ENHANCING CONTRASTIVE LEARNING FOR RETINAL IMAGING VIA ADJUSTED AUGMENTATION SCALES

Zijie Cheng^{*,1} *Boxuan Li*^{1} *André Altmann*^{1} *Pearse A Keane* ^{2} *Yukun Zhou*^{*,2}

^{1} UCL Department of Medical Physics & Biomedical Engineering, United Kingdom
^{2} UCL Institute of Ophthalmology, United Kingdom
$\star$ Corresponding authors. Email: {rmapzch, yukun.zhou.19}@ucl.ac.uk

ABSTRACT

Contrastive learning, a prominent approach within self-supervised learning, has demonstrated significant effectiveness in developing generalizable models for various applications involving natural images. However, recent research indicates that these successes do not necessarily extend to the medical imaging domain. In this paper, we investigate the reasons for this suboptimal performance and hypothesize that the dense distribution of medical images poses challenges to the pretext tasks in contrastive learning, particularly in constructing positive and negative pairs. We explore model performance under different augmentation strategies and compare the results to those achieved with strong augmentations. Our study includes six publicly available datasets covering multiple clinically relevant tasks. We further assess the model's generalizability through external evaluations. The model pre-trained with weak augmentation outperform those with strong augmentation, improving AUROC from 0.838 to 0.848 and AUPR from 0.523 to 0.597 on MESSIDOR-2, and showing similar enhancements across other datasets. Our findings suggest that optimizing the scale of augmentation is critical for enhancing the efficacy of contrastive learning in medical imaging.

Index Terms— contrastive learning, augmentation scales, data distribution, retinal imaging

1. INTRODUCTION

Contrastive learning is a machine learning paradigm that trains models to distinguish between similar and dissimilar data points without relying on explicit labels. Despite being pre-trained only on unlabeled data, contrastive learning drives competitive pre-trained models compared to those pre-trained with supervised learning-based methods. In the natural image domain, it has demonstrated promising results in diverse tasks such as object detection [1], image classification [2], and video analysis [3]. Compared to generative learning, contrastive learning has shown greater effectiveness in various applications involving natural images [4, 5]. However, whether this observation extends to medical images remains underexplored.

Recent research has begun comparing contrastive learning and generative learning in medical AI. For instance, RETFound [6], a foundation model for retinal images, employed a generative learning strategy using the Masked Autoencoder [7] for model development and demonstrated superior performance compared to contrastive learning methods in retinal disease classification. Understanding the reasons behind this inconsistency and developing a simple yet efficient solution to improve contrastive learning for medical imaging is crucial.

The suboptimal performance of contrastive learning in medical imaging is likely due to inherent differences between the distributions of natural and medical images [8]. Natural images are colorful with varying pixel intensities, while medical images are usually grayscale and structurally similar, especially within the same organ or tissue type [9, 10]. This characteristic results in a denser distribution of medical images within the latent space compared to natural images [11]. We hypothesize that such a dense distribution degrades performance when applying contrastive learning methods to medical images. As shown in Figures 1(a) and 1(b), natural images under strong augmentations in contrastive learning are sparsely distributed in latent space, while different medical images tend to overlap heavily. Since the pretext of contrastive learning is to differentiate between positive pairs (augmented views of the same image) and negative pairs (augmented views of different images), the severe overlap in medical images makes the pretext task highly challenging, making it difficult for the model to converge effectively.

In this work, we propose a simple yet effective solution to enhance contrastive learning performance by reducing augmentation scales. The project pipeline is illustrated in Figure 1(c). We employ Distillation with No Labels (DINO) [4], and validated our solution on glaucoma and diabetic retinopathy diagnosis, using both internal and external evaluations. Our approach not only enhances feature clustering but also demonstrates improved diagnostic accuracy compared to models using stronger augmentation strategies.

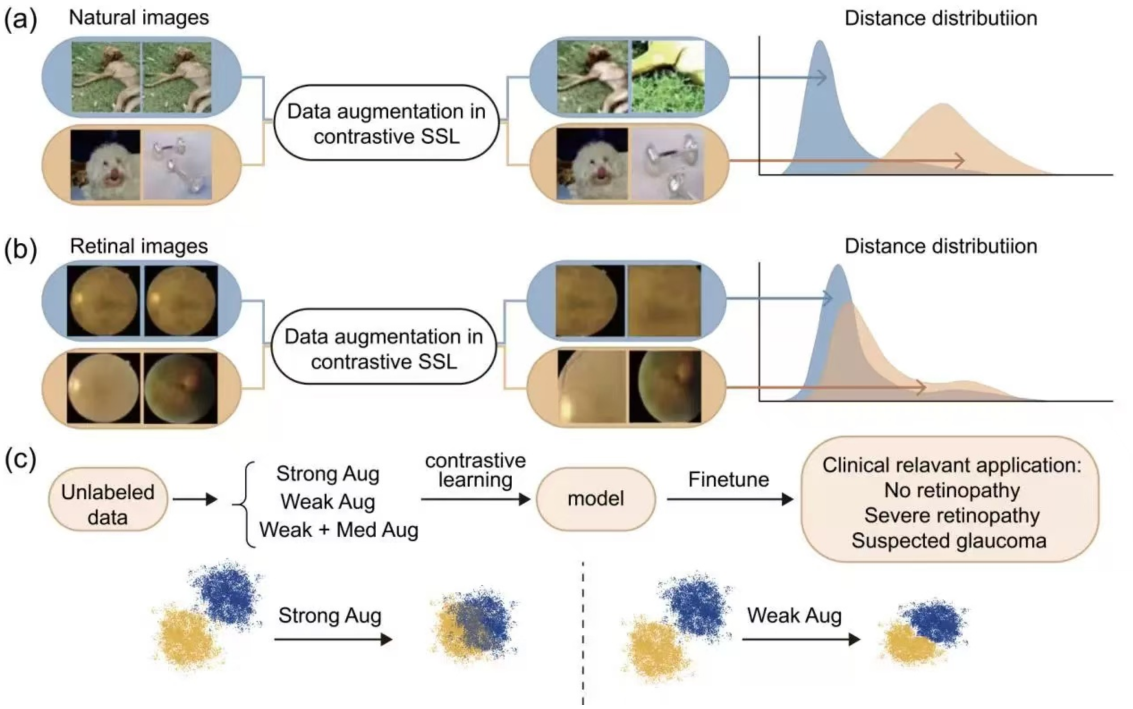

Fig. 1: Figures (a) and (b) illustrate the distribution of distances between positive pairs and negative pairs in both natural and medical image domains. Figure (c) presents the project pipeline: unlabeled data is used to pre-train contrastive learning models while investigating various augmentation strategies. The blue dots and yellow dots indicate augmented images from different original images. The goal of this approach is to enhance feature clustering and improve the accuracy of retinal disease diagnosis.

2. METHODS

2.1. Problem Definition

For contrastive learning, given a set of unlabeled retinal images $\mathcal{D} = \{x_i\}_{i=1}^{N}$, we create positive pairs $\mathcal{P}^+$ by randomly selecting an image $x_i \in \mathcal{D}$ and apply twice augmentation $\Phi_{t,s}$ respectively to get augmented data x_i^1 and x_i^2, where t indicates the augmentation type and s the scale range. While for negative pairs, we sample two images x_i and $x_j \in \mathcal{D}$ (with $i \neq j$) and apply the augmentation to each image, forming the negative pair $\mathcal{P}^- = (x_i^1, x_j^2)$. The distance between positive pairs and negative pairs can be measured by $Dis(\cdot)$:

$$\text{Dis}(\mathcal{P}^+) = |x_i^1 - x_i^2|, \tag{1}$$

$$\text{Dis}(\mathcal{P}^-) = |x_i^1 - x_j^2|. \tag{2}$$

The general training objective of contrastive learning is to train the model f to maximize the distance between negative pairs and to minimize that for positive pairs,

$$f = \text{argmax}\left(\text{Dis}(\mathcal{P}^-) - \text{Dis}(\mathcal{P}^+)\right). \tag{3}$$

When $\text{Dis}(\mathcal{P}^+)$ approximates $\text{Dis}(\mathcal{P}^-)$, it is challenging to train the model f to converge well. This issue is prominent in medical imaging due to less variation compared to natural images. For example, retinal images depict the anatomical tissue of retina, often showing similar structure and orientation [12], as shown in Figure 1(b). With strong augmentations Φ_{strong} (e.g., cropping the images into small patches), $\text{Dis}(\mathcal{P}^-)$ decrease while $\text{Dis}(\mathcal{P}^+)$ increases, which brings further challenges in achieving objects of equation 3 and may result in suboptimal model pre-training with contrastive learning, showing the poor performance in classifying the positive and negative pairs.

Such suboptimal model performance extends to downstream applications, where models are fine-tuned with labeled data $\mathcal{D}_l = \{x_i, y_i\}_{i=1}^{L}$ for diverse tasks like disease diagnosis, where x represents the data and y indicates the label. To improve the model's capability in clinically meaningful applications, we aim to optimize $\arg\max_{x_i, y_i \in \mathcal{D}_l} T\left(E(x_i), y_i\right)$, where $T(\cdot)$ is the score function and $E(\cdot)$ is the encoder of the model f.

Our strategy involves enhancing the contrastive learning performance by specifically decreasing $\text{Dis}(\mathcal{P}^+)$ while increasing $\text{Dis}(\mathcal{P}^-)$.

2.2. Scattering Data Distribution with Weak Augmentations

To achieve such a goal for retinal images, a straightforward strategy is to scale down the augmentation. An extreme case

Table 1: Data summary for the datasets used for disease diagnosis. Each dataset is split into training, validation, and testing sets.

Dataset	Country	Types	Training	Validation	Testing
Diabetic retinopathy					
MESSIDOR-2	France	5	972	246	526
IDRiD	India	5	329	84	103
APTOS2019	India	5	2048	514	1100
Glaucoma					
PAPILA	Spain	3	312	79	98
Multi-class disease					
JSIEC	China	39	534	150	316
Retina	NR	4	336	84	181

is to remove the augmentation so that Dis$(\mathcal{P}^+)$ achieves 0 and Dis$(\mathcal{P}^-)$ stays as a high value. However, pre-training without any augmentation hardly trains the model to learn generalizable and diverse features. Hence, we propose to proportionally scale down the augmentation, termed as Φ_{weak}, to ease the challenge of training model f to converge while also allowing the model to learn generalizable features. Additionally, we also investigate the effects of several augmentations that mimic the retinal image artefacts, including random bias field, and Gaussian blur. We combine it with Φ_{weak} to form $\Phi_{weak+med}$.

3. EXPERIMENT

3.1. Data

We evaluate the efficacy of different augmentation strategies using clinically meaningful tasks, including diabetic retinopathy (DR) diagnosis, glaucoma detection, and multi-class retinal disease classification.

For DR diagnosis, we include MESSIDOR-2 [13], IDRiD [14], and APTOS2019 [15]. The labels for DR are based on the International Clinical DR Severity Scale, covering five stages from no DR to proliferative DR. For glaucoma diagnosis, we use the PAPILA dataset [16], which has three categorical labels: non-glaucoma, early glaucoma (suspected glaucoma), and advanced glaucoma. For multi-class disease classification tasks, we use two datasets, JSIEC [17] containing 1,000 images with 39 categories of common retinal diseases and conditions, and Retina dataset [18] with labels for normal, glaucoma, cataract, and retinal disease. Data splitting details are shown in Table 1. The pre-training data are from Moorfields Eye Hospital [6].

3.2. Pre-training details

DINO [4], a representative and commonly used contrastive learning strategy, was used in the experiment. We first initialized the model with ImageNet weights and then pre-trained it using 1.4 million retinal images from Moorfields Eye Hospital. The data preprocessing, data quality control, model architecture, and hyperparameters (except for those related to augmentations) were standardized to ensure a fair comparison. The model was pre-trained using an NVIDIA A100 (80G). The details of Φ_{strong}, Φ_{weak}, and $\Phi_{weak+med}$ are listed in Table 2. Local crop indicates the range of cropping local patches and global crop represents that for cropping global patches. Color jittering involves random adjustments to image brightness and contrast to simulate variations in imaging conditions. Gaussian blur applies a Gaussian filter to smooth images, mimicking motion blur or out-of-focus effects. Random noise adds Gaussian noise to simulate sensor or acquisition noise. Random bias field introduces smooth, spatially varying intensity variations to mimic changes in light illumination direction.

Table 2: Various settings of augmentation types and scales. Augmentations not listed are consistent with the strong augmentations. For local and global crops, the range (e.g., (0.05, 0.4)) represents the cropping scales relative to the original image. The symbol p denotes the probability of applying a particular transformation. A ”×” indicates that the transformation is not applied.

	Local crop	Global crop	Color jitter	Blur	Noise	Bias field
Φ_{strong}	(0.05, 0.4)	(0.4, 1.0)	bright:0.4 contrast:0.4	×	×	×
Φ_{weak}	(0.2, 0.5)	(0.5, 1.0)	bright:0.2 contrast:0.2	×	×	×
$\Phi_{\text{weak+med}}$	(0.2, 0.5)	(0.5, 1.0)	bright:0.2 contrast:0.2	std:0.1 p:0.5	std:0.1 p:0.5	scale:0.1 p:0.5

We then compared these models by adapting them to downstream tasks, i.e., disease diagnosis. We evaluated the model performance with the Area Under the Receiver Operating Characteristic curve (AUROC) and the Area Under the Precision-Recall curve (AUPR). Each experiment is run five times with random seeds to obtain performance statistics.

3.3. Experiment Result

We first observed the clustering performance, i.e., how positive and negative pairs distribute, before and after reducing the augmentation scale. The results are shown in Figure 2. The model pre-trained with Φ_{weak} better separated these pairs. We also repeatedly augmented each image multiple times to create image groups, where positive pairs consist of images within the same group, and negative pairs are images from different groups. We found that positive pairs cluster more closely under weak augmentation in the t-SNE map [19].

In internal evaluation, as shown in Table 4, DINO with Φ_{weak} outperforms other augmentation strategies in most retinal disease classification tasks. On MESSIDOR-2, IDRiD, PAPILA, and JSIEC, the model with Φ_{weak} shows consistently better performance in both AUROC and AUPR metrics compared to with Φ_{strong}. Notably, on JSIEC, the model pre-trained with Φ_{weak} achieves a 10% higher AUPR than

Table 3: This table presents the external evaluation results on diabetic retinopathy (DR) datasets based on AUROC. For each dataset pair, the highest mean value among the different augmentation strategies is highlighted in bold.

Fine-tune data	APTOS2019		IDRiD		MESSIDOR-2	
Test data	IDRiD	MESSIDOR-2	APTOS2019	MESSIDOR-2	APTOS2019	IDRiD
Φ_{strong}	.784 ± .003	**.767 ± .006**	.745 ± .016	.742 ± .031	**.806 ± .023**	.740 ± .034
Φ_{weak}	**.790 ± .019**	.760 ± .007	**.749 ± .033**	**.760 ± .027**	.798 ± .038	**.744 ± .019**
$\Phi_{weak+med}$	.751 ± .014	.691 ± .002	.733 ± .055	.720 ± .020	.706 ± .077	.736 ± .034

Table 4: Model comparison on disease diagnosis with internal evaluation. Each column indicates the model pre-trained with varied data augmentation strategies. The highest value in each row is highlighted in bold.

	Φ_{strong}	Φ_{weak}	$\Phi_{weak+med}$
MESSIDOR-2			
AUROC	.838 (.834, .842)	**.848 (.844, .852)**	.823 (.814, .832)
AUPR	.523 (.513, .533)	**.597 (.583, .610)**	.523 (.487, .558)
APTOS2019			
AUROC	.933 (.932, .933)	.933 (.931, .935)	.924 (.924, .925)
AUPR	**.667 (.664, .670)**	.662 (.655, .669)	.637 (.635, .640)
IDRiD			
AUROC	.747 (.731, .763)	**.790 (.778, .802)**	.726 (.717, .734)
AUPR	.461 (.439, .482)	**.498 (.481, .514)**	.432 (.413, .452)
PAPILA			
AUROC	.791 (.778, .803)	**.816 (.799, .834)**	.792 (.782, .803)
AUPR	.637 (.628, .646)	**.671 (.646, .696)**	.628 (.615, .641)
JSIEC			
AUROC	.960 (.957, .963)	**.977 (.974, .979)**	.968 (.966, .970)
AUPR	.651 (.631, .670)	**.760 (.746, .773)**	.707 (.690, .725)
Retina			
AUROC	.781 (.774, .789)	.807 (.799, .815)	**.814 (.804, .823)**
AUPR	.594 (.580, .608)	**.632 (.615, .648)**	.626 (.612, .639)

with Φ_{strong}. However, when medical augmentation Φ_{med} is introduced, the model's performance decreases in most cases. On MESSIDOR-2 and IDRiD, $\Phi_{weak+med}$ reduces the model's performance by 2.5% and 6.4%, respectively, even making it lower than with Φ_{strong}.

For external evaluation, as shown in Table 3, Φ_{weak} also achieved a marginal improvement in terms of model generalizability compared to Φ_{strong}. When the model trained on IDRiD was externally evaluated on APTOS2019 and MESSIDOR-2, the model pre-trained with Φ_{weak} outperforms Φ_{strong} by 0.4% and 1.8%, respectively.

4. CONCLUSION

In this study, we aimed to improve the contrastive learning performance in the medical image domain. We proposed a hypothesis that the dense distribution of medical images might cause the suboptimal performance of contrastive learning, and validated in our experiments and validation. Our findings suggest that simply reducing augmentation scales to an appropriate level can improve the clustering performance and therefore enhance model performance. Additionally, when incorporating medical-specific augmentation Φ_{med} to Φ_{weak}, the collective augmentation again decreases Dis($\mathcal{P}^-$), while increase Dis($\mathcal{P}^+$), generating adverse effects on model performance. These offer key guidance into the model pre-training with contrastive learning for medical images.

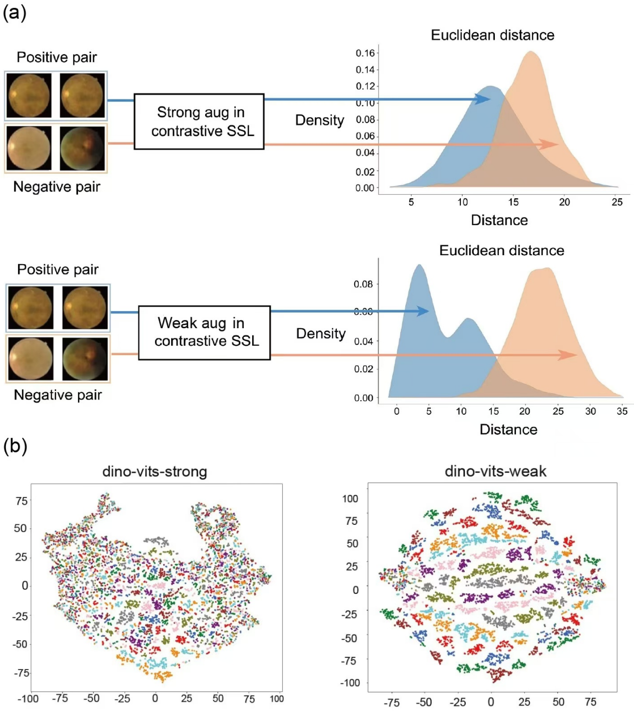

Fig. 2: We extract features using the DINO teacher model (encoder), pre-trained separately with strong and weak augmentations. First, we calculate the Euclidean distances between positive and negative pairs and compare their distance distributions in Figure (a). We also use a t-SNE map to visualize the feature clustering in Figure (b), where different colors represent augmented views from different images.

Although bringing insights, we acknowledge several limitations in this work that should be studied in future work. First, we only validated our hypothesis and solution on DINO; more contrastive learning strategies, such as DINOv2 [5], could be investigated. Second, some quantitative metrics describing the clustering performance have not been investigated, which will be proposed in future work to guide the augmentation scaling. Finally, some techniques like tailored loss functions adjusting the weights on positive and negative pairs will be studied. This work pioneered the optimization of contrastive learning in the medical domain and encouraged the tailored model training settings for medical images.

5. COMPLIANCE WITH ETHICAL STANDARDS

Only de-identified retrospective data were used in this research, without the active involvement of patients. All datasets used in the downstream applications are publicly available.

6. ACKNOWLEDGMENT

No funding was received for conducting this study. The authors have no relevant financial or non-financial interests to disclose.

7. REFERENCES

[1] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo, "Detco: Unsupervised contrastive learning for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 8392–8401.

[2] Jiaqi Zeng and Pengtao Xie, "Contrastive self-supervised learning for graph classification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, pp. 10824–10832, May 2021.

[3] Ankit Singh, Omprakash Chakraborty, Ashutosh Varshney, Rameswar Panda, Rogerio Feris, Kate Saenko, and Abir Das, "Semi-supervised action recognition with temporal contrastive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 10389–10399.

[4] Mathilde et al. Caron, "Emerging Properties in Self-Supervised Vision Transformers," May 2021, arXiv:2104.14294 [cs].

[5] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski, "Dinov2: Learning robust visual features without supervision," 2024.

[6] Yukun et al. Zhou, "A foundation model for generalizable disease detection from retinal images," *Nature*, vol. 622, pp. 156–163, Oct. 2023.

[7] Kaiming et al. He, "Masked Autoencoders Are Scalable Vision Learners," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, June 2022, pp. 15979–15988, IEEE.

[8] Yang Wen, Leiting Chen, Yu Deng, and Chuan Zhou, "Rethinking pre-training on medical imaging," *Journal of Visual Communication and Image Representation*, vol. 78, pp. 103145, 2021.

[9] Michael L Joy, "Mr current density and conductivity imaging: the state of the art," in *Conference Proceedings of the IEEE Engineering in Medicine and Biology Society*, 2004, pp. 5315–5319.

[10] Richard Legras, Alain Gaudric, and Kelly Woog, "Distribution of cone density, spacing and arrangement in adult healthy retinas with adaptive optics flood illumination," *PloS one*, vol. 13, no. 1, pp. e0191141, 2018.

[11] S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers, "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 820–838, 2021.

[12] Niall Patton, Tariq M. Aslam, Thomas MacGillivray, Ian J. Deary, Baljean Dhillon, Robert H. Eikelboom, Kanagasingam Yogesan, and Ian J. Constable, "Retinal image analysis: Concepts, applications and potential," 2006.

[13] E. Decencière *et al.*, "Feedback on a publicly distributed image database: the messidor database," *Image Analysis & Stereology*, vol. 33, no. 3, pp. 231–234, 2014.

[14] "Indian diabetic retinopathy image dataset (idrid)," https://ieee-dataport.org/open-access/indian-diabetic-retinopathy-image-dataset-idrid.

[15] "Aptos 2019 blindness detection," https://www.kaggle.com/competitions/aptos2019 blindness-detection/data.

[16] O. Kovalyk, J. Morales-Sánchez, R. Verdú-Monedero, et al., "Papila: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment," *Scientific Data*, vol. 9, no. 291, 2022.

[17] L.P. Cen, J. Ji, J.W. Lin, et al., "Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks," *Nature Communications*, vol. 12, pp. 4828, 2021.

[18] "Cataract dataset," https://www.kaggle.com/datasets/jr2ngb/cataractdataset.

[19] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.