Collaborative Multiscale Representation Learning for AMD Detection

Nayeon Jang[1], Seonghyeon Ko[1], Junghyun Bum[2], Duc-Tai Le[1], Chang-Hwan Son[3], and Hyunseung Choo[4*]

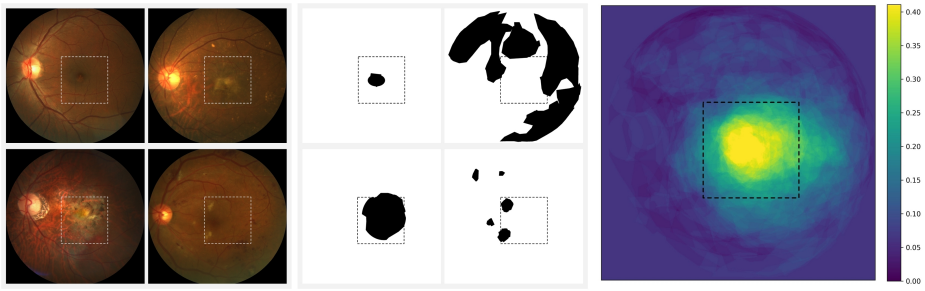
[1] Department of AI Systems Engineering, Sungkyunkwan University, Suwon, South Korea
{nayeonjang, shko0215, ldtai}@skku.edu
[2] Convergence Research Institute, Sungkyunkwan University, Suwon, South Korea
bumjh@skku.edu
[3] Department of Software Science & Engineering, Kunsan National University, Gunsan, South Korea
cson@kunsan.ac.kr
[4] Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon, South Korea
choo@skku.edu

Abstract. Age-related Macular Degeneration exhibits heterogeneous lesion morphologies and spatial distributions across global fundus and local macular region. Previous studies compress disparate lesion features into single scale representation, leading to representational bottleneck which dilutes peripheral features and allows extensive lesions to dominate subtle ones. To address this challenge, we propose a Collaborative Multi-scale Representation (CoMRep) learning. Dual-branch feature extractor produces multiscale features for global fundus and local macula. The subsequent Global-local Cross-scale Attention projects each feature into multiple latent representations, yielding diverse perspectives of lesion features with cross-scale attention. Collaborative Glocal Feature Aggregation utilizes perspective experts to refine diverse glocal features and integrate for final prediction. Experiments on the ADAM dataset demonstrate that the proposed CoMRep exceeds 0.73% AUC, 5.10% accuracy, and 3.19% recall compared to the previous state-of-the-art method with the same experimental setting. Qualitative evaluation further indicates that CoMRep captures heterogeneous AMD lesions with multiple glocal feature mixture. The code is available at https://github.com/jny0812/CoMRep.

Keywords: Age-related macular degeneration · Multi-scale learning · Cross-scale attention · Mixture-of-Experts.

1 Introduction

Age-related Macular Degeneration (AMD) is a leading cause of blindness, requiring early detection [1–3]. Recent advances in deep learning mitigate the large-



(a) AMD fundus images (b) Lesions masks (c) Lesion frequency heatmap

Fig. 1: Spatial distributions of AMD lesions on (a) fundus images and (b) corresponding lesion masks. (c) Lesion occurrence frequency heatmap aggregated over the dataset, showing the macula with dashed box.

scale screening burden, however, robust AMD detection is still challenging due to heterogeneous AMD lesions (e.g., drusen, pigmentary abnormalities) [4]. AMD lesions are diverse in morphology, boundary clarity, and spatial extent [5], as represented in Fig. 1 with fundus images 1(a) and corresponding lesion masks 1(b) of ADAM dataset [6]. The lesion frequency map 1(c) further indicates that lesions often extend beyond the macula. Previous single-scale approaches compress these characteristics into a single vector representation, leading to a representational bottleneck. This compression collapses lesion location and distribution into a global statistic, attenuating peripheral/off-macula features. In multi-lesion cases, high-saliency regions dominate the pooled representation, suppressing subtle lesions and biasing detection toward a single prominent phenotype.

Previous studies typically formulate AMD detection as single-scale image classification and improve performance by scaling either computation or supervision [6]. ForbiddenFruit and Zasti_AI boost accuracy with heavy ensembling over standard backbones, where the downstream decision still relies on a single-scale compressed features. This dilutes peripheral/off-macula signals and leads dominant large-area abnormalities to overshadow subtle lesions. VUNO EYE TEAM leverages pixel-level annotations across 15 lesion types [7] to enhance sensitivity for fine pathologies, however, dense lesion supervision incurs substantial annotation cost. Foundation models reduce annotation dependency by large-scale pretraining, yet downstream adaptation still employs a single-scale representation [8]. Previous approaches struggle to encode heterogeneous global-to-local AMD lesion features in one representation, especially off-macula distribution and subtle small-scale lesions under multi-lesion co-occurrence.

To address these challenges, we propose Collaborative Multi-scale Representation (CoMRep) learning to effectively integrate global fundus and local macula features. CoMRep initially yields AMD lesion features in global fundus and local macular region by dual branch feature extractor. Multi-scale features are divided into multiple latent representations and propagated to Global-local Cross-scale

Attention (GCA). GCA enables bidirectional interactions between global and local scale features, producing diverse glocal perspectives of AMD lesion features. Subsequent Collaborative Glocal Feature Aggregation (CGFA) employs a mixture-of-perspective experts to explicitly capture lesion heterogeneity while specializing in different lesion patterns. CoMRep finally aggregates the miscellaneous glocal representations into a unified one for robust AMD detection. The main contributions are as follows:

1. To the best of our knowledge, this study is the first to formulate AMD detection as a multi-scale problem, capturing heterogeneous lesion features across global fundus and local macula.
2. The proposed GCA performs bidirectional cross-attention among multiple global and local features to produce diverse perspectives of glocal representations while retaining complementary features across scale and location.
3. The proposed CGFA is a perspective-aware glocal feature aggregation method, collaboratively integrating disparate glocal representations into a single robust representation in an efficient manner.
4. The proposed CoMRep outperforms 4.5% accuracy and 3.2% recall and achieves the second-best AUC, compared to the previous state-of-the-art methods. Notably, CoMRep attains the top AUC under the same training dataset, without relying on additional data or dense annotations.

2 Methods

2.1 Multi-scale Feature Extraction

AMD-related macular lesions typically emerge around the fovea and exhibit marked variability in size and morphology across individuals [10]. CoMRep explicitly extracts a local macular RoI to preserve subtle lesion manifestations attenuated when using only full-field fundus images. Given a fundus image $I \in \mathbb{R}^{H\times W\times 3}$ and an annotated fovea coordinate (x, y), a square RoI centered at the fovea is defined. The side length s is $\rho \cdot \min(H, W)$ with a fixed length ρ. Accordingly, the RoI is specified by the top-left and bottom-right corners (x_1, y_1) and (x_2, y_2), respectively. We set ρ based on clinical observation, which the macular region typically spans approximately 5-5.5 mm and corresponds to roughly 30-40% of the retinal diameter in standard fundus photographs [11–13].

The dual-branch feature extraction module consists of a global feature extractor $F_g(\cdot)$ and a local feature extractor $F_l(\cdot)$, which take the global fundus image I_g and the cropped macular image I_l as input, respectively. The global feature vector $h_g = F_g(I_g)$ captures broadly distributed pathological context (e.g., widespread drusen), whereas the local feature vector $h_l = F_l(I_l)$ emphasizes fine-grained lesion patterns concentrated in the central macula. However, naively aggregating h_g and h_l through concatenation tends to collapse the fused embedding into a relatively homogeneous subspace, limiting its ability to represent heterogeneous lesion phenotypes. This representational bottleneck requires a mechanism that supports diverse interpretations of global and local features.

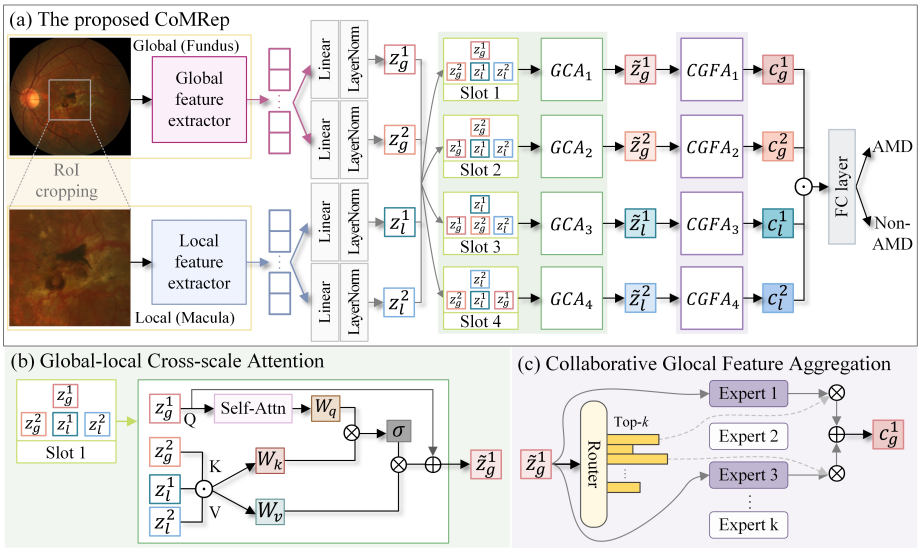


Fig. 2: Overall architecture (a) of the proposed CoMRep. The GCA and CGFA are illustrated in (b) and (c).

2.2 Global-local Cross-scale Attention

The GCA generates $2M$ multi-latent representations z_α^β by projecting each h_α into M latent subspaces to overcome limited dimension of single-vector fusion, where α and β denote the scale and representation indices, respectively:

$$z_\alpha^\beta = W_\alpha^\beta h_\alpha + b_\alpha^\beta, \quad \alpha \in \{g, l\}, \quad \beta \in \{1, 2, \ldots, M\}. \tag{1}$$

The learnable parameters W_α^β and b_α^β represent weights and bias, respectively.

GCA processes a four-step cross-attention with z_α^β, where a slot i selects one representation as the query and uses the remaining representations as references. z_i denotes the query token selected from $\{z_\alpha^\beta\}$ at step i, and $[z_j]_{j\neq i}$ denotes the concatenated reference tokens. z_i applies self-attention (SA) to emphasize its feature and suppress noise, generating a refined query q_i:

$$q_i = SA(z_i), \quad k_i = [z_j]_{i\neq j}, \quad v_i = [z_j]_{i\neq j}, \quad i \in \{1, 2, 3, 4\}, \tag{2}$$

GCA changes z_i assignment at each step and generates $\tilde{z}_\alpha^\beta$ through cross-attention over the remaining representations as references to refine each representation with cross-scale features, where d denotes the feature dimension:

$$\tilde{z}_\alpha^\beta = z_i + \text{softmax}\left(\frac{q_i k_i^\top}{\sqrt{d}}\right) v_i, \tag{3}$$

Repeating this procedure over four slots yields mutually complementary refinements of the multi-latent representations.

2.3 Collaborative Glocal Feature Aggregation

Top-k Router CGFA assigns the refined $\tilde{z}_\alpha^\beta$ a mixture of perspective experts with learned weights to accommodate heterogeneous lesion characteristics. The router computes expert logits g_α^β over E experts with given $\tilde{z}_\alpha^\beta$, where W_r and b_r are learnable router parameters:

$$g_\alpha^\beta = W_r \tilde{z}_\alpha^\beta + b_r, \tag{4}$$

CGFA activates only the Top-k experts according to g_α^β to improve computational efficiency and enforce selective aggregation. The router converts the selected logits into routing weights $\omega_{\alpha,\beta}$:

$$\omega_{\alpha,\beta} = \mathrm{softmax}\big(g_\alpha^\beta \mid \varepsilon_\alpha^\beta\big), \tag{5}$$

where, ε_α^β denotes the index set of the Top-k experts selected from g_α^β, and the softmax is normalized only within ε_α^β. Experts outside ε_α^β receive zero weight, and the selected weights sum to one.

Perspective expert network Each expert consists of a two-layer feed-forward network with GELU activation function. Given $\omega_{\alpha,\beta}$, CGFA aggregates the outputs of the selected experts as $m_{\alpha,\beta}$:

$$m_{\alpha,\beta} = \mathrm{GELU}\Big(W^{(2)}\,\mathrm{GELU}\Big(W^{(1)}\tilde{z}_\alpha^\beta + b^{(1)}\Big) + b^{(2)}\Big), \tag{6}$$

where $W^{(1)}$ and $W^{(2)}$ indicate the weights learned for each selected expert, and $b^{(1)}$ and $b^{(2)}$ are the corresponding bias vectors. Each expert produces its latent perspective representation by applying a GELU-based nonlinear transformation. This mechanism allows CGFA to expand representational capacity by selectively activating experts that match different combinations of lesion appearance.

The outputs of the Top-k selected experts are aggregated by the routing weights to form a collaborated representation c_α^β:

$$c_\alpha^\beta = \sum_{e \in \mathcal{E}_\alpha^\beta} \omega_{\alpha,\beta}\, m_{\alpha,\beta}, \tag{7}$$

c_α^β subsequently concatenated to construct a unified glocal feature vector c, which is fed into a classification head to produce the final AMD probability $\hat{y} = \sigma(Wc + b)$, where σ denotes the sigmoid function.

2.4 Loss Function

The proposed CoMRep detects AMD for each sample i. We utilize the Binary Cross-Entropy between the predicted probability $\hat{y}_i^k$ and the ground-truth y_i^k:

$$L_{cls} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_i^k \log\left(\hat{y}_i^k\right), \tag{8}$$

where N denotes the batch size and K is the number of classes. This direct supervision encourages the $2M$ glocal representations for accurate AMD detection.

We add an auxiliary load-balancing loss L_{aux} to prevent routing collapse into a few experts in CGFA:

$$L_{aux} = \frac{1}{B \cdot L} \sum_{b=1}^{B} \sum_{l=1}^{L} \left(E \cdot f_e - 1\right)^2, \tag{9}$$

where B, L, and E denote the numbers of blocks, layers, and experts, respectively. f_e is the average selections of expert e. L_{aux} penalizes deviations of the actual expert utilization from the uniform distribution.

The final loss function consists of a weighted sum of L_{cls} and L_{aux}, where λ is a weighting factor for L_{aux}:

$$L_{total} = L_{cls} + \lambda L_{aux}, \tag{10}$$

3 Experiments

3.1 Experimental Setup

The Automatic Detection challenge on Age-related Macular degeneration (ADAM) dataset contains 1,200 triplets of fundus image, AMD label and fovea coordinates collected at the Zhongshan Ophthalmic Center (Sun Yat-sen University) [6]. The dataset comprises two different image resolutions, 824 images with $2,124 \times 2,124$ pixels from Zeiss visucam 500 and 376 images with $1,444 \times 1,444$ pixels from Canon CR-2. The dataset is divided into training, validation and test sets each with 400 images. We adopt Contrast Limited Adaptive Histogram Equalization to all images to reduce illumination imbalance and enhance for heterogeneous AMD lesions. We utilize a 16-fold augmentation strategy to increase data diversity and improve model generalization [9]. All input images are resized to 224×224 and undergo normalization. The proposed method is trained for 150 epochs with batch size 16 and Adam optimizer with a learning rate of 1×10^{-5}. The method is implemented with PyTorch on a hardware environment with an Intel i7 CPU, 64GB RAM, and NVIDIA RTX 2060 GPU.

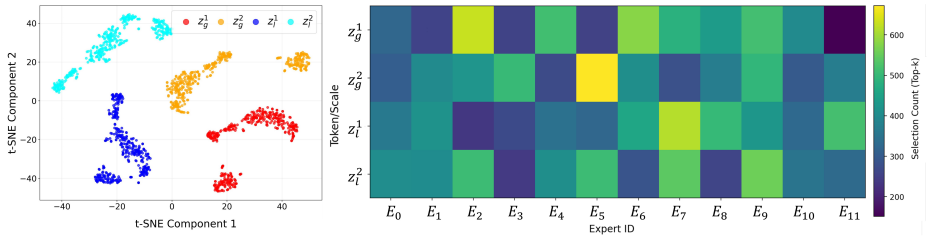
Existing AMD detection methods [6, 8] structurally rely on single-scale representations. We further compare the proposed CoMRep with general classification models [14–17] to isolate the benefit of multi-scale design from backbone-specific factors. The performance evaluation utilizes Area Under the ROC Curve (AUC), Accuracy (Acc) and Recall.

3.2 Experiment Results

The proposed CoMRep achieves Acc 5.10% and Recall 3.19% improvements, and the second-best AUC with 0.26% discrepancy compared to the state-of-the-art methods, as shown in Table 1. Under the same protocol without additional

Table 1: Performance comparison of AMD detection on the ADAM dataset. The best and second-best results are highlighted in **bold** and underlined, respectively.

Scale	Model	Add. Data	Method (Model)	AUC	Acc	Recall
Single	AMD	O	VUNO EYE TEAM [6]	**0.9714**	–	–
			ForbiddenFruit [6]	0.9592	–	–
			Muenai_Tim [6]	0.9399	–	–
		X	RETFound+logistic regression [8]	0.9619	0.8825	0.8988
			Zasti_AI [6]	0.9581	–	–
	Non-AMD	X	ResNet34 [14]	0.9121	0.8800	0.7528
			SwinT [15]	0.9106	0.8875	0.7640
			ConvNeXt-Tiny [16]	0.8788	0.8625	0.5843
			ViT-Base [17]	0.9059	0.8875	0.6517
Multi	AMD	X	**Ours(CoMRep)**	0.9689	**0.9275**	**0.9275**



(a) t-SNE (b) Top-k expert selection

Fig. 3: Analysis of heterogeneous representations and expert routing. (a) t-SNE visualization of four global–local latent representations, showing separable embedding structures across scales. (b) Expert selection frequency (Top-k) across multi-scale tokens, highlighting non-uniform routing behavior.

data, CoMRep further enhances AUC by 0.73% compared to the second-best method. These results indicate that the proposed method demonstrates strong data efficiency and robust generalization by accommodating heterogeneous lesion patterns, rather than relying on increased training data volume. CoMRep surpasses performance over standard single-scale baselines by up to AUC 10.25%, Acc 7.54%, and Recall 58.74%. The empirical gains highlight the effectiveness of collaborative multi-scale features.

We qualitatively analyze the proposed method to identify that each module operates as intended. The multi-latent features global and local scale distribute along different directions across scales and within each scale in Fig. 3(a). This disparate distribution demonstrates that the multi-latent representations captures the heterogeneous AMD lesion features. The subsequent GCA combines cross-scale features, yielding diverse glocal perspective of AMD lesions. CGFA takes glocal features and refines them with proper perspective experts, as shown in Fig. 3 (b). The heatmap depicts that the routing patterns over expert combinations. The global representations $\tilde{z}_g^1$ and $\tilde{z}_g^2$ undergo on small subset of experts.

Table 2: Performance under different numbers of experts and Top-k settings.

Metric	8 Experts		12 Experts		16 Experts	
	k=1	k=2	k=1	k=2	k=1	k=2
AUC	0.9521	0.9594	0.9642	**0.9689**	0.9243	0.9462
Accuracy	0.9125	0.9275	0.92	**0.9275**	0.87	0.91
F1-score	0.8166	0.8324	0.9156	**0.9267**	0.8389	0.8451
Recall	0.809	0.8571	0.9159	**0.9275**	0.7912	0.8539

Table 3: Ablation results for GCA and CGFA.

Branch	GCA	CGFA	AUC
Single	–	–	0.9175
Dual	–	–	0.9234
Dual	O	–	0.9372
Dual	–	O	0.9450
Dual	O	O	**0.9689**

Table 4: AUC performance according to macular crop size.

Crop size(%)	31	32	33	34	**35**	36	37	38	39
AUC	0.9603	0.9594	0.9615	0.9664	**0.9689**	0.9656	0.9579	0.9525	0.9586

The local representations $\tilde{z}_l^1$ and $\tilde{z}_l^2$ exhibit distributed selections across multiple experts. The macular-focused glocal features require various perspectives.

Table 2 reports performance under different numbers of experts and Top-k routing settings. Performance improves across all metrics when increasing k at a fixed E. The $E = 12$ setting shows notable gains of 0.47% in AUC and 1.11% in F1-score, mitigating single-path dependency. The best overall performance at $E = 12, k = 2$ surpasses the second-best result by 0.49%, 0.82%, 1.21%, 1.27% in AUC, Accuracy, F1-score, and Recall, respectively. Increased experts $E = 16$ degrade performance, yielding 2.27% AUC and 7.36% Recall lower than the best configuration. This indicates that overly dispersed routing reduces learning efficiency.

The Dual-branch baseline improves AUC by 0.59% over the Single-branch, but the gain remains limitation of insufficient alignment of cross-scale representations in Table 3. Dual-branch with GCA and CGFA indicate AUC 1.38% and 2.16% enhancements, respectively. CGFA yields the larger gain by fusing features through perspective-based expert selection. CoMRep achieves the best performance, improving AUC by 4.95% over the Dual-branch baseline and by 5.14% over the Single-branch with collaborative refinement under heterogeneous lesion patterns. Finally, Table 4 identifies a 35% macular RoI crop ratio as the best setting, achieving AUC gains of 0.89% over the average across the tested range and 1.72% over the worst setting; we therefore use 35% as the default crop size in subsequent experiments.

4 Conclusion

This study addresses the varying contributions of global context and macula-centered local features under AMD lesion heterogeneity, which makes prediction unstable. Simple feature combination or inadequate cross-scale fusion can further dilute lesion patterns. To address these challenges, we propose a collaborative multi-scale representation learning framework. A dual-branch design

extracts global and local features, refines them into collaborative multi-latent representations through GCA, and aggregates them using a mixture of perspective experts in CGFA. The proposed method achieves the best AUC among methods trained under the same protocol without additional data, achieving the second-best AUC, the best accuracy and recall. Although the proposed CoMRep highlights its efficient competitiveness, the multicenter validation is necessary to demonstrate its generalization capability. Future work will also integrate dynamic fovea detection which is not dependent on the fovea coordinates, further strengthen CoMRep applicability in the real-world clinical settings.

Acknowledgments. This work was supported in part by the Korea government (MSIT), IITP, Korea, under IITP-2026-RS-2020-II201821 (60%) and RS-2019 II190421 (10%); and by the MSS, Korea, under the Starting Growth Technological R&D Program (TIPS Program, RS-2024-00514724, 30%). Professors D.T. Le and H. Choo are with SKAI X Inc.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Wong, W.L., Su, X., Li, X., et al.: Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. The Lancet Global Health **2**(2), e106–e116 (2014)
2. Vision Loss Expert Group of the Global Burden of Disease Study, and the GBD 2019 Blindness and Vision Impairment Collaborators: Global estimates on the number of people blind or visually impaired by age-related macular degeneration: a meta-analysis from 2000 to 2020. Eye **38**, 2070–2082 (2024). https://doi.org/10.1038/s41433-024-03050-z
3. Gervelmeyer, J., et al.: Interpretable-by-Design Deep Survival Analysis for Disease Progression Modeling. In: Linguraru, M.G., et al. (eds.) Medical Image Computing and Computer Assisted Intervention-MICCAI 2024. LNCS **15010**. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72117-5_47
4. Peng, Y., et al.: DeepSeeNet: A deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs. Ophthalmology 126(4), 565–575 (2019)
5. Yao, H., et al.: Weakly supervised lesion localization of nascent geographic atrophy in age-related macular degeneration. In: Greenspan, H., et al. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2023. LNCS, vol. 14220, pp. (2023). Springer, Cham. https://doi.org/10.1007/978-3-031-43907-0_46
6. Fang, H., et al.: ADAM challenge: Detecting age-related macular degeneration from fundus images. IEEE Trans. Med. Imaging 41(10), 2828–2847 (2022)
7. Son, J., Shin, J.Y., Kim, H.D., Jung, K.-H., Park, K.H., Park, S.J.: Development and validation of deep learning models for screening multiple abnormal findings in retinal fundus images. Ophthalmology **127**(1), 85–94 (2020)
8. Sinichi, P., Bernabeu, M.O., Javidi, M.: Comparison of Classical and Deep Learning-Based Feature Representations for Age-Related Macular Degeneration.

In: Su, R., Frangi, A.F., Zhang, Y. (eds.) *Proceedings of 2024 International Conference on Medical Imaging and Computer-Aided Diagnosis (MICAD 2024)*. Lecture Notes in Electrical Engineering, vol. 1372, pp. 505–514. Springer Nature Singapore (2025). https://doi.org/10.1007/978-981-96-3863-5_46

9. Chakraborty, R., Pramanik, A.: DCNN-based prediction model for detection of age-related macular degeneration from color fundus images. Medical & Biological Engineering & Computing **60**(5), 1431–1448 (2022)
10. Ferris III, F.L., Wilkinson, C.P., Bird, A., Chakravarthy, U., Chew, E., Csaky, K., Sadda, S.R.: Clinical classification of age-related macular degeneration. Ophthalmology **120**(4), 844–851 (2013)
11. Carl Zeiss Meditec AG: VISUCAM NM/PRO Fundus Camera. Hanson Instruments (product page). Available: https://www.hansoninstruments.co.uk/zeiss-visucam-nm-pro-fundus-camera (accessed 18 Oct 2025)
12. Sampson, D.M., et al.: Towards standardizing retinal optical coherence tomography angiography: a review. Light: Science & Applications **11**(1), 63 (2022)
13. Tran, K., Mendel, T.A., Holbrook, K.L., Yates, P.A.: Construction of an inexpensive, hand-held fundus camera through modification of a consumer "point-and-shoot" camera. *Investigative Ophthalmology & Visual Science* **53**(12), 7600–7607 (2012). https://doi.org/10.1167/iovs.12-10449
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
15. 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), 10012–10022 (2021)
16. Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11976–11986 (2022)
17. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations (ICLR) (2021)
18. Jiang, Y., Shen, Y.: M4oE: A Foundation Model for Medical Multimodal Image Segmentation with Mixture of Experts. In: Linguraru, M.G., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. LNCS **15012**. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72390-2_58