JointViT: Modeling Oxygen Saturation Levels with Joint Supervision on Long-Tailed OCTA

Zeyu Zhang[1,2,*], Xuyin Qi[3], Mingxi Chen[4], Guangxi Li[5], Ryan Pham[3], Ayub Qassim[1], Ella Berry[1], Zhibin Liao[3], Owen Siggs[1], Robert Mclaughlin[3], Jamie Craig[1], and Minh-Son To[1]

[1] Flinders University
[2] The Australian National University
[3] The University of Adelaide
[4] Guangdong Technion - Israel Institute of Technology
[5] The University of Sydney

https://steve-zeyu-zhang.github.io/JointViT

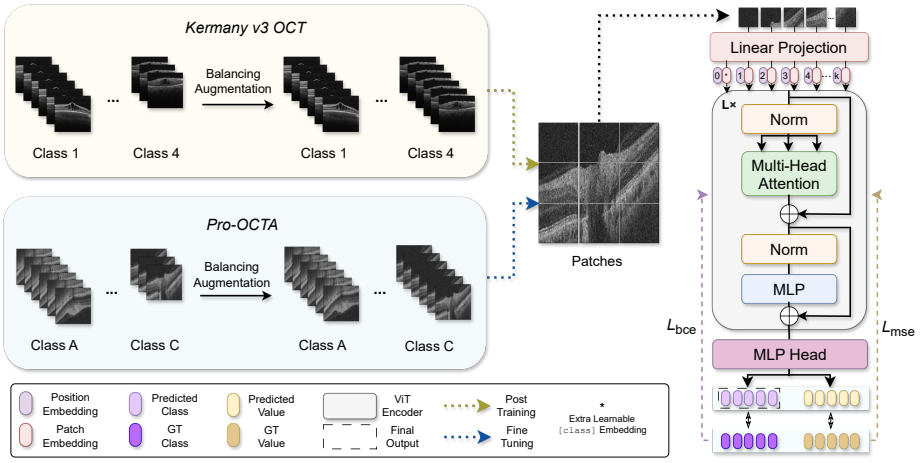


Fig. 1: The figure illustrates the pipeline of our proposed **JointViT**, which comprises a **balancing augmentation** and a plain Vision Transformer with a **joint loss** for supervision. The classes are denoted as numbers in Kermany v3 dataset [15] and alphabets in Prog-OCTA. *GT* abbreviates ground truth.

Abstract. The oxygen saturation level in the blood (SaO_2) is crucial for health, particularly in relation to sleep-related breathing disorders. However, continuous monitoring of SaO_2 is time-consuming and highly variable depending on patients' conditions. Recently, optical coherence tomography angiography (OCTA) has shown promising development in rapidly and effectively screening eye-related lesions, offering the potential for diagnosing sleep-related disorders. To bridge this gap, our paper

*Work done while being a student researcher at Flinders Health and Medical Research Institute, Flinders University.

presents three key contributions. Firstly, we propose **JointViT**, a novel model based on the Vision Transformer architecture, incorporating a **joint loss** function for supervision. Secondly, we introduce a **balancing augmentation** technique during data preprocessing to improve the model's performance, particularly on the long-tail distribution within the OCTA dataset. Lastly, through comprehensive experiments on the OCTA dataset, our proposed method significantly outperforms other state-of-the-art methods, achieving improvements of up to **12.28%** in overall accuracy. This advancement lays the groundwork for the future utilization of OCTA in diagnosing sleep-related disorders.

Keywords: Optical Coherence Tomography Angiography · Oxygen Saturation Levels · Long-Tailed Distribution.

1 Introduction

Microvascular dysfunction describes a constellation of vessel and endothelial changes including vessel destruction, abnormal vasoreactivity, and thrombosis that lead to tissue damage and progressive organ failure. Microvascular disease has been linked to an increased risk of stroke, cognitive decline, retinopathy, vascular nephropathy and heart failure. Since microvascular dysfunction seen in specific organs may be a manifestation of systemic disease, studying the microvasculature of the retina provides a window into systemic disease, and may serve as a surrogate marker of cumulative injury to the systemic microvasculature [1, 2]. Early detection of microvascular disease provides an opportunity for timely intervention and prevention of future complications.

Obstructive sleep apnoea (OSA) is the most common sleep-related disorder [3]. Microvascular endothelial dysfunction in OSA driven by vascular stresses during sleep has been linked to the development of hypertension and atherosclerosis, and treatment of OSA with nocturnal continuous positive airway pressure has been shown to improve endothelial function, improve patient's sleep-related symptoms, and reduce the risk of future cardiovascular morbidity [4]. Untreated, OSA is associated with adverse outcomes such as cardiovascular and cerebrovascular disease, cognitive decline, and depression.

Low blood oxygen levels during sleep may indicate the presence of OSA or a sleep-related breathing disorder [27]. Oxygen saturation level in the blood (SaO_2) indicates the percentage of hemoglobin binding sites in the bloodstream occupied by oxygen molecules [10]. This metric serves as a crucial indicator of the body's ability to adequately oxygenate tissues and organs [29]. Typically, SaO_2 is acquired by a pulse oximeter, which can monitor blood oxygen levels continuously during sleep [17]. However, as a non-invasive method, it may yield results with some variations and errors that may not fully reflect actual oxygen levels [25]. Conducting sleep studies also involves monitoring patients all night long with the help of these devices [25]. This process is time-consuming and demands continuous monitoring in a clinical setting.

Optical coherence tomography (OCT) is a non-invasive imaging test that utilises infrared light to obtain 2D and 3D visualisation of structures with micrometre resolution. OCT angiography (OCTA) is an extension of the OCT technique that extracts blood flow information of the retina, which allows both vessel structure and haemodynamic parameters to be measured [5] [32]. OCTA scans can be acquired in a matter of seconds non-invasively without the use of an intravenous contrast [5]. Unlike OCT, which primarily provides structural images of the eye, OCTA allows for the visualization of blood flow within the retina and choroid without the need for contrast agents [7].

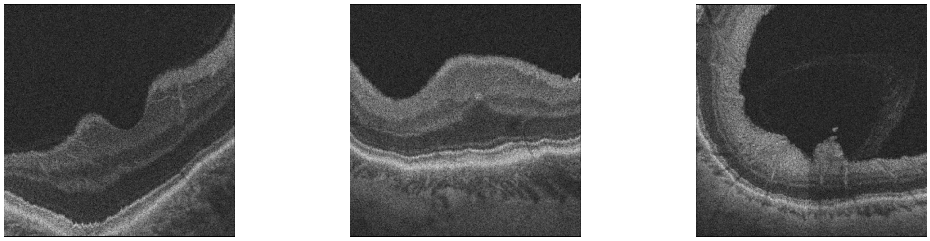
The ability of OCTA to image the microvasculature of the retina opens the door to screening not only for sleep-related disordered conditions but also for various other systemic conditions [37]. However, since the OCTA data is acquired in hospitals where there are more patients with sleep-related disorders and fewer healthy instances, this leads to a long-tailed distribution, making the prediction considerably challenging. Moreover, directly modeling concrete SaO_2 values based on a limited and long-tailed OCTA dataset appears to be impractical. Instead, predicting SaO_2 categories is much more robust and will benefit the diagnosis of sleep-related breathing disorders. In pursuit of this objective, we have proposed a novel approach grounded in Vision Transformer [6] architecture to leverage OCTA data for predicting SaO_2 categories. Our contributions can be succinctly summarized in the following three points.

- We proposed a novel model named **JointViT** built upon the plain Vision Transformer architecture, which incorporates a well-designed **joint loss** function that leverages both SaO_2 categories and SaO_2 values for supervision.
- We introduced a **balancing augmentation** technique within the data preprocessing phase to enhance the model's performance specifically on the long-tail distribution within the OCTA dataset.
- We further conducted comprehensive experiments on the OCTA dataset, which demonstrated that our proposed method has significantly outperformed other state-of-the-art methods, yielding improvements of up to **12.28%** in accuracy. This advancement lays the groundwork for the future utilization of OCTA in diagnosing sleep-related disorders.

2 Related Works

Medical Imaging Recognition Recognizing medical imaging is a crucial and fundamental task within medical image analysis, playing a vital role in computer vision for medical imaging. The long-term goal of the medical imaging community is to design an effective model architecture and develop an efficient learning strategy to acquire a robust representation of medical imaging. M3T [14] proposed a synergistic method combining 3D CNN, 2D CNN, and Transformer [36] to accurately classify Alzheimer's disease in 3D MRI images, achieving the highest performance in multi-institutional validation databases and

demonstrating the feasibility of efficiently integrating CNN and Transformer for 3D medical images. MedViT [22] proposed a hybrid CNN-Transformer method to enhance robustness against adversarial attacks and improve generalization ability in medical image diagnosis. FCNlinksCNN [28] proposed a deep learning method to delineate unique Alzheimer's disease signatures from multimodal inputs, integrating MRI, age, gender, and Mini-Mental State Examination score. It achieves accurate diagnosis by linking a fully convolutional network with a multilayer perceptron, producing precise visualizations of individual Alzheimer's disease risk. MedicalNet [4] proposed a heterogeneous 3D network method based on 3D-ResNet [12] to extract general medical three-dimensional (3D) features, achieving accelerated training convergence and improved accuracy in various 3D medical imaging tasks, including lung segmentation, pulmonary nodule classification, and liver segmentation. COVID-ViT [8] proposed 2D and 3D forms of Vision Transformer [6] deep learning architecture to classify COVID from non-COVID based on 3D CT lung images, achieving superior performance compared to DenseNet [13].



(a) Low (b) Borderline Low (c) Normal

Fig. 2: The figure illustrates OCTA instances corresponding to each level of SaO_2 in Prog-OCTA dataset.

Long-Tailed Image Recognition Addressing the long-tailed problem in computer vision remains a challenging and critical task due to the severe class imbalance, where a few classes dominate while many others are underrepresented. This leads to biased predictions favoring the major classes, posing significant hurdles for accurate model performance. OLTR [19] proposed a dynamic meta-embedding method and introduced a dataset for Open Long-Tailed Recognition (OLTR) to address imbalanced classification, few-shot learning, and open-set recognition simultaneously. Logit adjustment [24] presented a method utilizing a novel balanced cross entropy (Bal-CE) loss to adjust logits based on label frequencies, encouraging a large relative margin between logits of rare and dominant labels. LiVT [39] proposed a novel method utilizing a balanced binary cross entropy (Bal-BCE) loss to address the challenges of long-tailed recognition by training Vision Transformer from scratch with long-tailed data. This approach achieves significant improvements without requiring additional data.

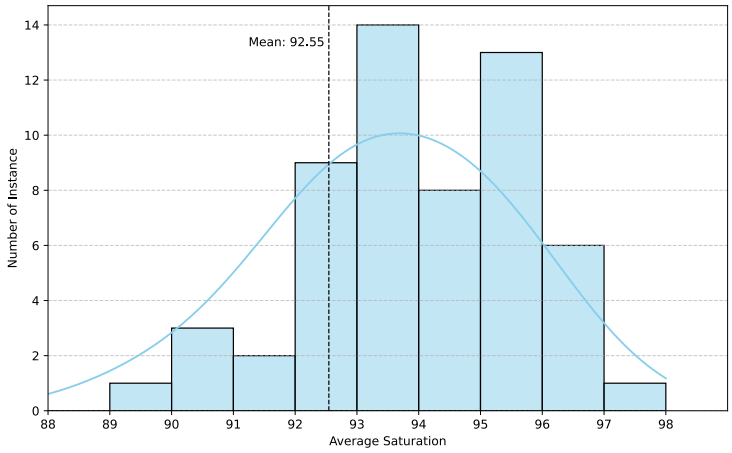


Fig. 3: The figure illustrates the SaO_2 value of patients in Prog-OCTA dataset, and the distribution of the dataset is imbalanced and apparently has a lower average SaO_2 than the normal people.

OCTA in AI for Health Optical Coherence Tomography Angiography (OCTA) is an enhancement of OCT, enabling noninvasive imaging of ocular structures with improved visualization of retinal and anterior eye anatomy compared to conventional methods [16]. OCTA-Net [20] proposed a split-based coarse-to-fine vessel segmentation method and introduced a new dataset, ROSE, focusing on retinal OCTA images, in order to address challenges in automated segmentation of retinal vessels, achieving improved performance over traditional and other deep learning methods and enabling analysis of retinal microvasculature for studying neurodegenerative diseases. Xie et al. [38] proposed a standardized OCTA analysis framework, utilizing deep learning models to extract geometrical parameters of retinal microvasculature, which enabled the comparison of these parameters among healthy controls and subjects with Alzheimer's Disease or mild cognitive impairment (MCI), demonstrating the potential of OCTA as a diagnostic tool for these conditions. Polar-Net [18] proposed a novel deep-learning framework, involving the transformation of OCTA images from Cartesian to polar coordinates, to facilitate region-based analysis akin to the ETDRS grid commonly used in clinical practice. Additionally, it integrates clinical prior information and adapts to assess the importance of retinal regions, thereby enhancing interpretability and performance in detecting Alzheimer's disease.

Oxygen Saturation Prediction Arterial oxygen saturation (SaO_2) serves as a crucial and frequently used parameter in diagnosing sleep apnea, sleep-related hypoxemia, and other types of sleep-related breathing disorders [1, 33]. SelANet [1] proposed a selective prediction method, using confidence scores from estimated respiratory signals to classify sleep apnea occurrence samples with high confidence and rejecting uncertain samples, achieving notable improvements in

classification performance. Mathew et al. [23] proposed a convolutional neural network based noncontact method to estimate blood oxygen saturation using smartphone cameras without direct contact with individuals, thus minimizing the risk of cross contamination. Res-SE-ConvNet [21] proposed a residual-squeeze-excitation-attention based convolutional network method, designed to determine the severity level of hypoxemia solely using Photoplethysmography (PPG) signals, achieving high accuracy and demonstrating the potential for aiding patients in receiving faster clinical decision support. ZleepNet [3] proposed a deep convolutional neural network model for sleep apnea detection, aiming to reduce the number of sensors required for signal detection. It achieved superior accuracy compared to traditional machine learning methods on publicly available sleep datasets.

3 Datasets

Prog-OCTA The dataset is acquired by ophthalmologists from patients who undergo ophthalmic investigations at enrollment and then every six months during monitoring. Prog-OCTA contains 57 OCTA instances aligned with fine-grained average oxygen saturation (SaO_2) labeled as ground truth. Since the dataset originates from patients with sleep disorders, the distribution of the dataset tends to have a lower average SaO_2, as illustrated in fig. 3.

Table 1: The classification of SaO_2 shown in this table is widely accepted by health-related research and proposed by well-established guidelines [30, 33].

SaO_2 (%)	**Classification**
96 - 100	Normal
93 - 95	Borderline Low
89 - 92	Low

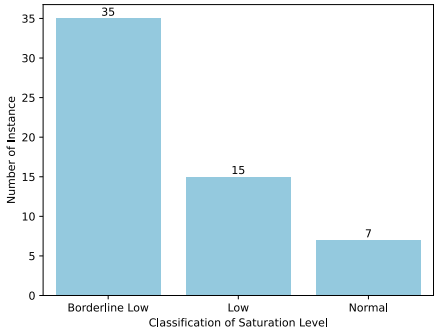
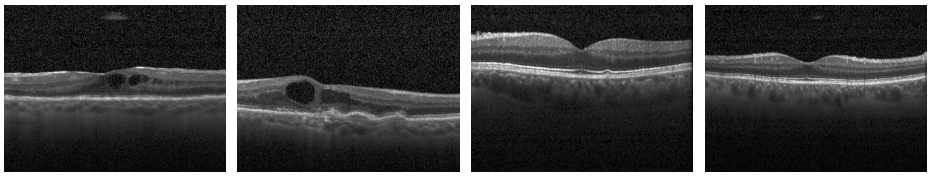


Fig. 4: The figure shows the long-tailed and imbalanced distribution of SaO_2 classes in Prog-OCTA, with the borderline low class being predominant.

However, directly predicting the average SaO_2 is not straightforward, and may not adequately reflect patients' saturation status for sleep disorder diagnosis. Additionally, it may lack representation power and robustness based solely on the instance number among each discrete SaO_2 value. Hence, we convert the discrete SaO_2 values to the oxygen saturation levels according to widely accepted referential standards in table 1 [30, 33].

After refactoring the SaO_2 values into classes according to table 1, we have organized the dataset into three classes. The distribution of each class is depicted in fig. 4. This illustrates that the distribution of SaO_2 is severely long-tailed

and imbalanced. The borderline low categories have significantly more instances than the combination of the other two classes, reflecting the actual SaO_2 status observed in the dataset, and an example for each class is provided in fig. 2.



(a) DME (b) CNV (c) Drusen (d) Normal

Fig. 5: The figures depict four categories of OCT included in Kermany v3 [15]: Diabetic macular edema (DME), characterized by fluid accumulation in the macula due to diabetes; CNV (Choroidal Neovascularization), involving abnormal growth of blood vessels in the retina; Drusen, indicated by the accumulation of deposits comprised of lipids and proteins in the retina; and normal instances.

Kermany v3 The Kermany database v3 [15] is a publicly available dataset comprising 109,312 structural retinal OCT images. It is widely used in OCT-related clinical studies and AI for health research [9, 31, 34]. It comprises four different categories: diabetic macular edema (DME), choroidal neovascularization (CNV), Drusen, and normal instances. The dataset distribution is shown in table 2 and fig. 6, and an example for each class is provided in fig. 5. The Kermany v3 dataset was subsequently utilized for post-training our JointViT model, aimed at enhancing the inherent representation derived from COVID-ViT [8] pretraining.

Table 2: The table displays the number of instances in each class within the training set and testing set of the Kermany database v3 [15].

SETS	**CLASSES**				
	CNV	DME	Drusen	Normal	**Total**
Train	37206	11349	8617	51140	**108312**
Test	250	250	250	250	**1000**

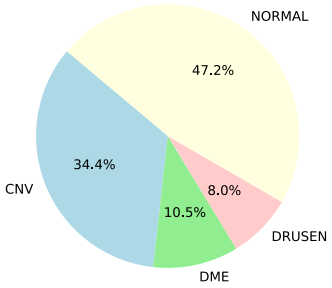


Fig. 6: The pie chart shows the proportion of different classes of OCT in the Kermany database v3 [15].

4 Methodology

Balancing Augmentation To address the challenges posed by the long-tailed distribution and severe class imbalance in the Prog-OCTA dataset and Kermany v3 dataset [15], we propose the integration of ***balancing augmentation*** into the data preprocessing pipeline. Unlike conventional data balancing techniques such as upsampling and downsampling [40], which often introduce biases or compromise critical data information while achieving class balance, balancing augmentation employs ***random cropping***, ***flipping***, and ***rotation*** at arbitrary angles. This approach introduces increased variability into the original dataset during the balancing process without introducing additional biases.

The balancing augmentation technique is implemented on the minority classes of both datasets, augmenting them to match the number of instances in the majority class, as illustrated in fig. 1, as part of the data preprocessing phase. This augmentation process serves to rectify imbalances within the dataset, particularly addressing long-tailed distribution issues from an input perspective. By ensuring a balanced representation of classes, the model's training process becomes more robust and less biased towards the majority class, thus enhancing its ability to generalize effectively across diverse data distributions. Employing this technique ultimately fosters improved model performance and contributes to more equitable and reliable predictions across all classes.

JointViT Backbone We utilize a Vision Transformer (ViT) [6] as our backbone, combined with a linear classification head. The ViT encoder is pretrained on COVID-ViT [8] and further fine-tuned via transfer learning for our downstream task. To augment the learned representation of ViT, we additionally perform post-training on the Kermany v3 OCT dataset [15] as an initialization for our SaO_2 prediction task on Prog-OCTA.

During post-training, the augmented OCT images in the Kermany v3 dataset [15] are divided into k patches of size 16×16 and processed through self-attention layers, incorporating patch and position embeddings. Unlike OCT images, OCTA images in Prog-OCTA are 3D volumes. Therefore, they are initially converted into 2D slices, then they are performed balancing augmentation and divided into patches, which is similar to the post-training process. The reason behind converting 3D OCTA volumes into 2D slices is twofold. Firstly, there is a lack of pretrained models designed specifically for 3D Vision Transformers [41] compared to well-established pretraining models for 2D Vision Transformers [11,26]. Secondly, utilizing raw 3D volumes as input would significantly increase computational complexity and demand greater computational resources compared to using 2D slices.

Unlike the vanilla Vision Transformers, which uses image categories as supervision, we employ both SaO_2 categories Y_{cls} and exact SaO_2 values Y_{val} as joint supervision. For classification, we utilize binary cross-entropy loss L_{bce}, which outperforms ordinary cross-entropy loss when combined with Vision Transformers [35,39]. For regression, we adopt mean squared error loss L_{mse} for supervision. The weight of each loss in the joint loss is controlled by coefficient λ, as shown

in eq. (1). Assuming the model is denoted as T and the input image as X, the joint loss L can be expressed as follows.

$$L = \lambda L_{\text{bce}}(T(X), Y_{\text{cls}}) + (1 - \lambda) L_{\text{mse}}(T(X), Y_{\text{val}}) \tag{1}$$

Additionally, while our MLP head predicts both classes and values for SaO_2, our approach differs from a multitasking model that handles both classification and regression tasks. We choose to focus solely on predicting the SaO_2 classes as the desired output. It is worth noting that accurately predicting precise SaO_2 values with limited data is impractical. Instead, we assign minimal weight to the mean squared error (MSE) loss in the joint loss function to prioritize the performance of classification.

5 Experiments

5.1 Evaluation Matrices

For a fair comparison, we used the test set accuracy to evaluate the overall classification performance of the methods, meanwhile using sensitivity and specificity to assess how the model handles both minority and majority classes in the long-tailed Prog-OCTA dataset.

Table 3: We compared our proposed **JointViT** on Prog-OCTA dataset with well-established 2D and 3D methods which are widely used in medical imaging recognition. The results demonstrate that our method significantly outperforms others in predicting saturation levels using imbalanced OCTA data.

Models	Test Acc. ↑	Sensitivity ↑	Specificity ↑
M3T [14]	$61.40^{\pm 2.48}$	$33.33^{\pm 0.00}$	$66.67^{\pm 0.00}$
I3D [2]	$61.40^{\pm 2.48}$	$33.33^{\pm 0.00}$	$66.67^{\pm 0.00}$
MedViT [22]	$61.40^{\pm 2.48}$	$33.33^{\pm 0.00}$	$66.67^{\pm 0.00}$
FCNlinksCNN [28]	$57.89^{\pm 7.44}$	$31.31^{\pm 2.86}$	$64.96^{\pm 2.42}$
MedicalNet (3D ResNet-10) [4]	$63.16^{\pm 0.00}$	$37.96^{\pm 3.47}$	$68.95^{\pm 1.72}$
MedicalNet (3D ResNet-18) [4]	$63.16^{\pm 4.30}$	$40.74^{\pm 8.59}$	$71.62^{\pm 5.52}$
MedicalNet (3D ResNet-34) [4]	$61.40^{\pm 2.48}$	$33.33^{\pm 0.00}$	$66.67^{\pm 0.00}$
MedicalNet (3D ResNet-50) [4]	$61.40^{\pm 2.48}$	$33.33^{\pm 0.00}$	$66.67^{\pm 0.00}$
MedicalNet (3D ResNet-101) [4]	$61.40^{\pm 2.48}$	$33.33^{\pm 0.00}$	$66.67^{\pm 0.00}$
MedicalNet (3D ResNet-152) [4]	$61.40^{\pm 2.48}$	$33.33^{\pm 0.00}$	$66.67^{\pm 0.00}$
MedicalNet (3D ResNet-200) [4]	$61.40^{\pm 2.48}$	$33.33^{\pm 0.00}$	$66.67^{\pm 0.00}$
3D DenseNet-121 [13]	$66.67^{\pm 2.48}$	$45.37^{\pm 3.46}$	$72.58^{\pm 0.82}$
3D DenseNet-161 [13]	$66.67^{\pm 2.48}$	$47.22^{\pm 2.27}$	$75.06^{\pm 2.14}$
3D DenseNet-169 [13]	$66.67^{\pm 2.48}$	$43.70^{\pm 7.97}$	$72.90^{\pm 2.37}$
3D DenseNet-201 [13]	$64.91^{\pm 4.96}$	$39.81^{\pm 4.72}$	$71.47^{\pm 3.40}$
COVID-ViT [8]	$61.40^{\pm 6.40}$	$36.63^{\pm 7.76}$	$61.66^{\pm 8.13}$
JointViT (Ours)	$\mathbf{70.17}^{\pm 8.90}$	$\mathbf{62.51}^{\pm 8.44}$	$\mathbf{82.83}^{\pm 7.07}$

5.2 Comparative Studies

To thoroughly evaluate our proposed JointViT, we conducted comprehensive experiments comparing it with other well-established methods, as demonstrated in table 3. Inspired by the principle of K-fold cross-validation, we divided the Prog-OCTA dataset randomly into 3 folds. Each time, we used two folds for training and one fold for testing, repeating the experiments three times for each model. This approach ensures a fair and robust evaluation, particularly when dealing with imbalanced and limited data. The experiments are conducted with an optimal λ set to 0.99, and the ablation of different values of λ has been conducted in the following section. The results displayed in table 3 demonstrate that our performance significantly outperforms other methods by up to 12.28% in overall accuracy. This achievement represents the state-of-the-art performance in modeling SaO_2 using long-tailed OCTA.

Given the long-tailed nature of the data, all models exhibit a pattern of higher specificity than sensitivity. This is because they excel at identifying common classes but face challenges in recognizing less frequent ones. However, our approach demonstrates not only higher specificity compared to other methods but, more crucially, significantly higher sensitivity. This indicates that our model not only effectively handles major classes but also possesses superior capability in addressing long-tailed classes compared to compared methods.

5.3 Ablation Studies

We performed experiments by varying the joint loss coefficient λ to examine its influence on both the joint loss and the overall performance of the model. The outcomes are detailed in table 4 and illustrated in fig. 7. Notably, the results indicate that the model achieves optimal performance when λ is set to 0.99.

Table 4: Ablation results in the table shows the impact of varying joint loss coefficient (λ) values on model performance, which indicates when λ is set to be 0.99, the model achieved optimal performance.

Models	Test Acc. ↑	Sensitivity ↑	Specificity ↑
COVID-ViT [8]	$61.40^{\pm 6.40}$	$36.63^{\pm 7.76}$	$61.66^{\pm 8.13}$
JointViT ($\lambda = 0.8$)	$61.40^{\pm 4.71}$	$33.35^{\pm 5.44}$	$67.41^{\pm 7.27}$
JointViT ($\lambda = 0.9$)	$65.72^{\pm 2.72}$	$48.89^{\pm 4.20}$	$66.27^{\pm 3.13}$
JointViT ($\lambda = 0.95$)	$64.91^{\pm 4.21}$	$54.67^{\pm 6.11}$	$70.31^{\pm 3.82}$
JointViT ($\lambda = 0.96$)	$63.61^{\pm 6.67}$	$50.43^{\pm 4.51}$	$76.20^{\pm 2.10}$
JointViT ($\lambda = 0.97$)	$66.67^{\pm 2.46}$	$57.22^{\pm 7.31}$	$64.82^{\pm 6.20}$
JointViT ($\lambda = 0.98$)	$66.67^{\pm 4.50}$	$51.32^{\pm 3.42}$	$70.42^{\pm 6.26}$
JointViT ($\lambda = 1$)	$63.16^{\pm 0.00}$	$35.92^{\pm 3.21}$	$76.81^{\pm 2.29}$
JointViT ($\lambda = 0.99$)	$\mathbf{70.17}^{\pm 8.90}$	$\mathbf{62.51}^{\pm 8.44}$	$\mathbf{82.83}^{\pm 7.07}$

We have additionally integrated ablation analysis concerning the *joint loss* within the proposed JointViT framework. This involved substituting our joint loss with a modified binary cross entropy loss, balanced BCE loss (*Bal-BCE*) [39], which tailors specifically for addressing the challenges inherent in long-tailed image recognition problems. The outcomes of this ablation study are presented in table 5, which shows that our joint loss significantly outperforms well-established Bal-BCE loss designs tailored for long-tailed image recognition scenarios, elucidating that our proposed joint loss formulation stands as the optimal selection for tackling the long-tailed distribution of Prog-OCTA. This underscores the efficacy and superiority of our devised joint loss scheme in addressing the intricacies of long-tailed image recognition tasks, particularly within the OCTA recognition domain. To clarify, in the ablation study of JointViT without the joint loss, only the BCE loss is utilized. In the case of the ablation study on JointViT with Bal-BCE loss, when $\lambda = 1$, the loss exclusively consists of the Bal-BCE loss, and setting $\lambda = 0.99$ substitutes the BCE loss with the Bal-BCE loss in the joint loss function.

Table 5: The table presents the results of conducting ablations involving the integration of different loss functions within the joint loss framework. The results indicate that our jointly designed loss function, which combines BCE and MSE losses, consistently achieves superior performance compared to alternative configurations.

Models	Test Acc. ↑	Sensitivity ↑	Specificity ↑
COVID-ViT [8]	$61.40^{\pm 6.40}$	$36.63^{\pm 7.76}$	$61.66^{\pm 8.13}$
JointViT w/o joint loss	$63.16^{\pm 0.00}$	$35.92^{\pm 3.21}$	$76.81^{\pm 2.29}$
JointViT w/ Bal-BCE ($\lambda = 1$)	$61.40^{\pm 2.48}$	$20.47^{\pm 0.82}$	$53.80^{\pm 0.83}$
JointViT w/ Bal-BCE ($\lambda = 0.99$)	$64.91^{\pm 2.48}$	$45.63^{\pm 17.55}$	$71.08^{\pm 11.96}$
JointViT w/ joint loss ($\lambda = 0.99$)	$\mathbf{70.17}^{\pm 8.90}$	$\mathbf{62.51}^{\pm 8.44}$	$\mathbf{82.83}^{\pm 7.07}$

In our comparative studies, we further explored the efficacy of utilizing the other backbones through ablation analysis, especially compared with the second best backbone in comparative studies. The findings, as depicted in table 6, underscore that our approach leveraging a ViT backbone continues to exhibit superior performance compared to alternative backbone architectures. This reaffirms the efficacy of our method in achieving optimal results across diverse backbone configurations.

Table 6: We further explored the efficacy of employing the second-best backbone in comparative studies. Our findings reveal that our approach which used a ViT backbone maintains its position as the top-performing method.

Models	Test Acc. ↑	Sensitivity ↑	Specificity ↑
2D DenseNet-161 w/ bal-aug & joint loss	$59.65^{\pm 4.96}$	$26.26^{\pm 7.36}$	$60.17^{\pm 8.17}$
3D DenseNet-161 w/ bal-aug & joint loss	$55.09^{\pm 3.46}$	$44.82^{\pm 6.89}$	$68.37^{\pm 10.46}$
JointViT	$\mathbf{70.17}^{\pm 8.90}$	$\mathbf{62.51}^{\pm 8.44}$	$\mathbf{82.83}^{\pm 7.07}$

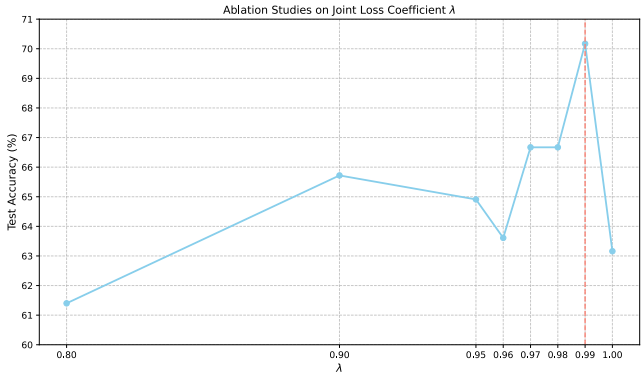


Fig. 7: The figure visualizes the relationship between joint loss coefficient (λ) values and overall model performance, which indicates the optimal λ value is 0.99.

We have additionally assessed the effects of post-training on our training approach. The outcomes illustrated in table 7 indicate that incorporating post-training with OCT images from the Kermany database v3 [15] yields notably superior performance compared to the absence of post-training. This observation underscores the efficacy of employing post-training as a valuable initialization step for subsequent OCTA recognition tasks.

Table 7: The table demonstrates the significant performance improvement achieved by incorporating post-training with OCT images from the Kermany database v3 [15], compared to the absence of post-training, thereby validating its efficacy as an initialization step for downstream OCTA recognition tasks.

Models	Test Acc. ↑	Sensitivity ↑	Specificity ↑
COVID-ViT [8]	$61.40^{\pm 6.40}$	$36.63^{\pm 7.76}$	$61.66^{\pm 8.13}$
JointViT w/o post-training	$57.89^{\pm 0.00}$	$37.45^{\pm 7.76}$	$65.85^{\pm 4.40}$
JointViT w/ post-training	$\mathbf{70.17}^{\pm 8.90}$	$\mathbf{62.51}^{\pm 8.44}$	$\mathbf{82.83}^{\pm 7.07}$

Eventually, we proceeded to incorporate ablations of our data preprocessing technique, specifically focusing on balancing augmentation (bal-aug). The results, as shown in table 8, underscore the efficacy of balancing augmentation. Notably, our model demonstrates remarkable performance improvement with the inclusion of balancing augmentation.

Table 8: The table showcases the efficacy of balancing augmentation in enhancing model performance, with notable improvements observed in various metrics.

Models	Test Acc. ↑	Sensitivity ↑	Specificity ↑
COVID-ViT [8]	$61.40^{\pm 6.40}$	$36.63^{\pm 7.76}$	$61.66^{\pm 8.13}$
JointViT w/o bal-aug	$64.91^{\pm 2.48}$	$45.63^{\pm 17.55}$	$71.08^{\pm 11.96}$
JointViT w/ bal-aug	$\mathbf{70.17}^{\pm 8.90}$	$\mathbf{62.51}^{\pm 8.44}$	$\mathbf{82.83}^{\pm 7.07}$

6 Discussion and Conclusion

Oxygen saturation level (SaO_2) significantly impacts health, particularly indicating conditions like sleep-related hypoxemia, hypoventilation, sleep apnea, and other related disorders. However, continuous monitoring of SaO_2 is time-consuming and subject to variation due to patients' status instability. On the other hand, Optical coherence tomography angiography (OCTA) images excel in speed and effectively screening eye-related lesions, showing promise in assisting with the diagnosis of sleep-related disorders. Therefore, we propose a novel method called JointViT, which incorporates a joint loss utilizing both SaO_2 values and categories for supervision. Additionally, we employ a balancing augmentation technique during data preprocessing to handle the long-tail nature of OCTA datasets. Our method significantly outperformed other established medical imaging recognition methods, paving the way for the future utilization of OCTA in diagnosing sleep-related disorders.

References

1. Bark, B., Nam, B., Kim, I.Y.: Selanet: decision-assisting selective sleep apnea detection based on confidence score. BMC Medical Informatics and Decision Making **23**(1), 190 (2023)
2. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
3. Chaw, H.T., Kamolphiwong, T., Kamolphiwong, S., Tawaranurak, K., Wongtanawijit, R., et al.: Zleepnet: A deep convolutional neural network model for predicting sleep apnea using spo 2 signal. Applied Computational Intelligence and Soft Computing **2023** (2023)
4. Chen, S., Ma, K., Zheng, Y.: Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625 (2019)
5. De Carlo, T.E., Romano, A., Waheed, N.K., Duker, J.S.: A review of optical coherence tomography angiography (octa). International journal of retina and vitreous **1**, 1–15 (2015)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
7. Ferrara, D., Waheed, N.K., Duker, J.S.: Investigating the choriocapillaris and choroidal vasculature with new optical coherence tomography technologies. Progress in retinal and eye research **52**, 130–155 (2016)
8. Gao, X., Khan, M.H.M., Hui, R., Tian, Z., Qian, Y., Gao, A., Baichoo, S.: Covidvit: Classification of covid-19 from 3d ct chest images based on vision transformer model. In: 2022 3rd International Conference on Next Generation Computing Applications (NextComp). pp. 1–4. IEEE (2022)
9. George, N., Shine, L., Ambily, N., Abraham, B., Ramachandran, S.: A two-stage cnn model for the classification and severity analysis of retinal and choroidal diseases in oct images. International Journal of Intelligent Networks **5**, 10–18 (2024)

10. Hafen, B.B., Sharma, S.: Oxygen saturation. In: StatPearls [Internet]. StatPearls Publishing (2022)
11. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
14. Jang, J., Hwang, D.: M3t: three-dimensional medical image classifier using multi-plane and multi-slice transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20718–20729 (2022)
15. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. cell **172**(5), 1122–1131 (2018)
16. Le, P.H., Patel, B.C.: Optical coherence tomography angiography. In: StatPearls [Internet]. StatPearls Publishing (2022)
17. Li, H., Gabb, V., Morrison, H., Blackman, J., Biswas, B., Kendrick, A., Coulthard, E.: Detecting sleep-related breathing disorders using overnight pulse oximetry in patients with dementia and mild cognitive impairment. Alzheimer's & Dementia **19**, e073303 (2023)
18. Liu, S., Hao, J., Xu, Y., Fu, H., Guo, X., Liu, J., Zheng, Y., Liu, Y., Zhang, J., Zhao, Y.: Polar-net: A clinical-friendly model for alzheimer's disease detection in octa images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 607–617. Springer (2023)
19. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2537–2546 (2019)
20. Ma, Y., Hao, H., Xie, J., Fu, H., Zhang, J., Yang, J., Wang, Z., Liu, J., Zheng, Y., Zhao, Y.: Rose: a retinal oct-angiography vessel segmentation dataset and new model. IEEE transactions on medical imaging **40**(3), 928–939 (2020)
21. Mahmud, T.I., Imran, S.A., Shahnaz, C.: Res-se-convnet: A deep neural network for hypoxemia severity prediction for hospital in-patients using photoplethysmograph signal. IEEE Journal of Translational Engineering in Health and Medicine **10**, 1–9 (2022)
22. Manzari, O.N., Ahmadabadi, H., Kashiani, H., Shokouhi, S.B., Ayatollahi, A.: Medvit: a robust vision transformer for generalized medical image classification. Computers in Biology and Medicine **157**, 106791 (2023)
23. Mathew, J., Tian, X., Wong, C.W., Ho, S., Milton, D.K., Wu, M.: Remote blood oxygen estimation from videos using neural networks. IEEE journal of biomedical and health informatics (2023)
24. Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S.: Long-tail learning via logit adjustment. arXiv preprint arXiv:2007.07314 (2020)
25. Okunlola, O.E., Lipnick, M.S., Batchelder, P.B., Bernstein, M., Feiner, J.R., Bickler, P.E.: Pulse oximeter performance, racial inequity, and the work ahead. Respiratory care **67**(2), 252–257 (2022)
26. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning

robust visual features without supervision. Transactions on Machine Learning Research (2023)
27. Orabona, R., Corda, L., Giordani, J., Bernardi, M., Maggi, C., Mazzoni, G., Pedroni, L., Uccelli, S., Zatti, S., Sartori, E., et al.: Sleep-disordered breathing and pregnancy outcomes: The impact of maternal oxygen saturation. International Journal of Gynecology & Obstetrics **164**(1), 140–147 (2024)
28. Qiu, S., Joshi, P.S., Miller, M.I., Xue, C., Zhou, X., Karjadi, C., Chang, G.H., Joshi, A.S., Dwyer, B., Zhu, S., et al.: Development and validation of an interpretable deep learning framework for alzheimer's disease classification. Brain **143**(6), 1920–1933 (2020)
29. Rhodes, C.E., Denault, D., Varacallo, M.: Physiology, oxygen transport. In: StatPearls [Internet]. StatPearls Publishing (2022)
30. Sateia, M.J.: International classification of sleep disorders-: highlights and modifications. Chest **146**(5), 1387–1394 (2014)
31. Song, A., Lusk, J.B., Roh, K.M., Hsu, S.T., Valikodath, N.G., Lad, E.M., Muir, K.W., Engelhard, M.M., Limkakeng, A.T., Izatt, J.A., et al.: Roboctnet: Robotics and deep learning for referable posterior segment pathology detection in an emergency department population. Translational Vision Science & Technology **13**(3), 12–12 (2024)
32. Spaide, R.F., Fujimoto, J.G., Waheed, N.K., Sadda, S.R., Staurenghi, G.: Optical coherence tomography angiography. Progress in retinal and eye research **64**, 1–55 (2018)
33. Summer, J., Singh, A.: What are normal oxygen levels during sleep? Sleep Foundation, `https://www.sleepfoundation.org/physical-health/what-are-normal-oxygen-levels-during-sleep` (2023), [Accessed 20-02-2024]
34. Talcott, K.E., Valentim, C.C., Perkins, S.W., Ren, H., Manivannan, N., Zhang, Q., Bagherinia, H., Lee, G., Yu, S., D'Souza, N., et al.: Automated detection of abnormal optical coherence tomography b-scans using a deep learning artificial intelligence neural network platform. International Ophthalmology Clinics **64**(1), 115–127 (2024)
35. Touvron, H., Cord, M., Jégou, H.: Deit iii: Revenge of the vit. In: European conference on computer vision. pp. 516–533. Springer (2022)
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
37. Venkatesh, R., Pereira, A., Aseem, A., Jain, K., Sangai, S., Shetty, R., Yadav, N.K.: Association between sleep apnea risk score and retinal microvasculature using optical coherence tomography angiography. American journal of ophthalmology **221**, 55–64 (2021)
38. Xie, J., Yi, Q., Wu, Y., Zheng, Y., Liu, Y., Macerollo, A., Fu, H., Xu, Y., Zhang, J., Behera, A., et al.: Deep segmentation of octa for evaluation and association of changes of retinal microvasculature with alzheimer's disease and mild cognitive impairment. British Journal of Ophthalmology **108**(3), 432–439 (2024)
39. Xu, Z., Liu, R., Yang, S., Chai, Z., Yuan, C.: Learning imbalanced data with vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15793–15803 (2023)
40. Zhang, Z., Ahmed, K.A., Hasan, M.R., Gedeon, T., Hossain, M.Z.: A deep learning approach to diabetes diagnosis. arXiv preprint arXiv:2403.07483 (2024)
41. Zhu, Z., Ma, X., Chen, Y., Deng, Z., Huang, S., Li, Q.: 3d-vista: Pre-trained transformer for 3d vision and text alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2911–2921 (2023)