technologies

Article

Segmentation of Retinal Blood Vessels Using Focal Attention Convolution Blocks in a UNET

Rafael Ortiz-Feregrino, Saul Tovar-Arriaga *, Jesus Carlos Pedraza-Ortega and Juvenal Rodriguez-Resendiz

Faculty of Engineering, Universidad Autónoma de Querétaro, Santiago de Querétaro 76010, Mexico; rafaortizferegrino@gmail.com (R.O.-F.); carlos.pedraza@uaq.mx (J.C.P.-O.); juvenal@uaq.edu.mx (J.R.-R.)
* Correspondence: saul.tovar@uaq.mx

Abstract: Retinal vein segmentation is a crucial task that helps in the early detection of health problems, making it an essential area of research. With recent advancements in artificial intelligence, we can now develop highly reliable and efficient models for this task. CNN has been the traditional choice for image analysis tasks. However, the emergence of visual transformers with their unique attention mechanism has proved to be a game-changer. However, visual transformers require a large amount of data and computational power, making them unsuitable for tasks with limited data and resources. To deal with this constraint, we adapted the attention module of visual transformers and integrated it into a CNN-based UNET network, achieving superior performance compared to other models. The model achieved a 0.89 recall, 0.98 AUC, 0.97 accuracy, and 0.97 sensitivity on various datasets, including HRF, Drive, LES-AV, CHASE-DB1, Aria-A, Aria-D, Aria-C, IOSTAR, STARE and DRGAHIS. Moreover, the model can recognize blood vessels accurately, regardless of camera type or the original image resolution, ensuring that it generalizes well. This breakthrough in retinal vein segmentation could improve the early diagnosis of several health conditions.

Keywords: retinal blood vessels; artificial intelligence; convolutional neural networks; attention module; segmentation

Citation: Ortiz-Feregrino, R.; Tovar-Arriaga, S.; Pedraza-Ortega, J.C.; Rodriguez-Resendiz, J. Segmentation of Retinal Blood Vessels Using Focal Attention Convolution Blocks in a UNET. *Technologies* **2023**, *11*, 97. https://doi.org/10.3390/technologies11040097

Academic Editor: Pietro Zanuttigh

Received: 1 June 2023
Revised: 29 June 2023
Accepted: 5 July 2023
Published: 13 July 2023

1. Introduction

AI plays a prominent role in various fields, including programming-assistance tools such as OpenIA Copilot [1], protein prediction using Deep Mind and the model in [2], congenital disease prediction [3], and lesion segmentation in medical imaging, as demonstrated by recent research in X-ray imaging for COVID-19 [4]. These examples illustrate how AI can achieve metrics comparable to an expert's, making it a promising solution for automating daily tasks that intelligent algorithms can efficiently tackle.

In recent years, deep learning (DL), a subfield of machine learning (ML), has experienced significant growth within the AI domain. This expansion is well-founded because DL models do not require direct guidance from an expert nor the manual modification of complex hyperparameters to achieve suitable performance. Instead, many previously labeled examples are sufficient to initiate the learning process. This attribute of neural networks enables the exploration of complex domains, such as medicine, without requiring the presence of a domain expert at all times.

In our study, we employed CNN models [5] and attention modules based on those incorporated by visual transformers [6] to segment retinal blood vessels. The retina, a delicate layer at the back of the eye, plays a crucial role in our vision, as its connections go directly to the brain [7]. The segmentation of retinal blood vessels can aid in detecting degenerative diseases such as diabetic retinopathy [7], cardiovascular problems [8], and many other congenital conditions. Unfortunately, these diseases have a high prevalence worldwide [9].

The UNET model has emerged as a popular choice among developers and researchers for addressing segmentation challenges across diverse domains, for example, segmenting lesions caused by diabetic retinopathy [10], the segmentation of brain tumors using a modified model [11,12], and the segmentation of other skin lesions [13]. However, its versatility continues. The model can also segment images taken by UAVs and perform activities such as those described in [14]. These models are built upon the UNET architecture, which is highly regarded for its adaptability and straightforward customization capabilities, rendering it an ideal candidate for exploring novel concepts. Adaptations of the model include adding convolutions and supplementary connections, as demonstrated in [13], and incorporating comprehensive attention blocks, as showcased in the referenced article, to mentioned a few.

Most object prediction, segmentation, and classification models that use images as inputs are primarily based on CNNs [5]. Meanwhile, transformers [15], a type of architecture mainly used in NLP, are now improving the existing metrics by their distinct way of "paying attention". As a result, image-based transformer models [16] have emerged and are proving to achieve comparable, or at times better, results than traditional CNNs. However, these models come with two significant challenges: the requirement of big training data and the intensive computational power they consume during training. Some techniques, such as transfer learning (TL) [17], are used when the dataset is limited. The source and target datasets should have similar domains to maximize TL benefits.

Another commonly used approach is to take the attention mechanism of transformers as an independent module that provides some of the benefits of transformers without requiring a vast amount of training data. The approach that can be seen in [18] combines MHSA with convolutions to generate a bottleneck transformer (BT) that can be viewed as an attention module. This type of implementation using transformers and convolutions is the basis of this work, along with a focus inspired by the focal transformer presented by [19]; this allowed us to improve state-of-the-art metrics using the U-Net network as a backbone and adding modules that we call a "Focal Attention Convolution Block" (FACB).

This study's proposed FACB offers a distinctive characteristic of seamless integration into any CNN as an additional module. This integration does not require altering the backbone of the underlying models, making it similar to a plug-and-play component. This approach provides the advantages of the attention mechanisms seen in vision transformers without necessitating extensive modifications to the base model.

The FACB consists of two main parts. Firstly, the initial stage of the FACB captures information from the input data at various levels, which we refer to as windows. These windows can have different dimensions, such as small, medium, or large, and can be singular or multiple. Ultimately, the output of these windows is concatenated, providing information from different regions of the input data. This concatenated information is then passed to the second block of the FACB, which comprises several attention modules operating in parallel to process the input information. The versatility and compatibility of the FACB with matrix operations enable its implementation at any stage within a CNN architecture. This flexibility allows researchers to seamlessly incorporate the FACB module into existing CNNs, enhancing the network's capabilities without significant structural changes.

The remainder of this paper is organized as follows. Section 2 presents the proposed method's main idea, the databases used, the preprocessing, data augmentation, and details about the FACB and its use in the UNET model. Section 3 presents a comparative table with the state of the art and images of the inference of the proposed model. Finally, the discussion and conclusions are presented in Sections 4 and 5, respectively.

In recent years, there have been significant advancements in the field of medical imaging, particularly in the area of retinal segmentation. Retinal segmentation is the process of identifying and separating different structures within the retina, which can be critical for diagnosing and treating a variety of eye diseases.

The automated segmentation of blood vessels has long been a challenging task in computer vision and artificial intelligence research. Among various AI approaches, ML models have proven effective in segmenting and classifying these structures. There has been a growing adoption of DL models in recent years in this domain [20]. DL models offer a significant advantage in generalizing across diverse domains [21]. Talking specifically about DL, vision transformers [22] are becoming important in the area but the ideal conditions to apply these models are seldom present. Instead of using a complete transformer, we can take the part that pays attention. Attention modules in a convolutional network allow for capturing local and global spatial relationships in images more efficiently. Unlike full transformers, which require computing all interactions between pair elements, attention modules in a convolutional network can selectively focus on relevant regions and reduce computational complexity. This is especially beneficial for images, where visual information is highly structured and spatial relationships between pixels play a crucial role. Attention modules in a convolutional network also enable the better interpretation and visualization of results by providing attention maps highlighting the image's most important regions. In addition, attention modules can be easily incorporated into existing convolutional network architectures, facilitating their implementation and leveraging both approaches' benefits. Attention modules in a convolutional network offer an efficient and effective way to model spatial relationships in images, transcending computational limitations and taking advantage of the intrinsic visual structure of images.

Khanal, A. et al. [23] proposed a stochastic training scheme for deep neural networks that robustly balances precision and recall. Their method yielded a better balance of precision and recall relative to state-of-the-art techniques, resulting in higher F1 scores. However, their method can be misleading for unbalanced datasets. Gegundez-Arias, et al. [24] present a new method for vascular tree segmentation. The method outperforms other U-Net-based methodologies in terms of accuracy, requiring fewer hyperparameters and lower computational complexity. However, the major limitation of the practical integration is the limited number of examples available for network training. Galdran, A. et al. [25] reflect on the need to construct algorithmically complex methodologies for retinal vessel segmentation. It suggests that minimalistic models, adequately trained, can attain results that do not significantly differ from more complex approaches. The authors suggest that research should switch to modern high-resolution datasets rather than rely on old datasets.

In the work of Tang, P. et al. [26], the main contributions include a novel ensemble model based on multiproportional red and green channels that outperform other existing methods concerning two primary performance metrics (segmentation accuracy and AUC) and the first instance of the red channel providing performance gains. In the work of Ma, Y. et al. [27], WA-Net was developed to improve the segmentation accuracy of retinal blood vessels. Cross-training between datasets was performed to verify the model's generalization performance, and the results showed that WA-Net extracts more detailed blood vessels and has a superior performance. However, there are still some limitations, such as the need for more effective data augmentation and a long computational time due to the introduction of weight normalization.

Tuyet, V.T.H. et al. [28] proposed a method for retinal vessel segmentation using three periods: a salient edge map in the retinal vessel image, feature extraction using CNN in a salient map, and segmentation based on the pixel level of the Sobel operator in saliency. The Jaccard index value of the proposed method was found to be higher than other approaches. Therefore, the number of layers or operators for the salient region map can be improved in the future. Park, K.-B. et al. [29] proposed M-GAN, which outperformed previous studies with respect to accuracy, IoU, F1 score, and MCC. It derived balanced precision and recall together through the FN loss function. The proposed method with an adversarial discriminator showed better segmentation performance than a method without a discriminator.

Compared to state-of-the-art methods, the proposed method of Zhuo, Z. et al. [30] achieved extremely competitive performances on the DRIVE and STARE datasets. The

results of cross-training show that the method has strong robustness and is faster than other CNN-based algorithms. However, the proposed method does not have a network structure that reduces the number of parameters while guaranteeing effective segmentation. Zhuo, Z. et al. [31] present a novel retinal vessel segmentation architecture that combines a U-Net with generative adversarial networks and a weighted feature matching loss. This architecture was evaluated on three retinal segmentation datasets (DRIVE, CHASE-DB1, and STARE). It showed improved performance compared to previous methods, with higher confidence scores, F1 score, sensitivity, specificity, accuracy, AUC-ROC, mean-IOU, and SSIM. However, the model suffers from a high false-positive rate.

2. Materials and Methods

Demonstrating the generalization of an artificial neural network model is a crucial aspect of its training. It requires diverse data to ensure the model recognizes and learns from patterns outside the training dataset. To achieve this, we utilized many publicly available datasets, each presenting unique characteristics in terms of their composition, such as variations in image resolution, centering, and color saturation, as depicted in Figure 1. These differences result from the use of various devices and the work of different individuals, resulting in images that differ significantly while all within the same domain (the retina).

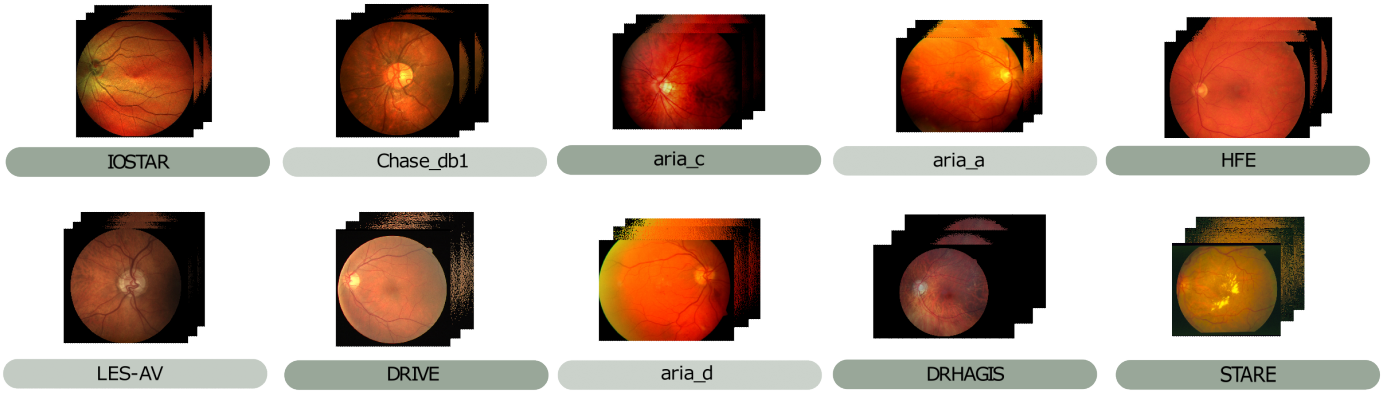

Figure 1. Although the images belong to the same retina domain, these datasets possess unique properties, as demonstrated by differences such as varying degrees of centering toward the optic nerve or the macula, mixed image resolutions, and differences in color saturation and brightness levels.

The blocks of a general methodology using DL models are very similar regardless of the task to be performed. In our case, as can be seen in Figure 2, the general development of our implementation, which contains the classic blocks of preprocessing and training/validation but inside each block, has specific processes.

2.1. Preprocessing

One of the main reasons data preprocessing is essential is the nature of deep neural networks. These networks are designed to learn functional patterns and representations from input data. However, the quality of the input data can vary widely and may contain noise, redundancy, outliers, or even irrelevant information. Preprocessing helps mitigate these problems by performing normalization, denoising, and the reduction of dimensionality. It is crucial to find a balance in data preprocessing and not to exaggerate when homogenizing the data since a model also benefits from the natural noise with which the data were taken.

2.1.1. Data Cleaning

The preprocessing block can be considered the kitchen where the food of our model is prepared. It is fundamental when handling various datasets because if garbage enters the model, we cannot expect good results no matter how robust the architecture is. Therefore, we must at least clean the data and try to normalize it as much as possible while maintaining the essence of each database.

The images contained in the datasets are very different in their composition, with some having a better resolution than others. In addition, the difference in saturation in the red channel is noticeable, so finding a balance when preprocessing the images is vital since we must consider the diversity of these images. RGB retinal images usually tend to have a reddish hue in their composition, which causes a considerable saturation in the red channel. The blue channel has the slightest presence in the image and can be considered the opposite of the red channel. Finally, the green channel is the one that best preserves the balance of all the channels, thus providing more precise and sharper information. Therefore, it was decided to only use the green channel. To keep the original information of the images the same, we chose to perform histogram equalization on the channel we used, i.e., the green one. After the equalization, we looked for an algorithm that would improve the sharpness and brightness without altering the images too much, so according to its performance, we decided to use CLAHE [32], which gives us a general enhancement of the image without altering too much of the original information of the images. The preprocessing applied to an image can be seen in the central block in Figure 2.

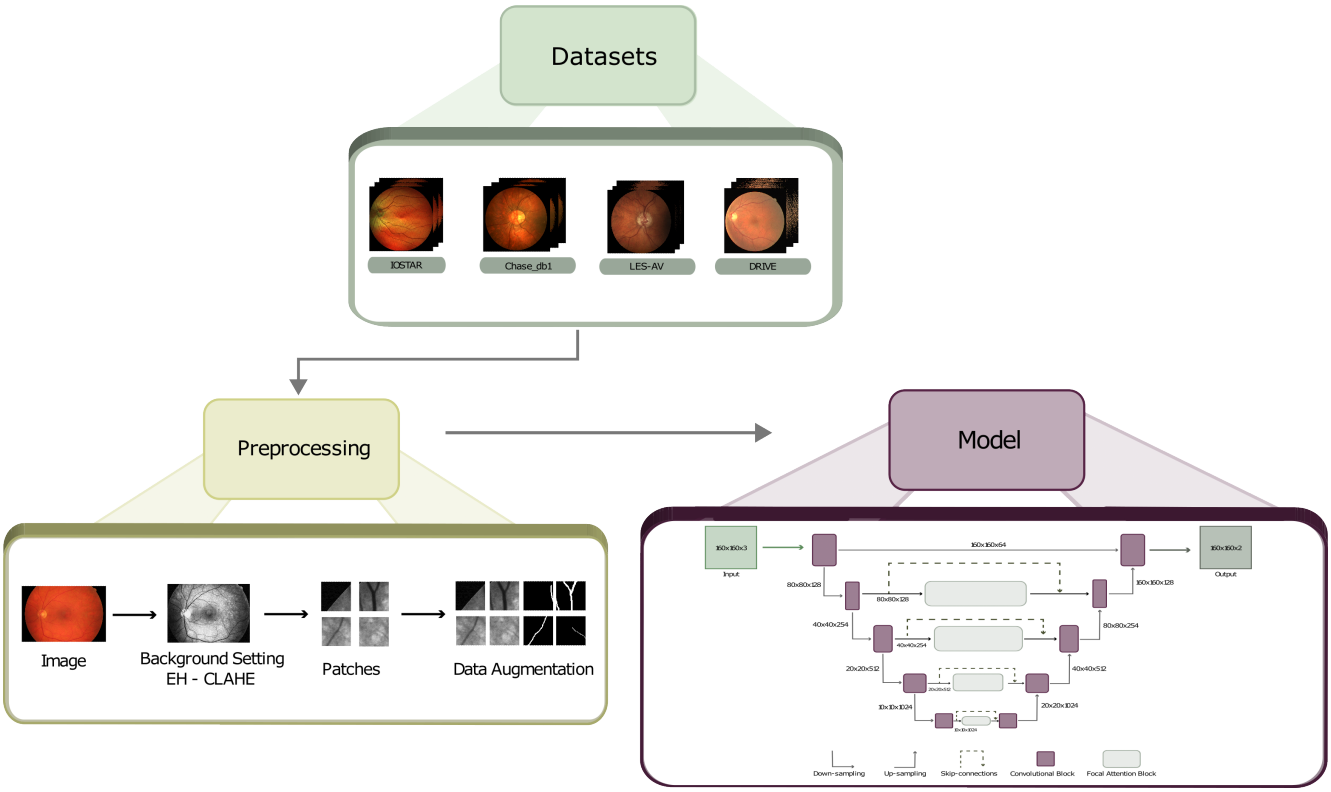

Figure 2. General methodology. The first block refers to the collection of the datasets. The central block represents the preprocessing applied to each of the datasets. The last block refers to the training and validation of the proposed model.

2.1.2. Data Augmentation

A disadvantage of public datasets is the small number of examples they contain, so an increase in data has to be implemented. For this reason, we divided each image into small segments called patches with a size of (160 × 160 × 3) and obtained 25,402 patches

for training and 2900 for validation, trying to maintain a ratio of 90–10. An example of the patches can be seen in Figure 2.

In addition to the generation of patches as a method of data augmentation, it was decided to perform an automatic augmentation prior to model input. This means that before delivering an example to the model, it goes through a process that can rotate the image, displace it, increase its color saturation, add noise, and remove a part of the image, in order to force the model to generalize as much as possible and avoid overtraining. However, this process does not always happen for all the input examples since the aim is to generate the most significant difference between each data point. The decision as to which items undergo this process and which do not is randomly determined for each epoch.

2.2. The Model

When addressing specific tasks in the medical domain, such as disease prediction in medical imaging or lesion segmentation, choosing a good DL model becomes critical. A well-selected model has the potential to improve the accuracy of diagnosis and treatment, which, in turn, can lead to better clinical outcomes and more effective medical care.

2.2.1. Model Overview

As mentioned before, transformers are becoming very relevant in work with images, despite how inconvenient their training can be. Thus, using small modules that provide some of the benefits is a viable approach. In this work, the principal idea is the incorporation of a module consisting of main blocks. The first one is in charge of extracting as much information from the input as possible using different regions. These regions start from the center of the input and can extend to the edges; the idea is to collect as much information in the input directions, which contain valuable information from different points, and the second block pays attention with its different attention modules to the information the first block provides. The general idea of the focal attention convolution block (FACB) can be seen in Figure 3.

Figure 4 shows the block called CBAM bottleneck multihead self attention (CB-BMHSA), which incorporates an attention module inspired by the focal attention (FT) [19], which delivers fine and granular information to an MHSA. For our implementation, we took the FT idea of generating information in different regions of the image in parallel and used a CBAM [33] attention module to obtain spatial and channel information and then pass the information to a BT [18], which is a variant of the multihead self attention (MHSA); however, this was designed to work with convolutions, which means that there is no need to flatten the input features as a traditional MHSA would need.

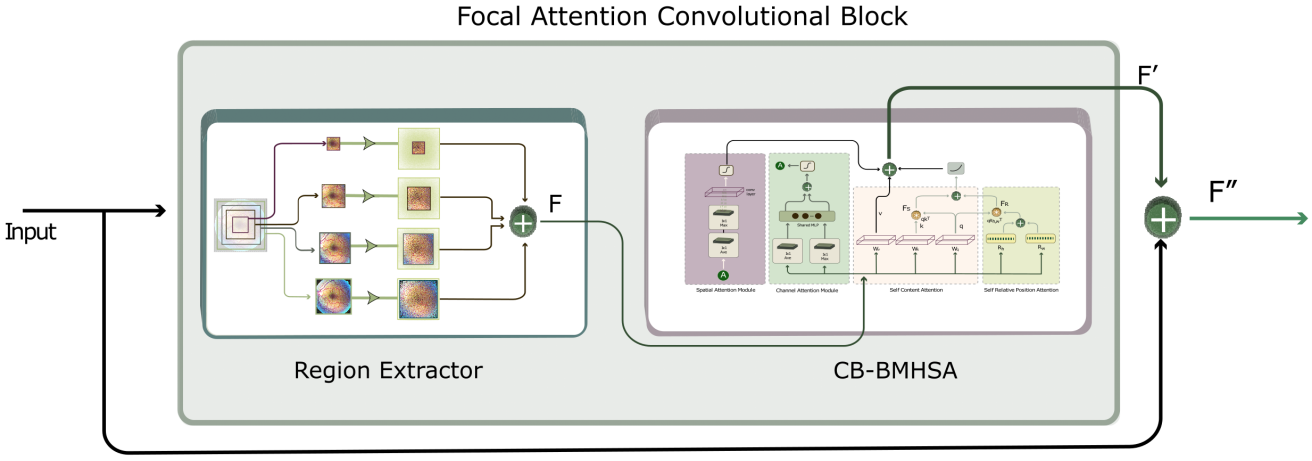

Figure 3. General FACB that contains two principal blocks.

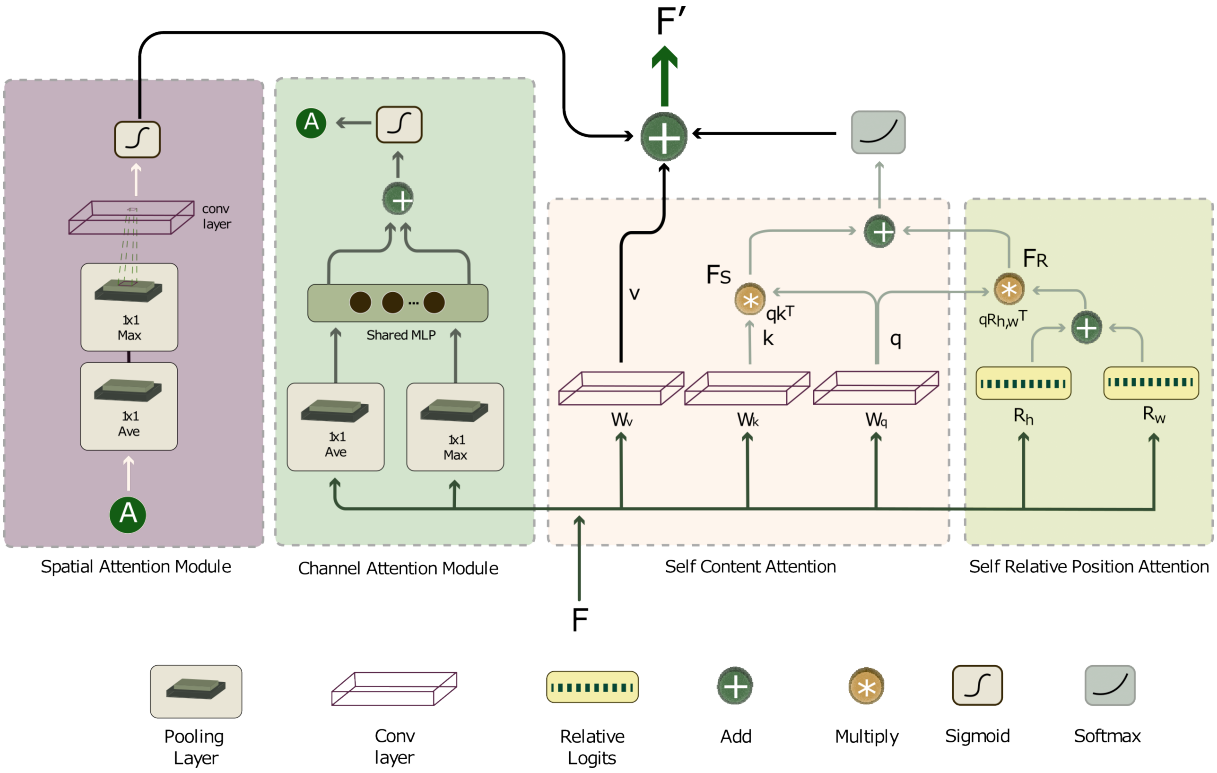

Figure 4. The CB-BMHSA attention block.

The input information of the CB-BMHSA comes from a block that we call the region extractor (RE). This block was inspired by the FAB used in the transformer in [19]. The block presented in the article, as mentioned earlier, has the disadvantage that it does not work with convolutions because it is 100% adapted to the use with transformers, so we decided to generate an alternative that can be coupled to the use of CNNs. In Figure 5, we can observe this block in more detail. The RE obtains the information from a window called Pr, which delimits the region of information to be extracted from the input features. Then, the window data are grouped into Pw sets using AvgPooling to generate subwindows with information from different regions of the input that will be summed to have general and granular details. Finally, when summing each subwindow, we use zero padding to avoid dimensionality problems. This procedure can be repeated for N number of windows as shown in Equation (1). It should be noted that FACB was designed for features with square dimensions. In addition, the FACB can be implemented on any CNN, extracting information from different regions of the input and paying attention to that information using different attention modules.

$$\sum_{n=1}^{N} ZeroPadding(AvgrPooling(Pr^{[n]})) \quad (1)$$

FACB can be implemented in any part of a traditional CNN, as the only requirement is that the input dimensions are square (m × m). Typically, classical CNN models always use square inputs and outputs, making the FACB an enhancer that allows the ability to pay attention to be added to models such as RESNET-50, RESNET-100 [34], VGG-16 [35], EfficientNet [36], Mobilnet [37], GoogleNet [38], and of course, the model that we use as a base in this article, the U-NET [39]. The U-NET was chosen because it is a well-studied model that has a wide variety of variants. These models have been applied to the segmentation of lesions, blood vessels, and other parts of the retina, which are components that can be used to compare and demonstrate that the FACB helps to enhance the performance of the model without modifying the main skeleton. The final model can be seen in Figure 6, to which FACB was added in the intermediate connections from the second downsampling.

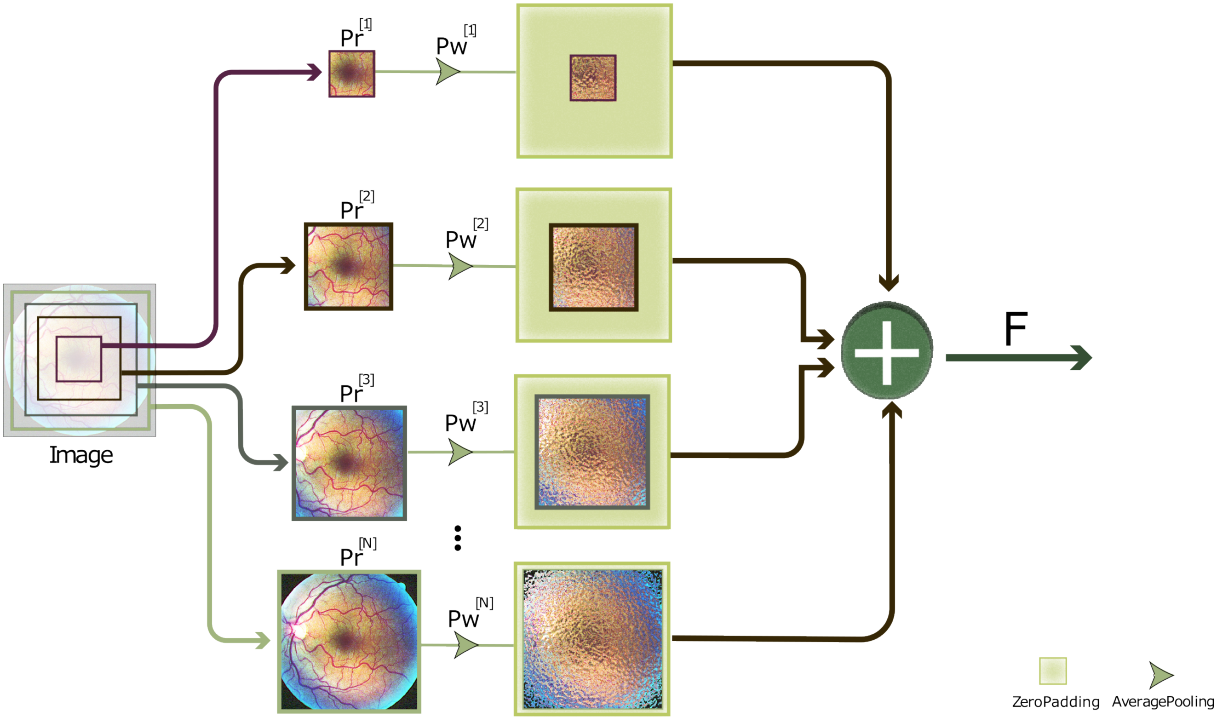

Figure 5. RE block that feeds the CB-BMHS block.

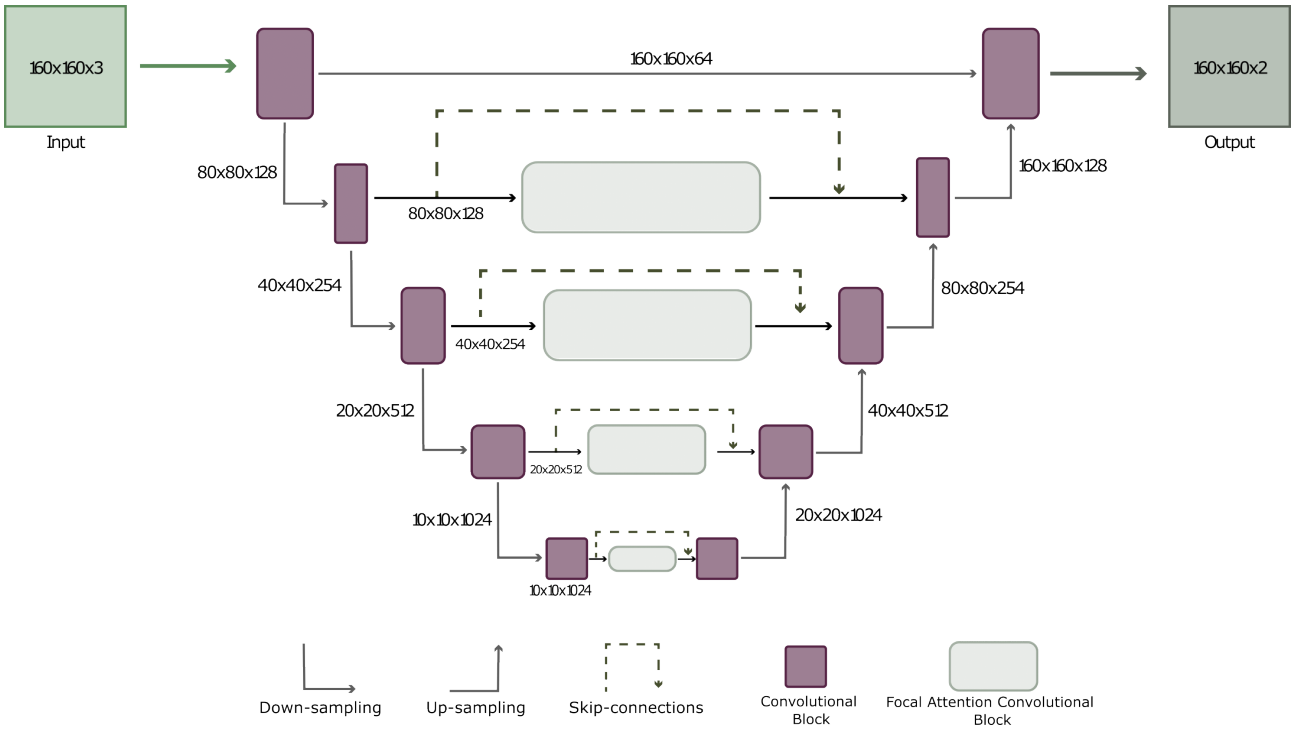

Figure 6. The complete U-Net architecture using FACB at the four points downstream.

2.2.2. UNET with FACB

When training a transformer either for natural language processing or a visual transformer for working with images, we know beforehand that this will require high computational power and a large amount of data. Similarly, adding attention modules increases the processing time of the model: not to the level of a transformer but more than that of a conventional CNN. Therefore, different configurations of the U-NET were tested

with the FACB to determine the cost benefit. In this way, we noticed that the results of the metrics were similar when using the FACB in the whole model as to when it was applied from the second output of the convolution block; however, the computational cost changed a lot, i.e., when using the first option, the training time increased by 50%.

3. Results

The model was evaluated with different metrics that would give us important information about its performance. Several authors only consider metrics such as accuracy or AUC, i.e., metrics that, for this task, do not reflect the reality of the prediction since the number of pixels in the images that represent an absence of blood vessels is much higher than the number of pixels that contain blood vessels. This means that, although our model did not detect any blood vessels in the image, the accuracy rate and the sensitivity were high, since it focuses on the true negatives, which are more abundant than true positives. Thus, the accuracy, recall, and F1 score better reflect the model's performance speaking exclusively for this task. Table 1 shows a comparison of our model and the bests in the state of the art. As can be seen, not all the models were trained on several datasets, and some were only trained on a single dataset, leaving uncertainty as to whether the model would generalize or would only perform well on that dataset.

Table 1. The table shows a comparison of the best models and our proposal. As seen at least in one metric, the proposed model achieved better results than the state of the art. Moreover, the table indicates a good generalization over all datasets. The best results are in bold.

Dataset	Author	Accuracy	AUC	Precision	Recall	Specificity	F1
DRIVE	Park, K.-B. et al. [29]	97.06	98.68	83.02	83.46	98.36	83.24
	Galdran, A. et al. [25]	-	98.1	-	-	-	-
	Chen, D. et al. [40]	96.22	**98.78**	-	**85.76**	**99.32**	81.60
	UNET with FACB	**97.9**	93.6	**91.7**	88.1	99.0	**89.9**
CHASEDB1	Park, K.-B. et al. [29]	97.36	98.59	-	-	-	81.1
	Galdran, A. et al. [25]	-	98.47	-	-	-	-
	Chen, D. et al. [40]	98.12	**99.25**	-	84.93	**99.66**	82.73
	UNET with FACB	**99.1**	97.0	94.8	**94.4**	99.5	**94.6**
HRF	Park, K.-B. et al. [29]	97.61	**98.52**	79.72	-	-	79.72
	Galdran, A. et al. [25]	-	98.25	-	-	-	-
	Tang, P. et al. [26]	96.31	98.43	-	76.53	98.66	77
	UNET with FACB	**97.7**	91.3	**87.2**	**83.8**	**98.9**	**85.4**
STARE	Park, K.-B. et al. [29]	**98.76**	**99.73**	84.17	83.24	**99.38**	83.7
	Galdran, A. et al. [25]	-	98.28	-	-	-	-
	Chen, D. et al. [40]	97.96	99.53	-	87.93	99.37	88.36
	UNET with FACB	97.9	93.6	**91.7**	**88.1**	99	**89.9**
LES-AV	Galdran, A. et al. [25]	-	**97.34**	-	-	-	-
	UNET with FACB	99.3	97.1	94.7	94.6	99.6	94.6
IOSTAR	Guo, C. et al. [41]	97.13	**98.73**	-	80.82	98.54	-
	Li, X. et al. [42]	95.44	96.23	-	73.22	98.02	-
	Wu, H. et al. [43]	97.06	98.65	-	82.55	98.30	-
	UNET with FACB	**99.3**	97.1	94.7	**94.6**	**99.6**	94.6
ARIA (mean)	Tajbakhsh, N. et al. [44]	-	-	-	-	-	72
	UNET with FACB	97.3	89.9	86.4	81.4	96.1	**83.2**

The metrics that interest us, as mentioned above, are precision, recall, and F1 score. As can be seen, our model achieved superior results as compared to the state-of-the-art models in almost all the datasets for these metrics. The only datasets for which the model did not achieve similar numbers to the others were Arias A-C-D, but this reaffirms what we mentioned earlier: the accuracy and AUC for these datasets are competitive metrics

but do not reflect the performance of the model since the rest of the metrics had much lower results.

In Figures 7 and 8, we can observe the inference on an image of each dataset. This inference shows us the outstanding performance of our model in the most challenging regions, such as the areas closest to the optic nerve, the bifurcations and intersections, the points where the path ends, the centralization towards the optic nerve or the macula, and the thinnest blood vessels and arteries. The comparison column provides valuable information on the differences between the prediction and the actual output. The yellow color represents the true positives of the segmentation, the red pixels are the false positives, and the green pixels are the false negatives. This final column allows us to directly compare the results in Table 1 and the model prediction. If we only look at the prediction and mask columns individually, it is difficult to infer whether the result is consistent with what is presented in the metrics output.

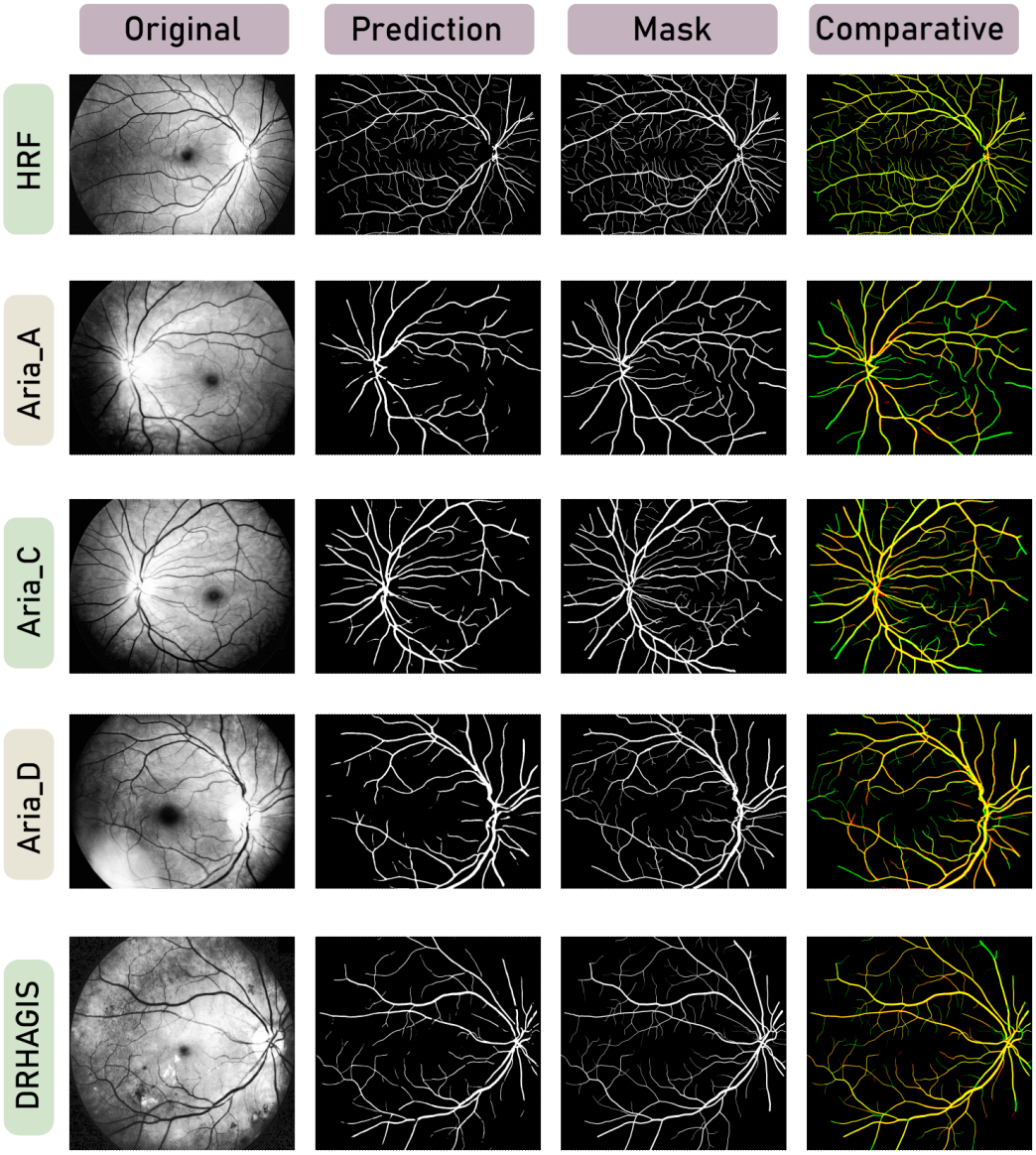

Figure 7. Model predictions part 1.

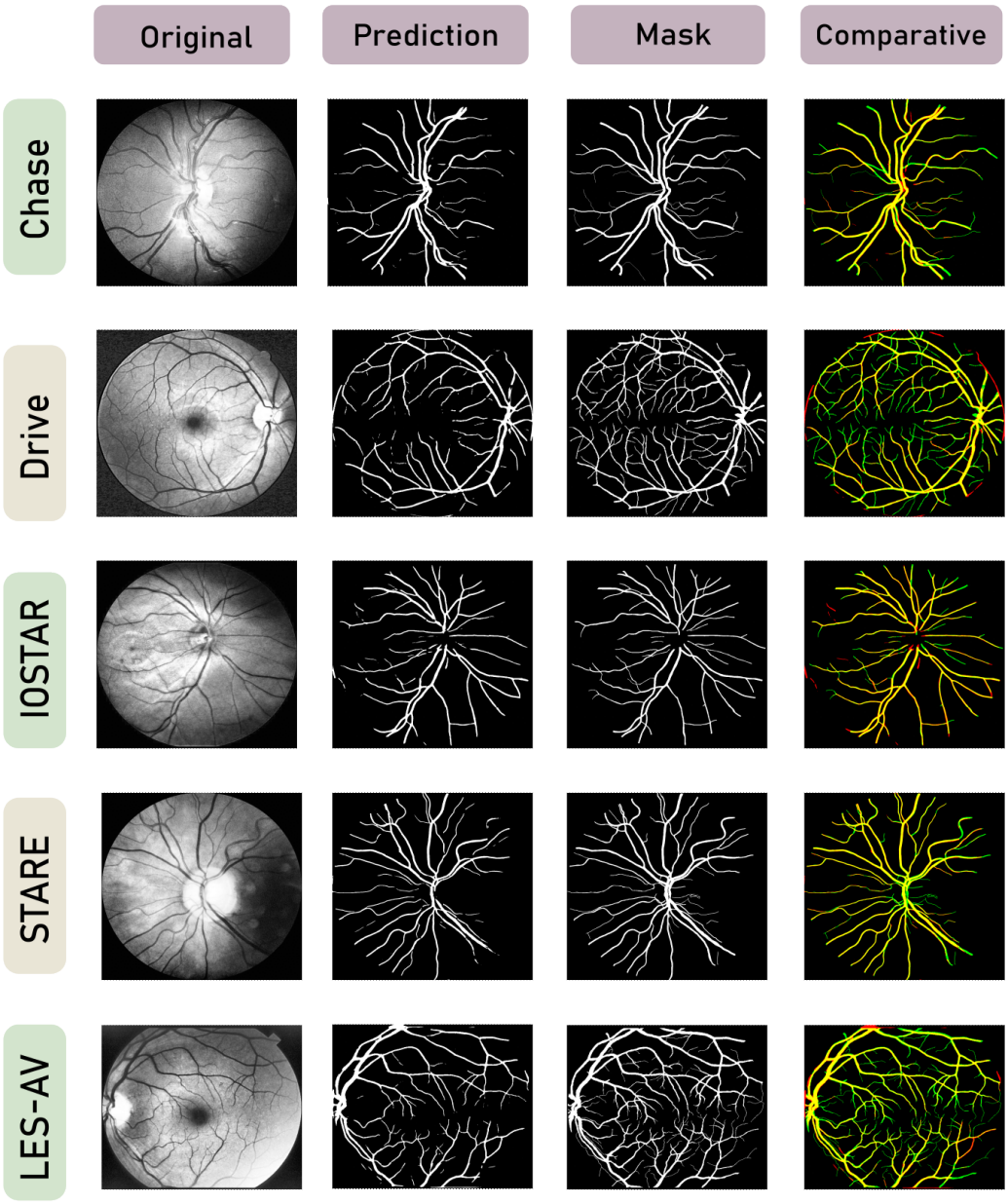

Figure 8. Model predictions part 2.

4. Discussion

The results of this paper indicate that the model has excellent performance in the most challenging regions of the datasets, such as the areas closest to the optic nerve, the bifurcations, and the intersections. This indicates that the FACB helps the UNET model pay attention to the segmentation's critical areas. However, the accuracy and AUC metrics for the Arias A-C-D datasets were lower than in the other datasets, reaffirming that these metrics only sometimes reflect the model's performance. Nevertheless, the inference on the images also showed outstanding results, indicating that the model can accurately detect blood vessels and arteries even in the thinnest areas. Further research should focus on exploring the implications of these results and investigating potential future research directions. There are two crucial considerations to bear in mind when utilizing the FACB. Firstly, incorporating multiple FACBs into neural networks results in a substantial increase in memory requirements. Consequently, integrating such models becomes feasible with

access to robust hardware resources. Secondly, training the model on high-resolution images exceeding 200 pixels poses a formidable challenge due to the corresponding surge in memory demands. To tackle this issue, it is vital to explore strategies, such as the one proposed in this article, which involve partitioning images into smaller patches. This approach enables the successful implementation of FACBs by mitigating the excessive memory demands associated with larger image resolutions.

5. Conclusions

This research shows that our model performs in the most challenging regions for retinal vessel segmentation. This study has implications for previous retinal vessel segmentation studies and provides potential future research directions. This approach in retinal vein segmentation could improve the early diagnosis of several health conditions. Moreover, the model's ability to recognize blood vessels accurately regardless of camera type or the original image resolution suggests that it generalizes well and could be used in various applications. This was achieved by combining the existing attention modules and the regional extractor, and combining the best of the two classic convolutional networks and care delivery approaches. Another significant consideration is that some datasets are more challenging than others. They demand the use of all the possible datasets available to train and evaluate the model in order to achieve generalization. It is possible to enhance segmentation performance by conducting fine-tuning on each dataset. However, the ultimate objective is to achieve generalizability regardless of the image type presented to the model.

The overarching goal for future research is to achieve robustness and generalizability regardless of the image type presented in the model. Continued efforts in fine-tuning techniques, dataset augmentation, and advancements in model architectures and attention mechanisms can contribute to developing highly accurate and versatile DL models for medical image analysis.

Combining the strengths of attention mechanisms and CNNs, we can develop models that capture relevant features, focus on informative regions, improve interpretability, and achieve more accurate segmentation results. These advancements can contribute to early disease diagnosis, precision medicine, and improved patient care in medical imaging.

Future research efforts can focus on optimizing the architecture and training strategies for attention-guided CNN models. This may involve exploring different attention mechanisms, designing novel network architectures that effectively fuse attention and convolutional operations, and developing specialized loss functions that encourage accurate vessel segmentation. Additionally, investigating the transferability and generalizability of attention-guided CNN models across different datasets and imaging modalities can further improve their practical applicability in various medical image analysis tasks.

Author Contributions: Conceptualization, R.O.-F.; Methodology, R.O.-F.; Software, R.O.-F.; Validation, S.T.-A. and J.C.P.-O.; Formal analysis, R.O.-F.; Investigation, R.O.-F.; Resources, S.T.-A., J.C.P.-O. and J.R.-R.; Data curation, R.O.-F.; Writing original draft preparation, R.O.-F.; Writing review and editing, S.T.-A., J.C.P.-O. and J.R.-R.; Visualization, R.O.-F. and J.R.-R.; Supervision, S.T.-A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this work are publicly accessible. For more information, see the respective articles of each dataset. In addition, the complete source code can be obtained from the following repository: https://github.com/FereBell/Focal-Attention-Convolution-Blocks (accessed on 15 June 2023).

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

References

1. Dakhel, A.M.; Majdinasab, V.; Nikanjam, A.; Khomh, F.; Desmarais, M.C.; Jiang, Z.M. GitHub Copilot AI Pair Programmer: Asset or Liability? *J. Syst. Softw.* **2023**, *203*, 111734. [CrossRef]
2. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef] [PubMed]
3. Yu, Z.; Wang, K.; Wan, Z.; Xie, S.; Lv, Z. Popular Deep Learning Algorithms for Disease Prediction: A Review. *Clust. Comput.* **2023**, *26*, 1231–1251. [CrossRef] [PubMed]
4. Arias-Garzón, D.; Alzate-Grisales, J.A.; Orozco-Arias, S.; Arteaga-Arteaga, H.B.; Bravo-Ortiz, M.A.; Mora-Rubio, A.; Saborit-Torres, J.M.; Serrano, J.Á.M.; De La Iglesia Vayá, M.; Cardona-Morales, O.; et al. COVID-19 Detection in X-ray Images Using Convolutional Neural Networks. *Mach. Learn. Appl.* **2021**, *6*, 100138. [CrossRef] [PubMed]
5. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions. *J. Big Data* **2021**, *8*, 53. [CrossRef]
6. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale 2021. *arXiv* **2020**, arXiv2010.11929.
7. Front Matter. In *Computational Retinal Image Analysis*; Elsevier: Amsterdam, The Netherlands, 2019; pp. i–iii, ISBN 978-0-08-102816-2.
8. Poplin, R.; Varadarajan, A.V.; Blumer, K.; Liu, Y.; McConnell, M.V.; Corrado, G.S.; Peng, L.; Webster, D.R. Prediction of Cardiovascular Risk Factors from Retinal Fundus Photographs via Deep Learning. *Nat. Biomed. Eng.* **2018**, *2*, 158–164. [CrossRef] [PubMed]
9. Cheloni, R.; Gandolfi, S.A.; Signorelli, C.; Odone, A. Global Prevalence of Diabetic Retinopathy: Protocol for a Systematic Review and Meta-Analysis. *BMJ Open* **2019**, *9*, e022188. [CrossRef] [PubMed]
10. Sambyal, N.; Saini, P.; Syal, R.; Gupta, V. Modified U-Net Architecture for Semantic Segmentation of Diabetic Retinopathy Images. *Biocybern. Biomed. Eng.* **2020**, *40*, 1094–1109. [CrossRef]
11. Rehman, M.U.; Cho, S.; Kim, J.H.; Chong, K.T. BU-Net: Brain Tumor Segmentation Using Modified U-Net Architecture. *Electronics* **2020**, *9*, 2203. [CrossRef]
12. Rehman, M.U.; Cho, S.; Kim, J.; Chong, K.T. BrainSeg-Net: Brain Tumor MR Image Segmentation via Enhanced Encoder–Decoder Network. *Diagnostics* **2021**, *11*, 169. [CrossRef]
13. Anand, V.; Gupta, S.; Koundal, D.; Nayak, S.R.; Barsocchi, P.; Bhoi, A.K. Modified U-NET Architecture for Segmentation of Skin Lesion. *Sensors* **2022**, *22*, 867. [CrossRef] [PubMed]
14. Zou, K.; Chen, X.; Zhang, F.; Zhou, H.; Zhang, C. A Field Weed Density Evaluation Method Based on UAV Imaging and Modified U-Net. *Remote. Sens.* **2021**, *13*, 310. [CrossRef]
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceeding of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
16. Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. Swin Transformer V2: Scaling Up Capacity and Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 24 June 2022.
17. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A Comprehensive Survey on Transfer Learning. *Proc. IEEE* **2020**, *109*, 43–76. [CrossRef]
18. Srinivas, A.; Lin, T.-Y.; Parmar, N.; Shlens, J.; Abbeel, P.; Vaswani, A. Bottleneck Transformers for Visual Recognition. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 25 June 2021; IEEE: Piscatway, NJ, USA, 2021; pp. 16514–16524.
19. Yang, J.; Li, C.; Zhang, P.; Dai, X.; Xiao, B.; Yuan, L.; Gao, J. Focal Self-Attention for Local-Global Interactions in Vision Transformers. *arXiv* **2021**, arXiv2107.00641.
20. Moccia, S.; De Momi, E.; El Hadji, S.; Mattos, L.S. Blood Vessel Segmentation Algorithms—Review of Methods, Datasets and Evaluation Metrics. *Comput. Methods Programs Biomed.* **2018**, *158*, 71–91. [CrossRef]
21. Ciecholewski, M.; Kassjański, M. Computational Methods for Liver Vessel Segmentation in Medical Imaging: A Review. *Sensors* **2021**, *21*, 2027. [CrossRef]
22. Maurício, J.; Domingues, I.; Bernardino, J. Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. *Appl. Sci.* **2023**, *13*, 5521. [CrossRef]
23. Khanal, A.; Estrada, R. Dynamic Deep Networks for Retinal Vessel Segmentation. *Front. Comput. Sci.* **2020**, *2*, 35. [CrossRef]
24. Gegundez-Arias, M.E.; Marin-Santos, D.; Perez-Borrero, I.; Vasallo-Vazquez, M.J. A New Deep Learning Method for Blood Vessel Segmentation in Retinal Images Based on Convolutional Kernels and Modified U-Net Model. *Comput. Methods Programs Biomed.* **2021**, *205*, 106081. [CrossRef]
25. Galdran, A.; Anjos, A. State-of-the-art retinal vessel segmentation with minimalistic models. *Front. Nat.* **2022**, *12*, 6174 [CrossRef]
26. Tang, P.; Liang, Q.; Yan, X.; Zhang, D.; Coppola, G.; Sun, W. Multi-Proportion Channel Ensemble Model for Retinal Vessel Segmentation. *Comput. Biol. Med.* **2019**, *111*, 103352. [CrossRef]
27. Ma, Y.; Li, X.; Duan, X.; Peng, Y.; Zhang, Y. Retinal Vessel Segmentation by Deep Residual Learning with Wide Activation. *Comput. Intell. Neurosci.* **2020**, *2020*, 1–11. [CrossRef] [PubMed]

28. Tuyet, V.T.H.; Binh, N.T. Improving Retinal blood vessels Segmentation via Deep Learning in Salient Region. *SN Comput. Sci.* **2020**, *1*, 248. [CrossRef]
29. Park, K.-B.; Choi, S.H.; Lee, J.Y. M-GAN: Retinal Blood Vessel Segmentation by Balancing Losses Through Stacked Deep Fully Convolutional Networks. *IEEE Access* **2020**, *8*, 146308–146322. [CrossRef]
30. Zhuo, Z.; Huang, J.; Lu, K.; Pan, D.; Feng, S. A Size-Invariant Convolutional Network with Dense Connectivity Applied to Retinal Vessel Segmentation Measured by a Unique Index. *Comput. Methods Programs Biomed.* **2020**, *196*, 105508. [CrossRef]
31. Kamran, S.A.; Hossain, K.F.; Tavakkoli, A.; Zuckerbrod, S.L.; Sanders, K.M.; Baker, S.A. RV-GAN: Segmenting Retinal Vascular Structure in Fundus Photographs Using a Novel Multi-Scale Generative Adversarial Network. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September 27–1 October 2021; Volume 12908, pp. 34–44.
32. Yadav, G.; Maheshwari, S.; Agarwal, A. Contrast Limited Adaptive Histogram Equalization Based Enhancement for Real Time Video System. In Proceedings of the 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Delhi, India, 24–27 September 2014; IEEE: Piscatway, NJ, USA, 2014; pp. 2392–2397.
33. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
34. 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 (CVPR), Boston, MA, USA, 7–12 June 2015.
35. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
36. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv* **2020**, arXiv:1905.11946.
37. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
38. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. *arXiv* **2014**, arXiv:1409.4842.
39. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015.
40. Chen, D.; Yang, W.; Wang, L.; Tan, S.; Lin, J.; Bu, W. PCAT-UNet: UNet-like Network Fused Convolution and Transformer for Retinal Vessel Segmentation. *PLoS ONE* **2022**, *17*, e0262689. [CrossRef]
41. Guo, C.; Szemenyei, M.; Yi, Y.; Xue, Y.; Zhou, W.; Li, Y. Dense Residual Network for Retinal Vessel 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. 1374–1378.
42. Li, X.; Jiang, Y.; Li, M.; Yin, S. Lightweight Attention Convolutional Neural Network for Retinal Vessel Image Segmentation. *IEEE Trans. Ind. Inf.* **2021**, *17*, 1958–1967. [CrossRef]
43. Wu, H.; Wang, W.; Zhong, J.; Lei, B.; Wen, Z.; Qin, J. SCS-Net: A Scale and Context Sensitive Network for Retinal Vessel Segmentation. *Med. Image Anal.* **2021**, *70*, 102025. [CrossRef]
44. Tajbakhsh, N.; Lai, B.; Ananth, S.; Ding, X. ErrorNet: Learning Error Representations from Limited Data to Improve Vascular Segmentation. In Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 3–7 April 2020.