This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. 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/JBHI.2026.3718754

Automatic Classification of Retinal OCT Images Based on Multi-Perspective Collaborative Self-Distillation and Category-Aware Contrastive Learning

Wendi Zhang, Ying Fan, Siheng Zhou, Fei Shi, Xinjian Chen, *Senior Member*, *IEEE*, Xun Xu, Dehui Xiang, Tao Peng, Weifang Zhu

***Abstract*—The automatic classification of retinal optical coherence tomography (OCT) images holds significant value for assisting in the diagnosis of related fundus diseases. However, it often faces challenges such as class imbalance and high inter-class similarity. In this paper, we propose a multi-perspective collaborative self-distillation network (MCSD-Net) for OCT image classification, which enables the model to learn through multi-perspective collaborative self-distillation mechanism and category-aware contrastive learning strategy. Specifically, a multi-perspective collaborative self-distillation mechanism is designed, which comprises structural distillation based on multi-stage feature fusion and historical distillation based on linear growth strategy, enhancing knowledge diversity from both structural and historical perspectives. A category-aware contrastive learning (CACL) strategy is designed, which constructs class-balanced batches to alleviate class imbalance and employs contrastive learning to strengthen the representation of intra-class and inter-class features. Additionally, a direction-aware attention module (DAAM) is incorporated to extract features from horizontal and vertical directions, thereby improving feature discriminability. The proposed MCSD-Net is evaluated on a private myopic tractional maculopathy (MTM) retinal OCT dataset and two public OCT datasets (OCT2017 and OCTDL). Experimental results demonstrate that the proposed MCSD-Net outperforms other state-of-the-art self-distillation based methods.**

***Index Terms*—Medical image classification, self-knowledge distillation, contrastive learning, attention mechanism, retinal OCT image**

This work was supported in part by the National Natural Science Foundation of China (62371326, 62271337, U20A20170), in part by Natural Science Foundation of Jiangsu Province (BK20231310), in part by the National Key Research and Development Program of China (2018YFA0701700), and in part by Priority Academic Program Development of Jiangsu Higher Education Institutions (YX11900123) (Wendi Zhang and Ying Fan contributed equally to this work.) (Corresponding author: Weifang Zhu.)

Wendi Zhang, Fei Shi, Xinjian Chen, Dehui Xiang and Weifang Zhu are with School of Electronic and Information Engineering, Soochow University, Suzhou 215006, China (email: wtzhangwd@stu.suda.edu.cn; shifei@suda.edu.cn; xjchen@suda.edu.cn; xiangdehui@suda.edu.cn; wfzhu@suda.edu.cn).

Ying Fan and Xun Xu are with Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, and National Clinical Research Center for Eye Diseases, Shanghai 200080, China (e-mail: mdfanying@sjtu.edu.cn, drxuxun@sjtu.edu.cn).

Siheng Zhou is with Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200080, China (sihengzhi@163.com).

Tao Peng is with School of Future Science and Engineering, Soochow University (e-mail: tpeng@suda.edu.cn).

I. Introduction

RETINAL diseases such as diabetic retinopathy (DR), age-related macular degeneration (AMD), and high myopia-related fundus pathologies are the main causes of irreversible vision impairment and even blindness. Early diagnosis and timely intervention of these diseases are crucial for the prognosis of vision[1]. For instance, myopic tractional maculopathy (MTM) mainly manifests as traction-related macular damage in highly myopic eyes, including macular retinal schisis, macular retinal detachment (RD), and macular holes (MH)[2], [3]. The incidence rate of MTM among eyes with high myopia is 7.3% to 34.4%[4]. In the ATN grading system for myopic maculopathy proposed by Ruiz et al.[5], MTM is classified into six grades based on the extent of macular retinal schisis, the degree of detachment, and the morphological characteristics of macular holes (representative examples of each grade are shown in Fig. 1), providing a standardized basis for clinical diagnosis.

With the advancement of imaging technologies, optical coherence tomography (OCT)[6] enables rapid acquisition of high-resolution three-dimensional retinal images in a non-invasive manner, which can reveal the structural details and pathological changes of tissue layer of the retina well, providing critical information for the early detection and dynamic monitoring of retinal diseases. Automatic classification of retinal diseases based on OCT images has emerged as a key direction to improve the performance and efficiency of computer-aided diagnosis. However, it still faces several challenges. First, obtaining high-quality annotated data relies heavily on the experiences of ophthalmologists, resulting in substantial time and costs for labeling. This constraint limits the scale of available training data and consequently affects the generalization and robustness of the models. Second, the distribution of different kinds of diseases often shows significant imbalance, causing models to be biased toward common categories with large samples, while under-performing on rare categories with limited data[7], [8]. Moreover, different kinds of diseases may exhibit similar morphological features in OCT images with blurred inter-class

Authorized licensed use limited to: Soochow University. Downloaded on October 14,2026 at 07:10:33 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. 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 Journal of Biomedical and Health Informatics. 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/JBHI.2026.3718754

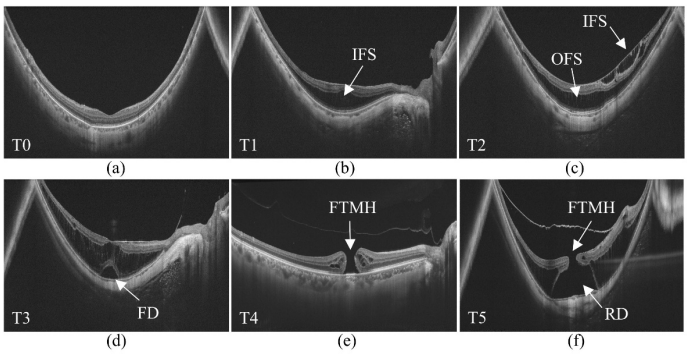

Fig. 1. Typical images of each grade of MTM. (a) T0: Normal; (b) T1: IFS or OFS; (c)T2: IFS and OFS; (d) T3: FD; (e) T4: FTMH; (f) T5: FTMH and RD. IFS denotes inner foveoschisis, OFS denotes outer foveoschisis, FD denotes foveal detachment, FTMH denotes full-thickness macular hole, and RD denotes retinal detachment.

boundaries, which increases the challenge of the classification task and the risk of misjudgment. Therefore, optimizing retinal OCT image classification algorithms to address the challenges of limited annotated data, class imbalance, and high inter-class similarity is of great clinical significance for the diagnosis of retinal diseases, including MTM.

Deep learning has achieved great success in the detection and classification of retinal diseases. Many convolutional neural networks have been developed to diagnose retinal diseases based on OCT images[9], [10], [11], [12], [13]. In recent years, Transformers have further advanced the state of the art, showing strong capabilities in medical image-based computer-aided diagnosis[14], [15], [16], [17]. However, these traditional deep learning approaches still face significant challenges. On one hand, these models typically contain large number of parameters, increasing computational costs and deployment complexity, which restricts their application in resource-constrained environments. On the other hand, these models rely heavily on large amounts of high-quality annotated data, resulting in limited generalization ability.

To address the challenges of high computation cost and heavy reliance on large-scale annotated data in traditional CNN and Transformer based models, self-distillation[18], [19] strategy has been proposed as an effective solution. By embedding dynamic auxiliary structures at one or each stage of the model and utilizing intermediate features to generate soft labels, self-distillation enables the simultaneous optimization of feature extraction and knowledge refinement, which can enhance the model's ability of feature capture and generalization under limited data conditions. Although current self-distillation methods have achieved notable results, most of them only distill from a single perspective[20], [21], [22], which may not fully exploit complementary feature information, and most structural self-distillation methods adopt single-stage distillation that tends to induce redundant structural features, thereby limiting the diversity of feature learning.

Contrastive learning[23], [24], [25] has been proposed to address the high inter-class similarity problem, which constructs positive and negative sample pairs and computes contrastive loss between them to pull positive samples closer while pushing negative samples apart, encouraging the model to capture fine-grained inter-class differences. However, in the case of class imbalance, traditional random sampling tends to result in batch sample homogenization, which compromises the negative sample constraints in contrastive learning and making it difficult for the model to learn effective discriminative features.

In this paper, a novel multi-perspective collaborative self-distillation network (MCSD-Net) is proposed, which collaboratively guides the model to achieve effective learning through multi-perspective self-distillation mechanism and category-aware contrastive learning (CACL) strategy. Specifically, a multi-perspective collaborative self-distillation mechanism is designed to conduct collaborative knowledge distillation from both structural and historical perspectives. Different from traditional single-stage structural or historical self-distillation methods, we integrate features of each stage of the model to mitigate feature redundancy and enhance the utilization efficiency of limited samples. A category-aware contrastive learning strategy is designed at the sample level by constructing class-balanced batch to enforce inter-class constraints, thereby facilitating the differentiation between intra-class and inter-class features. Due to the influence of tractional lesions and vascular lesions, the structural layers of the retina and the lesions in the OCT images often show morphological changes in the horizontal and vertical directions. Therefore, a direction-aware attention module (DAAM) is proposed to enhance the feature extraction capability. Extensive experiments on multiple retinal OCT image datasets comprehensively validate the effectiveness of the proposed MCSD-Net.

The main contributions of this work are summarized as follows:

1) We propose a novel network termed MCSD-Net, which leverages multi-perspective collaborative self-distillation mechanism and category-aware contrastive learning strategy to achieve information synergy and optimized representation learning.
2) The proposed multi-perspective collaborative self-distillation mechanism improves the utilization efficiency of limited data through structural distillation based on multi-stage feature fusion and historical distillation based on linear growth strategy.
3) At the sample level, a category-aware contrastive learning strategy is designed by forming class-balanced batch samples, which actively narrows the feature distances among same category, thereby improving its discriminative ability for similar categories even in the class imbalance cases.
4) A direction-aware attention module is designed, which strengthens the learning of texture features from both horizontal and vertical directions, thereby enhancing the feature learning ability of the model.
5) Extensive experiments are conducted on three retinal OCT image datasets, demonstrating the effectiveness of the proposed MCSD-Net, which outperforms other state-

Authorized licensed use limited to: Soochow University. Downloaded on October 14,2026 at 07:10:33 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. 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 Journal of Biomedical and Health Informatics. 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/JBHI.2026.3718754

of-the-art self-distillation based approaches.

The rest of this paper is organized as follows: Section II reviews the related works on knowledge distillation, self-distillation and contrastive learning. Section III presents our proposed MCSD-Net in details. Section IV provides experimental setup and discusses the experimental results. Section V concludes our work and discusses the limitation and future work.

II. Related Works

A. Traditional Knowledge Distillation

Traditional knowledge distillation typically involves a high-performance teacher model and a lightweight student model. It transfers knowledge from the teacher to the student via soft targets, intermediate feature matching, or other forms of supervision, thereby enhancing the model's generalization ability while maintaining lower computational costs. Niyaz et al.[26] combined knowledge distillation with mutual learning, proposing an online distillation framework that employs a single-teacher and multi-student, which yielded improved performance on both MSI-versus-MSS classification and polyp detection tasks. Akhter et al.[27] utilized Vision Transformers to distill diagnostic knowledge from high-resolution chest X-ray images, thereby enhancing the diagnostic efficacy of low-resolution ones. For skin disease classification, Wang et al.[28] proposed a self-supervised diverse knowledge distillation (SSD-KD) method replacing the traditional single relational modeling block with dual relational blocks, assisting the student model to capture both intra-instance and inter-instance relational information from the teacher. Hao et al.[29] proposed the one-for-all KD framework (OFA-KD), where intermediate features are projected into the logits space and an adaptive target enhancement scheme is introduced to prevent the student from being disturbed by irrelevant information, significantly improving the distillation performance between heterogeneous architectures.

B. Self-Distillation

Self-distillation improves training efficiency and model performance by leveraging built-in auxiliary structures for within-network, single-phase knowledge transfer. Zhang et al. proposed self-distillation networks such as BYOT[18], SCAN[30] and TECNN[19], in which shallow classifiers are constructed upon the intermediate features of different layers and knowledge distillation is utilized to enhance the accuracy of shallow classifiers and achieve internal knowledge propagation. Xu et al.[21] designed a self-distillation mechanism that enforces consistency in global feature and posterior distributions across different distorted versions of the same training data. Yun et al.[31] proposed class-wise self-knowledge distillation (CS-KD), a regularization technique that mitigates overfitting in deep neural networks by aligning the predictive distribution of samples sharing the same label. Ji et al.[32] proposed a feature refinement via self-knowledge distillation (FRSKD), which employs an auxiliary self-teacher network to transfer refined knowledge to the classifier network through joint soft-label and feature-map distillation. Kim et al.[20] proposed progressive self-knowledge distillation (PS-KD), a regularization method that adaptively adjusts soft training targets by blending ground-truth labels with the model's own historical predictions. Yang et al.[33] integrated self-knowledge distillation and Mixup image augmentation into a unified framework (MixSKD), mutually distilling feature maps and probability distributions between original images and mixup images pairs, thereby guiding the network to learn cross-image knowledge by modelling supervisory signals from mixup images. Ni et al.[22] proposed reverse self-distillation (RSD), in which the supervisory information from shallow layers is used to supervise deeper layers, achieving good classification performance on the CIFAR100 and TinyImageNet datasets. Li et al.[34] proposed a dual-teacher self-distillation framework that integrates historical and structural information to establish a collaborative supervision mechanism for more effective knowledge transfer. Hu et al.[35] proposed a hierarchical self-distillation (HSD) strategy to alleviate gradient vanishing and generate robust hierarchical features for fine-grained categorization, using predictions from deeper layers as soft labels to progressively supervise shallower ones.

C. Contrastive Learning

In medical image analysis, contrastive learning is commonly integrated with self-supervised or semi-supervised frameworks to tackle issues such as annotations scarce, class imbalance, and high inter-class similarity. Wang et al.[36] proposed a semantically-relevant contrastive learning strategy, which evaluates instance relevance to mine additional positive pairs, thereby increasing positive diversity and learning more informative representations for histopathology image classification. Zhang et al.[37] proposed a triplet attention and dual-pool contrastive learning (DCL) method for multi-label medical image classification, in which DCL includes dual-pool contrastive training for improving the discrimination of label embeddings and dual-pool contrastive inference for reducing the clinical risks caused by misclassification. Pan et al.[38] proposed a contrastive CutMix data augmentation method, constructing augmented samples with semantically consistent labels to improve model performance of long-tailed recognition. Yu et al.[39] employed feature enhanced contrastive learning alongside a dual-consistency strategy to address under-segmentation in semi-supervised gland segmentation tasks.

III. Methodology

A. Overview of the Proposed Method

Fig. 2 illustrates the overall architecture of the proposed MCSD-Net. ResNet-50[40] pre-trained on ImageNet is adopted as the backbone. Given an input image $X \in \mathbb{R}^{C\times H\times W}$, feature maps $F_i \in \mathbb{R}^{C_i\times\frac{H}{2^i}\times\frac{W}{2^i}}$ $(i = 1,2,3,4)$ with four different resolutions are extracted through the four sequential stages.

Authorized licensed use limited to: Soochow University. Downloaded on October 14,2026 at 07:10:33 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. 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 Journal of Biomedical and Health Informatics. 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/JBHI.2026.3718754

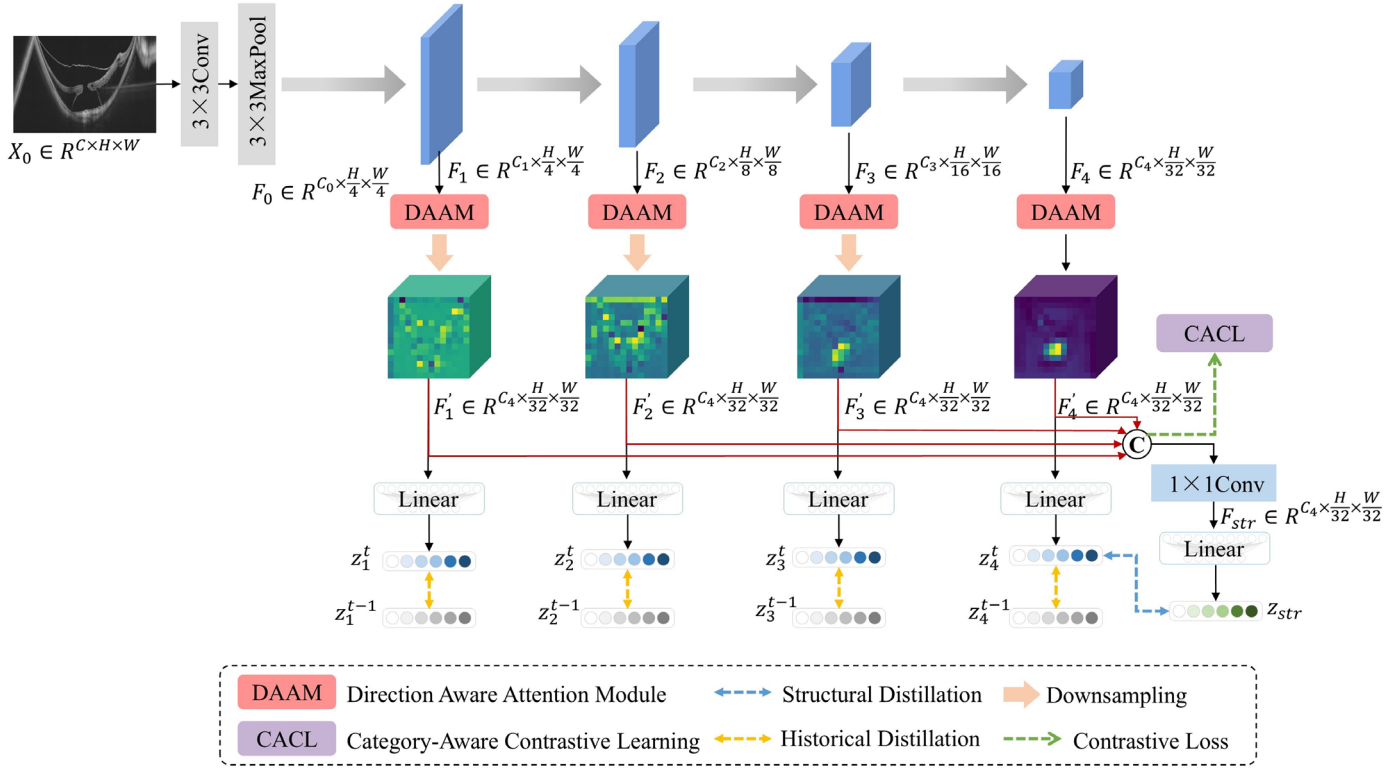

Fig. 2. The framework of the proposed MCSD-Net. The backbone consists of four stages. Auxiliary branches are attached to extract features from each stage, with each branch incorporating a DAAM and being supervised by historical information. Features from all stages are fused (indicated by the red lines) to generate structural knowledge, which subsequently supervises the final stage of the backbone. The fused features are also utilized by the CACL strategy to perform contrastive learning. During the inference phase, only the backbone is employed for feature extraction to perform OCT classification.

Each F_i is subsequently processed by DAAM to capture channel-wise inter-dependencies along with texture features in the horizontal and vertical directions. Given that the feature maps from the shallower stages contain rich low-level information such as image contours and textures, an independent auxiliary branch is attached to each of the first three stages of the backbone for knowledge transfer. Each auxiliary branch consists of a downsampling module followed by a fully connected (FC) layer. The downsampling module, composed of a convolutional layer and a batch normalization layer, adjusts the spatial size of the shallow feature maps to match the resolution of the output feature of the fourth stage. The FC layer then produces the shallow logits z_i $(i = 1,2,3)$.

During the training phase, the backbone and the auxiliary branches are jointly optimized through mutual learning based on the multi-perspective collaborative self-distillation mechanism and the CACL strategy. The multi-perspective collaborative self-distillation mechanism facilitates knowledge transfer from two perspectives: structural information and historical information. The multi-stage feature fusion based structural distillation strategy integrates feature maps from different stages to transfer detailed information to the backbone, thereby strengthening its perceptual capacity. The linear growth based historical distillation strategy dynamically guides the feature learning in the current epoch by leveraging a fusion of the predictions of the previous epoch and the ground truth. The CACL strategy constructs category balanced batches and calculates the contrastive loss between anchor samples and their positive/negative pairs within each batch. This encourages feature representations to be more compact within the same class and more separable between different classes, thereby enhancing the model's discriminative capability for inter-class features. By optimizing the distill ation loss and the contrast loss, the classification performance of the model is improved. During the inference phase, only the backbone is retained for feature extraction, and the final classification result is output by the backbone's classifier.

B. Multi-Perspective Collaborative Self-Distillation Mechanism

To improve the utilization efficiency of limited data and enhance both the training efficiency and performance of the network, we propose a multi-perspective collaborative self-distillation mechanism, which integrates multi-stage feature fusion based structural distillation and linear growth based historical distillation.

1) Structural Distillation Based on Multi-stage Feature Fusion: During the feature extraction process, repeated downsampling and convolutional operations may lead to the loss of fine-grained or detailed features. To mitigate this issue, the downsampling module and FC layer based auxiliary branch is adopted to provide multi-scale feature maps to the backbone, enriching it with detailed information (e.g., edges, textures). Traditional structural distillation usually guides the backbone using individual stage features, which may result in

Authorized licensed use limited to: Soochow University. Downloaded on October 14,2026 at 07:10:33 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. 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 Journal of Biomedical and Health Informatics. 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/JBHI.2026.3718754

information fragmentation or feature redundancy, making it difficult to focus on the most discriminative cues[18], [19]. Thus, we fuse multi-stage and multi-scale features to generate soft labels that guide the classifier of the backbone, thereby improving its discriminative capability and performance. Specifically, given an input image $X \in \mathbb{R}^{C\times H\times W}$, the network gets feature maps $F_i \in \mathbb{R}^{C_i\times\frac{H}{2^i}\times\frac{W}{2^i}}$ $(i = 1,2,3,4)$ through four stages of feature extraction. Auxiliary branches downsample the shallow and deep feature maps to align the spatial dimensions. These aligned feature maps are then concatenated along the channel dimension to get feature F_{str}, which is followed by a 1×1 convolution for dimensionality reduction and a FC layer to produce the logits z_{str}. The softmax function is applied to obtain the soft target probability distribution p_{str}, which serves as a structured guidance for the output distribution p_4 of the classifier. The structural distillation is optimized by using the Kullback-Leibler (KL) divergence loss, defined as,

$$\mathcal{L}_{STR} = \mathcal{L}_{KL}(p_{str}\|p_4) = T_{str}^2 \sum_{m=1}^{M} p_{str}^m \cdot log\frac{p_{str}^m}{p_4^m} \quad (1)$$

where M denotes the number of categories, T_{str} represents the temperature parameter for structural distillation.

2) Historical Distillation Based on Linear Growth Strategy: As the number of training epochs increases, the network may experience the phenomenon of catastrophic forgetting. Historical information, however, often embodies meaningful guidance that can enhance the current optimization process. To mitigate knowledge forgetting and improve performance, a linear growth based historical distillation mechanism is introduced, which leverages past prediction information to guide the current network.

Specifically, a linear weighted combination of the prediction from the (t-1)-th epoch p_i^{t-1} and the ground truth y is constructed as the historical target distribution $p_{his,i}(i = 1,2,3,4)$,

$$\alpha_t = \alpha_T \times \frac{t}{Epochs} \quad (2)$$

$$p_{his,i} = (1-\alpha_t)\cdot y + \alpha_t \cdot p_i^{t-1}, i = 1,2,3,4 \quad (3)$$

where α_t denotes the weight of the historical information of the t-th epoch, α_T is the initial weight of α_t, y is the ground truth, and Epochs represents the total number of training rounds.

At the beginning of the training, it is hoped that the network will learn more useful information from the ground truth y. As the number of training rounds increases, the proportion of historical information is gradually increased through a linear growth strategy, promoting the network to achieve comprehensive utilization of historical information and true labels. The optimization process for historical distillation at the t-th epoch can be formulated as:

$$\mathcal{L}_{HIS} = \sum_{i=1}^{S}\mathcal{L}_{KL}(p_{his,i}\|p_i) = \sum_{i=1}^{S}\left(T_{his}^2\sum_{m=1}^{M} p_{his,i}^m \cdot log\frac{p_{his,i}^m}{p_i^m}\right) \quad (4)$$

where p_i is the probability distribution obtained by applying the softmax function to z_i $(i = 1,2,3,4)$, S denotes the stage of the network, M represents the number of categories, and T_{his} represents the temperature parameter for historical distillation.

C. Category Aware Contrastive Learning

To further relieve the issues of class imbalance and high inter-class similarity in retinal OCT image classification, a category aware contrastive learning (CACL) strategy is proposed. In the case of class imbalance, traditional random sampling tends to result in batches with low class diversity, which undermines the positive/negative pair constraints essential for contrastive learning. Therefore, a category-aware batch construction method is designed, which increases the probability of diverse categories within each batch and alleviates the dominant bias of majority classes, so as to facilitate stable inter-class contrastive constraints. The procedure is detailed in Algorithm 1. Assuming a target batch with size of B and a total of M classes, a Set is first initialized to record the selected classes. A random probability δ is generated for sampling. If δ is less than the threshold (e.g., 0.5) or the Set is empty, a class will be randomly selected from the M classes and added to Set. If δ is greater than or equal to the threshold, a class from the current Set will be randomly chosen. Subsequently, if the selected class has available samples, one sample will be randomly drawn from its index list and added to the current batch. Otherwise, a new target class will be selected. Repeats above processes until the batch size reaches B.

In order to fully leverage the semantic information from different stages, features F_i $(i = 1,2,3,4)$ from four stages of the backbone are batch normalized respectively,

$$\hat{F}_i = BatchNorm(F_i), i = 1,2,3,4 \quad (5)$$

Then $\hat{F}_i$ $(i = 1,2,3,4)$ are concatenated along the channel dimension to generate a high-dimensional representation $F \in \mathbb{R}^{B\times Dim}$, which integrates multi-stage information,

$$F = Concat\left[\hat{F}_1, \hat{F}_2, \hat{F}_3, \hat{F}_4\right] \quad (6)$$

where B is the batch size and Dim is the feature dimension after concatenation. L_2 normalization is adopted to the high-dimensional representation F, yielding the final feature representation,

$$f_j = \frac{F_j}{\|F_j\|_2}, j = 1,2,\ldots,\mathrm{B} \quad (7)$$

where f_j denotes the normalized feature of the j-th sample within a batch.

For any given sample j in the batch, samples sharing the same class label are treated as positive samples, while those with different labels are considered negative samples. The category-aware contrastive loss $\mathcal{L}_{CACL}$ is computed as follows:

$$\mathcal{L}_{CACL} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{p\in\mathcal{P}_j} log\frac{exp(f_j\cdot f_p/\tau)}{\sum_{n\in\mathcal{A}_j} exp(f_j\cdot f_n/\tau)+\epsilon} \quad (8)$$

where $\mathcal{P}_j$ represents the set of samples belonging to the same class as sample j excluding itself, $\mathcal{A}_j$ denotes the set of all samples in the batch except j. p and n are indices for positive

Authorized licensed use limited to: Soochow University. Downloaded on October 14,2026 at 07:10:33 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. 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 Journal of Biomedical and Health Informatics. 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/JBHI.2026.3718754

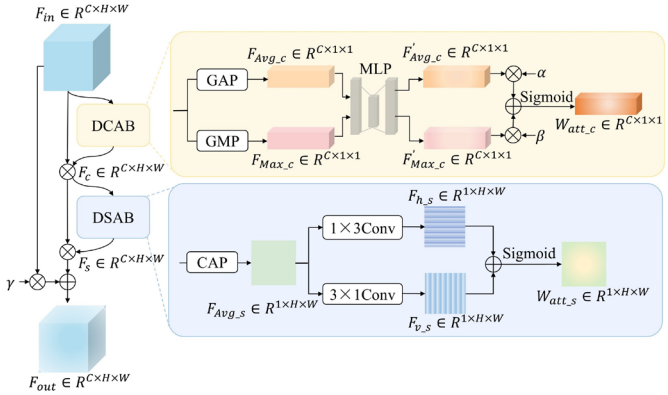

Fig. 3. The architecture of the DAAM.

Algorithm 1 Category-Aware Batch Construction

Input: Dataset D; class_indices; Batch size B; Number of classes M; Threshold δ;
Initialize: data = {}; Set = {};
1: **while** length(data) < B **do**
2: **if** Set == ∅ or δ < 0.5 **then**
3: selected_class = Random(M);
4: Set ← Set ∪ {selected_class};
5: **else**
6: selected_class = Random(Set);
7: **end if**
8: **if** classes_indices[selected_class] ≠ ∅ **then**
9: idx ← Random(classes_indices[selected_class]);
10: data.append(D[idx]);
11: **end if**
12: **end while**
13: Output: data

and negative samples, respectively. τ is the temperature coefficient for the contrastive learning. ϵ is a small constant to ensure numerical stability during computation.

D. Direction Aware Attention Module

Given that retinal lesions often exhibit direction-specific pathological characteristics along horizontal and vertical orientations, the DAAM is designed with attention mechanisms across both channel and spatial dimensions to enhance lesion-related features. The structure of the DAAM is shown in Fig. 3, which mainly consists of a Dynamic Channel Attention Block (DCAB) and a Direction-aware Spatial Attention Block (DSAB). The DCAB emphasizes inter-dependencies between channels to strengthen feature representations for specific semantics. The DSAB strengthens the learning of horizontal and vertical pathological features in the OCT image, thereby boosting lesion identification capability of the model.

Specifically, for the DCAB, the input feature map $F_{in} \in \mathbb{R}^{C\times H\times W}$ is first squeezed via global average pooling (GAP) and global max pooling (GMP) to get the aggregated features $F_{Avg_c} \in \mathbb{R}^{C\times 1\times 1}$ and $F_{Max_c} \in \mathbb{R}^{C\times 1\times 1}$ respectively. Then the aggregated features are fed to a multi-layer perceptron (MLP) composed of two FC layers and a non-linear activation function respectively. Learnable parameters α and β are introduced to perform adaptively weighted summation of the two features. The combined feature is finally passed through a Sigmoid activation function to generate the channel attention weights W_{att_c},

$$W_{att_c} = \sigma\left(\alpha \cdot MLP\left(GAP(F_{in})\right) + \beta \cdot MLP\left(GMP(F_{in})\right)\right) \tag{9}$$

where $\sigma(\cdot)$ denotes the Sigmoid function, $GAP(\cdot)$ represents GAP and $GMP(\cdot)$ represents GMP.

Multiply the input feature map F_{in} with the channel attention weights W_{att_c} to generate the channel attention feature map $F_c \in \mathbb{R}^{C\times H\times W}$,

$$F_c = F_{in} \otimes W_{att_c} \tag{10}$$

where $\otimes$ denotes the element-wise multiplication.

F_c is taken as the input feature of the DSAB. First, channel average pooling (CAP) is performed to aggregate the channel information of the feature map and obtain feature F_{Avg_s}. Then, feature F_{Avg_s} is subjected to 1×3 horizontal convolution and 3×1 vertical convolution respectively to obtain the horizontal feature map F_{h_s} and the vertical feature map F_{v_s}. F_{h_s} and F_{v_s} are summed and activated by Sigmoid function to generate the spatial attention weights W_{att_s},

$$W_{att_s} = \sigma\left(Conv_{1\times 3}\left(CAP(F_c)\right) + Conv_{3\times 1}\left(CAP(F_c)\right)\right) \tag{11}$$

where $\sigma(\cdot)$ denotes the Sigmoid function, $CAP(\cdot)$ represents CAP, $Conv_{1\times 3}$ and $Conv_{3\times 1}$ denote the horizontal and vertical convolution kernels respectively.

Multiply the channel attention map F_c with W_{att_s} to generate the spatial attention feature map $F_s \in \mathbb{R}^{C\times H\times W}$,

$$F_s = F_c \otimes W_{att_s} \tag{12}$$

Finally, the output feature map F_{out} of the DAAM is obtained by computing a weighted sum of the input feature map F_{in} and the spatial attention map F_s, following by a ReLU activation function,

$$F_{out} = ReLU(F_s + \gamma \cdot F_{in}) \tag{13}$$

where γ is a learnable parameter for adjusting the proportion of the input feature map F_{in}.

E. Loss Function

The proposed model is optimized under the joint guidance of the classification loss, structural distillation loss, historical distillation loss, and category-aware contrastive learning loss. Cross-entropy loss is employed as the classification loss:

$$\mathcal{L}_{CLS} = \sum_{s=1}^{S}\sum_{m=1}^{M} y_m log p_m \tag{14}$$

where S denotes the stage of the network, M represents the number of categories, and y_m represents the ground truth, p_m is the output of the model.

Consequently, the total loss function can be defined as

Authorized licensed use limited to: Soochow University. Downloaded on October 14,2026 at 07:10:33 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. 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 Journal of Biomedical and Health Informatics. 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/JBHI.2026.3718754

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{CLS} + \lambda_1 \times \mathcal{L}_{STR} + \lambda_2 \times \mathcal{L}_{HIS} + \lambda_3 \times \mathcal{L}_{CACL} \quad (15)$$

where $\mathcal{L}_{CLS}$ denotes the classification cross-entropy loss, $\mathcal{L}_{STR}$, $\mathcal{L}_{HIS}$ and $\mathcal{L}_{CACL}$ represent the structural distillation loss, historical distillation loss and category-aware contrastive learning loss, respectively. λ_1, λ_2 and λ_3 are the weights of structural distillation loss, historical distillation loss and category-aware contrastive learning loss, respectively. In our experiments, λ_1, λ_2 and λ_3 are all set to 1, which means the four items in Equation (15) are set to equal weights.

IV. Experiments

A. Datasets

To verify the effectiveness of the proposed method, extensive experiments have been conducted on three retinal OCT image datasets.

1) The MTM Dataset: This dataset is provided by Shanghai General Hospital, which consists of 4,117 OCT B-scan images from 246 eyes with myopic tractional maculopathy. The collection and analysis of image data were approved by the Medical Ethics Committee of Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, according to the Declaration of Helsinki. Written informed consent was obtained from all study participants. The current study was registered at http://www.clinicaltrials.gov, with a registration number of NCT03446300. According to the ATN grading system, the images are classified into six grades (T0 to T5) under the supervision of a senior ophthalmologist. Typical images of each grade are shown in Fig. 1. Due to the limited number of samples in grades T4 and T5, data augmentations including horizontal flipping, random rotation (within the range of 10° to 20°) and their combination have been applied to increase the sample diversity. The final data distribution of each grade is shown in TABLE I. Taking eye as the unit, a stratified 5-fold cross-validation strategy is used to evaluate the proposed MCSD-Net, which ensures that each fold maintains the identical class distribution. To the best of our knowledge, this is the first work focusing on the automatic grading of MTM.

2) The OCT2017 Dataset[41]: The OCT2017 dataset contains 109,312 horizontal macular single-line OCT B-scan images collected from 4,686 patients, spanning four diagnostic categories, including: choroidal neovascularization (CNV), diabetic macular edema (DME), drusen, and normal. The dataset is divided into a training set of 108,312 images and a test set of 1,000 images (250 images per class). In our experiments, the training set is divided into training and validation subsets according to the ratio of 8:2.

3) The OCTDL Dataset[42]: The OCTDL dataset is collaboratively provided by Erlangen-Nurnberg Pattern Recognition Lab of Friedrich-Alexander University in Germany and Ural Federal University in Russia. It consists of 2,064 OCT B-scan images from 821 patients, spanning seven diagnostic categories, including: AMD, DME, epiretinal membrane (ERM), retinal artery occlusion (RAO), retinal vein occlusion (RVO), vitreomacular interface disease (VID), and normal (NO). The dataset is randomly divided into training, validation and test sets according to the ratio of 6:1.5:2.5.

TABLE I
Distribution of the MTM dataset

	T0	T1	T2	T3	T4	T5	Total
Eyes	65	51	69	37	15	9	246
OCT Images	1096	958	1240	693	364	156	4507

B. Implementation Details and Evaluation Metrics

All experiments are implemented using PyTorch 1.10.0 and CUDA 11.4 and NVIDIA GTX 3090 GPUs with 24 GB of memory. For MTM and OCTDL datasets, the total training epoch is 100, the batch size is set to 8, the initial learning rate is set to 0.001 and decayed by a factor of 10 every 30 epochs. For OCT2017 dataset, the total training epoch is 100, the batch size is set to 32, the initial learning rate is 0.001 and decayed by a factor of 10 every 15 epochs. The stochastic gradient descents (SGD) strategy is used as the optimizer with momentum 0.9 and weight decay 5e-4.

Accuracy (ACC), F1-score, area under the ROC curve (AUC) and specificity (Spe) are adopted as evaluation metrics for classification performance evaluation. In addition, model parameters (Params), FLOPs, training time (the total training time for the five-fold cross-validation experiments) and inference speed (the average latency per image) are reported to evaluate the complexity and computational cost of the models.

C. Comparison Experiments

To evaluate the performance of the proposed MCSD-Net, it is compared with several self-distillation based methods, including the progressive self-distillation method PS-KD[20], BYOT[18], the feature refinement-based self-distillation method FRSKD[32], self-distillation for efficient and compact neural networks TECNN[19], reverse self-distillation RSD[22], data distortion-guided self-distillation DDGSD[21], class-wise self-distillation CS-KD[31], self-distillation from image mixup MixSKD[33], and dual-teacher self-distillation DTSKD[34]. We also compare MCSD-Net with recent OCT task-specific models such as RETFound[43]and VisionFM[44]. The comparison experimental results on three datasets are presented in TABLE II, III and IV, and the t-SNE visualization results for each dataset are illustrated in Figures 5, 6, and 7, respectively.

As shown in TABLE II, on the MTM dataset, the propose MCSD-Net outperforms all other methods for all evaluation metrics and achieves the smallest standard deviation, indicating superior and more stable performance. PS-KD only relies on historical information and lacks effective modeling of inter-sample similarity, which limits its performance in classification tasks with subtle inter-sample differences. FRSKD introduces an auxiliary self-teacher network focusing on the retention and utilization of local detailed information. However, in our retinal OCT classification task, due to the high similarity of features among different categories, relying solely on local information can easily lead the model to be confused when distinguishing similar samples. BYOT, TECNN and RSD employ structural distillation strategies by incorporating additional shallow classifiers to facilitate knowledge transfer between deep and shallow layers, which improves the performance to some extent. Nevertheless, these

Authorized licensed use limited to: Soochow University. Downloaded on October 14,2026 at 07:10:33 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. 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 Journal of Biomedical and Health Informatics. 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/JBHI.2026.3718754

TABLE II
COMPARISON RESULTS ON THE MTM DATASET (%)

Methods	ACC	F1-score	AUC	Spe	Params (M)	FLOPs (G)	Training Time (h)	Inference (ms/img)
PS-KD[20]	82.34±4.22	82.22±4.38	96.29±1.49	95.32±1.05	**25.55**	4.58	**4.72**	3.99
BYOT[18]	86.34±5.94	86.57±5.27	97.10±1.97	96.54±1.04	33.18	74.10	21.33	3.79
FRSKD[32]	87.53±3.23	87.54±3.33	97.55±1.24	96.53±0.82	25.56	**4.13**	7.89	3.99
TECNN[19]	86.50±6.06	86.45±5.71	96.93±2.40	96.58±1.15	45.80	84.68	25.05	4.98
RSD[22]	86.23±4.54	85.95±4.44	96.63±2.90	96.34±1.07	43.01	79.33	21.62	5.16
DDGSD[21]	83.17±4.33	82.92±4.39	96.45±1.59	95.29±0.93	**25.55**	4.58	7.91	**3.19**
CS-KD[31]	86.22±3.04	85.93±3.00	96.72±1.46	96.02±0.68	25.56	**4.13**	5.83	3.59
MixSKD[33]	88.35±2.92	88.44±2.80	97.44±0.88	96.99±0.64	96.46	10.55	14.28	8.17
DTSKD[34]	87.96±2.67	88.09±3.26	96.90±1.95	96.59±0.76	73.15	16.19	6.99	8.58
RETFound[43]	82.90±4.48	82.54±4.60	96.17±2.15	95.12±1.22	303.31	59.69	11.69	7.03
VisionFM[44]	79.75±3.65	79.02±3.70	95.08±2.23	94.39±0.78	86.14	17.86	9.62	5.71
MCSD-Net	**91.36±1.46**	**91.24±1.28**	**98.26±0.55**	**97.61±0.29**	55.85	16.89	8.48	4.91

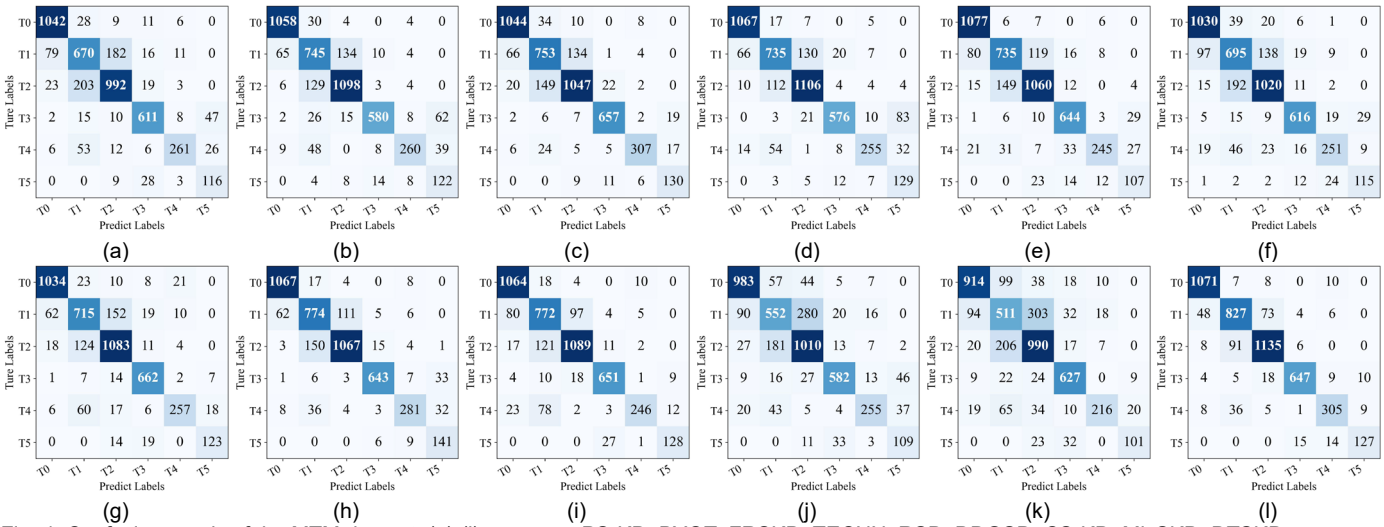

Fig. 4. Confusion matrix of the MTM dataset. (a)-(l) represent PS-KD, BYOT, FRSKD, TECNN, RSD, DDGSD, CS-KD, MixSKD, DTSKD, RETFound, VisionFM and the proposed MCSD-Net.

methods primarily focus on network structure optimization, lacking mechanisms for handling sample similarity. The independent stage-wise guidance mechanism may lead to information fragmentation. CS-KD, DDGSD and MixSKD adopt sample distillation strategies. CS-KD overemphasizes intra-class consistency, potentially suppressing discriminative features of intra-class samples. DDGSD employs an indiscriminate augmentation strategy, resulting in insufficient diversity of augmented features. MixSKD utilizes mixed images for distillation, which may disrupt the subtle morphological information present in retinal OCT images. DTSKD introduces a dual-teacher mechanism combining structural and historical information, providing complementary supervisory information. However, it lacks specific measures to address feature similarity and class imbalance.

Due to the limited size of the MTM dataset and the challenges such as high inter-class similarity and class imbalance, the classification performances of OCT task-specific foundation models RETFound and VisionFM, which rely on a large amount of training data, are not satisfactory on the MTM dataset.

The proposed MCSD-Net effectively addresses the challenges of high inter-class similarity, subtle feature differences and class imbalance in retinal OCT classification tasks by integrating multi-scale feature extraction, category-aware contrastive learning strategy and multi-perspective collaborative self-distillation mechanism. It demonstrates superior classification performance and stability on the MTM dataset.

In terms of the model complexity and computational cost, as can be seen from TABLE II, although the proposed MCSD-Net adopts a multi-branch architecture with multiple optimization strategies, its parameters, computational cost (FLOPs) and training time remain at a moderate level, significantly lower than those of Transformer-based models such as RETFound and VisionFM. Since only the backbone is used during inference, the inference speed of MCSD-Net is also at an intermediate level. Overall, the proposed MCSD-Net achieves a favorable balance between classification performance and model efficiency.

Fig. 4 shows the confusion matrices of different methods on the MTM dataset. As can be seen from Fig. 4, the proposed MCSD-Net achieves excellent performance across most categories. Particularly, it outperforms other methods in distinguishing the highly similar T1 and T2 grades, demonstrating its effectiveness in capturing subtle pathological differences. For minority classes such as T4 and T5, the proposed MCSD-Net can alleviate classification bias caused by class imbalance, which is second only to FRSKD.

Authorized licensed use limited to: Soochow University. Downloaded on October 14,2026 at 07:10:33 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. 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 Journal of Biomedical and Health Informatics. 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/JBHI.2026.3718754

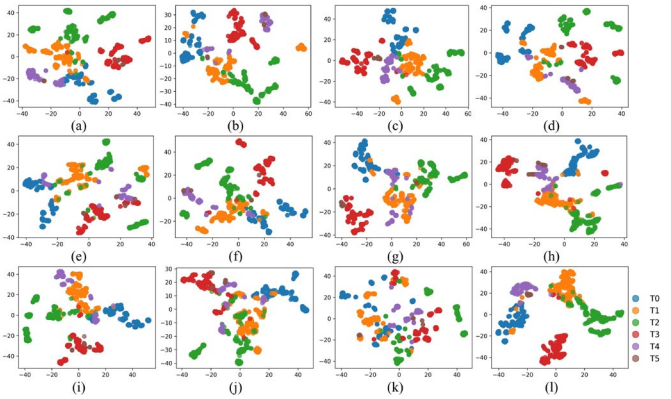

Fig. 5. The t-SNE visualization results of the MTM dataset. (a)-(l) represent PS-KD, BYOT, FRSKD, TECNN, RSD, DDGSD, CS-KD, MixSKD, DTSKD, RETFound, VisionFM and the proposed MCSD-Net.

TABLE III
COMPARISON RESULTS ON THE OCT2017 DATASET (%)

Methods	ACC	F1-score	AUC	Spe
PS-KD[20]	94.70	94.63	99.87	98.23
BYOT[18]	95.50	95.52	99.85	98.50
FRSKD[32]	96.70	96.65	99.94	98.50
TECNN[19]	96.40	96.40	99.91	98.80
RSD[22]	96.60	96.57	99.92	98.87
DDGSD[21]	94.80	94.75	99.82	98.27
CS-KD[31]	94.90	94.85	99.61	98.30
MixSKD[33]	93.90	93.85	99.89	97.97
DTSKD[34]	97.00	96.99	99.95	99.00
RETFound[43]	96.20	96.16	99.85	98.73
VisionFM[44]	95.60	95.53	99.73	98.53
MCSD-Net	**97.80**	**97.80**	**99.92**	**99.27**

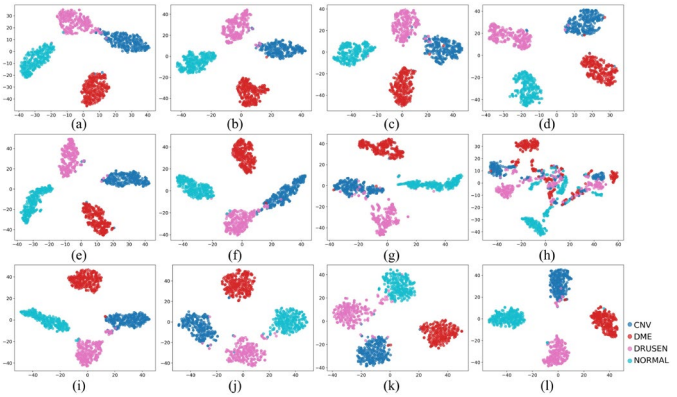

Fig.6. The t-SNE visualization results of the OCT2017 dataset. (a)-(l) represent PS-KD, BYOT, FRSKD, TECNN, RSD, DDGSD, CS-KD, MixSKD, DTSKD, RETFound, VisionFM and the proposed MCSD-Net.

In summary, MCSD-Net achieves a balanced optimization across different MTM grades and substantially improves grading robustness.

From the t-SNE visualization shown in Fig. 5, it can be seen that in the feature distribution generated by the proposed MCSD-Net, the features within same categories are closer, while the distance between different categories is greater. In contrast, the distributions of other methods are relatively scattered, with many overlapping parts between different classes, indicating that these methods cannot focus well on the feature differentiation within and between classes.

As can be seen from TABLE III, all methods achieve good classification performance on the OCT2017 dataset. The proposed MCSD-Net outperforms other methods on the majority of evaluation metrics. The t-SNE visualization results presented in Fig. 6 indicate that different classes are well-separable in the two-dimensional feature space for most of the methods. In the t-SNE plot of the proposed MCSD-Net, various types of samples form more compact and clearly bounded clusters, demonstrating that our model learns more discriminative feature representations. In contrast, the clusters for some classes in other comparison methods appear relatively diffuse (e.g. BYOT, FRSKD and TECNN), or exhibit a certain degree of overlaps between different classes (e.g. CS-KD and MixSKD), suggesting that these categories are more challenging to be distinguished.

TABLE IV
COMPARISON RESULTS ON THE OCTDL DATASET (%)

Methods	ACC	F1-score	AUC	Spe
PS-KD[20]	81.92	81.68	95.43	96.99
BYOT[18]	84.75	83.79	95.32	97.46
FRSKD[32]	84.18	84.02	95.54	97.36
TECNN[19]	80.79	80.37	95.05	96.80
RSD[22]	78.15	75.01	93.55	96.36
DDGSD[21]	82.86	82.51	95.30	97.14
CS-KD[31]	83.80	82.53	95.50	97.30
MixSKD[33]	73.45	70.73	92.32	95.57
DTSKD[34]	73.26	68.89	89.80	95.54
RETFound[43]	83.19	83.06	95.51	96.87
VisionFM[44]	66.48	60.79	77.80	94.41
MCSD-Net	**84.93**	**84.19**	**95.57**	**97.49**

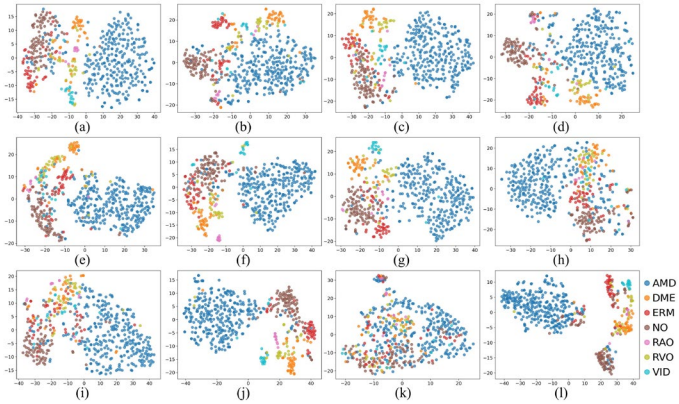

Fig.7. The t-SNE visualization results of the OCTDL dataset. (a)-(l) represent PS-KD, BYOT, FRSKD, TECNN, RSD, DDGSD, CS-KD, MixSKD, DTSKD, RetFound, VisionFM and the proposed MCSD-Net.

As shown in TABLE IV, the proposed method also achieves the best results in all four metrics on the OCTDL dataset. The t-SNE visualization shown in Fig. 7 demonstrates that, compared to other methods, MCSD-Net exhibits superior for inter-class separation and intra-class compactness, further validating its advantages in feature representation learning.

D. Ablation Experiments

Several ablation experiments are conducted on the MTM dataset to investigate the contribution of key components in the proposed method, including the selection of backbone networks, hyperparameter settings, the attention module, self-distillation and contrastive learning strategies, stage-wise feature fusion for structural distillation, and growth strategies in historical distillation.

Authorized licensed use limited to: Soochow University. Downloaded on October 14,2026 at 07:10:33 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. 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 Journal of Biomedical and Health Informatics. 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/JBHI.2026.3718754

TABLE V
ABLATION EXPERIMENT RESULTS OF BACKBONES (%).

Model	ACC	F1-score	AUC	Spe
ResNexT50[45]	87.38±3.86	87.20±3.89	96.00±1.75	96.67±0.92
ConvNexT-B[46]	84.88±3.40	85.01±3.39	95.55±2.55	95.93±1.13
OverLoCK-B[47]	77.01±6.63	75.36±2.01	91.55±8.87	93.26±5.52
ViT-B[48]	70.75±4.98	70.27±5.20	90.88±2.85	91.62±1.35
Swin-Transformer-B[49]	87.54±3.56	87.14±4.06	97.44±1.09	96.59±0.75
TransNexT-B[50]	89.57±3.19	89.63±3.21	97.48±0.61	97.14±1.02
MCSD-Net	**91.36±1.46**	**91.24±1.28**	**98.26±0.55**	**97.61±0.29**

1) Ablation Study on Backbones: To validate the influence of different backbone architectures on model performance, we replace ResNet50 with several alternatives, including the CNN-based ResNeXt50[45], ConvNeXt-B[46] and OverLoCK-B[47], and the Transformer-based ViT-B[48], Swin-Transformer-B[49], and TransNeXt-B[50]. As shown in TABLE V, ResNet50 achieves an accuracy of 91.36%±1.46%, substantially surpassing all other backbones, with optimal performance across all four metrics and the smallest standard deviations, reflecting superior stability. For the CNN-based backbones, ResNeXt50 suffers from insufficient inter-channel correlation, making it incompatible with the proposed multi-stage feature fusion mechanism. ConvNeXt-B, relying on depthwise separable convolutions and layer normalization, is more suitable for large-scale feature learning, and its effectiveness diminishes under the limited data scale of this study. OverLoCK-B introduces feature redundancy due to its intricate network hierarchy. The Transformer-based backbones tend to lose local details and are susceptible to overfitting when data are insufficient, resulting in performance deterioration. TransNeXt-B achieves competitive performance, benefiting from its aggregated attention and convolutional design that enhance local-global feature integration. However, its hybrid architecture is less tailored to the subtle pathological variations in retinal OCT images, limiting its ability to capture fine-grained lesion-specific features. Overall, compared to other CNN- and Transformer-based architectures, ResNet50 aligns more naturally with the design of MCSD-Net, striking an effective balance between feature extraction efficiency and classification performance. Accordingly, ResNet50 is adopted as the backbone network in this study.

2) Hyperparameter Selection: The proposed MCSD-Net involves four main hyperparameters, including the temperature coefficient T_{str} for the structural distillation loss, the temperature coefficient T_{his} for the historical distillation loss, the temperature coefficient τ for the contrastive learning loss, and the initial weight α_T for the historical information. We perform ablation experiments to evaluate the impact of different values for these parameters. Fig. 8 illustrates the influence of each hyperparameter on model performance. As shown in Fig. 8(a) to (c), the classification accuracy initially increases and then decreases with rising temperature coefficients, reaching its optimum when $T_{str} = 1$, $T_{his} = 0.1$, $\tau = 0.3$. With these settings, the model achieves the best mean accuracy with smallest standard deviation, indicating that the corresponding loss functions contribute most effectively to model performance. Fig. 8(d) shows the model performances with different initial weights α_T. It can be seen from Fig. 8(d) that the best performance is achieved when $\alpha_T = 0.8$. When

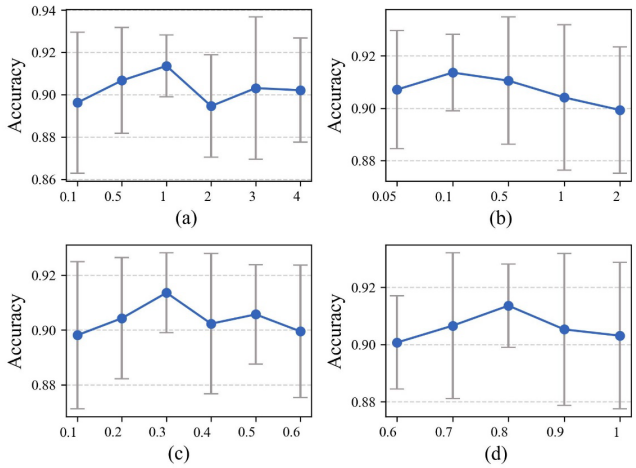

Fig. 8. Ablation experiments of the hyperparameter. (a) Structural distillation loss temperature coefficient T_{str}; (b) Historical distillation loss temperature coefficient T_{his}; (c) Contrast learning loss temperature coefficient τ; (d) Initial weight of historical information α_T. The blue dots represent the mean accuracy, and the whiskers represent the standard deviations.

α_T increases or decreases, the accuracy will decline. This indicates that if the weight is too large, the model may overly rely on historical information, while if it is too small, the model may fail to fully utilize the historical information. Setting $\alpha_T = 0.8$ achieves an optimal balance between historical information and the ground truth, thereby obtaining optimized generalization ability.

3) Ablation Study on the DAAM: TABLE VI presents the ablation experimental results for the DAAM, along with comparisons against classical attention mechanisms. Taking the proposed MCSD-Net without DAAM as the baseline in this ablation study. The baseline model achieves an accuracy of 88.91%±2.53%. The incorporation of either the DCAB or the DSAB individually with the baseline leads to a performance improvement, indicating that both attention mechanisms effectively enhance feature representation. Furthermore, combining DSAB and DCAB yields superior performance. Specifically, the DCAB+DSAB configuration achieves the best accuracy of 91.36%±1.46%. This result suggests that sequentially applying channel attention to capture global feature interdependencies, followed by spatial attention to refine local regional features, provides an optimal pathway for performance gain. In summary, the design of the DAAM significantly strengthens the capability for feature capture and discrimination, leading to substantial performance enhancement.

To further validate the superiority of the DAAM, it is compared with classical attention mechanisms, including SE[51], CBAM[52], and ECA[53]. The results shown in TABLE VI indicate that the proposed DAAM outperforms all these classical attention modules across all metrics, with smaller standard deviations, demonstrating superior performance and stability.

4) Ablation Study on Self-distillation and Contrastive Learning Strategies: Taking ImageNet pre-trained ResNet50 with DAAM as the baseline in this ablation study. To validate the effectiveness of the proposed multi-perspective collaborative self-distillation and CACL strategies and to

Authorized licensed use limited to: Soochow University. Downloaded on October 14,2026 at 07:10:33 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. 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 Journal of Biomedical and Health Informatics. 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/JBHI.2026.3718754

TABLE VI
ABLATION EXPERIMENT RESULTS OF THE DAAM (%).

Models	ACC	F1-score	AUC	Spe
Baseline*	88.91±2.53	88.65±2.39	97.02±1.66	96.97±0.54
DCAB	89.59±2.53	89.27±2.33	97.97±0.86	97.09±0.74
DSAB	89.27±2.60	88.92±2.55	97.48±1.44	97.03±0.80
DSAB+DCAB	90.39±1.90	90.23±1.83	97.43±0.94	97.51±0.57
DCAB+DSAB	**91.36±1.46**	**91.24±1.28**	**98.26±0.55**	**97.61±0.29**
SE[51]	89.96±3.14	89.79±3.05	97.68±0.83	97.10±0.83
CBAM[52]	90.31±2.13	90.20±2.14	97.46±1.05	97.17±0.65
ECA[53]	88.19±2.66	88.74±2.52	97.00±1.42	96.74±0.74

*Taking ImageNet pre-trained ResNet50 with DAAM as Baseline in this ablation experiment.

TABLE VII
ABLATION EXPERIMENT RESULTS OF SELF-DISTILLATION AND CONTRASTIVE LEARNING STRATEGIES (%).

$\mathcal{L}_{STR}$	$\mathcal{L}_{HIS}$	$\mathcal{L}_{CACL}$	ACC	F1-score	AUC	Spe
-	-	-	86.94±4.30	87.09±4.20	97.37±1.64	96.63±1.21
✓	-	-	88.06±3.19	87.96±3.22	96.91±1.69	96.79±0.88
-	✓	-	88.57±4.51	88.34±4.69	97.48±1.45	96.75±1.33
-	-	✓	88.13±3.26	88.06±3.21	95.92±2.52	96.79±0.98
✓	✓	-	90.09±3.42	89.94±3.37	97.59±1.65	97.30±0.76
✓	-	✓	90.18±2.49	90.21±2.42	97.95±1.17	97.39±0.52
-	✓	✓	89.38±3.53	89.33±3.24	97.79±0.99	97.31±0.78
✓	✓	✓	**91.36±1.46**	**91.24±1.28**	**98.26±0.55**	**97.61±0.29**

*The first line is entirely marked with "-" to indicate the baseline, which represents the proposed MCSD-Net without DAAM.

demonstrate their complementary benefits, we conduct a series of ablation experiments. The results are summarized in TABLE VII. The first row shows the performance of the baseline model, which does not employ any of the proposed strategies, achieving an accuracy of 86.94%±4.30%. Rows two to four present the results of individually introducing each learning strategy. Each strategy alone provides a performance improvement over the baseline. Rows five to seven show the performance when combining any two of the three strategies, leading to further gains for all evaluation metrics. Finally, when three strategies are applied simultaneously, the model achieves the best results on all four metrics, accompanied by smallest standard deviations. This demonstrates that the proposed learning strategies not only enhance classification performance but also improve the stability of the model.

5) Ablation Study on Stage-wise Feature Fusion for Structural Distillation: Features extracted from shallow and deep stages of the model encode information at different levels of abstraction. TABLE VIII shows the ablation experimental results of structural distillation under various combinations of features from different stages. The results show that fusing features from all four stages yields the best performance, indicating that full-stage feature fusion effectively integrates the detailed information from shallow features with the semantic information from deep ones, thus strengthening the model's feature representation capability and stability. In contrast, models that utilize only a subset of stages yield limited performance gains, suggesting that effective structural distillation cannot be achieved by relying on information from a single shallow or deep stage alone. Consequently, the four-stage feature fusion strategy is adopted to fully leverage multi-scale information and enhance model performance.

6) Ablation Study on Growth Strategies in Historical Distillation: TABLE IX presents the ablation results for different growth strategies within the historical distillation mechanism. The linear growth strategy achieves the best performance across all four metrics, accompanied by smallest standard deviations. This suggests that its linear and continuous weight increase characteristic is more compatible with the feature iteration of historical distillation, effectively avoiding issues such as abrupt changes, excessively rapid increase or periodic oscillations observed in other strategies, thereby enhancing performance while ensuring stability. In contrast, the step growth, exponential growth, and cosine growth strategies yield slightly lower overall performance with larger variances, struggling to strike a balance between model performance and stability. These results demonstrate that the linear growth strategy is better suited to the proposed historical distillation mechanism and can effectively improve both model performance and robustness.

TABLE VIII
ABLATION EXPERIMENT RESULTS OF STAGE-WISE FEATURE FUSION FOR STRUCTURAL DISTILLATION (%).

F1	F2	F3	F4	ACC	F1-score	AUC	Spe
-	✓	✓	-	90.33±2.55	90.08±2.56	97.50±1.69	97.31±0.76
-	-	✓	✓	89.85±1.89	89.59±1.69	97.59±0.82	97.28±0.54
-	✓	✓	✓	90.25±2.81	90.08±2.84	97.53±1.10	97.50±0.83
✓	✓	✓	-	91.25±1.86	91.12±1.95	98.20±0.79	97.59±0.65
✓	✓	✓	✓	**91.36±1.46**	**91.24±1.28**	**98.26±0.55**	**97.61±0.29**

TABLE IX
ABLATION EXPERIMENT RESULTS OF GROWTH STRATEGIES IN HISTORICAL DISTILLATION (%).

Growth Strategy	ACC	F1-score	AUC	Spe
Step Growth	89.66±2.38	89.59±2.17	97.33±1.78	97.43±0.78
Exponential Growth	90.32±1.98	90.07±1.69	97.35±2.27	97.42±0.64
Cosine Growth	89.66±2.13	89.11±2.20	96.89±1.80	97.15±0.54
Linear Growth	**91.36±1.46**	**91.24±1.28**	**98.26±0.55**	**97.61±0.29**

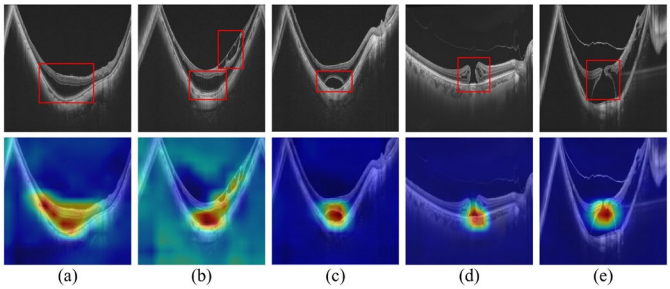

Fig. 9. Grad-CAMs visualization of different grades on the MTM dataset. (a) T1; (b) T2; (c) T3; (d) T4; (e) T5.

E. Visualization

Grad-CAM[54] is used for the feature visualization analysis of different grades in the MTM dataset, which is shown in Fig. 9. The results demonstrate that the proposed method effectively learns the salient regions in various grades images and accurately focuses on the characteristic lesion features specific to each grade.

Specifically, for the T1 grade image exhibiting an outer foveoschisis (indicated by the red box in Fig. 9(a)), the corresponding region in the heatmap shows a significant activation distribution, indicating that the model successfully identifies and concentrates on this critical region.

Authorized licensed use limited to: Soochow University. Downloaded on October 14,2026 at 07:10:33 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. 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 Journal of Biomedical and Health Informatics. 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/JBHI.2026.3718754

The T2 grade image contains both inner and outer foveoschisis. Both regions display high responses in the heatmap, confirming that the model also captures the lesion characteristics of this grade. The most typical feature of T3 grade is foveal retinal detachment. As observed in Fig. 9(c), the model's attention is highly concentrated on the area of foveal retinal detachment, enabling successful classification for this grade. For T4 and T5 grade images, the proposed model also exhibits excellent ability to focus on typical lesions. These observations confirm that our method can identify the characteristic lesions of each grade well, thereby achieving superior grading performance.

V. Discussion

In this study, a novel classification technique based on self-distillation learning was applied for retinal OCT image classification. We proposed the MCSDNet as an automatic classification model. It is designed to automate the grading and classification of retinal diseases, thereby aiding clinicians in making more efficient and standardized diagnostic assessments.

MCSD-Net achieves comprehensive performance improvement through its core components: (1) the multi-perspective collaborative self-distillation mechanism breaks single-perspective limitations by fusing multi-stage structural features and dynamically balancing historical prediction with ground truth, maximizing internal feature utilization under limited annotations; (2) the CACL strategy addresses class imbalance and high inter-class similarity by constructing class-balanced batches and fusing multi-stage features for contrastive learning, making intra-class samples more compact and inter-class samples more separable; (3) the DAAM module targets directional characteristics of retinal lesions, with DCAB capturing inter-channel dependencies and DSAB enhancing horizontal/vertical texture extraction.

Compared with state-of-the-art self-distillation methods (e.g., PS-KD, FRSKD, DTSKD) and recent OCT task-specific foundation models (e.g., RETFound, VisionFM), the advantages of MCSD-Net have been fully validated through comparison experiments. As shown in TABLE II, III and IV, comparison experiments reveal that single-dimensional focused methods such as PS-KD and FRSKD suffer from insufficient feature diversity, while DTSKD, despite integrating structural and historical, lacks targeted solutions for class imbalance and high inter-class similarity, leading to suboptimal performance. In contrast, the proposed MCSD-Net, with the synergistic effect of its three core components, outperforms other methods across accuracy, F1-score, AUC and specificity on MTM, OCT2017 and OCTDL datasets. Meanwhile, MCSD-Net achieves a favorable balance between performance and efficiency, with moderate parameters, computational complexity, training time and inference latency, and significantly outperforms recent OCT task-specific foundation models like RETFound and VisionFM in both effectiveness and practicality. Ablation experiments further confirm the necessity and effectiveness of each component. Removing any of the multi-perspective collaborative self-distillation mechanism, the CACL strategy, or the DAAM module leads to a significant decline in model performance, underscoring their indispensability in addressing the core challenges of OCT image classification. In clinical practice, the proposed model can assist ophthalmologists in efficient retinal disease classification, thereby reducing manual diagnosis burden and improving diagnostic consistency, which is crucial for early intervention and prognosis management of retinal diseases.

Although the proposed MCSD-Net has achieved superior performances in retinal OCT image classification tasks, there are still some limitations. On one hand, the DAAM primarily focuses on feature extraction in two fixed directions, horizontal and vertical dimensions. While effective in enhancing discrimination of certain lesion patterns, its representational capacity for more complex or asymmetric morphological features should be further improved. In the future work, the DAAM could be extended by exploring content-adaptive mechanisms for dynamically selecting feature extraction orientations based on input characteristics, thereby improving recognition capability for complex lesion patterns. On the other hand, although the current self-distillation mechanism integrates multi-stage features and historical information, its way of knowledge transfer is relatively fixed and fails to dynamically adjust the utilization proportion of knowledge from different sources according to the training process. Our future work will investigate dynamic and adaptive self-distillation mechanisms, which would enable the model to leverage its historical knowledge and multi-stage features at different training phases adaptively and efficiently, ultimately further enhancing feature learning and knowledge transfer efficacy. Moreover, the proposed MCSD-Net improves generalization via offline training strategies and lacks online adaptive capability to cope with multi-center induced domain shifts in practical clinical deployment. Inspired by test-time adaptation strategies[55], lightweight inference-stage adaptive modules can be explored to boost the cross-domain robustness of our model in future research.

VI. Conclusion

In this paper, we propose a multi-perspective collaborative self-distillation network named MCSD-Net for the automated classification of retinal OCT images, which includes multi-perspective collaborative self-distillation mechanism, CACL strategy and DAAM module. By integrating multi-stage feature fusion based structural self-distillation and linear growth based historical self-distillation, the multi-perspective collaborative self-distillation mechanism enhances the efficiency of utilizing the network's own features, thereby enhancing the feature learning ability of the model. The CACL strategy constructs class-balanced batches to improve the model's ability to distinguish between similar samples, thereby mitigating the challenges posed by high inter-class similarity. The DAAM performs feature extraction from both horizontal and vertical directions, enhancing the model's discriminative capability for salient features. Extensive experiments on the private MTM dataset, as well as on the public OCT2017 and OCTDL datasets, demonstrate that the proposed MCSD-Net achieves superior classification performance, outperforming other state-of-the-art self-distillation methods.

Authorized licensed use limited to: Soochow University. Downloaded on October 14,2026 at 07:10:33 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. 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 Journal of Biomedical and Health Informatics. 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/JBHI.2026.3718754

REFERENCES

[1] R. Bourne et al., "Trends in prevalence of blindness and distance and near vision impairment over 30 years: An analysis for the global burden of disease study," *The Lancet Global Health*, vol. 9, no. 2, pp. 130–143, Feb. 2021.

[2] K. Ohno-Matsui et al., "International photographic classification and grading system for myopic maculopathy," *American Journal of Ophthalmology*, vol. 159, no.5, pp. 877-883.e7, May 2015.

[3] B. Parolini et al., "The new myopic traction maculopathy staging system," *European Journal of Ophthalmology*, vol. 31, no. 3, pp. 1299–1312, May. 2021.

[4] R. Zhang and C. Wu, "Advances in classification, diagnosis and treatment of myopic maculopathy," *Yan Ke Xue Bao*, vol. 36, no. 3, pp. 219–226, 2021.

[5] J. Ruiz-Medrano, J. A. Montero, I. Flores-Moreno, L. Arias, A. García-Layana, and J. M. Ruiz-Moreno, "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, Mar. 2019.

[6] J. M. Schmitt, "Optical coherence tomography (OCT): A review," *IEEE J. Select. Topics Quantum Electron.*, vol. 5, no. 4, pp. 1205–1215, Aug. 1999.

[7] C. Mosquera, L. Ferrer, D. H. Milone, D. Luna, and E. Ferrante, "Class imbalance on medical image classification: towards better evaluation practices for discrimination and calibration performance," *European Radiology*, vol. 34, pp. 7895–7903, 2024.

[8] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 27, pp. 2–54, 2019.

[9] S. A. Kamran, S. Saha, A. S. Sabbir, and A. Tavakkoli, "Optic-Net: A novel convolutional neural network for diagnosis of retinal diseases from optical tomography images," in *18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, Dec. 2019, pp. 964–971.

[10] L. Huang, X. He, L. Fang, H. Rabbani, and X. Chen, "Automatic classification of retinal optical coherence tomography images with layer guided convolutional neural network," *IEEE Signal Process. Lett.*, vol. 26, no. 7, pp. 1026–1030, Jul. 2019.

[11] L. Fang, C. Wang, S. Li, H. Rabbani, X. Chen, and Z. Liu, "Attention to lesion: Lesion-aware convolutional neural network for retinal optical coherence tomography image classification," *IEEE Trans. Med. Imaging*, vol. 38, no. 8, pp. 1959–1970, Aug. 2019.

[12] Y. Rong et al., "Surrogate-assisted retinal OCT image classification based on convolutional neural networks," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 1, pp. 253–263, Jan. 2019.

[13] R. Rasti, H. Rabbani, A. Mehridehnavi, and F. Hajizadeh, "Macular OCT classification using a multi-scale convolutional neural network ensemble," *IEEE Trans. Med. Imaging*, vol. 37, no. 4, pp. 1024–1034, Apr. 2018.

[14] C. Playout, "Focused attention in transformers for interpretable classification of retinal images," *Medical Image Analysis*, vol. 82, p. 102608, 2022.

[15] S. Chen et al., "FIT-Net: Feature interaction transformer network for pathologic myopia diagnosis," *IEEE Trans. Med. Imaging*, vol. 42, no. 9, pp. 2524–2538, Sep. 2023.

[16] H. Wang et al., "OCTFormer: An efficient hierarchical transformer network specialized for retinal optical coherence tomography image recognition," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–17, 2023.

[17] A. Bhattarai et al, "Analyzing adjacent B-scans to localize sickle cell retinopathy in OCTs," *Medical Image Computing and Computer Assisted Intervention*, vol. 15003, pp. 574–584, 2024.

[18] L. Zhang, J. Song, A. Gao, and J. Chen, "Be your own teacher: Improve the performance of convolutional neural networks via self-distillation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3712–3721.

[19] L. Zhang, C. Bao, and K. Ma, "Self-distillation: Towards efficient and compact neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 8, pp. 4388–4403, Aug. 2022.

[20] K. Kim, B. Ji, D. Yoon, and S. Hwang, "Self-knowledge distillation with progressive refinement of targets," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 6547–6556.

[21] T. Xu and C. Liu, "Data-distortion guided self-distillation for deep neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 5565–5572.

[22] S. Ni, X. Ma, M. Zhu, X. Li, and Y.-D. Zhang, "Reverse self-distillation overcoming the self-distillation barrier," *IEEE Open J. Comput. Soc.*, vol. 4, pp. 195–205, 2023.

[23] J. Zhou, X. Dong, and Q. Liu, "Clustering-guided twin contrastive learning for endomicroscopy image classification," *IEEE J. Biomed. Health Inform*, vol. 28, no. 5, pp. 2879–2890, May. 2024.

[24] F. Zhao, Z. Wang, H. Du, X. He, and X. Cao, "Self-supervised triplet contrastive learning for classifying endometrial histopathological images," *IEEE J. Biomed. Health Inform*, vol. 27, no. 12, pp. 5970–5981, Dec. 2023.

[25] P. Khosla et al., "Supervised contrastive learning," in *Proceedings of the 34th International Conference on Neural Information Processing System*, 2020, pp. 18661–18673.

[26] U. Niyaz and D. R. Bathula, "Augmenting knowledge distillation with peer-to-peer mutual learning for model compression," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, Mar. 2022, pp. 1–4.

[27] Y. Akhter, R. Ranjan, R. Singh, and M. Vatsa, "Low-resolution chest X-ray classification via knowledge distillation and multi-task learning," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, May. 2024, pp. 1–5.

[28] Y. Wang, Y. Wang, J. Cai, T. Lee, C. Miao and Z. Wang,

Authorized licensed use limited to: Soochow University. Downloaded on October 14,2026 at 07:10:33 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. 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 Journal of Biomedical and Health Informatics. 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/JBHI.2026.3718754

"SSD-KD: A self-supervised diverse knowledge distillation method for lightweight skin lesion classification using dermoscopic images," *Medical Image Analysis*, 2023.

[29] Z. Hao et al., "One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023, pp. 79570–79582.

[30] L. Zhang, Z. Tan, J. Song, J. Chen, C. Bao, and K. Ma, "SCAN: A scalable neural networks framework towards compact and efficient models," in *NeurIPS 2019*, 2019, pp. 4029–4038.

[31] S. Yun, J. Park, K. Lee, and J. Shin, "Regularizing class-wise predictions via self-knowledge distillation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13873–13882.

[32] M. Ji, S. Shin, S. Hwang, G. Park, and I.-C. Moon, "Refine myself by teaching myself : Feature refinement via self-knowledge distillation," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10659–10668.

[33] C. Yang et al., "MixSKD: Self-knowledge distillation from mixup for image recognition," in *Computer Vision – ECCV 2022: 17th European Conference*, 2022, pp. 534–551.

[34] Z. Li, X. Li, L. Yang, R. Song, J. Yang, and Z. Pan, "Dual teachers for self-knowledge distillation," *Pattern Recognition*, vol. 151, p. 110422, Jul. 2024.

[35] Y. Hu, X. Jiang, X. Liu, X. Luo, Y. Hu, and X. Cao, "Hierarchical self-distilled feature learning for fine-grained visual categorization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 3, pp. 4005–4018, May. 2025.

[36] X. Wang et al., "Transformer-based unsupervised contrastive learning for histopathological image classification," *Medical Image Analysis*, vol. 81, p. 102559, Oct. 2022.

[37] Y. Zhang, L. Luo, Q. Dou, and P.-A. Heng, "Triplet attention and dual-pool contrastive learning for clinic-driven multi-label medical image classification," *Medical Image Analysis*, vol. 86, p. 102772, May. 2023.

[38] H. Pan, Y. Guo, M. Yu, and J. Chen, "Enhanced long-tailed recognition with contrastive CutMix augmentation," *IEEE Trans. on Image Process*, vol. 33, pp. 4215–4230, 2024.

[39] J. Yu et al., "Semi-supervised gland segmentation via feature-enhanced contrastive learning and dual-consistency strategy," *IEEE J. Biomed. Health Inform*, vol. 29, no. 6, pp. 4314–4324, Jun. 2025.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778.

[41] D. S. Kermany et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122-1131.e9, Feb. 2018.

[42] M. Kulyabin, "OCTDL: Optical coherence tomography dataset for image-based deep learning methods," *IEEE Dataport*, p. 365, Apr. 2024.

[43] Y. Zhou, M.A. Chia, S.K. Wagner et al., "A foundation model for generalizable disease detection from retinal images," *Nature*, vol. 622, pp. 156–163, 2023.

[44] J. Qiu et al., "Development and validation of a multimodal multitask vision foundation model for generalist ophthalmic artificial intelligence," *NEJM AI*, vol. 1, no.12, 2024.

[45] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5987-5995.

[46] Z. Liu, H. Mao, C. -Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11966-11976.

[47] M. Lou and Y. Yu. "OverLoCK: An overview-first-look-closely-next ConvNet with context-mixing dynamic kernels," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP*R), 2025, pp. 128-138.

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, 2017, pp. 6000–6010.

[49] Z. Liu et al, "Swin transformer: hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9992-10002.

[50] D. Shi, "TransNeXt: Robust foveal visual perception for vision transformers," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 17773-17783.

[51] J. Hu, L. Shen and G. Sun, "Squeeze-and-excitation networks," in 2*018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132-7141.

[52] S. Woo, J. Park, JY. Lee, and I.S. Kweon, "CBAM: Convolutional block attention module," in *15th European Conference on Computer Vision (ECCV 2018)*, 2018, pp. 3-19.

[53] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11531-11539.

[54] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Why did you say that? Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[55] Li, Xiang et al. "Beyond visual prompt: Unlocking parameter-efficient online adapter for continual test-time adaptation in medical image segmentation." *Inf. Fusion*, vol. 133, p. 104256, Feb. 2026.

Authorized licensed use limited to: Soochow University. Downloaded on October 14,2026 at 07:10:33 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.