Metadata-enhanced contrastive learning from retinal optical coherence tomography images

Robbie Holland[a], Oliver Leingang[b], Hrvoje Bogunović[b,c], Sophie Riedl[b], Lars Fritsche[d], Toby Prevost[e], Hendrik P. N. Scholl[f,g], Ursula Schmidt-Erfurth[b], Sobha Sivaprasad[h,i], Andrew J. Lotery[j], Daniel Rueckert[a], and Martin J. Menten[a,k], on behalf of the PINNACLE consortium

[a]*BioMedIA, Imperial College London, London, United Kingdom*
[b]*Laboratory for Ophthalmic Image Analysis, Medical University of Vienna, Vienna, Austria*
[c]*Christian Doppler Lab for Artificial Intelligence in Retina, Medical University of Vienna, Vienna, Austria*
[d]*Department of Biostatistics, University of Michigan, Ann Arbor, MI, United States*
[e]*Nightingale-Saunders Clinical Trials & Epidemiology Unit, King's College London, London, United Kingdom*
[f]*Institute of Molecular and Clinical Ophthalmology Basel, Basel, Basel-Stadt, Switzerland*
[g]*Department of Ophthalmology, Universitat Basel, Basel, Basel-Stadt, Switzerland*
[h]*Institute of Ophthalmology, University College London, London, United Kingdom*
[i]*Moorfields National Institute for Health and Care Biomedical Research Centre, Moorfields Eye Hospital, London, United Kingdom*
[j]*Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, Southampton, Hampshire, United Kingdom*
[k]*Institute for AI and Informatics in Medicine, Technical University of Munich, Munich, Bavaria, Germany*

Abstract

Deep learning has potential to automate screening, monitoring and grading of disease in medical images. Pretraining with contrastive learning enables models to extract robust and generalisable features from natural image datasets, facilitating label-efficient downstream image analysis. However, the direct application of conventional contrastive methods to medical datasets introduces two domain-specific issues. Firstly, several image transformations which have been shown to be crucial for effective contrastive learning do not translate from the natural image to the medical image domain. Secondly, the assumption made by conventional methods, that any two images are dissimilar, is systematically misleading in medical datasets depicting the same anatomy and disease. This is exacerbated in longitudinal image datasets that repeatedly image the same patient cohort to monitor their disease progression over time. In this paper we tackle these issues by extending conventional contrastive frameworks with a novel metadata-enhanced strategy. Our approach employs widely available patient metadata to approximate the true set of inter-image contrastive relationships. To this end we employ records for patient identity, eye position (i.e. left or right) and time series information. In experiments using two large longitudinal datasets containing 170,427 retinal optical coherence tomography (OCT) images of 7,912 patients with age-related macular degeneration (AMD), we evaluate the utility of using metadata to incorporate the temporal dynamics of disease progression into pretraining. Our metadata-enhanced approach outperforms both standard contrastive methods and a retinal image foundation model in five out of six image-level downstream tasks related to AMD. We find benefits in both a low-data and high-data regime across tasks ranging from AMD stage and type classification to prediction of visual acuity. Due to its modularity, our method can be quickly and cost-effectively tested to establish the potential benefits of including available metadata in contrastive pretraining.

---

1. Introduction

Deep learning promises to revolutionise medical imaging by automating screening, diagnosis and monitoring of diseases (Davenport and Kalakota, 2019). Traditionally, the successful application of supervised deep learning has required large amounts of labelled data for training. This does not translate well to the medical domain where labelled images are scarce as their annotation typically requires the time and cost of a trained clinician. ImageNet, commonly used to train and benchmark algorithms for natural image processing, contains millions of labelled images (Deng et al., 2009), while the median size of datasets featured in the Medical Image Computing & Computer Assisted Intervention (MICCAI) conference in 2019 ranged from 120 to 180 subjects (Kiryati and Landau, 2021). There is a large disparity between the growing amount of medical data and the resources needed to annotate it for the research and development of deep learning tools in medicine (Willemink et al., 2020).

Pretraining with contrastive frameworks enables deep learning models to reach the same level of performance using fewer labelled training samples (Chen et al., 2020b). They work by contrasting two augmented views of each image per batch, assuming their similarity depends on whether or not they originate from the same image. However, they are prone to provide misleading learning signals when the true set of inter-image relationships are unknown. As illustrated in (Khosla et al., 2020), removing negative pairs of images from the same ImageNet class improves downstream performance. In contrast, large volumes of medical images are not typically annotated with gold-standard class labels for disease stage or type, but are commonly accompanied by metadata such as the date of the medical scan and the patient's anonymised identity. This widely available information has the potential to indicate the true set of inter-image relationships used in contrastive frameworks in the medical domain.

Optical coherence tomography (OCT) enables low-cost, non-invasive imaging of the eye. This has led to the accumulation of large, metadata-enriched retinal datasets that approach the size of those for natural images. Hence, ophthalmology is ideally suited for adapting advances in self-supervised learning to the medical domain. A particularly challenging task is the diagnosis and prognosis of age-related macular degeneration (AMD) in retinal OCT. AMD, the leading cause of irreversible blindness in the elderly, is a progressive disease and is that is projected to increase in prevalence by nearly 50% from 196 million cases worldwide in 2020 to 288 million by 2040 (Wong et al., 2014).

In this paper we extend contrastive pretraining with strategies leveraging time series of widely available metadata in longitudinal retinal imaging datasets (see Figure 1). We conduct our analysis across seven diverse downstream tasks related to AMD in two large datasets totalling 170,427 OCT images of 7,912 patients collected in the scope of the PINNACLE study (Sutton et al., 2022). Overall, our contributions and key findings include:

- We adapt the set of transformations and pretraining protocols for two standard contrastive pretraining methods in datasets of retinal OCT images.

- We then introduce metadata-enhanced learning, leveraging widely available patient metadata to address known issues with standard contrastive methods. To this end, our approach flexibly encodes known temporal dynamics of disease progression into pretraining using widely

available metadata. Specifically, we use the scan date, anonymous patient identifier and eye laterality to create more informative contrastive positive pairs that correct misleading false negative pairs that arise systematically in standard contrastive frameworks.

- We found that models pretrained with standard contrastive methods consistently outperformed models pretrained on ImageNet and RadImageNet, and matched a retinal foundation model trained on as many as 15x retinal OCT images. Moreover, models pretrained with metadata-enhanced learning, redefining contrastive relationships between longitudinal images acquired within 1 year intervals, consistently outperformed both standard contrastive methods and the retinal foundation model. Notably, to classify Healthy vs. Early AMD and Late vs. Early AMD metadata-enhanced extensions required 20x and 100x fewer labelled data, respectively, to recover the performance of the foundation model.

2. Related Work

2.1. Self-supervised learning

Self-supervised pretraining has gained momentum in recent years as a method for creating generalisable representations of unlabelled data. Pretrained models can then be finetuned on supervised tasks resulting in faster training and requiring fewer annotations to reach good performance. It works by solving so-called pretext tasks that create learning signals without human supervision. Initially pretext tasks were derived from ad-hoc heuristics such as predicting image rotation (Komodakis and Gidaris, 2018) and colourisation (Zhang et al., 2016) but these were surpassed by more general context matching tasks such as solving jigsaw puzzles (Noroozi and Favaro, 2016) and predicting one local context from another (Oord et al., 2018).
Contrastive frameworks (Chen et al., 2020a; He et al., 2020; Grill et al., 2020; Chen and He, 2021) currently advance the state-of-the-art in self-supervised learning. In their seminal work, Chen et al. 2020a proposed SimCLR which creates a context matching task by combining a set of image transformations with a contrastive loss. They define similarity at the image level, such that two augmented views of the same image create a positive contrastive pair and views from different images make negative pairs. Models then learn to maximise similarity in representation space between positive pairs, and minimise similarity between negative pairs. They find that a diverse set of transformations, including random crops, rotations and colour shifts, helps in building more generalisable representations than could be learned from any one pretext task.
Khosla et al. 2020 show that standard contrastive learning on ImageNet is degraded by misleading negative pairs consisting of images belonging to the same class, and restore performance by instead making these pairs positive. In Bootstrap Your Own Latent (BYOL) Grill et al. 2020 remove negative pairs altogether and employ imbalanced student and teacher networks to prevent a degenerate solution. Overall, choosing an appropriate and diverse set of transformations as well as the correct handling of false negative pairs are crucial for the successful application of contrastive learning.

2.2. Self-supervised learning in the medical domain

Early applications of self-supervised learning to medical images also relied on ad-hoc pretext tasks such as image registration (Li and Fan, 2017) or region-specific tasks such as locating anatomical landmarks in the heart (Bai et al., 2019), distance between random patches in the brain (Spitzer et al., 2018) and estimating retinal thickness in the eye (Holmberg et al., 2020). Chen et al. 2019 use a generic context restoration task to improve plane detection in ultrasound, landmark detection
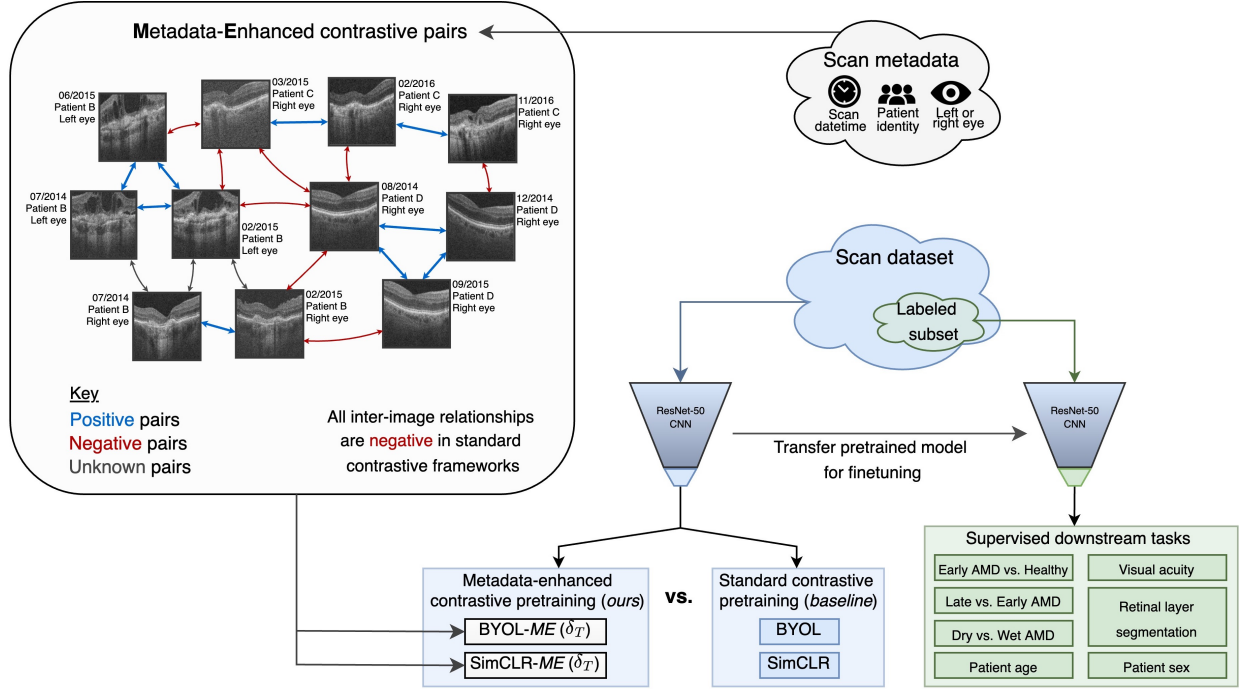
Figure 1: In this paper we pretrain models with existing contrastive frameworks (BYOL and SimCLR) and our metadata-enhanced contrastive versions on large datasets of unlabelled OCT images. Our method addresses existing weaknesses of standard frameworks by leveraging metadata widely available in the clinical workflow. To this end we employ information of patient identity, eye position (i.e. left or right) and time series information to indicate the set of true inter-image contrastive relationships. We benchmark pretraining strategies by quantifying improvements on seven downstream tasks related to the clinical assessment of AMD.

in CT and tumour segmentation in MRI.

There are very few applications of standard contrastive learning to medical datasets (Ghesu et al., 2022). Many transformations used to generate views of natural images are not compatible with medical images. In contrast to colour images medical scans are often single-channel rendering two of the most performant augmentations in SimCLR, colour shift and greyscale, inapplicable. Additionally, medical scans often have a strong prior on their content. For example, studies typically collect scans targeting a single anatomical region or medical condition. Views generated from different images that feature similar disease-related or anatomical content will nonetheless form misleading negative pairs and degrade the learning signal.

To address these problems others have used domain specific information. Chen et al. 2021 argue that views generated from any two frames of the same ultrasound video are highly correlated. By switching their contrastive relationship from negative to positive they show improvement in detection of COVID-19 and pneumonia. Others generate more informative positive pairs by choosing bilateral views to improve classification of conditions in dermatological photographs (Azizi et al., 2021), multi-modal views between genetic information and fundus images for classifying cardiac disease (Taleb et al., 2022), views from the same temporal longitude for improving lung segmentation (Zeng et al., 2021a) and views with the same disease label for classifying pleural effusion in chest X-ray (Vu et al., 2021). Two studies create positive pairs from spatially proximal 2D slices to improve volumetric segmentation of cardiac MRI (Zeng et al., 2021b; Chaitanya et al., 2020).

Ciga et al. 2022 limit negative pairs to histopathological images with different staining and tissue types for improved detection of cancer.

2.3. Deep learning for retinal imaging and AMD

Most existing applications of deep learning to retinal images are fully supervised (Bogunović et al., 2017; Lee et al., 2017; Schmidt-Erfurth et al., 2018; De Fauw et al., 2018). There have been very few applications of self-supervised learning to retinal images. Holmberg et al. 2020 use a multi-modal approach to improve detection of retinopathy by pretraining models to use fundus photographs to predict retinal thickness maps derived from OCT. Srinivasan et al. 2021 find that contrastive pretraining on ImageNet boosted downstream analysis of diabetic retinopathy in small fundus imaging datasets.

Self-supervised pretraining for AMD has so far included one work using fundus photographs (Yellapragada et al., 2022) and another employing a temporal ordering task to learn late AMD features from longitudes in OCT (Rivail et al., 2019). More recently (Zhou et al., 2023) released RETFound, a foundation model for retinal images trained on over 700,000 OCT images. However, there are no studies comprehensively testing the potential role of metadata for improving self-supervised learning in the domain of retinal OCT.

3. Materials and Methods

Our proposed method is summarised in Figure 1. In this section we describe the two longitudinal retinal OCT datasets in section 3.1. We then revisit standard contrastive learning and introduce our proposed metadata-enhanced framework in section 3.2. Finally, we specify experiments to evaluate our pretrained models and the benefit of our metadata-enhanced modifications in section 3.3.

3.1. OCT datasets

In this paper we use two datasets collected as part of the PINNACLE project (Sutton et al., 2022). They each contain longitudinal data consisting of retinal OCT scans of patients with AMD. For both datasets images were acquired using Topcon 3D OCT devices (Topcon Corporation, Tokyo, Japan). The first dataset, collected at the Southampton Eye Unit, consists of 48,825 OCT scans and followed 6,368 eyes from 3,498 patients. A subset of 2,037 images are healthy control scans curated from 570 patients. In Figure 3a we show the number of times eyes were scanned. The average eye was scanned 7.7 times over an average duration of 1.9 years, with longitudes featuring AMD spanning 2.3 years. The second dataset, collected at Moorfields Eye Hospital, is larger containing 121,602 OCT scans of 7,336 eyes from 3,844 patients. The longitudinal dimension is also greater, with longitudes containing on average 16.6 scans spanning 3.5 years. We extracted the mediolateral 2D slice centred at the fovea and resampled to $208{\times}256$ pixels with a pixel size of $7.0{\times}23.4\ \mu m^2$, half the median resolution. We elect to maintain these as separate datasets to test whether any benefits from our pretraining methods generalise to different hospitals and collection sites.

Each image is accompanied by metadata detailing which eye (i.e. left or right eye) is depicted, the patient's anonymised identity and the exact date and time of scanning. This metadata was reserved for metadata-enhanced pretraining while other demographic information recording the patient's age and sex are instead used for downstream evaluation.
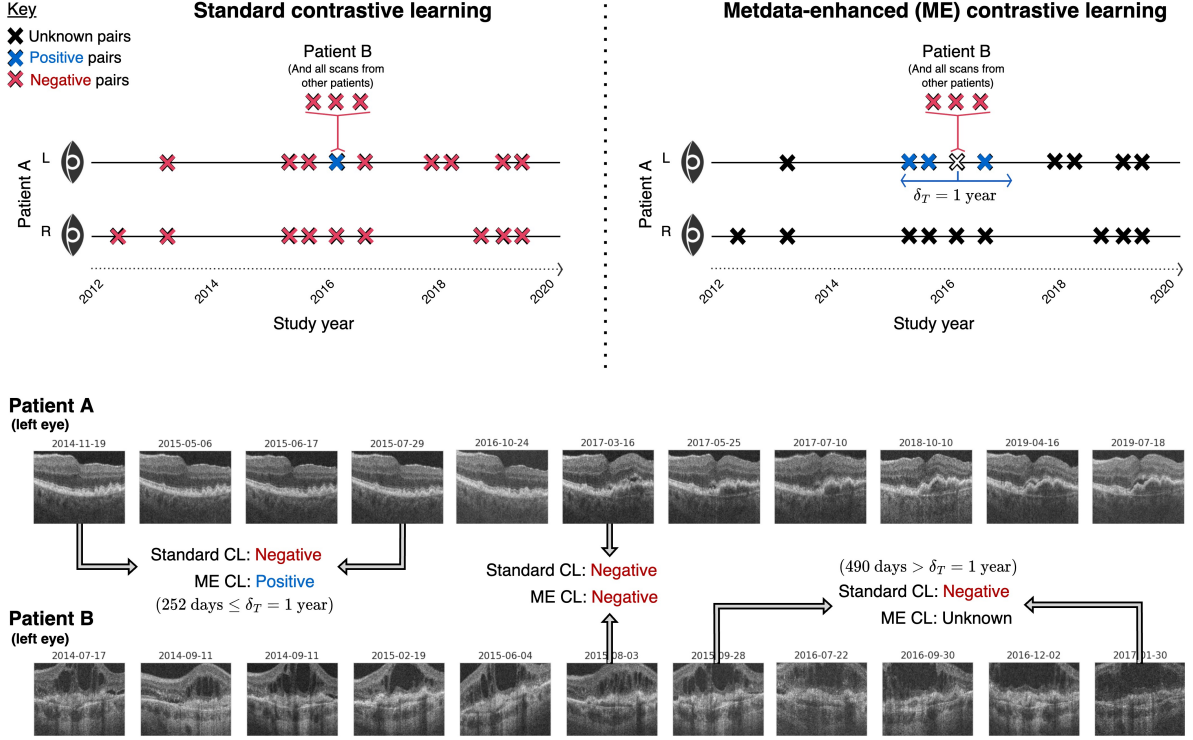
Figure 2: Our method enhances standard contrastive pretraining with widely available medical metadata to correct many of the misleading negative pairs that arise systematically in standard frameworks. Moreover, by introducing inter-image positive pairs we combine artificial image transformations with natural ones that already exist between images acquired closely in time (controlled by a δ_T parameter). Our method also removes contrastive pairs with unknown relationships and retains negative pairs containing images originating from different patients.

(a) Distribution of longitude lengths

(b) Distribution of time intervals between scans of the same longitude
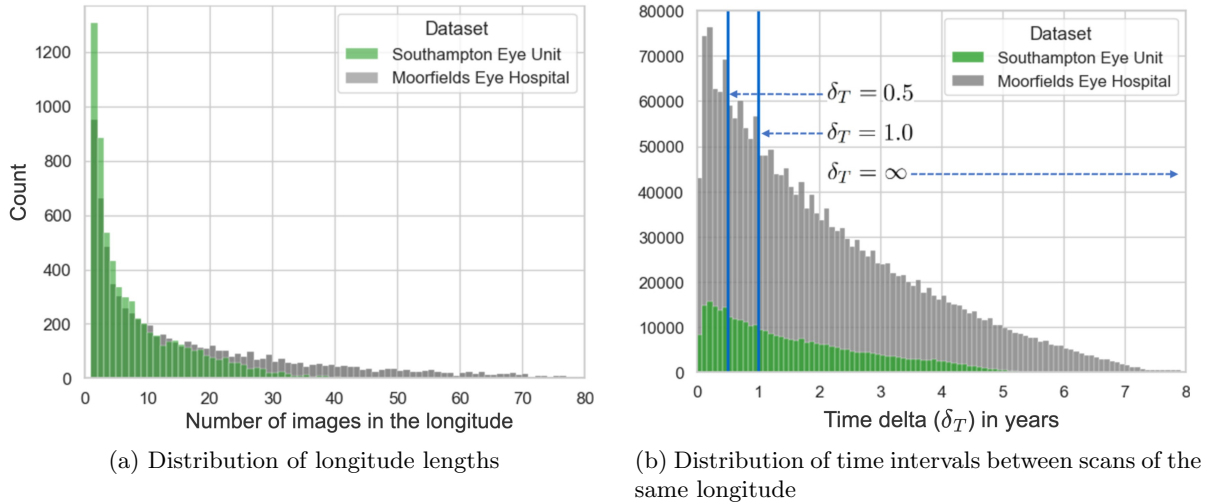
Figure 3: The distribution of longitude lengths (left) in the Southampton and Moorfields datasets and the frequency distribution of time intervals δ_T in years between all pairs of longitudinal scans from the same eye (right), with the width of each bin covering a duration of one month.

6

AMD stage labels, indicating whether a patient has early or late stage AMD, were extracted automatically from the electronic health records. For 39% of scans it was possible to extract an AMD stage label in this manner, and the remaining 61% were left unlabelled. Late stage AMD was further characterised into either wet (choroidal neovascularization) or dry (geographic atrophy) AMD. Clinical information also included visual acuity, a metric for assessing functional quality of vision, which measured the logarithm of the minimum angle of resolution (LogMAR) perceptible to the patient. We rescale this measurement via a linear mapping to find the Letter Score. This is a more intuitive metric which indicates how many letters, from 0 to 95, could be read by the patient on a letter chart (Beck et al., 2003). Finally, we obtained segmentations of eleven retinal layers using the Iowa Reference Algorithms (Retinal Image Analysis Lab, Iowa Institute for Biomedical Imaging, USA), a graph-cut-based segmentation algorithm (Li et al., 2005; Garvin et al., 2009; Abràmoff et al., 2010).

3.2. Metadata-enhanced contrastive learning

In this study, we first extend SimCLR (Chen et al., 2020a) as it is the most widely adopted and paradigmatic contrastive method (Chaitanya et al., 2020; Azizi et al., 2021; Ciga et al., 2022; Taleb et al., 2022). For a second standard baseline we include BYOL (Grill et al., 2020) for its exclusive use of positive pairs. As alluded to in section 2.2 standard contrastive learning suffers from two issues in medical applications. The first is that several transformations used by SimCLR and BYOL, such as hue, saturation and colour dropping, are not applicable in single-channel medical scans. The second is that many negative contrastive pairs will in fact feature highly similar views since all OCT images depict the same anatomy and limited set of pathologies, especially if both images originate from the same eye or patient.

To address these issues we redefine contrastive inter-image relationships as depicted in Figure 2. To this end we leverage the longitudinal medical metadata described in section 3.1. Let L longitudinal scans of an eye with position l (i.e. left or right eye) from patient u be $^u e^l = \{(^u x_i^l, {}^u t_i^l)\}_{i=1}^L$ where x_i and t_i denote the scan image data and acquisition timestamp, respectively. We define two scans as being in temporal proximity if the time between their acquisition dates falls between δ_T^{min} and δ_T. Finally, we define a function S for the contrastive relationship between any two scans as:

$$S_{\delta_T^{min},\,\delta_T}(^u x_i^l, {}^v x_j^m) = \begin{cases} \boldsymbol{+} & (u = v) \wedge (l = m) \wedge (\delta_T^{min} \leq |t_i^u - t_j^v| \leq \delta_T) \\ \boldsymbol{-} & u \neq v \\ \boldsymbol{?} & \text{otherwise} \end{cases} \tag{1}$$

This equation encodes three assumptions. Firstly, we assume that longitudinal images of the same eye acquired within a set time window δ_T differ minimally in their progression of AMD and healthy ageing (see Figure 2). These cases are labelled as positive pairs for the purpose of contrastive learning. To observe evidence of this assumption refer to Figure A.9. The second assumption states that pairs of images from different individuals are likely to be dissimilar and we label these pairs as negative accordingly. Furthermore, scans outside the δ_T window may be similar in cases of slow AMD progression but semantically different in cases of fast disease progression. Consequently, we remove these as contrastive pairs altogether. Similarly, disease progression is known to be correlated in fellow eyes with a temporal delay. Therefore, we also remove pairs of images from fellow eyes due to their unknown relationship. The third assumption states that pairs of images from different individuals are more likely to be dissimilar with respect to disease state, and we label these pairs as negative accordingly.

7

PyTorch Code 1: Metadata-enhanced wrapper for standard contrastive losses. This function uses the patient identifier, eye side, and scan date of each image to construct more informative positive and negative pairs for standard contrastive methods. This customizable approach allows for flexible encoding of in-domain knowledge into pretraining. In this paper, we choose to define positive pairs as images of the same eye acquired within a short temporal interval defined by δ_T. We also use metadata to exclude negative pairs of images originating from either eye of the same patient.

```python
def standard_contrastive_loss(positive_pairs, negative_pairs):
    # Any contrastive loss such as SimCLR, BYOL, MoCo, DINO...
    ...

def metadata_enhanced_contrastive_loss(embeddings, metadata, d_T_min, d_T):
    patient_ids, eye_sides, scan_dates = metadata

    # Use metadata to identify positive and negative pairs
    # (excluding unknown pairs by default)
    time_intervals = (scan_dates - scan_dates.T).abs()
    temporal_mask = (d_T_min < time_intervals) & (time_intervals <= d_T_max)
    same_patient_mask = (patient_ids == patient_ids.T)
    same_eye_mask = same_patient_mask & (eye_sides == eye_sides.T)

    positive_pairs = embeddings[torch.where(same_eye_mask & temporal_mask)]
    negative_pairs = embeddings[torch.where(~same_patient_mask)]

    # Calculate contrastive loss as normal on redefined pairs
    loss = standard_contrastive_loss(positive_pairs, negative_pairs)
    return loss
```

Using Equation 1 to redefine inter-image relationships, we introduce metadata-enhanced (ME) variants of BYOL and SimCLR as BYOL-ME(δ_T) and SimCLR-ME(δ_T), respectively. To create augmented views we retain the set of transformations used in standard contrastive learning. However, instead of creating two views from each image we now generate positive pairs from two related images by augmenting each once. When creating batches we ensure that every image is part of at least one positive pair. While we chose SimCLR and BYOL in this paper as two of the most widely adopted and researched contrastive frameworks, our positive and negative pair reformulation is applicable to any contrastive framework and can be calibrated for any progressive disease.

Our metadata-enhanced framework can be implemented by a simple extension to existing contrastive losses (see PyTorch Code 1). By redefining positive and negative pairs using a customisable set of rules, existing pipelines can be modified at minimal cost.

3.3. Experimental protocol

3.3.1. Contrastive pretraining protocol

Both standard and metadata-enhanced contrastive pretraining variants, for both SimCLR, BYOL, use a ResNet50 (4x) backbone with half precision, and were randomly initialised before training according to the He protocol (He et al., 2016). In both datasets we allocate 4,800 samples for validation, 4,800 for testing and the remainder for pretraining. Models then train with the AdamW optimiser using a learning rate of $5 \cdot 10^{-4}$ and momentum of 0.9 for a fixed number of 120,000 optimisation steps. We use linear warmup for the first 1,200 steps and decay the learning rate with the cosine schedule without restarts. The batch size is 384 and we use the model after the final training step for downstream evaluation. Additional details are listed in Appendix A.2. Due to the sensitivity of contrastive frameworks to the chosen transformations we closely follow those used by BYOL (Grill et al., 2020), transferring as many as are applicable to retinal OCT. We first, with probability 0.8, augment each scan by varying image brightness and contrast using maximum relative changes of 0.4. Then we rotate by a random angle up to 8 degrees, crop centrally to resolution 188×236 to remove rotation border artefacts, flip horizontally with probability 0.5, randomly crop to a scale between 0.25 and 1.0 with aspect ratio between 3/4 and 4/3, and finish by resizing to 192×192. We chose not to use Gaussian blur as we, in accordance with (Azizi et al., 2021), found it obfuscated local textural features in the retinal images. We also omit hue, saturation and colour dropping as these augmentations do not have any effect on greyscale images. In our experiments we first pretrain multiple ResNet50 models using both standard SimCLR and BYOL in addition to their metadata-enhanced extensions defined in section 3.2. For both SimCLR-ME and BYOL-ME we test three values of δT: 0.5 years, 1.0 years and no time limit (i.e. all available longitudinal scans). These values result in including 18.8%, 35.1% and 100% of the potentially positive pairs respectively, or 469,334, 874,029 and 2,493,568 in total as can be seen in Figure 3b. We set δ_T^{min} to 0.02 years to exclude poor quality pairs that are likely due to a scan retake.

3.3.2. Pretrained baselines

To fairly assess the benefit of metadata-enhanced learning we compare against three pretrained baselines. The first baseline is a model pretrained by (Chen et al., 2020a) in SimCLR on natural images in ImageNet. Our second baseline uses a more medically aligned model pretrained to classify 165 different conditions on RadImageNet (Mei et al., 2022). We also compare our methods against RETFound, a foundation model for retinal images trained on over 700,000 OCT images (Zhou et al., 2023). RETFound was trained as a Masked Autoencoder (He et al., 2022) and uses a ViT-Large backbone. We use the version specifically trained on OCT data, which was found to be highly performant in retinal disease prediction, including AMD, in external datasets. These baselines, pretrained on natural images, medical images and external retinal OCT images, offer increasing levels of in-domain pretraining against which to compare in-dataset contrastive learning, and our metadata-enhanced extensions.

3.3.3. Finetuning protocol

To compare different pretrained models we adopt the standard linear evaluation finetuning protocol used by (Chen et al., 2020a). We first freeze the pretrained encoder and replace any dense projection layers used in pretraining with a single linear layer that projects from the latent space to the dimension of the supervised label. More specifically, for the segmentation task this involves a 11x192 dimension matrix and for all other problems the output is size 1. For classification problems

9

we then apply a sigmoid activation function and for segmentation and regression problems we use tanh activation and scale labels to the range [-1,1]. We use a batch size of 1024 and ensure we use the same train, validation and test splits as for pretraining. We then train all models for 3000 epochs using the early stopping if downstream validation performance does not improve for 600 steps. To fit the model we use an AdamW optimiser with momentum of 0.9, weight decay of $1 \cdot 10^{-2}$ and a learning rate of $5 \cdot 10^{-4}$ using the cosine annealing schedule after 50 linear warmup epochs. These hyperparameters are the same as were used to evaluate RETFound in (He et al., 2022). Moreover, for all models pretrained on external dataset we also resize the input image to the size they were trained on, which was 224x224. In training we employ a weaker form of the augmentations described in the pretraining protocol. Finally, we report performance on the test set of the model checkpoint with the best validation downstream performance.

3.3.4. Downstream tasks

To evaluate pretrained models, including pretrained baselines and models pretrained with standard and metadata-enhanced contrastive learning, we test them on the following seven downstream tasks:

1. **AMD stage and type classification tasks** to test the ability of pretrained models to learn disease progression and disease type. The following tasks are tested:
   (a) Early/Intermediate AMD vs. Healthy eye classification.
   (b) Late vs. Early/Intermediate AMD classification.
   (c) Late stage Dry vs. Wet AMD classification.
2. **Prediction of functional vision assessment score**, which involves learning any clinical features visible in the retina that may degrade the patient's quality of vision.
3. **Prediction of patient demographic information**, including patient sex classification and age regression to incorporate learning healthy ageing and demographic characteristics.
4. **Dense retinal surface segmentation**, where models simultaneously regress the column-wise location of eleven retinal cell layers along each OCT A-scan as in (He et al., 2021). Unlike the aforementioned image-level tasks, this task requires learning pixel-level structural variations which are known to be affected by pathology, healthy ageing and differences in patient sex.

For all tasks we tested scenarios where fewer labelled samples are available for training. To do this we sample exponentially increasing amounts of labelled data for training, starting from just 20 training samples and increasing to 10,000 (for subset selection protocol refer to Appendix A.4). Performance on classification tasks is measured in area under the receiver operator characteristic curve (AUC) and regression and segmentation tasks in mean average error (MAE). We repeat all downstream experiments using five different seeds and report performance in a low-data regime finetuned on 100 labelled samples, or 1% of the data, against a full high-data regime using a maximum of 10,000 labelled samples. Our entire workflow and results are repeated across both retinal OCT datasets. In total our results summarise 4,620 finetuning runs evaluating 11 different pretraining strategies.

4. Results

4.1. Self-supervised pretraining on retinal OCT data is crucial for downstream performance

Models pretrained on retinal OCT images markedly outperformed models pretrained on natural images such as ImageNet, as shown in Table 1 and Figure 4. Features extracted from RadImageNet, containing 1.35 million radiological images, were more transferable than those from ImageNet, but were still outperformed by all retinal OCT pretrained models. Full end-to-end finetuning (shown in Table A.2) enables models with out-of-domain pretraining to reduce the disparity in performance with models pretrained on OCT images. Furthermore, standard contrastive methods pretrained on our datasets match the performance of RETFound despite using 13x fewer parameters and training on as much as 15x fewer unlabelled images. For example, in age regression on the Southampton dataset, standard BYOL underperforms compared to RETFound in the low-data regime of (7.21 vs. 6.96 MAE) but performs better in the high-data regime (5.91 vs. 6.18 MAE).

In linear evaluation, RETFound, which was trained as a Masked Autoencoder to reconstruct images on a patch-level, outperformed all other methods on retinal layer segmentation. This reflects other works suggesting that image-level representation learning tasks, such as contrastive learning, do not generalise well to dense pixel-level tasks like segmentation (Wang et al., 2021; Li et al., 2022; Hu et al., 2021; Chaitanya et al., 2020).

Table 1: We compare linear evaluation performance of different pretraining strategies across seven retinal downstream tasks on the Southampton and Moorfields datasets. We find metadata-enhanced pretraining strategies outperformed all other methods in all but one image-level task in both the low-data and high-data regime on both datasets. Best performance is highlighted in bold, and second best is underlined.

Southampton Eye Unit

Pretraining variant		Early vs. Healthy (AUC) ↑		Late vs. Early (AUC) ↑		Visual acuity (MAE Letters)↓		Dry vs. Wet (AUC) ↑		Patient Age (MAE years)↓		Patient Sex (AUC) ↑		Segmentation (MAE μm) ↓	
# finetuning labels		100	8299	100	10000	100	10000	100	10000	100	10000	100	10000	100	10000
ImageNet SimCLR		0.551	0.587	0.568	0.537	18.4	18.5	0.570	0.620	7.50	7.51	0.516	0.518	98.3	86.4
RadImageNet pretrained		0.775	0.712	0.676	0.745	20.7	15.4	0.480	0.536	7.75	6.97	0.529	0.627	177	<u>62.0</u>
RETFound (foundation model)		0.900	0.941	0.732	0.792	13.9	12.3	0.798	0.885	6.96	6.18	0.619	0.737	**59.8**	**40.1**
SimCLR	Standard	0.916	0.946	0.772	0.829	14.5	12.2	0.807	0.843	6.92	6.06	0.592	0.721	102	74.6
BYOL	Standard	0.892	0.940	0.750	0.828	14.0	11.7	0.828	0.883	7.21	**5.91**	0.587	0.746	<u>94.4</u>	70.8
SimCLR	*Ours ($\delta_T = \infty$)*	0.915	0.943	0.781	0.841	14.9	12.3	0.802	0.891	6.88	6.03	0.648	<u>0.755</u>	97.3	76.8
	Ours ($\delta_T = 1.0$)	**0.930**	<u>0.950</u>	0.770	0.832	14.4	11.9	0.827	<u>0.893</u>	<u>6.74</u>	5.97	<u>0.65</u>	0.743	98.3	74.2
	Ours ($\delta_T = 0.5$)	0.922	0.946	0.772	0.836	14.7	12.0	0.813	0.892	6.82	6.02	**0.659**	0.751	98.0	75.5
BYOL	*Ours ($\delta_T = \infty$)*	0.923	0.947	0.761	0.831	14.0	11.8	<u>0.842</u>	**0.901**	6.70	5.97	0.629	0.748	96.7	84.2
	Ours ($\delta_T = 1.0$)	<u>0.929</u>	**0.951**	<u>0.789</u>	<u>0.845</u>	<u>13.4</u>	<u>11.6</u>	**0.849**	0.873	**6.68**	5.93	0.641	**0.763**	96.1	83.7
	Ours ($\delta_T = 0.5$)	0.923	0.946	**0.790**	**0.847**	**13.0**	**11.5**	0.836	0.849	6.76	<u>5.92</u>	0.618	0.746	96.3	81.3

Moorfields Eye Hospital

| Pretraining variant		Early vs. Healthy (AUC) ↑		Late vs. Early (AUC) ↑		Visual acuity (MAE Letters)↓		Dry vs. Wet (AUC) ↑		Patient Age (MAE years)↓		Patient Sex (AUC) ↑		Segmentation (MAE μm) ↓	
ImageNet SimCLR		—	—	0.520	0.460	17.7	17.8	—	—	8.24	8.17	0.479	0.468	165	108
RadImageNet pretrained		—	—	0.549	0.524	16.7	15.5	—	—	7.57	7.27	0.497	0.559	153	88.1
RETFound (foundation model)		—	—	0.698	0.804	**13.9**	12.4	—	—	7.38	6.63	0.569	0.693	<u>89.3</u>	**54.0**
SimCLR	Standard	—	—	0.733	0.836	14.9	12.0	—	—	7.41	6.59	0.556	0.736	122	<u>73.0</u>
BYOL	Standard	—	—	0.704	0.835	15.0	**11.9**	—	—	7.34	6.61	0.582	0.754	<u>121</u>	73.3
SimCLR	*Ours ($\delta_T = \infty$)*	—	—	0.722	0.866	14.8	12.4	—	—	**6.89**	6.46	0.612	**0.79**	185	101
	Ours ($\delta_T = 1.0$)	—	—	0.748	0.864	14.7	12.1	—	—	7.05	6.47	**0.618**	<u>0.776</u>	179	92.6
	Ours ($\delta_T = 0.5$)	—	—	0.740	<u>0.871</u>	14.8	12.1	—	—	7.00	<u>6.32</u>	0.618	0.776	176	93.5
BYOL	*Ours ($\delta_T = \infty$)*	—	—	0.748	0.865	14.6	12.3	—	—	6.97	6.43	0.589	0.762	194	101
	Ours ($\delta_T = 1.0$)	—	—	<u>0.765</u>	0.868	14.5	<u>12.1</u>	—	—	7.09	6.46	0.610	0.770	190	102
	Ours ($\delta_T = 0.5$)	—	—	**0.768**	**0.876**	<u>14.3</u>	12.2	—	—	<u>6.94</u>	**6.31**	0.612	0.767	188	98.8
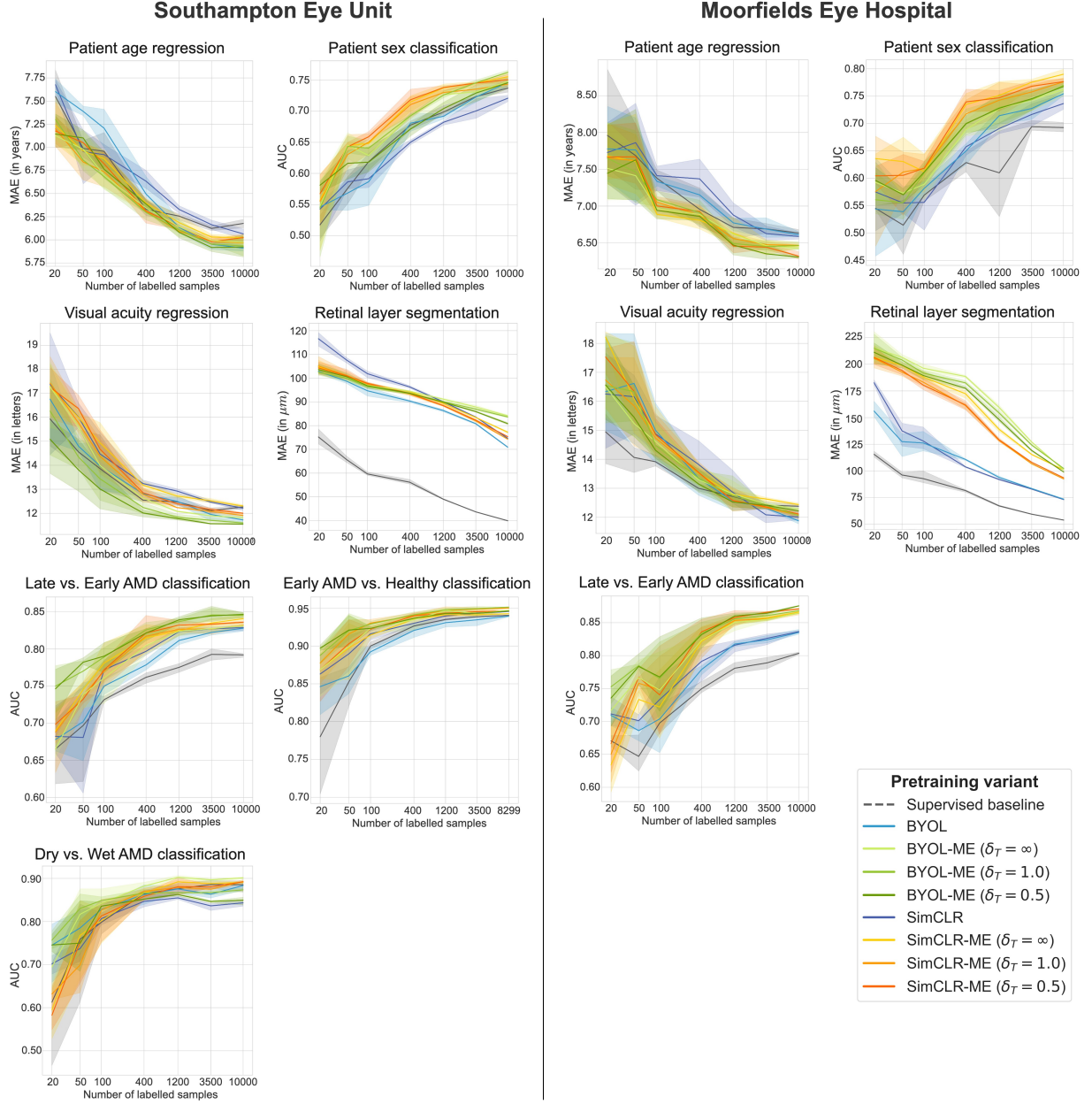
Figure 4: Results of linear evaluation on downstream tasks using seven logarithmically spaced amounts of labelled finetuning samples (with 95% confidence intervals). For clarity we omit the ImageNet and RadImageNet models, which were consistently outperformed by the RETFound baseline (model colour key in bottom right). In all tasks except segmentation, and visual acuity on Moorfields data, metadata-enhanced contrastive pretraining extending BYOL and using $\delta_T \leq 1.0$ outperforms standard contrastive learning and RETFound, especially in scenarios with fewer labelled data.

4.2. Metadata-enhanced pretraining outperforms all other forms of pretraining on image-level retinal tasks

We find that metadata-enhanced BYOL, which does not use any negative pairs, matched or outperformed metadata-enhanced SimCLR in all scenarios (except Sex classification in both datasets).

This leads us to conclude that misleading negative pairs still degrade performance, despite the removal of many using metadata. Furthermore, for BYOL we find that $\delta_T \leq 1.0$ typically performs better than $\delta_T = \infty$. For example, BYOL-ME($\delta_T \leq 1.0$) surpasses BYOL-ME($\delta_T = \infty$) in the both low and high-data regimes for Late vs. Early classification in both datasets. These results support our hypothesis that defining positive pairs over short time intervals is beneficial for image-level retinal analysis.

Overall, BYOL-ME $\delta_T \leq 1.0$ outperforms all pretrained baselines and standard contrastive learning approaches on every image-level task in the Southampton dataset, and all but one in the Moorfields dataset. For example, to classify Late vs. Early AMD on the Moorfields dataset we are able to surpass the performance of RETFound finetuned on all 10,000 labelled samples (0.80 AUC) using only 400 samples, or 2.5% (0.83 AUC). Similarly, we need only 100, or 1%, to recover the same performance on the Southampton dataset (0.79 AUC). Given all 10,000 labelled samples, our approach surpasses all baselines by achieving 0.85 AUC and 0.88 AUC in the Southampton and Moorfields datasets, respectively.

5. Discussion

In this paper we demonstrated the benefits of incorporating widely available metadata into self-supervised contrastive learning. Our approach uses metadata to address known issues of contrastive methods. To this end, we used metadata indicate the true set of inter-image contrastive relationships in two longitudinal datasets of unlabelled OCT images. Our metadata-enhanced approach surpassed both standard contrastive learning and RETFound, the foundation model for retinal OCT, in both low-data and high-data regimes. Our most performant variant, BYOL-ME($\delta_T \leq 1.0$), surpassed all pretrained baselines in all image-level tasks on the Southampton dataset, and all but one on the Moorfields dataset. Our strategy also enabled models using 100x fewer labelled samples to achieve the same performance as the most performant pretrained baselines in classifying Late vs. Early AMD. These results demonstrate that conditioning pretraining with widely available, yet typically unused, metadata can address known weaknesses of established contrastive frameworks in medical domains.

Our study represents the first to fully establish the wide ranging potential of metadata in self-supervised pretraining in retinal OCT. Previous works in medical domains (Chen et al., 2021; Azizi et al., 2021; Taleb et al., 2022; Zeng et al., 2021a; Vu et al., 2021) have either focussed on evaluating multiple downstream tasks or on varying labelled subset sizes. Our work uses seven tasks, two datasets and seven labelled subset sizes to provide a comprehensive picture of in-domain self-supervised pretraining. Furthermore while related works focused on improving a single contrastive loss, we used both SimCLR and BYOL to ensure that pretraining with metadata represents a general improvement to constrastive learning as a whole, rather than addressing the flaws of a single framework. Overall, our experiments more closely resemble those used in studies on natural images where more exhaustive evaluations are standard practise.

While our metadata-enhanced framework indicates the true contrastive relationship between two images, there exist edge cases where it will make incorrect assignments. Without access to the downstream task labels, misleading contrastive pairs containing images from different patients that, by chance, exhibit the same severity and type of AMD will not become positive pairs using our approach. Moreover, using a fixed temporal cutoff δ_T to determine positive pairs cannot encode variable rates of disease progression between eyes and individuals. A more informative set of metadata might allow for accurate negative and positive pairs between fellow eyes. This may

13

lead to further boost to performance in smaller datasets, but in ours these pairs constitute only 0.022% of the total. Moreover, for segmentation tasks that require encoding of local image features contrastive frameworks using high-level context matching tasks had limited value for the pixel-level segmentation task.

To address some of these issues future work could, for each contrastive pair of images, compare the auto-generated metadata relationship reported by Equation 1 to a fully unsupervised similarity score computed using standard contrastive pretraining. Then, by highlighting cases where both BYOL and SimCLR disagree with Equation 1 we could begin to identify any systematic errors in our formulation. Furthermore, future work can begin to investigate the potentially utility to prognostic predictions of disease progression through embedding temporal dynamics into pretraining Shen et al. (2024).

6. Conclusion

In this work we comprehensively demonstrated the benefits of incorporating medical metadata into self-supervised contrastive learning. Our metadata-enhanced extension of existing contrastive frameworks addresses known issues with contrastive methods while allowing researchers to flexibly encode the temporal dynamics of disease progression into pretraining. Using this approach we outperformed both standard contrastive pretraining and a foundation model for retinal images across two datasets on six diverse image-level tasks in both a low-data and high-data regime. We found benefits in downstream tasks relevant to the screening and management of AMD in retinal OCT, ranging from classification of AMD stage and type to estimation of functional endpoints. These promising findings motivate a shift in focus towards conditioning contrastive learning on medical datasets with other modalities, which can enable accurate retinal image analysis using few, gold-standard annotations.

7. Acknowledgements

References

Abràmoff, M.D., Garvin, M.K., Sonka, M., 2010. Retinal imaging and image analysis. IEEE reviews in biomedical engineering 3, 169–208.

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., et al., 2021. Big self-supervised models advance medical image classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3478–3488.

Bai, W., Chen, C., Tarroni, G., Duan, J., Guitton, F., Petersen, S.E., Guo, Y., Matthews, P.M., Rueckert, D., 2019. Self-supervised learning for cardiac mr image segmentation by anatomical position prediction, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 541–549.

Beck, R.W., Moke, P.S., Turpin, A.H., Ferris III, F.L., SanGiovanni, J.P., Johnson, C.A., Birch, E.E., Chandler, D.L., Cox, T.A., Blair, R.C., et al., 2003. A computerized method of visual acuity testing: adaptation of the early treatment of diabetic retinopathy study testing protocol. American journal of ophthalmology 135, 194–205.

Bogunović, H., Montuoro, A., Baratsits, M., Karantonis, M.G., Waldstein, S.M., Schlanitz, F., Schmidt-Erfurth, U., 2017. Machine learning of the progression of intermediate age-related macular degeneration based on oct imaging. Investigative ophthalmology & visual science 58, BIO141–BIO150.

Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E., 2020. Contrastive learning of global and local features for medical image segmentation with limited annotations. Advances in Neural Information Processing Systems 33, 12546–12558.

Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., Rueckert, D., 2019. Self-supervised learning for medical image analysis using image context restoration. Medical image analysis 58, 101539.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020a. A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR. pp. 1597–1607.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E., 2020b. Big self-supervised models are strong semi-supervised learners. Advances in neural information processing systems 33, 22243–22255.

Chen, X., He, K., 2021. Exploring simple siamese representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15750–15758.

Chen, Y., Zhang, C., Liu, L., Feng, C., Dong, C., Luo, Y., Wan, X., 2021. Uscl: Pretraining deep ultrasound image diagnosis model through video contrastive representation learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 627–637.

Chuang, C.Y., Robinson, J., Lin, Y.C., Torralba, A., Jegelka, S., 2020. Debiased contrastive learning. Advances in neural information processing systems 33, 8765–8775.

Ciga, O., Xu, T., Martel, A.L., 2022. Self supervised contrastive learning for digital histopathology. Machine Learning with Applications 7, 100198.

Davenport, T., Kalakota, R., 2019. The potential for artificial intelligence in healthcare. Future healthcare journal 6, 94.

De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., et al., 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature medicine 24, 1342–1350.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee. pp. 248–255.

Garvin, M.K., Abramoff, M.D., Wu, X., Russell, S.R., Burns, T.L., Sonka, M., 2009. Automated 3-d intraretinal layer segmentation of macular spectral-domain optical coherence tomography images. IEEE transactions on medical imaging 28, 1436–1447.

Ghesu, F.C., Georgescu, B., Mansoor, A., Yoo, Y., Neumann, D., Patel, P., Vishwanath, R., Balter, J.M., Cao, Y., Grbic, S., et al., 2022. Self-supervised learning from 100 million medical images. arXiv preprint arXiv:2201.01283.

Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al., 2020. Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems 33, 21271–21284.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16000–16009.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9729–9738.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

He, Y., Carass, A., Liu, Y., Jedynak, B.M., Solomon, S.D., Saidha, S., Calabresi, P.A., Prince, J.L., 2021. Structured layer surface segmentation for retina oct using fully convolutional regression networks. Medical image analysis 68, 101856.

Holmberg, O.G., Köhler, N.D., Martins, T., Siedlecki, J., Herold, T., Keidel, L., Asani, B., Schiefelbein, J., Priglinger, S., Kortuem, K.U., et al., 2020. Self-supervised retinal thickness prediction enables deep learning from unlabelled data to boost classification of diabetic retinopathy. Nature Machine Intelligence 2, 719–726.

Hu, H., Cui, J., Wang, L., 2021. Region-aware contrastive learning for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16291–16301.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D., 2020. Supervised contrastive learning. Advances in Neural Information Processing Systems 33, 18661–18673.

Kiryati, N., Landau, Y., 2021. Dataset growth in medical image analysis research. Journal of imaging 7, 155.

Komodakis, N., Gidaris, S., 2018. Unsupervised representation learning by predicting image rotations, in: International Conference on Learning Representations (ICLR).

Lee, C.S., Baughman, D.M., Lee, A.Y., 2017. Deep learning is effective for classifying normal versus age-related macular degeneration oct images. Ophthalmology Retina 1, 322–327.

Li, H., Fan, Y., 2017. Non-rigid image registration using fully convolutional networks with deep self-supervision. arXiv preprint arXiv:1709.00799 .

Li, H., Li, Y., Zhang, G., Liu, R., Huang, H., Zhu, Q., Tao, C., 2022. Global and local contrastive self-supervised learning for semantic segmentation of hr remote sensing images. IEEE Transactions on Geoscience and Remote Sensing 60, 1–14.

Li, K., Wu, X., Chen, D.Z., Sonka, M., 2005. Optimal surface segmentation in volumetric images-a graph-theoretic approach. IEEE transactions on pattern analysis and machine intelligence 28, 119–134.

Mei, X., Liu, Z., Robson, P.M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K.E., Yang, T., et al., 2022. Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. Radiology: Artificial Intelligence 4, e210315.

Noroozi, M., Favaro, P., 2016. Unsupervised learning of visual representations by solving jigsaw puzzles, in: European conference on computer vision, Springer. pp. 69–84.

Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 .

Rivail, A., Schmidt-Erfurth, U., Vogl, W.D., Waldstein, S.M., Riedl, S., Grechenig, C., Wu, Z., Bogunovic, H., 2019. Modeling disease progression in retinal octs with longitudinal self-supervised learning, in: International Workshop on PRedictive Intelligence In MEdicine, Springer. pp. 44–52.

Schmidt-Erfurth, U., Waldstein, S.M., Klimscha, S., Sadeghipour, A., Hu, X., Gerendas, B.S., Osborne, A., Bogunović, H., 2018. Prediction of individual disease conversion in early amd using artificial intelligence. Investigative ophthalmology & visual science 59, 3199–3208.

Shen, C., Menten, M.J., Bogunović, H., Schmidt-Erfurth, U., Scholl, H., Sivaprasad, S., Lotery, A., Rueckert, D., Hager, P., Holland, R., 2024. Spatiotemporal representation learning for short and long medical image time series. arXiv preprint arXiv:2403.07513 .

Spitzer, H., Kiwitz, K., Amunts, K., Harmeling, S., Dickscheid, T., 2018. Improving cytoarchitectonic segmentation of human brain areas with self-supervised siamese networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 663–671.

Srinivasan, V., Strodthoff, N., Ma, J., Binder, A., Müller, K.R., Samek, W., 2021. On the robustness of pretraining and self-supervision for a deep learning-based analysis of diabetic retinopathy. arXiv preprint arXiv:2106.13497 .

Sutton, J., Menten, M.J., Riedl, S., Bogunović, H., Leingang, O., Anders, P., Hagag, A.M., Waldstein, S., Wilson, A., Cree, A.J., et al., 2022. Developing and validating a multivariable prediction model which predicts progression of intermediate to late age-related macular degeneration—the pinnacle trial protocol. Eye , 1–9.

Taleb, A., Kirchler, M., Monti, R., Lippert, C., 2022. Contig: Self-supervised multimodal contrastive learning for medical imaging with genetics, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20908–20921.

Vu, Y.N.T., Wang, R., Balachandar, N., Liu, C., Ng, A.Y., Rajpurkar, P., 2021. Medaug: Contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation, in: Machine Learning for Healthcare Conference, PMLR. pp. 755–769.

Wang, X., Zhang, R., Shen, C., Kong, T., Li, L., 2021. Dense contrastive learning for self-supervised visual pretraining, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3024–3033.

Willemink, M.J., Koszek, W.A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L.R., Summers, R.M.,

16

Rubin, D.L., Lungren, M.P., 2020. Preparing medical imaging data for machine learning. Radiology 295, 4.

Wong, W.L., Su, X., Li, X., Cheung, C.M.G., Klein, R., Cheng, C.Y., Wong, T.Y., 2014. 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, e106–e116.

Yellapragada, B., Hornauer, S., Snyder, K., Yu, S., Yiu, G., 2022. Self-supervised feature learning and phenotyping for assessing age-related macular degeneration using retinal fundus images. Ophthalmology Retina 6, 116–129.

Zeng, D., Kheir, J.N., Zeng, P., Shi, Y., 2021a. Contrastive learning with temporal correlated medical images: A case study using lung segmentation in chest x-rays, in: 2021 IEEE/ACM International Conference On Computer Aided Design (ICCAD), IEEE. pp. 1–7.

Zeng, D., Wu, Y., Hu, X., Xu, X., Yuan, H., Huang, M., Zhuang, J., Hu, J., Shi, Y., 2021b. Positional contrastive learning for volumetric medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 221–230.

Zhang, R., Isola, P., Efros, A.A., 2016. Colorful image colorization, in: European conference on computer vision, Springer. pp. 649–666.

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., et al., 2023. A foundation model for generalizable disease detection from retinal images. Nature , 1–8.

Appendix A. Supplementary material

Appendix A.1. Unfrozen finetuning results

To test finetuning performance with unfrozen pretrained weights we use the same evaluation protocol described in Section 3.3.3 but lower the batch size to 384 and weight decay to $1.5 \cdot 10^{-6}$. We find unfreezing can allow inferior pretraining strategies to catch up to stronger models during finetuning, and also improve performance on retinal layer segmentation (see Table A.2 and Figure A.5). However, during full finetuning, while all other models effectively converged, RETFound faced specific challenges in convergence likely due to its use of ViT-Large. Ultimately, for RETFound we found better performance using the linear evaluation protocol. Overall, finetuning most in-domain pretrained models with unfrozen weights performed comparably or slightly worse on image-level tasks than with the linear evaluation protocol.

Appendix A.2. Contrastive pretraining details

In BYOL we choose the update coefficient $\tau = 0.9995$ in accordance with the authors recommendations for smaller batch sizes. We also found that for metadata-enhanced SimCLR setting $Q = 20$ using the configuration described in (Chuang et al., 2020) increased stability during pretraining.

Appendix A.3. Finetuning details

Pretraining speeds up the rate of convergence during finetuning. Accordingly, we found that on some tasks most pretrained models achieve optimal performance very early during finetuning. To address this validation epochs are initially run frequently at every four training steps. We then scheduled this interval to increase to 2 epochs at 100 steps, 3 epochs at 500 steps, and 5 epochs at 1000 steps to improve efficiency by skipping redundant checks later in training.

Table A.2: Downstream performance of unfrozen pretrained models. In most cases models performed comparably to the frozen linear evaluation. Best performance is highlighted in bold, and second best is underlined.

Southampton Eye Unit

Pretraining variant	Early vs. Healthy (AUC) ↑		Late vs. Healthy (AUC) ↑		Visual acuity (MAE LogMAR)↓		Dry vs. Wet (AUC) ↑		Patient Age (MAE years)↓		Patient Sex (AUC) ↑		Segmentation (MAE μm) ↓	
#finetuning labels	100	8299	100	10000	100	10000	100	10000	100	10000	100	10000	100	10000
No pretraining	0.835	0.917	0.682	0.808	16.3	12.4	0.694	0.825	8.17	9.55	0.594	0.701	31.4	17.2
ImageNet SimCLR	0.841	0.902	0.694	0.796	16.2	12.4	0.708	0.833	7.45	6.15	0.612	0.708	30.7	17.9
RadImageNet pretrained	0.879	<u>0.944</u>	0.662	0.804	18.1	12.1	0.709	<u>0.871</u>	7.40	6.10	0.562	0.723	69.3	16.4
RETFound (foundation model)	0.919	0.935	0.748	0.771	13.4	12.6	0.749	0.812	**7.18**	6.78	0.490	0.659	**22.9**	28.9
SimCLR — Standard	0.903	0.941	0.771	**0.849**	13.6	11.7	0.778	0.860	7.35	6.07	0.608	0.742	32.3	16.0
BYOL — Standard	0.903	0.936	0.754	0.835	13.4	**11.5**	0.763	0.869	7.64	6.08	0.626	0.730	<u>29.2</u>	16.3
SimCLR — Ours ($\delta_T = \infty$)	0.905	0.942	0.779	0.836	14.0	11.6	0.791	0.859	7.39	**5.98**	0.634	**0.763**	36.8	<u>15.8</u>
SimCLR — Ours ($\delta_T = 1.0$)	**0.929**	0.938	0.765	0.838	13.8	<u>11.5</u>	0.725	0.849	7.31	<u>5.99</u>	0.633	0.746	36.1	16.1
SimCLR — Ours ($\delta_T = 0.5$)	0.914	0.941	0.772	0.827	14.0	11.6	0.784	**0.874**	7.45	6.08	<u>0.638</u>	<u>0.761</u>	36.0	16.1
BYOL — Ours ($\delta_T = \infty$)	0.915	0.937	**0.783**	<u>0.845</u>	13.2	11.8	**0.821**	0.864	7.25	6.16	0.637	0.748	34.5	16.0
BYOL — Ours ($\delta_T = 1.0$)	<u>0.924</u>	0.937	<u>0.779</u>	0.835	<u>13.1</u>	11.6	<u>0.817</u>	0.841	7.24	6.16	**0.639**	0.733	34.1	15.9
BYOL — Ours ($\delta_T = 0.5$)	0.920	**0.948**	0.765	0.832	**13.0**	11.7	0.802	0.853	<u>7.19</u>	6.11	0.611	0.731	35.6	**15.8**

Moorfields Eye Hospital

Pretraining variant	Early vs. Healthy (AUC) ↑		Late vs. Healthy (AUC) ↑		Visual acuity (MAE LogMAR)↓		Dry vs. Wet (AUC) ↑		Patient Age (MAE years)↓		Patient Sex (AUC) ↑		Segmentation (MAE μm) ↓	
No pretraining	—	—	0.639	0.809	16.2	12.8	—	—	8.09	6.90	0.511	0.684	41.5	23.5
ImageNet SimCLR	—	—	0.592	0.815	15.9	12.7	—	—	8.16	6.90	0.494	0.682	41.2	25.2
RadImageNet pretrained	—	—	0.607	0.819	15.3	12.3	—	—	7.46	6.72	0.514	0.672	63.1	22.1
RETFound (foundation model)	—	—	0.632	0.706	14.3	13.0	—	—	7.62	7.13	0.531	0.635	**36.2**	39.2
SimCLR — Standard	—	—	0.725	0.835	13.7	11.7	—	—	7.24	6.68	0.588	0.744	42.3	21.2
BYOL — Standard	—	—	0.710	0.828	**13.5**	11.8	—	—	7.24	**6.40**	0.573	0.749	42.1	21.5
SimCLR — Ours ($\delta_T = \infty$)	—	—	0.711	0.849	14.0	<u>11.7</u>	—	—	7.08	6.57	**0.627**	<u>0.771</u>	53.4	21.1
SimCLR — Ours ($\delta_T = 1.0$)	—	—	0.722	0.854	14.0	11.9	—	—	**7.05**	6.55	<u>0.616</u>	0.770	50.6	<u>21.0</u>
SimCLR — Ours ($\delta_T = 0.5$)	—	—	<u>0.739</u>	**0.858**	<u>13.7</u>	**11.6**	—	—	7.18	6.55	0.605	0.755	48.7	21.1
BYOL — Ours ($\delta_T = \infty$)	—	—	0.719	0.849	13.9	11.9	—	—	<u>7.06</u>	<u>6.50</u>	0.582	0.745	47.7	21.2
BYOL — Ours ($\delta_T = 1.0$)	—	—	0.734	0.851	13.9	11.9	—	—	7.16	6.52	0.592	**0.771**	47.4	**20.9**
BYOL — Ours ($\delta_T = 0.5$)	—	—	**0.741**	<u>0.856</u>	13.8	11.9	—	—	7.14	6.55	0.594	0.756	46.2	21.1

Appendix A.4. Sampling and training on labelled subsets

Labelled subsets for finetuning have exponentially increasing sizes. To choose these sizes we round down the number of labelled samples to the nearest 10 labels if the amount is less than 100, and we otherwise round down to the nearest 100. The largest amount is not changed so as to estimate performance with 100% of the available labels.

For classification tasks we stratify our labelled subsets by uniformly sampling a proportionate amount of scans from each label class. This prevents the labelled subset from featuring only one class or exhibiting a class balance unrepresentative of the original population. Finally to enable fair comparison between different models we ensure that each is finetuned on the same set of randomly drawn labelled subsets.

Appendix A.5. Robustness tests

Our metadata pairings defined in Equation 1 permit only positive relationships between images from the same longitude. This introduces a theoretical degenerate solution in which images from the same eye are mapped to the same latent feature point. In analogy to BYOL (Grill et al., 2020) we find that this degenerate solution reliably does not occur (see Figure A.6). Moreover, as expected relationships between embeddings of images of the same eye over time respect clinical expectations. Namely, embeddings of images that were taken further apart in time, over which more progression can occur, are typically more distant. While this trend holds in most cases, it highlights two exceptions. Firstly, some images taken over a small window are in fact distant in feature space. This reflects that AMD can progress very rapidly in a short space of time, such as conversion to the late stage. Conversely, some patients progress very slowly and travel a relatively short distance in feature space even over a long time period.
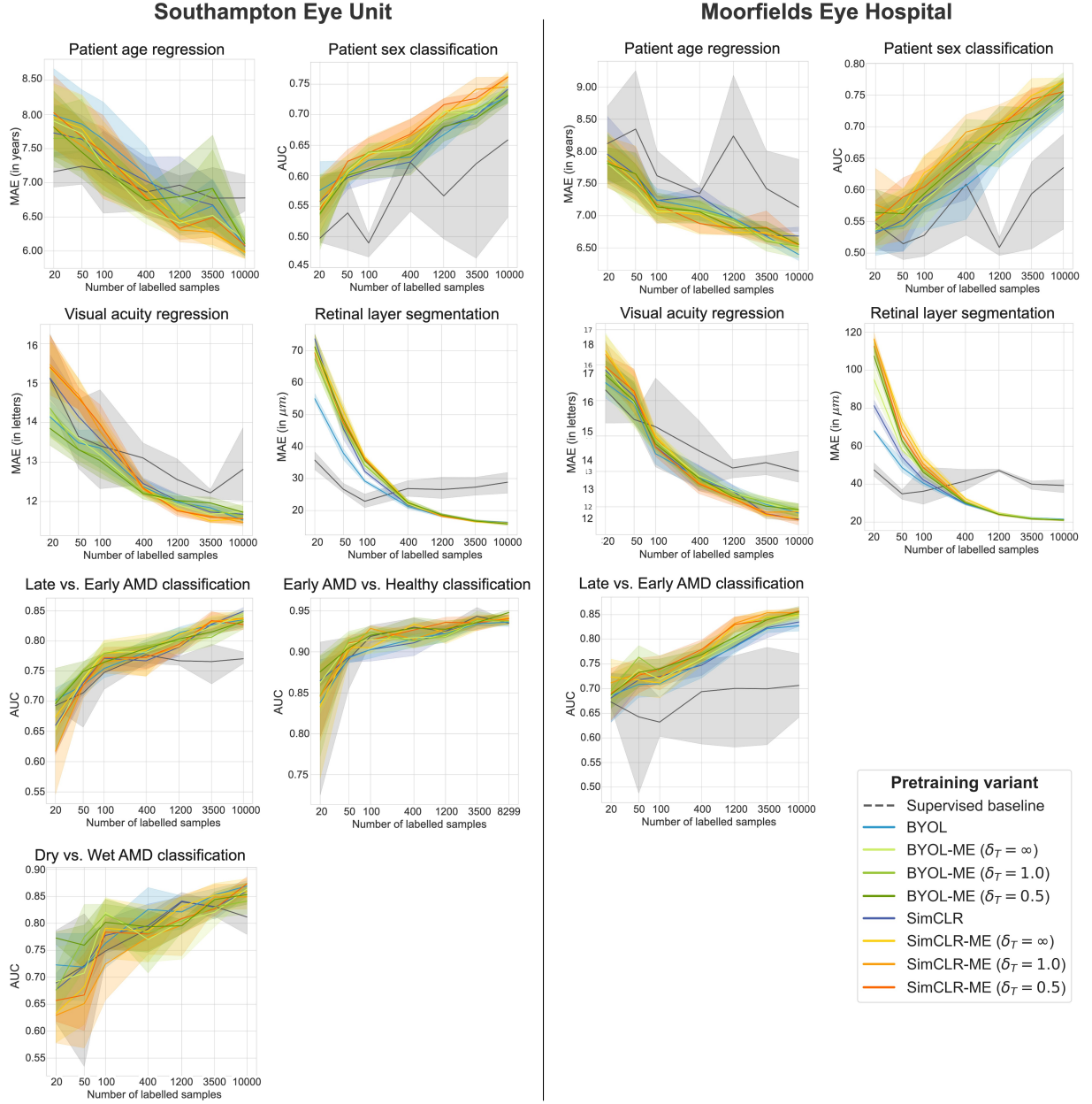
Figure A.5: Performance of finetuning models with fully unfrozen weights on downstream tasks on both datasets. Depicted is the performance (with 95% CIs) against varying sizes of the labelled subsets used for finetuning.

In a second robustness test, metadata-enhanced learning did not improve performance on predicting the time of acquisition, a task unrelated to disease state (see Table A.3).

Standard BYOL   **BYOL-ME ($\delta_T = 1.0$)**

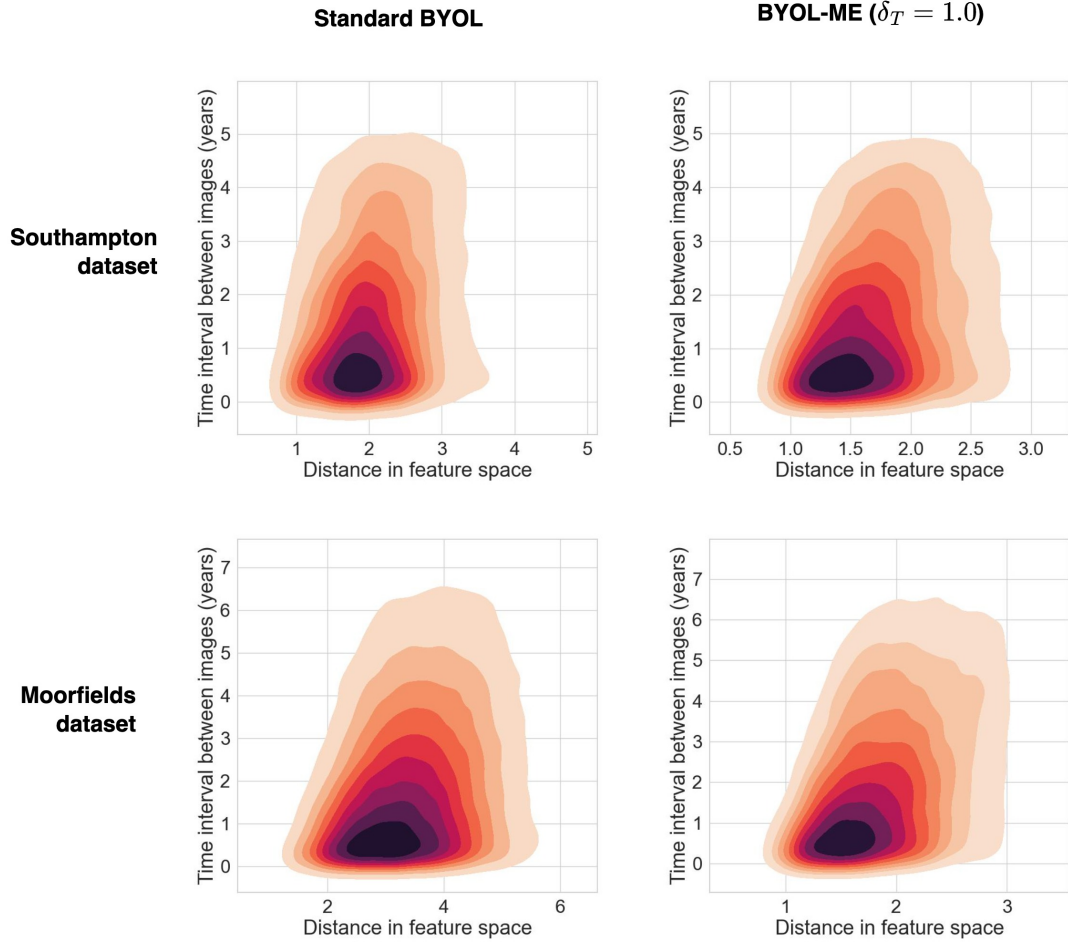Southampton dataset

Moorfields dataset

Figure A.6: Distributions of times between longitudinal images of the same eye against distance measured in the feature space on unseen test data. Our metadata-enhanced learning explicitly encourages that images of the same eye should, on a small time interval, show similar disease-related features and be separated by a smaller distance in the feature space. Emergence of this relationship can already be seen for standard contrastive learning.
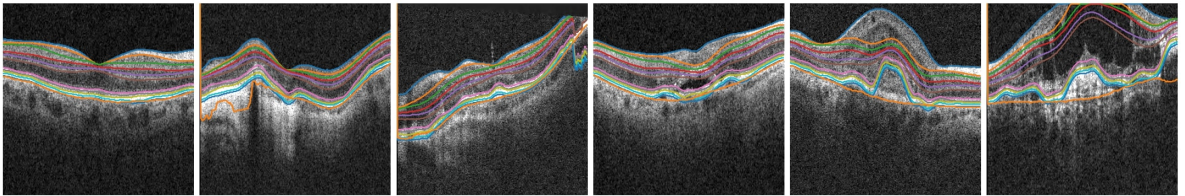

Figure A.7: Performance of the IOWA tool on example OCT images from the Southampton dataset with no pathology, hypertransmission, subretinal fluid and intraretinal fluid. In most cases it is robust to large deformations and presence of fluid.
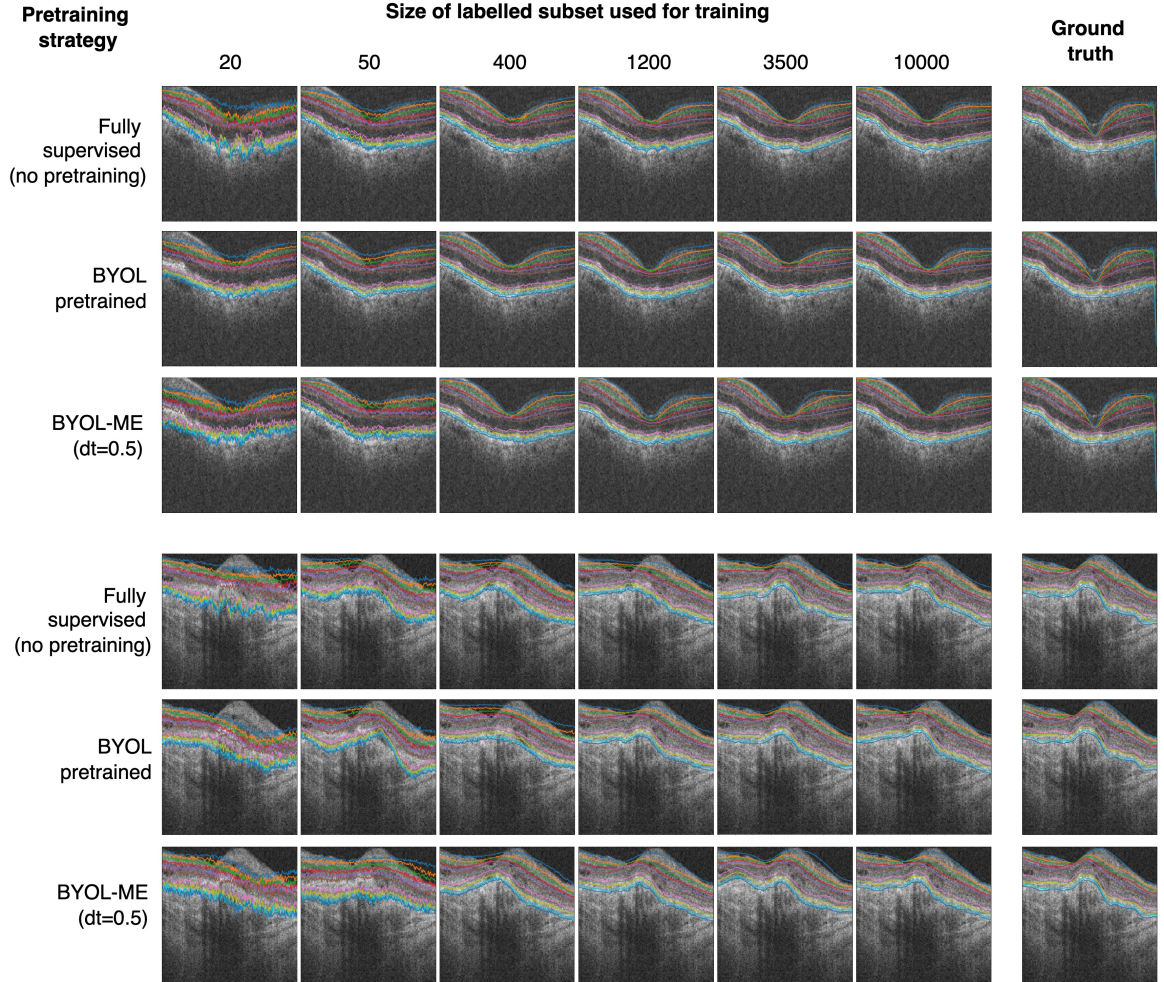
Figure A.8: Performance on downstream segmentation of retinal layers using no pretraining (fully supervised), pretraining with BYOL and with BYOL-ME($\delta_T = 0.5$) finetuned using varying sizes of subsets of the labelled data.

Table A.3: MAE in hours on test set of models on predicting the time of day of image acquisition. Models were finetuned on 10,000 images from the Southampton dataset.

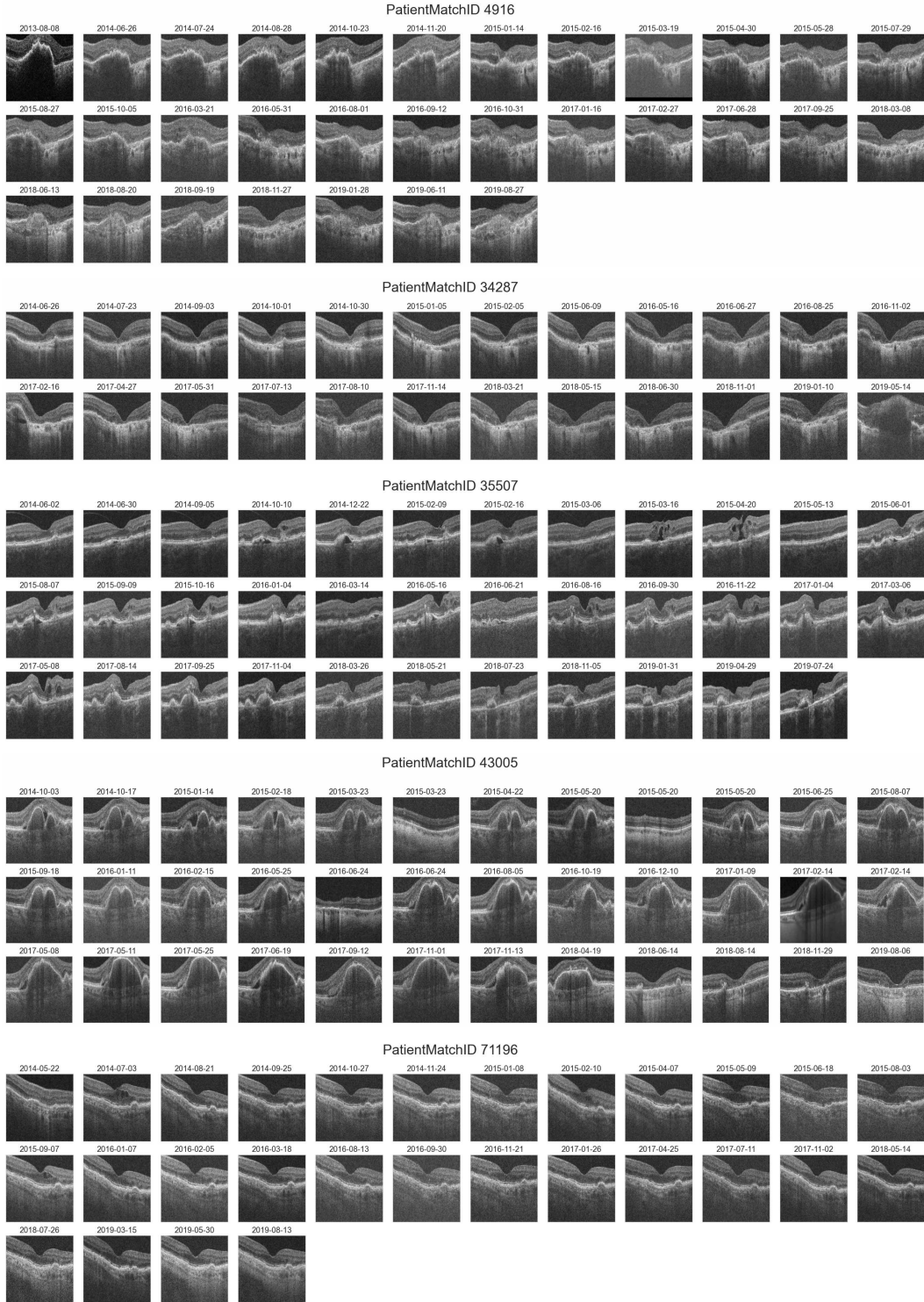| Model	SimCLR	BYOL	SimCLR-ME (δ_T=1.0)	BYOL-ME (δ_T=1.0)
MAE	2.65	2.65	2.65	2.66

21

Figure A.9: Longitudinal scanning series examples from the Southampton dataset. Most feature growing hypertransmission (seen as signal below the retina) which is evidence for growing atrophy of the photoreceptors.