applied sciences

MDPI

Article

Polarformer: Optic Disc and Cup Segmentation Using a Hybrid CNN-Transformer and Polar Transformation

Yaowei Feng [1], Zhendong Li [1,2,*], Dong Yang [1], Hongkai Hu [1], Hui Guo [1,2] and Hao Liu [1,2]

[1] School of Information Engineering, Ningxia University, Yinchuan 750021, China
[2] Collaborative Innovation Center for Ningxia Big Data and Artificial Intelligence Co-Founded by Ningxia Municipality and Ministry of Education, Yinchuan 750021, China
* Correspondence: lizhendong13@mails.ucas.ac.cn

Abstract: The segmentation of optic disc (OD) and optic cup (OC) are used in the automatic diagnosis of glaucoma. However, the spatially ambiguous boundary and semantically uncertain region-of-interest area in pictures may lead to the degradation of the performance of precise segmentation of the OC and OD. Unlike most existing methods, including the variants of CNNs (Convolutional Neural Networks) and U-Net, which limit the contributions of rich global features, we instead propose a hybrid CNN-transformer and polar transformation network, dubbed as Polarformer, which aims to extract discriminative and semantic features for robust OD and OC segmentation. Our Polarformer typically exploits contextualized features among all input units and models the correlation of structural relationships under the paradigm of the transformer backbone. More specifically, our learnable polar transformer module optimizes the polar transformations by sampling images in the Cartesian space and then mapping them back to the polar coordinate system for masked-image reconstruction. Extensive experimental results present that our Polarformer achieves superior performance in comparison to most state-of-the-art methods on three publicly available datasets.

Keywords: deep learning; multi-model learning; medical segmentation; transformer; attention

Citation: Feng, Y.; Li, Z.; Yang, D.; Hu, H.; Guo, H.; Liu, H. Polarformer: Optic Disc and Cup Segmentation Using a Hybrid CNN-Transformer and Polar Transformation. *Appl. Sci.* **2023**, *13*, 541. https://doi.org/10.3390/app13010541

Academic Editor: Jan Egger

Received: 12 November 2022
Revised: 18 December 2022
Accepted: 25 December 2022
Published: 30 December 2022

1. Introduction

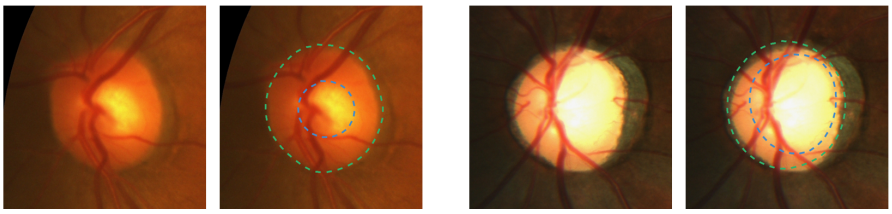
The number of people diagnosed with glaucoma is increasing. According to one forecast [1], the number of glaucoma patients will reach 111.8 million in the future. As a result, early detection is essential for treatment to preserve vision, thereby, avoiding permanent vision loss. In the diagnosis method of glaucoma, OD and OC segmentation is typically performed first as an antecedent task in most glaucoma diagnostic approaches. Given that OD and OC segmentation is costly, tedious and burdensome, accurate segmentation of OD/OC is of great significance in assisting doctors in disease assessment and diagnosis [2].

It is hoped that algorithms may be used to facilitate detection and even diagnosis, which has motivated works on fundus image-segmentation tasks. Figure 1 shows the retinal fundus images of a healthy eye and a glaucoma-suspicious eye, which are the main structures of the fundus captured by the fundus camera, including OD and OC. Compared with the two types of images, glaucoma will cause pathological areas in the background image of the eye and will change the physiological structure of the optic nerve.

The images are clear; however, normal (healthy) images cannot be easily distinguished from abnormal (glaucoma) images with the naked eye. Therefore automated systems need to be developed. Recently, OD and OC segmentation approaches have evolved from conventional craft methods into deep-learning methods. Conventional methods are mostly based on circular transformation for detecting the boundaries [3–5]. However, in these methods, most of the features used are designed by hand, which not only requires strong prior knowledge but also requires a great deal of computation.

Thus, these methods are intractable for segmenting a subtle OC. In response to this problem, deep-learning methods [6–9] play an important role and have achieved promising

performance. However, these deep-learning methods only used original fundus images. According to our observations, a valid geometrical constraint is that the closed nested structures of OD and OC are both of approximately elliptical shape, and there is variability of the object (size, position, etc.). However, representing this information and these constraints is difficult to implement within the network.



(a) Healthy eye (b) Glaucoma-suspicious eye

Figure 1. The region enclosed by the green dotted circle is the optic disc (OD); the central bright zone enclosed by the blue dotted circle is the optic cup (OC); and the region between them is the neuroretinal rim. (**a**) An example of a healthy eye. (**b**) An example of a glaucoma-suspicious eye.

With these in mind, M-Net [6] converts images to polar coordinates to learn representations for improving the performance of segmentation. Under the polar coordinate system, this geometric constraint can be easily converted into spatial relationships and can present an ordered layer structure. However, these similar methods only treat the center of the image as the polar origin in the transformation and coordinate system, while we formulate a Polar Origin Predictor capable of automatically detecting the centroid of the OC as the polar origin.

The previous methods take polar transformation as a preprocessing process with no parameters learned. Since the hyperparameter is not optimal for segmentation evaluation and effective training, they cannot exploit all supervisory signals incorporating the extracted features. To solve this problem, we formulate a polar converter, which conducts the original image, and its polar origin performs a polar transform. Consequently, a learnable polar transformer module (LPTM) is proposed, which consists of a polar origin detector and a polar converter. It is trained for transforming the original images into the polar coordinate system. Thus, the network enables learning the polar representation.

Another challenge is that the blurred edges, shape and size of the OC and OD vary among patients. Features of the background region (such as blood vessels and exudates) easily interfere with the foreground (the OD and OC regions). Accordingly, it is necessary to introduce sufficient contextual information and global information under different receptive fields to search for relationships between features.

The image features extracted by the methods [6–11] mentioned before mostly focus on local information, which cannot accurately capture the global contextual information, particularly from the OD and OC regions, thus, limiting the model's generalizability. The underlying reason is that convolutions have a limited field of perception. In other words, the network focuses more on local features when extracting features and cannot effectively consider the large-range contextual information.

To achieve this, we design a transformer module to capture global contextual information, which enhances the correlation between feature information by using the transformer's unlimited effective receptive fields. Unlike the existing transformer models, our Polarformer incorporates the CNN-based module and the transformer-based module under the deep neural network after the LPTM. For the local visual features, the CNN-based module extracts the local visual features with CNNs to obtain more discriminative features, and then we combine the feature pyramid network (FPN) [12] to improve their spatial resolution, which handles multi-scale feature information.

For the global contextual information, our transformer-based module introduces global contextual information that leverages unlimited effective receptive fields to enhance useful features and suppress useless feature responses, thereby, distilling spatial relationships.

In summary, we propose a hybrid CNN-transformer and polar transformation network to solve the problem mentioned above. Our Polarformer aims to extract discriminative and semantic features for robust OD and OC segmentation, thus, developing a more accurate segmentation model. The remainder of this paper is organized as follows: We present related works on the segmentation of the OD and OC in Section 2. Section 3 gives the specific details of our model. In Section 4, we give our experimental results, compare them with other methods and discuss our approach. In Section 5, we conclude our work.

2. Related Work

2.1. Optic Disc and Cup Segmentation

In the past few years, CNNs have made progress on the optic disc (OD) and optic cup (OC) segmentation tasks. For example, Liu et al. [13] detected OD by training a segmentation network to detect OD from fundus images based on CNN. Mohan et al. [14] combined the FCN [15] network with atrous convolution to achieve automatic segmentation of OD and reliable detection of diseases, such as glaucoma. Furthermore, Sevastopolskyet al. [8] modified a U-Net convolutional neural network for easier and faster OD and OC segmentation tasks; however, it still operates in two stages.

Subsequently, Fu et al. [6] proposed M-Net, which considered the relationship between OD and OC. Their method used OD and OC simultaneously and presented a multi-label loss function to generate the final segmentation images. Zhang et al. proposed a generic medical segmentation framework called ET-Net [16], which extracts discriminative contextual features and selectively aggregates multi-scale information and embedding edge attention representations. This was developed to guide the segmentation process.

Yin et al. [17] presented a region proposal network based on Mask-RCNN localization to pay attention to accurate optic nerve head localization, which combines prior information to learn a discriminative feature representation for segmentation.

2.2. Polar Transformation Networks

Most recently, Salehinejad et al. [18] improved the diversity of the datasets by proposing a sampling method based on generating a new image for each pixel. Liu et al. [19] proposed DDNet, a method that learned rich contextual information from both the Cartesian domain and the polar domain. Zahoor et al. [20] applied a polar transform to convert the circular ROI into different rectangular tiles, which were adaptively thresholded to obtain the exact OD boundary.

Fu et al. [6] first localized the disc center by using the Active Contour Models and then transferred the original fundus image into the polar coordinate system based on the detected disc center. Jiang et al. [21] considered the rotation-invariant problem as a translation-invariant problem, and their method adopted the center loss function to learn rotation-invariant features. In order to achieve rotational invariance, Kim et al. [22] developed a deep network for original images in a polar coordinate system in classification tasks. Their method replaced convolution layers in conventional CNNs with cylindrical convolutional layers by using cylindrical sliding windows.

2.3. Transformer Models

The transformer is different from the traditional CNN and Recurrent Neural Network (RNN). It does not have a complex network structure. The core of the transformer is the self-attentive mechanism, which was applied for the first time in the area of natural language processing. Ashish Vaswani et al. [23] proposed a transformer for the first time, as using the self-attentive mechanism could obtain potential relationships between the input and output.

Due to the advantages of the transformer, researchers began to apply it to the segmentation field. Dosovitskiy et al. [24] interpreted an image as a sequence of patches and processed it using a standard transformer encoder. Liu et al. [25] presented the Swin Transformer, which brought greater efficiency by limiting the self-attention computation to non-overlapping local windows while also allowing for cross-window connection. Carion et al. [26] introduced DETR, a new design for transformer-based object detection systems and bipartite matching loss for direct set prediction.

Our motivation is to propose a hybrid CNN-transformer and polar transformation network by taking advantage of the transformer's unlimited effective receptive fields to introduce global contextual information. To our knowledge, there has been no work considering polar representation with a transformer in the fundus medical task. Although some methods have combined polar transformation with classification tasks by using neural networks, OD and OC segmentation use polar transformation only for a preprocessing step.

3. The Proposed Method

Unlike U-Net-based OD and OC segmentation methods where the global context feature relationship cannot be modeled, our basic idea is to globally optimize the feature extraction process and strengthen correlations between features in parallel to improve the discrimination ability of each pixel's representation. As illustrated in Figure 2, there are two major components in our Polarformer. (1) A learnable polar transformation module (Section 3.1), which performs a differentiable log-polar transform. (2) A CNN-transformer module (Section 3.2), which extracts image features with high resolution and aggregates global self-attention at the end. Finally, we apply a segmentation head to output each class's confidence scores and convert the final predictions back to the Cartesian coordinate system. The combination of two components forms our method and is described in the following sections.

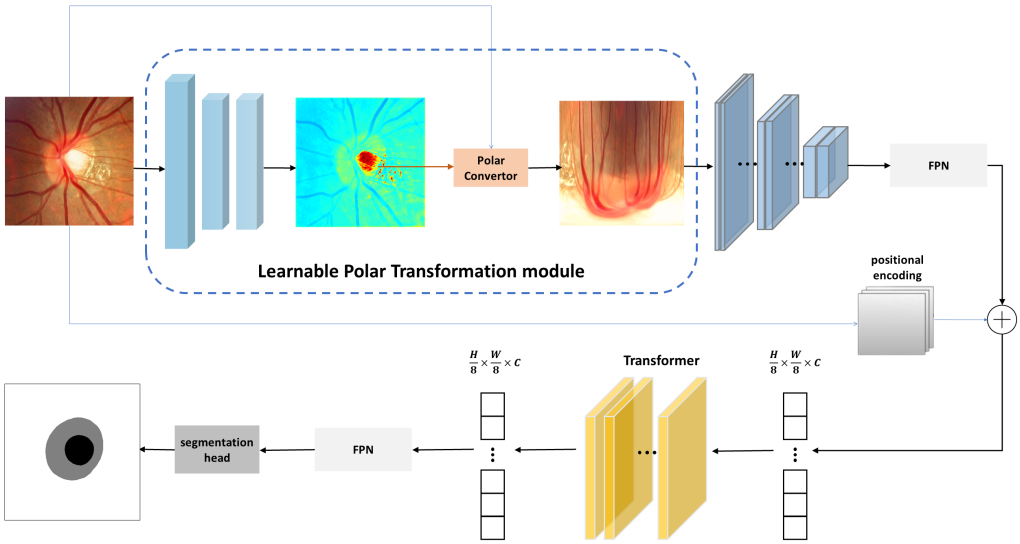


Figure 2. The architecture of our proposed Polarformer.

3.1. Learnable Polar Transformation Module

Regarding the imbalanced class distributions in fundus images, OC pixels are more likely to be misclassified. M-Net [6] segmented the OC and OD in the polar coordinate space to alleviate this problem. However, M-Net takes polar transformation as a pre-processing process without learned parameters. With the hyper-parameter being not optimal for segmentation evaluation and optimal training, features extracted by them cannot exploit the full supervision. To overcome the limitations of using the original images, we designed a learnable polar transformation module (LPTM). As visualized in Figure 3, our LPTM

consists of a polar origin detector, which detects the origin of the polar transform, and a polar converter transforms the original images into a polar representation.

3.1.1. Polar Origin Detector

When performing polar coordinate transformation, the polar origin is an important parameter that determines the final segmentation performance. In order to obtain a better polar representation, it is necessary to select an image center that is close to the segmented object and to then proceed to the next step based on the polar origin. Otherwise, the subsequent analysis in polar coordinates is likely to be inaccurate. Recent deep models [27] directly regress the coordinates of the target point through the fully connected layer of the network. However, these methods are not an optimal choice. Other methods predict heatmaps and take their argmax [28,29].

These are not the best positions because the backpropagation gradients are zero for all parts except the target point, which interferes with training. To be specific, we proceed as follows: According to [30], the green channel highlights the features of the OC. Thus, we take the green channel of the images and follow stepwise convolution to obtain a heat map. The polar origin detector consists of a sequence of standard convolution blocks followed by a 1×1 convolution. The polar origin is the pixel coordinate point with the highest intensity calculated from the heatmap predicted in the last layer of the model.

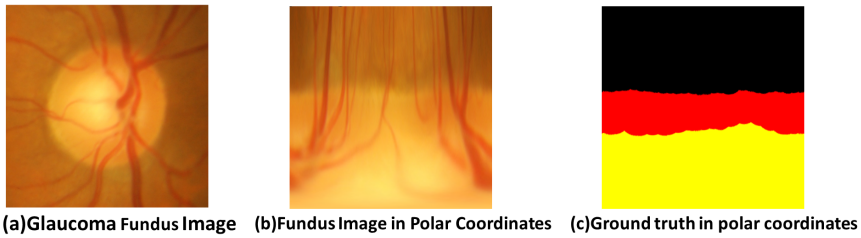


Figure 3. An example of segmentation using polar coordinates. (**a**) Glaucoma fundus image. (**b**) Fundus image in polar coordinates. (**c**) The ground truth in polar coordinates.

3.1.2. Polar Converter

In the original fundus datasets, the OC is contained in the OD, both are approximately elliptical, and there is variability of the object (size, position, etc.). However, the representation of this information and the constraints are difficult to implement in the network. From observations, under the polar coordinate system, this geometric constraint can be easily converted into spatial relationships and can present an ordered layer structure. Following [31], we employ log-polar coordinates for original datasets that effectively improve the cup-to-disk ratio, balance the dataset, prevent overfitting and improve the segmentation accuracy. Inspired by a Spatial Transformer Network (STN) [32], we formulate γ_i to denote the output coordinates, where X denotes the input:

$$x_i^s = x_0 + r^{x_i^t / W_0} \cos \frac{2\pi y_i^t}{H_0}, \tag{1}$$

$$y_i^s = y_0 + r^{x_i^t / W_0} \sin \frac{2\pi y_i^t}{H_0}, \tag{2}$$

where (x_0, y_0) is the origin input coordinates. W_0 and H_0 are the width and height of the output. γ is the maximum distance to the origin, which we set to $0.5\sqrt{H_0^2 + W_0^2}$ in our experiments. (x_i^s, y_i^s) denotes the source sample point coordinates, and (x_i^t, y_i^t) denotes the transformed log-polar coordinates.

3.2. CNN-Transformer Module

The distance between different regions has changed in the images' polar representation. As the input areas have a significant impact on the model output, without a larger receptive field, the network may obtain incorrect segmentation results by misleading local features. In the OD and OC segmentation task, these U-Net variants have limited effective receptive fields and only focus on the local information of images.

To solve this problem, we first exploit CNN to extract the local visual features to obtain more discriminative features. Next, we take advantage of the feature pyramids network to improve their spatial resolution, which handles multi-scale feature information. Then, we propose a transformer-based module, which is used to obtain better feature representation by aggregating long-range context information. Our module computes pairwise interactions (self-attention) between the optic cup and optic disc features, combining their features and generating contextualized features. The results of the experiment proved that transformer could obtain global contextual information.

3.2.1. Feature Pyramids Network

In our method, we use a CNN as a backbone to extract feature maps. Then, we place the rich semantic feature maps $\boldsymbol{X}_0 \in \mathbb{R}^{H_0 \times W_0 \times D_0}$, where D_0 is the number of color channels, W_0 represents the width and H_0 represents the height. Our FPN is shown in Figure 4. The FPN learns a feature map $\boldsymbol{P}^{out}$ from different resolution feature maps $\boldsymbol{P}^{in}$, where $\boldsymbol{P}^{in} = (\boldsymbol{P}_1^{in}, \boldsymbol{P}_2^{in}, \ldots)$, and feature map $\boldsymbol{P}_i^{in}$ is obtained by the encoder at layer $\boldsymbol{i}$. The size of the feature is (W, H), and the feature map size $\boldsymbol{P}_i^{in}$ is $(\frac{W}{2^i}, \frac{h}{2^i})$. In the network, we designed this as $\boldsymbol{P}^{out} = \text{upsample}_{\times 2}(\boldsymbol{P}_i^{in}) + Conv_i(\boldsymbol{P}_{i-1}^{in})$. As described above, $\boldsymbol{P}(\boldsymbol{X}_0) = 1/16$ of the origin. In this case, segmentation cannot be done accurately due to the coarseness of the data. Thus, we obtain upsampled feature maps $\boldsymbol{P}^{out}$ by upsampling the **input FPN**.

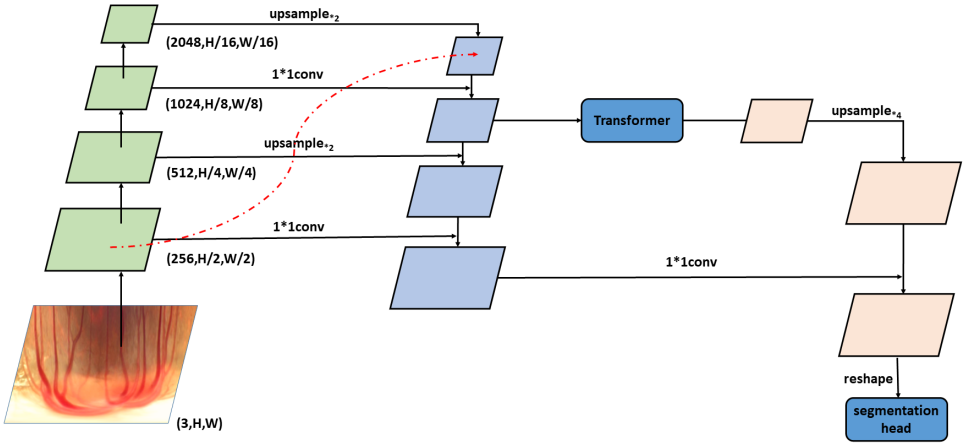


Figure 4. An overview of the CNN-transformer module.

When feeding the input to the transformer, due to its characteristics, the spatial resolution of the feature map can be kept unchanged, and thus the output feature map is also 1/8 the size of the input images. The problem is that segmentation is not possible at this spatial resolution; therefore, we have to upsample the feature maps and find the **output FPN**. The process is as follows:

$$\boldsymbol{P}^{out} = upsample_{\times 2}(\boldsymbol{P}_i^{in}) + Convi(\boldsymbol{P}_{i-1}^{in}), \tag{3}$$

$$\boldsymbol{g}^{out} = upsample_{\times 4}(\boldsymbol{g}^{in}) + Convi(\boldsymbol{P}^{out}). \tag{4}$$

3.2.2. Transformer Module

We obtain the local image features using the CNN backbone and the feature pyramid network. The transformer module builds global contextual information based on them. The transformer part is a conventional encoder–decoder structure to learn contextual information for the segmentation. In order to extract image features, [33,34] use a vision transformer [24] as the encoder. Inspired by these, our transformer module is followed [24] using a vision transformer, which consists of a number of stacked transformer layers. Each layer calculates the paired interaction between input units and outputs the upper and lower contextual $\boldsymbol{X}_{out}$ from the same number of the unit. Before inputting the transformer, the CNN-transformer module outputs the corresponding image feature maps.

The feature map is then flattened into 1-D patch embedding by adding a positional embedding. The visual features and position codes of each unit: $\boldsymbol{F^s} = \boldsymbol{F^v} + \boldsymbol{E}$. Then, $\boldsymbol{F^s}$ is input to the transformer encoder, which contains L layers of multi-head self-attention. We emphasize the mechanism of self-attention (SA) as the self-attention layer is the most important part of the encoder. The transformer architecture is shown in Figure 5. It contains a query $\boldsymbol{Q}$, a key $\boldsymbol{K}$ and a value $\boldsymbol{V}$ as input and outputs a refined feature as follows:

$$SA(z_i) = Softmax(\frac{q_i k^T}{\sqrt{d_h}})v, \tag{5}$$

where $[\boldsymbol{q},\boldsymbol{k},\boldsymbol{v}] = \mathbf{z}\boldsymbol{W}_{(\boldsymbol{qkv})}$, $\boldsymbol{W}_{(\boldsymbol{qkv})} \in \mathbb{R}^{D_0 \times 3D_h}$ is the projection matrix and vector $\boldsymbol{z}_i \in \mathbb{R}^{1 \times D_0}$, $\boldsymbol{q}_i \in \mathbb{R}^{1 \times D_h}$ are the i^{th} row of $\mathbf{z}$ and $\boldsymbol{q}$, respectively. The output FPN upsamples the transformer's output, and then the segmentation head of the module outputs confidence scores of each class in the mask in polar coordinates. Finally, the image is converted from polar coordinates to the image in Cartesian coordinates.

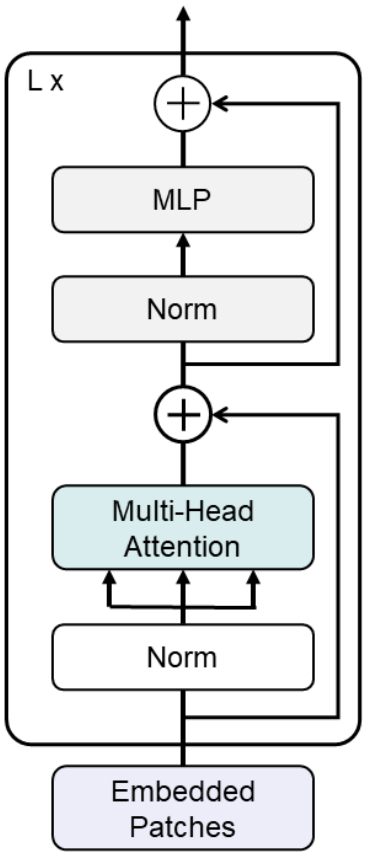


Figure 5. The architecture of the transformer.

4. Experiments and Results

4.1. Datasets and Evaluation Metrics

The experiment was conducted using three public datasets: REFUGE Challenge [35], DRISHTI-GS [36] and RIM-ONE v3 [37]. All images were cropped to a size of 576 × 576 pixels according to the approach proposed by [6]. REFUGE dataset : The REFUGE dataset contains 1200 images in total, which is divided into three parts. There are 400 images that can be used for training, 400 images for validation and 400 images for testing. DRISHTI-GS dataset : The DRISHTI-GS dataset contains 101 images: 51 images are randomly selected for training and 50 images for testing. RIM-ONE v3 dataset: There are 159 images in the data set. In this paper, we randomly selected 99 images for training and 60 images for testing.

Evaluation Metrics

Our test set was evaluated by the standard Dice coefficient as the evaluation metric. The overlap of the algorithm segmentation results and the ground truth labels was measured using the Dice score. For each image, we calculated the prediction result of the OD and OC Dice scores.

$$Dice = \frac{2|X \sqcap Y|}{|X| + |Y|}, \tag{6}$$

where X is the ground truth and Y is the prediction result.

4.2. Implementation Details

Each model was trained using PyTorch 1.10.0 on the NVIDIA GeForce RTX 3090 GPU. For all networks, the batch size was 4. In the comparative experiment, the learning rate for the SETR [33], TransU-Net [34] and our model was 0.0002 and for the other models was 0.0001. The iterations of all models are 10,000 iterations (27 epochs). After each epoch, we store the model with the best validation loss by using checkpoints.

In order to solve the problem of poorly trained deep convolutional neural network models due to small amounts of data, we first trained the network on ImageNet [38] and then fine-tuned the model on a small-scale dataset. In this way, we significantly reduced the time needed to train the model and achieve better results, initializing the weights of the convolutional layer with the pre-trained weights of our CNN backbone. For the training, we used a combinational average of pixel-wise cross-entropy loss and dice loss.

4.3. Comparisons with the State of the Art

We compared Polarformer with other methods, including U-Net variant methods, which utilize skip connections, and transformer-based methods, which introduce global contextual information. The methods with U-Net and variants include U-Net [39], U-Net++ [40], U-Net3+ [41] and Attention-based U-Nets [42]. The transformer-based methods include SETR [33] and TransU-Net [34]. We compared with BGA-Net [43], NENet [44], PraNet [45], M-Net [6], nnU-Net [46], which were proposed for OD and OC segmentation tasks. We compared DeepLabV3+ [47] as well. We compared them using RIM-ONE v3, REFUGE and DRISHTI-GS with their released source codes.

Table 1 tabulates the Dice scores of Polarformer compared with the state-of-the-art methods. From the table, it can be seen that our Polarformer achieved better results. SETR [33] and TransU-Net [34] are transformer-based methods. In terms of accuracy, we can see that these methods are slightly inferior to our method, indicating that more detailed features are learned under polar representation. The shape of the OC and OD are non-rotated ellipses, and their features contain structural information. Our model can learn more discriminative features of the OD and OC by learning long-range dependencies.

Furthermore, our proposed LPTM automatically globally optimizes the spatial relationship between the OD and OC and can be explicitly modeled in a prior way. In addition, our model has a greater advantage compared with M-Net [6]. The reason is that the polar transformation in [6] is a hyper-parameter that may fall into local optima. Clearly,

we observe that our Polarformer demonstrates the competitive performance of the OC segmentation compared with DeepLabV3+ [47] and U-Net [34], even if there is difficulty distinguishing and there is noise in the images.

Table 1. Comparisons of our approach compared with different state-of-the-art methods on the DRISHTI-GS dataset, RIM-ONE v3 dataset and REFUGE dataset.

Method	RIM-ONE v3		DRISHTI-GS		REFUGE	
	$Dice_{cup}$	$Dice_{disc}$	$Dice_{cup}$	$Dice_{disc}$	$Dice_{cup}$	$Dice_{disc}$
U-Net	0.837	0.948	0.830	0.945	0.835	0.951
U-Net3+	0.843	0.955	0.833	0.952	0.837	0.959
DeepLabV3+	0.857	0.961	0.842	0.951	0.855	0.943
AttU-Net	0.852	0.965	0.845	0.950	0.857	0.964
M-Net	0.862	0.952	0.859	0.948	0.864	0.952
PraNet	0.856	0.961	0.841	0.953	0.857	0.966
nnU-Net	0.865	0.966	0.862	0.960	0.876	0.965
SETR	0.877	0.965	0.880	0.954	0.878	0.955
TransU-Net	0.874	0.954	0.883	0.944	0.877	0.964
BGA-Net	0.872	0.967	0.898	0.975	✗	✗
NENet	✗	✗	0.840	0.963	✗	✗
Polarformer(R101)	0.888	0.968	0.893	0.974	0.890	0.975
Polarformer(eff-B4)	0.895	0.972	0.901	0.977	0.892	0.974

Figures 6–8 show the comparison between our method and other methods on REFUGE dataset, DRISHTI-GS dataset and RIM-ONE v3 dataset of the same image. The figures' black parts represent the OC , and the gray parts represent the OD. By comparing the ground truth and our visualization results in the figures, we can see that the segmentation results of different models have differences, particularly in the edge of the OD and OC. From the visualization of the experimental results, DeepLabV3+ had a better effect on the boundary segmentation of the OC. AttU-Net showed results that were not ideal. The U-Net methods were inaccurate , and the boundaries were chaotic. Compared with our method, this is because the method in this paper adopts the transformer module to take advantage of its unlimited effective receptive fields to introduce global contextual information, which makes the boundary segmentation more accurate.

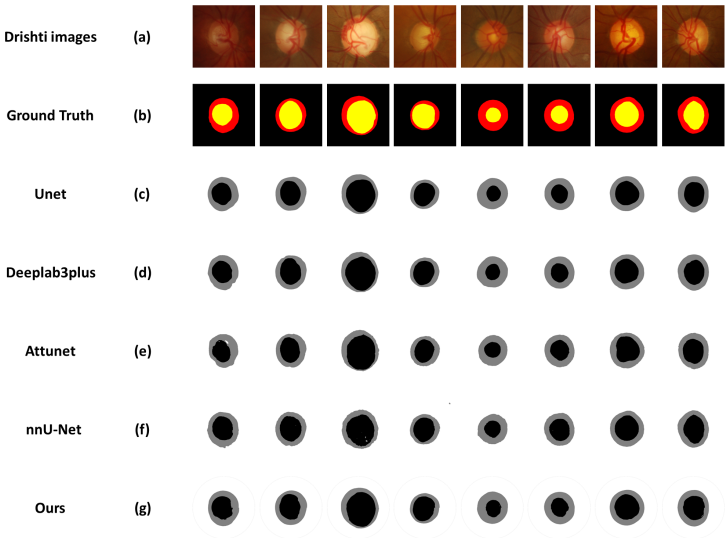


Figure 6. Visualization of OD and OC segmentation results on the DRISHTI-GS dataset. From top to bottom: (**a**) DRISHTI-GS images. (**b**) The ground truth. (**c**) The segmentation results of U-Net. (**d**) The segmentation results of DeepLabV3+. (**e**) The segmentation results of AttU-Net. (**f**) The segmentation results of nnU-Net. (**g**) The segmentation results of our Polarformer.

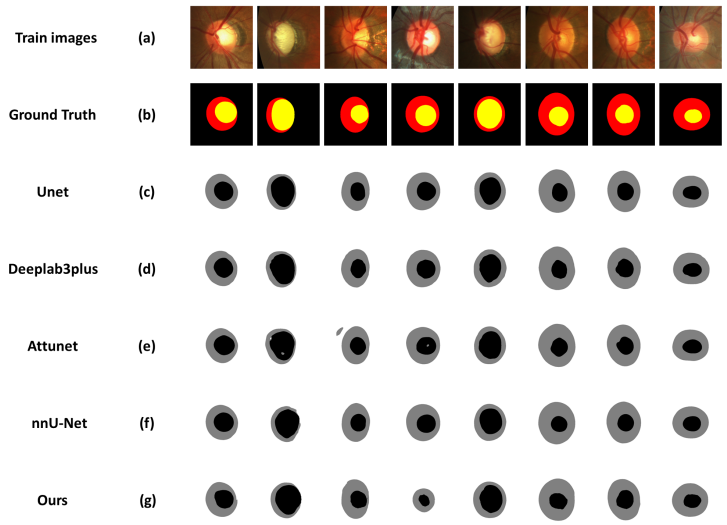


Figure 7. Visualization of OD and OC segmentation results on the REFUGE dataset. From top to bottom: (**a**) REFUGE images. (**b**) The ground truth. (**c**) The segmentation results of U-Net. (**d**) The segmentation results of Deeplabv3plus. (**e**) The segmentation results of AttU-Net. (**f**) The segmentation results of nnU-Net. (**g**) The segmentation results of our Polarformer.

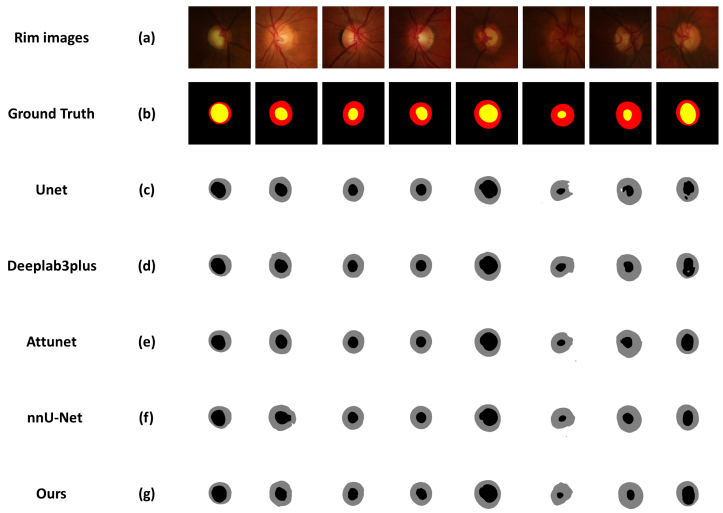


Figure 8. Visualization of OD and OC segmentation results on RIM-ONE v3 dataset. From top to bottom:(**a**) RIM-ONE v3 images. (**b**) The ground truth. (**c**) The segmentation results of Unet. (**d**) The segmentation results of Deeplabv3plus. (**e**) The segmentation results of AttU-Net. (**f**) The segmentation results of nnU-Net. (**g**) The segmentation results of our Polarformer.

4.4. Ablation Study

To further test our module, we performed ablation studies to evaluate the LPTM and Transformer Module as shown in Table 2. We used the OpenCV linear polar transformation implementation to compare with our Polarformer. In our experiments, we took ResNet-101 [48] and EfficientNet-B4 [49] as the CNN backbone. We used three layers of the transformer. In most cases, the Dice score of OD's Dice scores only changed by ± 0.005, so here we only report the Dice scores of OC.

Two conclusions can be made from Table 2. (1) The polar origin detector (POD) achieved higher performance than training without it. LPTM played a crucial role in performance by converting the original images through the POD and polar converter. (2) The combination of the transformer module had a different effect on the task when

compared with the combination without the module. According to the results, we the modules in our Polarformer indicate the contributions of the discriminative features.

Table 2. Ablation study on every module of our proposed method on the DRISHTI-GS dataset, RIM-ONE v3 dataset and REFUGE dataset.

Model	DRISHTI-GS		REFUGE		RIM-ONE v3	
	ResNet-101	Eff-B4	ResNet-101	Eff-B4	ResNet-101	Eff-B4
Baseline	0.881	0.870	0.873	0.876	0.868	0.872
w/o POD	0.887	0.890	0.884	0.887	0.882	0.885
w/o Transformer	0.878	0.882	0.885	0.880	0.879	0.882
Polarformer	0.893	0.901	0.890	0.893	0.888	0.895

4.5. Discussion

In the field of OD and OC image segmentation, the labeled fundus images are small. Overfitting problems often occur when training network models on small datasets. Thus, we use other datasets (e.g., ImageNet [38]) to pre-train the model weights. To investigate the effect of pre-training, we set up a comparison experiment. Table 3 compares the performance. From these results, we observe that the performance of using pre-trained weights showed an increase in Dice scores for both OD and OC.

Table 3. Comparisons of our approach (pre-trained weights) compared with the U-net (training from scratch) on the REFUGE dataset.

Method	REFUGE	
	Dice_{cup}	$\text{Dice}_{\text{disc}}$
U-Net (R101 scratch)	0.825	0.948
U-Net (R101 pretrain)	0.835	0.951
Ours (R101 scratch)	0.881	0.960
Ours (R101 pretrain)	0.892	0.974

5. Conclusions

In this paper, we developed a hybrid CNN-transformer and polar transformation network for OD and OC segmentation. We handled this segmentation task from a fundamentally new perspective, where the OD and OC were represented and segmented in the polar coordinate space rather than in the original image space. Specifically, we combined the CNN network and transformer to explore the polar representation of fundus images with the transformer to exploit contextualized features among all input units to greatly improve each pixel's representational discrimination ability.

The polar origin detector is of great importance for OD and OC segmentation. In addition, a polar converter was introduced to train the neural network and use the polar coordinate transformation of the original fundus datasets, which balances the view cup view disc ratio. As the shape of the OD and OC are elliptical, this transformation results in a reduction in the dimensions. Moreover, we developed a transformer module to take advantage of its unlimited effective receptive fields to introduce global contextual information. After that, we used a feature pyramid network not only to enable fusing different scale features but also to provide a multi-scale prediction for further segmentation tasks.

The performance of our method on three publicly available datasets verified the effectiveness of our approach. Our Polarformer also demonstrated cross-domain ability and had powerful performance in the few-shot scenario, which is a future direction of this research.

Author Contributions: Methodology, Y.F., Z.L., D.Y., H.G. and H.L.; Validation, Y.F., D.Y. and H.H.; Writing–original draft, Y.F.; Visualization, D.Y. and H.H.; Funding acquisition, Z.L., H.G. and H.L. All authors have read and agreed to the published version of the manuscript.

Funding: National Science Foundation of China under Grant: 62241603; Key Research and Development Program of Ningxia Hui Autonomous Region: 2022BEG03158; National Science Foundation of Ningxia under Grant: 2022AAC05006.

Institutional Review Board Statement: This study did not require ethical approval.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data are available in the public domain.

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

References

1. Almazroa, A.; Burman, R.; Raahemifar, K.; Lakshminarayanan, V. Optic disc and optic cup segmentation methodologies for glaucoma image detection: A survey. *J. Ophthalmol.* **2015**, *2015*, 180972. [CrossRef] [PubMed]
2. Lin, L.; Wang, Z.; Wu, J.; Huang, Y.; Lyu, J.; Cheng, P.; Wu, J.; Tang, X. Bsda-net: A boundary shape and distance aware joint learning framework for segmenting and classifying octa images. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 65–75.
3. Aquino, A.; Gegúndez-Arias, M.E.; Marín, 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. [CrossRef] [PubMed]
4. Bekkers, E.J.; Loog, M.; ter Haar Romeny, B.M.; Duits, R. Template matching via densities on the roto-translation group. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 452–466. [CrossRef] [PubMed]
5. Chakravarty, A.; Sivaswamy, J. Joint optic disc and cup boundary extraction from monocular fundus images. *Comput. Methods Programs Biomed.* **2017**, *147*, 51–61. [CrossRef] [PubMed]
6. 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. [CrossRef]
7. Li, S.; Sui, X.; Luo, X.; Xu, X.; Liu, Y.; Goh, R. Medical image segmentation using squeeze-and-expansion transformers. *arXiv* **2021**, arXiv:2105.09511.
8. Sevastopolsky, A. Optic disc and cup segmentation methods for glaucoma detection with modification of U-Net convolutional neural network. *Pattern Recognit. Image Anal.* **2017**, *27*, 618–624. [CrossRef]
9. Tan, J.H.; Acharya, U.R.; Bhandary, S.V.; Chua, K.C.; Sivaprasad, S. Segmentation of optic disc, fovea and retinal vasculature using a single convolutional neural network. *J. Comput. Sci.* **2017**, *20*, 70–79. [CrossRef]
10. Cheng, P.; Lin, L.; Huang, Y.; Lyu, J.; Tang, X. Prior guided fundus image quality enhancement via contrastive learning. In Proceedings of the 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), Nice, France, 13–16 April 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 521–525.
11. Huang, Y.; Zhong, Z.; Yuan, J.; Tang, X. Efficient and robust optic disc detection and fovea localization using region proposal network and cascaded network. *Biomed. Signal Process. Control* **2020**, *60*, 101939. [CrossRef]
12. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
13. Liu, Y.; Fu, D.; Huang, Z.; Tong, H. Optic disc segmentation in fundus images using adversarial training. *IET Image Process.* **2019**, *13*, 375–381. [CrossRef]
14. Mohan, D.; Kumar, J.H.; Seelamantula, C.S. High-performance optic disc segmentation using convolutional neural networks. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 4038–4042.
15. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
16. Zhang, Z.; Fu, H.; Dai, H.; Shen, J.; Pang, Y.; Shao, L. Et-net: A generic edge-attention guidance network for medical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, 13–17 October 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 442–450.
17. Yin, P.; Wu, Q.; Xu, Y.; Min, H.; Yang, M.; Zhang, Y.; Tan, M. PM-Net: Pyramid multi-label network for joint optic disc and cup segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, 13–17 October 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 129–137.
18. Salehinejad, H.; Valaee, S.; Dowdell, T.; Barfett, J. Image augmentation using radial transform for training deep neural networks. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), Calgary, AB, Canada, 15–20 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 3016–3020.
19. 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.

20. Zahoor, M.N.; Fraz, M.M. Fast optic disc segmentation in retina using polar transform. *IEEE Access* **2017**, *5*, 12293–12300. [CrossRef]
21. Jiang, R.; Mei, S. Polar coordinate convolutional neural network: From rotation-invariance to translation-invariance. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 355–359.
22. Kim, J.; Jung, W.; Kim, H.; Lee, J. Cycnn: A rotation invariant cnn using polar mapping and cylindrical convolution layers. *arXiv* **2020**, arXiv:2007.10588.
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3058.
24. 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 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
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, BC, Canada, 10–17 October 2021; pp. 10012–10022.
26. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–28 October 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
27. Toshev, A.; Szegedy, C. Deeppose: Human pose estimation via deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1653–1660.
28. Wei, S.E.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional pose machines. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4724–4732.
29. Zhang, F.; Zhu, X.; Dai, H.; Ye, M.; Zhu, C. Distribution-aware coordinate representation for human pose estimation. In Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 7093–7102.
30. Ali, R.; Sheng, B.; Li, P.; Chen, Y.; Li, H.; Yang, P.; Jung, Y.; Kim, J.; Chen, C.P. Optic disk and cup segmentation through fuzzy broad learning system for glaucoma screening. *IEEE Trans. Ind. Inform.* **2020**, *17*, 2476–2487. [CrossRef]
31. Segman, J.; Rubinstein, J.; Zeevi, Y.Y. The canonical coordinates method for pattern deformation: Theoretical and computational considerations. *IEEE TPAMI* **1992**, *14*, 1171–1183. [CrossRef]
32. Jaderberg, M.; Simonyan, K.; Zisserman, A.; kavukcuoglu K. Spatial transformer networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1213.
33. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.; et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890.
34. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
35. Orlando, J.I.; Fu, H.; Breda, J.B.; van Keer, K.; Bathula, D.R.; Diaz-Pinto, A.; Fang, R.; Heng, P.A.; Kim, J.; Lee, J.; et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Med. Image Anal.* **2020**, *59*, 101570. [CrossRef]
36. Sivaswamy, J.; Krishnadas, S.R.; Datt Joshi, G.; Jain, M.; Syed Tabish, A.U. Drishti-GS: Retinal image dataset for optic nerve head(ONH) segmentation. In Proceedings of the 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), Beijing, China, 28 April–2 May 2014; pp. 53–56. [CrossRef]
37. 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 2011 24th International Symposium on Computer-Based Medical Systems, Bristol, UK, 27–30 June 2011; IEEE Computer Society: Piscataway, NJ, USA, 2011; CBMS '11, pp. 1–6.
38. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 248–255.
39. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. *arXiv* **2015**, arXiv:1505.04597.
40. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–11.
41. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.W.; Wu, J. Unet 3+: A full-scale connected unet for medical image segmentation. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1055–1059.
42. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention u-net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999.
43. Luo, L.; Xue, D.; Pan, F.; Feng, X. Joint optic disc and optic cup segmentation based on boundary prior and adversarial learning. *Int. J. Comput. Assist. Radiol. Surg.* **2021**, *16*, 905–914. [CrossRef] [PubMed]

44. Pachade, S.; Porwal, P.; Kokare, M.; Giancardo, L.; Mériaudeau, F. NENet: Nested EfficientNet and adversarial learning for joint optic disc and cup segmentation. *Med. Image Anal.* **2021**, *74*, 102253. [CrossRef] [PubMed]
45. Fan, D.P.; Ji, G.P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; Shao, L. Pranet: Parallel reverse attention network for polyp segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, 4–8 October 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 263–273.
46. Isensee, F.; Jaeger, P.F.; Kohl, S.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **2021**, *18*, 203–211. [CrossRef]
47. Firdaus-Nawi, M.; Noraini, O.; Sabri, M.; Siti-Zahrah, A.; Zamri-Saad, M.; Latifah, H. DeepLabv3+ _encoder-decoder with Atrous separable convolution for semantic image segmentation. *Pertanika J. Trop. Agric. Sci* **2011**, *34*, 137–143.
48. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26–30 June 2016; pp. 770–778.
49. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019; pp. 6105–6114.