electronics

MDPI

Article

Attention-Enhanced Guided Multimodal and Semi-Supervised Networks for Visual Acuity (VA) Prediction after Anti-VEGF Therapy

Yizhen Wang [1], Yaqi Wang [2,*], Xianwen Liu [1], Weiwei Cui [3], Peng Jin [1], Yuxia Cheng [1] and Gangyong Jia [1,*]

[1] Department of Computer Science, Hangzhou Dianzi University, Hangzhou 310018, China; 221050001@hdu.edu.cn (Y.W.); 212050282@hdu.edu.cn (X.L.); 222050160@hdu.edu.cn (P.J.); yxcheng@hdu.edu.cn (Y.C.)
[2] College of Media Engineering, Communication University of Zhejiang, Hangzhou 310018, China
[3] School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK; acw499@qmul.ac.uk
* Correspondence: wangyaqi@cuz.edu.cn (Y.W.); gangyong@hdu.edu.cn (G.J.)

Abstract: The development of telemedicine technology has provided new avenues for the diagnosis and treatment of patients with DME, especially after anti-vascular endothelial growth factor (VEGF) therapy, and accurate prediction of patients' visual acuity (VA) is important for optimizing follow-up treatment plans. However, current automated prediction methods often require human intervention and have poor interpretability, making it difficult to be widely applied in telemedicine scenarios. Therefore, an efficient, automated prediction model with good interpretability is urgently needed to improve the treatment outcomes of DME patients in telemedicine settings. In this study, we propose a multimodal algorithm based on a semi-supervised learning framework, which aims to combine optical coherence tomography (OCT) images and clinical data to automatically predict the VA values of patients after anti-VEGF treatment. Our approach first performs retinal segmentation of OCT images via a semi-supervised learning framework, which in turn extracts key biomarkers such as central retinal thickness (CST). Subsequently, these features are combined with the patient's clinical data and fed into a multimodal learning algorithm for VA prediction. Our model performed well in the Asia Pacific Tele-Ophthalmology Society (APTOS) Big Data Competition, earning fifth place in the overall score and third place in VA prediction accuracy. Retinal segmentation achieved an accuracy of 99.03 ± 0.19% on the HZO dataset. This multimodal algorithmic framework is important in the context of telemedicine, especially for the treatment of DME patients.

Keywords: anti-VEGF; visual acuity, multimodal; diabetic macular edema; optical coherence tomography

check for updates

Citation: Wang, Y.; Wang, Y.; Liu, X.; Cui, W.; Jin, P.; Cheng, Y.; Jia, G. Attention-Enhanced Guided Multimodal and Semi-Supervised Networks for Visual Acuity (VA) Prediction after Anti-VEGF Therapy. *Electronics* **2024**, *13*, 3701. https://doi.org/10.3390/electronics13183701

Academic Editors: Bernard Fong and Loi Lei Lai

Received: 29 August 2024
Revised: 10 September 2024
Accepted: 13 September 2024
Published: 18 September 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Diabetic macular edema (DME) is the most common form of vision-threatening retinopathy in patients with diabetes [1,2]. Noninvasive imaging using optical coherence tomography allowed clinicians to detect mild diabetic macular edema to monitor progression and guide treatment [3]. Anti-vascular endothelial growth factor (VEGF) drugs were used to treat several ocular diseases that cause neovascular growth or swelling in the subretinal macula at the back of the eyes [4–6]. Ophthalmologists usually used a 1 + PRN regimen in which patients received one anti-VEGF injection in the first month and were observed every month thereafter, with repeat injections based on visual acuity (VA) and optical coherence tomography (OCT) imaging [7–12]. However, a large number of patients did not respond or respond adequately to this therapy. Different studies reported [13–16] that despite monthly anti-VEGF injections, 10% to 50% of patients do not respond or respond adequately. On the other hand, many patients with DME felt anxious due to the high cost of treatment [17,18]. Therefore, patients did not comply with the standard treatment

regimen, which usually led to a worse prognosis [19]. To our knowledge, the treatment effect could be greatly improved if we could predict the relevant index changes before treatment and tailor an individual treatment plan for them. Therefore, predicting VA and related indicators after anti-VEGF injections is of great value for effective DME treatment. However, existing methods for predicting VA after anti-VEGF treatment have two problems: 1. Existing VA prediction models usually rely heavily on manual intervention and expert knowledge to interpret the results, which limits their scalability and practicality in telemedicine environments. These models also often lack transparency in how predictions are generated from input data, making it difficult for clinicians to trust or act on their recommendations. 2. They are limited by the lack of integration between image-based data and clinical data, which reduces their accuracy and practicality in clinical settings. Many models also suffer from poor interpretability, especially in telemedicine scenarios, where a clear understanding of the predicted results is crucial for decision-making. Our proposed model addresses these limitations by adopting a multimodal approach that combines OCT images with clinical data, thereby improving the accuracy and interpretability of predictions.

In order to address the urgent need for an automated and rational solution to assist physicians in diagnosis, a specialized multimodal modeling algorithm was designed and presented in this paper. The algorithm was divided into two phases: the first one was to go from image modality to data modality, and the second one was to predict the final result from the data obtained in the previous phase combined with other clinical indicators of the patient. Before that, we conducted a correlation analysis for patients' clinical indicators, i.e., we explored the association of each indicator with post-treatment VA among multiple indicators, and the phenomenon we observed based on the results was that patients' pre-treatment and post-treatment central retinal thickness (CST) and pre-treatment VA had the highest degree of association for post-treatment VA. Therefore, in stage 1, we first designed and presented a targeted model to accurately segment the retina, after which we automated the calculation of the CST values based on the segmentation results. In stage 2, we used a machine learning approach as a regression task to combine the CST results with the pre-treatment VA to predict the post-treatment VA, and in this process, we also analyzed the importance of each index during the training phase, further confirming the feasibility of obtaining the CST from images and combining it with VA prediction. The main contributions of this paper are as follows:

- The method was based on a semi-supervised framework that can effectively address the time-consuming and laborious problem of retinal annotation of OCT images and the difficulty of measuring CST with patients' clinical information.
- This method explored the effect of patient clinical information, such as CST, on visual acuity values in patients with diabetic macular edema after anti-VEGF treatment, and demonstrated through specific data analysis and experiments that accurate calculation of patient CST can effectively improve the accuracy of VA prediction.
- The multimodal framework proposed in this method combined OCT images with patient information and demonstrated the frontier and effectiveness of the multimodal approach in the diagnosis and treatment of diabetic macular edema through competition rankings as well as experiments.

2. Related Work

Since the introduction of U-Net [20], 3D U-Net [21] for medical image segmentation in 2015, the feasibility of deep learning for medical image segmentation has been further demonstrated. And due to the success of transformer structure in the field of natural language processing, many transformer-based methods [22,23] have been proposed for application in medical image segmentation. All these methods required large-scale high-quality labeled data, yet for medical images, only experts can provide such data [24]. Semi-supervised segmentation is an effective solution to these problems, which combines self-training with consistency in the field of natural images [25], and it improves

the application of semi-supervised segmentation to medical images [26]. For example, Wu et al. [27] proposed a mutual consistency network (MC Net+) that utilizes uncertainty. And Li et al. [28] utilized regularization enhancement, which is beneficial for segmentation. The application of multimodal images and multimodal information can improve the accuracy of clinical diagnosis and downstream tasks such as segmentation and classification. For example, Huang et al. [29] and Chen et al. [30] used image information in combination with some other diagnostic information of the patient to handle the task. And Zhang et al. [31] used multimodal contrastive mutual learning and pseudo-label relearning to improve segmentation performance. Despite this, no multimodal VA prediction methods have been proposed.

Many methods have been proposed to automatically segment retinal layers, including digital image processing-based [32] and deep learning-based methods [33–36]. Several researchers have also explored diabetic macular edema by deep learning methods. For example, Alryalat et al. [37] segmented edematous fluid and fed the results into a classification model to predict the response of DME patients to anti-VEGF injections, i.e., good versus bad responders. And Liu et al. [38] used different generative adversarial networks to predict OCT images of short-term response to anti-VEGF treatment in DME patients. However, predicting only the patient's response is not enough; thus, subsequent works proposed exploring the prediction of some metrics after treatment. Zhang et al. [39] used a pure machine learning approach to predict the visual acuity of DME patients after anti-VEGF treatment using patient information provided by experts. However, this approach did not incorporate image modality information, and the required features such as post-treatment CST were measured manually by experts. Mirete et al. [40] analyzed changes in subfoveal choroidal thickness (SFCT) and its relationship with changes in macular central thickness (CME) in patients with type 2 diabetes after anti-VEGF drugs. Liu et al. [41] used an integrated model to predict post-treatment CST and VA one month after three loading doses of anti-VEGF injections in patients with DME, but did not explore the correlation between some of the pre-treatment and post-treatment metrics.

In spite of the fact that all of these works aimed to predict some indexes based on patient information, none of them effectively combined the information obtained from each modality. For instance, while U-Net-based models [20,21] have shown great promise in medical image segmentation, they require large-scale labeled datasets, which are often unavailable in clinical settings. Transformer-based models [22,23], although powerful, face similar challenges due to their data-hungry nature. Semi-supervised approaches such as MC-Net+ [27] leverage uncertainty to improve segmentation, yet still struggle to achieve high accuracy on unlabeled data. Our method combines the strengths of these approaches by utilizing a semi-supervised framework with attention mask data augmentation, improving segmentation performance even on challenging retinal OCT images without large labeled datasets.

3. Materials

HZO Dataset: The data were OCT scans of 300 eyes from 167 patients in Hangzhou Optometric Hospital, collected by the Heidelberg OCT acquisition machine in Germany. The data contain a total of 5000 OCT images, of which 1500 are expertly labeled retinal images and the rest are unlabeled images. In the training, we first processed the dataset into images of size 224×224, and used a variety of preprocessing techniques for each image, including normalization, random flipping, random rotation, and denoising. And then, we divided it into a training set of 3500 images, a validation set of 500 images, and a test set of 1000 images to evaluate the model, where the training set of 3500 images contained 500 annotated and 3000 unannotated images.

APTOS Dataset: The 2021 Ali Tianchi Asia-Pacific Ophthalmology Society Big Data Competition was organized by the Asia Pacific Tele-Ophthalmology Society (APTOS) and co-organized by Rajavithi Hospital (Thailand), Aravind Eye Hospital (India), and Sun Yat-sen University Zhongshan Eye Center (China). It was sponsored by the Medical Services

Department of the Ministry of Public Health of Thailand and supported by the Aliyun Tianchi platform. Training data and test data for the tournament were provided by the Asia Pacific Society of Ophthalmology in collaboration with Rajavati Hospital in Thailand and Aravind Eye Hospital in India. The data included the corresponding text information CSV files of each eye before and after treatment and the corresponding OCT images. We selected the training set of the rematch to evaluate the effectiveness of our algorithm; the data contained pre-treatment and post-treatment 2864 OCT images of 221 patients, but all unlabeled, as well as the corresponding CSV files. The data are publicly available at the following link: https://tianchi.aliyun.com/dataset/dataDetail?spm=5176.26982894.J_2044539870.1.30976959sz5jVL&dataId=127971 (accessed on 1 January 2020).

4. Methods

Our overall network framework for postoperative VA prediction is shown in Figure 1. We used CTCT [26] as the benchmark network for the segmentation network in our first stage, and we improved it accordingly. During stage 1, all OCT images (labeled and unlabeled) were inputted together to train both AmNet and Swin-unet [22] simultaneously. From the labeled images, attention-enhanced images were derived from the AmNet branch. Next, the enhanced images were fed into the AmNet branch again to create predictions. Moreover, the supervised loss L_{sup} was calculated based on the ground truth (GT) and the outputted predictions of the labeled images. With respect to unlabeled images, P1 and P2 output probability maps and their corresponding binary mask maps were obtained from the AmNet and Swin-Unet branches, respectively. MC Dropout (Monte Carlo) then computed the respective uncertainty maps to guide the cross-supervised loss L_{un}. In stage 2, we obtained the segmentation results of the retina using the trained AmNet network (no attention mask modules) and computed the CST on the images, and then the corresponding features of the patients pre- and post-treatment were combined to predict the final VA value using a machine learning algorithm.

4.1. Attention Mask Data-Augmentation

Inspired by WSDAN [42] and the challenge of OCT images, we propose a local branch with an Encoder–Decoder structure and an attention mask module, called the AmNet. As shown in Figure 1, the proposed AmNet consists of four encoders, an attention block, and four decoders, ensuring continuous structure in the spatial dimension. To face the problems of retinal lesions including structure changes, breaks and blurring, our proposed data enhancement method can fit the real data well and enhance the regularization of the network. Algorithm 1 demonstrates the operation process of attention mask data augmentation. Specifically for each image, as shown in Figure 2, we obtained its feature map from the last encoder and obtained the corresponding attention maps from the attention block, and then randomly selected one of the attention maps. For the selected attention map, we randomly generated multiple mask regions to obtain the augment map A_{mask}, and we chose 8×8 pixels for the mask size. Finally, we can obtain the enhanced image x_k, as follows:

$$x_k = x * A_{mask}, \tag{1}$$

where x is the input image, and $*$ denotes the elemental point multiplication of two tensors. Figure 3 shows the image after applying the attention mask enhancement. Comparison with the original image showed that in the retinal region, especially the borders of the retina, our data enhancement method masks in these areas (black squares in the figure).

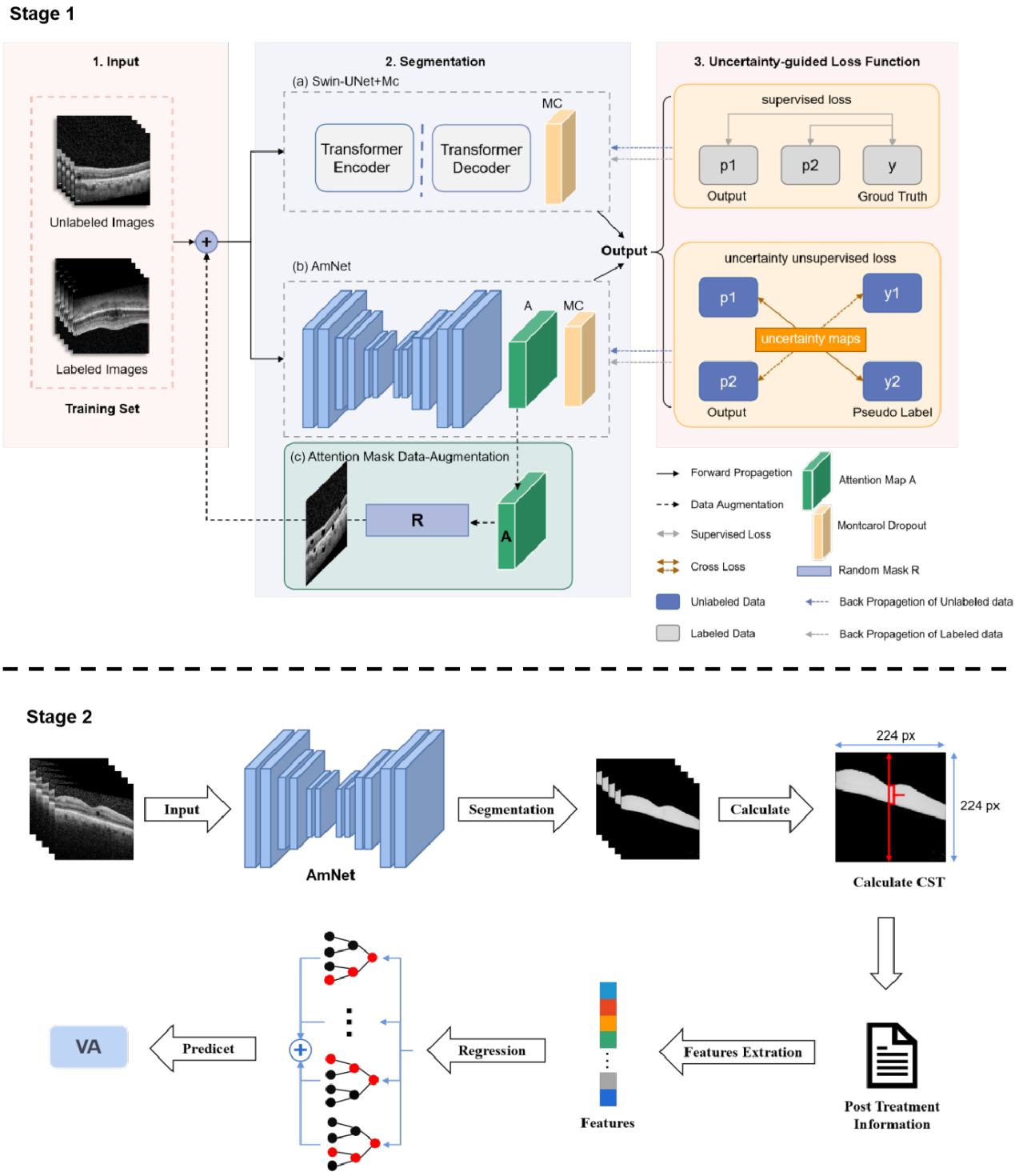

Figure 1. The architecture of our multimodal network. In the first stage, we trained the retinal segmentation network using a small number of labeled images and a large number of unlabeled images. In the second stage, we used the trained AmNet (no attention mask modules) to generate the segmentation map and calculate the thickness of the central macular concavity of the retina, which was combined with other textual information to form features for input into the random machine learning algorithm to predict the final VA.

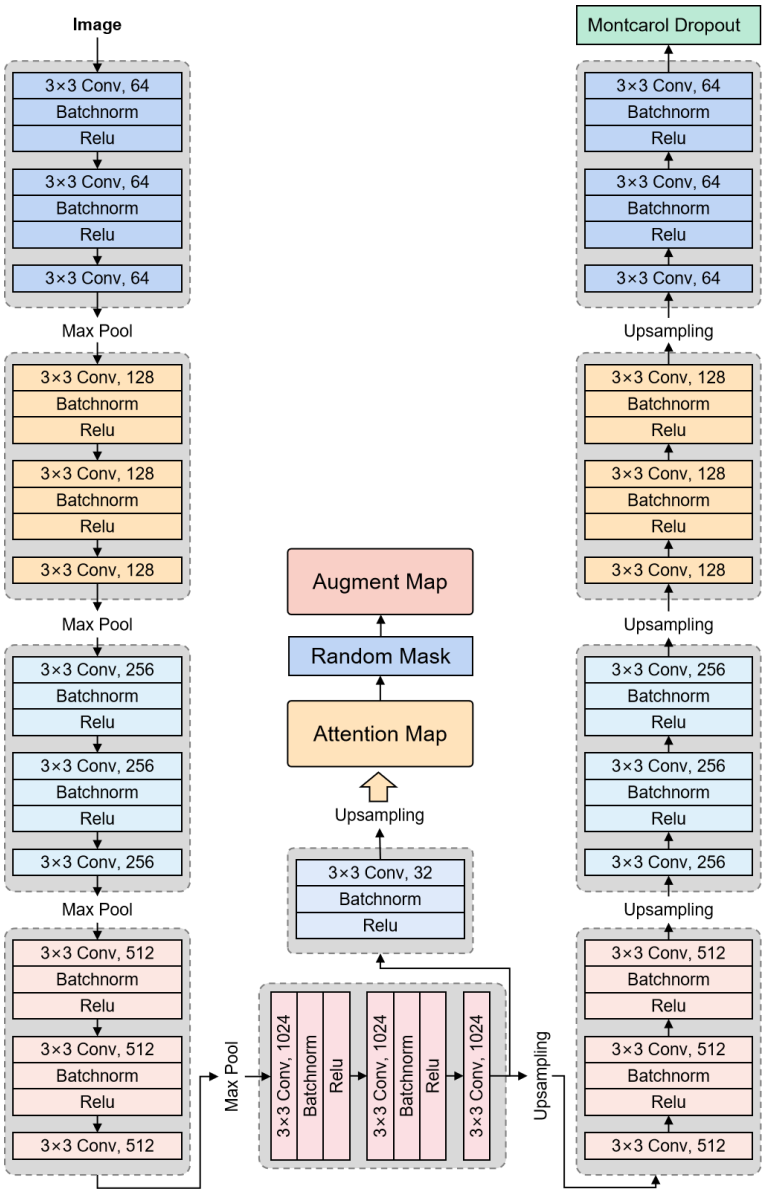

Figure 2. The AmNet branch for segmenting input images.

Algorithm 1: Attention mask data augmentation algorithm.

Input:
Input Images D
The elemental point multiplication of two tensors $*$
for *each* x_i *in batch from* D **do**
$x_i^f \leftarrow Feature(x_i)$
$A_{map} \leftarrow Attention_Maps_Corrsponded(x_i^f)$
$A_{map}^S \leftarrow Randomly_Selected(A_{map})$
$A_{mask} \leftarrow Mask_Regions_Generated\left(A_{map}^S\right)$
$x_i^k \leftarrow x_i * A_{mask}$
D_k.**append**(x_i^k)
end
Output: The enhanced images D_k

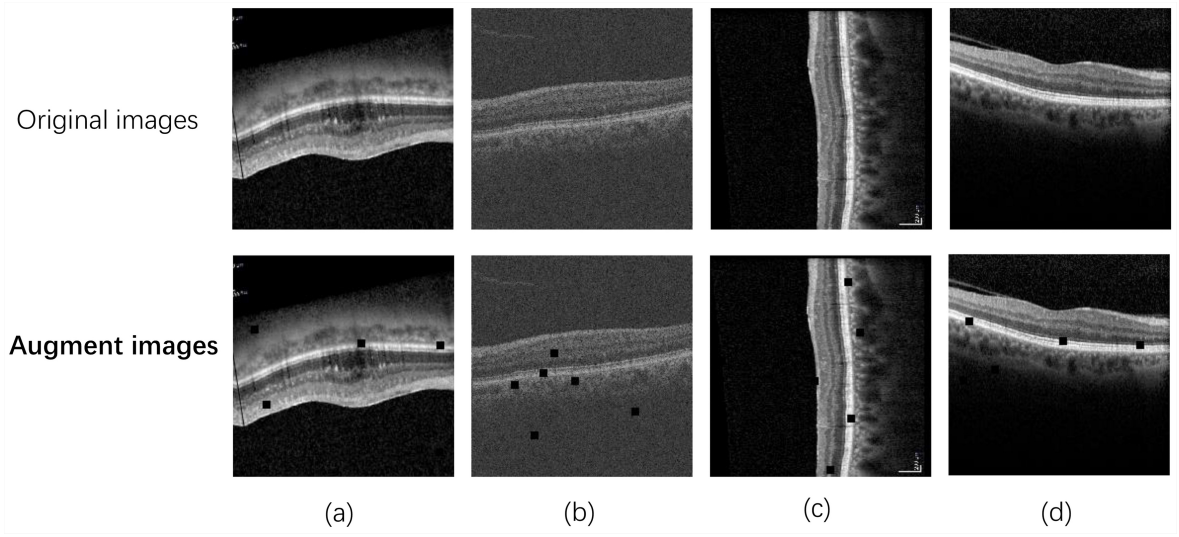

Figure 3. Enhanced OCT images. The first row is the original image; the second row is the image enhanced with the attention mask. In the enhanced graph, the black square is the result of our data enhancement method; it can be clearly seen that there was a black obscured block in the border part of the retina.

4.2. Uncertainty-Guided Loss Function

The poor quality of pseudo-labeling in semi-supervised learning will result in the wrong information propagating into the network, which will affect the training of the model if no constraint is imposed. Algorithm 2 demonstrates the operation process of the uncertainty-guided loss function. To overcome this problem, in Equation (3), we proposed a loss function based on uncertainty to guide pseudo-labeling by the MC module. To estimate the prediction uncertainty, we used Monte-Carlo dropout [43]. Figure 4 shows the uncertainty plot calculated by the model using MC dropout. The input data with additional noise were forward-propagated N times to obtain the prediction entropy:

$$U = -\sum \frac{1}{N} \sum_{n} p_n log(\frac{1}{N} \sum_{n} P_n). \tag{2}$$

where P_n is the probability vector of the n_{th} prediction. After obtaining the entropy map, we added it to the mean squared error (MSE), which can prevent the impact of pixels with high uncertainty on model training, as follows:

$$L_{un}(y_i^{'} - y_i) = \frac{\sum_i (U_i < H) \left\| y_i^{'} - y_i \right\|^2}{\sum_i (U_i < H)}, \tag{3}$$

where H is the set uncertainty threshold and returns 1 when the uncertainty of the corresponding voxel is less than H, otherwise it returns 0. $y_i^{'}$ is the probability corresponding to the network prediction result and y_i is the value corresponding to the network segmentation mask. In Figure 5, we show the computational visualization of uncertainty unsupervised loss.

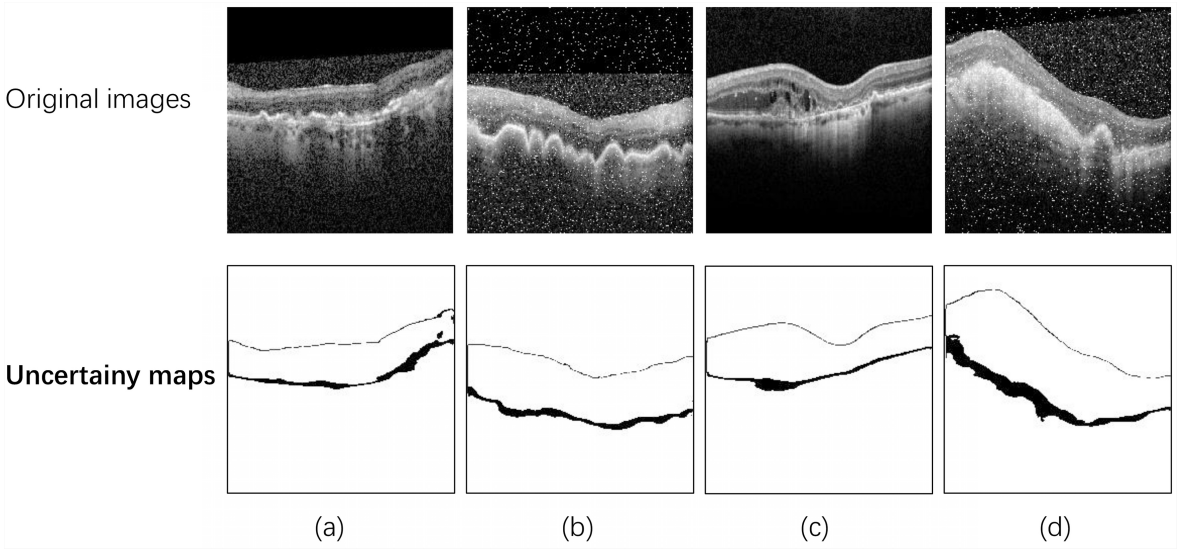

Figure 4. Uncertainty maps of several OCT image patches. The uncertainty of the model for retinal segmentation reveals that some of the uncertainties were exactly where the model segmentation was wrong, where black represents the high uncertainty of the model for the prediction.

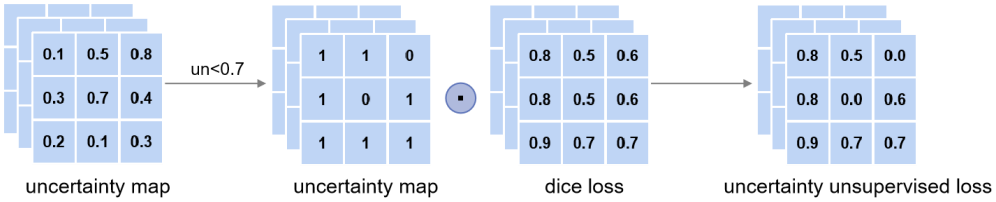

Figure 5. Uncertainty unsupervised loss calculation process.

Algorithm 2: Uncertainty-guided loss function/pseudo-labeling guided by MC module

Input:

Probability vector of prediction PV

Set uncertainty threshold H

Value corresponding to the network segmentation mask y_n, probability corresponding to the network prediction result $y_n^{'}$

for *each prediction* ***from*** 1 ***to*** N **do**

$A_{PV_n} \leftarrow \frac{1}{N}(\sum_{i=1}^{n} PV_i)$

%%A_{PV_n} means the Average value of PV_i

$U_n \leftarrow -A_{PV_n} \ln(A_{PV_n})$

$Sum_{U_i} \leftarrow \sum_{n=1}^{i}(is_true(U_n < H))$

%%*is_true* is a function which returns 1 when the result is true, otherwise 0

$L_{un}(y_n^{'} - y_n) \leftarrow \frac{Sum_{U_i}\left\|y_n^{'} - y_n\right\|^2}{Sum_{U_i}}$

end

4.3. Macular Central Recess Thickness Calculation

In the first stage, we used the modal information of the patient picture to obtain the retinal segmentation results, and in the second stage, we used the obtained retinal segmentation results to perform the extraction of the modal feature of the thickness of the central retinal notch, as we believed that this modal feature was useful for the subsequent VA prediction of the patient. This step included both the pre-treatment and post-treatment retinal segmentation result calculation. Figure 6 shows the location of our calculations.

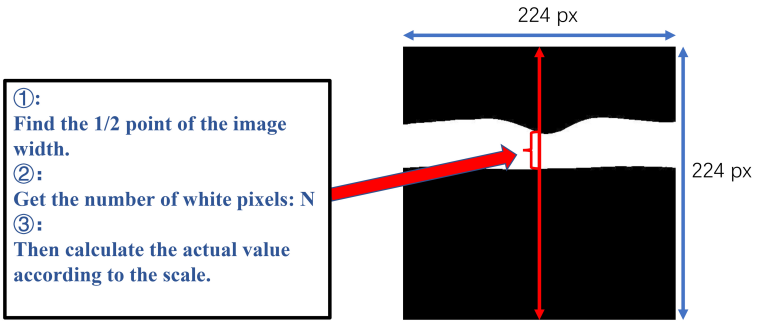

Figure 6. Schematic diagram of CST calculation. We counted the white pixel points at the selected position and transformed them into standard unit values by the result of the segmentation

4.4. Prediction VA

We first analyzed the correlations of each feature in the dataset, as shown in Figure 7. In the later sections, we used preCST to refer to pre-treatment CST and CST to refer to post-treatment CST. The highest correlations with postoperative VA were preVA (0.83), preCST (0.38), and CST (0.4). To predict the final VA, we took a machine learning approach and used XGBoost [44] as our algorithmic tool, with default settings except for hyperparameter tuning, as shown in Figure 8. We visualized and analyzed the importance of the features during the training process, and it was obvious that preVA, preCST, and CST were much more important than the other features for predicting VA, which verifies that our method of calculating retinal sulcus thickness from the patient's image modality and combining it with other modal information is effective and feasible.

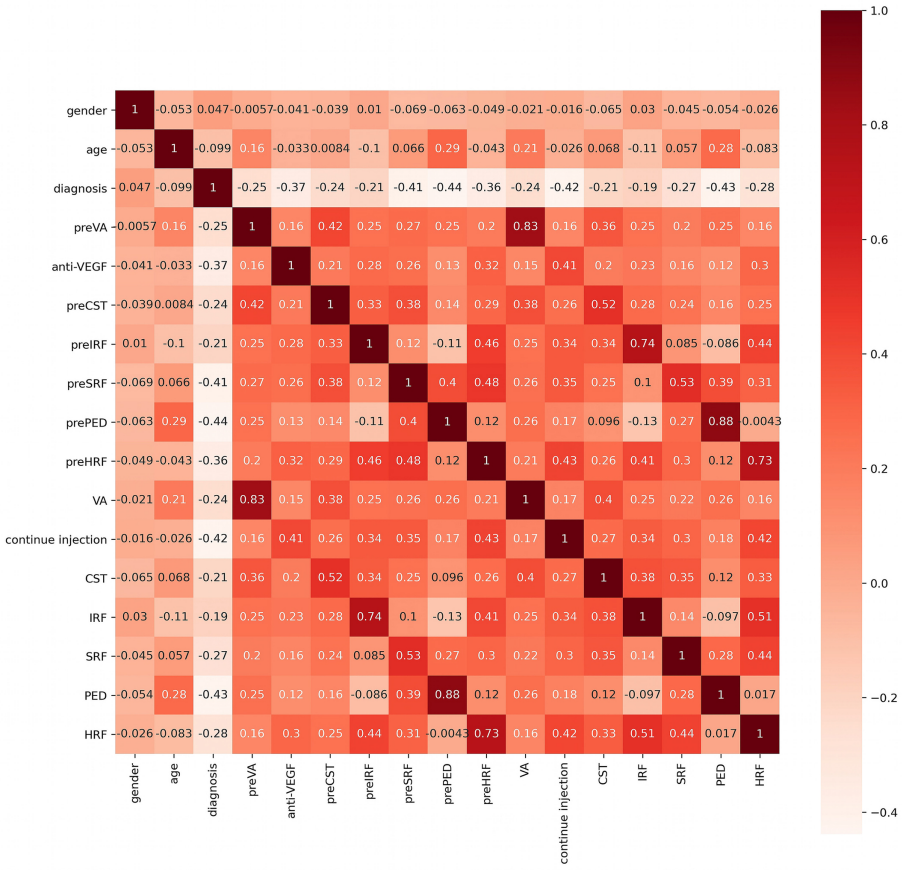

Figure 7. Correlation analysis between individual patient information. In the heatmap, the darker the color, the higher the degree of correlation.

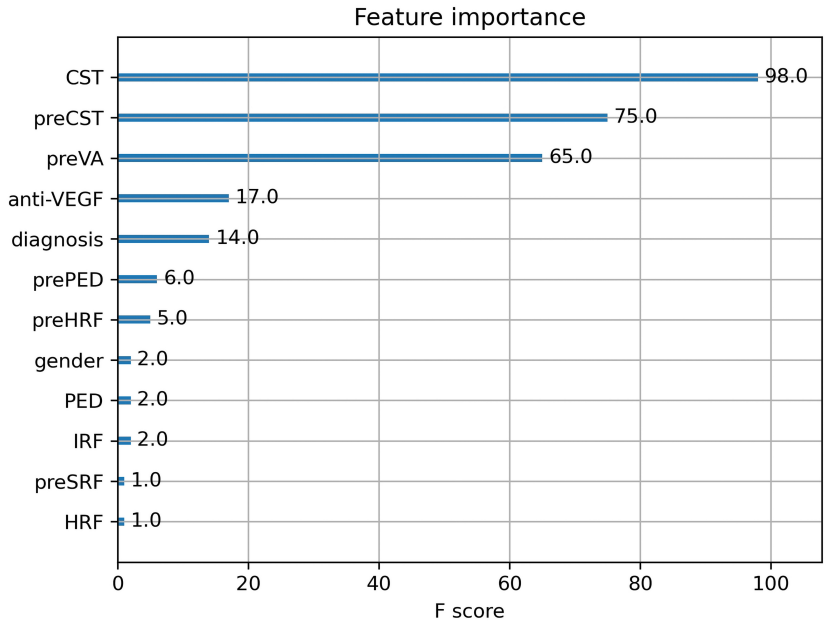

Figure 8. Histogram of feature importance scores. The features with higher scores indicated that the machine learning algorithm split more frequently in this process, and we selected features according to their importance.

5. Experiment

5.1. Implementation Details

We implemented our algorithmic model using Pytorch 1.2.0 and an Nvidia RTX 2080s GPU (Nividia, Santa Clara, CA, USA). The weights of all networks are learned by the Adam optimizer, and the weight decay is the default value. The learning rate was set to 0.0001. In addition, the image input size was 224 × 224 and online data enhancement techniques, including random flipping and random rotation, were used to further mitigate the risk of overfitting. The machine learning part used XGBoost for training and prediction, and the parameters were searched to select the optimal combination. For a fair comparison, all compared networks were implemented on the same computer with hyperparametric optimization.

5.2. Evaluation

Dice was a commonly used metric in medical image segmentation, which was an ensemble similarity metric usually used to calculate the similarity of two samples with a value threshold of [0, 1]. We used it to evaluate the effect of retinal segmentation on the HZO dataset. The Dice coefficient was calculated as follows:

$$Dice = \frac{2 \times (pred \cap true)}{pred \cup true}, \tag{4}$$

where pred is the result of the model output and true is the true annotation.

For the APTOS dataset, the preCST and CST were calculated using a ±7.5% confidence interval, and the calculation was correct if it was within this confidence interval; if the actual value of VA was not greater than 1, the predicted result was in the interval of [VA-0.05, VA + 0.05] was considered correct; if the actual value of VA was greater than 1, the predicted result was in the interval of [VA × 0.925, VA × 1.075] was considered correct.

To assess the prediction quality of the model systematically, three more evaluation metrics were used, including mean absolute error (MAE) ($MAE = \frac{1}{n}\sum_{i=1}^{n}\left|l_i' - l_i\right|$), root

mean square error (RMSE) ($RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(l_i' - l_i)^2}$), and R^2($R^2 = \frac{\sum_{i=1}^{n}(l_i' - l_i)^2}{\sum_{i=1}^{n}(l_i - \bar{l})^2}$). The true outcome is 1 and $\bar{l}$ is the mean of l. l' is the predicted value and i is the number of patients. The lower the MAE and RMSE values, the closer the predicted CST and preCST and VA values were to the true values. The coefficient of determination R^2 was used to show the goodness of fit of the model.

6. Results

We first experimentally compared our proposed algorithm with some semi-supervised medical segmentation schemes on the HZO dataset, including UAMT [45], CPS [25], and CTCT [26]. In addition, we used CTCT as our baseline. As can be seen from Table 1, our algorithm has improved Dice scores in this step of retinal segmentation compared to these methods.

In Figure 9, a few challenging OCT images show blurred or broken retinal borders, such as Figure 9a–c. According to the segmentation results, our proposed method achieved precise segmentation results for this type of image. Specifically, the retina boundary was broken on the right side of Figure 9a,b and in the middle area of Figure 9c. Our algorithm successfully segmented these regions, whereas the baseline network could not. As shown in Figure 9d,e, due to edema and changes in retinal morphology, it was hard to distinguish the tissue boundary from the surrounding area, making segmentation difficult. In this case, our algorithm delivered very close segmentation results. Although there was severe deformation of the retina in Figure 9d, most of the regions were still segmented.

Table 1. Results of semantic segmentation. We compared the Dice results of retinal segmentation with UAMT, CPS, and CTCT on the HZO dataset.

	Method	Dice
Segment	UAMT [45]	97.55 ± 0.4%
	CPS [25]	98.23 ± 0.12%
	CTCT [26]	98.75 ± 0.03%
	Ours	**99.03 ± 0.19%**

Since we used CTCT as the baseline, the next experimental comparison analysis will be performed afterward based on the segmentation results and CTCT. In Table 2, we compare the preCST and CST calculation accuracy of our algorithm with CTCT on the APTOS dataset. PreCST and CST accuracy were 67.788% and 68.807%, which were 5.3% and 2.2% higher than CTCT, respectively. In addition, our algorithm performs better on other metrics, such as preCST (MAE, RMSE, and R2 were 49, 108.805, and 0.577, respectively) and CST (MAE, RMSE, and R2 were 37.365, 87.932, and 0.618, respectively). This result comes from the accuracy improvement of retinal segmentation, although it was only about 0.4% higher on segmented Dice; due to the scale of central retinal thickness measurement, sometimes a few pixels of error can affect the accuracy, so fine segmentation of retinal thickness was crucial, and this result influenced the subsequent VA prediction.

Table 2. The accuracy of preCST and CST predictions. We compared the computational accuracy of preCST and CST with CTCT on the APTOS dataset and calculated the corresponding MAE, RMSE, and R^2 to demonstrate the superiority of our model.

	Method	Accuracy	MAE (μm)	RMSE (μm)	R^2
preCST	CTCT [26]	62.5%	57.99	110.789	0.561
	Ours	**68.807%**	**49**	**108.805**	**0.577**
CST	CTCT [26]	65.596%	45.846	90.584	0.595
	Ours	**67.788%**	**37.365**	**87.932**	**0.618**

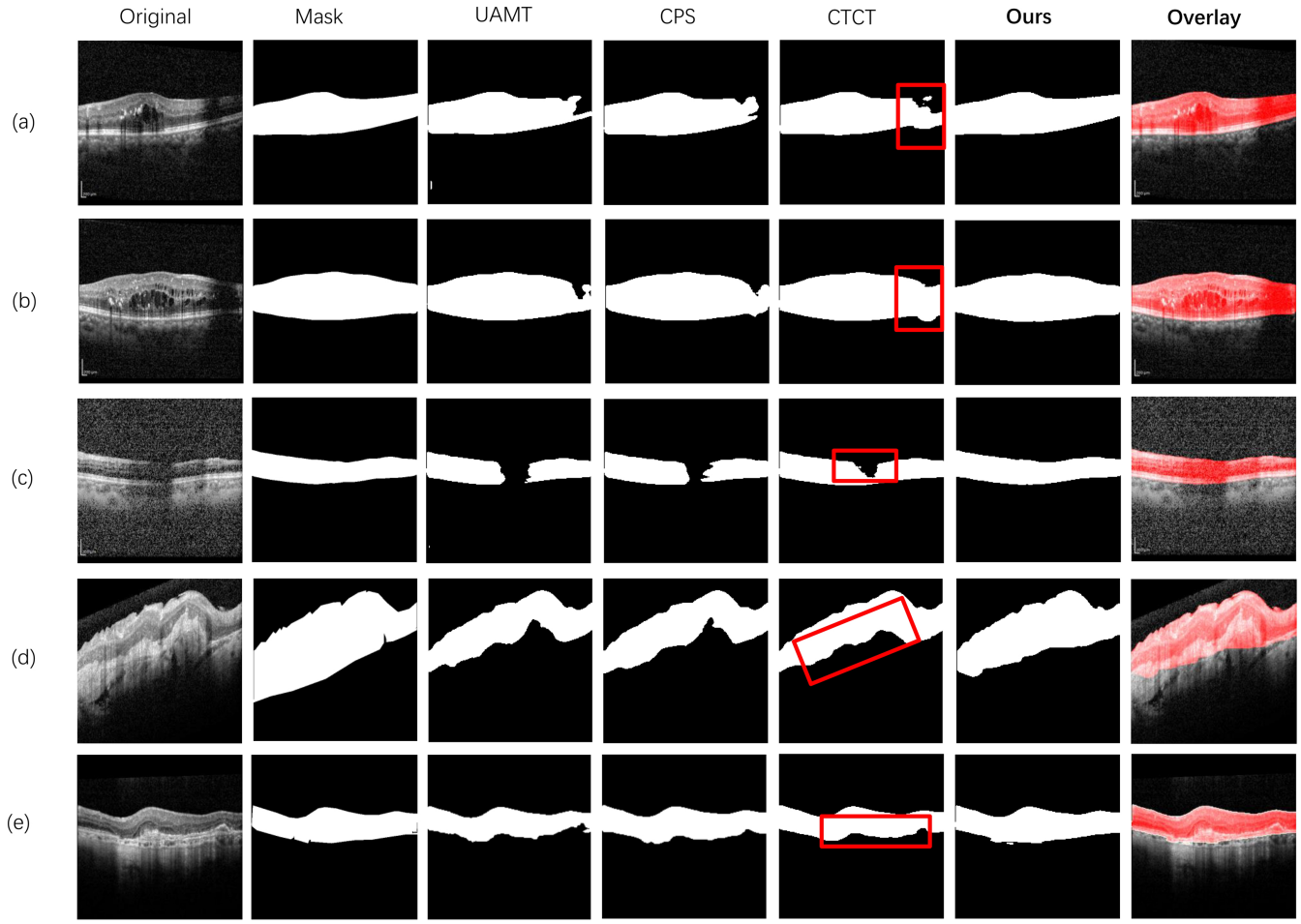

Figure 9. Comparison of the segmentation results. There are seven columns, including the original image, the labeled image, the UAMT segmentation result, the CPS segmentation result, the CTCT segmentation result, and the visualization of the segmentation result of our algorithm and the overlay on the original image. In the red boxes, the baseline model predicts incorrect results, while our algorithm produces satisfactory segmentation results.

The detailed regression analysis results for prediction and ground truth are shown in Figure 10. The curve fit of the calculation result of our algorithm was better than the baseline, which indicated that our calculation was closer to the ground truth. To illustrate the effectiveness of our algorithm, we also examined the computational error values of preCST and CST, as shown in Figure 11. It can be seen in the same sample that the left side of the corresponding position had a lower height than the right side, indicating that our calculation results compared to the baseline resulted in a significant reduction in error, , i.e., the accuracy of some samples with large errors was improved.

In Table 3, the CST results calculated by Table 2 are compared to the VA accuracy predicted by the other feature fusions; our algorithm still outperformed the baseline by about 5%, and the corresponding other indicators also outperformed the baseline (MAE, RMSE, and R2 were 0.106, 0.141, and 0.722, respectively). Moreover, we analyzed the distribution of computational errors in VA, as shown in Figure 12. It can be seen from a box plot inside the graph that the computational error of VA obtained by our algorithm was more concentrated in a smaller error range compared to the graph of the baseline.

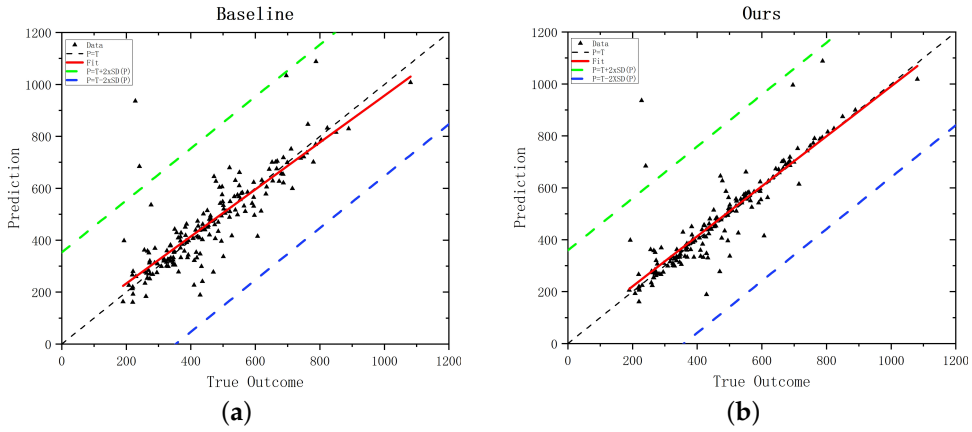

Figure 10. Regression analysis for prediction and ground truth. A red solid line represents the regression of prediction and annotation correlation (black triangles) from one input OCT image. Axis values of the black triangle show CST (or pre-CST) values of the ground truth and prediction. The black dashed line indicates that the predicted results are the same as the actual results. Blue or green dashed lines represent the center line plus or minus two times the predicted standard deviation. In general, the closer the solid red line is to the dashed black line, the better the model fits. P is the prediction; T is the true outcome, and SD is the standard deviation.

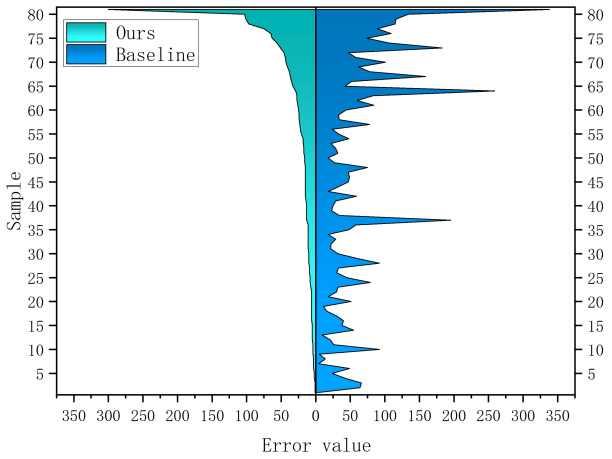

Figure 11. Comparison plots of the errors of preCST and CST with the real results. We selected some errors of preCST and CST with the real results. The left side shows the error distribution calculated by our algorithm and the right side shows the error distribution calculated by baseline; the horizontal axis shows the error values and the vertical axis shows the corresponding samples.

Table 3. The accuracy of VA prediction. We compared the computational accuracy of VA with CTCT on the APTOS dataset and calculated the corresponding MAE, RMSE, and R^2 to demonstrate the superiority of our model.

	Method	Accuracy	MAE	RMSE	R^2
VA	CTCT [26]	33.12%	0.112	0.150	0.688
	Ours	**38.15%**	**0.106**	**0.141**	**0.722**

In addition to improving accuracy at all steps, our multimodal algorithm utilized the most relevant features required by the competition. Table 4 lists the central retinal thickness and VA rankings of the top five teams in the competition at that time. A preCST prediction of our current multimodal framework was better than that of the teams at that time, and a CST prediction was similar to the highest, along with a final VA prediction that exceeded

the highest team at that time. Analyzing the results, our current algorithm demonstrated superior accuracy, and we believed it could also achieve better results in the competition.

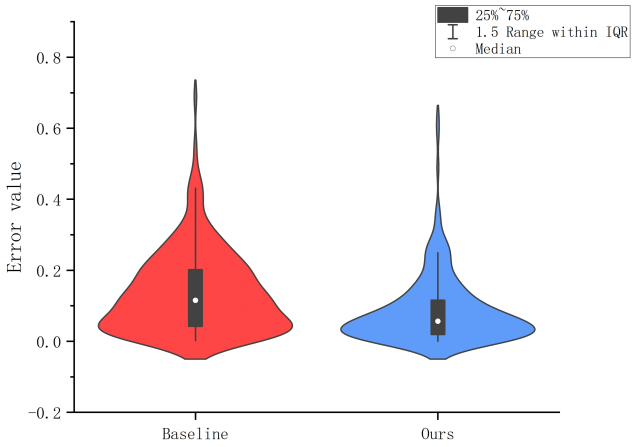

Figure 12. Error distribution plot of the VA prediction. The violin plot of the error distribution illustrates the VA prediction calculated by the baseline and our algorithm. The white points in the graph are the median of the error.

Table 4. Retinal central macula thickness scores for APTOS 2021. A ranking based on the final total score is shown in the first column, while accuracy scores for preCST, CST and VA are shown in the second and third columns.

TOP 5	preCST	CST	VA
Team 1	68.71%	69.30%	32.16%
Team 2	64.62%	68.42%	32.16%
Team 3	65.20%	64.33%	34.21%
Team 4	59.36%	59.94%	35.67%
Ours	**61.40%**	**65.20%**	**33.04%**

7. Discussion

In this work, we designed a multimodal algorithmic framework in two stages for predicting VA values in DME patients after receiving anti-VEGF treatment. According to our findings, the algorithmic framework accurately predicted post-treatment VA based on pre-treatment OCT images and clinical indicators.

Previous studies have shown that pre-treatment clinical indicators obtained from OCT images can only be used to estimate the likelihood of good or bad treatment, while accurate prediction of post-treatment indicators is considered infeasible. Therefore, we explored the degree of association of these parameters with VA and found a high degree of association for the metric CST, and we could obtain this value from OCT images. We designed a semi-supervised retinal segmentation framework based on OCT images, and the module designed for the morphological characteristics of the retina is well suited to ensure the accuracy of the model for segmentation, especially in images that suffer a high degree of morphological damage due to DME, which happens to be the most important. In the next stage, we input the CST values and the pre-treatment VA into the ML model to predict the post-treatment VA. In terms of VA accuracy prediction, it is also found that the accuracy of CST directly affects the prediction results according to the study results, which also shows that our proposed multimodal algorithm of calculating CST from OCT images and then predicting VA based on the results is effective and feasible. This will help physicians to develop better treatment plans for their patients, and more importantly, the prediction of our system is based on clinical information common to DME patients, such as OCT images, and does not require manual intervention to measure additional clinical indicators. This allows more physicians to use the system without the need for additional investment in time and manpower.

Nevertheless, our study has some limitations that should be acknowledged. For calculating CST based on the segmentation of retinal images in the first stage, our current method of calculation, which may misjudge images with distortions, may lead to large deviations in the results, and we will apply a well-designed method to obtain more accurate values for the calculation of the central retinal concave thickness. In addition, our model was trained on high-quality OCT images, and its performance may degrade when applied to low-quality images or datasets with different characteristics. In addition, although our method achieves high accuracy in predicting visual acuity based on retinal thickness, it may be less effective in cases where OCT images are severely distorted or have artifacts. Future work will focus on improving the robustness of the model by integrating more data from different clinical settings.

8. Conclusions

In conclusion, our study presents a novel approach that combines semi-supervised learning with multimodal data for VA prediction. This combination allows the model to effectively utilize both OCT images and clinical data, addressing the limitations of previous methods that rely solely on a single data source. The superiority of the algorithm was validated on a hospital dataset called HZO as well as on the 2021-APTOS dataset. The results of this study also showed the interpretability of our proposed method. In summary, the proposed multimodal algorithmic framework can help doctors and professionals to develop better treatment plans and improve the controllability and effectiveness of treatment.

Future work will focus on external validation of our model using datasets from other institutions and patient populations to further assess its generalizability. Additionally, we plan to explore the application of our method to other retinal diseases, such as age-related macular degeneration (AMD) and diabetic retinopathy, where OCT imaging is commonly used. By validating the model across a broader range of conditions, we aim to enhance its clinical utility and ensure its effectiveness in a wider variety of telemedicine applications.

Author Contributions: Conceptualization, Y.W. (Yizhen Wang); data curation, X.L. and W.C.; funding acquisition, Y.W. (Yaqi Wang) and G.J.; investigation, P.J. and X.L.; methodology, Y.W. (Yaqi Wang & Yizhen Wang); supervision, Y.C.; writing—original draft, Y.W. (Yizhen Wang); writing—review & editing, Y.W. (Yaqi Wang). All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the National Natural Science Foundation of China (No. 62206242), and the National Natural Science Foundation of China (No. U20A20386).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Center for Rehabilitation Medicine, Department of Ophthalmology, Zhejiang Provincial People's Hospital.

Data Availability Statement: The HZO dataset analyzed during the current study is not publicly accessible but is available upon reasonable request from the corresponding author. The APTOS-2021 dataset is now open. For more information, please see APTOS-2021: A dataset for predicting anti-VEGF treatment outcomes https://tianchi.aliyun.com/specials/promotion/APTOS (accessed on 1 January 2020).

Acknowledgments: Thanks to the Ali Tianchi Big Data Contest for providing data and platform support, and Hangzhou Optometric Hospital for providing data support.

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

References

1. Yau, J.W.; Rogers, S.L.; Kawasaki, R.; Lamoureux, E.L.; Kowalski, J.W.; Bek, T.; Chen, S.J.; Dekker, J.M.; Fletcher, A.; Grauslund, J.; et al. Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care* **2012**, *35*, 556–564. [CrossRef] [PubMed]
2. Yan, Y.; Jin, K.; Gao, Z.; Huang, X.; Wang, F.; Wang, Y.; Ye, J. Attention-based deep learning system for automated diagnoses of age-related macular degeneration in optical coherence tomography images. *Med. Phys.* **2021**, *48*, 4926–4934. [CrossRef] [PubMed]
3. Ye, X.; Wang, J.; Chen, Y.; Lv, Z.; He, S.; Mao, J.; Xu, J.; Shen, L. Automatic screening and identifying myopic maculopathy on optical coherence tomography images using deep learning. *Transl. Vis. Sci. Technol.* **2021**, *10*, 10. [CrossRef] [PubMed]

4. Régnier, S.; Malcolm, W.; Allen, F.; Wright, J.; Bezlyak, V. Efficacy of anti-VEGF and laser photocoagulation in the treatment of visual impairment due to diabetic macular edema: A systematic review and network meta-analysis. *PLoS ONE* **2014**, *9*, e102309. [CrossRef]
5. Iglicki, M.; González, D.P.; Loewenstein, A.; Zur, D. Next-generation anti-VEGF agents for diabetic macular oedema. *Eye* **2022**, *36*, 273–277. [CrossRef]
6. Jin, K.; Ye, J. Artificial intelligence and deep learning in ophthalmology: Current status and future perspectives. *Adv. Ophthalmol. Pract. Res.* **2022**, *2* , 100078. [CrossRef]
7. Lai, K.; Huang, C.; Li, L.; Gong, Y.; Xu, F.; Zhong, X.; Lu, L.; Jin, C. Anatomical and functional responses in eyes with diabetic macular edema treated with "1+ PRN" ranibizumab: One-year outcomes in population of mainland China. *BMC Ophthalmol.* **2020**, *20*, 229. [CrossRef]
8. Sugimoto, M.; Tsukitome, H.; Okamoto, F.; Oshika, T.; Ueda, T.; Niki, M.; Mitamura, Y.; Ishikawa, H.; Gomi, F.; Kitano, S.; et al. Clinical preferences and trends of anti-vascular endothelial growth factor treatments for diabetic macular edema in Japan. *J. Diabetes Investig.* **2019**, *10*, 475–483. [CrossRef]
9. James, D.G.; Mitkute, D.; Porter, G.; Vayalambrone, D. Visual outcomes following intravitreal ranibizumab for diabetic macular edema in a pro re nata protocol from baseline: A real-world experience. *Asia-Pac. J. Ophthalmol.* **2019**, *8*, 200–205.
10. Iglicki, M.; Loewenstein, A.; Barak, A.; Schwartz, S.; Zur, D. Outer retinal hyperreflective deposits (ORYD): A new OCT feature in naïve diabetic macular oedema after PPV with ILM peeling. *Br. J. Ophthalmol.* **2020**, *104*, 666–671. [CrossRef]
11. Xu, F.; Liu, S.; Xiang, Y.; Hong, J.; Wang, J.; Shao, Z.; Zhang, R.; Zhao, W.; Yu, X.; Li, Z.; et al. Prediction of the Short-term therapeutic effect of anti-VEGF therapy for diabetic macular edema using a generative adversarial network with OCT images. *J. Clin. Med.* **2022**, *11*, 2878. [CrossRef] [PubMed]
12. Lee, J.; Moon, B.G.; Cho, A.R.; Yoon, Y.H. Optical Coherence Tomography Angiography of DME and Its Association with Anti-VEGF Treatment Response. *Ophthalmology* **2016**, *123*, 2368–2375. [CrossRef] [PubMed]
13. Wong, T.Y.; Sun, J.; Kawasaki, R.; Ruamviboonsuk, P.; Gupta, N.; Lansingh, V.C.; Maia, M.; Mathenge, W.; Moreker, S.; Muqit, M.M.; et al. Guidelines on diabetic eye care: The international council of ophthalmology recommendations for screening, follow-up, referral, and treatment based on resource settings. *Ophthalmology* **2018**, *125*, 1608–1622. [CrossRef] [PubMed]
14. Jampol, L.M.; Bressler, N.M.; Glassman, A.R. Revolution to a new standard treatment of diabetic macular edema. *JAMA* **2014**, *311*, 2269–2270. [CrossRef]
15. Das, A.; McGuire, P.G.; Rangasamy, S. Diabetic macular edema: Pathophysiology and novel therapeutic targets. *Ophthalmology* **2015**, *122*, 1375–1394. [CrossRef] [PubMed]
16. Gonzalez, V.H.; Campbell, J.; Holekamp, N.M.; Kiss, S.; Loewenstein, A.; Augustin, A.J.; Ma, J.; Ho, A.C.; Patel, V.; Whitcup, S.M.; et al. Early and long-term responses to anti–vascular endothelial growth factor therapy in diabetic macular edema: Analysis of protocol I data. *Am. J. Ophthalmol.* **2016**, *172*, 72–79. [CrossRef]
17. Zur, D.; Iglicki, M.; Sala-Puigdollers, A.; Chhablani, J.; Lupidi, M.; Fraser-Bell, S.; Mendes, T.S.; Chaikitmongkol, V.; Cebeci, Z.; Dollberg, D.; et al. Disorganization of retinal inner layers as a biomarker in patients with diabetic macular oedema treated with dexamethasone implant. *Acta Ophthalmol.* **2020**, *98*, e217–e223. [CrossRef]
18. Rubino, A.; Rousculp, M.; Davis, K.; Wang, J.; Girach, A. Diagnosed diabetic retinopathy in France, Italy, Spain, and the United Kingdom. *Prim. Care Diabetes* **2007**, *1*, 75–80. [CrossRef]
19. Weiss, M.; Sim, D.A.; Herold, T.; Schumann, R.G.; Liegl, R.; Kern, C.; Kreutzer, T.; Schiefelbein, J.; Rottmann, M.; Priglinger, S.; et al. Compliance and adherence of patients with diabetic macular edema to intravitreal anti–vascular endothelial growth factor therapy in daily practice. *Retina* **2018**, *38*, 2293–2300. [CrossRef]
20. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*; Springer International Publishing: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
21. Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention*; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 424–432.
22. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv* **2021**, arXiv:2105.05537.
23. Wang, H.; Cao, P.; Wang, J.; Zaiane, O.R. Uctransnet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 27 February–2 March 2022; pp. 2441–2449.
24. Tajbakhsh, N.; Jeyaseelan, L.; Li, Q.; Chiang, J.N.; Wu, Z.; Ding, X. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Med. Image Anal.* **2020**, *63*, 101693. [CrossRef] [PubMed]
25. Chen, X.; Yuan, Y.; Zeng, G.; Wang, J. Semi-supervised semantic segmentation with cross pseudo supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2613–2622.
26. Luo, X.; Chen, J.; Song, T.; Wang, G. Semi-supervised medical image segmentation through dual-task consistency. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual Event, 2–9 February 2021; pp. 8801–8809.
27. Wu, Y.; Ge, Z.; Zhang, D.; Xu, M.; Zhang, L.; Xia, Y.; Cai, J. Mutual consistency learning for semi-supervised medical image segmentation. *Med. Image Anal.* **2022**, *81*, 102530. [CrossRef] [PubMed]

28. Li, X.; Yu, L.; Chen, H.; Fu, C.W.; Xing, L.; Heng, P.A. Transformation-Consistent Self-Ensembling Model for Semisupervised Medical Image Segmentation. *IEEE Trans. Neural Networks Learn. Syst.* **2021**, *32*, 523–534. [CrossRef]
29. Huang, S.C.; Shen, L.; Lungren, M.P.; Yeung, S. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10–17 October 2021; pp. 3942–3951.
30. Chen, R.J.; Lu, M.Y.; Weng, W.H.; Chen, T.Y.; Williamson, D.F.; Manz, T.; Shady, M.; Mahmood, F. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10–17 October 2021; pp. 4015–4025.
31. Zhang, S.; Zhang, J.; Tian, B.; Lukasiewicz, T.; Xu, Z. Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation. *Med. Image Anal.* **2023**, *83*, 102656. [CrossRef]
32. Lou, S.; Chen, X.; Han, X.; Liu, J.; Wang, Y.; Cai, H. Fast retinal segmentation based on the wave algorithm. *IEEE Access* **2020**, *8*, 53678–53686. [CrossRef]
33. Pekala, M.; Joshi, N.; Liu, T.A.; Bressler, N.; DeBuc, D.C.; Burlina, P. Deep learning based retinal OCT segmentation. *Comput. Biol. Med.* **2019**, *114*, 103445. [CrossRef]
34. He, Y.; Carass, A.; Liu, Y.; Jedynak, B.M.; Solomon, S.D.; Saidha, S.; Calabresi, P.A.; Prince, J.L. Structured layer surface segmentation for retina OCT using fully convolutional regression networks. *Med. Image Anal.* **2021**, *68*, 101856. [CrossRef] [PubMed]
35. Moradi, M.; Chen, Y.; Du, X.; Seddon, J.M. Deep ensemble learning for automated non-advanced AMD classification using optimized retinal layer segmentation and SD-OCT scans. *Comput. Biol. Med.* **2023**, *154*, 106512. [CrossRef]
36. Kugelman, J.; Allman, J.; Read, S.A.; Vincent, S.J.; Tong, J.; Kalloniatis, M.; Chen, F.K.; Collins, M.J.; Alonso-Caneiro, D. A comparison of deep learning U-Net architectures for posterior segment OCT retinal layer segmentation. *Sci. Rep.* **2022**, *12*, 14888. [CrossRef]
37. Alryalat, S.A.; Al-Antary, M.; Arafa, Y.; Azad, B.; Boldyreff, C.; Ghnaimat, T.; Al-Antary, N.; Alfegi, S.; Elfalah, M.; Abu-Ameerh, M. Deep learning prediction of response to anti-vegf among diabetic macular edema patients: Treatment response analyzer system (tras). *Diagnostics* **2022**, *12*, 312. [CrossRef]
38. Liu, S.; Hu, W.; Xu, F.; Chen, W.; Liu, J.; Yu, X.; Wang, Z.; Li, Z.; Li, Z.; Yang, X.; et al. Prediction of OCT images of short-term response to anti-VEGF treatment for diabetic macular edema using different generative adversarial networks. *Photodiagnosis Photodyn. Ther.* **2023**, *41*, 103272. [CrossRef] [PubMed]
39. Zhang, Y.; Xu, F.; Lin, Z.; Wang, J.; Huang, C.; Wei, M.; Zhai, W.; Li, J. Prediction of Visual Acuity after anti-VEGF Therapy in Diabetic Macular Edema by Machine Learning. *J. Diabetes Res.* **2022**, *2022*, 5779210. [CrossRef] [PubMed]
40. Udaondo Mirete, P.; Muñoz-Morata, C.; Albarrán-Diego, C.; España-Gregori, E. Influence of Intravitreal Therapy on Choroidal Thickness in Patients with Diabetic Macular Edema. *J. Clin. Med.* **2023**, *12*, 348. [CrossRef]
41. Liu, B.; Zhang, B.; Hu, Y.; Cao, D.; Yang, D.; Wu, Q.; Hu, Y.; Yang, J.; Peng, Q.; Huang, M.; et al. Automatic prediction of treatment outcomes in patients with diabetic macular edema using ensemble machine learning. *Ann. Transl. Med.* **2021**, *9*, 43. [CrossRef] [PubMed]
42. Hu, T.; Qi, H.; Huang, Q.; Lu, Y. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. *arXiv* **2019**, arXiv:1901.09891.
43. Kendall, A.; Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? *Adv. Neural Inf. Process. Syst.* **2017**, *30*. Available online: https://proceedings.neurips.cc/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html (accessed on 1 January 2020).
44. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
45. Yu, L.; Wang, S.; Li, X.; Fu, C.W.; Heng, P.A. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In *Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 605–613.

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