Enhancing Early Detection of Tractional Retinal Lesions in OCT via Self-Supervised Learning

Wei Zhang[1,*], Yu Lu[1,*], Han Jiang[2,*], Xiujuan Zhao[3], Qianying Liu[4], Bingding Huang[1], Zhaoshun Zhang[2], Zhicheng Dong[2], Liyilei Su[1]

[1] School of Artificial Intelligence, Shenzhen Technology University, Shenzhen, China
[2] School of Information Science and Technology, Tibet University, Lhasa, China
[3] Guangdong Provincial Clinical Research Center for Ocular Diseases, Sun Yat-sen University, Guangzhou, China
[4] School of Computing Science, University of Glasgow, Glasgow, United Kingdom
[*] Corresponding authors. Email: {zhangwei1,lvyu}@sztu.edu.cn, jiangh1003@163.com

2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3315-1557-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/BIBM66473.2025.11356400

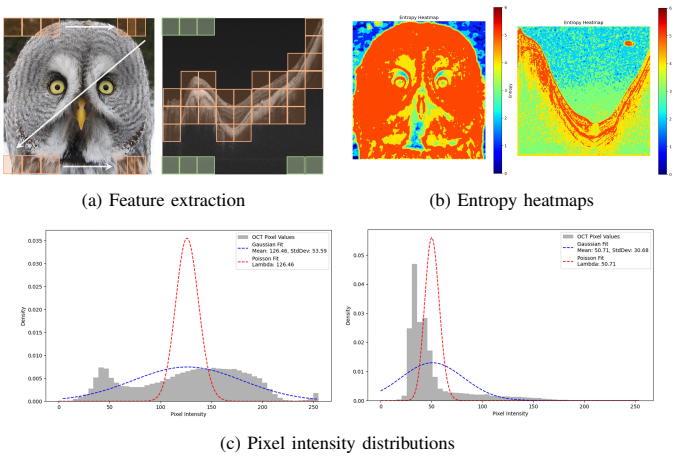
***Abstract*—Optical Coherence Tomography (OCT) plays a vital role in the early detection and monitoring of tractional retinal lesions (TRL), providing high-resolution visualization of retinal structures. However, automated TRL diagnosis remains challenging due to complex lesion morphology, large low-entropy background regions, and the scarcity of high-quality labeled data. Existing Self-Supervised Learning (SSL) approaches often treat all image patches equally, making them sensitive to background noise and limiting their ability to capture fine-grained lesion features. To address these issues, we propose Clustering Heterogeneous Masked Image Modeling (CH-MIM), a novel SSL framework tailored for OCT-based TRL analysis. Our method leverages a large-scale clinical dataset containing 11,861 OCT scans collected over five years, including 3,950 expert-annotated images across six TRL severity levels (T0–T5). CH-MIM introduces a Weighted Feature Space Clustering (WFSC) module to selectively mask high-entropy regions, effectively filtering out irrelevant background information. A heterogeneous progressive masking strategy combines binary, Gaussian, and Poisson noise masks to provide diverse, informative reconstruction tasks. Furthermore, a Consistency Regularization Module (CRM) enforces stable predictions across masking branches, improving representation robustness and transferability to downstream classification. Extensive experiments demonstrate that CH-MIM achieves a top-1 accuracy of 97.7% and top-5 accuracy of 99.8%, surpassing state-of-the-art supervised and self-supervised baselines. These results highlight the potential of CH-MIM as an effective pretraining strategy for automated TRL screening and its applicability to broader OCT-based retinal disease diagnosis.**

***Index Terms*—Self-Supervised Learning, Optical Coherence Tomography, Tractional Retinal Lesions (TRL), Feature Clustering, Consistency Regularization.**

I. Introduction

Optical Coherence Tomography (OCT) uses interferometric light reflection to provide high-resolution cross-sectional retinal images [1], [2]. It is indispensable for analyzing retinal layer structures and macular morphology, offering critical insights into tractional retinal lesions (TRL) [3] and supporting clinical decisions.

This research is supported by the National Natural Science Foundation of China (Grant No. 62266039), the Key Project of Guangdong Province General Higher Education Institutions in Specific Research Areas (Grant No. 2023ZDZX2055), and the Open Project of National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen 518060, P.R. China (Grant No. SZU-BDSC-OF2024-11).



(a) Feature extraction (b) Entropy heatmaps

(c) Pixel intensity distributions

Fig. 1: Comparison between natural and OCT images in terms of (a) feature extraction, (b) entropy distribution, and (c) pixel intensity characteristics.

Supervised deep learning has achieved remarkable progress in medical image analysis [4]–[8]. Multi-task frameworks perform joint lesion grading and segmentation [9], while transformer-based relation networks improve lesion localization [10]. Fusion models further enhance OCT disease detection [11] and pathological myopia diagnosis [12]. However, these models depend on large annotated datasets, which are costly to obtain.

Self-supervised learning (SSL) leverages abundant unlabeled OCT data for representation learning. Metadata-enhanced contrastive learning [13], masked autoencoding [14], and masked image modeling [15] have achieved promising results. Core ideas such as attention [16] and convolutional backbones [17], [18] continue to inspire SSL methods. For OCT, Theodoros Pissas et al. [19] extended Masked Autoencoders (MAE) using multimodal data, while Fatema-E Jannat et al. [20] employed SwinV2-based MAE to enhance generalization. Yukun Zhou et al. [21] proposed RETFound, bridging self-supervised pretraining and downstream OCT diagnosis, outperforming prior MAE-based methods.

However, as shown in Fig. 1, OCT scans contain extensive low-entropy regions (highlighted in green in Fig. 1(a) and

Authorized licensed use limited to: Shenzhen Technology University. Downloaded on February 09,2026 at 02:18:41 UTC from IEEE Xplore. Restrictions apply.

cyan in Fig. 1(b)) that weakly correlate with retinal structures. Existing patch-wise methods [19] (arrows in Fig. 1(a)) treat all patches equally, increasing computation and causing non-zero gradient projections in both lesion and background subspaces. During pre-activation, the mixture of key and background features hampers representation learning. Moreover, OCT pixel intensities often follow a Poisson rather than Gaussian distribution (see Fig. 1(c)), reflecting OCT's photon-counting characteristics. Thus, pretext tasks must account for both structural and statistical disparities.

To address these limitations, we propose CH-MIM (Clustering Heterogeneous Masked Image Modeling), a self-supervised framework for TRL detection. It introduces three key contributions. (1) A clustering-based masking module automatically identifies high-entropy feature regions using KMeans [22]. (2) A heterogeneous progressive masking strategy adapts to OCT's intensity distribution, generating stable gradient signals and diverse masking patterns beyond binary masking [21]. (3) A consistency regularization module enforces prediction stability across progressive masking stages.

We also construct the largest OCT dataset for pathological myopia, containing 11,861 scans from the Zhongshan Ophthalmic Center (7,911 unlabeled and 3,950 expert-annotated). It is the first to cover six TRL grades (T0–T5), reflecting real-world imbalance and clinical difficulty in differentiating adjacent stages.

The main contributions are summarized as follows:

- **CH-MIM:** a self-supervised framework for TRL detection, integrating clustering-based masking and heterogeneous progressive masking for effective feature separation.
- **Consistency regularization:** enforcing prediction stability across masking stages to improve robustness and generalization.
- **Comprehensive dataset:** a large-scale OCT dataset with six-grade TRL annotations (T0–T5), supporting pretraining and evaluation of self-supervised retinal models.

II. Method

A. Framework for Self-Supervised Lesion Classification

In the self-supervised learning task for tractional retinal lesions, key challenges include effective feature extraction, model learning enhancement, and feature consistency across different masking strategies. To tackle these challenges, this paper proposes CH-MIM, a self-supervised neural network architecture that enhances feature learning through three key modules: the Weighted Feature Space Clustering Neural Network (WFSC), the Heterogeneous Progressive Masking Mechanism (HPM), and the Consistency Regularization Module Mechanism (CRM). WFSC improves feature learning by selecting high-entropy lesion regions and filtering out the background. HPM integrates multiple masking strategies to progressively increase task complexity, while CRM enforces prediction consistency, enhancing stability and generalization. Together, these modules significantly improve the accuracy and robustness of lesion classification.

B. Weighted Feature Space Clustering Neural Network (WFSC)

To address low pixel intensity and background interference in OCT images, we propose the WFSC framework. WFSC consists of two key steps: feature weight calculation and feature space clustering. First, the Spatial and Channel Shortcut Attention Module (SCSAM) assigns dynamic importance weights to image blocks, using Channel Attention (CA) for inter-channel weighting and Spatial Attention (SA) for spatial relationships. Then, feature space clustering optimizes feature distribution using a weighted Euclidean distance metric, increasing lesion area clustering density while reducing background interference. This strategy enhances feature representation, reduces computational complexity, and optimizes OCT image reconstruction. The feature map $F \in \mathbb{R}^{C\times H\times W}$ input to SCSAM is processed as follows:

$$\begin{aligned} F_1 &= \mathrm{CA}(F) \otimes F + F, \\ F_2 &= \mathrm{SA}(F_1) \otimes F_1 + F_1, \end{aligned} \tag{1}$$

where $\otimes$ represents the element-wise multiplication of the two attention modules with the feature map. Specifically, each patch $F_i \in \mathbb{R}^{C\times H\times W}$ from the feature map is first passed through the Channel Attention module, which captures the global inter-channel dependencies via global statistical pooling. Then, the Multi-Layer Perceptron (MLP) generates adaptive weights, combined with a residual structure to retain the original information.

Through SCSAM, the model dynamically adjusts its focus based on the structural properties of the input data. The generated weights provide important structured priors for the clustering process in the high-dimensional feature space.

C. Heterogeneous Progressive Masking Mechanism (HPM)

The traditional 0/1 hard masking strategy often leads to the loss of critical information in OCT images, especially in fine structures such as blood vessels and the optic disc. To address this, we propose the HPM strategy, which gradually increases masking difficulty to improve lesion region learning. HPM consists of three stages: Stage 1 employs 0/1 hard masking to strengthen learning; Stage 2 applies Gaussian noise masking to capture global features while preserving local structures; Stage 3 uses Poisson noise masking to simulate light intensity variations, focusing on lesion brightness changes. This progressive approach enhances feature extraction at multiple levels, significantly improving the model's ability to classify lesion regions in OCT images.

D. Consistency Regularization Module (CRM)

In self-supervised learning, the lack of labeled data can introduce feature selection bias, particularly with heterogeneous progressive masking strategies, leading to inconsistent predictions and reduced model performance. To address this, we propose the CRM, which minimizes prediction discrepancies across different branches by aligning their outputs. This ensures robust feature learning and classification, even under varying masking and noise conditions. By constraining

Authorized licensed use limited to: Shenzhen Technology University. Downloaded on February 09,2026 at 02:18:41 UTC from IEEE Xplore. Restrictions apply.

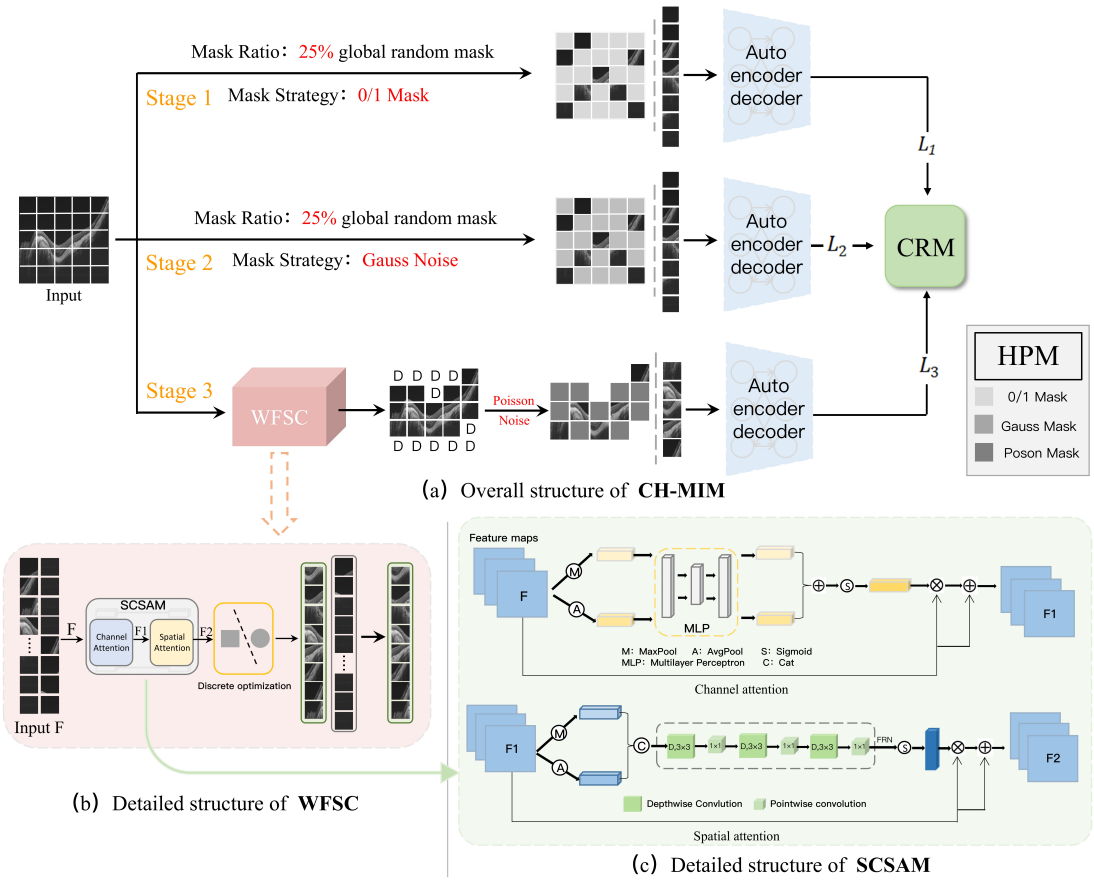


Fig. 2: Overview of the proposed CH-MIM framework for OCT-based tractional retinal lesion (TRL) analysis. (a) Overall architecture with three heterogeneous masking stages (0/1 Mask, Gaussian Noise Mask, and WFSC with Poisson Noise) followed by autoencoder-decoders, aggregated by the CRM module. (b) Structure of WFSC using SCSAM to select high-entropy patches. (c) Internal design of SCSAM combining channel and spatial attention for refined feature selection.

predictions through a consistency loss function, CRM prevents information loss, enhancing the model's stability, robustness, and accuracy in complex tasks.

$$\begin{aligned} L_{CRM} = & \frac{1}{N}\sum_{i=1}^{N}\left(pred_{1,i} - pred_{2,i}\right)^2 \\ & + \frac{1}{N}\sum_{i=1}^{N}\left(pred_{1,i} - pred_{3,i}\right)^2 \\ & + \frac{1}{N}\sum_{i=1}^{N}\left(pred_{2,i} - pred_{3,i}\right)^2 \end{aligned} \tag{2}$$

where $pred_{a_i}$ is the i-th prediction on the a-th branch.

III. Experiments

A. Experimental Settings

1) Dataset: In this study, we have systematically collected data over the past five years. The dataset primarily comprises 11,861 high-quality OCT images of pathological myopia, which can be further classified into three categories: Atrophic lesions (A), Tractional lesions (T), and Neovascular lesions (N). The main objective of our experiments is to conduct an in-depth six-class diagnostic analysis of Tractional lesions (T).

2) Evaluation Metrics: To comprehensively evaluate the classification performance on the six-class pathological myopia traction dataset, we adopt six commonly used metrics: maximum accuracy (**Max Acc**), Top-1 accuracy, Top-5 accuracy, precision (**P**), recall (**R**), and F1-score (**F1**).

Let TP, FP, FN, and TN denote true positives, false positives, false negatives, and true negatives, respectively. Given N samples and y_i, $\hat{y}_i$ representing the ground truth and predicted label for the i-th sample, the metrics are defined as follows:

a) Accuracy.:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Max Acc represents the highest accuracy achieved during training, while Top-1 and Top-5 correspond to the proportion of test samples for which the correct label is ranked first or within the top five predictions, respectively:

$$\text{Top-}k = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\left(y_i \in \text{Top-}k(\hat{y}_i)\right), \quad k \in \{1, 5\} \tag{4}$$

Authorized licensed use limited to: Shenzhen Technology University. Downloaded on February 09,2026 at 02:18:41 UTC from IEEE Xplore. Restrictions apply.

b) Precision.:

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

c) Recall.:

$$R = \frac{TP}{TP + FN} \tag{6}$$

d) F1-score.:

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{7}$$

All metrics are reported in percentage (%) form. The proposed CH-MIM framework is compared against existing supervised and self-supervised methods using these metrics, as shown in Table I.

3) Implementation Details: All algorithms are implemented in PyTorch and run on a platform with four NVIDIA 4090 GPUs. The dataset includes 11,861 TRL images, with 7,911 unlabeled images for self-supervised pretraining and 3,950 labeled images categorized by pathological grading (T0-T5) for supervised fine-tuning. Labeled images are split into a training set (3,075 images) and a test set (875 images) in a 77.8 percent:22.2 percent ratio, with five-fold cross-validation for stability and generalization. Images are resized to 224×224 pixels and augmented using random cropping, horizontal flipping, and color jittering for enhanced data diversity.

During pretraining, the model utilizes the CH-MIM backbone, dividing images into non-overlapping 14×14 patches, each 16×16 pixels. The encoder consists of 24 Transformer layers, while the decoder has 8 layers. The model is trained for 2000 epochs to extract general features. In the fine-tuning phase, the AdamW optimizer is employed with an initial learning rate of 0.001, a layer decay of 0.75, and 5 epochs of warm-up. Training lasts for 300 epochs with early stopping, monitored by the validation set, ensuring efficient and robust training.

B. Main Results

TABLE I: Performance comparison on pathological myopia traction six-class classification (%).

Model	Pretrain Dataset	Max Acc (↑)	Top-1 (↑)	Top-5 (↑)	Precision (↑)	Recall (↑)	F1 (↑)
ResNet [23]	ImageNet-1K [24]	92.2	91.9	96.0	88.5	87.5	87.8
DenseNet [25]	ImageNet-1K	92.9	92.6	96.7	88.1	89.0	88.4
ViT [23]	ImageNet-1K	92.7	92.5	96.5	90.0	86.8	88.0
SwinT [26]	ImageNet-1K	92.4	92.1	96.2	92.7	88.7	90.5
Conformer [27]	ImageNet-1K	91.8	91.7	95.8	89.6	87.4	87.6
MAE [14]	ImageNet-1K	93.6	93.6	97.1	93.1	93.0	93.0
	OCT^{TRL}	94.4	94.2	97.8	93.6	93.4	93.5
CAE [28]	ImageNet-1K	93.8	94.0	97.5	93.3	93.2	93.3
	OCT^{TRL}	94.7	94.5	98.0	94.0	94.1	94.1
SimMIM [15]	ImageNet-1K	93.7	93.5	97.3	93.3	93.3	93.1
	OCT^{TRL}	94.6	94.5	97.9	93.9	93.7	93.7
RETFound [21]	ImageNet-1K	93.6	93.6	97.1	93.1	93.0	93.0
	OCT^{TRL}	94.4	94.2	97.8	93.6	93.4	93.5
OCT-SelfNet [20]	ImageNet-1K	92.4	92.1	96.2	92.7	88.7	90.5
	OCT^{TRL}	93.2	93.5	97.1	93.2	90.2	92.3
0/1 Mask	OCT^{TRL}	95.4	95.2	98.5	95.4	95.4	95.3
Gauss Noise	OCT^{TRL}	96.1	95.8	99.2	96.0	95.9	96.0
WFSC+Poisson Noise	OCT^{TRL}	95.6	95.4	98.9	95.3	95.4	95.6
CH-MIM (Ours)	OCT^{TRL}	**97.9**	**97.7**	**99.8**	**97.1**	**97.0**	**97.0**

As shown in Table I, the CH-MIM model excels in unsupervised traction lesion diagnosis with a classification accuracy of 97.9 percent, a Top-1 accuracy of 97.7 percent, and a Top-5 accuracy of 99.8 percent, outperforming DenseNet and Transformer models like ViT and SwinT. OCT-based pretraining enhances MAE, CAE, and SimMIM Top-1 accuracy by 0.6 percent, 0.5 percent, and 1.0 percent, respectively, but still falls short of CH-MIM. Meanwhile, we conducted a performance comparison between CH-MIM, RETFound [21] with standard MAE for transfer tasks, and OCT-SelfNet, where ViT in MAE is replaced with SwinT. As observed, CH-MIM improved the performance by 3.5 and 4.7, respectively, further validating the necessity of designing pre-trained models for OCT retinal disease tasks as proposed in this paper.

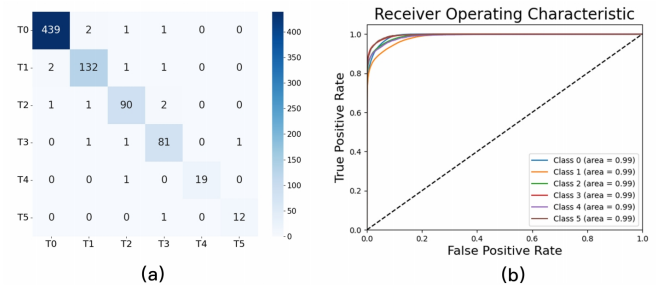


Fig. 3: (a) Confusion matrix for the tractional six-class diagnostic task; (b) ROC curves for the corresponding results.

Optimized WFSC feature extraction improves CH-MIM's Top-1 accuracy by 3.2 and 3.5 percentage points over CAE and MAE, respectively. As shown in Fig. 3, the confusion matrix and ROC analysis indicate near-perfect AUC values, while slight confusion between T2 and T3 highlights room for finer-grained feature learning and class differentiation.

Qualitative visualization. Fig. 4 presents heatmaps for traction levels T0–T5: (a) original B-scans; (b,c) SCSAM at native and 224×224; (d,e) CBAM at the same resolutions. SCSAM focuses on anatomically meaningful regions (foveal pit and layer boundaries) while suppressing background. After 224×224 patching, grid artifacts appear, yet SCSAM keeps compact, contour-aligned activations; CBAM is more diffuse and background-sensitive, especially for T3–T4. Across grades, SCSAM saliency scales with severity and yields clearer inter-grade separation, whereas CBAM shows greater overlap—supporting SCSAM as the more stable, discriminative masker.

C. Ablation Study

1) Parallel Networks Performance Validation: As shown in Table II, the effectiveness of the parallel network architecture with the CRM module, have been validated. The Stage 3 branch performs best in the six-class classification task, highlighting the advantage of WFSC guidance. Applying the pretraining weights from the three-branch model to a single-branch model results in decreased performance, suggesting that the three-branch structure is better suited for multi-channel information fusion.

2) SCSAM Performance Validation: As shown in Table III, SCSAM masking strategy significantly outperforms global random masking, improving Top-1 accuracy by 1.9% and

Authorized licensed use limited to: Shenzhen Technology University. Downloaded on February 09,2026 at 02:18:41 UTC from IEEE Xplore. Restrictions apply.

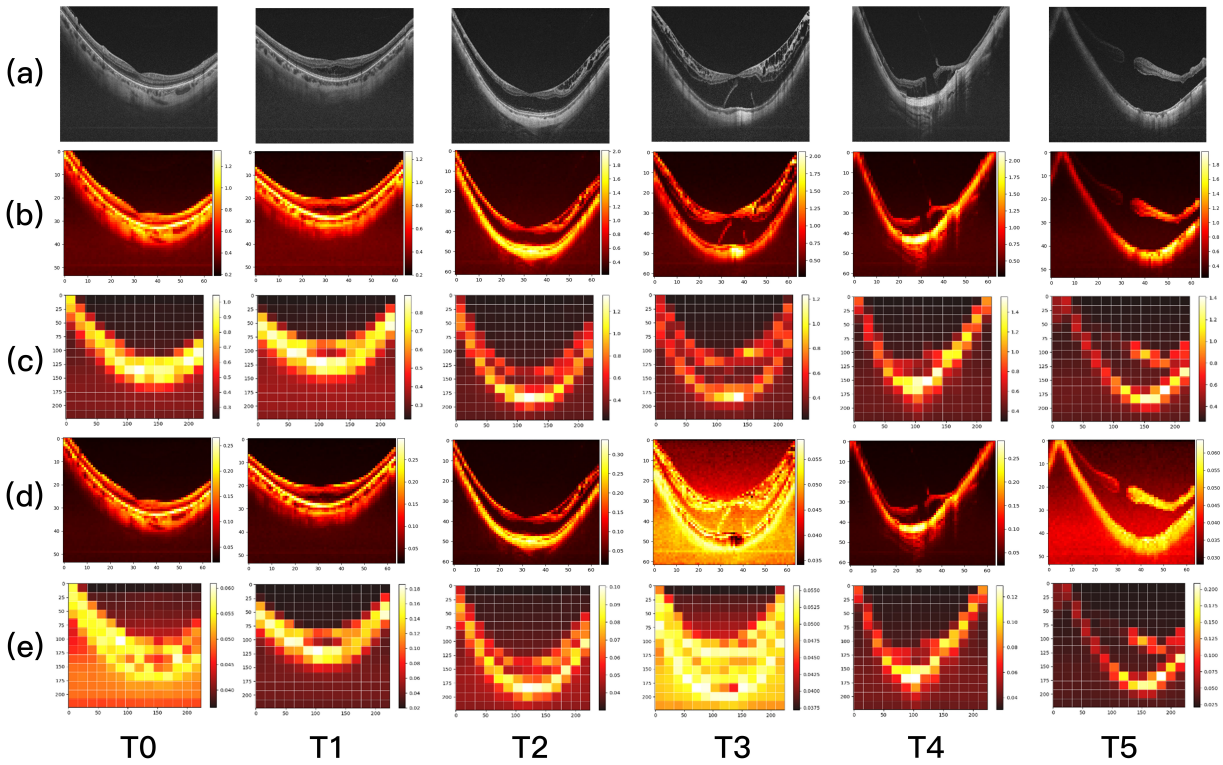


Fig. 4: Heatmap visualization across tractional levels T0–T5 using auxiliary maskers: (a) original OCT B-scans; (b,c) SCSAM at native and 224 × 224 resolutions; (d,e) CBAM at the same resolutions. Patch partitioning at 224 × 224 simulates the model input and clustering analysis.

TABLE II: Parallel network performance validation (%).

Stage	Acc (↑)	Top-1 (↑)	Top-5 (↑)	Pre (↑)	Recall (↑)	F1 (↑)
Stage1	96.9	95.8	99.1	95.8	95.8	95.6
Stage2	97.3	96.9	99.5	96.8	96.7	96.8
Stage3	**97.9**	**97.7**	**99.8**	**97.1**	**97.0**	**97.0**

TABLE III: SCSAM performance validation (%).

Auxiliary Masker	Acc (↑)	Top-1 (↑)	Top-5 (↑)	Pre (↑)	Recall (↑)	F1 (↑)
LinearAttention	95.4	94.2	98.2	94.7	94.5	94.6
Self-Attention [16]	96.2	95.3	98.6	95.4	95.3	95.3
CBAM [29]	97.1	96.8	99.3	96.5	96.6	96.5
SCSAM (Ours)	**97.9**	**97.7**	**99.8**	**97.1**	**97.0**	**97.0**

enhancing precision and recall. Further experiments confirm the importance of SCSAM in feature extraction. In self-supervised tasks, SCSAM achieves a Top-1 accuracy of 97.7%, significantly outperforming traditional attention mechanisms. It effectively reduces background noise and improves the model's ability to recognize foreground regions.

3) WFSC Performance Validation: As shown in Table IV, we use the global random mask (removing Cluster and SCSAM) as the baseline and gradually add the modules in WFSC. Although adding only SCSAM for dynamic ranking and manually setting the threshold improved performance by 0.7%, it struggled to adapt to different OCT images. After introducing Cluster to form the complete WFSC, performance was significantly improved by 1.3%, enabling more accurate lesion area extraction.

TABLE IV: WFSC performance validation (%).

Cluster	SCSAM	Acc (↑)	Top-1 (↑)	Top-5 (↑)	Pre (↑)	Recall (↑)	F1 (↑)
×	×	96.0	95.7	98.6	95.2	95.1	95.0
×	✓	96.6	96.4	99.4	95.8	95.8	95.7
✓	✓	**97.9**	**97.7**	**99.8**	**97.1**	**97.0**	**97.0**

IV. Discussion and Limitations

The proposed CH-MIM framework demonstrates strong potential for improving self-supervised OCT analysis and enabling early detection of tractional retinal lesions (TRL). By integrating weighted feature space clustering and progressive masking strategies, the model achieves superior accuracy over existing baselines. However, several limitations remain. First, the dataset used in this study, while large, is still limited in diversity and suffers from class imbalance, particularly for severe TRL grades, which may affect generalizability to underrepresented clinical cases. Second, CH-MIM is designed and evaluated primarily on single-modality 2D B-scan OCT images, whereas real-world TRL diagnosis often benefits from multimodal data sources, such as fundus photography or OCT angiography. Third, despite using self-supervised pretraining to reduce labeled data requirements, the multi-branch architecture and progressive masking mechanism increase computational cost, potentially limiting deployment in real-time or resource-constrained settings.

Future work will focus on expanding the dataset with more diverse and balanced samples, integrating multimodal

Authorized licensed use limited to: Shenzhen Technology University. Downloaded on February 09,2026 at 02:18:41 UTC from IEEE Xplore. Restrictions apply.

imaging modalities for more comprehensive feature representation, and exploring lightweight model adaptations to enhance computational efficiency. These improvements aim to further advance the clinical applicability of CH-MIM for automated and scalable retinal disease screening.

V. Conclusions

This study develops an OCT-based retinal image dataset for TRL and introduces the CH-MIM self-supervised pretraining framework for TRL diagnosis. The framework integrates WFSC, HPM, and CRM. WFSC enhances feature learning by selecting high-entropy lesion regions and filtering out the background, while HPM progressively increases task complexity. CRM improves model stability and generalization by enforcing prediction consistency. These strategies significantly enhance the model's ability to extract fine-grained features, particularly under low signal-to-noise ratios and non-uniform imaging conditions. Experimental results demonstrate that CH-MIM performs excellently in TRL diagnosis, providing a novel approach for the application of self-supervised learning in OCT medical imaging, especially in early disease detection. Future research will focus on optimizing the model's multimodal adaptability and computational efficiency, expanding its application to a broader range of medical imaging tasks.

References

[1] B. E. Bouma, J. F. de Boer, D. Huang, I.-K. Jang, T. Yonetsu, C. L. Leggett, R. Leitgeb, D. D. Sampson, M. Suter, B. J. Vakoc *et al.*, "Optical coherence tomography," *Nature Reviews Methods Primers*, vol. 2, no. 1, p. 79, 2022.

[2] P. E. Stanga, S. Pastor-Idoate, U. Reinstein, P. Vatas, U. Patel, S. Dubovy, D. Z. Reinstein, and O. Zahavi, "Navigated single-capture 3D and cross-sectional wide-field OCT of the mid and peripheral retina and vitreoretinal interface," *European Journal of Ophthalmology*, vol. 32, no. 3, pp. 1642–1651, 2022.

[3] D. Thomas and G. Duguid, "Optical coherence tomography—a review of the principles and contemporary uses in retinal investigation," *Eye*, vol. 18, no. 6, pp. 561–570, 2004.

[4] Y. Lu, Z. Xu, K. L. Yung, and W. H. Ip, "MSL-Net: A lightweight self-supervised multi-task framework for IoMT-enabled medical image segmentation and landmark localization," *Internet of Things*, p. 101709, 2025.

[5] Z. Xu, Y. Lu, W. Zhang, X. Li, S. Shi, and X. Fu, "Multi-Task Self-Supervised Learning for Automated Measurement of Left Ventricular Ejection Fraction in Echocardiography," in *Proceedings of the 2025 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2025, pp. 1–6.

[6] H. Liang, L. Li, Y. Lu, Q. Liu, H. Ge, and X. Fu, "LAGNet: Label Attention Graph Networks for Ocular Disease Classification Using Fundus Images," in *Proceedings of the 2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2024, pp. 1–5.

[7] Y. Li, C. Yang, H. Zeng, Z. Dong, Z. An, Y. Xu, Y. Tian, and H. Wu, "Frequency-Aligned Knowledge Distillation for Lightweight Spatiotemporal Forecasting," *arXiv preprint arXiv:2507.02939*, 2025.

[8] H. Zeng, Y. Li, R. Niu, C. Yang, and S. Wen, "Enhancing spatiotemporal prediction through the integration of mamba state space models and diffusion transformers," *Knowledge-Based Systems*, vol. 316, p. 113347, 2025.

[9] A. Foo, W. Hsu, M. L. Lee, G. Lim, and T. Y. Wong, "Multi-task learning for diabetic retinopathy grading and lesion segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 34, no. 08, 2020, pp. 13 267–13 272.

[10] S. Huang, J. Li, Y. Xiao, N. Shen, and T. Xu, "RTNet: Relation Transformer Network for Diabetic Retinopathy Multi-Lesion Segmentation," *IEEE Transactions on Medical Imaging*, vol. 41, no. 6, pp. 1596–1607, 2022.

[11] Z. Ai, X. Huang, J. Feng, H. Wang, Y. Tao, F. Zeng, and Y. Lu, "FN-OCT: Disease Detection Algorithm for Retinal Optical Coherence Tomography Based on a Fusion Network," *Frontiers in Neuroinformatics*, vol. 16, p. 876927, 2022.

[12] S. Chen, Z. Wu, M. Li, and Y. Zhu, "Fit-net: Feature interaction transformer network for pathologic myopia diagnosis," *IEEE Transactions on Medical Imaging*, vol. 42, no. 11, pp. 3021–3032, 2023.

[13] R. Holland, O. Leingang, H. Bogunović, S. Riedl, L. Fritsche, T. Prevost, H. P. Scholl, U. Schmidt-Erfurth, S. Sivaprasad, A. J. Lotery *et al.*, "Metadata-enhanced contrastive learning from retinal optical coherence tomography images," *Medical Image Analysis*, vol. 97, p. 103296, 2024.

[14] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16 000–16 009.

[15] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "SimMIM: A Simple Framework for Masked Image Modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 9653–9663.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Łukasz Kaiser, and I. Polosukhin, "Attention is All you Need," *Advances in Neural Information Processing Systems (NIPS 2017)*, vol. 30, 2017.

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper With Convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[18] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv preprint arXiv:1704.04861*, 2017.

[19] W. Su, P. Miao, H. Dou, and X. Li, "ScanFormer: Referring Expression Comprehension by Iteratively Scanning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 13 449–13 458.

[20] F.-E. Jannat, S. Gholami, M. N. Alam, and H. Tabkhi, "OCT-SelfNet: A Self-Supervised Framework with Multi-Modal Datasets for Generalized and Robust Retinal Disease Detection," *arXiv preprint arXiv:2401.12344*, 2024.

[21] Y. Zhou, M. A. Chia, S. K. Wagner, M. S. Ayhan, D. J. Williamson, R. R. Struyven, T. Liu, M. Xu, M. G. Lozano, P. Woodward-Court *et al.*, "A foundation model for generalizable disease detection from retinal images," *Nature*, vol. 622, no. 7981, pp. 156–163, 2023.

[22] J. B. McQueen, "Some methods of classification and analysis of multivariate observations," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 248–255.

[25] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.

[26] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10 012–10 022.

[27] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, and Q. Ye, "Conformer: Local Features Coupling Global Representations for Visual Recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 367–376.

[28] X. Chen, M. Ding, X. Wang, Y. Xin, S. Mo, Y. Wang, S. Han, P. Luo, G. Zeng, and J. Wang, "Context Autoencoder for Self-supervised Representation Learning," *International Journal of Computer Vision*, vol. 132, no. 1, pp. 208–223, 2024.

[29] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

Authorized licensed use limited to: Shenzhen Technology University. Downloaded on February 09,2026 at 02:18:41 UTC from IEEE Xplore. Restrictions apply.