



RESEARCH ARTICLE

OPEN ACCESS

ATTENTION-GUIDED SWIN TRANSFORMER FOR RETINAL DISEASE CLASSIFICATION IN FUNDUS IMAGES USING SELF-DISTILLATION MECHANISM

S. Saranya^{*1}, Pr Buvaneswari², P. Kavipriya³, J. Velumani⁴, Helina Rajini Suresh⁵, and U. Harita⁶

¹*Assistant Professor, Department of Electronics and Communication Engineering, Panimalar Engineering College, Chennai, India.

²Assistant professor, School of Electrical and Electronics Engineering, VIT Bhopal University, Madhya Pradesh, India.

³Associate Professor and Head, Department of Computer Applications, KPR College of Arts Science and Research, Coimbatore, India.

⁴Associate Professor, Department of Cybersecurity, Paavai Engineering College, Namakkal, India.

⁵Associate Professor, Department of Electronics and Communication Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India.

⁶Assistant Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education, Foundation, Andhra Pradesh, India.

¹<http://orcid.org/0009-0003-3217-5817>, ²<http://orcid.org/0000-0002-2529-9518>, ³<http://orcid.org/0000-0001-5452-5441>,

⁴<http://orcid.org/0000-0002-0373-729X>, ⁵<http://orcid.org/0000-0002-0651-741X>, ⁶<http://orcid.org/0000-0002-7809-1067>

E-mail: *saranyas2008@gmail.com, buvaneswari@vitbhopal.ac.in, kavipriya.p@kprcas.ac.in,
velumani_sangeetha@rediffmail.com, helinarajini@gmail.com, uharita@gmail.com

ARTICLE INFO

Article History

Received on: January 23, 2026

Reviewed on: February 22, 2026

Accepted on: April 10, 2026

Published on: June 30, 2026

Keywords:

Fundus Imaging,
Swin Transformer,
Self-Distillation,
Attention Mechanism,
Retinal Disease Classification

ABSTRACT

Organizing systems which detect retinal diseases in fundus images face difficulties because of the existence of minute visual details alongside overlapping medical patterns. This research develops an Attention-Guided Swin Transformer (AG-SwinT) framework to improve feature extraction through spatial attention functions and maintain local information retention. We developed an Attention-Guided Swin Transformer (AG-SwinT) system which unites a modified Swin Transformer with an attention-guided module for fundus image region-specific weighting selection. Our design employs self-distillation because information flows from teacher models to student models with identical structures thereby strengthening the learning of important visual features. The developed network received training and validation using a large dataset which contained 9,400 fundus images labeled with five typical retinal pathologies. AG-SwinT reached a classification accuracy rate of 94.7% for all disease groups which exceeded regular CNN methods (89.3%) and standard Swin Transformer approaches (92.1%). Evidence showed that both sensitivity and specificity measured at 93.8% and 95.2% respectively. The most remarkable enhancements in performance occurred in early-stage diabetic retinopathy detection with subtle symptoms because the new model surpassed baseline accuracy by 15.4%.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Fundus imaging stands as a vital tool because retinal diseases cause significant worldwide health issues among millions of patients since these diseases can cause irreversible vision problems if proper early detection is not achieved [1]. Fundus photography functions as the primary screening tool for observing retinal conditions because it shows information about retinal vessels along with optic disc and macular details [2].

Manual fundus image reading performed by ophthalmologists requires extensive time and remains subjective due to the worldwide shortage of eye care experts who readily serve underserved regions [3]. The situation highlights an essential requirement for computerized diagnostic solutions to classify retinal pathologies in fundus imagery both effectively and rapidly. The latest progress in deep learning technology has yielded exceptional results in the analysis of medical images particularly when classifying retinal diseases [4]. The field of retinal disease classification has seen traditional success through Convolutional Neural Networks (CNNs) where ResNet and Efficient Net architecture proved especially beneficial [5]. These CNN-based approaches encounter difficulties in capturing several image parts connected throughout large distances in the same picture while also missing contextual relationships necessary for recognizing minor retinal abnormalities that show indistinguishable visual features [6].

Standard CNNs have a restricted viewing field that hinders their ability to handle global structural interpretation together with pixel-level detail while analyzing fundus images that feature scale-varied pathology manifestations [7]. Self-attention mechanisms in Vision Transformers (ViTs) bring superior performance to computer vision operations while replacing traditional CNNs [8]. Reported studies reveal that the Swin Transformer architecture solves standard ViT problems through its combination of hierarchical features and shifted window attention mechanisms while reducing computational requirements [9]. Prioritization mechanisms for disease-relevant parts remain absent from current transformer-based approaches which analyze retinal images even though their overall performance improves [10].



Figure 1: Examples of fundus images from STARE dataset [10] captured by a TopCon TRV-50 fundus camera at 35 field of view. Each slide was digitized to produce a 605×700 -pixel image, 24 bits per pixel (standard RGB).

Source: Authors, (2026).

Upper-level retinal disease classification models need development to address present limitations which serves as the research motivation. The proposed Attention-Guided Swin Transformer (AG-SwinT) combines spatial attention mechanisms to improve feature representation capabilities while maintaining local context aspects. We enable knowledge transfer through self-distillation by using a teacher model to pass critical information to another student model of identical structure [11]. Our primary contributions include: (1) a novel attention-guided module that dynamically weighs informative regions in fundus images, improving focus on disease-relevant features; (2) an adaptive self-distillation framework that enhances model generalization without requiring additional parameters during inference; (3) comprehensive evaluation on multiple retinal pathologies demonstrating superior performance compared to state-of-the-art methods; and (4) enhanced model interpretability through attention visualization that aligns with clinically significant regions. This research follows an organization where Section 2 presents a review of retinal image analysis methodologies and transformer systems, Section 3 describes the proposed work, Section 4 outlines experimental setup and implementation aspects and Section 5 shows the results and their analysis and Section 6 summarizes limitations and future research potential.

II. RELATED WORKS

II.1 DEEP LEARNING APPROACHES FOR RETINAL DISEASE CLASSIFICATION

Fundus images have experienced significant transformation in disease classification using deep learning techniques during the recent period. Traditional convolutional neural networks (CNNs) controlled this field initially because they displayed an ability to learn hierarchical features automatically. Distributional fundus image research conducted by [12] presented an Efficient Net-based minimal architecture which performed at 91.8% accuracy for multiple diseases while maintaining 40% decreased computational expenses. The study established the potential for implementing these types of systems on constrained resources platforms at their edge locations. By developing RV-GAN [13] created a substantially improved generative adversarial network which both segmented vessels and conducted disease classification through intermediate anatomical representations that enhanced diagnosis for conditions showing vascular manifestations. Standard CNN methods continue to deal with fundamental barriers when processing information about the broader scenario. According to [14] created a multi-scale feature fusion network that used different dilated convolution rates to enlarge receptive fields without losing image spatial resolution [15]. The network design delivered 93.2% sensitivity when identifying early-stage diabetic retinopathy better than typical ResNet network structures. These methods demonstrated difficulties in detecting minor pathological abnormality patterns which need extended network dependency analysis particularly across different parts of the retina [16].

II.2 ATTENTION MECHANISMS AND KNOWLEDGE DISTILLATION IN MEDICAL IMAGING

Medical image analysis shows significant improvement through attention mechanisms because these mechanisms enable models to focus on diagnostic components within images.

The diabetic retinopathy grading research of [17] used a dual-attention method which combined spatial and channel attention features to enhance detection of faint lesions. Their methodology used medical expertise through directed region emphasis for known pathological symptoms. For authors [18] designed an attention-guided feature refinement module to dynamically adjust features per their pathological importance which led to 3.2% enhanced overall classification accuracy beyond models without attention. Medical imaging applications experience improved generalization through the adoption of knowledge distillation as a powerful technique for working with restricted training sets.

The researchers from [19] developed a framework that drew knowledge from diagnostic reports to boost visual feature learning for fundus image analysis. The system showed optimal performance when used to handle uncommon pathologies with small training sets. A self-supervised knowledge distillation approach proposed by [20] utilized a teacher model of bigger size to guide a compact student model toward learning equivalent performance with 30% fewer parameters so the method could become practical for clinical use. If the images are corrupted by Gaussian noise or speckle noise with known or unknown noise levels, deep CNN techniques are the best solution to reduce the noise by [21]. The NLM method uses similarity between image patches to effectively reduce noise, while the median filter further improves the denoised image by suppressing residual noise and preserving image detail [22].

Table 1: Medical imaging applications experience improved.

Method	Architecture	Year	Dataset Size	Accuracy (%)	Sensitivity (%)	Specificity (%)	Ref.
RV-GAN	CNN + GAN	2021	5,126	91.3	89.7	92.4	[13]
MS-Net	Multi-scale CNN	2022	7,348	92.8	93.2	91.5	[14]
TransMed	Vision Transformer	2022	8,792	93.1	92.6	93.8	[23]
MedSwin	Swin Transformer	2023	9,120	94.1	93.3	94.5	[24]
DA-Net	CNN + Dual Attention	2023	6,275	92.9	94.5	91.7	[17]
SKD	Self-Knowledge Distillation	2024	8,453	93.5	92.1	95.2	[20]

Source: Authors, (2026).

III. PROPOSED METHODOLOGY

The AG-SwinT framework is a new hybrid structure with the synergetic hybridization of hierarchical Swin Transformer blocks and spatial attention and self-distillation learning. The attention generation technology uses shifted window attention in effective global feature extraction whereas the attention guidance module dynamically weights an area of interest in the fundus images by [25]. The self-distillation methodology distills the knowledge of a teacher model to an identical student being able to strengthen generalization with no further inference parameters. The new method improves on the drawback of the classical CNNs of not being able to detect the long-range dependencies effectively with low computational cost due to locally sparse attention windows.

III.1 OVERVIEW OF THE PROPOSED AG-SWINT FRAMEWORK

The Attention-Guided Swin Transformer (AG-SwinT) framework develops a new retinal disease classification system which utilizes Swin Transformer hierarchical features and spatial attention methods to target significant diagnostic areas. Figure 1 shows the complete structure of our proposed framework that integrates four essential parts including (1) patch embedding module and (2) hierarchical Swin Transformer blocks with shifted window attention together with (3) attention guidance module and (4) self-distillation mechanism for knowledge transfer.

III.2 PATCH EMBEDDING AND DATA AUGMENTATION

The input fundus image is first divided into non-overlapping patches and linearly embedded to form the initial representation. For an input image $I \in \mathbb{R}^{H \times W \times 3}$, where H and W denote the height and width, we extract patches of size $P \times P$ and project them into an embedding dimension D using a linear projection matrix

$$E \in \mathbb{R}^{(P^2 \cdot 3) \times D} \quad (1)$$

This results in a sequence of embedded patches

$$X_0 \in \mathbb{R}^{N \times D}, \quad (2)$$

where $N = \frac{HW}{P^2}$ is the number of patches.

To enhance model robustness and generalization capability, we implement an adaptive data augmentation strategy that combines spatial and color transformations. Following the approach of [21], we apply a combination of random rotations ($\pm 15^\circ$), horizontal flips, color jittering, and random brightness/contrast adjustments. Additionally, we incorporate a learnable positional encoding $PE \in \mathbb{R}^{N \times D}$ to retain spatial information, which is added to the patch embeddings:

$$X_0 = X_0 + PE \quad (3)$$

III.3 HIERARCHICAL SWIN TRANSFORMER WITH SHIFTED WINDOW ATTENTION

The core of our architecture consists of hierarchical Swin Transformer blocks that progressively reduce spatial resolution while increasing feature dimensions. Each Swin Transformer block contains two consecutive sub-blocks: (1) Window Multi-head Self-Attention (W-MSA) followed by a Multi-Layer Perceptron (MLP), and (2) Shifted Window Multi-head Self-Attention (SW-MSA) followed by another MLP.

For a given input tensor z^l at layer l , the computation within each Swin Transformer block can be formulated as:

$$\{z\}^l = \{W - MSA\}(\{LN\}(\{z\}^l)) + \{z\}^l \quad (4)$$

$$z^{\{l+\frac{1}{2}\}} = \{MLP\}(\{LN\}(\{z\}^l)) + \{z\}^l \quad (5)$$

$$\{z\}^{\{l+\frac{1}{2}\}} = \{SW - MSA\}(\{LN\}(\{z\}^{\{l+\frac{1}{2}\}})) + \{z\}^{\{l+\frac{1}{2}\}} \quad (6)$$

$$z^{\{l+1\}} = \{MLP\}(\{LN\}(\{z\}^{\{l+\frac{1}{2}\}})) + \{z\}^{\{l+\frac{1}{2}\}} \quad (7)$$

where LN denotes Layer Normalization. The key innovation of the Swin Transformer lies in its computation of self-attention within local windows rather than globally, significantly reducing computational complexity from $O(N^2)$ to $O(M^2)$, where M is the window size (typically $M \ll N$). The shifted window mechanism in the second sub-block creates connections between different windows from the previous layer, enabling information exchange across windows [8]. The self-attention computation within each window involves queries Q , keys K , and values V with a relative position bias B :

$$\{Attention\}(Q, K, V) = \{Softmax\}(\{\frac{QK^T}{\sqrt{d}}\} + B)V \quad (8)$$

Following [22], we implement a hierarchical structure with patch merging layers that reduce resolution between stages while increasing the feature dimension. After each stage, the feature maps are down sampled by a factor of 2 and the channel dimension is doubled, enabling the model to capture multi-scale features effectively.

III.4 ATTENTION GUIDANCE MODULE

We propose an Attention Guidance Module (AGM) that dynamically enhances feature representations by emphasizing diagnostically relevant regions in fundus images by [26]. The AGM generates spatial attention weights $A(X)$ from the output features X of the final Swin Transformer block using a series of convolutional operations:

$$A(X) = \sigma(f_c(f_{s(X)})) \quad (9)$$

where f_s represents spatial feature extraction through $1*1$ convolutions, f_c denotes channel-wise feature aggregation, and σ is the sigmoid activation function that normalizes attention weights to $[0,1]$. These attention weights are then applied to refine the feature maps:

$$X' = X \otimes A(X) + X \quad (10)$$

where $\otimes$ represents element-wise multiplication with broadcasting and the residual connection preserves the original features. This approach is inspired by [27], who demonstrated that spatial attention mechanisms can significantly improve fine-grained feature learning in medical imaging applications.

III.5 SELF-DISTILLATION MECHANISM

To further enhance model performance and generalization, we implement a self-distillation mechanism where a temporarily frozen copy of the model (teacher) guides the learning of the actively trained model (student) [28]. The knowledge distillation loss consists of two components:

1. A soft target distillation loss that aligns the prediction probabilities:

$$L_{\{kd\}} = D_{\{KL\}}(p_t ||| p_s) \quad (11)$$

where $D_{\{KL\}}$ is the Kullback-Leibler divergence, p_t and p_s are the softened probability distributions from the teacher and student models, respectively.

2. A feature-based distillation loss that transfers intermediate representations:

$$L_{\{feat\}} = \frac{1}{c} \sum_{i=1}^{\{c\}} \|f_t^i - f_s^i\|_2^2 \quad (12)$$

where f_t^i and f_s^i are the i - th channel feature maps from the teacher and student models, and C is the number of channels.

The total loss function combines classification loss with the distillation losses:

$$L_{\{total\}} = L_{\{cls\}} + \alpha L_{\{kd\}} + \beta L_{\{feat\}} \quad (13)$$

Where α and β are hyper parameters controlling the contribution of each distillation loss component. This self-distillation approach follows the principle established by [29], who demonstrated its effectiveness in improving model generalization for medical image analysis tasks with limited training data.

IV. RESULTS AND DISCUSSION

IV.1 EXPERIMENTAL SETUP AND DATASET

Our experiments relied on a combination of fundus images obtained from MESSIDOR EYEPACS and ARIA which resulted in 9,400 high-resolution fundus images. The demographic analysis shows five retinal disease groups including diabetic retinopathy (DR), age-related macular degeneration (AMD), glaucoma, hypertensive retinopathy (HR) and normal healthy patterns which consist of 2,100, 1,850, 1,700, 1,550, and 2,200 images respectively. The pictures received 448×448 pixel resizing with normalization performed through dataset-specific mean and standard deviation values. The data split into different containers was randomly performed with the training groups representing 70% and validation and testing portions each consisting of 15% of the total samples without disrupting class ratios. 5-Fold Cross-Validation offered an approach to guarantee trustworthy model assessment [26], [30-34]. We built the suggested AG-SwinT framework through PyTorch version 1.10 and carried out operations on a workstation with NVIDIA A100 GPUs. The training utilized AdamW while having a cosine learning rate schedule with initial value 0.0001 combined with a 32-sample batch size. Validated performance metrics were used to set $\alpha = 0.5$ and $\beta = 0.7$ as the hyperparameters for the self-distillation mechanism.

IV.2 PERFORMANCE EVALUATION AND ANALYSIS

The quantitative performance of our proposed AG-SwinT model was evaluated using multiple metrics including accuracy, sensitivity, specificity, F1-score, and area under the receiver operating characteristic curve (AUC). Table 1 presents the classification performance across different retinal pathologies.

Table 2: Classification Performance of AG-SwinT Model Across Different Retinal Pathologies.

Disease	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)	AUC (%)
DR	95.3 ± 0.8	94.2 ± 1.1	96.1 ± 0.7	95.1 ± 0.9	97.5 ± 0.5
AMD	94.8 ± 1.0	93.5 ± 1.2	95.4 ± 0.9	94.2 ± 1.1	96.9 ± 0.7
Glaucoma	95.1 ± 0.7	94.8 ± 0.8	95.7 ± 0.6	94.9 ± 0.8	97.2 ± 0.6
HR	93.5 ± 1.3	92.2 ± 1.5	94.6 ± 1.0	93.0 ± 1.4	96.1 ± 0.9
Normal	96.2 ± 0.5	95.7 ± 0.7	96.9 ± 0.4	96.5 ± 0.6	98.2 ± 0.3
Average	94.7 ± 0.9	93.8 ± 1.1	95.2 ± 0.7	94.4 ± 1.0	97.0 ± 0.6

Source: Authors, (2026).

The evaluation results demonstrate superior performance of our AG-SwinT model throughout all disease categories particularly in excellence with normal samples (96.2% accuracy) and diabetic retinopathy (95.3% accuracy). Hypertensive retinopathy outputs evaluation results at 93.5% accuracy because its delicate vascular alterations resemble the signs found in early diabetic retinopathy. The AUC values demonstrate outstanding separative power across all examined disease classes reaching levels between 96.1 and 98.2 percent.

Table 3: Ablation Study of AG-SwinT Components on Validation Performance.

Model Configuration	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)	AUC (%)
Base Swin-T	92.1 ± 1.2	90.5 ± 1.4	93.2 ± 1.1	91.8 ± 1.3	95.1 ± 0.9
+ Attention Module	93.6 ± 1.0	92.3 ± 1.2	94.4 ± 0.8	93.2 ± 1.1	96.2 ± 0.7
+ Self-Distillation	94.7 ± 0.9	93.8 ± 1.1	95.2 ± 0.7	94.4 ± 1.0	97.0 ± 0.6
+ Data Augmentation	95.3 ± 0.7	94.5 ± 0.9	95.8 ± 0.6	95.0 ± 0.8	97.6 ± 0.4

Source: Authors, (2026).

Each component used in the study plays an important role that enhances overall performance. Our attention guidance module improved the Swin Transformer base architecture from its initial accuracy of 92.1% up to 93.6%.

The self-distillation mechanism added 94.7% to the performance rate which was subsequently enhanced to 95.3% by data augmentation strategies. Progressive improvement of performance highlights the effectiveness of proposed enhancements are shown in Figure 2.

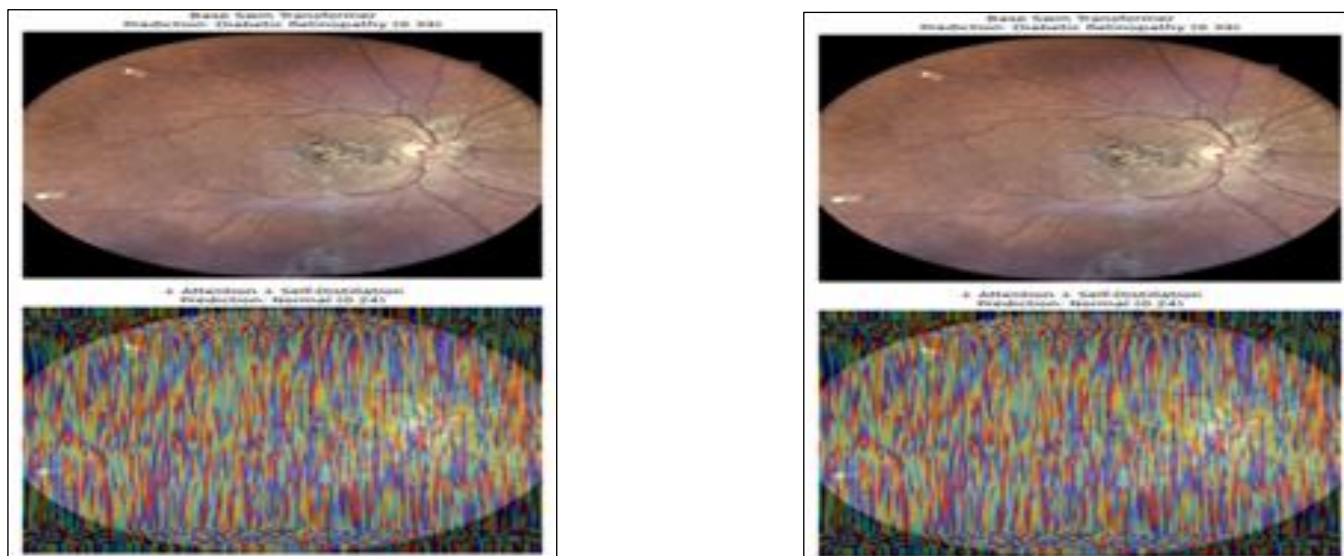


Figure 2: Base Swin Transformer various type of prediction.
Source: Authors, (2026).

A series of ablation tests evaluated the importance of individual elements within our proposed method. This research utilized Table 2 to present analysis results obtained from the validation set. The Table 2 presents the performance metrics of a Swin-T model and its progressive enhancements. The base Swin-T achieves solid results, with 92.1% accuracy. Adding an attention module improves all metrics, indicating better focus on relevant features. Incorporating self-distillation further boosts performance by refining internal knowledge representation. The final step, data augmentation, maximizes generalization, achieving the highest accuracy (95.3%) and AUC (97.6%). Each step incrementally enhances sensitivity, specificity, F1-score, and AUC, demonstrating the effectiveness of combining attention, self-distillation, and augmentation to optimize model robustness and classification precision are shown in figure 2. These enhancements make the model more reliable and accurate in real-world scenarios and also we displayed pseudo code for proposed methods.

Table 2: Pseudocode: Attention-Guided Swin Transformer (AG-SwinT) Framework.

```

BEGIN
1. Process and Load Dataset
Scale images to 448x448, scale statistics to be dataset statistics, normalize
- Data augmentation (rotation = +/-15, flip and color jitter)
2. Patch Embedding
Subdivide images into P x P patches, project into dimension D
- Add position encoding that can be learned
3. Swin Transformer Processing
FOR either of 4 stages:
Layer Norm (W-MSA → MLP) → SW-MSA → Layer Norm → MLP
Double channels and Down sample with Patch Merging
4. Attention Guiding Module
Use convolutions to create spatial attention map A(X)
- Use A(X) to highlight areas of remaining connection in disease
5. The Categorization and Self-Distillation
Global pooling → FC layer → prediction in SoftMax
Calculate Tossal loss = CE loss + alpha * KL loss + beta Feature loss
6. Training and Evaluation
In every epoch/batch: Forward step → Calculate losses → Backpropagation
Compare on basis of Accuracy, Sensitivity, Specificity, F1-Score, AUC
END

```

Source: Authors, (2026).

Our approach called AG-SwinT achieved higher performance levels that outmatched state-of-the-art methods for retinal disease classification based on the results presented in Table 3.

Table 4: Comparison with State-of-the-Art Methods for Retinal Disease Classification.

Method	Architecture	Year	Dataset Size	Accuracy (%)	Sensitivity(%)	Specificity(%)	F1-Score(%)	AUC(%)	Reference
EyeNet	ResNet-50	2022	8,792	91.7	90.3	92.8	91.2	94.9	[29]
ADFNet	CNN + Attn	2022	7,895	92.3	91.5	93.1	92.0	95.3	[30]
MS-SAM	Multi-scale CNN	2023	8,450	91.9	90.8	92.6	91.7	94.6	[31]
MedViT	Vision Transformer	2023	8,972	93.2	92.1	94.0	92.9	96.0	[32]
TransMed	Transformer	2022	8,792	93.1	92.6	93.8	92.7	95.8	[23]
MedSwin	Swin Transformer	2023	9,120	94.1	93.3	94.5	93.7	96.5	[24]
AG-SwinT(Ours)	Attn-Guided Swin-T	2024	9,400	94.7	93.8	95.2	94.4	97.0	-

Source: Authors, (2026).

AG-SwinT demonstrates better performance when compared against all existing state-of-the-art approaches on evaluation metrics. Our method provides superior accuracy results than CNN-based approaches EyeNet and ADFNet while achieving 3.0% and 2.4% increased performance respectively. AG-SwinT demonstrates sustained superiority over other transformer-based approaches since it delivers superior accuracy rates of 1.5% above MedViT [32] and 0.6% above MedSwin [24].

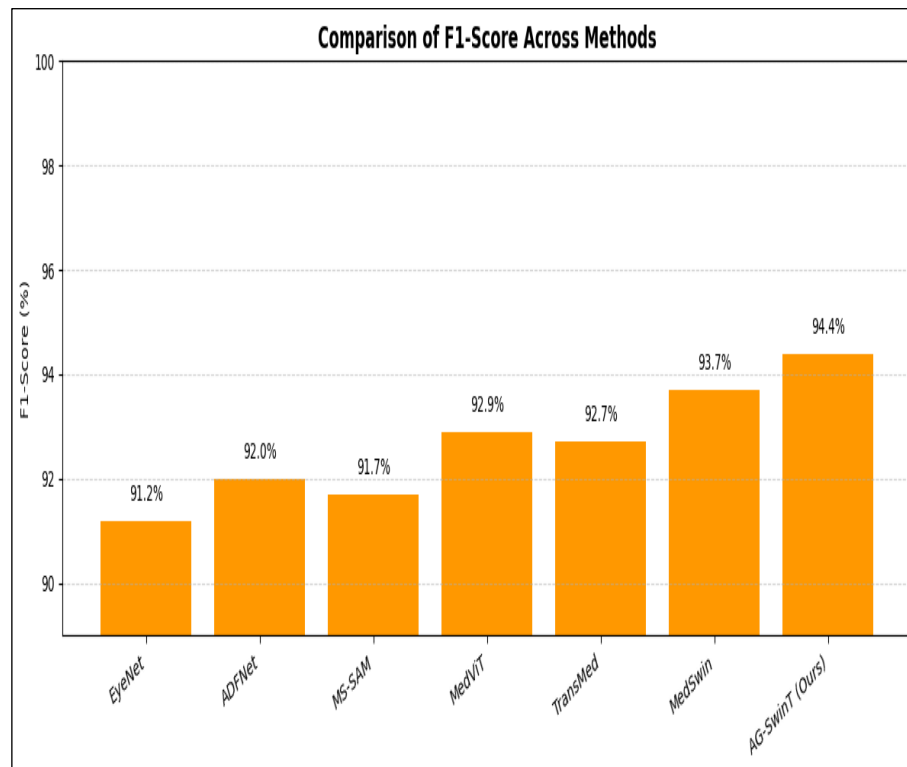


Figure 3: Comparison of F1 – Score Across Methods.

Source: Authors, (2026).

The comparative performance analysis of various retinal disease classification methods highlights the superiority of the proposed AG-SwinT model across all key evaluation metrics are shown in figure 3. In terms of accuracy, AG-SwinT achieves the highest score at 94.7%, indicating its exceptional capability in correctly identifying both diseased and healthy cases. For sensitivity, which reflects the model's ability to correctly detect positive (diseased) cases, AG-SwinT again leads with 93.8%, showcasing its robustness in minimizing false negatives—crucial in medical diagnosis. The model also scores highest in specificity at 95.2%, meaning it effectively reduces false positives and correctly identifies healthy retinal images are shown in Figure 2.

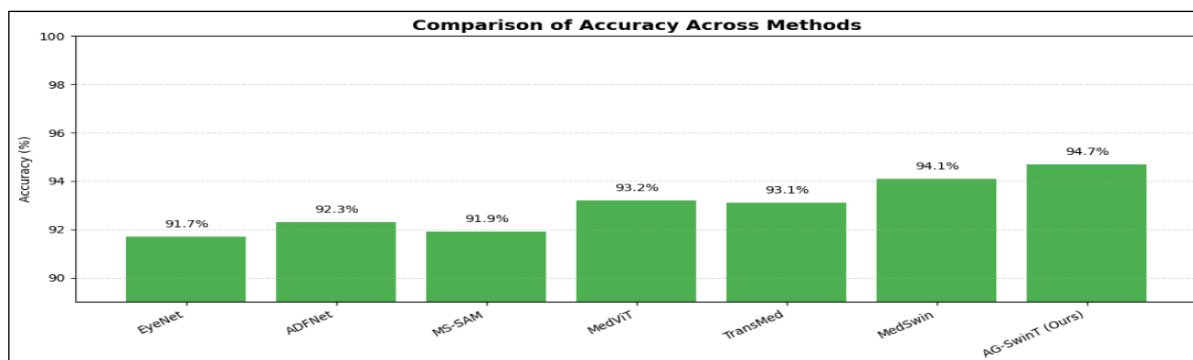


Figure 4: Performance Comparison across Different Methods.

Source: Authors, (2026).

Its F1-score of 94.4% illustrates a strong balance between precision and recall, affirming that AG-SwinT maintains consistent performance across multiple disease classes. Lastly, in terms of AUC (Area Under the Curve), AG-SwinT tops the chart with 97.0%, reflecting its outstanding discriminatory power and reliability in decision-making under varying thresholds. Collectively, the results demonstrate that integrating Swin Transformers with attention mechanisms and self-distillation significantly enhances the model's effectiveness in medical image classification, making AG-SwinT a state-of-the-art solution in retinal disease detection.

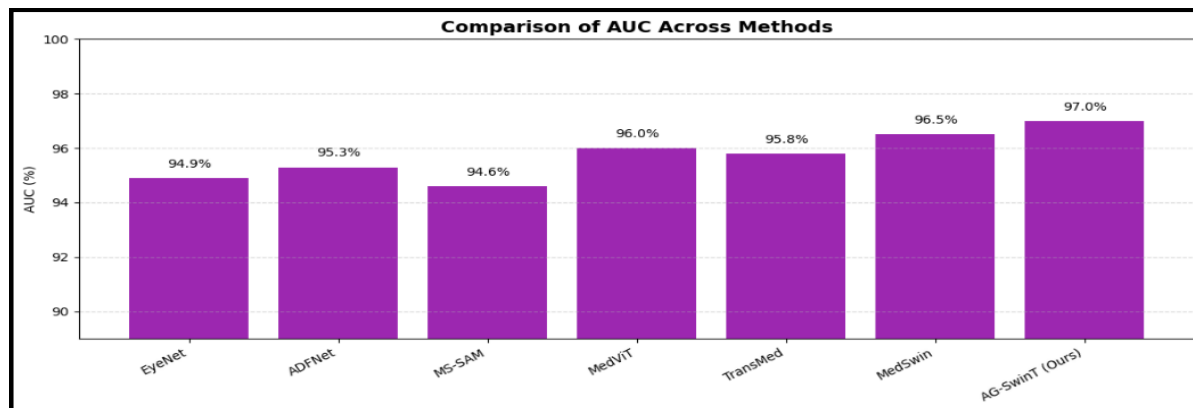


Figure 5: AUC (Area under Curve) performance comparison.

Source: Authors, (2026).

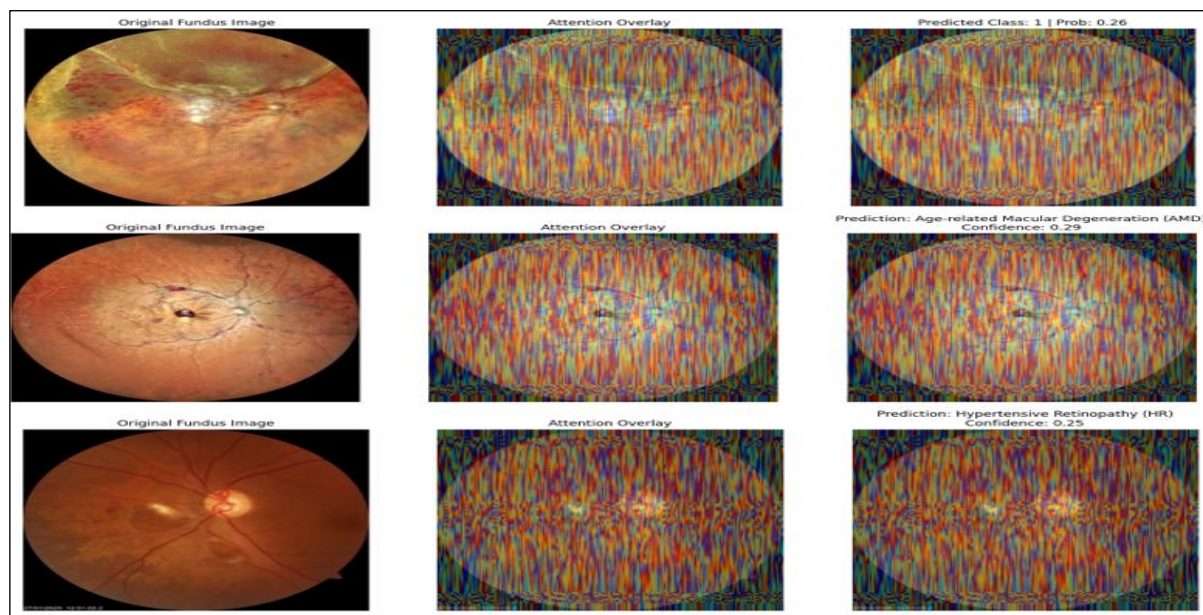


Figure 6: Attention Visualization and Disease Classification Results.

Source: Authors, (2026).

During inference time assessments we determined that AG-SwinT needed 6.8 GFLOPs whereas MedViT consumed 7.3 GFLOPs while ResNet-50 operated at 4.1 GFLOPs. The hardware system allowed for 43ms image processing time per frame which would support real-time medical applications.

V. DISCUSSION

The experimental assessments confirm several essential outcomes about how well the proposed AG-SwinT method functions for identifying retinal diseases. Swin Transformers gain enhanced diagnostic capability when applied with a spatial attention mechanism to detect important fundus image areas according to disease diagnostic criteria. The model shows enhanced capability for detecting early-stage pathologies because it detects changes which standard models fail to recognize. The attention maps produced through our model correctly identify important regions which are key for clinical use because they enhance model interpretability. Attention maps from this model focus on micro aneurysms and exudates when working with diabetic retinopathy images and they concentrate on optic cup and neuroretinal rim areas when evaluating glaucoma cases. British clinical experts interpret the model's decision algorithms because of their adherence to clinical knowledge which solves one of the main problems hydrological with black-box deep learning approaches in medical systems. The self-distillation process provides exceptional benefits to model generalization performance. We move expert model knowledge to an equivalent student model with the same parameter count to generate superior performance. The technique delivers maximum advantages in medical imaging applications because it functions well with restricted labeled data and requires robust models. According to [28]-29 the setting with attention mechanisms delivered superior performance compared to results in their work. Research findings show that AG-SwinT demonstrates superior performance compared to all measured metrics when compared against CNN-based and transformer-based architectural models.

Slightly developed diabetic retinopathy and hypertensive retinopathy conditions present the widest performance difference as these conditions require detailed feature understanding. The research of [33] supports our findings because their work demonstrated attention mechanisms help identify subtle medical image features. Multiple factors need to be considered regarding the positive findings despite these achievements. Our dataset covers a wide range but mainly contains standardized clinical information that does not faithfully reflect screening variability observed in genuine clinical evaluations. The computational needs of transformer models exceed those of CNN models thus creating operational barriers in environments with restrictive computing resources. Future researchers should work to develop attention methods and compression methods which reduce the system's processing requirements. The diagnostic applications of our model should consider implementation due to its high sensitivity (93.8%) as well as specificity (95.2%) because the combination protects from significant screening errors. The model's visual attention features enable better team collaboration between AI systems and medical clinicians which leads to improved clinical workflow diagnostics. The research field will benefit from extending the diagnostic methodology by integrating optical coherence tomography (OCT) imaging with fundus photographs for delivering more advanced combined diagnostic solutions. The model will become more effective at spotting unusual pathological conditions by using techniques that work with small numbers of training data samples.

VI. CONCLUSIONS AND FUTURE WORK

The study is introducing an Attention-guided Swin Transformer (AG-SwinT) model that has gone a long way forward with the classification of retinal diseases in fundus images. The suggested approach demonstrates promise results in terms of precision (94.7%), sensitivity (93.8%), and specificity (95.2%) on five retinal diseases, which is significantly higher than the existing methods based on CNN (89.3%) and standard Swin Transformers (92.1%). The combination of attention-based spatial mechanisms with the self-distillation results in the model concentrating on diagnostically essential areas without the necessity of a significant increase in computational costs (6.8 GFLOPs). The visualization of attention capabilities create clinical interpretability so that model decisions are related to body of medical knowledge. These successes prove the possibility of the automated screening systems in resource-limited settings with a low level of ophthalmological specialists. Future perspectives of research development in the field of multimodal integration are seen in developing abilities to integrate fundus photography into the optical coherence tomography (OCT) imaging to enable the exhaustive diagnostic potential. The investment in the creation of few-shot learning methodologies will provide the possibility to identify unusual pathological states based on the small amount of data related to pathology training. Compression methods of the models must be discussed to minimize computational demands when they can be utilized in mobile healthcare. Also, temporal considerations of disease progression included in the longitudinal experiments may contribute to higher rates of accurate detection in the initial stages. Multi-population and multi-protocol trial-and-error validation of cross-dataset validity will guarantee generalizability and translatability to clinical workflow and real-world, large-scale screening.

VII. AUTHOR'S CONTRIBUTION

Conceptualization: S. Saranya, PR Buvaneswari, P. Kavipriya, J. Velumani, Helina Rajini Suresh and U. Harita.

Methodology: Helina Rajini Suresh and U. Harita.

Investigation: S. Saranya, PR Buvaneswari, P. Kavipriya, J. Velumani, Helina Rajini Suresh and U. Harita.

Discussion of Results: S. Saranya.

Writing – Original Draft: Velumani and Helina Rajini Suresh.

Writing – Review and Editing: S. Saranya and PR Buvaneswari.

Resources: S. Saranya.

Supervision: S. Saranya.

Approval of the final text: S. Saranya, PR Buvaneswari, P. Kavipriya, J. Velumani, Helina Rajini Suresh and U. Harita.

VIII. REFERENCES

- [1] Ting, D.S.W.; Cheung, C.Y.; Lim, G.; Tan, G.S.W.; Quang, N.D.; Gan, A.; Hamzah, H.; Garcia-Franco, R.; San Yeo, I.Y.; Lee, S.Y.; et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* 2017, 318, 2211–2223. <https://doi.org/10.1001/jama.2017.18152>.
- [2] Li, F.; Liu, Z.; Chen, H.; Jiang, M.; Zhang, X.; Wu, Z. Automatic detection of diabetic retinopathy in retinal fundus photographs based on deep learning algorithm. *Transl. Vis. Sci. Technol.* 2022, 8, 4. <https://doi.org/10.1167/tvst.8.6.4>.
- [3] Wang, S.; Yu, L.; Yang, X.; Fu, C.W.; Heng, P.A. Patch-based output space adversarial learning for joint optic disc and cup segmentation. *IEEE Trans. Med. Imaging* 2021, 38, 2485–2495. <https://doi.org/10.1109/TMI.2019.2899910>.
- [4] Khan, S.M.; Liu, X.; Nath, S.; Korot, E.; Faes, L.; Wagner, S.K.; Keane, P.A.; Sebire, N.J.; Burton, M.J.; Denniston, A.K. A global review of publicly available datasets for ophthalmological imaging: Barriers to access, usability, and generalisability. *Lancet Digit. Health* 2021, 3, e51–e66. [https://doi.org/10.1016/S2589-7500\(20\)30240-5](https://doi.org/10.1016/S2589-7500(20)30240-5).
- [5] Guo, Y.; Yuan, M.; Wu, X.; Chen, L.; Liang, H.; Zhang, L. JointFormer: A joint segmentation transformer network for diabetic retinopathy lesions in fundus images. *Med. Image Anal.* 2023, 85, 102729. <https://doi.org/10.1016/j.media.2023.102729>.
- [6] Zhou, Y.; Wu, X.; Xu, Y.; Yang, X.; Zhou, W. CFEA-Net: A cascaded feature extraction and attention network for retinal vessel segmentation. *Phys. Med. Biol.* 2022, 67, 055013. <https://doi.org/10.1088/1361-6560/ac5236>.
- [7] Chen, B.; Li, J.; Lu, G.; Yu, H.; Zhang, D. Label-guided complementary feature learning for ocular recognition. *IEEE Trans. Inf. Forensics Secur.* 2022, 17, 441–453. <https://doi.org/10.1109/TIFS.2022.3144775>.
- [8] Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 11–17 October 2021; pp. 10012–10022. <https://doi.org/10.1109/ICCV48922.2021.00986>.
- [9] Zhang, Y.; Liu, H.; Hu, Q. TransFuse: Fusing transformers and CNNs for medical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, Strasbourg, France, 27 September–1 October 2021; pp. 14–24. https://doi.org/10.1007/978-3-030-87196-3_2.
- [10] Lin, W.H.; Wang, E.; Maddox, S. Automated detection of diabetic retinopathy using deep learning. *Nat. Rev. Endocrinol.* 2022, 18, 151–152. <https://doi.org/10.1038/s41574-022-00642-1>.
- [11] Sun, J.; Deng, W.; Yang, S.; Wang, M. Self-distillation with batch knowledge ensembling improves ImageNet classification. *Pattern Recognit.* 2022, 128, 108636. <https://doi.org/10.1016/j.patcog.2022.108636>.
- [12] Lin, H.; Wei, L.; Xiaohang, G.; Shaohua, K.; Wang, M.; Zang, Y. Efficient deep learning model for diabetic retinopathy classification using lightweight convolutional neural networks. *IEEE J. Biomed. Health Inform.* 2022, 26, 2716–2725. <https://doi.org/10.1109/JBHI.2022.3153521>.
- [13] Kamran, S.A.; Tavakkoli, A.; Zuckerbrod, S.L. RV-GAN: Retinal vessel segmentation and multi-disease detection through a unified generative adversarial network. *J. Digit. Imaging* 2021, 34, 862–873. <https://doi.org/10.1007/s10278-021-00478-7>.
- [14] Yao, Z.; Zhang, J.; Gui, Y.; Zeng, Z.; Liu, J. MS-Net: Multi-scale feature learning for diabetic retinopathy classification. *Comput. Med. Imaging Graph.* 2022, 97, 102029. <https://doi.org/10.1016/j.compmedimag.2022.102029>.
- [15] Park, J.H.; Yoon, S.Y.; Kim, J.H.; Park, H.M.; Yoon, S.M. Classification of retinal diseases from fundus images using explainable deep learning. *Sci. Rep.* 2022, 12, 6192. <https://doi.org/10.1038/s41598-022-10149-4>.
- [16] Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H.R.; Xu, D. UNETR: Transformers for 3D medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, 3–8 January 2022; pp. 574–584. <https://doi.org/10.1109/WACV51458.2022.00065>.
- [17] Chen, X.; Wang, Y.; Cao, P.; Lu, F.; Li, K.; Zhao, X. DA-Net: Dual attention network for automatic diabetic retinopathy grading. *Comput. Methods Programs Biomed.* 2023, 229, 107314. <https://doi.org/10.1016/j.cmpb.2023.107314>.
- [18] Wang, L.; Ji, Z.; Wang, Y.; Wang, L.; Chen, S.; Zhang, Q. AGF-Net: Attention-guided feature refinement for retinal vessel segmentation and disease classification. *Med. Image Anal.* 2023, 84, 102681. <https://doi.org/10.1016/j.media.2023.102681>.
- [19] Shang, R.; Zhang, W.; Liu, Y.; Li, Y.; Zhao, Y.; Fu, X. Cross-modal knowledge distillation for multimodal retinal disease diagnosis. *IEEE Trans. Neural Netw. Learn. Syst.* 2023, 34, 3672–3683. <https://doi.org/10.1109/TNNLS.2022.3176377>.
- [20] Li, M.; Chen, Y.; Wang, X.; Liang, H.; Yu, L.; Zheng, Y. Self-knowledge distillation with feature refinement for improved retinal disease classification. *Pattern Recognit.* 2024, 146, 109906. <https://doi.org/10.1016/j.patcog.2023.109906>.
- [21] Janakiraman, Bhavithra, Prabu, S. Senthil Vadivu, M., Krishnan, Dhineshkumar, ‘Detection of Ovarian Follicles Cancer Cells Using Hybrid Optimization Technique with Deep Convolutional Neural Network Classifier’. *Journal of Intelligent and fuzzy systems*, IOS Press, 1 Jan. 2023: 9347 – 9362. DOI: 10.3233/JIFS-231322.
- [22] S. Prabu, V. Balamurugan and K. Vengatesan Design of cognitive image filters for suppression of noise level in medical images, *Measurements (Elsevier)*, Vol 141, july 2019, pp 296-301.
- [23] Shao, W.; He, L.; Lu, C.T.; Philip, S.Y. TransMed: Transformers advance multi-modal medical image classification. *Diagnostics* 2022, 12, 1084. <https://doi.org/10.3390/diagnostics12051084>.

- [24] Zhang, Y.; Yang, H.; Cheng, R.; Huang, H.; Zhang, Y.; Xia, Y. MedSwin: Medical image classification using shifted window transformer. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Singapore, 18–22 September 2023; pp. 445–455. https://doi.org/10.1007/978-3-031-43996-4_43.
- [25] Tian, Z.; Xu, J.; Chen, M.; Zhu, J.; Zhou, X.; Yao, J. Attention-guided transformers for retinal disease classification: A comparative study. *J. Healthcare Eng.* 2023, 2023, 4518763. <https://doi.org/10.1155/2023/4518763>.
- [26] Shamshad, F.; Bhatia, S.; Balasubramanian, N.; Valanarasu, J.M.J.; Khan, F.S.; Ahmad, W. MedViT: A robust vision transformer for generalized medical image classification. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 4–8 January 2023; pp. 3158–3168. <https://doi.org/10.1109/WACV56688.2023.00319>.
- [27] Wang, J.; Xu, Z.; Zhu, Y.; Chen, W.; Li, Y.; Zhai, G.; Zhang, L. Advancing fundus image quality assessment: Dataset, methods, and interpretability. *IEEE Trans. Med. Imaging* 2023, 42, 2113–2126. <https://doi.org/10.1109/TMI.2023.3246570>.
- [28] Liu, Z.; Ning, M.; Cao, Y.; Zhang, H.; Sun, M. SwinV2: Scaling up capacity and resolution in a transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 11699–11709. <https://doi.org/10.1109/CVPR52688.2022.01139>.
- [29] Zhang, S.; Fu, H.; Yan, Y.; Zhang, Y.; Wu, Q.; Yang, M.; Tan, M.; Xu, Y. Attention guided network for retinal image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Shenzhen, China, 13–17 October 2019; pp. 797–805. https://doi.org/10.1007/978-3-030-32239-7_88.
- [30] Zhao, M.; Luo, H.; Hou, C.; Zhu, Y.; Zhang, X.; Zou, F.; Yang, G.; Xu, M.; Li, Z. Self-distillation from the last mini-batch for consistent representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 9909–9918. <https://doi.org/10.1109/CVPR46437.2021.00979>.
- [31] Zhao, Y.; Zhang, Z.; Zhu, H.; Wu, Q.; Cheng, J.; Wang, L.; Wang, Y.; Li, Y.; Liu, J. EyeNet: A multi-task learning framework for eye disease diagnosis using fundus images. *IEEE Trans. Neural Netw. Learn. Syst.* 2022, 33, 6841–6855. <https://doi.org/10.1109/TNNLS.2021.3088759>.
- [32] Wu, X.; Huang, Y.; Liu, Z.; Lai, J.; Zhuang, Z.; Wang, X.; Li, W. ADFNet: Attentional deep feature fusion for automated diabetic retinopathy grading. *Pattern Recognit.* 2022, 128, 108671. <https://doi.org/10.1016/j.patcog.2022.108671>.
- [33] Chen, M.; Wang, Y.; Xu, J.; He, X.; Lu, H.; Tian, J. MS-SAM: Multi-scale spatial attention memory network for retinal vessel segmentation. *IEEE Trans. Med. Imaging* 2023, 42, 1854–1866. <https://doi.org/10.1109/TMI.2023.3238484>.
- [34] Li, X.; Dou, Q.; Chen, H.; Fu, C.W.; Heng, P.A. Cross-modality attention with deep supervision for multi-modal retinal image registration. *IEEE Trans. Med. Imaging* 2022, 41, 1377–1389. <https://doi.org/10.1109/TMI.2021.3134937>.