Mathematical Biosciences and Engineering

MBE, 21(1): 1110–1124.
DOI: 10.3934/mbe.2024046
Received: 29 July 2023
Revised: 05 November 2023
Accepted: 28 November 2023
Published: 25 December 2023

http://www.aimspress.com/journal/MBE

Research article

DCT-Net: An effective method to diagnose retinal tears from B-scan ultrasound images

Ke Li[1,†], Qiaolin Zhu[3,†], Jianzhang Wu[2,3,†], Juntao Ding[1], Bo Liu[1], Xixi Zhu[3], Shishi Lin[3], Wentao Yan[3,*] and Wulan Li[1,*]

[1] The First Affiliated Hospital of Wenzhou Medical University, Wenzhou 325035, China
[2] Oujiang Laboratory (Zhejiang Lab for Regenerative Medicine, Vision and Brain Health), Wenzhou 325000, China
[3] The Eye Hospital, School of Ophthalmology & Optometry, Wenzhou Medical University, Wenzhou 325027, China

† These authors contributed to this work equally.

* **Correspondence:** Email: lwlwzmu@163.com, yanwentao0625@163.com.

Abstract: Retinal tears (RTs) are usually detected by B-scan ultrasound images, particularly for individuals with complex eye conditions. However, traditional manual techniques for reading ultrasound images have the potential to overlook or inaccurately diagnose conditions. Thus, the development of rapid and accurate approaches for the diagnosis of an RT is highly important and urgent. The present study introduces a novel hybrid deep-learning model called DCT-Net to enable the automatic and precise diagnosis of RTs. The implemented model utilizes a vision transformer as the backbone and feature extractor. Additionally, in order to accommodate the edge characteristics of the lesion areas, a novel module called the residual deformable convolution has been incorporated. Furthermore, normalization is employed to mitigate the issue of overfitting and, a Softmax layer has been included to achieve the final classification following the acquisition of the global and local representations. The study was conducted by using both our proprietary dataset and a publicly available dataset. In addition, interpretability of the trained model was assessed by generating attention maps using the attention rollout approach. On the private dataset, the model demonstrated a high level of performance, with an accuracy of 97.78%, precision of 97.34%, recall rate of 97.13%, and an F1 score of 0.9682. On the other hand, the model developed by using the public funds image dataset demonstrated an accuracy of 83.82%, a sensitivity of 82.69% and a specificity of 82.40%.

The findings, therefore present a novel framework for the diagnosis of RTs that is characterized by a high degree of efficiency, accuracy and interpretability. Accordingly, the technology exhibits considerable promise and has the potential to serve as a reliable tool for ophthalmologists.

Keywords: retinal tears; ultrasound image; automatic diagnosis; vision transformer; attention rollout

1. Introduction

Retinal tears arise from vitreous traction on the retina or degeneration and atrophy of the retina, and it is frequently observed in individuals who have acute posterior vitreous detachment [1]. The identification of retinal tears, which serve as a risk factor for the occurrence of retinal detachment, poses a significant challenge. In the absence of timely detection and intervention, 30–50% of the cases will progress to retinal detachment [2], a condition that leads to severe blinding. In most cases, retinal tears can be diagnosed by using indirect fundoscopy in conjunction with scleral pressure examination [3]. However, in situations where the patient's refracting media is murky, B-scan ultrasound emerges as a viable option among the limited alternative diagnostic tools available. Moreover, ultrasound is also more accessible and less expensive than other types like OCT and ultra-wide-field imaging. It is widely prevalent and available in many local hospitals and primary community clinics. However, conventional manual methods require the involvement of highly skilled physicians to prevent their potential oversight or misdiagnosis [4]. In this context, only a few of the large hospitals in China have professional sonographers, as is the case in other developing countries and regions. As a result, the development of a model capable of automatically diagnosing retinal tears is critical and urgent [5].

Deep learning represents the most effective approach to automating the development of diagnostic systems. Previous studies have proposed a multitude of models, with predominant focus on the utilization of convolutional neural networks (CNNs) [6,7]. For example, Li et al. [8] screened for notable peripheral retinal lesions (NPRLs) by using numerous models, such as InceptionResNetV2, InceptionV3, ResNet50 and VGG16. Furthermore, with an accuracy of 79.8%, a system based on seResNet50 was developed by Zhang et al. [9] to screen numerous types of NPRLs. However, the inability of the CNN to capture long-distance image features hinders its continued development. In this context, Dosovitskiy et al. [10] proposed the vision transformer (ViT) as a solution to this problem, using the excellent transformer [11] from natural language processing as a point of reference. Subsequently, ViT was observed to outperform CNNs in a multitude of tests after self-attention methods were substituted for convolutional processes. Accordingly, several researchers have made efforts to implement the model in the treatment of ophthalmic disorders, particularly, retinal issues. Jiang et al. [12] employed a ViT to automatically identify normal eyes, age-related macular degeneration, and diabetic macular edema, achieving a classification accuracy of 99.69%. Furthermore, a deep learning model based on a ViT was introduced by Wu et al. [13] to assess diabetic retinopathy, and it realized an accuracy of 91.4% and a kappa score of 0.935. However, studies that report on the automatic diagnosis of retinal tears are few.

The present study involved the collection and construction of a retinal tear dataset comprising 1831 images, with the aim of developing more effective diagnostic algorithms. Despite the widely acknowledged fact that ViT is data-driven and performs exceptionally well with ample training data,

our study encountered a hurdle due to the limited availability of data. Although the use of transfer learning has been demonstrated to be able to partially address this challenge, it should be noted that this approach may not be sufficient and could potentially lead to an increase in computational resources. Consequently, a hybrid structure was devised to introduce inductive bias and enhance the model's adaptability to our limited dataset. Furthermore, through experimental analysis, it has been observed that the utilization of deformable convolution [14] affords superior adaptability to the contour of lesions and yields improved performance. Thus, based on the aforementioned rationales, we proposed a novel framework called the deformable convolution and transformer network (DCT-Net) in the current study, which integrates the merits of deformable convolution and the vision transformer. The model was subjected to rigorous testing on two datasets to assess its overall performance and efficacy. Additionally, attention maps were generated in order to validate their interpretability. The current body of research on retinal tear diagnostic systems is limited, and our study has partially addressed this research gap.

To summarize, the main contributions of the present study can be succinctly stated as follows:

• A dataset comprising 1831 B-scan ultrasound images of retinal tears was assembled.

• A novel model that is more appropriate for small datasets of medical images is proposed. To our knowledge, this study represents the first investigation into the utilization of ViT-based architecture for the purpose of identifying retinal tears through the analysis of ultrasound images.

• The efficacy of the model in terms of lesion detection, as well as its commendable performance, are demonstrated through the analysis of two datasets.

2. Materials and methods

The contents of the current study can be categorized into three primary modules: data collection and preprocessing; model design and validation and interpretability analysis and external validation. The flowchart is illustrated in Figure 1.

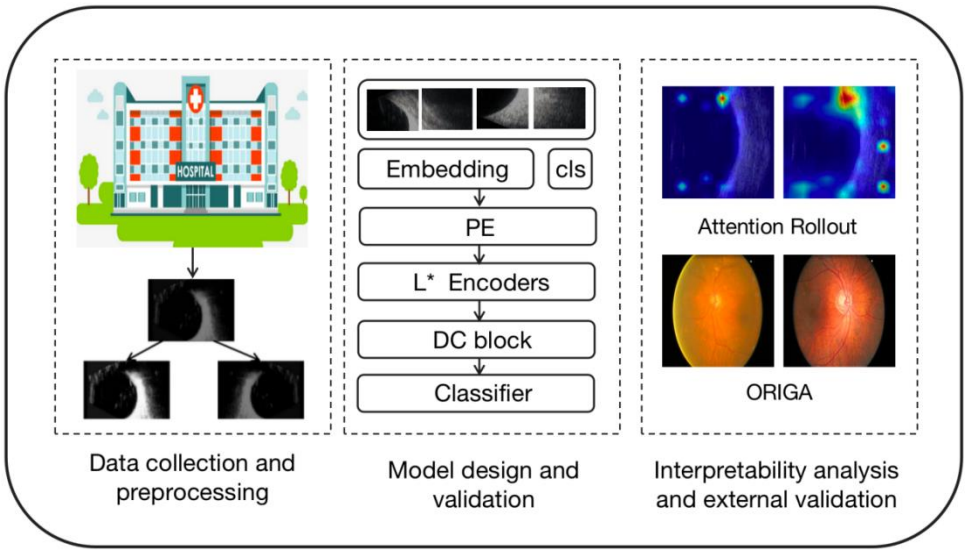

Figure 1. The flowchart of this research.

2.1. Datasets

2.1.1. Data collection

The investigation was carried out in adherence to the Protocol for the Declaration of Helsinki, as amended in 2013.

A comprehensive set of 1902 ultrasound B-scan images was collected for this retrospective study. These samples were obtained from the eye hospital of Wenzhou Medical University for the period from October 2017 to April 2022. All positive samples were verified by professional ophthalmologists. However, the images were collected from a variety of devices with varying resolutions and file types. Thus, to accommodate the model's input, each image underwent a resizing process to 224 × 224 pixels, and any blurry pixels were removed. Finally, 1831 samples (910 positive and 927 negative) were utilized for subsequent investigations.

2.1.2. Data augmentation

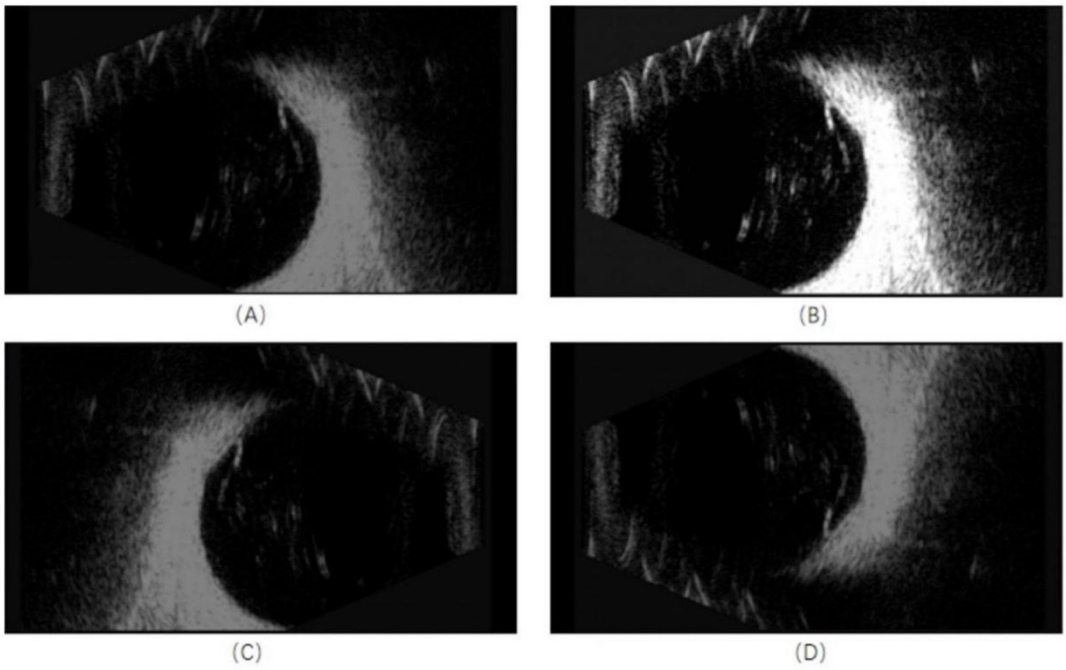

Figure 2. Three image augmentation methods. A. Original image; B. Brightness shift; C. Horizontal flip; D. Vertical flip.

Data augmentation is a data processing technique that is employed to enhance the quantity and diversity of training samples by transforming existing data. There are two distinct categories of data arguments, namely, augment online and augment offline. Typically, the former approach is utilized for larger datasets, wherein operations are executed on the data batch. Conversely, the latter approach is employed for smaller datasets, wherein operations are directly performed on the original data [15]. Accordingly, the offline method was selected as a result of the limited dataset available for our study. Various data augmentation techniques, including rotation, cropping,

brightness shift, contrast modification, horizontal flipping, vertical flipping, etc., can be employed for image augmentation [16,17]. However, not all enhancement techniques are universally applicable, because the labels of the image categories could be modified after enhancement. After conducting analysis, we opted to employ horizontal flip, vertical flip and brightness shift techniques in order to enhance the original dataset. Figure 2 illustrates the aforementioned augmentation operations.

2.2. DCT-Net

The ViT model is based on direct global relationship modeling and has demonstrated significant accomplishments in the extraction of global features through the use of a multi-head self-attention mechanism. However, it has limitations in its ability to effectively accommodate minuscule lesions, and it proves inadequate when confronted with a limited size of training data. In this context, convolution operations, specifically deformable convolutions, exhibit better adaptability to local detail characteristics. This study presents a novel approach that integrates the ViT and deformable convolution to realize the accurate detection of retinal tears with enhanced precision. Figure 3 presents a visual representation of the proposed model. Furthermore, the utilization of transfer learning technology was employed in this particular aspect to enhance network performance and expedite the training process.

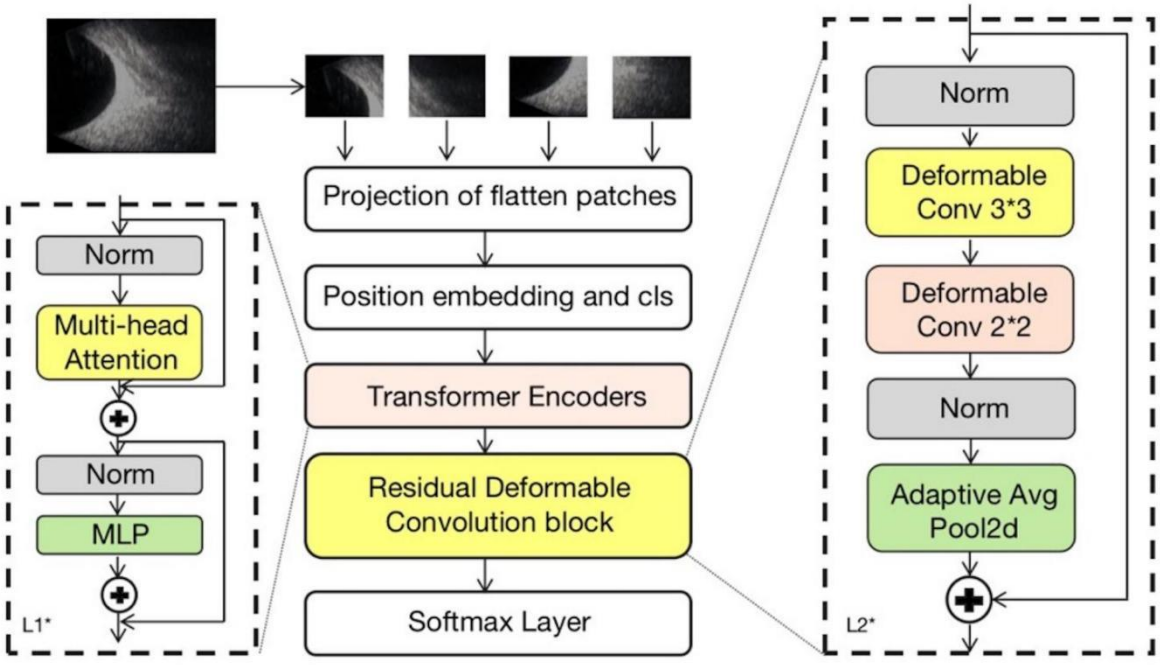

Figure 3. The proposed DCT-Net for retinal tear detection. After entering the classification model, the sample images were successively passed through the feature extractor and the residual deformable convolution block. Finally, the results were obtained as an output through a Softmax layer.

2.2.1. Transformer encoder

The input images ($H \times W \times C$) were split into *n* patches. After these patches were flattened, a linear projection layer was used to convert them to *D*-dimensional vectors. A class token was also appended, as illustrated in the BERT [18]. Following position embedding, the *D*-dimensional vectors were subsequently transmitted to the Transformer Encoder. Maintaining the dimensions of the vectors was crucial throughout the entire process.

In the Transformer Encoder, the input vectors undergo an initial step of layer normalization, which expedites the convergence of the network. The procedure is denoted by Eq (1) in terms of the mean and standard deviation of the input, respectively.

$$\mathrm{LayerNorm}(x_i) = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}} \tag{1}$$

The resulting output is used to compute the mutual attention by utilizing multi-head attention layers (as demonstrated in Eqs (2)–(4)). Subsequently, the Layer Norm and Multi-Layer Perceptron layer were employed to obtain the final outputs. The inclusion of residual connections in this process effectively mitigated the issue of gradient vanishing. To optimize the utilization of the transfer learning's weight, we employed an equal number of encoders as the conventional ViT model.

$$Q_i = QW_i^Q,\ K_i = KW_i^K,\ V_i = VW_i^V \tag{2}$$

$$\mathrm{head}_i = \mathrm{Attention}(Q_i, K_i, V_i) \tag{3}$$

$$\mathrm{Multihead}(Q, K, V) = \mathrm{Concact}(\mathrm{head}_1, \mathrm{head}_2, \ldots, \mathrm{head}_{12}) \tag{4}$$

$$y(P_0) = \sum_{P_n \in R} w(P_n) \cdot x(P_0 + P_n + \Delta P_n) \tag{5}$$

2.2.2 Deformable convolution block

The diagnosis of retinal tears using ultrasound images is highly dependent on the position and shape of the small lesion areas. However, the standard ViT is insufficient for acquiring such localized data. As a result of conventional convolution employing regular kernels, the receptive field remains constant and is ill-equipped to accommodate variations in edge shape. By appending a learnable offset to the standard convolution kernel, deformable convolution can modify the sampling area's shape, bringing it closer to the object's edge. The sampling procedure for deformable convolution and ordinary convolution is presented in Figure 4. Equation (5) illustrates the calculation process.

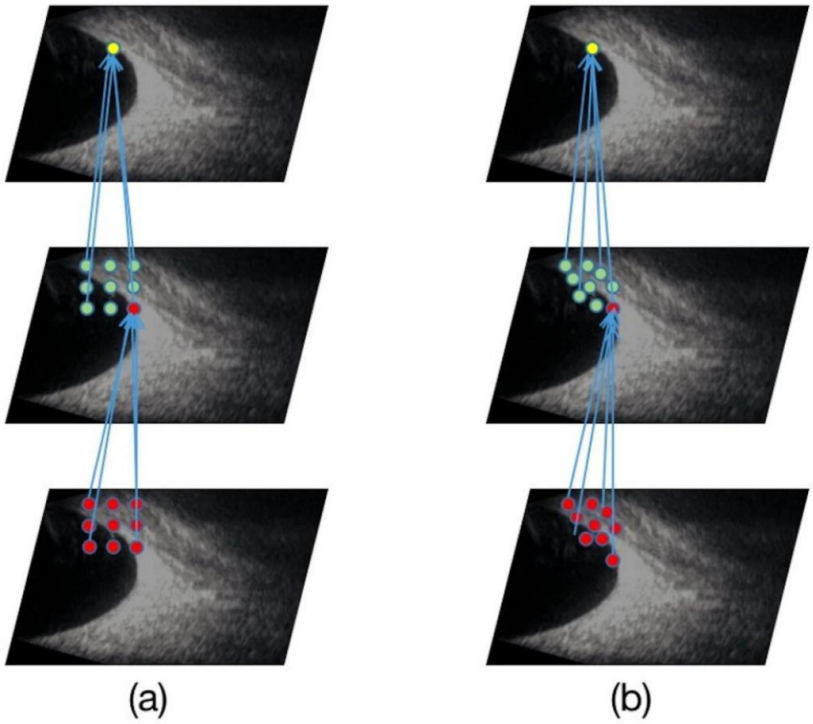

Figure 4. Sampling process. (a) Common convolution and (b) deformable convolution. The top image shows the result of employing the activation unit on objects. The middle image shows the result of the sampling process performed to obtain the top-level activation unit. The bottom image was used to obtain the sampling area for the middle image.

Subsequently, a residual deformable convolution block was devised in order to enhance the extraction of intricate features. Similar to the Transformer Encoder, the designed module initially employs a Batch Norm layer to convert inputs into data with a mean of 1 and a variance of 0. Two deformable convolutional layers were used to capture local concrete detail features. To enhance nonlinearity while minimizing computational workload, the convolutional kernel of the first layer was designed to be larger than that of the second layer. Subsequently, an adaptive average pooling layer was incorporated in order to enhance the efficacy of feature extraction and computational processes. Furthermore, the concept of residual connection was incorporated into the model design, drawing inspiration from Resnet [19]. This addition was made in order to mitigate the issue of gradient vanishing [20].

2.3. Interpretability analysis

The utilization of pooling layers in a CNN can lead to the merging of position information, potentially resulting in the loss of certain details during the generation of rough heat maps [21,22]. Our model effectively captures global features and is founded upon a self-attention mechanism. Moreover, it has the ability to deliver elaborate visualizations to an adequate degree [23]. However, attention-based networks are incompatible with the traditional Grad-CAM [24] method. This is attributed to the fact that the CNN permits the aggregation of feature map weights from multiple channels, whereas the ViT restricts the addition of distinct patches. Therefore, we adopted the

attention rollout method proposed by Samira Abnar [25]. Attention rollout in essence calculates the product of the attention matrix from the low level to the high level of the network. The concrete realization is achieved through the recursive calculation of each layer's tokens, computing information from the input layer to the higher level. Concurrently, the residual connection and the weight must be taken into account. It is represented by Eq (6).

$$\mathrm{AttentionRollout_L} = (\mathrm{A_L} + \mathrm{I})\mathrm{AttentionRollout_{L-1}} \tag{6}$$

where A_{L} is the attention matrix of the L layer and I is the identity matrix.

3. Results

3.1. Training strategy

The adoption of a transfer learning strategy was implemented with the aim of expediting the training process and enhancing the performance of the model. The pre-training process was conducted by using the ImageNet dataset, which comprises a vast collection of more than 1000 categories of nature images. The cross-entropy loss [26,27] was employed as the loss function in our study. This choice was made to address the issue of the sigmoid function's derivative form, which is susceptible to saturation and results in slow gradient updates. Furthermore, the Adam optimizer [28] was also utilized. The approach offers the benefits of rapid convergence and a relatively facile process for configuring hyperparameters.

Furthermore, an early stopping strategy was developed with the intention of mitigating the issue of overfitting. Following each iteration of training, a comprehensive evaluation was conducted on the designated test dataset. The training process was deemed to be complete once the accuracy on the test set ceased to exhibit substantial improvements and stabilized after approximately 10 epochs.

3.2. Performance on private datasets

In order to enhance the precision of an evaluation of the performance of the designed model, a set of widely recognized state-of-the-art (SOTA) models, viz. Alexnet [29], Inception v3 [30], Resnet101 [19], VGG16 [31] and ViT, were chosen as the baseline models. The preprocessing steps and training strategies remained consistent across all baselines, with the exception of Inception v3, which required an input size of 299 × 299 pixels.

Table 1 presents a comprehensive overview of the performance metrics for both the baseline models and the model that has been specifically designed for this study. The confusion matrix for multiple models on the test set is depicted in Figure 5. The number in each small square represents the corresponding number of images with the same predicted true label and it is the percentage of the total number of images under the true label. It is worth mentioning that within the category of CNN-based models, Inception v3 exhibited the highest level of performance, achieving an accuracy rate of 96.82%, an F1 score of 0.9605 and an AUC of 0.9828. The ViT model with the pure self-attention mechanism did not perform well; particularly, the performance was even worse than that of the CNN. Nevertheless, our designed model exhibited superior performance across all metrics, surpassing all other models, and only a mere 10 samples were classified incorrectly. To our

knowledge, the proposed model exhibited superior performance even as compared to human experts (with a sensitivity of 96%) [32].

Table 1. Performance comparison of DCT-Net with baseline models on the classification problem.

Model	Accuracy	Precision	Recall	F1 Score	AUC
Alexnet	95.11%	94.64%	95.68%	0.9456	0.9286
Inception V3	96.82%	96.55%	96.37%	0.9605	0.9828
Resnet101	96.74%	96.94%	96.42%	0.9599	0.9772
VGG16	96.52%	96.42%	96.66%	0.9595	0.9598
Vit	95.76%	95.66%	95.87%	0.9515	0.9444
DCT-Net	97.78%	97.34%	97.13%	0.9682	1.0000

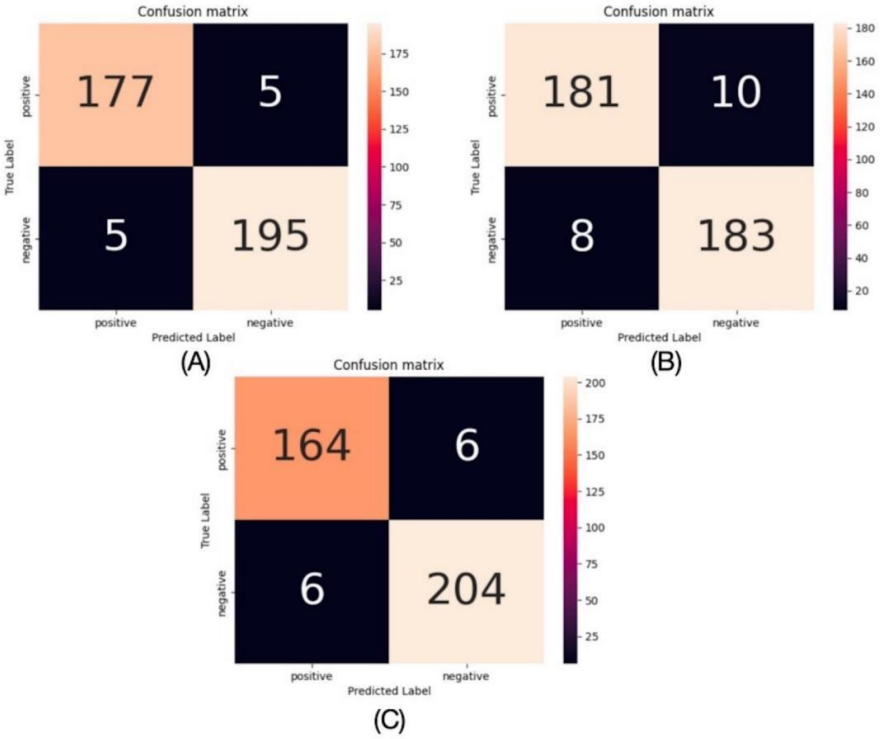

Figure 5. The confusion matrix for different models on retinal tear datasets. A. Inception V3; B. Vision transformer; C. DCT-Net.

3.3. External validation

As an external validation step, we utilized the ORIGA datasets in this section to ensure that the proposed model possesses exceptional generalizability and can adapt to various database types. The dataset comprised a total of 650 images depicting instances of glaucoma. In order to conduct a comparative analysis against other models documented in the literature [33–35], we used the original dataset without employing any augmentation techniques. Table 2 shows the results, where NMD denotes that the pre-training was performed by using a non-medical dataset, SOD denotes that the pre-training was performed by using a similar ophthalmic dataset and CT-Net denotes that common

convolution replaced the deformable convolution. The ViT did not perform well among them, most likely as a result of the limited dataset. On the other hand, the DCT-Net achieved the highest accuracy at 83.8%, demonstrating the best performance. Additionally, the significance of deformable convolution became apparent when it was compared to CT-Net.

Table 2. Performance comparison of the DCT-Net with others on the ORIGA dataset.

Model	Accuracy	Sensitivity	Specificity
CNN	70.4%	70.7%	74.8%
VGG	70.1%	69.8%	71.0%
GoogLeNet	71.8%	69.8%	73.5%
ResNet	71.5%	71.3%	71.7%
Chen [34]	70.8%	69.2%	71.0%
Shibata [35]	73.3%	73.2%	76.7%
NMD+CNN	74.5%	68.7%	80.7%
SOD+CNN	73.9%	80.9%	72.2%
NMD+Attention	74.9%	71.2%	77.7%
Xu [33]	76.6%	75.3%	77.2%
ViT	71.4%	74.0%	67.8%
CT-Net	**80.5%**	**81.7%**	**80.1%**
DCT-Net	**83.8%**	**82.7%**	**82.4%**

3.4. Interpretation

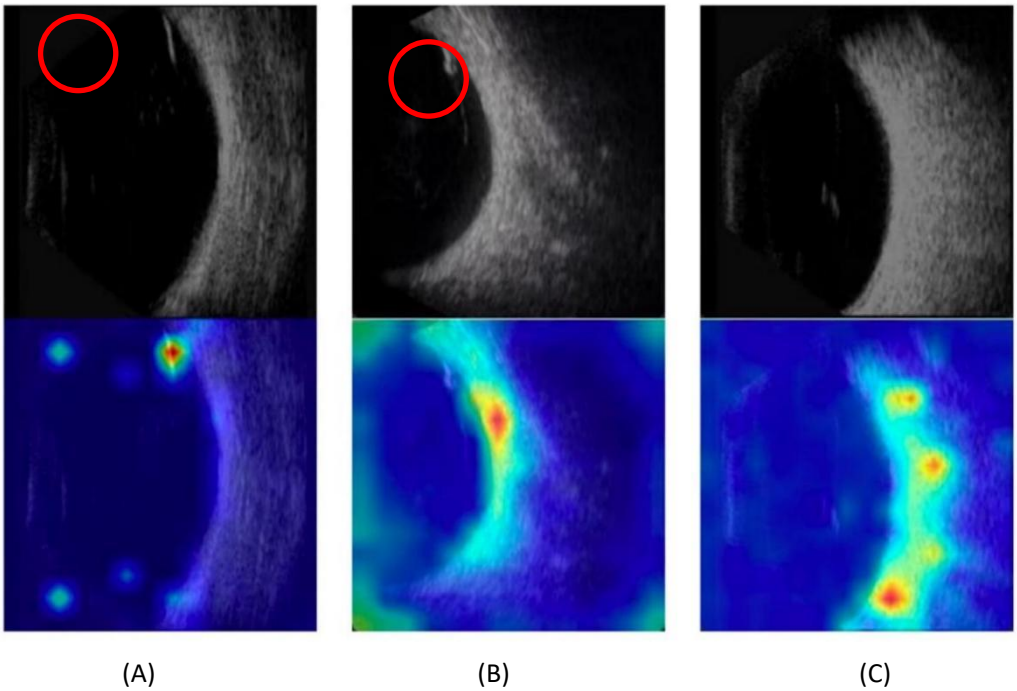

Figure 6. The attention maps for three samples. (A) and (B) are the lesion images and (C) is the normal image.

Models that are easily interpretable offer valuable insights into their inner workings, thereby benefiting both patients and clinicians. Figure 6 displays three attention maps that were generated by using our private dataset. We have used the red circle to mark the lesion parts in the original image. In the attention maps, higher intensity of color is indicative of a greater level of attention. The aforementioned images demonstrate a strong correspondence between the regions of heightened attention and the affected areas of the lesion. This indicated that the model possesses a well-defined operational framework and possesses exceptional interpretive qualities.

3.5. Hardware

The hardware configuration utilized in this study is as follows. The central processing unit (CPU) utilized in the system comprised a 7-core Intel$^{(R)}$ Xeon$^{(R)}$ CPU E5-2680 v4 operating at a frequency of 2.40 GHz. Additionally, the system incorporated a single graphics processing unit in the form of an RTX 3070ti with 8 GB of dedicated memory. The training process employed Python version 3.8, PyTorch framework version 1.10.0 for machine learning and CUDA version 11.3.

4. Discussion

CNNs have demonstrated remarkable performance on previous image processing tasks and are widely acknowledged as the SOTA approach. For instance, Yu et al. [36,37] employed CNNs for the purpose of detecting concrete cracks, achieving exceptional performance. Ragupathy and Karunakaran [38] proposed a CNN-based model for the detection of meningioma brain tumors. The model demonstrated promising performance metrics. However, due to the constraints imposed by the small convolutional kernel, CNNs may not be able to effectively extract global features. As shown in Table 1, it appears that the performance of the CNN-based model has encountered a bottleneck, making further improvements challenging. When comparing the CNN with the ViT, it can be observed that the ViT utilizes the attention mechanism to calculate the relationship between global pixels, thereby enabling a comprehensive global perspective. Numerous studies have substantiated the impressive efficacy of the ViT model [39]. However, our investigation revealed that the pure ViT did not perform well on small datasets of retinal tears (with the accuracy of 95.76%).

To enhance the efficacy of lesion detection on limited datasets, a novel architecture was initially devised, integrating the merits of convolution and attention mechanisms. As shown in Table 2, the utilization of global feature extraction techniques contributes to the generation of a relatively comprehensive latent space feature representation. Concurrently, as a result of incorporating the inductive bias of convolution, the proposed model demonstrates substantial enhancements on the limited public dataset, achieving an accuracy of 80.5%. Moreover, replacing ordinary convolutions with deformable convolutions has been found to yield more favorable outcomes, as evidenced by an accuracy rate of 83.8%. This phenomenon could potentially be attributed to the enhanced precision resulting from extracting both the location and shape of the lesion areas. From the perspective of external validation and interpretable analysis, the model possesses robustness and sufficient accuracy.

Notwithstanding the enhanced performance achieved in this study, certain constraints remain. First, ophthalmic ultrasound is highly dependent on the equipment, technique and examiner experience. However, the data collected for this study came from a variety of devices. This may

compromise the validity of the results. Second, all of the retinal tear images utilized in this study were procured from a single hospital. This may lead to an absence of diversity in the cases. Moreover, only retinal tears were included in our study. Ultrasound imaging can, in fact, be utilized to diagnose additional retinal disorders. Correspondingly, the value of the model can be enhanced through the incorporation of additional disease types. Finally, the incorporation of the residual deformable convolution module and the utilization of a ViT as the feature extractor resulted in an increased number of parameters for our model (Table 3). This results in increased demands on the environment in terms of model deployment.

Utilizing ultrasound to identify retinal tears is an extremely practical method. It is superior to alternative approaches when it comes to handling intricate clinical scenarios, such as ocular media opacity. However, the extraction of useful features via conventional machine learning methods is hampered by low resolution. Fortunately, the progress that has been made in deep learning enables the analysis of these images in an efficient manner. Our current research is, without a doubt, preliminary in nature. Moving forward, we aim to enhance the model's architecture and implement global vision technology that is more streamlined or possesses a reduced number of parameters. This will allow the effortless deployment of lightweight models across diverse environments. Furthermore, our objective is to enhance the quantity and range of samples gathered in order to prevent issues with model generalization that may arise from discrepancies in the training data. Lastly, we will collaborate with clinicians and conduct additional multicenter studies to precisely quantify the extent to which this model can benefit physicians.

Table 3. Parameters of the different models used in the study.

Model	Parameters (1×10^6)
Alexnet	57.01
Inception v3	25.12
Resnet101	42.5
VGG16	134.27
Vision Transformer	85.80
DCT-Net	138.36

5. Conclusions

A novel model was developed for the diagnosis of ophthalmological conditions in the current study. The model demonstrated superior performance on both our proprietary dataset and the glaucoma dataset that was publicly available. The framework is a comprehensive computing framework that exhibits superior performance and does not necessitate the generation of manually designed features. Overall, this technology provides significant practical value in the field of clinical application, particularly in the realm of automated diagnosis.

Use of AI tools declaration

The authors declare that they have not used artificial intelligence tools in the creation of this article.

Acknowledgments

We would like to thank all editors and reviews for their careful review and revision of the paper. This research was supported in part by the National Key R&D Program of China [2018YFA0701700].

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. N. E. Byer, Natural history of posterior vitreous detachment with early management as the premier line of defense against retinal detachment, *Ophthalmology*, **101** (1994), 1503–1513. https://doi.org/10.1016/s0161-6420(94)31141-9
2. J. Lorenzo-Carrero, I. Perez-Flores, M. Cid-Galano, M. Fernandez-Fernandez, F. Heras-Raposo, R. Vazquez-Nuñez, et al., B-scan ultrasonography to screen for retinal tears in acute symptomatic age-related posterior vitreous detachment, *Ophthalmology*, **116** (2009), 94–99. https://doi.org/10.1016/j.ophtha.2008.08.040
3. J. AMDUR, A method of indirect ophthalmoscopy, *Am. J. Ophthalmol.*, **48** (1959), 257–258. https://doi.org/10.1016/0002-9394(59)91247-4
4. K. E. Yong, Enhanced depth imaging optical coherence tomography of choroidal nevus: Comparison to B-Scan ultrasonography, *J. Korean Ophthalmol. Soc.*, **55** (2014), 387–390. https://doi.org/10.3341/jkos.2014.55.3.387
5. M. S. Blumenkranz, S. F. Byrne, Standardized echography (ultrasonography) for the detection and characterization of retinal detachment, *Ophthalmology*, **89** (1982), 821–831. https://doi.org/10.1016/S0161-6420(82)34716-8
6. H. C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, et al., Deep convolutional neural networks for Computer-Aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans. Med. Imaging*, **35** (2016), 1285–1298. https://doi.org/10.1109/TMI.2016.2528162
7. M. Chiang, D. Guth, A. A. Pardeshi, J. Randhawa, A. Shen, M. Shan, et al., Glaucoma expert-level detection of angle closure in goniophotographs with convolutional neural networks: the Chinese American eye study: Automated angle closure detection in goniophotographs, *Am. J. Ophthalmol.*, **226** (2021), 100–107. https://doi.org/10.1016/j.ajo.2021.02.004
8. Z. Li, C. Guo, D. Lin, Y. Zhu, C. Chen, L. Zhang, et al., A deep learning system for identifying lattice degeneration and retinal breaks using ultra-widefield fundus images, *Ann. Transl. Med.*, **7** (2019), 618. https://doi.org/10.21037/atm.2019.11.28
9. C. Zhang, F. He, B. Li, H. Wang, X. He, X. Li, et al., Development of a deep-learning system for detection of lattice degeneration, retinal breaks, and retinal detachment in tessellated eyes using ultra-wide-field fundus images: a pilot study, *Graefes Arch. Clin. Exp. Ophthalmol.*, **259** (2021), 2225–2234. https://doi.org/10.1007/s00417-021-05105-3
10. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., An image is worth 16x16 words: Transformers for image recognition at scale, preprint, arXiv: 2010.11929.

11. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, preprint, arXiv: 1706.03762.
12. Z. Jiang, L. Wang, Q. Wu, Y. Shao, M. Shen, W. Jiang, et al., Computer-aided diagnosis of retinopathy based on vision transformer, *J. Innov. Opt. Health Sci.*, **15** (2022), 2250009. https://doi.org/10.1142/S1793545822500092
13. J. Wu, R. Hu, Z. Xiao, J. Chen, J. Liu, Vision Transformer-based recognition of diabetic retinopathy grade, *Med. Phys.*, **48** (2021), 7850–7863. https://doi.org/10.1002/mp.15312
14. J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, et al., Deformable convolutional networks, preprint, arXiv: 1703.06211.
15. P. T. Jackson, A. A. Abarghouei, S. Bonner, T. P. Breckon, B. Obara, Style augmentation: data augmentation via style randomization, *CVPR Workshops*, **6** (2019), 10–11.
16. Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, in *Proceedings of the AAAI conference on artificial intelligence*, **34** (2020), 13001–13008.
17. C. Bowles, L. Chen, R. Guerrero, P. Bentley, R. Gunn, A. Hammers, et al., Gan augmentation: Augmenting training data using generative adversarial networks, preprint, arXiv: 1810.10863.
18. J. Devlin, M. W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in *Proceedings of naacL-HLT*, **1** (2019), 2.
19. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE conference on computer vision and pattern recognition,* (2016), 770–778.
20. S. Hochreiter, The vanishing gradient problem during learning recurrent neural nets and problem solutions, *Int. J. Uncertainty Fuzziness Knowledge Based Syst.*, **6** (1998), 107–116. https://doi.org/10.1142/S0218488598000094
21. P. Murugan, S. Durairaj, Regularization and optimization strategies in deep convolutional neural network, preprint, arXiv: 1712.04711.
22. C. C. J. Kuo, M. Zhang, S. Li, J. Duan, Y. Chen, Interpretable convolutional neural networks via feedforward design, preprint, arXiv: 1810.02786.
23. Z. Zhang, H. Zhang, L. Zhao, T. Chen, S. Ö. Arik, T. Pfister, Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding, preprint, arXiv: 2105.12723.
24. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Why did you say that? Visual explanations from deep networks via gradient-based localization, preprint, arXiv: 1610.02391.
25. S. Abnar, W. Zuidema, Quantifying attention flow in transformers, preprint, arXiv: 2005.00928.
26. M. C. Dickson, A. S. Bosman, K. M. Malan, Hybridised loss functions for improved neural network generalisation, preprint, arXiv: 2204.12244.
27. C. Ma, D. Kunin, L. Wu, L. Ying, Beyond the quadratic approximation: the multiscale structure of neural network loss landscapes, preprint, arXiv: 2204.11326.
28. S. J. Reddi, S. Kale, S. Kumar, On the convergence of adam and beyond, preprint, arXiv: 1904.09237.
29. A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, *Adv. Neural Inform. Process. Syst.*, **25** (2012), 2.
30. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2016), 2818–2826.

31. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, preprint, arXiv: 1409.1556.
32. J. Lorenzo-Carrero, I. Perez-Flores, M. Cid-Galano, M. Fernandez-Fernandez, F. Heras-Raposo, R. Vazquez-Nuñez, et al., B-scan ultrasonography to screen for retinal tears in acute symptomatic age-related posterior vitreous detachment, *Ophthalmology*, **116** (2009), 94–99. https://doi.org/1016/j.ophtha.2008.08.040
33. X. Xu, Y. Guan, J. Li, Z. Ma, L. Zhang, L. Li, Automatic glaucoma detection based on transfer induced attention network, *Biomed. Eng. Online*, **20** (2021), 1–19. https://doi.org/10.1186/s12938-021-00877-5
34. X. Chen, Y. Xu, S. Yan, D. W. K. Wong, T. Y. Wong, J. Liu, Automatic feature learning for glaucoma detection based on deep learning, in *Medical Image Computing and Computer-Assisted Intervention*, **18** (2015).
35. N. Shibata, M. Tanito, K. Mitsuhashi, Y. Fujino, M. Matsuura, H. Murata, et al., Development of a deep residual learning algorithm to screen for glaucoma from fundus photography, *Sci. Rep.*, **8** (2018),14665. https://doi.org/10.1038/s41598-018-33013-w
36. Y. Yu, M. Rashidi, B. Samali, M. Mohammadi, T. N. Nguyen, X. Zhou, Crack detection of concrete structures using deep convolutional neural networks optimized by enhanced chicken swarm algorithm, *Struct. Health Monit.*, **5** (2022), 2244–2263. https://doi.org/10.1177/14759217211053546
37. Y. Yu, B. Samali, M. Rashidi, M. Mohammadi, T. N. Nguyen, G. Zhang, Vision-based concrete crack detection using a hybrid framework considering noise effect, *J. Build Eng.*, **61** (2022), 105246. https://doi.org/10.1016/j.jobe.2022.105246
38. B. Ragupathy, M. Karunakaran, A fuzzy logic-based meningioma tumor detection in magnetic resonance brain images using CANFIS and U-Net CNN classification, *Int. J. Imaging Syst. Technol.*, **31**(2021), 379–390. https://doi.org/10.1002/ima.22464
39. Z. Jiang, L. Wang, Q. Wu, Y. Shao, M. Shen, W. Jiang, et al., Computer-aided diagnosis of retinopathy based on vision transformer, *J. Innov. Opt. Health Sci.*, **15** (2022), 2250009. https://doi.org/10.1142/S1793545822500092