A Region-Based Approach to Diabetic Retinopathy Classification with Superpixel Tokenization

Clément Playout[1,2][0000−0002−0978−3588], Zacharie Legault[1][0009−0003−5817−1830], Renaud Duval[2][0000−0002−3845−3318], Marie Carole Boucher[2][0000−0003−1988−3512], and Farida Cheriet[1][0000−0001−6170−5627]

[1] Polytechnique Montréal
[2] Hôpital Maisonneuve-Rosemont

Abstract. We explore the efficacy of a region-based method for image tokenization, aimed at enhancing the resolution of images fed to a Transformer. This method involves segmenting the image into regions using SLIC superpixels. Spatial features, derived from a pretrained model are aggregated segment-wise and input into a streamlined Vision Transformer (ViT). Our model introduces two novel contributions: the matching of segments to semantic prototypes and the graph-based clustering of tokens to merge similar adjacent segments. This approach leads to a model that not only competes effectively in classifying diabetic retinopathy but also produces high-resolution attribution maps, thereby enhancing the interpretability of its predictions.

Keywords: Diabetic retinopathy · Superpixel · Vision Transformers.

1 Introduction

Automatic diagnosis of diabetic retinopathy (DR) from retinal images is one of the most promising applications of deep learning in ophthalmology. Convolutional neural networks (CNNs), and more recently Vision Transformers (ViTs) [4], have shown impressive performance in DR detection and grading. ViTs break down the input image into a sequence of tokens extracted from non-overlapping square patches. This procedure, while computationally convenient, is a rather coarse-grained approach to downsampling the image. In most natural images, but especially in fundus imaging, anatomical structures such as lesions, vessels, and optic disc/cup, do not lend themselves very well to this kind of strategy given their irregular shapes and highly variable sizes. Microaneurisms notably are typically only a few pixels wide even in high resolution images, and their presence and number is a crucial element in clinical guidelines for DR grading. A square token might contain parts of several different structures, whose characteristics must all be encoded in a fixed-size vector. Several studies have underscored the efficacy of Vision Transformers (ViT) in diabetic retinopathy classification from fundus images [20,17,14] [3,6]. Transformers have demonstrated good scalability with larger datasets, prompting the development of foundation models pretrained on massive retinal datasets [**?**]. However, a common aspect of these approaches is that the input image resolution is constrained by memory limitations, typically ranging from a maximum of 512 × 512 to as low as 224 × 224. With emerging modalities such as Ultra Wide Field fundus imaging gaining clinical adoption and offering resolutions up to 4000 × 4000, it becomes crucial to devise new methodologies to scale these models to higher resolutions.

Departing from the grid-like slicing of input images, we propose an exploration of superpixel sampling. Concurrently to the present work, recent studies have presented promising outcomes by changing the tokenization process of images fed to ViTs[1]. Aasan et al.[1] suggest a straightforward superpixel sampling approach for input images, achieving performance comparable to standard ViT while enhancing interpretability. Similarly, Huang et al.[7] introduce the concept of "supertokens" wherein the network learns to aggregate similar tokens. Likewise, SPFormer [13] learns to group pixel features into clusters of superpixels, achieving competitive performance in classification tasks. Concerning the generated superpixels, the authors of [13] note that slightly superior clusters are obtained with the conventional SLIC algorithm [2], a direction we chose to follow.

Our model consists of two primary components: a feature extractor and a classification model. For the feature extraction task, we leveraged an EfficientNet-5 model [18](pretrained on our data), while for the classification aspect, a six-layer Transformer architecture is used. Our principal contributions lie in the novel interaction between these modules and a novel pooling procedure proposed within the Transformer framework. Figure 1 provides an overview of our model.

2 Methodology

2.1 Superpixels computation and segment-wise features aggregation

The SLIC algorithm [2] efficiently computes superpixels in a straightforward approach. It places cluster centers across the image in a grid, iteratively comparing neighboring pixels within a set radius to cluster centroids, considering both color and spatial proximity. Post-iteration, it updates the centroid positions similarly to k-means, using a compactness parameter in its distance calculation to balance the superpixels' boundary adherence with their regularity. The algorithm is customized via two hyperparameters: the desired number of superpixels N and their compactness (adherence of superpixel boundaries to the image's structures).

The resulting superpixels divide the image into segments arranged from top left to bottom right, with each pixel assigned to a segment $s \in \{0, ..., N\}$. Segments in the fundus image's black borders are merged, significantly reducing total segments. Nonetheless, the segment count is still referred to as N for simplicity. Utilizing our pretrained CNN, we extract a feature map $\mathcal{F} \in$

[1] Although most of the related works were either under review or in preprint stage at the time of writing, we still find it relevant to mention them.

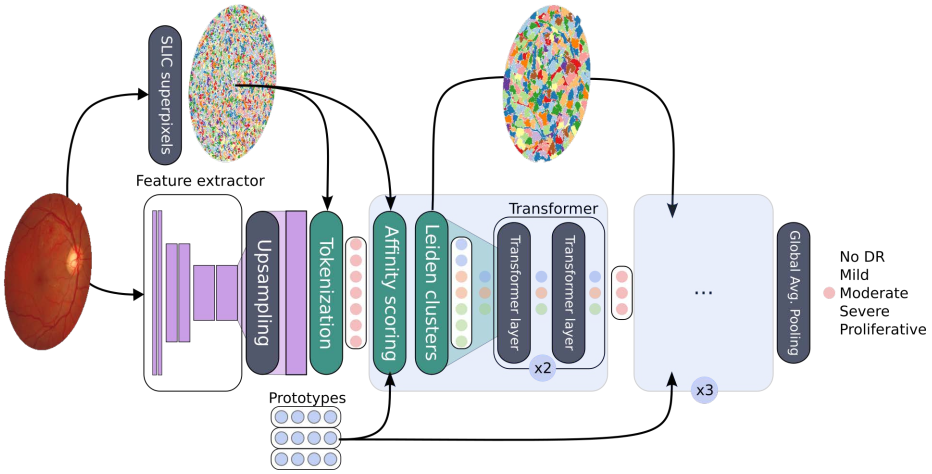

Fig. 1. Overview of the proposed method. Features are extracted using a conventional CNN and tokenized based on superpixel segments obtained using SLIC [2]. The sequence obtained is given to our proposed block (highlighted in grey). First, affinity between tokens (based on spatial proximity and prototypes matching) is estimated. The Leiden algorithm for community detection [19] is used to pool tokens. The sequence of tokens is then fed to two regular transformer encoder layers. This block is repeated three times.

$\mathbb{R}^{d\times(H/2^k)\times(W/2^k)}$ from an image $I \in \mathbb{R}^{3\times H\times W}$, where k corresponds to the number of pooling steps in the CNN and d the number of feature maps. As noted by Shlapentokh-Rothman et al. [16], we observe that downsampling S to $\frac{H}{2^k} \times \frac{W}{2^k}$ results in losing of the smallest superpixels. Hence, we opt to reinterpolate $\mathcal{F}$ to the resolution of S.

The image's tokenization is achieved by averaging the feature maps over each segment s.

$$\mathbf{x} = \underset{i,j}{\text{mean}}(\mathcal{F}_{:,i,j}) \text{ s.t. } \text{SLIC}(I)(i,j) = s \tag{1}$$

Note that instead of averaging, we could take any form of reductive function (e.g. max, sum, etc.).

The resulting sequence can be input into a standard Transformer architecture. Nevertheless, even with superpixels, the number N of tokens can become exceedingly large on high resolution images. Given that attention computation scales as $\mathcal{O}(N^2)$, we introduce a novel dynamic pooling layer specifically designed for region-based tokens, to limit the overall computational cost. The following sections describe this phase in our model.

2.2 Affinity-based recombination of tokens

Our rationale is based on the premise that superpixel segments containing similar information should be merged together. In the context of medical images, this notion of similarity is governed by two principles: spatial proximity and whether the segments cover homogeneous biomarkers within the image. For example, a significant portion of the retinal background can be merged. However, a segment containing hemorrhages should not be merged with the macula, even though they might be overlapping (see Figure 2).

We compute two adjacency matrices $A_{\text{sim}}, A_{\text{prox}} \in \mathbb{R}^{N \times N}$ representing similarity and proximity respectively between each token of the sequence. The first matrix indicates that superpixels are adjacent, while the second matrix identifies superpixels with similar semantic content. We ensure that both matrices are normalized using the following equation:

$$A_{\text{norm}} = D^{-1/2} \cdot A \cdot D^{1/2} \tag{2}$$

where D is the diagonal degree matrix.

To constrain simultaneously the spatial and semantic conditions, we defined the affinity matrix as:

$$A_{\text{affinity}} = A_{\text{sim}} \odot A_{\text{prox}} \tag{3}$$

where $\odot$ is the Hadamard product of matrices. The sequence of tokens can then be recombined following the equation:

$$\mathbf{x}' = A_{\text{affinity}} \cdot \mathbf{x} \tag{4}$$

Equation 4 is a token-wise recombination similar to a Transformer attention block. However, it does not change the sequence length. Our next contribution borrows an algorithm from the literature on graph computation to progressively cluster our segments based on the explicit rules contained in A_{affinity}.

2.3 Segment adjacency

Given the superpixel segmentation map S, our objective is to determine the adjacency matrix delineating the segments' interconnections. We propose a simple approach necessitating merely four convolutions utilizing the kernels $K_x = [0, 1]$, $K'_x = [1, 0]$, $K_y = [0, 1]^\top$, and $K'_y = [1, 0]^\top$. This method yields two tensors E_x and $E_y \in \mathbb{N}^{2 \times H \times W}$, which can be viewed as the indices of edges connecting neighboring superpixels horizontally and vertically, respectively.

Leveraging the scattering operations provided in the PyTorch Geometric library [2], we obtain a $N \times N$ adjacency matrix A_{prox}, where each entry $A_{\text{prox}}(i, j)$ denotes the number of adjacent pixels between segments i and j. Moreover, the diagonal entries correspond to the size of each segment. The normalized matrix (following Equation 2) can be construed as representing the connectivity strength between two segments, in the range [0, 1].

[2] https://pytorch-geometric.readthedocs.io/en/latest/modules/utils.html#torch_geometric.utils.to_dense_adj

2.4 Prototype matching

In addition to the proximity, our methodology relies on reinforcing the connection between similar segments. Instead of comparing all the pairs of segments, we match every token to a prototype (obtained from annotated image samples) among a set of K candidates. Tokens matching with the same prototype are considered as belonging to the same cluster. The similarity between the embedded point $\mathbf{x}_i$ and the prototype $\mathbf{p}_j$ is measured using the Student's t-distribution following the idea proposed in [21]:

$$q_{ij} = \frac{(1 + ||\mathbf{x}_i - \mathbf{p}_j||^2/\alpha)^{-\frac{\alpha+1}{2}}}{\sum_{j'}(1 + ||\mathbf{x}_i - \mathbf{p}_{j'}||^2/\alpha)^{-\frac{\alpha+1}{2}}} \tag{5}$$

We also experimented with the cosine similarity (similarly to [10]) but found little practical difference. However, we observed the importance of normalizing both in x_i and p_j in equation 5. The scores q_{ij} are gathered into an association matrix $Q \in \mathcal{R}^{N \times K}$. Q can be interpreted as a probability distribution over the set of prototypes given a specific token. Using a temperature parameter $\tau > 1$, we define

$$C = \underset{K}{\mathrm{softmax}}(Q \times \tau) \tag{6}$$

and finally

$$A_{\mathrm{sim}} = C \cdot C^T \tag{7}$$

2.5 Leiden clustering for dynamic pooling

With A_{prox} and A_{sim} computed, we can now compute A_{affinity} following Equation 3. A_{affinity} corresponds to an undirected graph, wherein the nodes denote segments, and the edges signify both semantic closeness and spatial adjacency. Unlike the prevalent pooling methodologies in the literature on graph neural networks, we don't know a priori the number of nodes in our graph, nor do we arbitrarily fix the number of clusters after reduction. In particular, we want the algorithm to be able to fuse tokens beyond immediate neighbors as long as they belong to the same subgraph, while retaining control over the approximate number of clusters, thereby mitigating the risk of excessively reducing our sequence size. This has motivated our exploration of more generalized graph algorithms, particularly within the realm of community detection in social networks. Traag et al. [19] introduced the Leiden algorithm for identifying clusters of nodes, also referred to as communities, within large graphs. Each node is allocated to a cluster with the objective of maximizing a metric called modularity, defined as:

$$\mathcal{H} = \frac{1}{2m} \sum_c (e_c - \gamma \frac{K_c^2}{2m}) \tag{8}$$

e_c represents the number of edges in community c, m denotes the total number of edges in the network, K_c is the sum of degrees of nodes within community c, and γ serves as a hyperparameter known as resolution. The aim is to maximize the discrepancy between the actual and expected numbers of edges within a community. For an in-depth understanding of the algorithm, we direct readers to the original publication [19]. The optimization of the modularity uses non-differentiable operations (iterative assignment of nodes to clusters followed by merging of these clusters); however A_{affinity} is trainable through Equation 3. Finally, tuning γ permits a fuzzy control over the token reduction factor. We empirically set it to have a reduction varying between 25% to 50% at each pooling step. We illustrate the effect of the Leiden pooling in Figure 2.

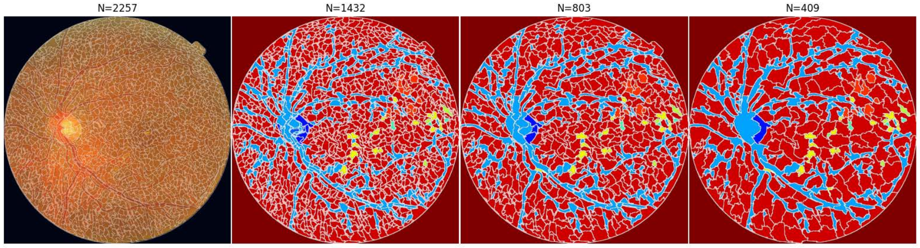

Fig. 2. Effect of the Leiden pooling on an image and groundtruths from MAPLES-DR. Note how our pooling approach preserves the semantic structures of the segments while merging similar ones. The labels include in particular "vessels, optic disc/cup, macula, hemorrhages, exsudates and cotton wool spot"

3 Experiments

3.1 Data

Our model training used the EyePACS [5] train split, which includes 35,126 images, setting aside 4,000 for validation. For testing, we utilized the DDR [12], Aptos [8], IDRiD [15] datasets, and the EyePACS test split, with image counts of 4,105, 3,662, 103, and 53,576, respectively. These 45° fundus images of varying resolutions were labeled into five categories: No-DR, Mild, Moderate, Severe, and Proliferative DR. We standardized the images to a 1024 × 1024 resolution, cropping out most of the black borders to focus on the circular ROI. Using Fast-SLIC[3], we generated 4096 superpixels per image, discarding about 40% in preprocessing (those that were too small or part of the black border).

3.2 Training procedure

Hyperparameter choices The Transformer was kept lightweight, with an embedding size of 176 and 16 attention heads, totalling 2.2 millions parameters. The

[3] `https://github.com/Algy/fast-slic`

input features were extracted from the third pooling layer of the EfficientNet-5 (27.3 millions parameters), which was kept frozen during the training of the ViT.

Prototype initialization Instead of sampling from a random distribution, we initialized the prototypes as the mean feature representations of segments from labelled retinal biomarkers. We exploited the 200 manually labelled images from the MAPLES-DR dataset [11], providing the masks for 14 types of biomarkers. Noting $Y(i,j) = \{1,\dots,14\}$ the groundtruth of an image, this initialization is formally obtained as

$$\begin{aligned} \mathbf{p}_m = \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}} \underset{s}{\text{mean}}(\mathcal{F}^{(i)}_{:,j,k}) \\ \text{s.t. } \text{SLIC}(I)(j,k) = s \text{ and } Y(j,k) = m \end{aligned} \tag{9}$$

where $\mathcal{D}$ is the MAPLES-DR dataset.

Target loss Given the ordered nature of diabetic retinopathy grades, we modeled the classification as a regression task. Our primary training loss, $\mathcal{L}_1$, was the mean squared error (MSE) between the network's output o_i and the true ordinal grade y_i. The discrete prediction was made by rounding o_i to the nearest integer. Additionally, drawing on similar works [21,9], we incorporated an unsupervised alignment loss, $\mathcal{L}_{\text{align}} = \text{KL}(C||Q)$, which uses the Kullback–Leibler divergence to ensure that segments align closely with their nearest prototype. Here, C is a "confident" version of Q, promoting clear segment-to-prototype associations. This alignment loss is calculated at each pooling layer ℓ, contributing to the total loss, as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_1 + \frac{\lambda}{L}\sum_{\ell=1}^{L}\mathcal{L}^{(\ell)}_{\text{align}} \tag{10}$$

where λ is a hyperparameter set to 0.01 in our experiments.

Optimization We trained our model using the AdamW solver, with a learning rate of 0.01 and cosine decay. Due to a limited hardware budget, we only trained the model for 10 epochs. Interestingly, we observed a very fast convergence on the validation set, especially for a non-pretrained ViT ($\kappa = 0.795$ after one epoch).

3.3 Results

Table 1 provides the comparative performance between our model and recently published papers. We also trained an EfficientNet-5 and various others CNN as baselines (we only include the former in Table 1 as it was the best performing). As we can see, our approach yields results slightly lower than the CNN, but higher than recently published models tested on the same data and based on ViTs like our proposed method.

Table 1. Comparative performance between our proposed approach, a CNN baseline (EfficientNet-5) and published results from recent ViT-based models.

Models	Resolution	Cohen's quadratic κ			
		EyePACS	Aptos	DDR	IDRiD
RetFOUND [22]	224	0.488	0.572	0.627	0.606
$\text{ViT}^{(16)}_{\text{base}}$ [14]	384	0.737	0.874	–	–
EfficientNet-5	1024	0.831	0.879	0.811	0.782
Ours	1024	0.819	0.861	0.759	0.796

In addition, we ran an ablation study to evaluate the effects of our specific contributions, summarized in Table 2. Note that beyond raw performance, ③ offers the benefit of requiring less GPU-memory than ①/② thanks to the pooling process. Running the training is however around 75% slower due to the relative complexity of the Leiden algorithm.

Table 2. Ablation study. Each variant was trained for 5 epochs.

Variants	Cohen's quadratic κ (EyePACS val)
1) CNN & Superpixels ViT	0.811
2) ① + Affinity recombination (eq. 4)	0.812
3) ① + ② + Leiden pooling (proposed model)	0.826

4 Discussion

Our model outperforms recent ViT models by using superpixel sampling and affinity pooling to capture intricate details necessary for accurate diabetic retinopathy classification. Our ablation study shows the beneficial impacts of affinity-based recombination and Leiden pooling on the κ score. Superpixels tokenization also offers great opportunities in term of model interpretability; we provide heatmaps in the supplementary material as well as in our code repository. Despite its computational demands, Leiden pooling reduces GPU memory use, a boon for processing high-resolution images. A current limitation is the need for a CNN for feature extraction, which is computationally expensive. Future work will seek more efficient feature extraction methods per segment. With this research, we propose a novel methological approach to classify images, combining conventional CNN, ViT and graph algorithms. Overall, our research presents a new avenue for training transformers on high-resolution images, competing well with traditional CNNs and ViTs in diabetic retinopathy classification, with potential wider medical imaging applications.

Code Availability Code is available in our repository https://github.com/ClementPla/RetinalViT/tree/prototype_superpixels. It contains links to all the datasets we use. Model weights will be shared upon request.

Acknowledgements We sincerely acknowledge Philippe Debanné and Dr. Fares Antaki for revising this manuscript. We also thank the Ophthalmology Research Fund from Université de Montréal (FROUM) and the Line Chevrette Fund for AI-Ophthalmology Research for supporting this project.

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

References

1. Aasan, M., Rivera, A.R., Kolbjornsen, O., Solberg, A.S.: A Spitting Image: Superpixel Transformers (Oct 2023), https://openreview.net/forum?id=Vy6sjPt2Vr
2. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(11), 2274–2282 (Nov 2012). https://doi.org/10.1109/TPAMI.2012.120, https://ieeexplore.ieee.org/document/6205760, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence
3. Adak, C., Karkera, T., Chattopadhyay, S., Saqib, M.: Detecting Severity of Diabetic Retinopathy from Fundus Images using Ensembled Transformers (Jan 2023). https://doi.org/10.48550/arXiv.2301.00973, http://arxiv.org/abs/2301.00973, arXiv:2301.00973 [cs]
4. 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 16x16 Words: Transformers for Image Recognition at Scale (Sep 2020), https://openreview.net/forum?id=YicbFdNTTy
5. Dugas, E., Jorge, J., Cukierski, W.: Diabetic retinopathy detection (2015), https://kaggle.com/competitions/diabetic-retinopathy-detection
6. Gu, Z., Li, Y., Wang, Z., Kan, J., Shu, J., Wang, Q.: Classification of Diabetic Retinopathy Severity in Fundus Images Using the Vision Transformer and Residual Attention. Computational Intelligence and Neuroscience **2023**, e1305583 (Jan 2023). https://doi.org/10.1155/2023/1305583, https://www.hindawi.com/journals/cin/2023/1305583/, publisher: Hindawi
7. Huang, H., Zhou, X., Cao, J., He, R., Tan, T.: Vision Transformer with Super Token Sampling (Jan 2024). https://doi.org/10.48550/arXiv.2211.11167, http://arxiv.org/abs/2211.11167, arXiv:2211.11167 [cs]
8. Karthik, M., Sohier, D.: APTOS 2019 Blindness Detection (2019), https://kaggle.com/c/aptos2019-blindness-detection
9. Khasahmadi, A.H., Hassani, K., Moradi, P., Lee, L., Morris, Q.: Memory-based graph networks. In: International conference on learning representations (2020), https://openreview.net/forum?id=r1laNeBYPB
10. Lee, D., Kim, S., Lee, S., Park, C., Yu, H.: Learnable Structural Semantic Readout for Graph Classification. In: 2021 IEEE International Conference on Data Mining (ICDM). pp. 1180–1185 (Dec 2021). https://doi.org/10.1109/ICDM51629.2021.00142, https://ieeexplore.ieee.org/abstract/document/9679111, iSSN: 2374-8486

11. Lepetit-Aimon, G., Playout, C., Boucher, M.C., Duval, R., Brent, M.H., Cheriet, F.: MAPLES-DR: MESSIDOR Anatomical and Pathological Labels for Explainable Screening of Diabetic Retinopathy (Jan 2024), https://arxiv.org/abs/2402.04258v1
12. Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., Kang, H.: Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Information Sciences **501**, 511 – 522 (2019). https://doi.org/https://doi.org/10.1016/j.ins.2019.06.011, http://www.sciencedirect.com/science/article/pii/S0020025519305377
13. Mei, J., Chen, L.C., Yuille, A., Xie, C.: SPFormer: Enhancing Vision Transformer with Superpixel Representation (Jan 2024). https://doi.org/10.48550/arXiv.2401.02931, http://arxiv.org/abs/2401.02931, arXiv:2401.02931 [cs]
14. Playout, C., Duval, R., Boucher, M.C., Cheriet, F.: Focused Attention in Transformers for interpretable classification of retinal images. Medical Image Analysis **82**, 102608 (Nov 2022). https://doi.org/10.1016/j.media.2022.102608, https://www.sciencedirect.com/science/article/pii/S1361841522002377
15. Porwal, P.: Indian Diabetic Retinopathy Image Dataset (IDRiD) (Apr 2018), https://ieee-dataport.org/open-access/indian-diabetic-retinopathy-image-dataset-idrid, publisher: IEEE
16. Shlapentokh-Rothman, M., Blume, A., Xiao, Y., Wu, Y., T V, S., Tao, H., Lee, J.Y., Torres, W., Wang, Y.X., Hoiem, D.: Region-Based Representations Revisited (Feb 2024), http://arxiv.org/abs/2402.02352, arXiv:2402.02352 [cs]
17. Sun, R., Li, Y., Zhang, T., Mao, Z., Wu, F., Zhang, Y.: Lesion-Aware Transformers for Diabetic Retinopathy Grading. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10933–10942 (Jun 2021). https://doi.org/10.1109/CVPR46437.2021.01079, iSSN: 2575-7075
18. Tan, M., Le, Q.V.: EfficientNet: Rethinking model scaling for convolutional neural networks. ArXiv **abs/1905.11946** (2019), https://api.semanticscholar.org/CorpusID:167217261
19. Traag, V.A., Waltman, L., van Eck, N.J.: From Louvain to Leiden: guaranteeing well-connected communities. Scientific Reports **9**(1), 5233 (Mar 2019). https://doi.org/10.1038/s41598-019-41695-z, https://www.nature.com/articles/s41598-019-41695-z, number: 1 Publisher: Nature Publishing Group
20. Wu, J., Hu, R., Xiao, Z., Chen, J., Liu, J.: Vision Transformer-based recognition of diabetic retinopathy grade. Medical Physics **48**(12), 7850–7863 (Dec 2021). https://doi.org/10.1002/mp.15312
21. Xie, J., Girshick, R., Farhadi, A.: Unsupervised deep embedding for clustering analysis. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48. pp. 478–487. ICML'16, JMLR.org, New York, NY, USA (Jun 2016)
22. Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., Kihara, Y., Allen, N., Gallacher, J.E.J., Littlejohns, T., Aslam, T., Bishop, P., Black, G., Sergouniotis, P., Atan, D., Dick, A.D., Williams, C., Barman, S., Barrett, J.H., Mackie, S., Braithwaite, T., Carare, R.O., Ennis, S., Gibson, J., Lotery, A.J., Self, J., Chakravarthy, U., Hogg, R.E., Paterson, E., Woodside, J., Peto, T., Mckay, G., Mcguinness, B., Foster, P.J., Balaskas, K., Khawaja, A.P., Pontikos, N., Rahi, J.S., Lascaratos, G., Patel, P.J., Chan, M., Chua, S.Y.L., Day, A., Desai, P., Egan, C., Fruttiger, M., Garway-Heath, D.F., Hardcastle, A., Khaw, S.P.T., Moore, T., Sivaprasad, S., Strouthidis, N., Thomas, D., Tufail, A., Viswanathan, A.C.,

Dhillon, B., Macgillivray, T., Sudlow, C., Vitart, V., Doney, A., Trucco, E., Guggeinheim, J.A., Morgan, J.E., Hammond, C.J., Williams, K., Hysi, P., Harding, S.P., Zheng, Y., Luben, R., Luthert, P., Sun, Z., McKibbin, M., O'Sullivan, E., Oram, R., Weedon, M., Owen, C.G., Rudnicka, A.R., Sattar, N., Steel, D., Stratton, I., Tapp, R., Yates, M.M., Petzold, A., Madhusudhan, S., Altmann, A., Lee, A.Y., Topol, E.J., Denniston, A.K., Alexander, D.C., Keane, P.A., UK Biobank Eye & Vision Consortium: A foundation model for generalizable disease detection from retinal images. Nature (Sep 2023). https://doi.org/10.1038/s41586-023-06555-x, https://doi.org/10.1038/s41586-023-06555-x