD and OC boundaries. In the encoder, we develop an adaptive context module (ACM) to enhance both local-range and long-range contextual representations. More specifically, to model adaptive spatial pyramid pooling (ASPP), a pyramid of ACM modules with muti-scale shifted window sizes is designed. To accurately recognize the OD and OC boundaries, it is necessary to extract and fuse features from different levels to capture a better long-range dependency context simultaneously. In the channel dimension, different levels of feature maps are concatenated together. Efficient boundary pixel attention (EBPA) is used to realize the inter-relationship between the boundary feature pixel points of the OD/OC and other pixel points. Furthermore, the boundary pixel points are weighted with the inter-relationship to extract the high-level semantic features, which further extracts and integrates the feature maps. To guide the decoder processing, richer contextual information is captured by upsampling via 4 × 4 convolution to obtain high-level semantic information, which is fused with the low-level semantic information via attentional feature fusion (AFF) to solve the blurred boundaries during the upsampling process. We obtain more effective segmented boundary features, which are significant for fundus image analysis.

3.2. Adaptive Context Module

In the encoder part, we develop a network based on the adaptive context module (ACM) to fully and effectively utilize both local-range and long-range semantic information interactions. In addition, ACM is used as a feature extraction tool to obtain multi-scale features via different shift windows in the hierarchy. We construct a one-short connection module (OCM) into the ACM to extract the local multi-scale information, as shown in Figure 2c. Two successive ACMs realize the partial window calculation self-attention, as shown in Figure 2b. Each window partitioning (WP) and shifted window partitioning (SWP) are used in ACMs to greatly reduce the computational complexity.

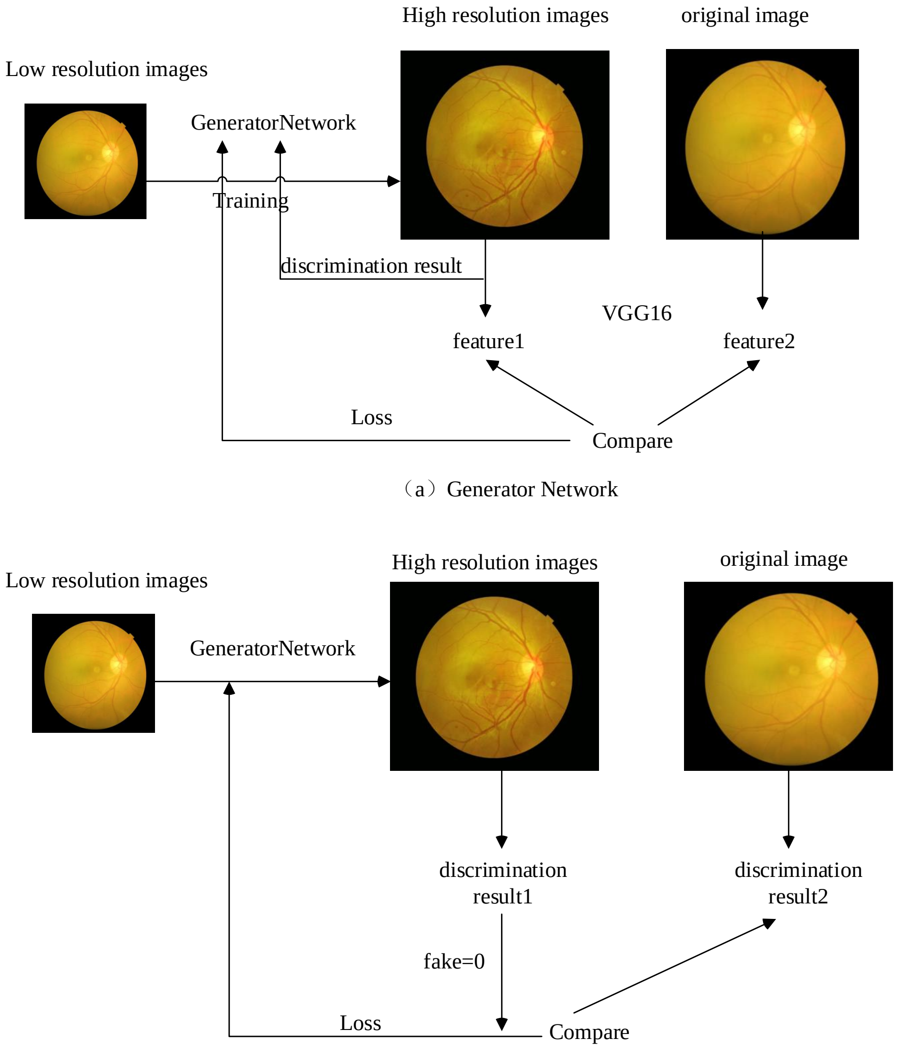


(a)The architecture of BEAC-Net

(b) Two successive ACMs

(c) The architecture of the OCM

Figure 2. (**a**) The overview of proposed BEAC-Net. (**b**) The architecture of two successive ACMs. (**c**) The detailed architecture of OCM.



（a）Generator Network

（b）Discriminator Network

Figure 3. SRGAN network for image enhancement.

When ACM receives the features $Y \in \mathbb{R}^{H \times W \times C}$ with a height of H, a width of W, and a channel of C, it will calculate the inputs of the MSA using two parallel branches, where Q is query, K is key, and V is value. The multi-head attention self-attention uses global attention to learn features with good generalization performance. In the left part, feature Y is split into non-overlapping windows with a size of $H \times W$ via (S)WP. Then, the features are reshaped as $Y_f \in \mathbb{R}^{M \times C}$. A full connected layer is applied to acquire query $Q \in \mathbb{R}^{M \times d}$, where in $d = C/k$, k denotes the head number. The output of the concatenation is the input of the right part. In the right part, feature Y is first utilized to extract local-scale information using OCM. Similar to DenseNet, all features are concatenated once in the last feature map, which keeps the input size constant and enables the expansion of the new output channel in Figure 2c. Two consecutive ACMs consist of a shifted-window-based MSA with two layers of MLPs. LN layers are used before each MSA block and each MLP, and residual connections are used after each MSA and MLP. OCM consists of two 1×1 convolution layers and three 3×3 depthwise separable convolution layers as dilation rates $r = \{1, 2, 3\}$. The features of all previous layers are sent as input to the separable 3×3 depthwise convolution. With $(y_0, y_1, \cdots, y_{m-1})$ as input, we obtain the following:

$$y_m = C_m([y_0, y, \ldots, y_{m-1}]) \tag{1}$$

where C_m is a multi-dimensional concatenate operation. We use the same operations on the left part for feature acquisition on the OCM to generate $y_l \in \mathbb{R}^{M\times C}$, where $K \in \mathbb{R}^{M\times d}$, $V \in \mathbb{R}^{M\times d}$ are acquired from Y_l. MSA is defined as follows:

$$MSA(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d}} + B\right)V \tag{2}$$

where $Q, K, V \in \mathbb{R}^{M\times d}$; d is the query/key dimension; and B is the bias matrix, of which values are obtained from $B \in \mathbb{R}^{M\times M}$.

3.3. Adaptive Spatial Pyramid Pooling

As the encoder module uses ACM and patch merge operations, the resolution of the extracted deep feature space is greatly reduced. Thus, learning from ASPP in Deeplab to capture multi-scale information parallel with multiple atrous rates, we design an adaptive spatial pyramid pooling (ASPP) module with four different shifted window sizes, including a 4 × 4 window, 6 × 6 window, 8 × 8 window, and 12 × 12 window, to capture multi-scale representation. In addition, we use the adaptive average pooling operation (AdaptiveAvgPool operation) on the input feature. As described above, smaller shifted windows aimed at capturing local information, and larger shifted windows aimed at capturing high-level features. Afterward, the results of the multi-scale representation are fed into an efficient criss-cross attention module that fuses and captures a generic representation in a nonlinear technique.

3.4. Efficient Boundary Pixel Attention

The fundus shows that the OC boundary regions are usually complex and blurred, and it is impossible to specify the boundary location. To solve this problem, we propose an efficient boundary pixel attention (EBPA) to collect contextual information in the horizontal and vertical directions to improve the boundary pixel feature capability, which effectively exploits the information of adjacent regions around the OD and OC boundaries to compensate for inadequate feature capture and to enhance the accuracy of segmented boundaries. We focus on acquiring pixel features near the OD and OC boundary region, and boundary detection provides powerful complementary information for semantic segmentation.

Figure 4 shows that the feature map X is shaped $R^{C\times W\times H}$. EBPA first uses two convolutional layers with 1×1 convolution on X and then generates two feature maps, $Q \in R^{C_1\times W\times H}$ and $K \in R^{C_2\times W\times H}$, where C_1, C_2 denotes the number of channels and $C_1, C_2 < C$. We use the affinity operation according to Q and K to generate attention map $B \in \mathbb{R}^{(W\times H)\times(H+W-1)}$. We can obtain a vector $Q_m \in R^{C_1}$ at each position u in the spatial dimension of Q. From the same horizontal and vertical directions at the corresponding location, the features are collected to the feature point m in $K^{C_2\times W\times H}$ to obtain $\Omega_m \in \mathbb{R}^{(H+W-1)\times C_2}$ simultaneously. $\Omega_{n,m} \in R^{C_2}$ is the n-th element of Ω_m. $d_{n,m} = Q_m \times \Omega_{n,m}^T$, $d_{n,m} \in D^{(H+W-1)\times(W\times H)}$ is the correlation degree between Q_m and $\Omega_{n,m}$, $i = [1,\ 2,\ \ldots,\ H+W-1]$. Then, the channel dimension obtains the attentional feature map B via a softmax operation. The other convolutional layer with 1×1 convolution is applied on X to generate $V \in R^{C_3\times W\times H}$ for feature adaptation. In the spatial dimension of V, a vector $V_m \in \mathbb{R}$ is obtained at each position of the feature point u and collects features from the same horizontal and vertical directions from the position of u to obtain the set $\Phi_m \epsilon \mathbb{R}^{(H+W-1)\times C}$. Finally, using the attentional feature map A on feature map V, the process is shown in the following equation: $\tilde{X_u} = \sum\limits_{i=0}^{H+W-1} A_{n,m}\phi_{n,m} + X_m$, where $\tilde{X_u}$ is a feature vector in $\tilde{X} \in \mathbb{R}^{(H+W-1)\times C_3}$ at position u, and $A_{n,m}$ is a scalar value at channel i and position u in A. Using all of the above operations, feature map $\tilde{X_m}$ with a larger receptive field of perception is obtained. Furthermore, the attentional feature map can be used so that contextual features can be selectively aggregated.

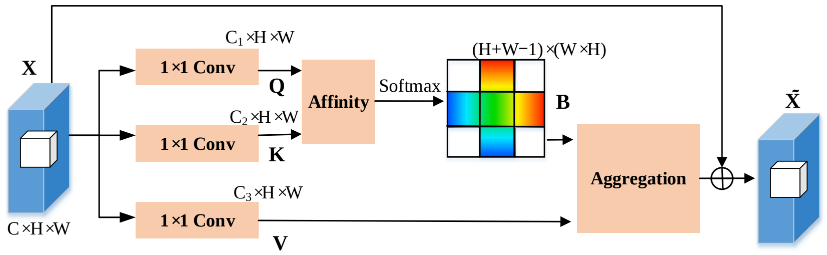


Figure 4. Structure of proposed efficient boundary pixel attention (EBPA).

3.5. Attentional Feature Fusion

In the decoder, to exploit the complementary spatial structure details and semantic information, we design an attentional feature fusion (AFF) module, which fuses different levels of features to address the potentially large inconsistencies in scale and semantics, as shown in Figure 5. Low-level features are rich in spatial detail, and high-level features are richer in semantic information. To this end, the AFF module selectively aggregates features at different levels based on different weights and optimizes feature maps at high levels and low levels. The whole procedure is summarized in Algorithm 1. The features obtained using the attention module are upsampled by 4 × 4 to obtain the high-level semantic information X, which is a feature with a larger receptive field, and the low-level semantic feature Y for feature fusion to obtain feature Z. X and Y first perform the initial feature fusion. After the sigmoid function, the output value is between 0 and 1. X and Y calculate the weighted average so that the group fusion weight is subtracted by 1.

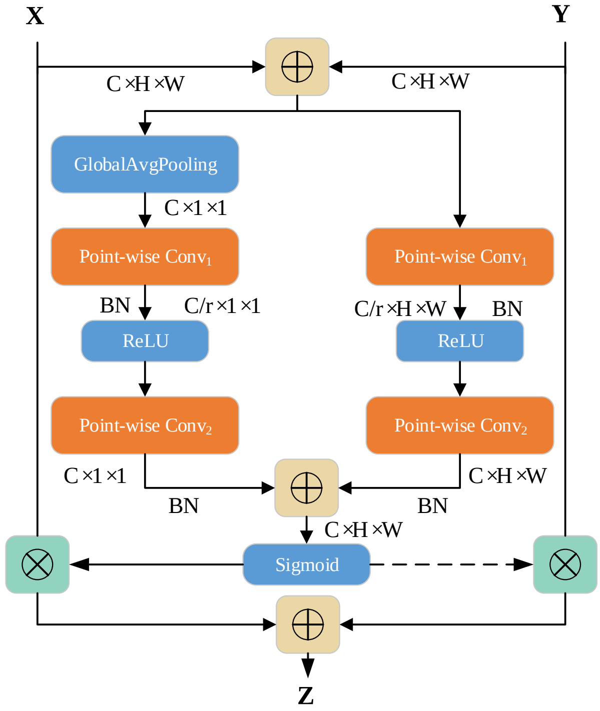


Figure 5. Structure of proposed attentional feature fusion (AFF).

The calculation formula is

$$A = F\left(X \uplus Y\right) \otimes X + \left(1 - F\left(X \uplus Y\right)\right) \otimes Y \tag{3}$$

where $F(X)$ is concerned with the scale of the channel via point-by-point convolution, and the channel attention of the point convolution local features is calculated using the formula $H(X)$:

$$H(X) = Y(PWConv_2(\sigma(Y(PWConv_1(X))))) \tag{4}$$

where $PWConv_1 1 \times 1$ point-wise convolution reduces the number of input feature x channels to $\frac{1}{r}$, B is the BatchNormalization layer, σ is the ReLU function that via $PWConv_2 1 \times 1$ convolution restores the channel numbers to the same number as the input channels, and r denotes the channel reduction ratio.

$M(X)$ is the channel attention formula for the global feature and differs from $H(X)$ in that a global average pooling (GAP) operation is first performed on the input A. $F(X)$ is expressed as

$$F(X) = \sigma\Big(H(X) \bigoplus M(X)\Big) \tag{5}$$

where $H(X)$ is the channel attention of the local features, $M(X)$ is the channel attention of the global feature, and σ denotes the ReLU function.

Algorithm 1 Description of the AFF

```
Input: Feature map X, Feature map Y, size = [B, L, C], $B = Batchsize$, $L = H * W$,
    $C = Channels$ //X is low-level feature, Y is high-level feature
Output: Feature map Z, size = [B, L, C]
$input = [B, L, C] \rightarrow input = [B, H, W, C]$ //$H = height$, $W = weight$ as the height
    and weight of fundus image
x = [B, C, H, W], y = [B, C, H, W]
xa = x + y
$x\_local = local\_att(xa)$ //Channel attention of local features
$x1 = k1 * xa$
$x2 = \delta(batchnormal(x1))$
$x\_local = batchnormal(k2 * x2)$
$x\_global = global\_att(xa)$ //Channel attention of global features
$x1 = AdaptiveAvgPool(xa)$
$x2 = k3 * x1$
$x3 = \delta(batchnormal(x2))$
$x\_global = batchnormal(k4 * x3)$
$xlg = x\_local + x\_global$
$w = sigmoid(xlg)$
$z_1 = 2 \times x \times w + 2 \times y \times (1 - w)$
$Z = tail(z_1)$
return $Z$
$In\ the\ above\ formulas, \delta\ refers\ to\ ReLU\ function,\ *\ denotes\ convolution\ operation,$
    $kidenotes\ convolutional\ filters\ with\ kernel\ size\ 3 \times 3,\ AdaptiveAvgPool\ denotes$
    $AdaptiveAvgPooling,\ sigmoid\ denotes\ sigmoid\ function.$
```

3.6. Loss Function

We observed that the fundus image in the dataset has some common characteristics; the background region (the area is black in the fundus image) and the foreground region (the OD and OC area) are imbalanced. Only one small part of the OC region is occupied. This problem affects the robustness and stability of the training model and the oscillations and anomalies of each evaluation metric. To improve the performance of the Dice coefficient and IoU, we conduct both multi-class cross-entropy and Dice Loss in the loss function.

(1) Multi-class cross-entropy: the multi-class cross-entropy is defined as

$$L_{ce} = \frac{1}{N}\sum_{i} L_i - \frac{1}{N}\sum_{i}\sum_{c=1}^{M} y_{ic} log(p_{ic}) \tag{6}$$

where M is the class number, $p_{ic} = \{0, 1\}$, and p_{ic} is the prediction when sample i belongs to c.

(2) Dice coefficient: Dice Loss training focuses more on mining the foreground area, which can alleviate the negative impact of the foreground–background area imbalance in the sample, which means that most of the area in the image does not contain the target. Furthermore, only a small part of the area contains the target, which is formulated as follows:

$$L_{Dice} = 1 - \frac{2\sum_{n=1}^{N} p_n y_n + \alpha}{\sum_{n=1}^{N} p_n^2 + \sum_{n=1}^{N} y_n^2 + \alpha} \tag{7}$$

where p denotes the prediction; y denotes the ground truth, which is usually used in semantic image segmentation; and α is set to a fixed value to achieve a non-zero denominator. We let the loss function calculate normally. We can also assign the maximum value of α in the denominator to avoid overfitting problems in the trained model.

Finally, we define the sum of the cross-entropy loss and Dice Loss together as follows:

$$L_{loss} = w_{ce} L_{ce} + w_{dice} L_{dice} \tag{8}$$

where w_{ce} is the weight of the cross-entropy loss, and w_{dice} is the weight of the Dice Loss. In this study, we use $w_{ce} = 0.4$ and $w_{dice} = 0.6$ in the training phase.

4. 66 Vision-Tech Dataset

66 Vision-Tech Dataset: Our fundus dataset is named the 66 Vision-Tech dataset, which is taken with a dilatation-free fundus camera (YZ50A1) from 66 VISION TECH Co., Ltd. The fundus camera has excellent performance with its near-infrared light source illumination. It can realize pupil precision alignment and split line fine focus. The currently available public datasets RIM-ONE-v3 [29] and DRISHTI-GS [30] are taken with Canon CR-2 and Nidek AFC-210 fundus cameras, respectively. The number of fundus images of the three datasets are shown in Table 1. Figure 6 shows the large differences in images, such as base color, brightness, texture, and contrast.

In Figure 6, OD occupies a small percentage of the fundus image. To improve the segmentation accuracy and prevent many irrelevant background areas from affecting the segmentation results, we pre-process the image with data expansion and contrast enhancement and locate the brightest point in the fundus image as the center of OD. Then, an external expansion twice the radius of the OD region is cropped out with the center. The cropped region contains both OD and OC intact. The label is also cropped by the corresponding area simultaneously.

Table 1. Comparison of amount among three datasets.

Dataset	Year of Publication	Total Number	Number of Training Sets	Number of Testing Sets
DRISHTI-GS	2017	101	50	51
RIM-ONE-v3	2021	159	140	19
66 Vision-Tech	2023	150	130	20

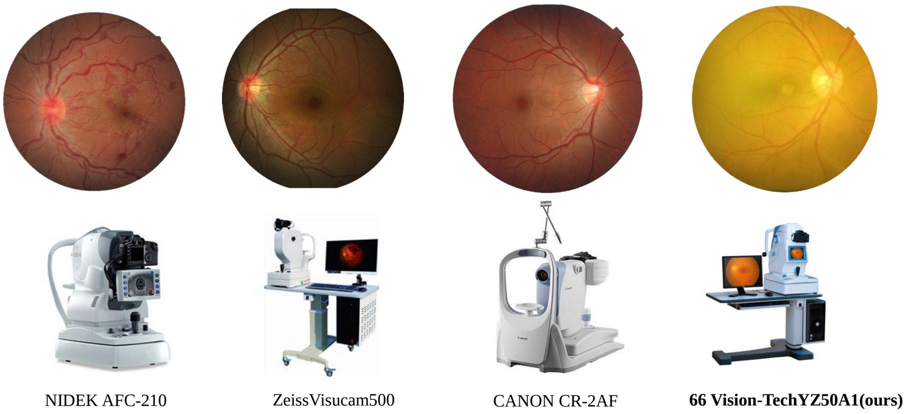


Figure 6. Fundus images taken by different fundus cameras.

5. Experiments

For better quantitative analysis, we conducted a statistical comparison and collected data on the following indicators: Dice, IoU, and HD. Due to the limited number in the fundus image dataset, we use two datasets, RIM-ONE-v3 and DRISHTI-GS, to train the model and use the 66 Vision-Tech dataset to test the generalization ability of the proposed model. The model requires a generalization ability to obtain excellent segmentation results because different fundus image datasets have different feature distributions.

5.1. Datasets

RIM-ONE-v3 dataset: The RIM-ONE dataset consists of three versions with image numbers 169, 455, and 159. In this paper, we use the dataset with five ophthalmologist annotations in the dataset. We used the division method of Wang et al. [31] to obtain the training and test images in the dataset.

DRISHTI-GS dataset: The Drishti-GS dataset contains 101 retinal images and mask annotations for the optical discs and optical cups used to detect glaucoma. The dataset has been divided into 50 training and 51 testing images. All the images have been annotated by four ophthalmologists.

66 Vision-Tech Dataset:The 66 Vision-Tech dataset contains a total of 160 fundus images, 140 for training and 20 for testing. The camera captured the original image size of 2000 × 1600 pixels, and due to the strong light setting and eyeless design, the overall color of the image was yellowish and slightly foggy. Therefore, the original image needed to be pre-processed, as shown in Figure 7.

1. Data annotation. Two ophthalmologists used Labelme to annotate 160 fundus images of the optic disk (OD) and optic cup (OC). In this paper, 140 were randomly selected as the training set and 20 were randomly selected as the test set using the division method following [31].
2. Data Augmentation. The fundus images are pre-processed for image defogging and contrast enhancement using the Automatic Color Enhancement (ACE) [32] algorithm, which corrects the final pixel values by calculating the degree of lightness and darkness of the target pixels and the surrounding pixels and their relationship to achieve contrast adjustment of the image. Due to the small number of images, in the experiment, we use random vertical flip, random horizontal flip, and random diagonal flip to every image [33], and one image is expanded to $2 \times 2 \times 2 = 8$ images by expanding the dataset.
3. Cropping ROI [34]. A small percentage of the OD area in the fundus images is used to obtain more accurate boundary segmentation of the optic disk and reduce the influence of background noise regions on the segmentation results. The center of the OD is detected by the brightest point in the fundus image as the center of the OD. Then, it performs an external expansion twice the radius of the cropped OD region

to obtain the full area of the OD and OC, which is sent to the BEAC-Net model for training and testing. In particular, the cropping images reduce the computing load on the computer and improve the computing efficiency. The labels are cropped into the corresponding ROI simultaneously.

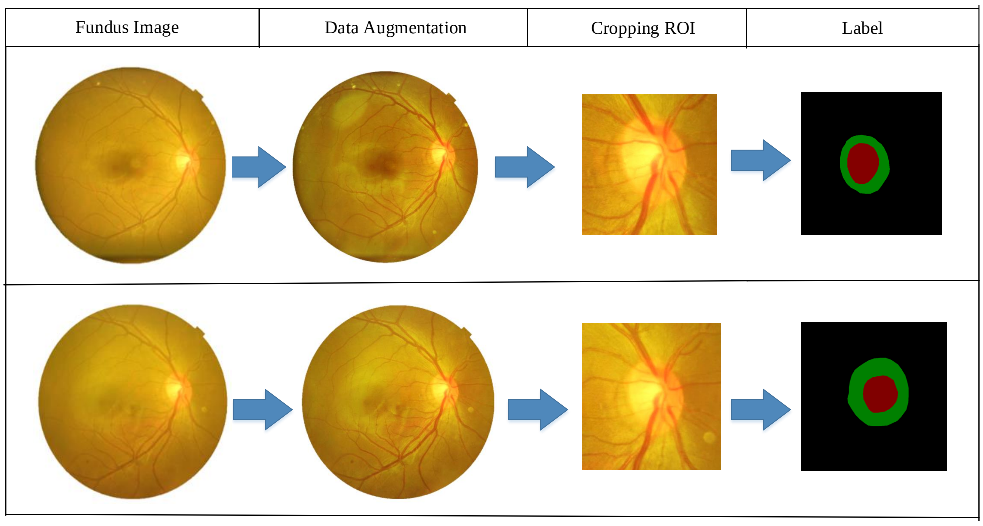


Figure 7. OD and OC segmentation in fundus images into three classes: disk (green), cup (red), and background (black).

5.2. Implementation Details

The BEAC-Net is implemented based on Python 3.6 and Pytorch 1.7.0 using Ubuntu 18.04 Linux and NVIDIA RTX3060 GPU for the experiments. For OD and OC partitioning, two public datasets were used as the training set to fit and optimize the model parameters. Subsequently, tests were performed on the internal dataset to evaluate the performance of the network. During training, the epochs are 300, and the popular SGD optimizer with a momentum value of 0.9 and weight decay value of 1×10^{-4} is used to optimize our model for backpropagation, along with a learning rate update strategy and an early stop mechanism. The learning rate decreases gradually from 0.001, and when the loss does not decrease after 10 training rounds, the network learning rate is reduced to half. During training, the batch size is set to 4, and the output segmented image size is 640 × 640 × 3.

5.3. Evaluation Metrics

We took a task-specific approach to the range of evaluation metrics, aiming to allow each experiment to be compared on the basis of having the same environmental settings. Accuracy, Dice coefficient ($Dice$), Intersection over Union (IoU),and Hausdorff Distance ($\mathrm{H_d}$) are our evaluation metrics used to measure the semantic performance for segmenting OD and OC relative to the ground truth. These parameters are formulated as follows:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{10}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{11}$$

$$\mathrm{H_d}(\mathrm{X,Y}) = \max\{\mathrm{d_{XY}}, \mathrm{d_{YX}}\} = \max\left\{max_{x \in X} min_{y \in Y} \mathrm{d}(\mathrm{x}, \mathrm{y}), max_{y \in Y} min_{x \in X} \mathrm{d}(\mathrm{x}, \mathrm{y})\right\} \tag{12}$$

where TP is true positive, FP is false positive, TN is true negative, and FN is false negative, respectively. X and Y are the number of pixels in the predicted and labeled binary mask images, respectively. The value of the Dice coefficient is between [0, 1]; the closer to 1, the better the segmentation result. The value of the IoU is also between [0, 1]; the larger the value of IoU, the more the region predicted to be the OD and OC overlaps with ground truth and the better the segmentation result.

5.4. Comparison with State-of-the-Art Models

To further validate the segmentation performance of the proposed BEAC-Net, we compared several current state-of-the-art models, including U-Net [8], Deeplabv3+ [15], M-Net [12], and Swin-Unet [35]. The same experimental environment is set up for all model training, while the same pre-processing operations are performed on the fundus image. Figures 8–10 show a visual comparison of the segmentation results between our BEAC-Net and other methods of OD and OC segmentation on three datasets. The segmentation results of the OC and OD show that the boundary part of the BECA-net is closest to the ground truth, indicating that the OC boundary pixels are correctly categorized, and it enhances the feature capture to collect the boundary contextual information via the EBPA module. The segmentation method based on the U-Net network with its U-shaped structure and jump connections can effectively transfer the high-level and low-level information, but the number of network layers is small and cannot capture boundary features. Deeplabv3+ predicts the feature map directly bilinear upsampling 16 times to the desired size, which does not have enough detailed information, and the segmentation results are affected by blood vessels and bright lesions. Swin-Unet uses the Swin Transformer to extract image feature information, and the structure can better capture different levels of features, but there are difficulties in preparing segmentation for OD with large noise backgrounds. The reason that these methods do not give the expected results in segmenting OC and OD is that the boundary features and the contextual information of the boundaries are not extracted completely.Our proposed BEAC-Net can make the boundaries of OD and OC regions more sensitive by introducing boundary awareness into the network, make the AFF module focus on fusing the multi-scale features, and eliminate the influence of blood vessels and background noise on the segmentation results. In Figure 10, BEAC-Net produces segmentation results that are closest to the ground truth.

5.5. Results on Cross-Dataset

In this part, we use the RIM-ONE-v3 and DRISHTI-GS datasets for testing and introduce the 66 Vision-Tech dataset to test the generalization ability of the BEAC-Net. Figure 11 shows the training results, including the accuracy curve, Dice curve, HD curve, and loss curve, for training the BEAC-Net network on the DRISHTI-GS dataset. As shown in the figure, the curves of the network converge faster in the early training period, indicating that the BEAC-Net network has a strong learning ability, and as the training rounds reach 50, the curves of the network training begin to level off with only a small fluctuation.

Table 2 shows the comparisons of the quantitative results of the RIM-ONE-v3 and DRISHTI-GS datasets. Our proposed model was compared with U-Net, Deeplabv3+, Encoder-Decoder CE-Net, M-Net, Ensemble CNN, U-shaped convolutional neural network, Robust, and Swin-Unet, which are widely used in the field of semantic segmentation. BEAC-Net outperforms the second-best Swin-Unet, with a Dice value of 0.8582 and IoU value of 0.8385 for OD segmentation on the RIM-ONE-v3 dataset, leading by 0.017 and 0.0284, respectively. In particular, the segmentation results of OC achieve excellent performance, with a Dice value of 0.8087 and IoU value of 0.7633 on the DRISHTI-GS dataset. Through extensive experiments, the validity of BEAC-Net was verified, and it was able to leverage the boundary information to provide supplementary information for better segmentation.

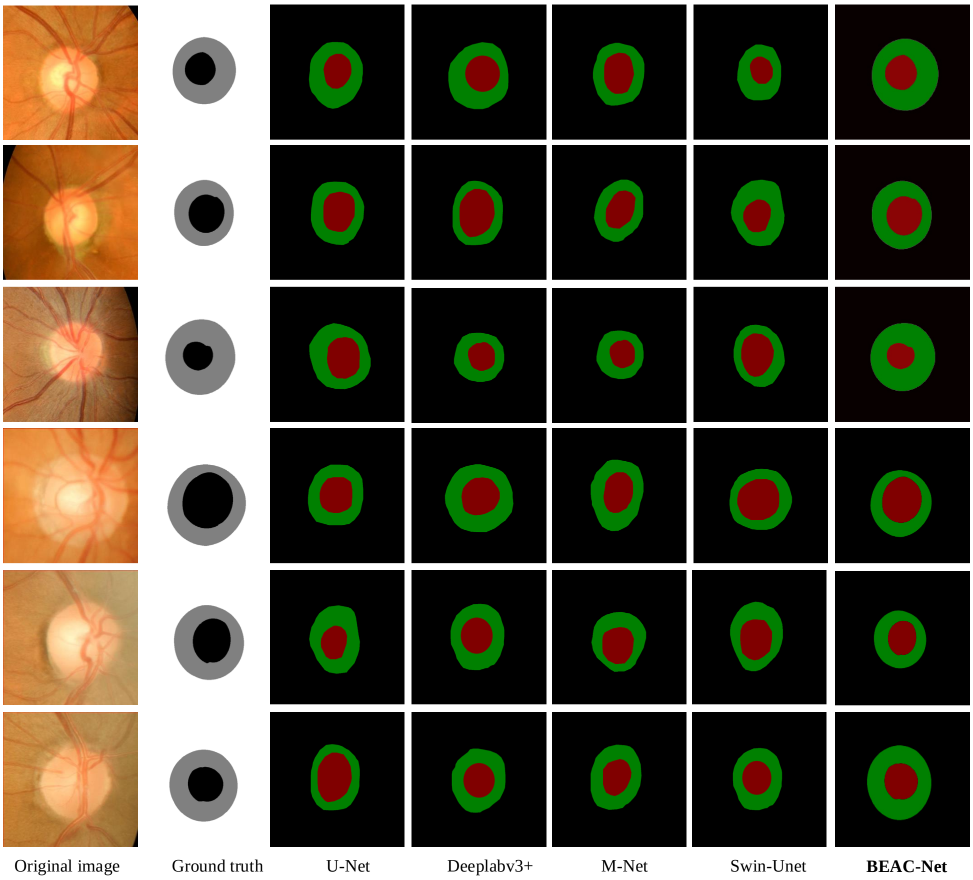


Figure 8. Results of OD and OC segmentation images on the RIM-ONE-v3 dataset, where green colors indicate OD, and red colors indicate OC.

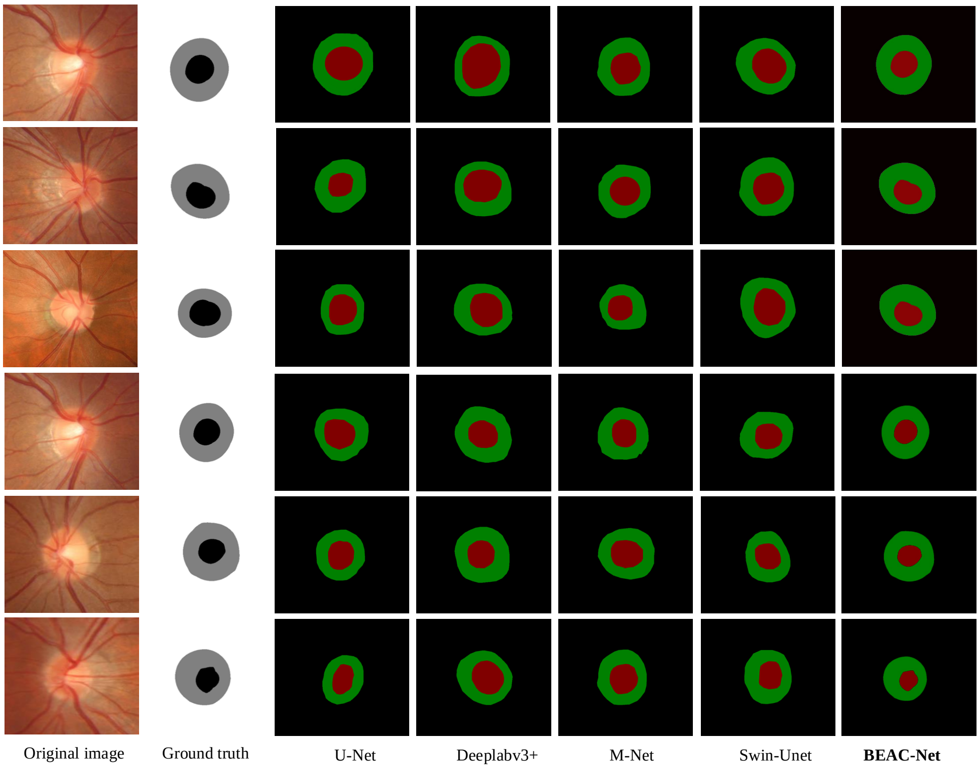


Figure 9. Results of OD and OC segmentation images on the DRISHTI-GS dataset, where green colors indicate OD, and red colors indicate OC.

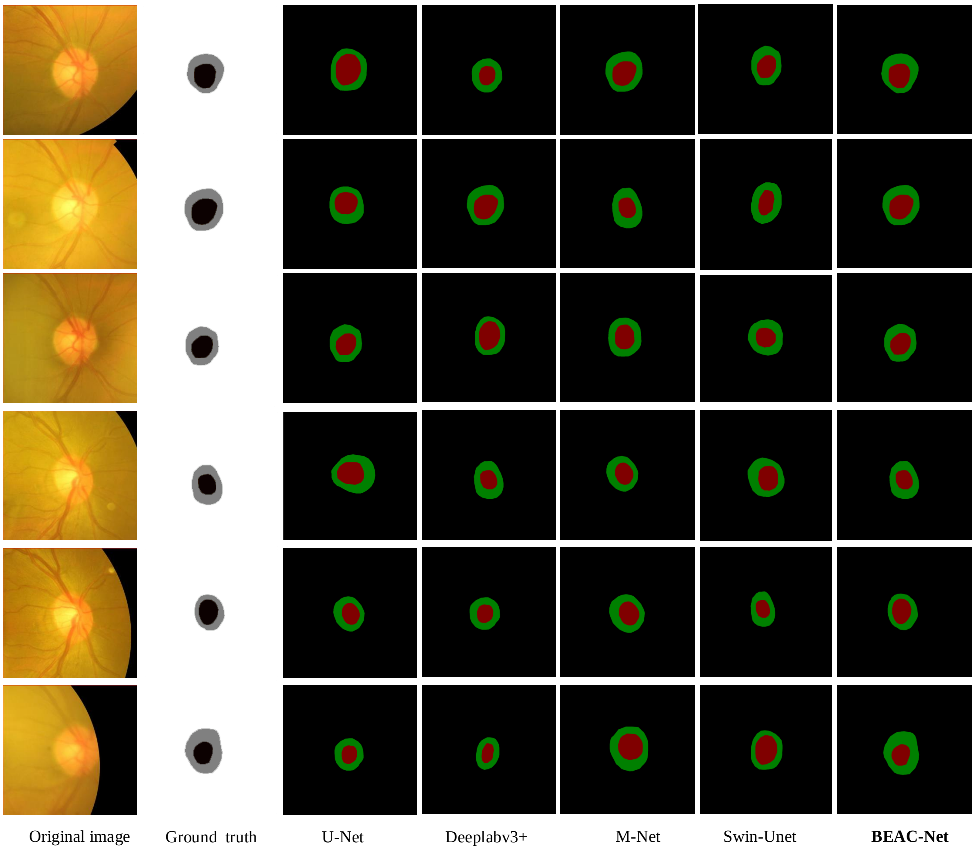


Figure 10. Results of OD and OC segmentation images on the 66 Vision-Tech dataset, where green colors indicate OD, and red colors indicate OC.

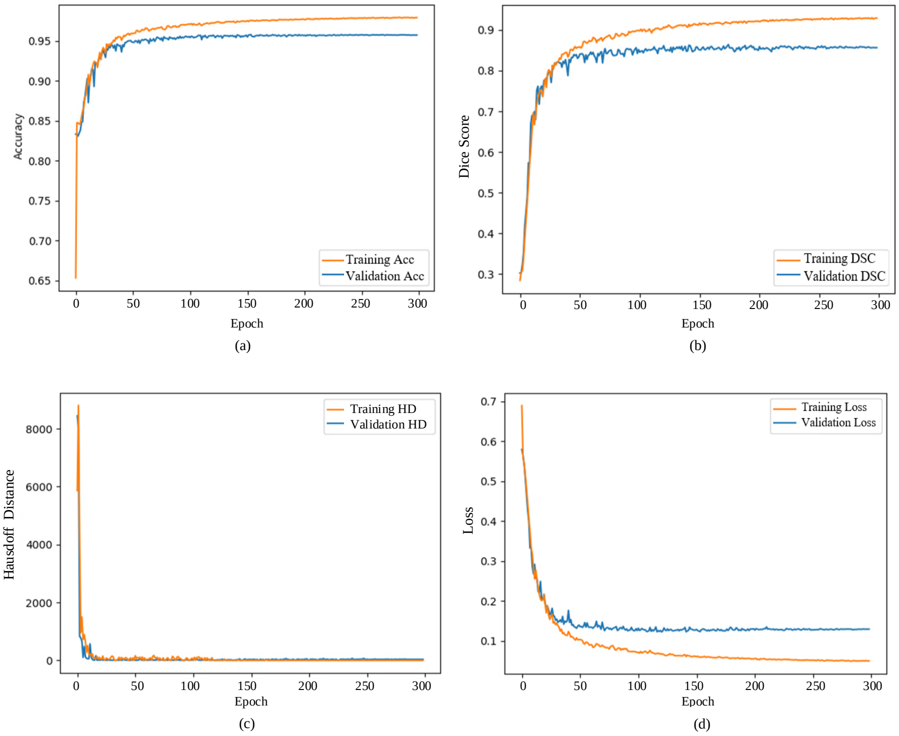


Figure 11. (**a**) Acc curve, (**b**) Dice score curve, (**c**) Hausdorff Distance curve, (**d**) Loss curve for DRISHTI-GS dataset.

We test the model's generalization ability across different datasets. In Table 3, on the 66 Vision-Tech dataset, our proposed BEAC-Net outperforms the other models. Compared to the segmentation results with the proposed BEAC network and other methods, the BEAC network performs best in the Dice and IoU metrics, reaching a Dice value of 0.8267 and IoU value of 0.8138 for the OD and a Dice value of 0.8057 and IoU value 0.7858 for the OC. Compared with Swin-Unet, leading by 0.0118 and 0.0086 for the OD and 0.0616 and 0.0043 for the OC, respectively, BEAC-Net is more sensitive to the segmentation boundary. We can see that BEAC-Net has the best HD performance for the segmentation of the OD and the OC, with the HD of the OD reaching 8.63 and the HD of the OC reaching 9.59, indicating that the ACM module implements the process of capturing both local-scale and long-range information to improve the segmentation accuracy of the pixels in the boundary region. The results show the good generalization properties of our proposed BEAC-Net model and demonstrate that it is an effective approach for a better presentation of image details in the segmentation results.

Table 2. Comparison of segmentation results on RIM-ONE-v3 and DRISHTI-GS dataset among different methods.

Method	**RIM-ONE-v3 Dataset**				**DRISHTI-GS Dataset**			
	OD		**OC**		**OD**		**OC**	
	Dice	**IoU**	**Dice**	**IoU**	**Dice**	**IoU**	**Dice**	**IoU**
U-Net [8]	0.7351	0.8206	0.7176	0.6633	0.7887	0.8206	0.7376	0.7533
Deeplabv3+ [15]	0.7467	0.8344	0.7236	0.6701	0.7861	0.8344	0.7336	0.7001
CE-Net [23]	0.7632	0.8478	0.7592	0.6732	0.7932	0.8478	0.7492	0.7532
M-Met [12]	0.7696	0.8114	0.7348	0.6900	0.8026	0.8114	0.7648	0.7300
Ensemble CNN [10]	0.8132	N/A	0.7240	N/A	0.8120	N/A	0.7740	N/A
U-shaped [33]	0.8344	N/A	0.7564	N/A	0.8361	N/A	0.7764	N/A
Robust [6]	0.8410	0.8256	0.7129	0.6633	0.8310	0.8256	0.7945	0.7429
Swin-Unet [35]	0.8412	0.8101	0.7332	0.6633	0.8582	0.8101	0.7822	0.7532
Our	0.8582	0.8385	0.7333	0.6633	0.8614	0.8385	0.8087	0.7633

Table 3. Comparison Between our model and previous methods on 66 Vision-Tech dataset.

Method	**OD**			**OC**		
	Dice	**IoU**	**HD**	**Dice**	**IoU**	**HD**
U-Net [8]	0.6948	0.7256	35.82	0.6515	0.7534	39.60
Deeplab [36]	0.7231	0.7512	29.74	0.6752	0.7655	33.52
Deeplabv3+ [15]	0.7554	0.7440	21.55	0.6836	0.7521	26.23
M-Met [12]	0.7675	0.7778	22.16	0.6872	0.7678	30.25
Trans-Unet [26]	0.7947	0.8065	18.75	0.7102	0.7763	21.85
Swin-Unet [35]	0.8149	0.8052	12.32	0.7441	0.7815	15.65
Our	0.8267	0.8138	8.63	0.8057	0.7858	9.59

5.6. Ablation Study

To validate the effectiveness of the different components of BEAC-Net, we conducted ablation studies on the 66 Vision-Tech dataset. EBPA, AFF, and ASPP influence and Dice Loss are discussed below; we selected modules that do not contain these as the baseline. Table 4 shows the ablation results. The Dice and IoU of segmentation OD are 0.7523 and 0.7636 when using the baseline, respectively, and adding any of the three modules results in a performance improvement. The OD and OC segmentation accuracy is optimized when all three modules are added.

(1) Efficient Boundary Pixel Attention

In the ablation experiments, we compare the baseline and add the EBPA modules, which can reduce the time and space complexity to $O\left(N\sqrt{N}\right)$ by successively stacking the two EBPA modules. EBPA can capture the optic disk and cup segmentation boundary context information in the fundus image from the horizontal and vertical directions and reduce the model parameters to obtain a lightweight model. In Table 4, the segmented OC Dice and IoU are improved by 0.0228 and 0.0345, respectively, compared with the baseline after adding EBPA. We perceive that better boundary segmentation results are obtained.

(2) Attentional Feature Fusion

We only use the simple linear operation concatenation when we fuse the high-level features and low-level features in the baseline. As shown in Table 4, the IoU of segmenting the OC using the concatenation module is 0.7108, and after replacing AFF from the baseline, the IOU reaches 0.7547, the segmentation accuracy improves to 0.0439, and HD decreases from 33.56 to 25.50.

(3) ASPP influence

ASPP is used to obtain multi-scale features with different shifted window sizes, and we compared adding ASPP to the baseline. In Table 4, there is a significant performance after adding ASPP; OD Dice and IoU are improved to 0.8148 and 0.7952, respectively. The reason is that by using different shifted windows rather than using different atrous convolution, the image extracted features have a better receptive field. After convolution, the image resolution decreases, and many details about the image boundaries are lost, while using shifted windows to obtain features instead of convolution ensures the same resolution of the feature map.

(4) Dice Loss

We combine the classification cross-entropy and Dice Loss assignment weights as the loss function to make the model converge faster in the training phase and fit the boundary shape in the ground truth more effectively. We train two BEAC-Net datasets, one with only the categorical cross-entropy (BEAC-Net + L_{ce}) as the loss function and the other with a combination of categorical cross-entropy and Dice Loss (BEAC-Net + L_{ce} + L_{dice}) as the loss function. Table 5 shows that after introducing the Dice Loss, the OC Dice and IoU values improved by 0.0098 and 0.0191, respectively, from 0.8059 to 0.8157 and 0.7767 to 0.7958.

Table 4. Comparison Between Our Model And Previous Methods On 66 Vision-Tech Dataset.

EBPA	AFF	ASPP	OD			OC		
			Dice	**IoU**	**HD**	**Dice**	**IoU**	**HD**
-	-	-	0.7523	0.7636	35.82	0.7426	0.7108	33.56
✓	-	-	0.7630	0.7841	22.23	0.7654	0.7453	25.43
-	✓	-	0.7837	0.7947	23.36	0.7742	0.7547	25.50
-	-	✓	0.8148	0.7952	21.47	0.7759	0.7557	25.21
✓	✓	-	0.8156	0.8017	15.23	0.7867	0.7712	17.78
✓	✓	✓	0.8267	0.8138	8.63	0.8057	0.7858	9.59

Table 5. Ablation study on the impact of the loss function.

Method	Dice		IoU	
	OD	**OC**	**OD**	**OC**
BEAC-Net + L_{ce}	0.8203	0.8059	0.8121	0.7767
BEAC-Net + L_{ce} + L_{dice}	0.8367	0.8157	0.8238	0.7958

6. Conclusions

In this paper, we propose a novel boundary-enhanced adaptive context network (BEAC-Net), which produces richer contextual information for OD and OC segmentation. BEAC-Net is based on ACM, which can enhance critical features to suppress background features to reduce the negative impact of the noise. The EBPA module is used to capture richer contextual information of the optic disk and cup segmentation boundaries in the fundus image in both horizontal and vertical directions. More importantly, BEAC-Net can integrate the feature maps from different levels via the AFF module adaptively. We also performed extensive experiments on the RIM-ONE-v3, DRISHTI-GS, and 66 Vision-Tech datasets to confirm the effectiveness of BEAC-Net. Since there are many types of diseases of the fundus, we will further investigate and actively explore BEAC-net applications in fundus image lesion segmentation to improve fundus image segmentation performance.

Author Contributions: Conceptualization, L.J., S.Y., X.T. and S.L.; methodology, L.J., S.L. and Y.J.; software, X.T.; validation, S.Y.; writing—original draft preparation, L.J.; writing—review and editing, L.J., S.Y. and X.T.; supervision, S.L. and Y.J. All authors have read and agreed to the published version of the manuscript.

Funding: This project has received funding from the National Natural Science Foundation of China (No. 62176264), Natural Science Foundation of Jiangsu Province (Higher Education Institutions) (20KJA520001), and Open Research Project of Zhejiang Lab (No. 2021KF0AB05).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and code used to support the findings of this study are available from the corresponding author upon request (2022010301@njupt.edu.cn).

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

References

1. Weinreb, R.N.; Leung, C.K.S.; Crowston, J.G.; Medeiros, F.A.; Friedman, D.S.; Wiggs, J.L.; Martin, K.R. Primary open-angle glaucoma. *Nat. Rev. Dis. Prim* **2004**, *363*, 1711–1720.
2. Herndon; Leon, W. Central corneal thickness as a risk factor for advanced glaucoma damage. *Arch. Ophthalmol.* **2004**, *122*, 17–21. [PubMed]
3. Razzak, M.I.; Naz, S.; Zaib, A. Deep learning for medical image processing: Overview, challenges and the future. In *Classification in BioApps: Automation of Decision Making*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 323–350.
4. Feng, Y.; Ji, Y.; Wu, F.; Gao, G.; Gao, Y.; Liu, T.; Liu, S.; Jing, X.Y.; Luo, J. Occluded Visible-Infrared Person Re-Identification. *IEEE Trans. Multimed.* **2023**, *25*, 1401–1413.
5. Liu, Q.; Hong, X.; Ke, W.; Chen, Z.; Zou, B. DDNet: Cartesian-polar Dual-domain Network for the Joint Optic Disc and Cup Segmentation. *arXiv* **2019**, arXiv:1904.08773.
6. Biswal, B.; Vyshnavi, E.; Sairam, M.V.S.; Rout, K. Robust Retinal Optic Disc and Optic Cup Segmentation via Stationary Wavelet Transform and Maximum Vessel Pixel Sum. *Inst. Eng. Technol.* **2020**, 14, 592–602.
7. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 640–651.
8. 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; pp. 234–241.
9. Qin, P.; Wang, L.; Lv, H. Optic Disc and Cup Segmentation Based on Deep Learning. In Proceedings of the Information Technology, Networking, Electronic and Automation Control Conference, Chengdu, China, 15–17 March 2019; pp. 1835–1840.
10. Zilly, J.; Buhmann, J.M.; Mahapatra, D. Glaucoma detection using entropy sampling and ensemble learning for automatic optic cup and disc segmentation. *Comput. Med. Imaging Graph.* **2017**, *55*, 28–41.
11. Liu, Z.; Yuan, H.; Shao, Y.; Liu, M. ResD-Unet Research and Application for Pulmonary Artery Segmentation. *IEEE Access* **2021**, *9*, 67504–67511.
12. Fu, H.; Cheng, J.; Xu, Y.; Wong, D.W.K.; Liu, J.; Cao, X. Joint Optic Disc and Cup Segmentation Based on Multi-label Deep Network and Polar Transformation. *IEEE Trans. Med. Imaging* **2018**, *37*, 1597–1605.
13. Tabassum, M.; Khan, T.M.; Arsalan, M.; Naqvi, S.S.; Mirza, J. CDED-Net: Joint Segmentation of Optic Disc and Optic Cup for Glaucoma Screening. *IEEE Access* **2020**, *8*, 102733–102747.
14. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Liu, W. CCNet: Criss-Cross Attention for Semantic Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Salt Lake City, UT, USA, 23–18 June 2018; pp. 603–612.
15. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 801–818.
16. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T. Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.

17. Aquino, A.; Gegundez-Arias, M.E.; Marin, D. Detecting the optic disc boundary in digital fundus images using morphological, edge detection, and feature extraction techniques. *IEEE Trans. Med. Imaging* **2010**, *29*, 1860–1869. [PubMed]
18. Chen, L. Weakly Supervised and Semi-Supervised Semantic Segmentation for Optic Disc of Fundus Image. *Symmetry* **2020**, *12*, 145.
19. Sukanya, R. Retinal Blood Vessel Segmentation and Optic Disc Detection Using Combination of Spatial Domain Techniques. *Int. J. Comput. Sci. Eng.* **2015**, *4*, 102–109.
20. Cheng, J.; Liu, J.; Xu, Y. Superpixel Classification Based Optic Disc and Optic Cup Segmentation for Glaucoma Screening. *IEEE Trans. Med. Imaging* **2013**, *32*, 1019–1032. [PubMed]
21. Feng, Y.; Yu, J.; Chen, F.; Ji, Y.; Wu, F.; Liu, S.; Jing, X.Y. Visible-Infrared Person Re-Identification via Cross-Modality Interaction Transformer. *IEEE Trans. Multimed.* **2022**, *early access*. [CrossRef]
22. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
23. Gu, Z.; Cheng, J.; Fu, H.; Zhou, K.; Hao, H.; Zhao, Y.; Zhang, T.; Gao, S.; Liu, J. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE Trans. Med. Imaging* **2019**, *38*, 2281–2292. [PubMed]
24. Wang, S.; Yu, L.; Li, K.; Yang, X.; Heng, P.A. DoFE: Domain-oriented Feature Embedding for Generalizable Fundus Image Segmentation on Unseen Datasets. *IEEE Trans. Med. Imaging* **2020**, *39*, 4237–4248. [PubMed]
25. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
26. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306.
27. Lin, A.; Chen, B.; Xu, J.; Zhang, Z.; Lu, G. DS-TransUNet: Dual Swin Transformer U-Net for Medical Image Segmentation. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 3178991.
28. Azad, R.; Heidari, M.; Shariatnia, M.; Aghdam, E.K.; Karimijafarbigloo, S.; Adeli, E.; Merhof, D. Transdeeplab: Convolution-free transformer-based deeplab v3+ for medical image segmentation. In Proceedings of the Predictive Intelligence in Medicine, Singapore, 22 September 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 91–102.
29. Fumero, F.; Alayon, S.; Sanchez, J.L.; Sigut, J.; Gonzalez-Hernandez, M. RIM-ONE: An open retinal image database for optic nerve evaluation. In Proceedings of the 24th International Symposium on Computer-Based Medical Systems (CBMS), Bristol, UK, 27–30 June 2011; pp. 1–6.
30. Sivaswamy, J.; Krishnadas, S.R.; Chakravarty, A.; Joshi, G.D.; Ujjwal. A Comprehensive Retinal Image Dataset for the Assessment of Glaucoma from the Optic Nerve Head Analysis. *JSM Biomed. Imaging Data Pap.* **2015**, *2*, 1004.
31. Wang, S.; Yu, L.; Yang, X.; Fu, C.W.; Heng, P.A. Patch-based output space adversarial learning for joint optic disc and cup segmentation. *IEEE Trans. Med. Imaging* **2019**, *38*, 2485–2495. [CrossRef]
32. Getreuer, P. Automatic color enhancement (ACE) and its fast implementation. *Image Process. Line* **2012**, *2*, 266–277. [CrossRef]
33. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Seattle, WA, USA, 13–19 June 2019; pp. 3146–3154.
34. Nieto-Castanon, A.; Ghosh, S.S.; Tourville, J.A.; Guenther, F.H. Region of interest based analysis of functional imaging data. *Neuroimage* **2003**, *19*, 1303–1316. [CrossRef]
35. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–24 October 2022; pp. 205–218.
36. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *40*, 834–848.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.