



Fusing CNN and Transformer for Diabetic Retinopathy Image Grading

Haitao Yao, Ke Pan, Lijun He, and Jianxin Zhang^(✉)

School of Computer Science and Engineering, Dalian Minzu University, Dalian
116600, China
jxzhang0411@163.com

Abstract. Convolutional neural networks (CNN) are widely used for diabetic retinopathy (DR) aided diagnosis, but the CNN approach suffers from insufficient global feature extraction capability. In this paper, a DR grading method based on the fusion of CNN and Transformer network is proposed to assist doctors in implementing DR image diagnosis. The proposed uses a dual-branch network architecture, utilizing the lightweight EfficientNet model in CNN to better extract local features from retinal images by balancing network depth, width and resolution. Meanwhile, in another branch, the Swin Transformer with good translation invariance and hierarchy is introduced, thus capturing global features of DR images with the powerful global modeling ability. Then, the local and global features of the dual-branch are fused at the end of the network to achieve a more robust DR image representation. The method achieves 86.14% DR grading accuracy on the Aptos-sigmaX10 dataset, an improvement of 2.99% and 1.09% compared to using only CNN or Transformer, respectively.

Keywords: Network fusion · EfficientNet · Swin Transformer · Diabetic retinopathy

1 Introduction

As an incurable chronic disease, diabetes will not only cause blood glucose metabolism problems, but also cause chronic damage to eyes, kidneys, heart and other organs. Diabetic Retinopathy (DR) is a typical set of lesions caused by damage to the microvasculature of the retina as a result of diabetes [1]. The DR proportion of adult diabetes patients accounts for more than 40% of the total number of diabetes patients, which is one of the most common causes of adult blindness [2]. DR is not clinically evident in the early stages of the disease, and by the time it is detected it has already produced severe retinal disease. As there are still no good clinical treatment options for patients with advanced DR, early prevention of retinopathy has become very important [3]. At present, the diagnosis of DR relies mainly on the recognition of professional ophthalmologists, but as the diagnosis of retinal images is a difficult and time-consuming task, there is an urgent need for computer-aided retinal image diagnosis [4].

With the development of deep learning technology and its outstanding performance in various fields, researchers have attempted to introduce deep learning methods represented by CNN into DR image assisted diagnosis. Among them, Pratt et al. [5] proposed the use of multilayer convolutional neural network (CNN) for five classification prediction of DR images, which greatly improved the accuracy of computer processing of DR image classification task and demonstrated the effectiveness of using deep learning for DR classification. On this basis, Dekhil et al. [6] proposed a network model with five convolutional layers to further improve DR image grading accuracy through image preprocessing and network lightweighting. To enhance the network's ability to extract features from retinal images, Torre et al. [7] propose an interpretable classifier. The method uses convolution to adapt the network architecture so that the receptive field is as close as possible to the original image size, and the classification results are interpreted by assigning a score to each point in the hidden and input spaces. The method also incorporates information from binocular retinal images. With the wide application of attention mechanism on deep learning, Zhou et al. [8] proposed a DR image grading network based on the attention mechanism to enable more attention to be paid to the information of lesions in the retina during grading. To detect small lesions in retinal images, Gu et al. [9] instead proposed a network based on a multi-channel attentional selection mechanism. The method introduces sorting losses to optimise the amount of information in each layer of the channel, and improves the accuracy of classification by combining fine-grained classification methods with multi-channel attentional acquisition of local features. In addition, to overcome the problem that the cross-entropy loss function is sensitive to noisy data and hyperparameter changes, Islam et al. [10] proposed a two-stage comparison method having a supervised comparison loss function, which also achieved better performance in DR grading.

Although CNNs have shown good results in DR image grading tasks, CNN methods also suffer from a lack of ability to capture global features [11]. With the rapid development of the Transformer model based on the self-attention [12] mechanism, researchers have also attempted to introduce the Transformer into DR image analysis, taking advantage of its strengths in global feature modeling to complete DR image classification prediction [13]. Based on the above content, this article aims to consider a scheme that can balance global and local modeling for DR image classification. More specifically, we propose a dual branch DR image classification method that combines CNN and Transformer, as shown in Fig. 1, to improve task performance by combining the advantages of CNN and Transformer. The method takes the form of a typical dual-branch network architecture, with a lightweight EfficientNet [14] model architecture as the CNN branch, which has a good balance of network depth, width and image resolution, allowing it to effectively extract local features from retinal images. In the other branch, the Swin Transformer [15] model is used, which uses a sliding window mechanism in the local window to process the image, giving the network excellent translation invariance and multi-level feature representation, able to consider global features of the DR image while taking into account local features. Finally,

at the end of the two-branch network, local and global features are fused to obtain a more robust representation of the DR image. The method achieved an accuracy of 86.14% on the DR dataset Aptos-sigmaX10, achieving better performance compared to the CNN and Transformer methods.

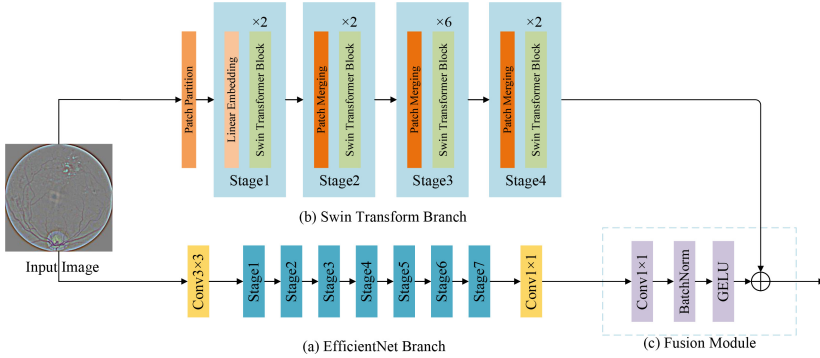


Fig. 1. CNN-Swin Transformer Network Structure.

2 Method

First, the principles of the EfficientNet and Swin Transformer models associated with this paper are presented in Sec. 2.1 and Sec. 2.2, then, a description of the specific model structure of this paper is given in Sec. 2.3, and the data pre-processing method is described in Sec. 2.4

2.1 EfficientNet

In CNNs, performance was usually optimised by adjusting the width and depth of the network model as well as the resolution of the images. Although this approach is simple and effective, it undoubtedly incurs further computational overhead. In order to achieve performance improvement while ensuring the portability of the model, EfficientNet [14] has brought a new approach that achieves a good balance between velocity and accuracy. EfficientNet uses a composite scaling approach, where the network width w , network depth d , and image resolution r are scaled uniformly by defining the parameter φ , achieving a high-precision and efficient balance through optimization of w , d , and r . With this in mind, EfficientNet is based on a baseline network structure, which can be scaled to generate a series of networks of different sizes by adjusting w , d and r . Reducing the number of parameters and computation while maintaining the performance benefits of the network.

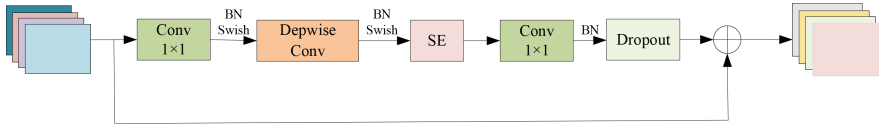


Fig. 2. MBConv structure diagram.

EfficientNet is mainly composed of a series of inverted bottleneck MBConv. The MBConv structure is shown in Fig. 2. The MBConv block, an EfficientNet-specific feature extraction structure, consists mainly of a 1×1 convolution kernel acting as an elevated tensor dimension and a $k \times k$ Depthwise Convolution for feature extraction. In addition, an channel attention module Squeeze-and-Excitation(SE) [16], a 1×1 convolution operation that acts to reduce the tensor dimension and a Dropout layer are included. Due to the use of the attention mechanism, EfficientNet also focuses more on relevant regions with more object detail than other models.

2.2 Swin Transformer

Transformer is a model based on a self-attentive mechanism that was originally applied to the field of natural language processing with advanced results. Subsequently, Dosovitskiy [17] et al. introduced the Transformer to the field of computer vision, proposing the Vision Transformer (ViT). ViT extracts image features by stacking Transformer blocks, which do not require convolution operations, and also performs well on computer vision tasks. ViT’s high computational volume is not suitable for the application. Swin Transformer [15] proposes a feature pyramid network architecture based on multi-scale hierarchical design, and designs a shift window to construct a multi-headed attention module, as shown in Fig. 3. This operation of paying attention in the local window enhances the Transformer’s ability to focus on finer local information. Also, this approach has enabled Swin Transformer to achieve higher accuracy and faster computational efficiency, achieving advanced results in multiple vision tasks.

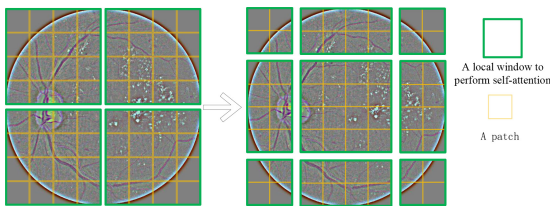


Fig. 3. Shifted Window.

Swin Transformer blocks are the most important component of the Swin Transformer. Swin blocks differ from ViT blocks in that they change the standard

Multi-Head Self-Attention (MSA) in ViT to Shifted Windows Multi-Head Self-Attention (W-MSA/SW-MSA), and by iteratively stacking Swin Transformer blocks, they not only improve the feature extraction of image blocks, but also establish dependencies between features of adjacent image blocks [18].

2.3 Overall Architecture

To more accurately detect DR patients, this paper proposes a deep learning dual-branch network model based on CNN and Transformer fusion from a model fusion perspective, enabling the network to make full use of both global and local features to achieve higher image classification results than both, providing ophthalmologists with better medical aid diagnostic references. The structure of the proposed fusion network model is shown in Fig. 1 and consists mainly of two branches, the CNN branch (a) and the Transformer branch (b), as well as a feature fusion module (c) for tensor adjustment.

In Fig. 1(a), the CNN model is chosen as EfficientNet-B4. The EfficientNet network is more effective in feature extraction by using a self-attentive module to focus on relevant regions with more informative features. EfficientNetB0-B7 is implemented by scaling adjustments on top of EfficientNet-B0. The network structure and parameters of the EfficientNet-B4 model, which was mainly used for the experiments, are shown in Table 1 for an input image resolution of 224×224 RGB images.

Table 1. The structure of EfficientNet-B4.

Stage	Operator	Resolution	Channels	Layers
1	Conv3 \times 3	224×224	48	1
2	MBCConv1, k3 \times 3	112×112	24	2
3	MBCConv6, k3 \times 3	112×112	32	4
4	MBCConv6, k5 \times 5	56×56	56	4
5	MBCConv6, k3 \times 3	28×28	112	6
6	MBCConv6, k5 \times 5	14×14	160	6
7	MBCConv6, k5 \times 5	14×14	272	8
8	MBCConv6, k3 \times 3	7×7	448	2
9	Conv1 \times 1 & Pooling & FC	7×7	1792	1

In Fig. 1(b), the Transformer model is chosen as swin-T. The Swin Transformer architecture is used, where the input RGB image (size $4 \times 4 \times 3$) is first processed into non-overlapping image blocks by Patch Partition, and then the image blocks are processed into feature dimension 48 ($4 \times 4 \times 3$). After processing, each image block has a dimension of $\frac{H}{4} \times \frac{W}{4} \times 48$. After linear transformation by Linear Embedding to expand the tensor dimension to an arbitrary size, the tensor is fed into the Swin blocks of the hierarchy to learn the image features

to obtain the complete image information. The tensor is first passed through a Linear Embedding layer, except for Stage 1, where a Patch Merging layer is applied to the feature map. The Patch Merging layer takes the small patches and combines them into one large patch to increase the field of perception and obtain more feature information from the image.

The feature fusion module corresponding to Fig. 1(c) contains mainly a 1×1 convolution of size, a BatchNorm layer and a GELU activation function, which serves to adapt the features to be fused. The CNN branch outputs a feature image with dimension $7 \times 7 \times 1792$ for an RGB image with input 224×224 and $7 \times 7 \times 768$ for the Transformer branch. The fusion method was chosen by adjusting the tensor to the same $7 \times 7 \times 768$ and then performing the fusion operation to obtain information with both local and global features of the image for DR grading.

In addition, the adopted classifier consists of a global average pooling layer and a fully connected layer, and is converted to classification probabilities using the Softmax function.

2.4 Data Preprocessing

Due to the different contrast of retinal images in the dataset, as well as the presence of black areas in most of the images, in addition to the problem of image blurring caused by differences in shooting height and angle, etc., which can affect the grading accuracy to a certain extent, data pre-processing of the DR images in the data is necessary for this reason.

Firstly, considering the large amount of black area noise present in the original DR image, the image was cropped to the maximum rectangle to obtain the effective image area. As the size specification of the cropped image is not uniform, the image is processed to a uniform size (pixels) after rounding the effective area of the retina using the binary method of circular cropping.

As in Eq. (1), based on the height and width of the image, the value of the binary method of locating the coordinates of the image centroid can be calculated.

$$\begin{cases} x = \frac{width}{2} \\ y = \frac{height}{2} \end{cases} \quad (1)$$

Based on the centre coordinates of the circle, the circle is drawn with the minimum height and width as the radius, as in Eq. (2), then a blank image of the same size as the original image is drawn, and the pixel values of the circular part of the original image are overlaid with this area in the blank image, resulting in a binary method circular cropped image.

$$radius = \frac{\min(width, height)}{2} \quad (2)$$

Thereafter, a Gaussian filter smoothing function is used to suppress other noise in the retinal image. Specifically, a Gaussian function is used as a template to do a convolution operation with the input image and a weighted average of the

image pixel values. The expression for the two-dimensional Gaussian function is given in Eq. (3):

$$G(x, y) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{3}$$

where x^2, y^2 represent the distance between the central pixel and the other pixel points in the field. σ represents the standard deviation. $G(x, y)$ indicates the weight of the calculated pixel point. When the size of the window template is $(2k + 1) \times (2k + 1)$ its value per pixel is calculated by Eq. (4)

$$H_{i,j} = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{(i-k-1)^2+(j-k-1)^2}{2\sigma^2}} \tag{4}$$

where i, j denote the coordinates of the ranks in the convolution kernel and k denotes the size of the convolution kernel.

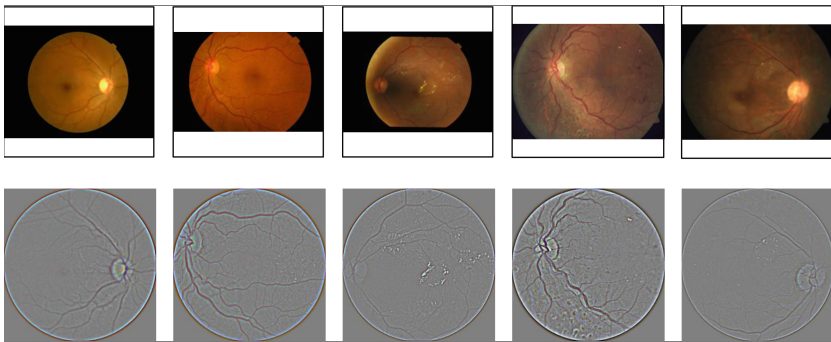


Fig. 4. Comparison chart before and after image pre-processing.

The resulting image in Fig. 4 shows the image in the dataset after image pre-processing. As can be seen from the figure, the retinal image features are enhanced to some extent after processing the cropped image by the Gaussian filter function. In addition, image enhancement strategies such as random level flipping and image rotation are used simultaneously to increase the generalisability of the network and to improve the training efficiency of the network.

3 Experiment Preparation

In Sect. 3.1 the dataset used in this experiment is first introduced, followed by Sect. 3.2 which describes the setup used in this experiment, and in Sect. 3.3 which describes the evaluation metrics used in this paper.

3.1 Dataset

The DR dataset Aptos_sigmaX10, publicly available on the Kaggle website¹ was used, which was filmed and provided by technicians at the Aravind Eye Hospital and examined and classified by specialist clinicians according to the actual degree of lesion.

Table 2. Distribution of the number of images in the Aptos dataset

Degree of lesion	Category labels	Number	Percentage
Non-Proliferative DR	0	1805	49.30%
Mild non-proliferative DR	1	370	10.10%
Moderate non-proliferative DR	2	999	27.30%
Severe non-proliferative DR	3	193	5.30%
Proliferative DR	4	295	8.10%

The dataset contains 3662 high-resolution colour retinal images, classified into five classes: Non-Proliferative DR, Mild non-proliferative DR, Moderate non-proliferative DR, Severe non-proliferative DR and Proliferative DR. We train in five classifications according to this scale. For the training data set the data distribution is shown in Table 2, with 1805 retinal images for Non-Proliferative DR (labeled 0), 370 retinal images for Mild non-proliferative DR (labeled 1), 999 retinal images for Moderate non-proliferative DR (labeled 2), 193 retinal images for Severe non-proliferative DR (labeled 3) and 295 retinal images for proliferative DR (labeled 4) 295 images. In this paper, the dataset was divided into a training set, a validation set and a test set in the ratio of 8:1:1 for model evaluation.

3.2 Implementation Details

The experiments were conducted in Ubuntu 20.04.2 on an Intel(R) Core(TM) i7-9700 CPU, RTX 2080 Ti GPU with 12 GB of video memory, and CUDA version is 11.4. The Transformer branch uses the officially published pre-trained weights for network initialization, and the CNN branch does not use migration learning for initialization. The model was trained using the AdamW [19] optimizer with a base learning rate set to $1e - 4$. The batch size was set to 16 and the number of training epochs to 100.

3.3 Evaluation Metrics

The DR image grading evaluation criteria used in the experiments were Confusion matrix. The confusion matrix can be used to reflect the relationship between

¹ Dataset courtesy of <http://www.kaggle.com>.

the predicted and true values, it mainly includes: TP (True Positive), FN (False Negative), FP (False Positive), TN (True Negative). Besides, precision, accuracy, and recall coefficient are also used to evaluate the fusion model.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

4 Experiment

To validate the effectiveness of the proposed method of grading diabetic retinopathy fusing CNN and Transformer, multiple sets of experimental tests were used on the dataset Aptos_sigmaX10.

4.1 Comparison of Different EfficientNet Models

The first set of experiments were selection experiments for different EfficientNet models. According to the results in Fig. 5, the EfficientNet-B4 model showed the best results with better feature extraction in the same DR dataset, achieving a grading accuracy of 83.15%, so in the subsequent use of EfficientNet-B4 for network fusion.

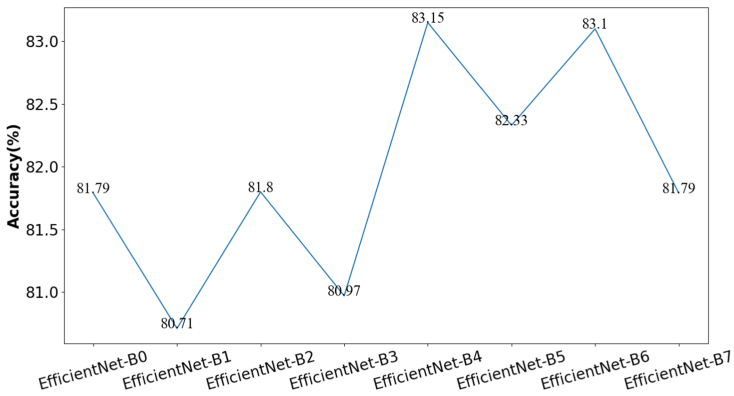


Fig. 5. Comparison of the grading effects of the EfficientNet series models.

4.2 Comparison of Different Swin Transformer Models

In addition to the ablation experiments carried out on EfficientNet, the same experiments was carried out using Swin Transformer for the dataset Aptos_sigmaX10, the results of which are shown in Fig. 6. The experimental results show that among the three Swin Transformer models, Swin-tiny can achieve the best accuracy results with a grading prediction accuracy of 85.06%. Therefore, we subsequently chose Swin-tiny as a complementary model to the global features of EfficientNet-B4.

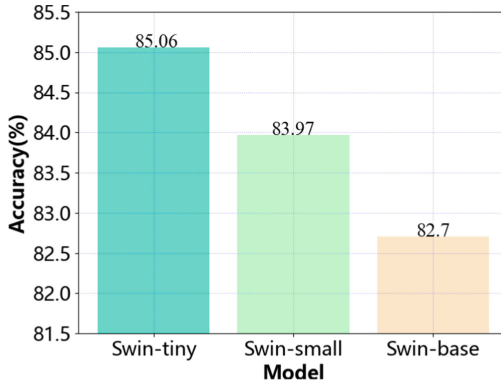


Fig. 6. Comparison of the grading effects of the Swin series models.

4.3 Fusion Model Precision Results

The third set of experiments aims to test the effectiveness of model fusion by comparing the fused model with a typical single network model in an experiment. The networks used for the tests mainly included ResNet, Vit, EfficientNet-B4 and Swin-tiny and the results of the comparison experiments obtained are shown in Table 3.

After an experimental comparison, the results showed that the best results were obtained with the model fusing CNN and Swin Transformer. The accuracy of the fusion model was improved by 1.74% compared to the ResNet network, and even more by 3.53% relative to the ViT network, and by 2.99% over the EfficientNet-B4 model, the best classification model in the EfficientNet family. In addition, the fusion model can be improved by 1.07% compared to the best performing Swin-tiny model in the Swin Transformer range. The results of the fusion model are not only higher than the accuracy of the EfficientNet model, but also higher than the accuracy of the Swin Transformer model, which proves the effectiveness of the fusion model of EfficientNet and Swin Transformer. it

Table 3. Comparative analysis table of the validity of the fusion model.

Models	Accuracy (%)
ResNet50 [20]	84.40%
ViT	82.61%
EfficientNet-B4	83.15%
Swin-tiny	85.06%
Ours	86.14%

can complement the global feature extraction ability of EfficientNet to a certain extent, and further improve the local feature extraction ability of the Swin Transformer model.

In addition, the confusion matrix results for the fusion model test results are given in Fig. 7. The figure shows that the fusion model correctly classified 179 of the 181 images non-Proliferative DR in the test set, 26 of the 37 images with mild non-proliferative DR, 88 of the 100 images with moderate non-proliferative DR, 7 of the 20 images with severe non-proliferative DR and 17 of the 30 images with proliferative DR. The best prediction results for non-Proliferative DR and the worst prediction results for severe non-proliferative DR may be due to the relatively large proportion of non-Proliferative DR images in the dataset, resulting in the model learning more information about that category during training; Additionally 50% of the severe non-proliferative lesions were misclassified as moderate non-proliferative lesions, mainly due to the small number of severe non-proliferative lesions in the dataset and the similarity of lesion information between moderate and severe non-proliferative lesions, such that a large number of images of severe non-proliferative lesions were judged as moderate non-proliferative lesions.

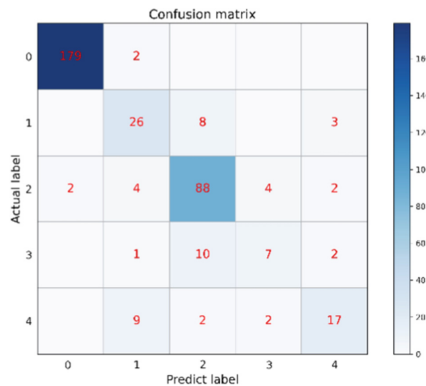


Fig. 7. Confusion Matrix for fusion model test results.

4.4 Comparison Results with Other Models

To further validate the validity of the model, the fusion model was compared with models proposed by other researchers on the dataset Aptos_sigmaX10 and the experimental results are shown in Table 4.

Table 4. Compared experimental results with other models model.

Reference	Year	Method	Accuracy (%)
Dekhil O et al. [6]	2019	Five-layer CNN	77%
Li et al. [21]	2019	CANet	83.20%
Yu et al. [22]	2021	Vision Transformer	85.50%
Islam M R et al. [10]	2022	Xception+CLAHE+t-SNE	84.36%
Ours	2023	EfficientNet+Swin Transformer	86.14%

As shown in Table 4, the proposed fusion model achieved 86.14% graded accuracy on this dataset, outperforming better than the results of the other four models. A 9.14% improvement in grading accuracy compared to the five-layer convolutional network model proposed by Dekhil et al. [6], Relative to the cross-disease attention network CANet proposed by Li et al. [21], this method uses the disease-specific attention module and the disease-dependent attention module to selectively learn useful features of the disease, achieving an accuracy of 83.20%, but still 2.94% lower than the proposed method, Comparing the two-stage comparison method with a supervised comparison loss function proposed by Islam et al. [10] also yields an accuracy gain of 1.78%. Furthermore, Yu et al. [22] attempted to use Vit for retinal disease classification tasks and achieve 85.50% accuracy performance by pre-training the Transformer model in a large fundus image database and then fine-tuning it in a downstream retinal disease classification task, but the results of this paper’s method still outperformed it by 0.64%. Taken together, the above comparative experimental results can effectively demonstrate the good competitive performance of the proposed fusion model.

5 Conclusion

Diabetic retinopathy is a common complication in diabetic patients. To address the shortcomings of convolutional neural networks for DR image grading, a DR grading method that fuses lightweight EfficientNet and Swin Transformer models is proposed to assist physicians in performing diagnosis, so as to better capture global and local features of DR images to improve the discriminative properties of the overall features. In future work, we will explore convolutional neural networks and Transformer architectures that are more suitable for this medical task and explore more effective fusion methods, such as feature interaction and combination in the middle of a dual-branch structure.

Acknowledgements. This work was supported in part by the Applied Basic Research Project of Liaoning Province under Grant 2023JH2/101300191, and the Young and Middle-aged Talents Program of the National Civil Affairs Commission.

References

1. Chen, C., Wu, S.Q., Wang, Y.F., et al.: Correlation between the duration of diabetes mellitus and diabetic retinopathy. *China Mod. Phys.* **58**(11), 77–80 (2020)
2. Yunlan, F.X.: Mechanism of NLRP3/IL-1 β pathway in proliferative diabetic retinopathy. *Int. J. Ophthalmol.* **19**(09), 1559–1562 (2019)
3. Selvachandran, G., Quek, S.G., Paramesran, R., et al.: Developments in the detection of diabetic retinopathy: a state-of-the-art review of computer-aided diagnosis and machine learning methods. *Artif. Intell. Rev.* **56**(2), 915–964 (2023)
4. Lachurej, J., Deorankar, A.V., Lachure, S., et al.: Diabetic retinopathy using morphological operations and machine learning. In: *IEEE International Advance Computing Conference (IACC)*, pp. 617–622 (2015)
5. Pratt, H., Coenen, F., Broadbent, D.M., et al.: Convolutional neural networks for diabetic retinopathy. *Procedia Comput. Sci.* **90**, 200–205 (2016)
6. Dekhil, O., Naglah, A., Shaban, M., et al.: Deep learning based method for computer aided diagnosis of diabetic retinopathy. In: *2019 IEEE International Conference on Imaging Systems and Techniques (IST)*, pp. 1–4. IEEE (2019)
7. de La Torre, J., Valls, A., Puig, D.: A deep learning interpretable classifier for diabetic retinopathy disease grading. *Neurocomputing* **396**, 465–476 (2020)
8. Zhou, Y., He, X., Huang, L., et al.: Collaborative learning of semi-supervised segmentation and classification for medical image. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2079–2088 (2019)
9. Gu, T.F., Hao, P.Y., Bai, C.W., et al.: Grading diabetic retinopathy in combination with multichannel attention. *Chin. J. Graph. Arts* **26**(07), 1726–1736 (2021)
10. Islam, M.R., Abdulrazak, L.F., Nahiduzzaman, M., et al.: Applying supervised contrastive learning for the detection of diabetic retinopathy and its severity levels from fundus images. *Comput. Biol. Med.* **146**, 105602 (2022)
11. Chen, H., Li, C., Wang, G., et al.: GasHis-transformer: a multi-scale visual transformer approach for gastric histopathological image detection. *Pattern Recogn.* **130**, 108827 (2022)
12. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: *Advances in Neural Information Processing Systems*, vol. 30 (2017)
13. Tsiknakis, N., Theodoropoulos, D., Manikis, G., et al.: Deep learning for diabetic retinopathy detection and classification based on fundus images: a review. *Comput. Biol. Med.* **135**, 104599 (2021)
14. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: *International Conference on Machine Learning*, pp. 6105–6114. PMLR (2019)
15. Liu, Z., Lin, Y., Cao, Y., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022 (2021)
16. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141 (2018)
17. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16 \times 16 words: transformers for image recognition at scale. *arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929)* (2020)

18. Liang, J., Cao, J., Sun, G., et al.: SwinIR: image restoration using swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1833–1844 (2021)
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint [arXiv:1711.05101](https://arxiv.org/abs/1711.05101) (2017)
20. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
21. Li, X., Hu, X., Yu, L., et al.: CANet: cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading. *IEEE Trans. Med. Imaging* **39**(5), 1483–1493 (2019)
22. Yu, S., et al.: MIL-VT: multiple instance learning enhanced vision transformer for fundus image classification. In: de Bruijne, M., et al. (eds.) MICCAI 2021, Part VIII. LNCS, vol. 12908, pp. 45–54. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87237-3_5