This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3260990

FIT-Net: Feature Interaction Transformer Network for Pathologic Myopia Diagnosis

Shaobin Chen, Zhenquan Wu, Mingzhu Li, Yun Zhu, Hai Xie, Peng Yang, Cheng Zhao, Yongtao Zhang, Shaochong Zhang, Xinyu Zhao,Tianfu Wang, Lin Lu*, Guoming Zhang*, Baiying Lei*,*Senior Member, IEEE*

***Abstract*— Automatic and accurate classification of retinal optical coherence tomography (OCT) images is essential to assist physicians in diagnosing and grading pathological changes in pathologic myopia (PM). Clinically, due to the obvious differences in the position, shape, and size of the lesion structure in different scanning directions, ophthalmologists usually need to combine the lesion structure in the OCT images in the horizontal and vertical scanning directions to diagnose the type of pathological changes in PM. To address these challenges, we propose a novel feature interaction Transformer network (FIT-Net) to diagnose PM using OCT images, which consists of two dual-scale Transformer (DST) blocks and an interactive attention (IA) unit. Specifically, FIT-Net divides image features of different scales into a series of feature block sequences. In order to enrich the feature representation, we propose an IA unit to realize the interactive learning of class token in feature sequences of different scales. The interaction between feature sequences of different scales can effectively integrate different scale image features, and hence FIT-Net can focus on meaningful lesion regions to improve the PM classification performance. Finally, by fusing the dual-view image features in the horizontal and vertical scanning directions, we propose six dual-view feature fusion methods for PM diagnosis. The extensive experimental results based on the clinically obtained datasets and three publicly available datasets demonstrate the effectiveness and superiority of the proposed method. Our code is avaiable at: https://github.com/chenshaobin/FITNet.**

***Index Terms*— Optical coherence tomography, pathologic myopia diagnosis, dual-scale Transformer, interactive attention, dual-view feature fusion.**

This work was supported partly by National Natural Science Foundation of China (Nos.U22A2024, 62106153, 62071309, 62101338, and 62271328), National Natural Science Foundation of Guangdong Province (No.2019A1515111205),Guangdong Basic and Applied Basic Research (Nos.2020A1515110605 and 2022A1515110704), Guangdong Provincial Key Laboratory (No.2020B121201001), Shenzhen Key Basic Research Project (Nos.JCYJ20220818095809021, KCXFZ20201221173213036, SGDX20201103095802007, JCYJ20190808155618806 and JCYJ20190808145011259), Shenzhen Science and Technology Program (No.JSGG20201102174200001), Shaobin Chen and Zhenquan Wu are contibuted equally, Corresponding authors(Lin Lu(e-mail:lulin888@126.com), Guoming Zhang(e-mail:13823509060@163.com), Baiying Lei (e-mail: leiby@szu.edu.cn)).

S. Chen, M. Li, Y. Zhu, H. Xie, P. Yang, C. Zhao, Y. Zhang, T. Wang and B. Lei are with School of Biomedical Engineering, Shenzhen University Medical School, National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Marshall Laboratory of Biomedical Engineering, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, Shenzhen University, Shenzhen,518060, China.

Z. Wu, X. Zhao and L. Lu are with State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat sen University, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Guangdong Provincial Clinical Research Center for Ocular Diseases, Guangzhou, 510060, China.

Z. Wu, S. Zhang, X. Zhao and G. Zhang are with Shenzhen Eye Hospital, Jinan University, Shenzhen Eye Institute, Shenzhen, 518040, China.

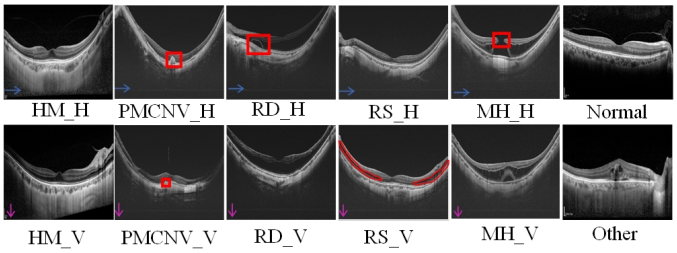


Fig. 1. OCT images contain four main types of pathological changes of PM (e.g.,RS, MH, RD, PMCNV), normal and other non-high myopia abnormalities. The underline H and V represent the horizontal and vertical scanning directions, respectively.

I. Introduction

MYOPIA is a quite common eye disease worldwide, high myopia (HM) indicates myopia without pathologic changes has continued to rise in many countries, especially in Asia. The myopia with pathologic changes (i.e., pathologic myopia (PM)) is the fourth most common cause of irreversible blindness in the world [1]. The myopia boom has affected populations around the world [2]. Myopia poses further vision challenges as HM increases the risk of pathologic changes such as cataracts, glaucoma, retinal detachment (RD), and myopic macular degeneration. All of these conditions can lead to irreversible vision loss [3]. The factors leading to pathological myopic retinopathy mainly include traction, atrophy, and neovascularization. The retinoschisis (RS), and macular hole (MH), RD are caused by traction, while macular atrophy is usually caused by pathologic myopic choroidal neovascularization (PMCNV). These are the most common macular degeneration and vision-threatening factors in PM, which need to be identified and treated in time [1], [4].

As a high-resolution retinal imaging method, optical coherence tomography (OCT) is a non-invasive imaging technique that visualizes retinal cross-sections. Hence, it is an essential tool for ophthalmologists to detect many types of myopic maculopathy. The examples of the pathological abnormalities that can be seen through OCT images are shown in Fig. 1.

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on May 18,2023 at 16:06:16 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3260990

Although OCT devices facilitate the diagnosis of ophthalmic diseases, reading OCT images often requires an experienced ophthalmologist. There are obvious differences in the position, shape, and size of the lesion structure from images in different scanning directions. Also, multiple pathological abnormalities may appear at the same time. Fig.1 shows that some diseases cannot be accurately diagnosed by a single direction scanning image. Ophthalmologists usually need to combine the images in the horizontal and vertical scanning directions to make accurate PM judgments. Therefore, it is desirable to develop a new automated diagnostic model to address these challenges.

In recent years, as a typical deep learning model, convolutional neural network (CNN) has been widely used in the automatic diagnostic analysis of retinal OCT images due to its powerful performance [5]. Artificial intelligence (AI) models have achieved good performance in diagnosing retinal diseases via OCT images such as diabetic retinopathy, macular edema, RD, and choroidal neovascularization (CNV) in age-related macular degeneration (AMD) [6]–[8]. Rong *et al.* [9] used denoised OCT images and masks extracted by thresholding and morphological dilation, which generate a large number of surrogate images for training CNN models for classification. Reza *et al.* [10] proposed a wavelet-based CNN applying wavelet transform to extract effective feature information in retinal OCT images. However, it is more complex to diagnose PM via OCT images since the lesion areas are too small to be diagnosed. Also, most research processes the entire image without considering the meaningful lesion area in the retinal layer [11].

Due to the local connection of CNN and the sliding window mechanism of the convolution kernel [12], CNN only focuses on the local features of the image and is difficult to enhance the remote dependence capability. Compared to CNN, Transformer [13] can capture distant features, which means that it can easily capture global information. As a new type of convolution-free neural network, Transformer uses the self-attention mechanism to extract intrinsic features, which has more application potential. For example, Vision Transformer (ViT) [14] applies a pure Transformer layer directly to image patches and then classifies the entire image. Transformer layers have shown promising accuracy and efficiency in classification and segmentation tasks [15]–[17].

To combine the advantages of CNN and Transformer, Hugo *et al.* [18] used convnet as a teacher to introduce a teacher-student strategy for Transformer. It relies on a distillation token that interacts with class token through a self-attention layer to ensure that the student learns from the teacher through attention. Furthermore, Chen *et al.* [19] studied the multi-scale feature representation of Transformer model and proposed a double-branch Transformer to combine two different size image patches to generate stronger image features. The proposed cross attention is used to achieve the feature fusion between two different branches. Inspired by the above works, we design a new method named feature interaction Transformer network (FIT-Net) to fuse both CNN and Transformer. This network embeds a dual-scale Transformer (DST) block in the CNN architecture. Specifically, we use ResNet50 [20] as the backbone convolutional network to obtain four multi-scale features of the image. They are paired to construct the remote dependence relationship through the Transformer encoder, and the interaction between the paired multi-scale features is realized through interactive attention (IA) units. Finally, the global class token (GCT) extracted from the image features of different scales is fused with the local class token (LCT) obtained from the convolutional network. Accordingly, the combination of the local features of CNN and the global features of traditional Transformer can achieve better classification performance.

Among the existing OCT image studies related to PM [21]–[24], they mainly focus on a single direction without combining the horizontal and vertical OCT images for analysis, which is particularly critical in clinical practice. By merging the horizontal and vertical directions to extract features, it can effectively reduce the misjudgment rate. In this study, we utilize our FIT-Net to achieve two mutually progressive classification tasks: preliminary screening of PM (Task 1) and a refined classification of PM for RS, MH, RD, and PMCNV (Task 2). Our contributions can be summarized in the following three aspects:

1) A novel FIT-Net is proposed to obtain the global and local class tokens of images from the features of different scales. To enhance the OCT image representation ability, we effectively fuse the GCT with the LCT.
2) To address the issue of weak ability of extracting local features for Transformer, we design a DST block and an IA unit. The class token of the image feature at different scales can exchange the learned abstract information of images, which enriches different feature representations.
3) Due to the obvious differences in the position, shape, and size of the lesion structure on the images of different scanning directions, we propose to use the feature fusion method to fuse the features of the dual-view OCT image, which can effectively reduce the misdiagnosis rate.

II. Related Work

A. OCT image classification tasks via CNN

In recent years, retinal OCT image classification tasks via CNN have been widely used in fundus diseases such as diabetes and AMD [25]–[30]. For example, Debjani *et al.* [24] fine-tuned a pre-trained GoogLeNet [31] for the task of classifying OCT images of diabetic macular edema (DME) and dry AMD. Lee *et al.* [8] used an independent VGG16 [32] network to achieve binary classification of OCT images of normal and AMD. Kermany *et al.* [33] fine-tuned the InceptionV3 [34] pre-trained on the ImageNet dataset to effectively solve the four-class OCT image problem for macular degeneration and diabetic retinopathy. Xu *et al.* [35] created an innovative fusion of color fundus images and OCT images to achieve the four-classification task of AMD and polypoid choroidal vascular disease through a two-channel deep CNN. Reza *et al.* [36] proposed a multi-scale convolutional mixture of expert ensemble model to identify normal retina, and two common types of macular pathologies. The above works show that traditional CNN methods have been widely used in OCT

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on May 18,2023 at 16:06:16 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3260990

image classification tasks, but the classification performance still needs further improvement.

In [22], the authors proposed a novel lesion-aware CNN for retinal OCT image classification. It designed a lesion detection network to generate soft attention maps from the whole OCT image, and the attention map is incorporated into the classification network. Accordingly, utilizing the retinopathy regions in the OCT images can guide the CNN to achieve more accurate classification results. Similarly, Huang *et al.* [23] designed a layer-guided CNN to identify normal retina and three common types of macular degeneration. However, they had to train an additional segmentation network to extract lesions or layer regions to assist classification, which also requires clinicians to spend a lot of time on labeling and is clinically inefficient [37].

CNN has also made a series of advances in OCT images related to PM [38]. The network based on the ResNet architecture has a better feature extraction ability for OCT images. For example, Lin's team developed an artificial intelligence system via deep learning from OCT macular images to identify 4 pathological changes in patients with PM [21]. This study used InceptionResNetV2 [39] to construct four independent binary classifiers to identify RS, MH, RD, and PMCNV. He *et al.* [40] proposed a novel modality-specific attention network for multi-modal retinal image classification, which effectively combined modality-specific diagnostic features of fundus and OCT images to achieve the nine fundus diseases classification task. In these studies, image features of different scanning directions were not comprehensively considered in the refined classification of pathological changes in PM, which is particularly critical in clinical practice.

B. ViT related work

Transformer is a network architecture based entirely on attention mechanisms, which eliminates convolutional operations and has achieved significant success in machine translation [13]. In recent years, inspired by Transformers, convolution-free models that rely only on Transformer layers have become popular in computer vision. ViT [14] is the first to use Transformer layer for image classification tasks, which is comparable to or even surpass that of CNN. Recently, ViT has been extended from image classification tasks in computer vision to image segmentation, image generation, and object detection tasks [16], [41]–[44]. Many variants based on ViT have also been proposed recently. For example, Hugo *et al.* [18] introduced a teacher-student strategy specific to Transformers in the proposed data-efficient image Transformer. The strategy relies on a distillation token to ensure that the Transformer student network can learn from a CNN teacher network through attention. However, since Transformers lack some inductive biases inherent in CNNs—such as translation equivalence and locality, traditional Transformers cannot generalize well on small datasets [45]. Also, Transformer lacks the ability to extract local information, and there is still a gap in performance with existing CNNs. Therefore, combining Transformers with convolutions may be a more straightforward way to introduce locality into traditional Transformers [46]–[48].

TABLE I
PARAMETER INFORMATION IN FIT-NET

Scale	Feature size	Conv1×1	Patch size	Patch dim	MLP dim
Scale 1	112×112×64	112×112×4	16	1024	2048
Scale 2	56×56×256	56×56×8	8	512	1024
Scale 3	28×28×512	28×28×16	4	256	512
Scale 4	14×14×1024	14×14×32	2	128	256

III. PROPOSED METHOD

A. Dual-scale Transformer Block

Fig. 2 shows our proposed FIT-Net, which mainly consists of three collaborative parts, i.e., the backbone, DST block, and IA unit. This method is mainly based on ResNet50 and ViT. In order to fully mine the effective information representation in multi-scale features to achieve better classification results, we embed a DST block in the backbone convolutional network. As shown in Fig. 2, four sets of multi-scale features extracted by the backbone convolutional network are each passed through a 1×1 convolution kernel to reduce the dimension of the features, and then used in pairs as the input of the DST block. We first perform feature linear projection on the input features of different scales, which is similar to ViT except that the input image features are divided into a series of patch token at the specified patch size. Meanwhile, additional class token is also concatenated to the patch representation to form a complete feature sequence. Finally, position embedding information is also added to the feature sequence. Since the class token in the feature sequence after the first Transformer encoder has learned the abstract representation information in the patch feature, we use the class token in one scale feature as the agent for information exchange with the patch feature sequences of another scale feature. To this end, we use the IA unit to achieve this purpose. The dimension sizes of multi-scale features used in proposed FIT-Net in this study as well as the dimension sizes of input feature dimensions, patch size, dimension of vector sequence, and dimension sizes of hidden layers in multi-layer perceptron (MLP) in the two DST blocks are shown in Table I.

The Transformer encoder consists of a series of block structures, where each block contains a multi-head attention (MHA) module and a feed-forward network (FFN). In the self-attention layer proposed in [13], the query (Q), key (K), and value (V) matrices are calculated according to the sequence of input vector($X \in \mathbb{R}^{N \times D}$):

$$Q = XW_Q, K = XW_K, V = XW_V \tag{1}$$

where W_Q, W_K, W_V are learnable parameters obtained by linear transformation. We use an inner product operation to match each set of query vectors $q \in \mathbb{R}^D$ with N key vectors $k \in \mathbb{R}^D$(i.e., matrix K). A Softmax function is used to scale and normalize these inner products to obtain N weights. Finally, the output of attention is a weighted sum of a set of vectors $v \in \mathbb{R}^D$ (i.e., matrix V). This implementation is defined as follows:

$$\text{Attention } (Q, K, V) = \text{Softmax}\left(QK^T/\sqrt{D}\right)V \tag{2}$$

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on May 18,2023 at 16:06:16 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3260990

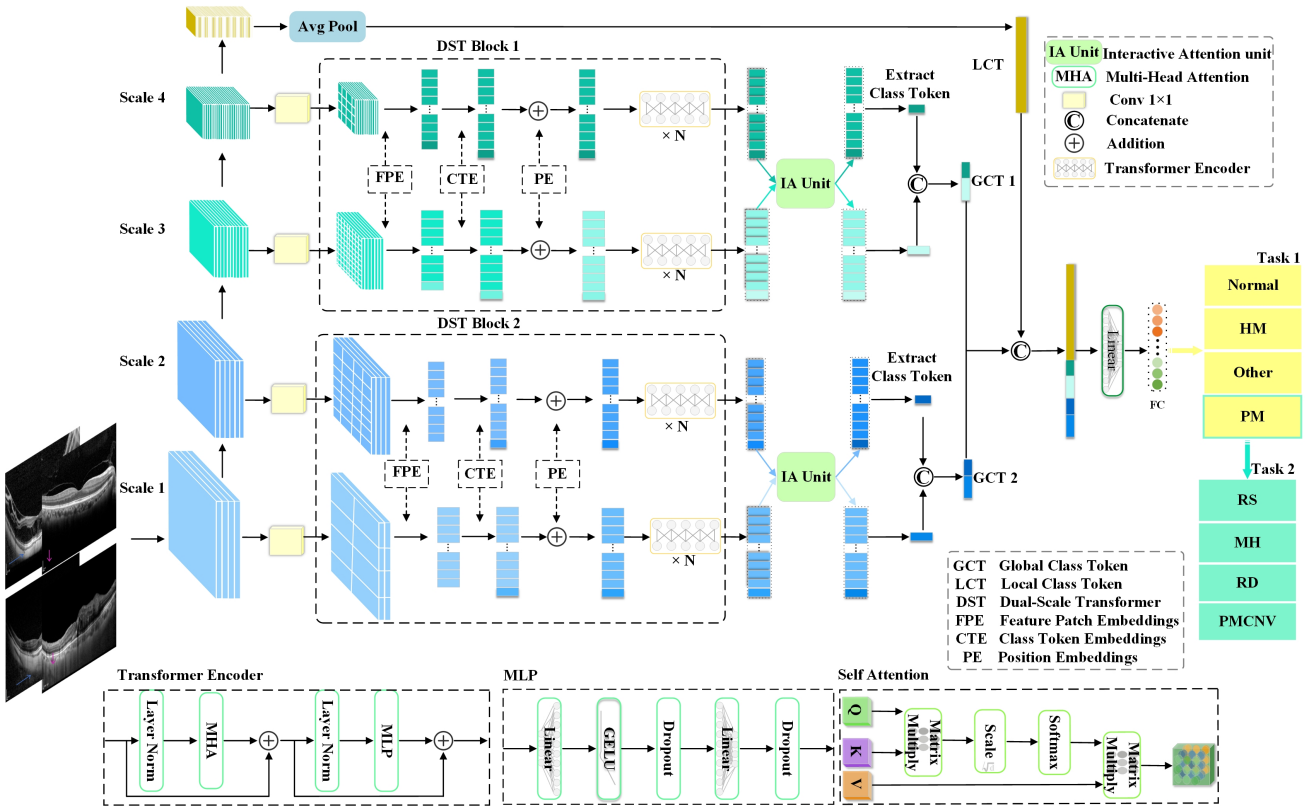


Fig. 2. The framework of our proposed FIT-Net, which mainly consists of three collaborative parts, i.e., the backbone, DST block, and IA unit. The network is mainly used for two progressive tasks.The dimension information of the four groups of multi-scale features Scale1, Scale2, Scale3 and Scale4 is shown in Table I.The LCT is generated by the backbone, while the GCT is generated by DST block.

The MHA divides the input Q, K, and V into h heads. After this processing, the vector dimensions of Q, K and V in each head will become $1/h$ of the original dimension, that is D/h. The MHA mechanism combines information learned from different heads. Each head is processed by a self-attention layer, and finally their outputs are concatenated together. The concatenated sequence is re-projected into an $N\times D$ vector by a linear layer so that the final output of MHA is consistent with the dimension of the input vector sequence. FFN is composed of two linear layers separated by a GELU [49] activation function. The first linear layer expands the dimension of the input, and the second linear layer reduces the dimension to the original input dimension. Layer normalization operations are applied before FFN and MHA, while skip-connections operations are applied after each FFN and MHA.

B. Interactive Attention Unit

In our proposed IA unit, we exchange the class token in different scale feature sequences and then pass the fused sequences through the next Transformer encoder. The abstract information learned by feature sequences of different scales can be exchanged with each other and enriches the representation of different patches. The interaction between the class token and patch representation of feature sequences at different scales can more effectively fuse features at different scales. The following sub-section details a high-dimensional input feature as an example to introduce the IA unit. The implementation is shown in Fig. 3.

For the high-dimensional input feature sequence, it is first necessary to fuse its class token with the patch features in the low-dimensional feature sequence to form a new feature sequence. The corresponding operation is also performed on the low-dimensional input feature sequence, and the implementation is defined as follows:

$$\begin{aligned} f_{cls}^{\prime l} &= F\left(f_{cls}^{h}\right), f_{cls}^{\prime h} = F\left(f_{cls}^{l}\right) \\ f^{\prime l} &= \left[\ f_{cls}^{\prime l} \| f_{\text{patch}}^{l}\ \right], f^{\prime h} = \left[\ f_{cls}^{\prime h} \| f_{patch}^{h}\ \right], \end{aligned} \tag{3}$$

where $F(.)$ is the linear projection function of dimension alignment, f_{cls}^{h} and f_{cls}^{l} are the class token of the high-dimensional and low-dimensional input feature sequences, respectively, f_{patch}^{l} and f_{patch}^{h} are the patch feature sequences in the low-dimensional and high-dimensional feature sequences, respectively, and $f^{\prime h}$ and $f^{\prime l}$ both represent the new fusion feature sequence. $\|$ represents the operation of splicing the patch feature sequence and class token into a complete feature sequence. Since the dimensions of the high-dimensional and low-dimensional input feature sequences are different, it is necessary to unify the dimensions of the class tokens of different scales through $F(.)$ before forming a new feature sequence. After that, our proposed IA unit will act on $f^{\prime h}$ and $f^{\prime l}$. The converted class token $f_{cls}^{\prime h}$ and $f_{cls}^{\prime l}$ will serve as unique query for the fusion feature sequences $f^{\prime h}$ and $f^{\prime l}$, respectively. We also apply MHA in the proposed IA unit, which is defined as multi-head interactive attention (MIA). In MIA, we do not use the FFN in the conventional Transformer

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on May 18,2023 at 16:06:16 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3260990

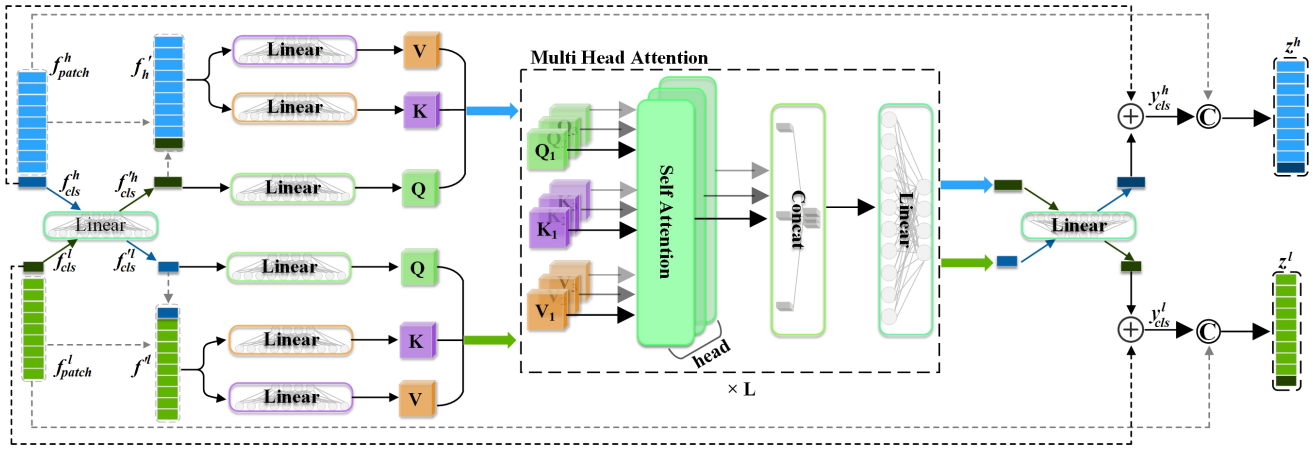


Fig. 3. The proposed IA unit. The class token of the high-dimensional input feature sequence is used as a query token, and information interaction is performed with the patch feature sequence of the low-dimensional input feature sequence through IA. The low-dimensional input feature sequence follows the same operation, also exchanges information with high-dimensional patch representations by exchanging class token.

encoder. The specific implementation is defined as follows:

$$\begin{aligned} Q^h &= f_{cls}^{\prime h} W_Q, \quad K^h = f^{\prime h} W_K, V^h = f^{\prime h} W_V \\ MIA\left(f^{\prime h}\right) &= \text{softmax}\left(Q^h {K^h}^T / \sqrt{D/H}\right) V^h \end{aligned} \tag{4}$$

$$\begin{aligned} Q^l &= f_{cls}^{\prime l} W_Q, \quad K^l = f^{\prime l} W_K, V^l = f^{\prime l} W_V \\ MIA\left(f^{\prime l}\right) &= \text{softmax}\left(Q^l {K^l}^T / \sqrt{D/H}\right) V^l, \end{aligned} \tag{5}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{N \times (D/H)}$ are parameters that can be learned in the network, and D and H are the embedded feature dimension and the number of heads in the Transformer encoder, respectively. The final output of the IA unit is given as:

$$y_{cls}^h = f_{cls}^h + G\left(MIA\left(LN\left(f^{\prime l}\right)\right)\right), z^h = \left[\ y_{cls}^h \quad \| f_{patch}^h\ \right], \tag{6}$$

$$y_{cls}^l = f_{cls}^l + G\left(MIA\left(LN\left(f^{\prime h}\right)\right)\right), z^l = \left[\ y_{cls}^l \quad \| f_{patch}^l\ \right], \tag{7}$$

where *G(.)* is the dimension-aligned linear projection function that matches *F(.)*, and its purpose is to keep the output dimension of MIA consistent with the dimension of the class token of the original input feature; y_{cls}^h and y_{cls}^l represent the class token processed by the IA unit; z^h and z^l represent the processed fused feature sequence; *LN* represents layer normalization operation.

C. Dual-view Feature Fusion

In order to effectively fuse the dual-view image features of the horizontal and vertical scanning directions to realize the refined classification task of PM in Task 2, this study continues to explore the fusion method of dual-view features based on the proposed FIT-Net. Dual-view OCT images are input into FIT-Net to obtain three class tokens, including one LCT from the backbone and two GCTs output by IA unit. This study explores six feature fusion methods for these three class tokens, and the following will be introduced in detail. **Concatenation Fusion (CF)**: The six class tokens obtained from dual-view OCT images are directly concatenated to obtain the fused class token. **Addition Fusion (AF)**: After the three class tokens of the single view are concatenated, the concatenated class tokens are added bit by bit. **Concatenation with MLP Fusion (CMF)**: The six class tokens obtained from the dual-view OCT images are concatenated and then processed by a MLP to obtain the fused class tokens. **Addition with MLP Fusion (AMF)**: On the basis of AF, a MLP is used to obtain the fused class token. **Deep Concatenation Fusion (DCF)**: After the three class tokens of dual-view are concatenated one by one, each is processed by a MLP, and then the processed class tokens are concatenated again to generate the fused class token. **Deep Addition Fusion (DAF)**: After the three class tokens of each view are added bit by bit in a one-to-one correspondence, each is processed by a MLP, and then the processed class tokens are added bit by bit again to generate the fused class token. The specific implementation is shown in Fig.4.

IV. Experiments and Results

A. Dataset

In this representative prospective study, a total of 9984 qualified OCT images were collected from Zhongshan Ophthalmic Center of Sun Yat-sen University from 2011 to 2018. The images can be grouped into four categories: Normal, high myopia without pathological changes (HM), PM, and other eye diseases (Other), with numbers of 4525, 2282, 2760 and 417, respectively. In the experiment, Task 1 and Task 2 will be implemented on this dataset. The dataset used in Task 2 has a total of 5042 OCT data for the two categories of HM and PM. The PM category is divided into four categories: RS, MH, RD, and PMCNV, with numbers of 2246, 728, 810 and 502, respectively. In this dataset, each eye is matched with a horizontal and vertical scanning image. The distribution of the types of pathological changes in PM is shown in Fig.5. As seen from the data distribution in Fig.5, some PM patients have multiple factors leading to pathological myopic retinopathy. Most PM patients with MH and RD will be accompanied by symptoms of RS at the same time. The proportion of data

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on May 18,2023 at 16:06:16 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3260990

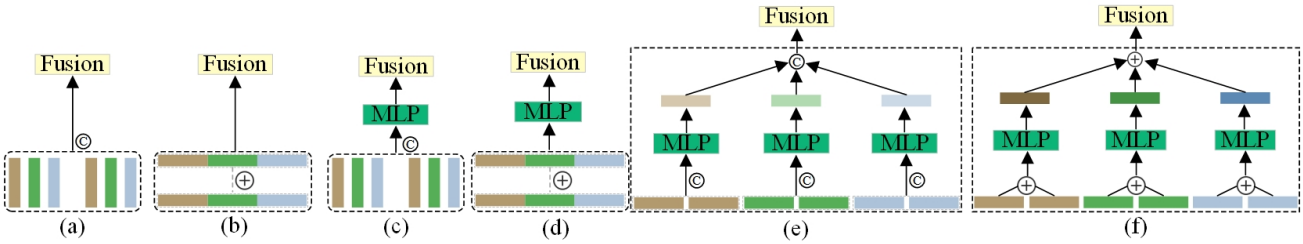


Fig. 4. Dual-view feature fusion method. (a)-(f) represent CF, AF, CMF, AMF, DCF and DAF, respectively.

with three or more pathological factors at the same time is relatively small.

Furthermore, the FIT-Net is further evaluated on three different datasets acquired by the Heidelberg SD-OCT imaging system. The first one is obtained at Noor Eye Hospital in Tehran, which include OCT image data of 50 normal, 48 dry AMD, and 50 DME patients with 1585, 1639, and 1105 image numbers for the three categories, respectively [36].The dataset is available at Dataset A. Dataset B is acquired from a Custom-made swept-source OCT (SS-OCT) imaging system designed and built in Dep. of Biomedical Engineering, University of Basel [50]. This dataset includes 124 volumes obtained after image quality assessment with automatic cropping. Each subject contains 30 to 300 B-scan with 300 x 300 pixels resolution. The participants included people with DME (40 subjects), Healthy (50 subjects), and Non-diabetic patients (34 subjects) with AMD, CNV, or Macular Hole (MH) disorders, the number of images is 7722, 10301 and 5826, respectively. The dataset is available at Dataset B. Dataset C is a publicly available dataset containing 108309 OCT B-scans (Spectralis OCT, Heidelberg Engineering, Germany) images collected from 4,686 patients at various medical institutions, including the Shiley Eye Institute of the University of California San Diego (UCSD). This dataset can be subdivided into four categories including CNV, DME, drusen, and normal, with numbers of 37205, 11348, 8616, and 51140, respectively. In addition, the dataset provided 1000 images (633 patients, 250 images for each category) independent of the training set for validation of model performance. For more details on the Dataset C, please refer to [33].

B. Experimental Settings

The proposed model is implemented on the PyTorch platform and trained on an NVIDIA TITAN RTX GPU (24GB) using a 10-fold cross-validation strategy. The method proposed in this study and other comparison methods use stochastic gradient descent optimizer with a momentum factor of 0.9 to optimize model parameters in the training process. The number of categories in Task 1 and Task 2 is unbalanced. In order to reduce the imbalance problem of the dataset, we train the proposed FIT-Net with a focal loss [51] function. Lin *et al.* [51] proposed a focal loss as an extended version of the cross-entropy loss. By down-weighting easy-to-classify samples, the focal loss will focus on hard-to-classify samples. The focal loss function is defined as:

$$L_{\text{Focal }(p,y)} = -y(1-p)^{\gamma}\log(p) - (1-y)p^{\gamma}\log(1-p), \quad (8)$$

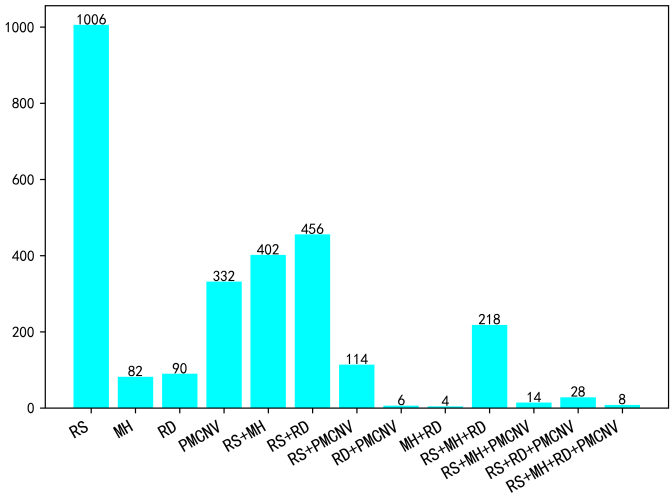


Fig. 5. The distribution statistics of datasets in Task 2

where y and p denote the ground truth label and predictive probability, respectively. The parameter γ controls the weight of easy samples. Note that when γ = 0, the focal loss becomes the cross-entropy loss. As the value of γ increases, the loss function becomes smaller for higher predictive possibility p. In the experiment, we set the value of γ as 2. The initial learning rate for training is set to 0.001, and a cosine learning rate decay [52] strategy is used to update the learning rate in this study. We also set the batch size to 128. Task 1 is a multi-class classification task and this study evaluates the classification performance for each category independently. Specifically, for class i, $h = TP_i, FP_i, TN_i, FN_i$ represents the number of true positives, false positives, true negatives, and false negatives, respectively. In this study, seven evaluation metrics are calculated for each category: accuracy (ACC), area under the receiver operating characteristic (ROC) curve (AUC), precision (Pre), sensitivity (Sen/Recall), specificity (Spec), F1 score (F1), Kappa:

$$\text{Pre}_i = \frac{TP_i}{TP_i + FP_i}, \text{Sen}_i = \frac{TP_i}{TP_i + FN_i} \quad (9)$$

$$\text{F1}_i = \frac{2 \times Pre_i \times Recall_i}{Sen_i + Recall_i}, \text{Spec}_i = \frac{TN_i}{TN_i + FP_i} \quad (10)$$

Furthermore, for an overall comparison of all categories, we calculate a weighted average of the evaluation metrics. The weighted evaluation metric $M \in Pre, Sen, F1, Spec, Kappa, ACC, AUC$ in the multi-class

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on May 18,2023 at 16:06:16 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3260990

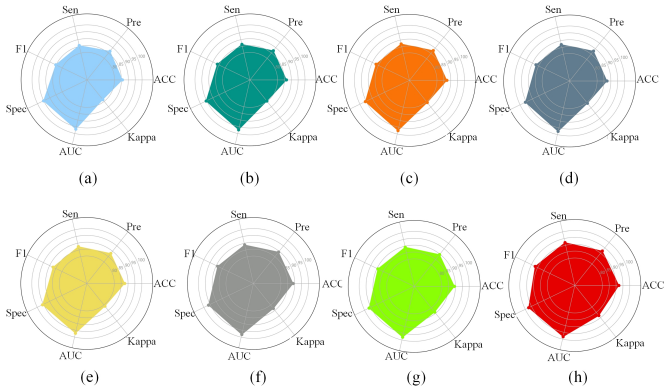


Fig. 6. Radar charts of various evaluation metrics for different methods. (a)-(h) represent the results of ViT, ResNet50, CrossViT-Ti, DeiT, E-Net, Swin-T, ConvNeXt-T and proposed FIT-Net, respectively.

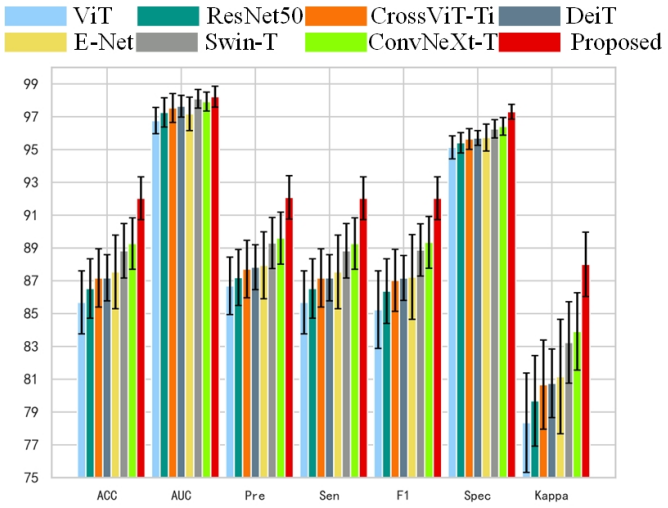


Fig. 7. Histogram of the performance of different models under various evaluation metrics.

classification task is defined as:

$$M_{\text{weighted}} = \sum_{i=1}^{4} w_i \times M_i \tag{11}$$

where w_i represents the proportion of the number of samples of category i in the total number of samples.

The clinical dataset used for Task 2 contains four lesion categories in a multi-label manner, and each sample can be assigned multiple lesion categories simultaneously. We treat this multi-label classification problem as four binary classification problems, and each consists of one lesion category of PM and HM. In each binary classification task, the data of different scanning directions are selected in the unit of the patient's eye for experimental exploration. Similarly, this study calculates seven evaluation metrics for each binary classification task in Task 2.

C. Task 1: the preliminary screening of PM

In order to verify the effectiveness of the proposed method in this study, we select seven networks for comparison, including 3 CNN networks: ResNet50 [20], EfficientNet-B0 (E-Net) [53] and ConvNeXt-T [54], and 4 Transformer-related models: ViT [14], CrossViT-Ti [19], DeiT [18] and Swin-T [16]. Among them, ViT uses the ViT-Base/16 architecture, DeiT uses the ViT-Base/16 architecture as the student model, and uses ResNet50 as the teacher model. The experimental results are summarized in Table II. To more intuitively compare the performance of each model under different metrics and the comprehensive classification performance of each model, the results of various evaluation metrics of different methods are drawn into radar charts and histogram in this study, as shown in Fig.6 and Fig.7. It is observed from the radar charts and histogram that our method achieves the best performance in multiple evaluation metrics. Compared with the current state-of-the-art model, the accuracy and sensitivity have increased by 2.76%, the accuracy has increased by 2.48%, and the Kappa has increased by 4.08%. Other metrics have also improved. This proves that the FIT-Net designed in this study can facilitate the network to extract discriminative features, which can obtain the best classification performance in Task 1.

To further explore the performance of different methods in four categories in Task1, ROC curves for four categories are drawn: Normal, HM, PM, and Other, respectively, as shown in Fig.10. The red line represents the proposed FIT-Net in this study, and it can be seen that FIT-Net has better overall performance to discriminate the four categories. As can be seen from Fig.10(a), the results of AUC of seven comparison methods and FIT-Net are relatively close to the Normal category, which indicates that most CNN models can extract the characteristics of Normal. Although the differences of AUC among methods are not particularly obvious, the ROC curve of the proposed FIT-Net is closer to the upper left corner of the coordinate axis, which indicates that the performance of this method is relatively better than that of other methods. In task 1, compared with other comparison methods, FIT-Net has significantly improved the evaluation indicators such as ACC, Pre, Sen, F1, Spec and Kappa, while the improvement in the evaluation indicator of AUC is only within 1%. We further compared FIT-Net's ROC curve with those of seven comparison methods in pairs to explore the implicit correlation between the curves. As a completely nonparametric method, the Delong test algorithm is widely used to compare areas under two or more ROC curves [55], [56]. We also used the Delong test algorithm to compare the ROC curve of FIT-Net under each category with the seven comparison methods one by one, and conducted statistical significance analysis on the results to obtain p values, as shown in Table III. The null hypothesis given here is: the comparison between the ROC curve of FIT-Net under each category and the seven comparison methods is not statistically significant. Generally, the threshold of significance level is set as 0.05, and the null hypothesis can be rejected with sufficient reasons only when p value is less than or equal to 0.05. From the results in Table III, there are sufficient reasons to reject some of the null hypotheses, although there are some p values greater than 0.05. Therefore, the comparison between the ROC curve of FIT-Net under each category and the seven comparison methods has a certain statistical significance, and from the visualization

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on May 18,2023 at 16:06:16 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3260990

TABLE II
ALGORITHM COMPARISON OF TASK 1 (%)

Method	ACC	Pre	Sen	F1	Spec	AUC	Kappa
ViT	85.68±1.92	86.69±1.75	85.68±1.92	85.24±2.36	95.13±0.70	96.76±0.80	78.35±3.03
ResNet50	86.53±1.81	87.20±1.71	86.53±1.81	86.37±1.97	95.41±0.62	97.26±0.89	79.68±2.76
CrossViT-Ti	87.17±1.78	87.71±1.75	87.17±1.78	87.03±1.89	95.64±0.63	97.53±0.88	80.67±2.72
DeiT	87.18±1.41	87.83±1.37	87.18±1.41	87.17±1.37	95.69±0.45	97.64±0.66	80.75±2.09
E-Net	87.54±2.24	87.94±2.04	87.54±2.24	87.23±2.58	95.73±0.82	97.17±1.02	81.16±3.49
Swin-T	88.83±1.66	89.30±1.56	88.83±1.66	88.88±1.59	96.26±0.56	98.09±0.57	83.24±2.48
ConvNeXt-T	89.27±1.57	89.60±1.59	89.27±1.57	89.34±1.58	96.41±0.54	97.92±0.58	83.91±2.36
Ours	**92.03±1.30**	**92.08±1.32**	**92.03±1.30**	**92.03±1.30**	**97.30±0.45**	**98.22±0.64**	**87.99±1.96**

TABLE III
TASK 1:DELONG TEST ANALYSIS OF DIFFERENT ROC CURVES

Category	Method	ViT	ResNet50	CrossViT-Ti	DeiT	E-Net	Swin-T	ConvNext-T
Normal	p-value	0.237	0.378	0.890	0.841	0.323	<0.001	0.438
HM	p-value	<0.001	<0.001	0.595	0.441	<0.05	0.335	0.218
PM	p-value	<0.001	<0.001	0.766	<0.05	0.063	<0.001	<0.05
Other	p-value	<0.001	0.440	0.444	0.448	0.130	<0.05	0.347

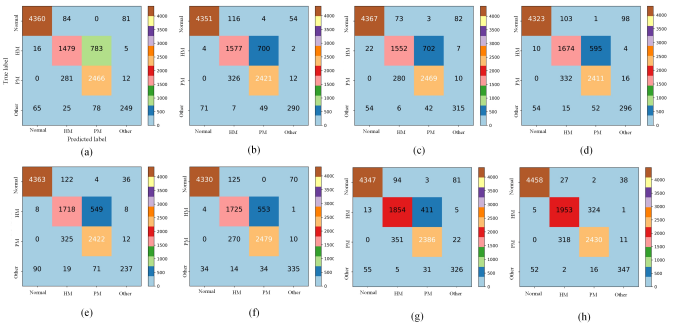


Fig. 8. Confusion matrix plots the related network with our proposed method on Task 1. The numbers in the confusion matrix diagram represent the number of categories. (a)-(h) represent the results of ViT, ResNet50, CrossViT-Ti, DeiT, E-Net, Swin-T, ConvNeXt-T and proposed FIT-Net, respectively.

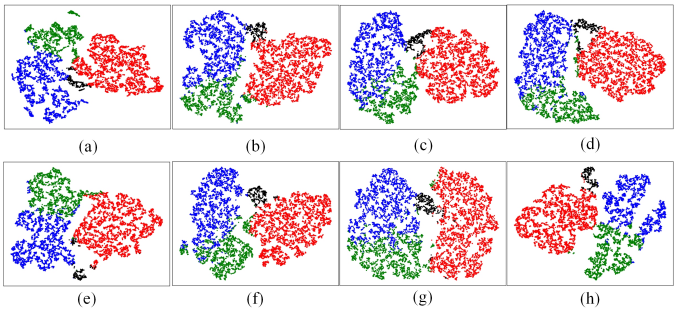


Fig. 9. t-SNE visualization results of different networks in Task 1. (a)-(h) represent the results of ViT, ResNet50, CrossViT-Ti, DeiT, E-Net, Swin-T, ConvNeXt-T and the proposed FIT-Net, respectively.

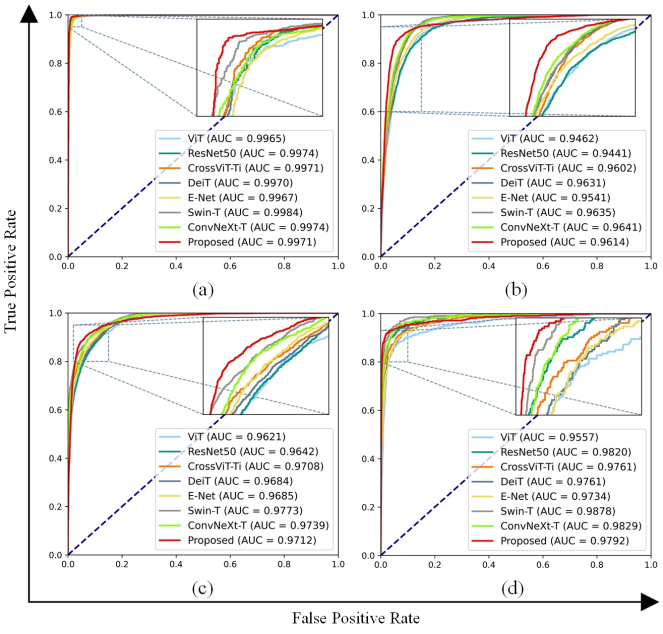


Fig. 10. ROC curves of the proposed method and the comparative method under different categories. (a)-(d) represent the results of four categories: Normal, HM, PM, and Other.

results of the classification indicators in Fig.7, FIT-Net has better classification performance.

The confusion matrix diagram in Fig.8 shows the classification performance of the comparison network and our proposed method to separate the four categories. We can see more intuitively that HM and PM are more likely to be confused with each other, and the performance of the seven comparison methods is poor, and the proposed FIT-Net has better classification performance in these two categories. Furthermore, we visualize the features learned by the last feature extraction layer of different networks and implement t-SNE [57] visualization, as shown in Fig.9. The effectiveness of the proposed FIT-Net can be further demonstrated from the t-SNE visualization results.

D. Ablation Study

As mentioned above, our proposed FIT-Net consists of DST block and IA unit. We conduct ablation experiments on these modules to verify that the designed modules are effective in terms of improving the classification performance. The results of the ablation experiments are shown in TableIV.It can be seen that when only DST block 1(DST 1) or DST block 2(DST 2) is used, the accuracy of the network is similar to that of

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on May 18,2023 at 16:06:16 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3260990

TABLE IV
ABLATION EXPERIMENT RESULTS IN TASK 1 (%)

LCT	DST1	DST2	IA	ACC	Pre	Sen	F1	Spec	AUC	Kappa
✓				86.53±1.81	87.20±1.71	86.53±1.81	86.37±1.97	95.41±0.62	97.26±0.89	79.68±2.76
	✓			86.46±0.36	87.57±0.74	86.46±0.36	86.17±0.59	95.39±0.20	97.25±0.64	79.56±0.63
		✓		87.20±1.02	87.86±1.10	87.20±1.02	87.00±1.12	95.64±0.37	97.54±0.56	80.67±1.56
✓	✓			88.56±0.46	89.10±0.47	88.56±0.46	88.47±0.51	96.12±0.18	97.69±0.62	82.76±0.72
✓		✓		88.59±0.40	89.16±0.54	88.59±0.40	88.49±0.45	96.15±0.16	97.80±0.61	82.83±0.64
✓	✓		✓	89.82±0.89	90.19±1.06	89.82±0.89	89.81±0.91	96.55±0.31	98.04±0.60	84.69±1.35
✓		✓	✓	89.93±0.85	90.26±0.94	89.93±0.85	89.94±0.79	96.63±0.26	98.02±0.64	84.87±1.24
✓	✓	✓		91.05±0.84	91.17±0.88	91.05±0.84	91.04±0.87	96.96±0.31	98.08±0.74	86.51±1.28
✓	✓	✓	✓	**92.03±1.30**	**92.08±1.32**	**92.03±1.30**	**92.03±1.30**	**97.30±0.45**	**98.22±0.64**	**87.99±1.96**

backbone (LCT). Without IA unit, adding one of the DST block (LCT + DST 1, LCT + DST 2) on the basis of backbone increases the accuracy by 2%. On the basis of backbone, either DST block 1 embedded with IA unit (LCT + DST 1+IA) or DST block 2 embedded with IA unit(LCT+ DST 2+IA), network accuracy is at least 3.2% higher than baseline. For the two DST blocks (LCT+ DST 1+ DST 2) without embedded IA unit, the network accuracy is improved by 4.5% compared with backbone. The accuracy of our proposed FIT-Net (LCT+ DST 1+ DST 2+IA) is 5.5% higher than that of backbone. Therefore, the proposed DST block and IA unit have good feature extraction ability, which can effectively improve the accuracy of Task1. Eq. (8) is our used focal loss function. Different parameters γ correspond to different loss functions. We set the value of parameter γ to 0, 0.5, 1, 2, and 3, respectively, to explore the performance of our method in Task 1. The experimental results are shown in Table V.It can be seen that the classification performance is the worst when γ is set to 0, that is, when the original cross-entropy loss function is used. With the gradual increase of γ, the classification performance gradually increases, but when γ is 3, the classification performance is lower than that when γ is 2. Therefore, we set gamma parameter as 2 in the experiment.

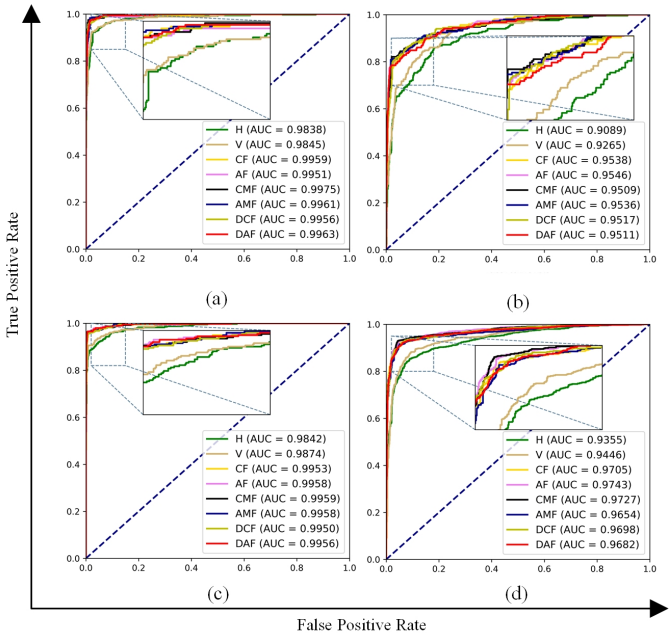


Fig. 11. ROC curve comparison between fusion method and single-view method only. (a)-(d) represent four lesion categories such as MH, PMCNV, RD, and RS, respectively.

E. Task 2:Pathological change type fine classification task

In Task 2, we separately analyze the classification performance of the four most common visual threat factors in PM, including RS, MH, RD, and PMCNV, in OCT images with two different scanning directions: Horizontal(H), Vertical (V). In order to further explore whether the classification performance of feature fusion of dual-view OCT images is better than that of using only single-view OCT images, this study designs six different feature fusion methods, as discussed in Section III.D. The experimental results are summarized in TableVI. In order to more intuitively compare the performance of the six fusion methods proposed in this study and the method using only a single view on different classification metrics, the experimental results are plotted as a histogram shown in Fig.12. It can be seen that among the four classifiers for the lesion categories, the feature fusion of dual-view OCT images in the horizontal and vertical scanning directions improves all classification metrics as compared with OCT images with a single view. It can also be seen from the ROC curve in Fig.11 compared with the method using only single-view OCT images that the six methods designed in this study to fuse the features of dual-view OCT images have better performance. In order to further comprehensively compare the performance of the six fusion methods, the average of each classification metric of the four classifiers is calculated under each fusion method, and the results are presented in TableVII. It can be seen from the results in Tables VI and VII that PMCNV achieves the best performance in the fusion method using AMF, while datasets MH, RD and RS all achieve the best performance in the fusion method using CMF. To validate the ability of the designed module to extract features, we also conduct experiments by class activation mapping using the output features of the proposed FIT-Net to verify that the extracted features can effectively distinguish the predicted classes. The Grad-CAM [58] visualization is shown in Fig.13. It can be seen that our proposed method is effective to learn the features of salient regions in dual view images of four different types of pathological changes, and pay attention to the features of specific diseases.

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on May 18,2023 at 16:06:16 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3260990

TABLE V
ABLATION ANALYSIS OF DIFFERENT LOSS FUNCTIONS IN TASK 1(%)

γ	ACC	Pre	Sen	F1	Spec	AUC	Kappa
0	90.41±1.60	90.63±1.46	90.41±1.60	90.39±1.61	96.75±0.54	98.12±0.61	85.55±2.40
0.5	90.73±1.47	90.81±1.51	90.73±1.47	90.70±1.50	96.84±0.52	98.15±0.70	86.01±2.23
1	90.89±1.29	91.01±1.38	90.89±1.29	90.88±1.30	96.91±0.45	98.03±0.86	86.28±1.95
2	**92.03±1.30**	**92.08±1.32**	**92.03±1.30**	**92.03±1.30**	**97.30±0.45**	**98.22±0.64**	**88.00±1.96**
3	91.27±1.28	91.48±1.16	91.27±1.28	91.29±1.29	97.06±0.44	98.17±0.50	86.87±1.92

TABLE VI
TASK 2: PERFORMANCE COMPARISON OF UNIDIRECTIONAL AND DUAL-VIEW FEATURE FUSION METHODS (%)

DataSet	Method	ACC	AUC	Pre	Sen	F1	Spec	Kappa
PMCNV	H	90.59±2.15	91.51±2.95	80.12±8.76	64.95±7.67	71.29±6.18	96.23±2.11	65.77±7.37
	V	91.31±1.91	92.75±3.23	81.52±9.16	68.51±9.25	73.80±6.16	96.32±2.02	68.67±7.15
	CF	94.33±2.04	95.51±2.42	91.15±6.27	76.09±9.30	82.65±6.59	98.34±1.14	79.31±7.76
	AF	94.97±1.93	95.76±2.98	94.08±5.58	77.25±10.18	84.40±6.58	98.86±1.11	81.47±7.64
	CMF	94.97±1.79	95.74±2.16	89.61±4.29	81.65±9.01	85.19±5.67	97.90±0.89	82.18±6.71
	AMF	95.12±1.85	95.64±2.47	91.20±5.73	80.91±6.51	85.64±5.38	98.25±1.17	82.71±6.48
	DCF	94.40±1.53	95.34±2.19	88.72±7.11	79.68±5.79	83.70±4.27	97.64±1.57	80.34±5.18
	DAF	94.69±2.05	95.47±2.66	91.83±6.31	77.71±9.27	83.87±6.52	98.42±1.22	80.73±7.69
MH	H	96.21±1.79	98.47±0.9	92.26±4.74	92.25±4.12	92.19±3.67	97.46±1.63	89.69±4.85
	V	96.48±0.93	98.43±0.97	93.80±3.50	91.72±3.10	92.67±1.83	97.99±1.30	90.36±2.44
	CF	98.61±1.12	99.52±0.50	97.54±2.57	96.75±2.58	97.13±2.30	99.21±0.83	96.21±3.04
	AF	98.28±1.37	99.51±0.40	96.28±3.37	96.78±2.98	96.50±2.73	98.77±1.12	95.36±3.65
	CMF	98.68±1.17	99.73±0.21	97.20±3.70	97.53±1.94	97.33±2.30	99.04±1.32	96.45±3.08
	AMF	98.48±1.24	99.59±0.32	95.75±4.08	98.33±1.84	96.97±2.37	98.51±1.56	95.96±3.21
	DCF	98.28±1.18	99.58±0.32	97.28±2.72	95.67±3.12	96.44±2.39	99.12±0.88	95.30±3.17
	DAF	98.28±1.32	99.61±0.32	96.21±4.23	96.97±3.16	96.51±2.59	98.69±1.52	95.37±3.47
RD	H	95.93±1.15	98.35±0.70	95.76±2.43	88.42±2.82	91.92±2.25	98.60±0.80	89.20±3.02
	V	96.90±1.06	98.73±0.73	97.53±2.66	90.58±3.56	93.86±2.09	99.13±1.03	91.79±2.79
	CF	98.45±0.95	99.54±0.28	98.38±2.89	95.83±2.89	97.03±1.77	99.39±1.11	95.99±2.42
	AF	98.52±0.57	99.50±0.38	97.88±2.44	96.56±2.26	97.16±1.02	99.21±0.91	96.16±1.41
	CMF	98.78±0.78	99.62±0.26	99.04±1.61	96.33±2.69	97.64±1.48	99.65±0.58	96.81±2.00
	AMF	98.58±0.84	99.57±0.28	98.58±2.17	96.08±2.69	97.28±1.56	99.48±0.80	96.32±2.13
	DCF	98.32±0.91	99.52±0.32	98.14±2.85	95.58±3.25	96.77±1.70	99.30±1.09	95.64±2.31
	DAF	98.71±0.79	99.53±0.40	98.57±1.88	96.58±2.43	97.54±1.45	99.48±0.70	96.67±1.99
RS	H	88.20±1.48	93.97±1.08	90.55±3.21	85.29±3.29	87.75±1.58	91.06±3.36	76.39±2.96
	V	89.09±2.15	94.78±1.36	88.76±3.18	89.48±3.47	89.05±2.16	88.70±3.57	78.17±4.30
	CF	93.41±1.35	97.31±0.87	94.57±1.83	92.07±2.90	93.26±1.45	94.74±1.88	86.83±2.70
	AF	93.86±1.32	97.48±0.76	94.61±2.26	92.97±2.01	93.75±1.34	94.74±2.25	87.71±2.63
	CMF	94.30±1.28	97.46±0.93	95.46±1.91	92.96±1.87	94.17±1.32	95.62±1.92	88.59±2.57
	AMF	93.28±1.20	97.09±1.16	94.82±1.98	91.53±2.85	93.10±1.29	95.00±2.04	86.56±2.40
	DCF	93.11±1.47	97.24±0.93	93.46±2.31	92.69±3.44	93.00±1.56	93.51±2.49	86.21±2.94
	DAF	92.93±1.46	97.08±0.88	93.61±2.02	92.06±2.15	92.81±1.48	93.78±2.17	85.85±2.91

Finally, we collected an additional 200 OCT images from 100 eyes as a pure test set to further verify the validity of our model. The number of HM, RS, MH, RD and PMCNV eyes is 13, 74, 26, 32 and 27 respectively, and there is one horizontal scan and one vertical scan image for each eye. We use these data to explore the performance of our model in Task 2, and the results are shown in TableVIII. In the four categories of RS, MH, RD and PMCNV, the method of feature fusion of dual-view OCT images in both horizontal and vertical scanning directions has better performance than that of OCT images using only a single view.And CMF has the best performance among the six fusion methods.

F. Robustness to Three Public Datasets

In order to further verify the effectiveness of our proposed method in image classification task, we apply our method to three public datasets. The three publicly available datasets address the classification of diseases such as AMD, DMA, CNV, and MH. Consistent with our proprietary dataset, we experiment with our method and seven other comparative methods. Same as [36], [50] and [33], the experiment was carried out with the 5-fold cross-validation strategy for dataset A, and the experiment was carried out with the 10-fold cross-validation strategy for dataset B and dataset C. The experimental results of dataset C is verified on the pure test set. As shown in Table IX, our method achieves the better classification performance. In dataset A, dataset B and dataset C, the classification accuracy has reached 99.03%, 99.80% and 93.12%, respectively. It can be seen that our method not only achieves the best performance in the PM classification task, but also achieves the best performance in other medical image classification tasks. The above results further prove the generalization ability of the proposed method.

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on May 18,2023 at 16:06:16 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3260990

TABLE VII
TASK 2: COMPARISON OF FOUR CLASSIFIERS UNDER DIFFERENT FUSION METHODS(%)

Method	ACC	AUC	Pre	Sen	F1	Spec	Kappa
H	92.73±3.44	95.58±2.97	89.67±5.82	82.73±10.55	85.79±8.55	95.84±2.88	80.26±9.92
V	93.44±3.34	96.17±2.51	90.40±6.00	85.07±9.60	87.34±8.02	95.53±4.07	82.25±9.46
CF	96.20±2.35	97.97±1.68	95.41±2.84	90.19±8.32	92.52±5.91	97.92±1.88	89.58±7.04
AF	96.41±2.03	98.06±1.56	**95.71±1.49**	90.89±8.02	92.96±5.10	97.90±1.83	90.18±6.01
CMF	**96.68±2.06**	**98.14±1.65**	95.33±3.54	**92.12±6.27**	**93.58±5.03**	**98.05±1.54**	**91.01±6.06**
AMF	96.37±2.26	97.97±1.69	95.09±2.64	91.71±6.70	93.25±4.69	97.81±1.68	90.39±5.91
DCF	96.03±2.32	97.92±1.76	94.40±3.72	90.90±6.59	92.48±5.28	97.39±2.33	89.37±6.44
DAF	96.15±2.43	97.92±1.75	95.06±2.56	90.83±7.82	92.68±5.38	97.59±2.24	89.66±6.63

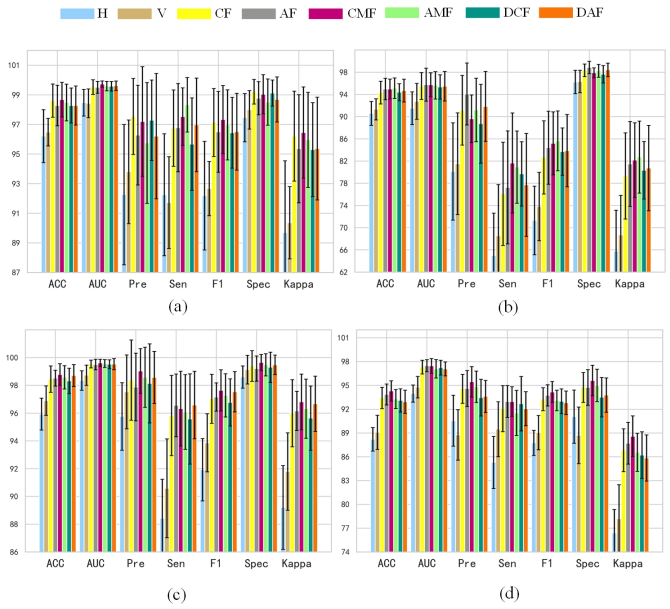


Fig. 12. Histogram of performance comparison between fusion methods and single view only. (a)-(d) represent four binary classifiers, such as MH, PMCNV, RD, and RS, respectively.

TABLE VIII
TASK 2: SUPPLEMENT EXPERIMENT WITH ADDITIONAL TEST SET(%)

Method	ACC	Pre	Sen	F1	Spec
H	91.51	99.26	88.97	93.76	96.15
V	89.56	100.0	85.37	91.76	100.0
CF	93.08	100.0	90.55	94.98	100.0
AF	92.77	100.0	90.32	94.79	100.0
CMF	**94.31**	**100.0**	**92.51**	**96.03**	**100.0**
AMF	94.05	100.0	92.39	95.86	100.0
DCF	93.46	99.61	91.73	95.47	98.08
DAF	94.02	100.0	92.03	95.79	100.0

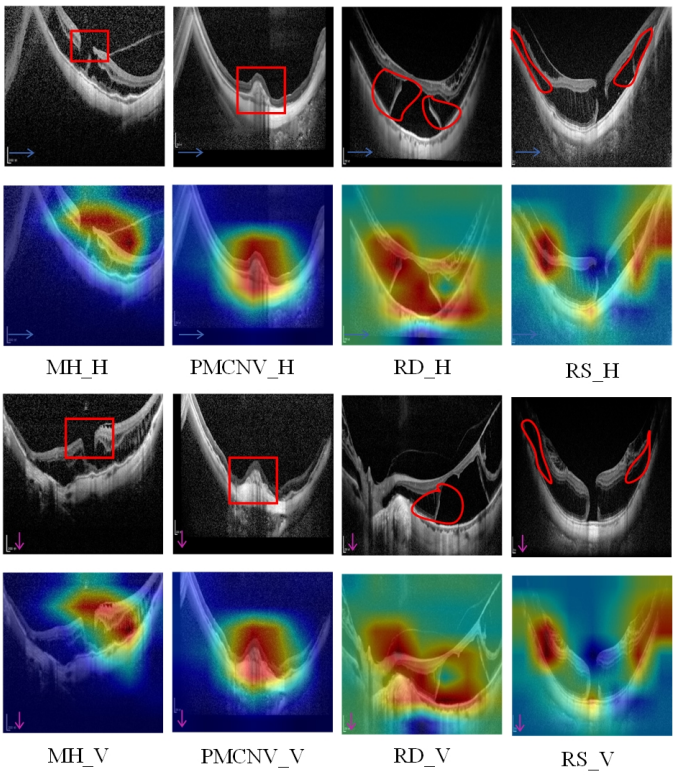


Fig. 13. Grad-CAMs visualization of different types of pathological changes using CMF.

G. Statistical analysis of results

Furthermore, this study calculates the confidence of each method when making corresponding category decision under four datasets, and perform statistical analysis. Specifically, for a k-class classification task, the last network layer of the model will output the predicted probability P_i (i=1, 2, $\cdots$,k) of each category for each input sample. Finally, the category j corresponding to the maximum probability is selected as the predicted category of the model. Then the probability of the predicted category j of the model for the input sample is P_j ($j \leq k$). In each fold test experiment, we summarize the predicted category probability P_j of each comparison model for all test samples, and draw the result as a violin diagram, as shown in Fig.14. The corresponding width of the violin diagram at a given confidence value indicates how often that confidence value occurs in the statistical sample. The wider the width of the violin diagram, the higher the frequency of the corresponding confidence value in the sample, conversely, the narrower the width, the lower the frequency of the corresponding confidence value in the sample. The white dots in the violin plot represent the median of the data. When the median is near the top, it means that most data values are large. The black bar in the center of the violin represents the range of quartile, while the top and bottom of the bar represent the third quartile and the first quartile, respectively. The height of the black bars reflects the volatility of the confidence value to some extent, and a smaller height means that most confidence value are distributed over a smaller range.

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on May 18,2023 at 16:06:16 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3260990

TABLE IX
EXPERIMENTAL RESULTS OF THE THREE PUBLIC DATASETS (%)

Dataset	Method	ACC	AUC	Pre	Sen	F1	Spec	Kappa
Dataset A	ViT	95.48±0.85	99.67±0.13	95.71±0.78	95.48±0.85	95.48±0.85	97.64±0.46	93.09±1.31
	ResNet50	94.98±0.47	99.52±0.18	95.12±0.45	94.98±0.47	94.95±0.49	97.41±0.27	92.33±0.74
	CrossViT-Ti	95.46±0.53	99.53±0.15	95.56±0.51	95.46±0.53	95.46±0.53	97.63±0.26	93.07±0.78
	DeiT	95.69±0.71	99.67±0.08	95.75±0.72	95.69±0.71	95.68±0.73	97.77±0.37	93.41±1.10
	E-Net	94.78±0.60	99.48±0.22	94.90±0.54	94.78±0.60	94.75±0.63	97.25±0.35	92.00±0.96
	Swin-T	96.27±0.26	99.82±0.07	96.42±0.21	96.27±0.26	96.26±0.27	98.05±0.18	94.30±0.42
	ConvNeXt-T	96.26±0.47	99.71±0.07	96.31±0.46	96.26±0.47	96.26±0.48	98.11±0.26	94.29±0.75
	MCME [36]	-	99.80	99.39±1.2	99.36±1.33	99.34±1.3	-	-
	Proposed	99.03±0.58	99.93±0.04	99.03±0.58	99.03±0.58	99.02±0.58	99.50±0.30	98.52±0.89
Dataset B	ViT	96.42±0.67	99.69±0.17	96.48±0.64	96.42±0.67	96.41±0.67	98.18±0.36	94.48±1.03
	ResNet50	96.64±1.29	99.65±0.25	96.68±1.29	96.64±1.29	96.62±1.31	98.25±0.68	94.81±2.00
	CrossViT-Ti	98.17±0.35	99.91±0.03	98.19±0.34	98.17±0.35	98.17±0.36	99.07±0.20	97.18±0.55
	DeiT	97.58±0.76	99.82±0.12	97.61±0.74	97.58±0.76	97.58±0.76	98.79±0.38	96.27±1.17
	E-Net	96.72±0.59	99.72±0.11	96.73±0.59	96.72±0.59	96.71±0.59	98.33±0.30	94.93±0.91
	Swin-T	98.39±0.32	99.91±0.04	98.41±0.32	98.39±0.32	98.39±0.32	99.21±0.15	97.52±0.50
	ConvNeXt-T	98.49±0.40	99.93±0.03	98.51±0.39	98.49±0.40	98.49±0.41	99.24±0.24	97.68±0.62
	Proposed	99.80±0.16	99.99±0.01	99.80±0.16	99.80±0.16	99.80±0.16	99.90±0.08	99.69±0.25
Dataset C	ViT	90.70	99.70	91.88	90.70	90.34	96.90	87.60
	ResNet50	90.10	99.59	91.88	90.10	89.87	96.70	86.80
	CrossViT-Ti	91.30	99.58	92.38	91.30	91.06	97.10	88.40
	DeiT	92.70	99.88	93.85	92.70	92.56	97.57	90.27
	E-Net	91.80	99.79	93.31	91.80	91.61	97.27	89.07
	Swin-T	93.00	99.72	93.98	93.00	92.90	97.67	90.67
	ConvNeXt-T	92.30	99.69	93.52	92.30	92.21	97.43	89.73
	Kermany *et al.* [33]	82.90	98.73	85.36	82.90	81.42	94.30	77.20
	Proposed	95.30	99.82	95.69	95.30	95.31	98.43	93.73

In general, our method performs better in the four datasets. In the black bar at the center of the violin, our method has a lower height, which indicates that our method has less fluctuation and more concentrated confidence when making decisions. It is most evident in our private dataset, dataset A and dataset B. Compared with other methods, the confidence of our method has a larger median, and the confidence interval is concentrated on a higher interval. Taking our private dataset as an example, the confidence of a large part of the data is concentrated on the interval above 0.980, and a small part of the data is lower than 0.875.

In this study, t test was further conducted on the predicted category confidence of the proposed FIT-Net and the other seven comparison methods, so as to analyze whether the comparison of classification performance between each comparison method and the FIT-Net has statistical significance. The null hypothesis given here is: the classification performance comparison between each comparison method and the FIT-Net is not statistically significant. Assuming that X is the set of probability P_j of the predicted category j of the FIT-Net for each test set sample, X_i $(i \leq 7)$ is the set of probability P_j of the predicted category j of the comparison method for each test set sample. X and X_i are the two groups of statistics in the test, and t test is conducted. According to the specific distribution of the two groups of test statistics, the p value can be obtained. The p-value is the probability of obtaining a test result that is at least as extreme as the one actually observed, assuming that the original hypothesis is correct. In statistics, the p-value obtained by statistics according to the significance test is generally set as the threshold of significance level 0.05, and the null hypothesis can be rejected only when the p-value is less than or equal to 0.05, then the hypothesis test will indeed have significance level. As can be seen from the results in TableX, in the four datasets, the p-values obtained by the t test of FIT-Net and the other 7 comparison methods are all less than 0.05, which indicates that the probability of the occurrence of the null hypothesis is very small and there are sufficient reasons to reject the null hypothesis. Therefore, the classification performance comparison between the FIT-Net and these 7 comparison methods has statistical significance.

H. Model complexity comparison

We further explored the performance of large parameter network models in task 1, including CrossViT-B [19], Swin-L [16], ConvNext-L [54], and the classification results are shown in Table XI. We further analyze the complexity of the model and compare the number of model parameters and floating point operations (FLOPs) between different methods. The comparison results of the model are shown in Table XII , and ACC indicates the average classification accuracy of task1. The model with a small number of parameters don't get better performance. Compared with ResNet50, the complexity of E-Net and CrossViT-Ti is smaller, and the model accuracy increases by only 1%. Generally speaking, as the model parameters increases, the accuracy of the model will also increase. Compared with CrossViT-Ti, the parameters of CrossViT-B have increased by more than 10 times, and the accuracy rate has increased by 1.25%. Compared with Swin-T, the parameters of Swin-L has increased several times, and the accuracy rate has increased by 0.47%. It can be seen that the increase in the number of parameters of these models has not brought about a significant increase in performance. On the contrary, the more complex the model, the overfitting phenomenon will often occur in the process of model training, resulting in the decline of the accuracy of the model.

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on May 18,2023 at 16:06:16 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3260990

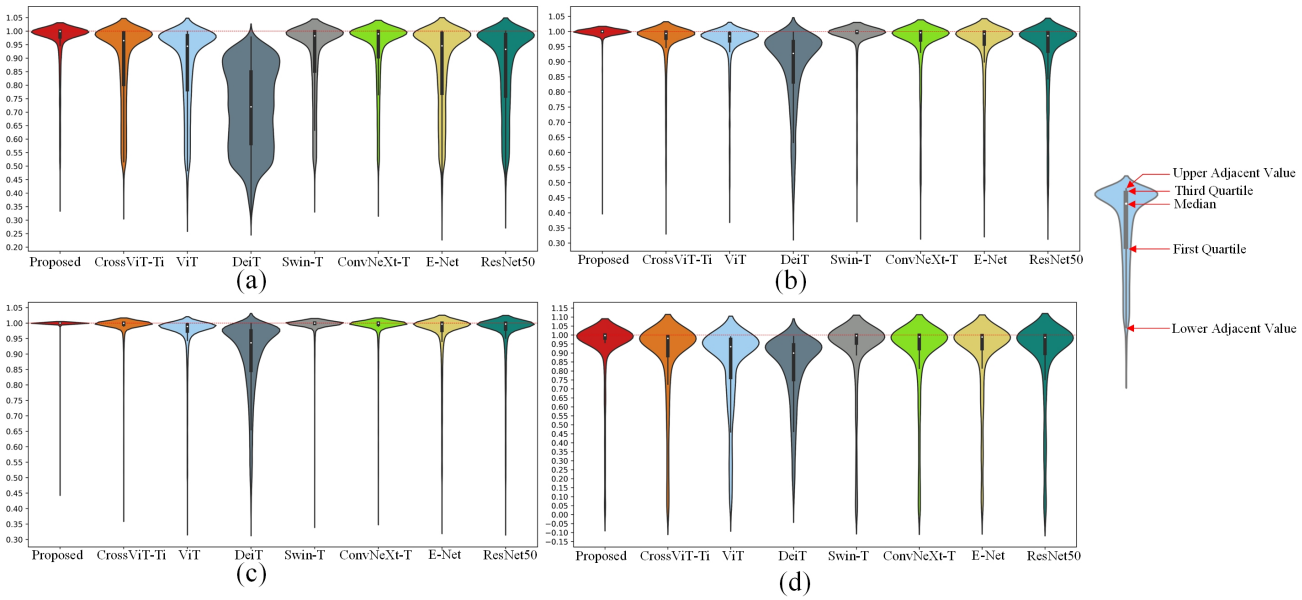


Fig. 14. The violin diagram is drawn according to the confidence analysis of each method when making the corresponding category decision under four datasets.(a)-(d) represent the results of private dataset, dataset A, dataset B and dataset C, respectively.

TABLE X
STATISTICAL SIGNIFICANCE ANALYSIS BETWEEN THE PROPOSED METHOD AND THE COMPARISON METHOD

Dataset	Method	ViT	ResNet50	CrossViT-Ti	DeiT	E-Net	Swin-T	ConvNeXt-T
Ours	p-value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.001	<0.001
Dataset A	p-value	<0.001	<0.001	<0.001	<0.0001	<0.001	<0.001	<0.001
Dataset B	p-value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.001	<0.0001
Dataset C	p-value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.001	<0.0001

TABLE XI
CLASSIFICATION PERFORMANCE OF LARGE PARAMETER NETWORK MODEL IN TASK 1

Method	ACC	AUC	Pre	Sen	F1	Spec	Kappa
CrossViT-B	88.42±1.84	97.77±0.61	88.94±1.56	88.42±1.84	88.35±1.82	96.08±0.58	82.58±2.69
Swin-L	89.30±0.72	98.17±0.42	89.68±0.76	89.30±0.72	89.32±0.71	96.42±0.24	83.94±1.09
ConvNext-L	87.88±2.09	97.67±0.75	88.36±1.96	87.88±2.09	87.88±2.09	95.94±0.70	81.82±3.10

TABLE XII
MODEL COMPLEXITY ANALYSIS

Method	input	#param	FLOPs	ACC
ViT	224×224	85.80M	16.86G	85.68
ResNet50	224×224	23.52M	4.12G	86.53
CrossViT-Ti	224×224	6.73M	1.29G	87.17
CrossViT-B	224×224	103.88M	20.11G	88.42
DeiT	224×224	85.81M	16.94G	87.18
E-Net	224×224	4.01M	398.09M	87.54
Swin-T	224×224	27.52M	4.36G	88.83
Swin-L	224×224	195.0M	34.04G	89.30
ConvNeXt-T	224×224	27.82M	4.47G	89.27
ConvNext-L	224×224	196.24M	34.4G	87.88
Ours	224×224	95.07M	6.69G	92.03

Compared with ConvNeXt-T, the number of parameters of ConvNeXt-L has increased several times, but its accuracy has dropped by 1.39%. The model accuracy of ViT is lower than that of ResNet50, but the model complexity is greatly improved. The model complexity of DeiT is similar to that of ViT, but the accuracy is 0.65% higher than that of ResNet50. The model complexity of Swin-T and ConvNeXt-T is similar to that of ResNet50, and the model accuracy is 2.3% higher. Although the parameters of our proposed model are not the smallest, compared with models with large parameters such as CrossViT-B, Swin-L and ConvNeXt-L, the parameters of our model are relatively small, and the performance is the best among all compared models, and the accuracy rate is 4.15% higher than that of the largest parameter model ConvNeXt-L.

I. Anti-jamming performance of model

We explore the proposed method to evaluate the anti-jamming performance of the model when different levels of speckle noise are added to OCT images. In each test dataset, we add different proportions of speckle noise to the data to explore the performance changes of our proposed method, which uses four interference of different proportions of speckle noise(e.g., the proportions of 1%, 3%, 5% and 10%). The experimental results of the four datasets are shown in Fig. 15. As can be seen from Fig.15, our method has the best performance on dataset B, even when 10% noise is added, the classification accuracy only decreases by 0.21%. In general,

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on May 18,2023 at 16:06:16 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3260990

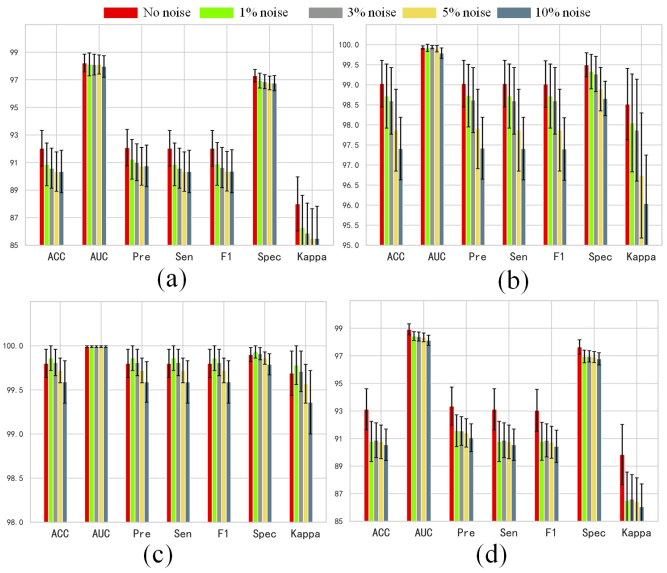


Fig. 15. Performance of the proposed method under different levels of noise. (a)-(d) represent the results of private dataset, dataset A, dataset B, dataset C.

as the proportion of noise increases, the performance of the model will gradually decrease. However, it can be seen from the experimental results that with the increase of noise ratio, the performance of our method will not be significantly degraded. Under the addition of 10% interference noise, the classification accuracy of our dataset, dataset A and dataset C is only decreased by 1.68%, 1.62% and 2.57%, respectively. Therefore, the anti-interference ability of our method is quite strong.

V. Conclusion

In this paper, we propose a novel FIT-Net for two mutually progressive classification tasks for Task 1 and Task 2. Our proposed FIT-Net makes full use of the Transformer encoder's ability to build long-range dependencies, and uses the Transformer encoder extracted by CNN to mine the effective information in the multi-scale features. At the same time, the interaction of information between different scale features is realized by embedding the IA unit, which effectively makes up for the shortcomings of CNN that only pays attention to the local features of the image, thereby achieving better classification performance. The experimental results on two classification tasks on the private clinical OCT dataset show that the proposed FIT-Net outperforms several well-known classical deep learning methods. Our method also achieves better performance on three public datasets. The experimental results of Task 2 also fully demonstrate that the fusion of dual-view OCT image features can effectively improve the diagnostic accuracy of four pathological changes in PM. The proposed method can help ophthalmologists quickly and accurately predict the likely occurrence of PM and further pathologic changes through easily accessible OCT images, which improves workflow and reduces clinical errors in the absence of well-trained ophthalmologists.

References

[1] J. Ruiz-Medrano, J. A. Montero, I. Flores-Moreno, L. Arias, *et al.*, "Myopic maculopathy: current status and proposal for a new classification and grading system (atn)," *Progress in retinal and eye research*, vol. 69, pp. 80–115, 2019.

[2] Y.-L. Wong and S.-M. Saw, "Epidemiology of pathologic myopia in asia and worldwide," *The Asia-Pacific Journal of Ophthalmology*, vol. 5, no. 6, pp. 394–402, 2016.

[3] T. Y. Wong, A. Ferreira, R. Hughes, G. Carter, *et al.*, "Epidemiology and disease burden of pathologic myopia and myopic choroidal neovascularization: an evidence-based systematic review," *American journal of ophthalmology*, vol. 157, no. 1, pp. 9–25, 2014.

[4] N. Leveziel, V. Caillaux, S. Bastuji-Garin, M. Zmuda, *et al.*, "Angiographic and optical coherence tomography characteristics of recent myopic choroidal neovascularization," *American Journal of Ophthalmology*, vol. 155, no. 5, pp. 913–919, 2013.

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[6] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature medicine*, vol. 24, no. 9, pp. 1342–1350, 2018.

[7] W. Lu, Y. Tong, Y. Yu, Y. Xing, *et al.*, "Deep learning-based automated classification of multi-categorical abnormalities from optical coherence tomography images," *Translational vision science & technology*, vol. 7, no. 6, pp. 41–41, 2018.

[8] C. S. Lee, D. M. Baughman, and A. Y. Lee, "Deep learning is effective for classifying normal versus age-related macular degeneration oct images," *Ophthalmology Retina*, vol. 1, no. 4, pp. 322–327, 2017.

[9] Y. Rong, D. Xiang, W. Zhu, K. Yu, *et al.*, "Surrogate-assisted retinal oct image classification based on convolutional neural networks," *IEEE journal of biomedical and health informatics*, vol. 23, no. 1, pp. 253–263, 2018.

[10] R. Rasti, A. Mehridehnavi, H. Rabbani, and F. Hajizadeh, "Automatic diagnosis of abnormal macula in retinal optical coherence tomography images using wavelet-based convolutional neural network features and random forests classifier," *Journal of biomedical optics*, vol. 23, no. 3, p. 035005, 2018.

[11] M. Sonka and M. D. Abràmoff, "Quantitative analysis of retinal oct," pp. 165–169, 2016.

[12] Y. Song, J. Y.-C. Teoh, K.-S. Choi, and J. Qin, "Dynamic loss weighting for multiorgan segmentation in medical images," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2023.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[15] K. Han, A. Xiao, E. Wu, J. Guo, *et al.*, "Transformer in transformer," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15 908–15 919, 2021.

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[17] L. Mou, Y. Zhao, H. Fu, Y. Liu, *et al.*, "Cs2-net: Deep learning segmentation of curvilinear structures in medical imaging," *Medical image analysis*, vol. 67, p. 101874, 2021.

[18] H. Touvron, M. Cord, M. Douze, F. Massa, *et al.*, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 347–10 357.

[19] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 357–366.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] Y. Li, W. Feng, X. Zhao, B. Liu, *et al.*, "Development and validation of a deep learning system to screen vision-threatening conditions in high myopia using optical coherence tomography images," *British Journal of Ophthalmology*, vol. 106, no. 5, pp. 633–639, 2022.

[22] L. Fang, C. Wang, S. Li, H. Rabbani, *et al.*, "Attention to lesion: Lesion-aware convolutional neural network for retinal optical coherence tomography image classification," *IEEE transactions on medical imaging*, vol. 38, no. 8, pp. 1959–1970, 2019.

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on May 18,2023 at 16:06:16 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3260990

[23] L. Huang, X. He, L. Fang, H. Rabbani, *et al.*, "Automatic classification of retinal optical coherence tomography images with layer guided convolutional neural network," *IEEE Signal Processing Letters*, vol. 26, no. 7, pp. 1026–1030, 2019.

[24] S. P. K. Karri, D. Chakraborty, and J. Chatterjee, "Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration," *Biomedical optics express*, vol. 8, no. 2, pp. 579–592, 2017.

[25] F. G. Venhuizen, B. van Ginneken, B. Bloemen, M. J. van Grinsven, *et al.*, "Automated age-related macular degeneration classification in oct using unsupervised feature learning," in *Medical Imaging 2015: Computer-Aided Diagnosis*, vol. 9414. SPIE, 2015, pp. 391–397.

[26] G. Lemaître, M. Rastgoo, J. Massich, C. Y. Cheung, *et al.*, "Classification of sd-oct volumes using local binary patterns: experimental validation for dme detection," *Journal of ophthalmology*, vol. 2016, 2016.

[27] Y. Sun, S. Li, and Z. Sun, "Fully automated macular pathology detection in retina optical coherence tomography images using sparse coding and dictionary learning," *Journal of biomedical optics*, vol. 22, no. 1, p. 016012, 2017.

[28] Y.-Y. Liu, M. Chen, H. Ishikawa, G. Wollstein, *et al.*, "Automated macular pathology diagnosis in retinal oct images using multi-scale spatial pyramid and local binary patterns in texture and shape encoding," *Medical image analysis*, vol. 15, no. 5, pp. 748–759, 2011.

[29] A. Albarrak, F. Coenen, Y. Zheng, *et al.*, "Age-related macular degeneration identification in volumetric optical coherence tomography using decomposition and local feature extraction," in *Proceedings of 2013 international conference on medical image, understanding and analysis*, 2013, pp. 59–64.

[30] S. Farsiu, S. J. Chiu, R. V. O'Connell, F. A. Folgar, *et al.*, "Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography," *Ophthalmology*, vol. 121, no. 1, pp. 162–172, 2014.

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[33] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, *et al.*, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[35] Z. Xu, W. Wang, J. Yang, J. Zhao, *et al.*, "Automated diagnoses of age-related macular degeneration and polypoidal choroidal vasculopathy using bi-modal deep convolutional neural networks," *British Journal of Ophthalmology*, vol. 105, no. 4, pp. 561–566, 2021.

[36] R. Rasti, H. Rabbani, A. Mehridehnavi, and F. Hajizadeh, "Macular oct classification using a multi-scale convolutional neural network ensemble," *IEEE transactions on medical imaging*, vol. 37, no. 4, pp. 1024–1034, 2017.

[37] Y. Song, L. Yu, B. Lei, K.-S. Choi, *et al.*, "Data discernment for affordable training in medical image segmentation," *IEEE Transactions on Medical Imaging*, pp. 1–1, 2022.

[38] K. Ohno-Matsui, T. Y. Lai, C.-C. Lai, and C. M. G. Cheung, "Updates of pathologic myopia," *Progress in retinal and eye research*, vol. 52, pp. 156–187, 2016.

[39] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[40] X. He, Y. Deng, L. Fang, and Q. Peng, "Multi-modal retinal image classification with modality-specific attention network," *IEEE Transactions on Medical Imaging*, vol. 40, no. 6, pp. 1591–1602, 2021.

[41] L. Yuan, Y. Chen, T. Wang, W. Yu, *et al.*, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 558–567.

[42] D. A. Hudson and L. Zitnick, "Generative adversarial transformers," in *International conference on machine learning*. PMLR, 2021, pp. 4487–4499.

[43] N. Carion, F. Massa, G. Synnaeve, N. Usunier, *et al.*, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[44] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 873–12 883.

[45] K. Han, Y. Wang, H. Chen, X. Chen, *et al.*, "A survey on vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[46] B. Graham, A. El-Nouby, H. Touvron, P. Stock, *et al.*, "Levit: a vision transformer in convnet's clothing for faster inference," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 259–12 269.

[47] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, *et al.*, "Bottleneck transformers for visual recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 519–16 529.

[48] H. Wu, B. Xiao, N. Codella, M. Liu, *et al.*, "Cvt: Introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 22–31.

[49] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[50] M. Tajmirriahi, Z. Amini, A. Hamidi, A. Zam, *et al.*, "Modeling of retinal optical coherence tomography based on stochastic differential equations: Application to denoising," *IEEE Transactions on Medical Imaging*, vol. 40, no. 8, pp. 2129–2141, 2021.

[51] T.-Y. Lin, P. Goyal, R. Girshick, K. He, *et al.*, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[52] T. He, Z. Zhang, H. Zhang, Z. Zhang, *et al.*, "Bag of tricks for image classification with convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 558–567.

[53] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[54] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, *et al.*, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 976–11 986.

[55] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, pp. 837–845, 1988.

[56] X. Sun and W. Xu, "Fast implementation of delong's algorithm for comparing the areas under correlated receiver operating characteristic curves," *IEEE Signal Processing Letters*, vol. 21, no. 11, pp. 1389–1393, 2014.

[57] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[58] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, *et al.*, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

Authorized licensed use limited to: Sungkyunkwan University. Downloaded on May 18,2023 at 16:06:16 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.