Morph-SSL: Self-Supervision With Longitudinal Morphing for Forecasting AMD Progression From OCT Volumes

Arunava Chakravarty, Taha Emre, Oliver Leingang, Sophie Riedl, Julia Mai, Hendrik P. N. Scholl, Sobha Sivaprasad, Daniel Rueckert, *Fellow, IEEE*, Andrew Lotery, Ursula Schmidt-Erfurth, and Hrvoje Bogunović, for the PINNACLE Consortium

***Abstract*— The lack of reliable biomarkers makes predicting the conversion from intermediate to neovascular age-related macular degeneration (iAMD, nAMD) a challenging task. We develop a Deep Learning (DL) model to predict the future risk of conversion of an eye from iAMD to nAMD from its current OCT scan. Although eye clinics generate vast amounts of longitudinal OCT scans to monitor AMD progression, only a small subset can be manually labeled for supervised DL. To address this issue, we propose Morph-SSL, a novel Self-supervised Learning (SSL) method for longitudinal data. It uses pairs of unlabelled OCT scans from different visits and involves morphing the scan from the previous visit to the next. The Decoder predicts the transformation for morphing and ensures a smooth feature manifold that can generate intermediate scans between visits through linear interpolation. Next, the Morph-SSL trained features are input to a Classifier which is trained in a supervised manner to model the cumulative probability distribution of the time to conversion with a sigmoidal function. Morph-SSL was trained on unlabelled scans of 399 eyes (3570 visits). The Classifier was evaluated with a five-fold cross-validation on 2418 scans from 343 eyes with clinical labels of the conversion date. The Morph-SSL features achieved an AUC of 0.779 in predicting the conversion to nAMD within the next 6 months, outperforming the same network when trained end-to-end from scratch or pre-trained with popular SSL methods. Automated prediction of the future risk of nAMD onset can enable timely treatment and individualized AMD management.**

***Index Terms*— Self-supervised learning, disease progression, age-related macular degeneration, retina, longitudinal OCT.**

Manuscript received 11 January 2024; revised 25 March 2024; accepted 10 April 2024. Date of publication 18 April 2024; date of current version 3 September 2024. This work was supported in part by the Wellcome Trust Collaborative Award (PINNACLE) under Grant 210572/Z/18/Z and in part by Austrian Science Fund (FWF) under Grant 10.55776/FG9. *(Corresponding author: Arunava Chakravarty.)*

Arunava Chakravarty, Taha Emre, Oliver Leingang, Sophie Riedl, Julia Mai, and Ursula Schmidt-Erfurth are with the Department of Ophthalmology and Optometry, Medical University of Vienna, 1090 Vienna, Austria (e-mail: arunava.chakravarty@meduniwien.ac.at; taha.emre@meduniwien.ac.at; oliver.leingang@meduniwien.ac.at; sophie.riedl@meduniwien.ac.at; julia.mai@meduniwien.ac.at; ursula.schmidt-erfurth@meduniwien.ac.at).

Hendrik P. N. Scholl is with the Institute of Molecular and Clinical Ophthalmology Basel, 4031 Basel, Switzerland, and also with the Department of Ophthalmology, University of Basel, 4001 Basel, Switzerland (e-mail: Hendrik.Scholl@usb.ch).

Sobha Sivaprasad is with the NIHR Moorfields Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust, EC1V 2PD London, U.K. (e-mail: sobha.sivaprasad@nhs.net).

Daniel Rueckert is with BioMedIA, Imperial College London, SW7 2AZ London, U.K., and also with the Institute for AI and Informatics in Medicine, Klinikum rechts der Isar, Technical University of Munich, 80333 Munich, Germany (e-mail: daniel.rueckert@tum.de).

Andrew Lotery is with the Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, SO17 1BJ Southampton, U.K. (e-mail: a.j.lotery@soton.ac.uk).

Hrvoje Bogunović is with the Department of Ophthalmology and Optometry and the Christian Doppler Laboratory for Artificial Intelligence in Retina, Medical University of Vienna, 1090 Vienna, Austria (e-mail: hrvoje.bogunovic@meduniwien.ac.at).

Digital Object Identifier 10.1109/TMI.2024.3390940

I. Introduction

AGE-RELATED macular degeneration (AMD) is a leading cause of blindness in the elderly population [1]. Although asymptomatic in its early and intermediate stages, it gradually progresses to a late stage leading to irreversible vision loss. Early or intermediate AMD (iAMD) is primarily characterized by the presence of drusen. Additionally, the Retinal Pigment Epithelium (RPE) and Photoreceptor (PR) layers degenerate over time and are associated with Hyper-reflective Foci (HRF). The late stage is characterized by significant vision loss either due to the presence of Geographic Atrophy (GA) called dry AMD, the presence of choroidal neovascularisation (CNV) called neovascular AMD (nAMD), or a combination of both. nAMD is caused by the abnormal growth of blood vessels that leak fluid into the retina [2] which can be effectively treated with intravitreal anti-VEGF injections. If patients at a higher risk of conversion to nAMD can be identified in the iAMD stage itself, then potential future vision loss could be avoided through frequent monitoring and early treatment. However, the rate of progression varies widely across patients. There are no reliable biomarkers in the iAMD stage to differentiate between slow and fast progressors making it difficult for clinicians to determine the precise risk and timing of conversion. Thus, deep learning (DL)-based methods to predict the future risk of conversion to nAMD can play a critical role in enabling patient-specific disease management.

Optical Coherence Tomography (OCT) provides a 3D view of the retinal tissue and comprises a series of cross-sectional 2D image slices called B-scans. In clinical practice, a longitudinal series of OCT scans is routinely acquired over

multiple patient visits to assess and monitor AMD progression. It generates a large amount of retrospective imaging data that can potentially be used to train DL models. However, due to the time, effort, and clinical expertise required, manual Ground Truth (GT) labels are rarely available for supervised training. Self-Supervised Learning (SSL) offers a way to address this issue by training DL networks to solve *pretext* tasks on unlabelled training data to learn useful feature representations.

In this work, we propose a novel SSL method specifically adapted to longitudinal datasets called Morph-SSL. It involves morphing an OCT scan from one visit to a future visit scan of the same eye. We surmise that the change between the features extracted from two visits should reflect the structural deformation and the intensity changes between them. Morph-SSL is employed to develop a prognostic model to predict future conversion from iAMD to nAMD within the next t months from a single current OCT scan. t can be any continuous time-point up to a maximum of 18 months. We refer to this task as TTC, predicting the probability distribution of the *Time-to-Conversion*. Once an Encoder has been trained with Morph-SSL, a 3-layer Classifier is trained for the TTC task on limited labelled data. The Encoder and Classifier can further be fine-tuned jointly. Our key contributions are:

(i) We propose Morph-SSL, a novel SSL method to learn representations that capture temporal changes in the retinal tissue from unlabelled longitudinal datasets. With minimal constraints on the unlabelled training data, Morph-SSL requires at least two visits per eye and can also use scans acquired at irregular intervals. The learned feature manifold is enforced to be smooth with meaningful notions of distance and direction, such that linear interpolation between two scans in the feature space leads to a gradual transition between them in the image space.

(ii) We model the Cumulative Distribution Function (CDF) of the probability of the TTC with a sigmoidal function over time. It allows using continuous GT labels of conversion time during training, ensures the monotonic non-decreasing property of the CDF, and can predict the conversion risk for arbitrary continuous time-points at test time.

(iii) We propose a score $r \in [0, 1]$ that quantifies the future risk of eyes to develope nAMD and can categorize them into low and high risk groups for conversion. r can play a crucial role in personalized treatment by identifying the high risk patients for early treatment and more frequent monitoring.

(iv) We develop an efficient CNN network to process entire OCT volumes instead of individual 2D B-scans. We explore (a) *S3DConv* block to replace 3D convolutions with three groups of 2D convolutions oriented in the three orthogonal planes; (b) concatenation-based (instead of additive) skip connections to have the same output channel size with fewer convolutions; (c) Layer Normalization instead of Batch Normalization to allow training with a batch size of 1.

II. Related Work

A. Self-Supervised Learning

It offers a way to overcome the paucity of labelled datasets for supervised training. SSL learns feature representations from unlabelled data by training the network on a pretext task that does not need manual labels. SSL-trained models can either be utilized for off-the-shelf feature extraction or to provide initial weights for fine-tuning on the desired *downstream* task with limited labelled training data. Recent SSL methods employ pretext tasks based on image reconstruction or Contrastive Learning (CL). Reconstruction-based methods train networks to predict the original image from its distorted version and have been applied to X-ray, CT, MRI and ultrasound images [3], [4]. The distortions involve transformations such as non-linear intensity mapping, local shuffling, and in-painting in Model-Genesis [3] and randomly swapping patches in the image [4].

CL has been applied to chest X-ray, dermatology [5], histology [6], MRI [7] and ultrasound [8] images. CL trains networks using random batches comprising two data-augmented versions per image, called *positive pairs*. While positive pairs are pulled closer, the features of different images in the batch called *negative pairs* are pushed apart. However, the images in a negative pair can still be semantically similar (same pathology or disease stage), resulting in many *False Negative pairs*. Their impact can be reduced by training with large batch sizes (1024 for chest X-rays, 512 for dermatology images in [5] and 128 for histology image patches in [6]). Since large batch sizes do not scale well to 3D images due to limited GPU memory, existing methods learn features at a 2D, slice-level for 3D MRI volumes [7], or for individual frames in ultrasound videos [8] where neighboring slices/frames of the same 3D image are excluded from negative pairs. The recently proposed *Non-Contrastive* methods overcome the problem of *False Negative pairs*. They do not maximize the negative pair separation but only ensure that they do not collapse onto the same feature representation. VICReg [9] keeps the standard deviation of each feature dimension over a batch above a threshold. Barlow Twins [10] forces the cross-correlation between two batch of features extracted from the two images in each positive pair to be close to the identity matrix. BYOL [11] prevents feature collapse using slightly different network weights to extract features for the two views in the positive pair, where the second network weight is computed as the moving average of past weights.

CL and Non-Contrastive SSL have been adapted for retinal OCT to learn features for 2D B-scans with training batch sizes of 128 in [12] and 384 in [13]. Another method learns features for central B-scans by predicting the time interval between two input scans from random visits of the same patient [14]. In contrast, *Morph-SSL with a novel image morphing-based pretext task can be trained with a batch size of 1 to reduce GPU memory usage, allowing us to learn feature representations for entire 3D OCT volumes instead of 2D B-scans.*

B. Time to Conversion Prediction

Existing methods either employ Color Fundus Photographs (CFP) or OCT imaging for TTC prediction. CFP is a 2D image of the retinal surface and lacks a cross-sectional view of the retina. A 9-grade AREDS disease severity scale [15] further stratifies the iAMD stage in CFPs, where each

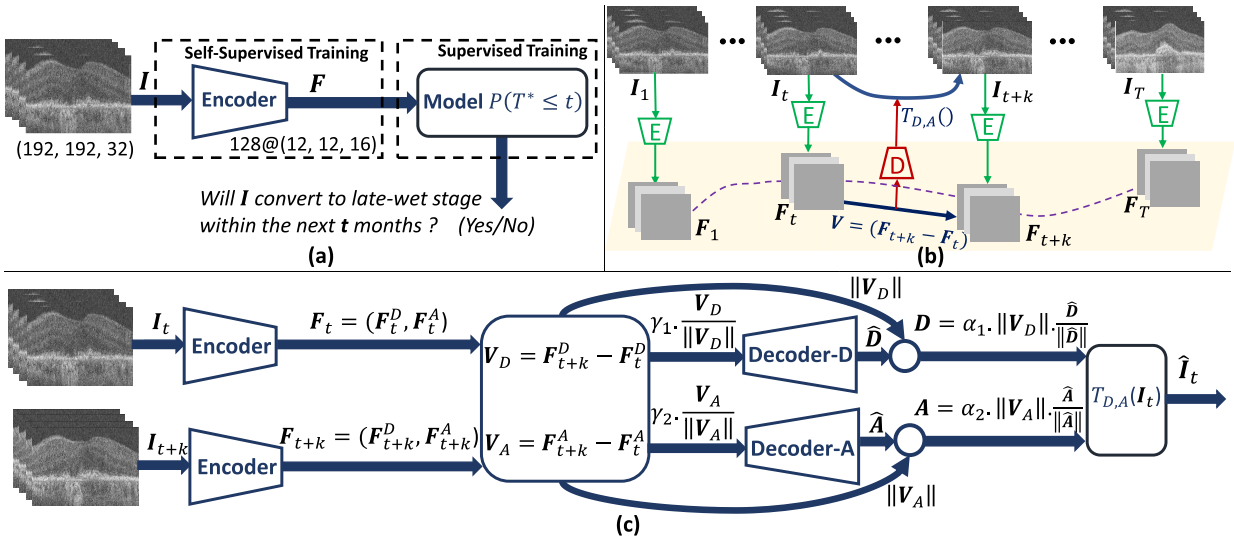

Fig. 1. (a) Overview of our 2-stage training framework: Stage-1 involves self-supervised pre-training of an Encoder using Morph-SSL. Stage-2 involves supervised training of a classifier for the downstream task of predicting future nAMD conversion within the next *t* months. Optionally, an end-to-end fine-tuning of the pre-trained Encoder and the classifier can be performed. The motivation behind Morph-SSL is shown in (b) and its details in (c). Please refer to Fig. 2(c) for the Encoder and Fig. 3(b) for the classifier architecture.

successive stage has been linked to an increased 5-year risk of conversion to advanced AMD (from 1% in grade 1 to about 50% risk in grade 9). However, no such severity scale exists for the relatively new OCT imaging. The establishment of imaging biomarkers in the iAMD stage preceding the onset of nAMD remains a topic of ongoing investigation, as evidenced by many recent [2], [16], [17], [18] and ongoing clinical studies [19]. In [16], morphological changes such as hyperreflective foci (HRF) and pigment epithelial detachment (PED) were observed one month before the manifestation of choroidal neovascularization (CNV). Additionally, an increase in drusen area or thickness [2], [17], HRF [17] and the presence of a thick Double-Layer Sign (characterized by the visible separation of the retinal pigment epithelium and Bruch's Membrane with different reflectivity) [18] have emerged as potential biomarkers indicative of an elevated risk of converting to nAMD within a two-year time frame. Moreover, an accelerated thinning and appearance changes in the choroid was also observed in [20] prior to nAMD onset within a one year follow-up period.

Some *CFP-based methods* predict the AREDS severity scale [21], [22]. Two-year conversion of nAMD was predicted with an ensemble of such predictions combined with features from drusen segmentation and demographic data [22]. The CNN-LSTM based methods in [23] and [24] require images from multiple past visits, hence cannot be used for patients visiting for the first time. The input CFPs from visits at irregular intervals are handled by scaling the input image features with visit time intervals [23] or using a time-aware LSTM network [24]. In [25], a Generative Adversarial Network was used to generate synthetic CFP images for future time-points. Combining CFP with genetic features can improve performance [26], but such information is not readily available in eye clinics. While CFP-based methods can predict long-term conversion, they are not sensitive to short-term conversion risks within 2 years, required for effective clinical intervention. Because CFPs lack a 3D view of the retina, they cannot capture subtle changes in retinal layers or extract accurate lesion volumes.

Many *OCT-based methods* first extract a set of handcrafted quantitative biomarkers to capture the distribution, appearance and volume of lesions like drusen, HRF and retinal layers such as RPE and PR. These biomarkers combined with other demographic [27] or genetic data [28] are input to an LSTM [27], Cox proportional hazards model [28], or an L1-penalized Poisson model [29] to predict the TTC. The biomarkers are extracted with automated segmentation methods that are often inaccurate and require voxel-level labels to train. Moreover, handcrafted biomarkers may not adequately capture the subtle retinal changes related to disease progression.

Another approach directly uses the OCT scans as input. To reduce the compute and GPU memory, most existing methods operate on individual B-scans with 2D CNNs. Some methods in [14] and [12] only use the central B-scan that passes through the macula, ignoring the remaining B-scans in the volume. While [14] employed this strategy for TTC prediction, [12] used it on other tasks such as predicting age, sex and visual acuity from OCT. Alternatively, 2D CNNs can be applied independently to every B-scan in the volume. During inference, the predictions from each B-scan is pooled, either by taking the average [30] or maximum [13] to obtain the volume-level prediction. During training, the same GT for the conversion time is used for every B-scan in the volume, even if only a few of them have the biomarkers indicative of progression risk, resulting in noisy training labels. *In contrast to these methods, in this work we explore a full 3D approach by developing a compact 3D-CNN network to effectively capture the spatial information across the individual B-scans*. Other than our work, the only other 3D-CNN method is found in [31], which employs two prediction networks to predict the conversion risk within six months, one using raw

OCT volumes and the other using retinal layer and lesion segmentation maps.

III. Method

Our primary contribution is Morph-SSL, a new Self-Supervised Learning method designed to leverage the wider availability of unlabeled longitudinal training data. It learns feature representations that are sensitive to the morphological characteristics in an input OCT scan which are indicative of future disease progression. Morph-SSL uses pairs of unlabeled scans acquired at irregular (but known) time-intervals from each subject to solve the *pretext* task of morphing the scan from the prior visit to the next as detailed in III-A. Considering implementation challenges such as GPU memory and the computation time required in a 3D convolutional network, we also developed a lightweight CNN architecture (in Fig. 2) for processing 3D OCT volumes.

In order to demonstrate the effectiveness of our learned representations, we chose a clinically relevant downstream task of predicting the future risk of conversion of eyes (currently in the iAMD stage) to the late nAMD stage in Section III-B. The Cumultative Distribution Function (CDF) of the time-to-conversion (TTC) is modeled as a sigmoid function over time. The CDF parameters are predicted using a classifier that employs the Morph-SSL trained encoder.

The overall pipeline combining the *unsupervised* representation learning and the *supervised* downstream conversion prediction stage is depicted in Fig. 1(a). First, a Fully Convolutional Encoder is trained with Morph-SSL to project an input OCT scan $\mathbf{I}$ to a convolutional feature map $\mathbf{F}$. This stage only utilizes unlabeled longitudinal image-pairs for training. Next, $\mathbf{F}$ is input to a classifier for the downstream task of predicting the future conversion to nAMD within the next t months. This task involves modeling the CDF of conversion time using a sigmoid distribution whose parameters are predicted by the classifier. By freezing the Morph-SSL pre-trained weights for the encoder, the classifier is trained in a supervised manner on limited labeled data. Optionally, an additional end-to-end finetuning can also be performed using the Morph-SSL pre-trained weights for network initialization.

A. Self-Supervised Learning

1) Motivation: Let $\{\mathbf{I}_t | 1 \leq t \leq T\}$ represent a set of 3D OCT scans of an eye acquired over T visits, in which $\mathbf{I}_t$ is obtained on the t^{th} visit. The Encoder projects each $\mathbf{I_t}$ to a feature map $\mathbf{F}_t$ of size 128@12×12×16. $\mathbf{F}_t$ can be interpreted as 128-dimensional features for overlapping 3D image patches in $\mathbf{I}_t$, with the patch size defined by the effective receptive field of the Encoder. As AMD progresses over successive visits, $\mathbf{F}_t$ traces a trajectory (denoted by the violet dotted line in Fig. 1(b)) that is locally linear between nearby visits $\mathbf{I}_t$, and $\mathbf{I}_{t+k}$ (assuming a smooth feature manifold) but maybe non-linear over the entire AMD progression. Let $\mathcal{T}_{D,A}(.)$ denote a transformation which morphs $\mathbf{I}_t$ to look similar to $\mathbf{I}_{t+k}$, with parameters $\mathbf{D}$ and $\mathbf{A}$. As $\mathbf{I}_t$ morphs into $\mathbf{I}_{t+k}$ in the image space, $\mathbf{F}_t$ should be linearly displaced to $\mathbf{F}_{t+k}$ by $\mathbf{V}_t = \mathbf{F}_{t+k} - \mathbf{F}_t$ in the feature manifold. This observation motivates our pretext task for Morph-SSL which employs an Encoder-Decoder architecture. The Encoder projects scans from two nearby visits, $\mathbf{I}_t$ and $\mathbf{I}_{t+k}$ to their features $\mathbf{F}_t$ and $\mathbf{F}_{t+k}$. The Decoder uses the displacement $\mathbf{V}_t$ as input to predict $\mathbf{D}$ and $\mathbf{A}$ of the morphing transformation $\mathcal{T}_{D,A}$. Our pretext task ensures that the displacements in the learned feature manifold capture the corresponding appearance changes in the image space.

$\mathcal{T}_{D,A}$ comprises a spatial deformation with the 3-channel $\mathbf{D}$ and an additive intensity transformation with the 1-channel $\mathbf{A}$, both of the same spatial size as $\mathbf{I}_t$. Each voxel at location $\boldsymbol{p}$ in $\mathbf{I}_t$ is displaced to the location $\boldsymbol{p} + \mathbf{D}(\boldsymbol{p})$, where $\mathbf{D}(\boldsymbol{p})$ is a 3-dimensional displacement vector representing the translations along the height, width and depth direction. Additionally, $\mathbf{A}(\boldsymbol{p})$ captures the intensity changes at each location $\boldsymbol{p}$, caused by newly formed pathologies in $\mathbf{I}_{t+k}$ such as fluids or drusen. Thus, the transformed image is $\widehat{\mathbf{I}}_t = \mathcal{T}_{D,A}(\mathbf{I}_t) = \Phi(\mathbf{I}_t; \mathbf{D}) + \mathbf{A}$, where Φ is the spatial deformation applied in a differentiable manner similar to the registration methods in [32] and [33] based on the Spatial Transformer Networks [34]. $\mathbf{A}$ has a single color channel (similar to $\mathbf{I}_t$) to model the additive intensity transformation since OCTs are grayscale images.

2) Morph-SSL Framework: The details are depicted in Fig. 1(c). $\mathbf{F_t}$ is split into two subspaces $\mathbf{F}_t^D$ and $\mathbf{F}_t^A$ of 64 channels each. The Decoder has two sub-networks, *Decoder-D* and *Decoder-A* that operate on $\mathbf{F}_t^D$ and $\mathbf{F}_t^A$ feature maps respectively, to predict $\mathbf{D}$ and $\mathbf{A}$. A notion of semantically meaningful directions and distance is incorporated. The amount of deformation between $\mathbf{I}_t$ and $\mathbf{I}_{t+k}$ should be proportional to the Euclidean distance $||\boldsymbol{V}_D|| = ||\boldsymbol{F}_{t+k}^D - \boldsymbol{F}_t^D||_2$, while the nature and location of the deformation should be captured by the direction alone, represented by the unit vector $\boldsymbol{V}_D/||\boldsymbol{V}_D||$. This property is enforced by our Decoder architecture in Fig. 1(c). Only the direction information $\gamma_1.(\boldsymbol{V}_D/||\boldsymbol{V}_D||)$ is input to *Decoder-D* and its output $\widehat{\mathbf{D}}$ is normalized and scaled to obtain $\mathbf{D} = \alpha_1.||\boldsymbol{V}_D||.(\widehat{\mathbf{D}}/||\widehat{\mathbf{D}}||)$. This ensures that $||\mathbf{D}|| = \alpha_1.||\mathbf{V}_D||$. Both γ_1, α_1 are learnable parameters (positive scalar weights) employed for numerical stability during training. A similar scheme is employed to predict $\mathbf{A}$. The direction $\gamma_2.(\boldsymbol{V}_A/||\boldsymbol{V}_A||)$ is input to *Decoder-A* and its output scaled to $\mathbf{A} = \alpha_2.||\boldsymbol{V}_A||.(\widehat{\mathbf{A}}/||\widehat{\mathbf{A}}||)$, where γ_2, α_2 are learnable positive weights (see Fig. 1(c)).

a) Loss function: The Encoder-Decoder network is trained to minimize the Mean Squared Error (MSE) between $\widehat{\mathbf{I}}_t$ and $\mathbf{I}_{t+k}$ by directly comparing their voxel intensities ($\mathcal{L}_{mse}$) as well as their feature maps extracted with a CNN ($\mathcal{L}_{prc}$). $\mathcal{L}_{mse}$ alone leads to blurred reconstructions which is remedied by using the additional *perceptual loss* $\mathcal{L}_{prc}$ [35], [36]. The OCT scans have a dark noisy background region both above and below the retinal tissue. We define the region between the Inner Limiting Membrane (ILM) and the Bruchs Membrane (BM) along with a small margin below the BM (to include the choroid) as the region of interest (ROI) containing the retinal tissue. The binary ROI mask $\mathbf{R}_t$ for the scan $\mathbf{I}_t$ is extracted automatically (see Section IV, *Preprocessing* section for details). While registering $\mathbf{I}_t$ to $\mathbf{I}_{t+k}$, we aim to morph the ROI in $\mathbf{I}_t$ to the corresponding retinal tissue region in $\mathbf{I}_{t+k}$. The

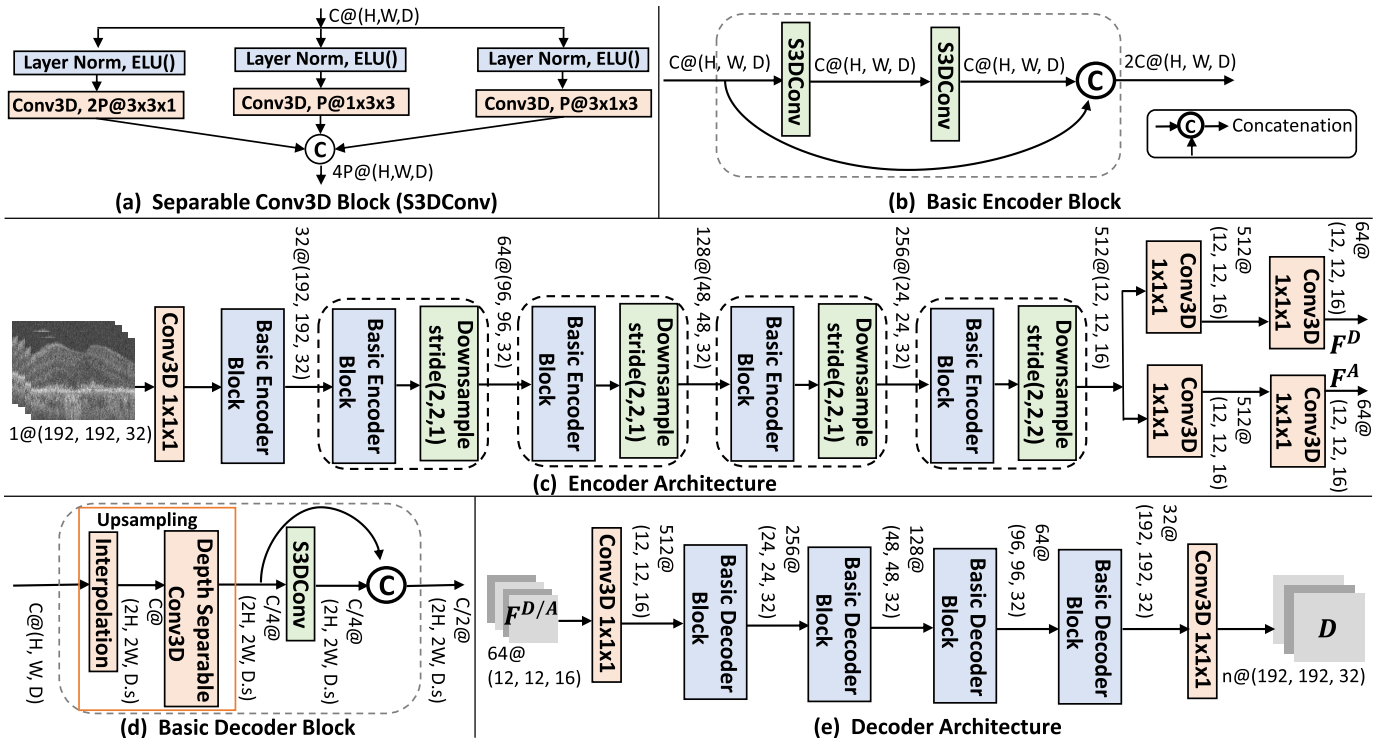

Fig. 2. The S3DConv block in (a) is used as the basic convolution operation in our entire Encoder and Decoder architecture. A series of Basic Encoder Blocks detailed in (b) constitutes our Encoder architecture as shown in (c). The Decoder-D and Decoder-A accept $\mathbf{F}^D$ and $\mathbf{F}^A$ as input respectively and employ a similar architecture comprising a series of Basic Decoder Blocks in (d), except the number of output channels in the last layer ($n = 3$ for Decoder-D and $n = 1$ for Decoder-A) as depicted in (e).

background noisy region is ignored while computing $\mathcal{L}_{mse}$ and $\mathcal{L}_{prc}$ using $\mathbf{R}_t$ (for $\mathbf{I}_t$) and $\mathbf{R}_{t+k}$ (for $\mathbf{I}_{t+k}$), obtained during pre-processing. Before computing the loss, the background regions are masked out through element-wise multiplication $\mathbf{I}_{t+k} = \mathbf{I}_{t+k} \odot \mathbf{R}_{t+k}$ and $\widehat{\mathbf{I}}_t = \widehat{\mathbf{I}}_t \odot \Phi(\mathbf{R}_t; \mathbf{D})$. The Encoder does not require the binary masks at inference time as they are only used to compute the loss.

$\mathcal{L}_{mse}$ has two terms. First, the MSE is computed with the only spatially deformed image $\Phi(\mathbf{I}_t; \mathbf{D})$. Next, $\mathbf{A}$ is fitted to the residual difference left after the spatial deformation, $\mathbf{U} = \mathbf{I}_{t+k} - \Phi(\mathbf{I}_t; \mathbf{D}).detach()$. The $detach()$ indicates that the gradients are not allowed to backpropagate through $\mathbf{U}$ which is computed on the fly and treated as the GT for $\mathbf{A}$. This two-step design is to ensure that $\mathbf{D}$ accounts for most of the reconstruction and avoid trivial solutions where $\mathbf{D}$ is an identity transformation (0 displacement for all voxels) while $\mathbf{A}$ tries to learn the entire difference $\mathbf{I}_{t+k} - \mathbf{I}_t$. Thus,

$$\mathcal{L}_{mse} = \frac{\lambda_1}{|\Omega|}.||\mathbf{I}_{t+k} - \Phi(\mathbf{I}_t; \mathbf{D})||_2^2 + \frac{\lambda_2}{|\Omega|}.||\mathbf{U} - \mathbf{A}||_2^2, \quad (1)$$

where $|\Omega|$ is the total number of voxels in the image and the relative weights $\lambda_1 = 10^1$, $\lambda_2 = 10^2$ were set empirically.

The MSE in the voxel intensity space encourages smoothed reconstructions, blurring the edges and textural content in $\widehat{\mathbf{I}}_t$. A perceptual loss term $\mathcal{L}_{prc}$ addresses this issue by extracting convolutional feature maps for $\mathbf{I}_{t+k}$ and $\widehat{\mathbf{I}}_t$ with a *Comparator* CNN network ψ and computing the MSE in the extracted feature space (rather than the raw voxel intensities) as

$$\mathcal{L}_{prc} = \frac{1}{3}\sum_{j=1}^{3}\frac{1}{|\Omega|}||\psi_j(\mathbf{I}_{t+k}) - \psi_j(\widehat{\mathbf{I}}_t)||_2^2. \quad (2)$$

$\psi_j(\mathbf{I})$ denotes the feature map from the j^{th} layer of ψ for the input $\mathbf{I}$. Typically, the first few layers of a pre-trained network such as VGG-16 are used for ψ [35], [36]. In the absence of a suitable pre-trained 3D CNN network for OCT volumes, we define ψ to have an architecture identical to the first 3 layers of our Encoder. Inspired by BYOL [11], ψ maintains a separate copy of its network weights which is updated with an exponential moving average of the past Encoder weights (for the first 3 layers) as the training proceeds. Although initially ψ is randomly initialized, the quality of its feature maps improves gradually during training. Thus, we eliminate the need for an existing pre-trained network for ψ.

Additional *regularization* loss terms are also incorporated to obtain an anatomically feasible $\mathcal{T}_{D,A}$. $\mathbf{D}$ is encouraged to be diffeomorphic by penalizing it to be smooth with $\mathcal{L}_{smth}$ and prevent folding with $\mathcal{L}_{fld}$. The $\mathcal{L}_{smth} = \sum_{p\in\Omega}||\nabla\mathbf{D}(\boldsymbol{p})||_2^2$ was defined as in [32], where the spatial gradient $\nabla\mathbf{D}(\mathbf{p})$ is computed at all voxel positions through discrete numerical approximation. $\mathcal{L}_{fld}$ as defined in [33], penalizes the anatomically infeasible deformations where the retinal tissue folds onto itself. Finally, the sparsity of $\mathbf{A}$ is ensured with an L1-regularization $\mathcal{L}_{add} = \sum_{p\in\Omega}|\mathbf{A}(\boldsymbol{p})|$. Thus, the total loss is

$$\mathcal{L} = \mathcal{L}_{mse} + \lambda_3.\mathcal{L}_{prc} + \lambda_4.\mathcal{L}_{smth} + \lambda_5.\mathcal{L}_{fld} + \lambda_6\mathcal{L}_{add}, \quad (3)$$

where $\lambda_3 = 10^1$, $\lambda_4 = 10^{-1}$, $\lambda_5 = 10^6$ and $\lambda_6 = 10^{-5}$ are empirically fixed, based on their relative importance and also to scale the different loss terms to a similar range. The range

of the $\mathcal{L}_{fld}$ is orders of magnitude lower than the other terms, thus requiring a significantly larger scaling weight.

3) Network Architecture: The *separable 3D Convolution Block* (S3DConv) depicted in Fig. 2(a) replaces 3D convolutions throughout our Encoder and Decoder Networks to reduce computation and network parameters. It employs 2D convolution filters in the three orthogonal planes. While 50% of the filters are 3×3×1 that operate on individual B-scans, the remaining are an equal number of 1×3×3 and 3×1×3 filters to capture contextual information across the neighboring B-scans. Using Layer Normalization instead of Batch Normalization allows training with a batch size of 1. The *pre-activation* strategy [37] ensures that the normalization and ELU activations are applied after the skip connections, at the beginning of the next S3DConv block for better gradient backpropagation.

The *Encoder* depicted in Fig. 2(c) has a series of five Basic Encoder Blocks interleaved with downsampling. The Basic Encoder Block comprises two S3DConv Blocks followed by a concatenation based skip connection (see Fig. 2(b)). Here, each S3DConv has C input and output channels by setting $P = C/4$ in Fig. 2(a). The downsampling is performed with a strided 3×3×3 depthwise-separable convolution [38]. It applies a separate 3×3×3 convolution (with 1 input and output channel) to each of the C input channels individually and their outputs are concatenated together. It is implemented in Pytorch by setting *groups* = 1 in the Conv3D layer. Due to large voxel spacing across the B-scans, downsampling along this direction is only performed in the final block with a stride of $(2, 2, 2)$ to ensure a roughly isotropic receptive field. All previous downsampling layers use a $(2, 2, 1)$ stride to only halve the height and width dimensions. The last Encoder block is followed by two parallel pathways, each consisting of two 1×1×1, 3D convolutional layers to obtain the final 64 channel $\mathbf{F}^D$ and $\mathbf{F}^A$.

Decoder: Both *Decoder-D* and *Decoder-A* (in Fig. 1(c)) have the same architecture as shown in Fig. 2(e), except for the number of output channels in the last 1×1×1, convolution layer (1 channel for **A** and 3 to predict **D** respectively). The Decoder architecture employs a series of Basic Decoder Blocks. They map a $C@(H, W, D)$ input feature map to a $\frac{C}{2}@(2H, 2W, D.s)$ output, where s is the upsampling factor across B-scans ($s = 2$ in the first block, 1 otherwise). As depicted in Fig. 2(d), it comprises an upsampling layer followed by a S3DConv whose outputs are concatenated with a skip connection. The upsampling layer performs two operations. First, the input of size $C@(H, W, D)$ is upsampled to $C@(2H, 2W, D.s)$ using trilinear interpolation. Next, a Depth-separable 3×3×3 convolution is employed, which divides the C input channels into $\frac{C}{4}$ groups of 4 channels each. A separate convolution filter is applied to each group to compress them to a single channel resulting in a $\frac{C}{4}@(2H, 2W, D.s)$ output.

B. Downstream TTC Estimation Task

The problem setting of the Downstream TTC task for an eye is depicted in Fig. 3(a). An OCT is acquired at each visit (red dots) occuring at irregular time intervals. The eye remains in the early/iAMD stage up to the visit at time T^- and is first diagnosed to have progressed to nAMD at time T^+. The exact time of conversion T^* is unknown as patients are monitored at discrete time-points but lies in $T^- < T^* \leq T^+$. We treat T^* as a continuous random variable and aim to model its CDF, $P(T^* \leq t)$ (y-axis in Fig. 3(a)). $P(T^* \leq t)$ is the probability that the eye has converted within the time-point t. The binary GT for $P(T^* \leq t)$ is 0 for $0 \leq t \leq T^-$, 1 for $t \geq T^+$ and unknown in the range $T^- < t < T^+$. We propose to model $P(T^* \leq t)$ with a sigmoidal distribution over time as

$$p_t = P(T^* \leq t) = 1/\left[1 + exp\left\{-\left(\frac{t-b}{a+0.05}\right)\right\}\right], \quad (4)$$

where b is an estimate of T^* and a controls the slope of the sigmoidal CDF. A steep slope (small a) would indicate a fast progression rate around T^* and viceversa.

1) Classifier Architecture: The scalars a and b are predicted with the classifier in Fig. 3(b). The SSL-trained feature map **F** of the input OCT scan is fed to the classifier. **F** is mapped to a single channel feature map **M** through a series of three 1×1×1 convolutional layers. A Class Activation Map (CAM) can be computed as a weighted sum of all channels in the final convolutional feature map, which in our case is **M** with a single channel. Thus, **M** can be interpreted as a saliency map for our classifier (see Fig. 4) which motivates how a and b are computed.

The b is obtained through the Global Average Pooling (GAP) of **M** denoted by $\widehat{b}$, scaling it by a non-negative learnable scalar weight α_1 and taking the reciprocal $b = 1/(\alpha_1 \cdot \widehat{b})$. We hypothesize that images predicted to convert soon (with small b) should lead to higher activations on the saliency map **M**.

The a is obtained by computing the spatial entropy of **M** denoted by $\widehat{a}$, scaling it by non-negative learnable scalar α_2 and applying the sigmoid activation. We hypothesize that low entropy (certain locations in **M** have high activations while others take very small values) indicates the detection of some salient regions in the OCT which may correlate to a sudden disease progression around the conversion event leading to a steep slope (small a). The spatial entropy is computed by first normalizing M to sum to 1, $\mathbf{M}'(i) = \mathbf{M}(i)/\sum_{p\in\Omega}\mathbf{M}(p)$ and then computing the entropy as $H = -\sum_{i\in\Omega}\mathbf{M}'(i).log\mathbf{M}'(i)$, where Ω represents each spatial position in **M**.

2) Loss Function: A maximum time interval of 18 months (normalized to [0,1]) was considered as longer durations are not useful for clinical intervention. The T^* for scans that do not convert within 18 months is unknown. For each scan, the classification loss $\mathcal{L}_{cls}$ consists of the average Binary Cross Entropy loss (BCE) computed for two time-points as

$$\mathcal{L}_{cls} = \begin{cases} \mathcal{L}_{ce}\left(p_{T^+}, 1\right) + \mathcal{L}_{ce}\left(p_{T^-}, 0\right), \text{ if } 0 \leq T^+, T^- \leq 1 \\ \mathcal{L}_{ce}\left(p_0, 0\right) + \mathcal{L}_{ce}\left(p_1, 0\right), \text{ if } T^+, T^- > 1 \\ \mathcal{L}_{ce}\left(p_0, 1\right) + \mathcal{L}_{ce}\left(p_1, 1\right), \text{ if } T^+ = 0, \end{cases} \quad (5)$$

where p_t at time t is computed using Eq. 4. $\mathcal{L}_{ce}$ denotes half of BCE loss to compute the average BCE over the two

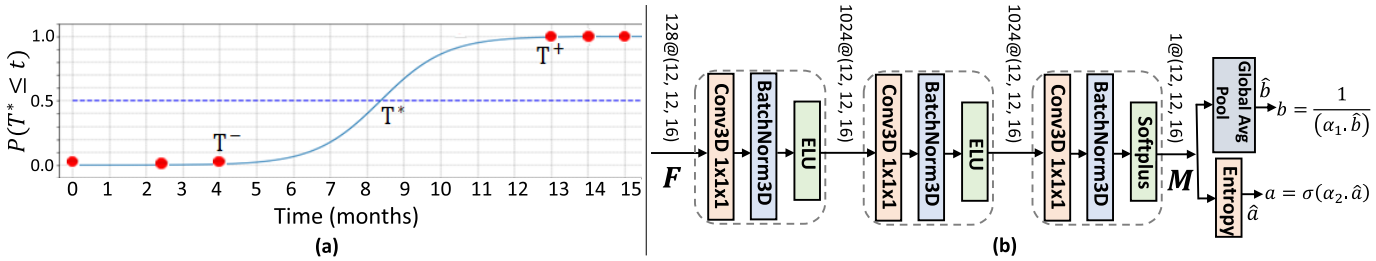

Fig. 3. Overview of the TTC Task. (a) CDF of the conversion time T^* can be best modeled using a sigmoidal function. Exact T^* is unknown due to the discrete nature of the visits (red dots) but occurs between the first visit where the eye has converted (T^+) and the visit (T^-) just before it. (b) The Classifier Network to predict the parameters of the sigmoidal function.

time-points. The first condition in Eq. 5 occurs when both T^- and T^+ occur within 18 months (1 after normalization) and $\mathcal{L}_{ce}$ is computed at these two time-points with the GT labels 0 at T^- and 1 at T^+. Since the sigmoidal function is monotonically non-decreasing, minimizing the loss at these two points automatically improves p_t for all t because $p_{T^-} \approx 0$ also ensures equal or lower predictions before T^- and $p_{T^+} \approx 1$ enforces equal or higher predictions after T^+. The second condition in Eq. 5 represents the scenario where the conversion (if the scan ever converts) occurs after 18 months and exact T^+ and T^- are unknown. Here, $\mathcal{L}_{ce}$ is computed at $t = 0$ and 1 with a GT label of 0 in both cases. The last condition in Eq. 5 occurs when the input OCT scan is the first visit of conversion and the GT label remains 1 throughout the 18-month interval. In addition to $\mathcal{L}_{cls}$, two regularization terms are also employed. Thus, the total loss

$$\mathcal{L}_{tot} = \mathcal{L}_{cls} + \gamma_1 ||a||_2^2 + \gamma_2 ||\mathbf{M} \odot (1 - \widehat{\mathbf{R}})||_1, \qquad (6)$$

where $\gamma_1 = \gamma_2 = 0.1$. An L2-regularization of a is performed for numerical stability. Moreover, higher activations in $\mathbf{M}$ outside the retina defined by the binary mask $\widehat{\mathbf{R}}$ are penalized. $\widehat{\mathbf{R}}$ is the ROI mask of the input scan $\mathbf{R}$ resized to 12×12×16.

IV. Experiments

A. Dataset

A private longitudinal dataset was created from the Fellow Eyes of a real-world retrospective cohort of OCT scans from the PINNACLE consortium [19] collected from the University Hospital Southampton and Moorfields Eye Hospital. The images were acquired using Topcon scanners with an average 3.6 ± 5.7 months interval between successive visits. A subset of the dataset was manually labelled for the TTC task and the remaining were used to train Morph-SSL.

The *SSL Dataset* had 3570 unlabelled OCT scans from multiple visits of 399 eyes with at least 3 visits per eye. Whenever treatment information was available, the visits after the first anti-VEGF injection were removed to ensure that most scans in the dataset are in the iAMD stage.

The *TTC Dataset* with 343 Eyes (2418 OCT Volumes) was manually examined by clinical experts for the downstream task. In our experiments, each OCT scan was considered independently and the corresponding GT labels for T^+ (and T^-) were obtained as the time-interval between the current visit and the manually identified first visit of conversion (and the visit just before it). All Scans after the first visit of conversion were removed to focus on the iAMD stage and the earliest indicators of nAMD in the first conversion visit.

B. Preprocessing:

The top and bottom boundaries delineating the retinal tissue called the Inner Limiting Membrane (ILM) and the Bruch's Membrane (BM) were extracted using the automated method in [39]. Thereafter, the curvature of the retinal surface was flattened by shifting each A-scan by an offset such that the BM lies on a straight plane similar to [39]. The binary ROI mask of the retina contained the region from 26 μm above the ILM to 169 μm (to include the choroid) below the BM. Both the OCT and its ROI mask were then cropped to the central 3×3 mm^2 en-face region. This region has been correlated with the onset of GA and neovascularization [20]. Finally, the volume was resized to 192×192×32 and its intensity linearly scaled to $[-1,1]$.

As an additional preprocessing for the *SSL Dataset*, the enface projections of all visits of an eye were aligned to its first visit using the unsupervised affine registration method in [20]. This step ensures that the Morph-SSL features capture the structural changes caused by AMD progression instead of image misalignment. The step is not performed for the *TTC Dataset* where each visit's scan is considered independently.

C. Experimental Setup

Morph-SSL was trained on image pairs formed from two random visits of the same eye, acquired within two years from each other. The *SSL Dataset* was randomly divided into 350 eyes (14078 image pairs) for training, 25 eyes (640 image pairs) for validation and the remaining 24 eyes (600 image pairs) for a qualitative evaluation of the learned features (see Fig. 5).

A stratified five-fold evaluation was conducted for the TTC task to reduce the bias of a specific train-test data split. The *TTC Dataset* was randomly divided into 5 mutually exclusive parts at the eye level. The experiments were repeated 5 times, each time considering one part as the held out test set while the remaining dataset was randomly divided into 85% for training and 15% for validation. The performance was evaluated for predicting the conversion to nAMD within $t = 0, 6, 12$ and 18 months, where t = 0 indicates that the input image is the first visit of conversion. We evaluated prediction scores using the area under the receiver operating characteristic curve (AUC), and

we evaluated binary predictions using balanced accuracy calculated as (Sensitivity+Specificity)/2 after thresholding the prediction scores at an operating point that maximized the Youden's J statistic. We performed both scan-level and eye-level evaluations. The scan-level performance was evaluated on all scans in the test set by treating each patient visit as an independent sample. The average performance across the five-folds is reported in the Appendix, Tables VI-X and the Delong test was employed for statistical significance between AUCs using the pyroc 0.20 library [40]. The eye-level performance assessment employed a *bootstrapping* based approach, incorporating 1000 random eye-level re-samplings of the test set in each fold. Each re-sampling of the test-set contained only one OCT scan (by randomly selecting a patient visit) from each eye. This re-sampling process was repeated 1000 times for each of the five folds, resulting in a total of $5 \times 1000 = 5000$ sample estimates for each performance metric (AUC and balanced accuracy). The mean and standard deviation across these 5000 sample estimates are reported in Tables I-V and the statistical significance was ascertained with the Wilcoxon Signed Rank Test.

D. Implementation Details

All experiments were implemented in Python 3.8.5 with Pytorch 1.8.1 on a server, using a single NVIDIA A100, 40 GB GPU. An implementation of the proposed method is available at: https://github.com/arunava555/Morph-SSL. Both the Morph-SSL and downstream TTC training employed similar Data Augmentations comprising random 3D translations (up to 15% of the image size along each axis), random horizontal flip (with 0.5 probability), Gaussian blurring ($\mu = 0$, random $\sigma \in [0, 0.9]$) and Gaussian noise ($\mu = 0$, $\sigma = 0.001$). For Morph-SSL, both scans in the training image pair were translated and flipped identically, while other augmentations were applied independently. During both training stages, Adam optimizer [41] was used ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay $= 10^{-5}$ for Morph-SSL, 10^{-2} for TTC) with a cyclic learning rate schedule [42] where the learning rate was linearly varied from lr_{min} (10^{-6} for Morph-SSL, 10^{-5} for TTC) to $lr_{max} = 10^{-4}$ and back to lr_{min} in each epoch. The validation performance was monitored to save the best network weights using minimum loss for Morph-SSL and highest average AUC for TTC.

Morph-SSL trained with a batch size of 1 for 160 epochs, 2000 batch updates per epoch, required 23 GB GPU memory. The downstream training was performed for 400 epochs of 500 batch updates. A batch size of 6 was employed when the Encoder and Classifier were fine-tuned together on the TTC task, requiring 28 GB GPU memory. Training the Classifier alone required 4GB of GPU for a batch size of 16. During inference, the proposed method requires 30.487 GFLOPs and takes an average of 0.04 seconds per image using GPU and 2.2 seconds per image using CPU alone for the downstream conversion prediction task.

During Morph-SSL based pre-training, the size of the encoder's output and the tunable weights of the loss terms in Eqs. 1 and 3 were empirically fixed by conducting a preliminary hyperparameter search. The large amount of time required in training multiple 3D models prevented a thorough hyperparameter grid search. So, multiple models were trained with different hyperparameter configurations on a small subset of 100 image pairs from the entire *SSL Dataset*. They were manually selected to cover a range of morphological changes, from small to moderate changes in the drusen structure, to large changes in the retinal thickness due to the presence of abnormalities such as PED. Reducing the size of the Encoder's output feature map below $128@12 \times 12 \times 16$ was found to have an adverse impact on the quality of the reconstructed images. Initially, each loss term was scaled in powers of 10 to balance their values to a similar range. Next, the weight of each loss term was varied one at a time in orders of 10 (keeping the other loss weights fixed). The output image reconstructions from the trained models were visually found to provide better image reconstructions when the scale of the loss values were in the following order: $\mathcal{L}_{prc} >$ deformation term in $\mathcal{L}_{mse} >$ additive term in $\mathcal{L}_{mse} > \mathcal{L}_{fld} > \mathcal{L}_{add} > \mathcal{L}_{smth}$. Here, the first three terms guide the network toward better reconstruction, while the weights of the last three regularization terms are kept relatively low to enable it to learn large transformations.

V. Results

A. Results on the TTC Task

1) Impact of Morph-SSL : In Table I, rows 1-3, we evaluate 3 training setups: (a) end-to-end training from random weight initialization (RI); (b) freeze the Morph-SSL trained Encoder weights and only train the classifier on the TTC task (FR); (c) use the Morph-SSL trained Encoder weights and the learned classifier weights from (b) to initialize and perform end-to-end finetuning of the Encoder and Classifier on the TTC task (FN).

The Morph-SSL features showed significant performance improvement, even without fine-tuning, over end-to-end training form scratch (row 1 vs 3). Further fine-tuning on the TTC task (row 1 vs. 2) did not lead to a statistically significant performance improvement, except for $t = 18$. This indicates that the initial Morph-SSL trained weights are very close to the optimal network weights for the TTC task. Overall, a good performance is observed in identifying the scans that have just converted to nAMD ($t = 0$) or are about to convert within 6 months. However, the AUC drops progressively as we consider larger time-intervals for forecasting into the future. This may indicate that often, distinct morphological changes signaling imminent nAMD conversion appear unexpectedly only a few months before conversion rather than gradually over a long period. Similar trends are also observed for the scan-level performance reported in the Appendix, Table VI.

A few examples of the Saliency Maps $\mathbf{M}$ of the proposed method are shown in Fig. 4 for OCT scans that convert at different time intervals in the future. It shows that the trained model is sensitive to abnormalities in the outer retina such as drusen, PED and HRF, which are known to be associated with AMD progression [2], [16], [17].

2) Comparison With TTC regression: Conversion prediction can also be posed as a regression task for predicting the TTC. In order to compare our approach against regression,

TABLE I

EYE-LEVEL EVALUATION (MEAN±STD. DEVIATION) OF MORPH-SSL PRE-TRAINING AFTER FREEZING WEIGHTS (FR) AND END-TO-END FINE-TUNING (FN), COMPARED WITH RANDOM NETWORK INITIALIZATION (RI). THE PROPOSED METHOD OF MODELING THE TIME-TO-CONVERSION IS COMPARED AGAINST REGRESSION IN ROWS 4-5. THE VALUES HIGHLIGHTED WITH * ARE **NOT** STATISTICALLY DIFFERENT FROM PROPOSED-FR (ROW 1) WITH $p > 0.05$

SL. No.		0 month		6 month		12 month		18 month	
		AUC	Bal Acc.	AUC	Bal Acc.	AUC	Bal Acc.	AUC	Bal Acc.
1	Proposed-FR	$\mathbf{0.856 \pm 0.05}$	0.817 ± 0.05	0.777 ± 0.05	0.753 ± 0.04	0.733 ± 0.07	0.729 ± 0.05	0.712 ± 0.09	0.728 ± 0.07
2	Proposed-FN	0.851 ± 0.05	$\mathbf{0.818 \pm 0.05}^*$	$\mathbf{0.779 \pm 0.05}^*$	$\mathbf{0.758 \pm 0.04}$	$\mathbf{0.734 \pm 0.07}^*$	$\mathbf{0.735 \pm 0.05}$	$\mathbf{0.719 \pm 0.09}$	$\mathbf{0.737 \pm 0.07}$
3	Proposed-RI	0.785 ± 0.08	0.763 ± 0.07	0.711 ± 0.06	0.711 ± 0.05	0.688 ± 0.08	0.701 ± 0.06	0.676 ± 0.1	0.702 ± 0.07
4	Regression-FR	—	0.681 ± 0.06	—	0.660 ± 0.04	—	0.633 ± 0.07	—	0.583 ± 0.08
5	Regression-FN	—	0.706 ± 0.08	—	0.680 ± 0.07	—	0.654 ± 0.07	—	0.589 ± 0.08

TABLE II

EYE-LEVEL PERFORMANCE (MEAN ± STD.DEV) FOR ABLATION ON THE ENCODER-DECODER ARCHITECTURE. EACH NETWORK IS PRE-TRAINED WITH MORPH-SSL AND EVALUATED BY EITHER FREEZING (FR) WEIGHTS OR END-TO-END FINETUNING (FN) ON THE DOWNSTREAM TASK. THE BEST VALUE IN EACH COLUMN IS HIGHLIGHTED IN BOLD. THE STATISTICAL SIGNIFICANCE OF ROWS 2-4 IS COMPARED WITH ROW 1 AND ROWS 6-8 WITH ROW 5 AND THE VALUES HIGHLIGHTED WITH * ARE **NOT** STATISTICALLY DIFFERENT WITH $p > 0.05$

SL. No.		0 month		6 month		12 month		18 month	
		AUC	Bal Acc.	AUC	Bal Acc.	AUC	Bal Acc.	AUC	Bal Acc.
1	Proposed-FR	$\mathbf{0.856 \pm 0.05}$	0.817 ± 0.05	0.777 ± 0.05	0.753 ± 0.04	0.733 ± 0.07	0.729 ± 0.05	0.712 ± 0.09	0.728 ± 0.07
2	3D Convolutions-FR	0.840 ± 0.06	0.804 ± 0.06	0.760 ± 0.06	0.740 ± 0.05	0.718 ± 0.08	0.720 ± 0.06	0.705 ± 0.09	0.721 ± 0.07
3	BatchNorm-FR	0.842 ± 0.05	0.811 ± 0.05	0.756 ± 0.05	0.740 ± 0.04	0.724 ± 0.06	$0.726 \pm 0.05^*$	0.719 ± 0.08	0.734 ± 0.06
4	Additive skip conn-FR	0.829 ± 0.06	0.794 ± 0.05	0.753 ± 0.05	0.738 ± 0.04	0.713 ± 0.07	0.716 ± 0.05	0.688 ± 0.08	0.711 ± 0.06
5	Proposed-FN	0.851 ± 0.05	$\mathbf{0.818 \pm 0.05}$	$\mathbf{0.779 \pm 0.05}$	$\mathbf{0.758 \pm 0.04}$	$\mathbf{0.734 \pm 0.07}$	$\mathbf{0.735 \pm 0.05}$	0.719 ± 0.09	$\mathbf{0.737 \pm 0.07}$
6	3D Convolutions-FN	0.844 ± 0.05	0.807 ± 0.05	0.763 ± 0.05	0.740 ± 0.04	0.722 ± 0.08	0.722 ± 0.05	0.708 ± 0.09	0.723 ± 0.07
7	BatchNorm-FN	0.834 ± 0.06	0.802 ± 0.05	0.757 ± 0.05	0.741 ± 0.04	0.727 ± 0.07	0.728 ± 0.05	$\mathbf{0.722 \pm 0.07}^*$	$0.735 \pm 0.05^*$
8	Additive skip conn-FN	0.823 ± 0.06	0.792 ± 0.06	0.754 ± 0.06	0.738 ± 0.05	0.714 ± 0.08	0.721 ± 0.05	0.688 ± 0.09	0.713 ± 0.06

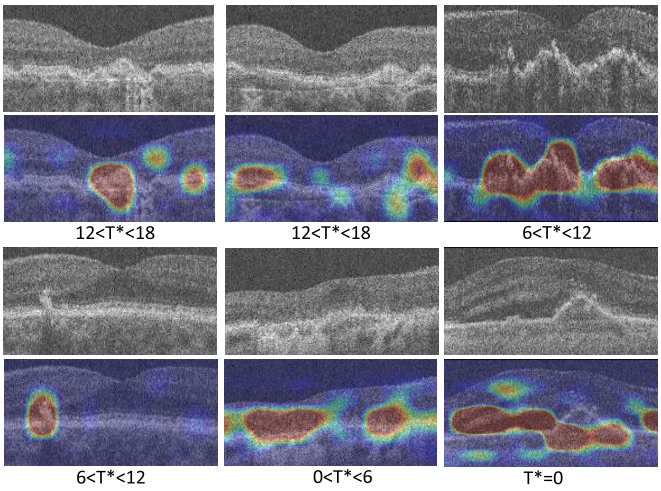

Fig. 4. Examples of saliency maps with frozen Morph-SSL weights.

we use the same Encoder (pre-trained with Morph-SSL) and the downstream classification network architecture (in Fig. 3(b)) but the prediction obtained through GAP of the final single channel output $\mathbf{M}$ is treated as the time-to-conversion from the current visit. The GT was computed as the mean of T^- and T^+ and used for training with a MSE loss. The output prediction is binned into 6-month intervals to obtain the binary prediction of nAMD conversion within $t = 0, 6, 12, 18$ months for comparison with our method (see row 4,5 in Table I for eye-level and Appendix, Table VI for scan-level performance). Since, the resulting predictions are binary and not continuous scores, AUC could not be computed and only balanced accuracy has been reported. A large drop in performance is observed between this regression-based approach as compared to our proposed method of modeling the CDF (with a sigmoid function over time) trained with the BCE loss at different time-points in Eq. 5. This performance gap can be primarily attributed to the inability of the regression-based approach to utilize training samples that do not convert within the time-period of 18 months as their GT label for the conversion time is unknown while these samples are also used for training in our formulation (second condition in Eq. 5). Another problem with the regression-based formulation is the unavailability of the exact conversion date T^* between T^+ and T^- leading to noisy labels. Furthermore, in some cases, formulating regression problems as a classification task has been shown to yield better performance. This has been linked to the ability of the cross-entropy-based classification loss to learn high-entropy (more diverse) feature representations as compared to regression with a MSE loss [43].

3) Impact of the Encoder Architecture : In this work, we propose an efficient CNN for 3D input images that optimizes the amount of computation and trainable parameters. Our solution involves two modifications: (i) substituting $3 \times 3 \times 3$, 3D convolutions with S3DConv which applies 3×3 convolutions along the three orthogonal spatial orientations (see Fig. 2(a)); (ii) using concatenation-based skip connections [44] instead of the additive residual skip connections in the Basic Blocks in Fig. 2(b), (d) which are used at each scale of our Encoder and Decoder architectures. We also employ Layer Normalization instead

TABLE III

EYE-LEVEL PERFORMANCE (MEAN ± STD. DEVIATION) FOR ABLATION ON THE DOWNSTREAM TTC CLASSIFICATION LOSS (ROWS 2-3) AND THE CLASSIFIER ARCHITECTURE (ROWS 4-5). THE MORPH-SSL PRE-TRAINED ENCODER WEIGHTS ARE FROZEN AND ONLY THE CLASSIFIER IS TRAINED IN EACH EXPERIMENT. THE BEST VALUE IN EACH COLUMN IS HIGHLIGHTED IN BOLD. ROWS 2-5 ARE COMPARED WITH ROW 1 AND THE VALUES WHICH ARE **NOT** STATISTICALLY SIGNIFICANT ($p > 0.05$) ARE HIGHLIGHTED WITH A *

SL. No.		0 month		6 month		12 month		18 month	
		AUC	Bal Acc.	AUC	Bal Acc.	AUC	Bal Acc.	AUC	Bal Acc.
1	Proposed	0.856 ± 0.05	0.817 ± 0.05	$\mathbf{0.777 \pm 0.05}$	$\mathbf{0.753 \pm 0.04}$	$\mathbf{0.733 \pm 0.07}$	0.729 ± 0.05	$\mathbf{0.712 \pm 0.09}$	$\mathbf{0.728 \pm 0.07}$
2	no $\|\|a\|\|_2^2$	$0.852 \pm 0.05^*$	$0.818 \pm 0.05^*$	0.765 ± 0.05	0.746 ± 0.04	0.723 ± 0.08	0.722 ± 0.05	0.699 ± 0.09	0.717 ± 0.07
3	no $\|\|\mathbf{M} \odot (1 - \widehat{\mathbf{R}})\|\|_1$	0.845 ± 0.06	0.811 ± 0.05	0.762 ± 0.06	0.741 ± 0.05	0.723 ± 0.08	0.721 ± 0.06	$0.707 \pm 0.09^*$	$0.727 \pm 0.07^*$
4	Multilabel Classifier	$\mathbf{0.879 \pm 0.05}$	$\mathbf{0.842 \pm 0.05}$	$0.774 \pm 0.06^*$	$0.752 \pm 0.05^*$	0.724 ± 0.07	0.722 ± 0.05	0.700 ± 0.08	0.718 ± 0.07
5	Separate a prediction	$0.853 \pm 0.05^*$	$0.816 \pm 0.06^*$	$0.776 \pm 0.05^*$	$0.751 \pm 0.04^*$	$0.732 \pm 0.07^*$	$\mathbf{0.730 \pm 0.05}^*$	$0.710 \pm 0.08^*$	$0.727 \pm 0.06^*$

of Batch Normalization which enables stable training with small batch sizes on limited GPU memory. Substituting $3 \times 3 \times 3$ convolutions with S3DConv in our Encoder-Decoder architecture reduces computation from 189 GFLOPs (with full 3D convolutions) to 71 GFLOPs and the 6,622,457 trainable network parameters to 2,702,329 during the Morph-SSL based pre-training. To evaluate the impact of the concatenation-based skip connections, the proposed Basic Encoder and the Basic Decoder Block in Fig. 2(b), (d) were substituted with the additive residual connection-based alternatives depicted in the Appendix, Fig. 7(a), (b). These modified Basic Encoder and Decoder Blocks require double the number of convolutional filters in the two S3DConv layers inside them compared to our concatenation-based skip connections for the same number of input and output channels, thereby requiring significantly more computation (203G vs 71G FLOPs) and network parameters (7,025,881 vs 2,702,329).

Although full 3D convolutions require a significant amount of computation and network parameters, S3DConv performs better in both settings with frozen Morph-SSL pre-trained weights (rows 1,2 in Table II) and fine-tuning (rows 5,6 in Table II) at all time-points. Full 3D convolutions are more prone to over-fitting when trained on limited labelled data for the downstream conversion prediction task. The performance of Layer Normalization (pre-trained with a batch size of 1) and Batch Normalization (pre-trained with a batch size of 2, limited by GPU memory) is presented in Table II (rows 1,3 for frozen Morph-SSL pre-trained weights and rows 3,5 with fine-tuning). Overall, Layer Normalization outperforms Batch Normalization under both settings at $t = 0, 6, 12$ whereas Batch Normalization performs better at $t = 18$. The proposed concatenation-based skip connections also outperform the residual additive skip connections for both the frozen and fine-tuned models with a statistically significant difference at all time-points (rows 1,4 and rows 5,8 in Table II). Our architectural choices also exhibited better performance in the scan-level evaluations reported in Appendix, Table VII.

4) Impact of the Loss Terms for the TTC Task : An ablation of the auxiliary loss terms in Eq. 6 is evaluated in rows 2, 3 of Table III at an eye-level (and Table VIII at a scan-level). The removal of L-2 regularization on the slope parameter a (row 1 vs 2) causes a minor drop in the AUC and Balanced Accuracy across all time points except $t = 0$, where no statistically significant difference is observed.

Removing the loss term which penalizes high activations outside the retinal tissue leads to a small drop in both AUC and Balanced Accuracy for all time-points (row 3 vs 1). However the difference at $t = 18$ was not statistically significant. Similar overall trends are also observed in the scan-level performance reported in the Appendix, Table VIII.

5) Impact of Our TTC Formulation: We propose to model the CDF of the TTC with a sigmoidal function. An alternative way is to pose it as multi-label classification with each class indicating if the image converts within a discrete time-point [13], [14], [30]. We compare our eye-level performance against multi-label classification in Table III, row 4 by modifying the last layer of our classifier architecture to produce a 4 channel ouptut (instead of 1), to which GAP is applied followed by a sigmoid activation to obtain the predictions for the 4 time-points. Both in terms of AUC and the Balanced accuracy, the proposed method clearly outperforms multi-label classification at $t = 12, 18$. At $t = 6$, the slightly better performance of our method was not statistically significant while the multi-label classification performed better at $t = 0$. Similar performance trends are also observed at the scan-level in the Appendix, Table VIII row 4. Although the performance of both methods are similar, our approach guarantees the monotonic increasing property of the CDF (e.g., the probability of an eye to convert within 12 months cannot be lower than the conversion probability within 6 months) which is not the case with multi-label classification. Across the 5 folds, the multi-label classifier is inconsistent in some cases, with higher prediction scores for a previous time-point compared to the next for a given input scan, 16 cases between $t = 0, 6$ months, 60 cases between $t = 6, 12$ and 84 cases with inconsistencies between $t = 12, 18$ months. Additionally, once trained, our model can predict conversion risk at any time-point within 18 months by varying t in eq 4, unlike multi-label classification that can predict conversion risk only at predefined discrete time intervals used during training.

6) Architecture Design to Predict Slope : Spatial entropy of $\mathbf{M}$ was used to predict the slope a of the sigmoid function. We compared this design choice against one which predicts two channels. One channel is used similar to $\mathbf{M}$ to compute b while a GAP is applied to the second channel for obtaining a. Although this new architecture requires extra network parameters, it did not result in a statistically significant difference in performance (see rows 1, 5 in Table III).

TABLE IV

EYE-LEVEL PERFORMANCE (MEAN±STD. DEVIATION) TO BENCHMARK THE PROPOSED METHOD (FINE-TUNED FROM MORPH-SSL PRE-TRAINED WEIGHTS) AGAINST STANDARD 3D NETWORKS (FINE-TUNED FROM THEIR AVAILABLE WEIGHTS PRE-TRAINED ON THE KINETICS DATASET) AND HANDCRAFTED BIOMARKERS WITH A RANDOM FOREST CLASSIFIER. THE BEST PERFORMANCE IN EACH COLUMN IS HIGHLIGHTED IN BOLD. THE PERFORMANCE DIFFERENCE OF ROWS 2-4 COMPARED TO THE PROPOSED METHOD (ROW 1) WAS FOUND TO BE STATISTICALLY SIGNIFICANT WITH ($p < 0.05$) FOR ALL TIME-POINTS

SL. No.		0 month		6 month		12 month		18 month	
		AUC	Bal Acc.	AUC	Bal Acc.	AUC	Bal Acc.	AUC	Bal Acc.
1	Proposed-Finetune	$\mathbf{0.851 \pm 0.05}$	$\mathbf{0.818 \pm 0.05}$	$\mathbf{0.779 \pm 0.05}$	$\mathbf{0.758 \pm 0.04}$	$\mathbf{0.734 \pm 0.07}$	$\mathbf{0.735 \pm 0.05}$	$\mathbf{0.719 \pm 0.09}$	$\mathbf{0.737 \pm 0.07}$
2	I3D [45]	0.784 ± 0.08	0.776 ± 0.06	0.715 ± 0.07	0.716 ± 0.05	0.672 ± 0.07	0.693 ± 0.05	0.667 ± 0.08	0.697 ± 0.06
3	X3D [46]	0.774 ± 0.07	0.769 ± 0.06	0.710 ± 0.06	0.716 ± 0.05	0.682 ± 0.07	0.700 ± 0.05	0.679 ± 0.08	0.710 ± 0.06
4	Biomarkers+Random Forest	0.752 ± 0.07	0.738 ± 0.05	0.695 ± 0.07	0.699 ± 0.06	0.664 ± 0.08	0.678 ± 0.06	0.629 ± 0.09	0.666 ± 0.06

TABLE V

EYE-LEVEL AREA UNDER THE ROC CURVE (MEAN±STD. DEVIATION) TO COMPARE SSL METHODS UNDER DIFFERENT TRAINING CONFIGURATIONS BY: EITHER TRAINING ON ONE-THIRD OR THE ENTIRE TRAINING DATASET; EITHER FREEZING SSL-TRAINED WEIGHTS OR FINETUNING END-TO-END. THE VALUES HIGHLIGHTED WITH * IN EACH COLUMN ARE **NOT** STATISTICALLY DIFFERENT ($p > 0.05$) COMPARED TO THE PROPOSED METHOD TRAINED WITH IDENTICAL DATA (EITHER ONE-THIRD OR THE ENTIRE DATASET) AND PROTOCOL (EITHER FREEZE SSL WEIGHTS OR FINETUNE). THE BEST PERFORMANCE IN EACH COLUMN IS HIGHLIGHTED IN **BOLD**

	One-third Training data				Entire Training data			
	0 month	6 month	12 month	18 month	0 month	6 month	12 month	18 month
Proposed-Freeze	0.835 ± 0.06	0.764 ± 0.06	0.727 ± 0.08	0.705 ± 0.10	$\mathbf{0.856 \pm 0.05}$	0.777 ± 0.05	0.733 ± 0.07	0.712 ± 0.09
Proposed-Finetune	$\mathbf{0.846 \pm 0.05}$	$\mathbf{0.780 \pm 0.06}$	$\mathbf{0.729 \pm 0.07}$	$\mathbf{0.714 \pm 0.09}$	0.851 ± 0.05	$\mathbf{0.779 \pm 0.05}$	$\mathbf{0.734 \pm 0.07}$	0.719 ± 0.09
Model Genesis-Freeze [3]	0.785 ± 0.07	0.712 ± 0.06	0.671 ± 0.07	0.663 ± 0.08	0.801 ± 0.06	0.718 ± 0.06	0.680 ± 0.07	0.668 ± 0.09
Model Genesis-Finetune [3]	0.772 ± 0.06	0.716 ± 0.07	0.688 ± 0.09	0.661 ± 0.10	0.823 ± 0.06	0.763 ± 0.05	$0.733 \pm 0.07^*$	$\mathbf{0.728 \pm 0.09}$
Time prediction-Freeze [14]	0.586 ± 0.07	0.557 ± 0.08	0.527 ± 0.09	0.522 ± 0.11	0.690 ± 0.09	0.605 ± 0.08	0.575 ± 0.10	0.567 ± 0.12
Time prediction-Finetune [14]	0.626 ± 0.13	0.610 ± 0.12	0.582 ± 0.11	0.561 ± 0.12	0.735 ± 0.08	0.669 ± 0.07	0.644 ± 0.07	0.608 ± 0.09
Barlow Twins-Freeze [10]	0.765 ± 0.06	0.687 ± 0.06	0.630 ± 0.08	0.598 ± 0.10	0.742 ± 0.07	0.692 ± 0.07	0.667 ± 0.07	0.654 ± 0.09
Barlow Twins-Finetune [10]	0.747 ± 0.07	0.671 ± 0.07	0.656 ± 0.08	0.643 ± 0.10	0.749 ± 0.08	0.686 ± 0.07	0.669 ± 0.08	0.648 ± 0.09
VICReg-Freeze [9]	0.767 ± 0.07	0.710 ± 0.06	0.688 ± 0.07	0.674 ± 0.07	0.817 ± 0.05	0.744 ± 0.05	0.713 ± 0.06	0.688 ± 0.07
VICReg-Finetune [9]	0.808 ± 0.06	0.741 ± 0.06	0.711 ± 0.06	0.703 ± 0.07	0.824 ± 0.06	0.750 ± 0.05	0.714 ± 0.07	0.698 ± 0.08

7) Comparison With State-of-the-Art 3D Networks : An alternative to SSL is to fine-tune standard CNN networks after initializing them with the already available pre-trained weights. We compared our performance against two popular 3D-CNN networks, I3D [45] and X3D [46]. Their last fully connected layer was modified to predict a and b in Eq. 4. Both networks were initialized with pre-trained weights trained on the Kinetics video dataset and fine-tuned end-to-end on our task. Our Morph-SSL trained Encoder significantly outperformed both of these networks (see rows 1-3 in Table IV for eye-level and Appendix, Table IX for scan-level performance) in terms of both AUC and Balanced accuracy across all time-points.

8) Comparison With Handcrafted Biomarker Based Method : Recent clinical research has linked the spatial distribution of drusen and HRF to the future progression of AMD [16], [17]. Similar to [17], we segmented the drusen and HRF automatically using the Iowa Reference Algorithm [39] with modified smoothness constraints [47] for drusen and a deep learning-based method [48] for HRF. Thereafter, the volume of drusen, HRF, and their areas in the enface (surface) projection were computed in 14 spatial sectors (entire scan, central 1 mm disc, central 3 mm disc, central 6 mm disc, peri-fovea, para-fovea, peri-nasal, para-nasal, peri-superior, para-superior, peri-inferior, para-inferior, peri-temporal and para-temporal sectors) based on the ETDRS grid commonly used in clinical research. This resulted in a 14 (sectors) $\times$ 4 (area and volume of HRF and drusen) $=$ 56 dimensional feature vector. A multi-output random forest comprising 350 decision trees (selected based on best validation performance) was trained to predict the conversion to nAMD for the different time-points. The results presented in Table IV (and Appendix, Table IX), rows 1 and 4, indicate the superiority of the proposed method over handcrafted biomarkers. This highlights the inadequacy of the current clinically known prognostic features in predicting nAMD conversion and motivates the use of DL networks that directly learn the prognostic imaging features from raw OCT scans instead of hand-crafting them.

9) Comparison With Other SSL Methods : We compare the performance of Morph-SSL against other state-of-the-art SSL methods at an eye-level in Table V and scan-level in Appendix, Table X. The same 3D U-net and the transformations for the reconstruction task were employed for Model Genesis as reported in [3]. The time interval prediction task [14] was originally developed for the central B-scans alone, however we implemented a 3D version using our Encoder architecture for a fair comparison. The latest CL methods, VICReg [9] and Barlow Twins [10] could not be trained in 3D due to their large batch size requirements. They were used to train a ResNet-50 with a batch size of 128 following [13]. The positive image pairs were constructed by selecting B-scans (from the same position) from two random visits of the same eye within 18 months and applying the data augmentations

used in [13]. The Classifier was modified to handle a 2048×32 (feature dimensions $\times$ B-scans) input. First, a 1D convolution layer with 32 input and 1 output feature channel was used to obtain a 2048 dimensional feature for the entire OCT volume. This was followed by two fully connected layers with 1024 and 2 neurons respectively, to get the predictions for a and b in Eq. 4. The SSL methods were compared under different training setups by: (a) using the SSL-trained features off-the-shelf and only training the Classifier (Freeze) vs. initialization with the SSL-trained network weights for end-to-end fine-tuning (Finetune), and (b) training on the entire vs. one-third of the supervised training data. To evaluate performance in a small data regime, one-third of the training data in each fold of the *TTC Dataset* was randomly selected and kept consistent across all SSL methods.

Small Data Regime: Morph-SSL outperforms all benchmark methods (under identical Freeze/Finetune setup) across all 4 time-points in Table V. All differences were statistically significant. Overall, finetuning improves performance over frozen weights for all SSL methods except for Model-Genesis at $t = 0, 18$ and Barlow Twins at $t = 0, 6$.

Entire Training Data: In the Freeze setup, again Morph-SSL outperforms all benchmark methods in terms of AUC with a statistically significant difference across all 4 time-points. In the fine-tuning setup, Morph-SSL still clearly outperforms time-interval prediction [14], Barlow twins [10] and VICReg [9] across all time-points. Although Morph-SSL outperformed Model-Genesis at $t = 0, 6$, the AUC difference was not statistically significant for $t = 12$ (p-value 0.43), while Model-Genesis performed better at $t = 18$.

Overall, Morph-SSL shows better performance than other methods, particularly in scenarios where features are used off-the-shelf or in a small data regime with limited labeled data for fine-tuning. When trained on the entire dataset, Morph-SSL was found to learn strong features with good performance on the TTC task with minimal effect of further fine-tuning.

B. Risk Score for Progression to nAMD

As part of current clinical practice, patients with iAMD undergo regular eye examinations and nAMD must be treated at the earliest sign of onset to prevent vision loss. However, biomarkers such as drusen volume and hyper-reflective foci (HRF) are insufficient to reliably identify the patients at a higher risk of conversion. As a result, an automated method to stratify iAMD eyes into low and high risk groups for future conversion to nAMD can help clinicians prioritize patients in the high risk group for early treatment and frequent monitoring. An ideal risk score should a) be a single time-independent scalar value; b) be bounded in the range [0, 1]; c) be inversely proportional to the predicted time to conversion b. We formulate such a risk score by modifying Eq. 4 as $r = 2/\left[1 + exp\left\{\frac{b}{a+0.05}\right\}\right]$. The test predictions for a and b were obtained from the five folds to compute r for each OCT scan. The scans were then stratified into 3 groups with low risk ($0 \leq r \leq 0.33$), moderate risk ($0.33 < r \leq 0.67$) and high risk ($0.67 < r \leq 1$). A population-level survival function for these groups is plotted in Fig. 6 using the Kaplan–Meier estimator on the GT conversion time. It depicts the mean and standard deviation of the survival probability for each group, computed across 1000 re-samplings using eye-level bootstrapping. Each scan within a re-sampling was independent and came from a different eye as only one OCT scan (from a random visit) was selected per eye during bootstrapping. A log-rank t-test between the curves was performed in a pairwise manner among the three risk groups and the median p-value across all bootstrap re-samplings was used to determine the statistical significance. The difference between the survival curves of the low-risk and the high-risk groups was found to be highly statistically significant with a p-value of 0.007. The difference between the medium-risk vs. the high-risk group was not significant (p-value$=$ 0.263) while the difference between the low-risk and the medium-risk groups was also significant with a p-value$<$ 0.05 (p-value$=$ 0.034). Overall, r is effective in stratifying eyes coming from low and high risk groups.

1) Intra-Eye Consistency: AMD is a degenerative disease where the retinal tissue progressively deteriorates, so the predicted risk scores from scans across multiple visits of the same patient should be monotonically increasing over time. However, this consistency is not explicitly enforced by the downstream classifier for conversion prediction which uses single OCT scans as input and treats each image as an independent sample (although the Morph-SSL pretraining employs pairs of visits to learn the feature embedding). Therefore to quantitatively assess the intra-eye consistency in predicted risk scores, we computed the eye-level concordance index (eCI). It involved constructing a set of all possible pairs of visits $(\mathbf{I}_t, \mathbf{I}_{t+k})$ for each eye. Each scan was independently fed into our trained model to derive the corresponding pair of risk scores (r_t, r_{t+k}). The eCI was then calculated as the fraction of the visit pairs (out of the total number of all possible pairs for the eye), in which the predicted risk scores adhered to the desired ordering $r_{t+k} \geq r_t$. Despite that the classifier treated each eye as an independent sample, the average eCI across all eyes in all folds was 0.73, indicating a moderately good consistency in the risk score predictions.

C. Interpolation in the Morph-SSL Feature Space

Given a pair of scans $\mathbf{I}_t$, $\mathbf{I}_{t+k}$ from two visits of the same eye, we extract their features $\mathbf{F}_t$ and $\mathbf{F}_{t+k}$, and generate an intermediate feature through linear interpolation as $\mathbf{F}'_\rho = \mathbf{F}_t + \rho.(\mathbf{F}_{t+1} - \mathbf{F}_t)$, where $\rho \in [0, 1]$. By using $\mathbf{F}_t$ and $\mathbf{F}'_\rho$ (instead of $\mathbf{F}_{t+k}$) as inputs to the Morph-SSL trained Decoder, we can predict the transformation that morphs $\mathbf{I}_t$ to artificially generate the intermediate OCT scan for $\mathbf{F}'_\rho$ (see Fig. 1(c)). The qualitative results in Fig. 5 depict four intermediate scans (along each column) by varying ρ. A gradual smooth transition between $\mathbf{I}_t$ and $\mathbf{I}_{t+k}$ is observed with the generated scans. Such a smooth feature embedding is enforced by our Decoder architecture which explicitly correlates the direction of the feature displacement $\mathbf{F}'_\rho - \mathbf{F}_t$ to the *type*, and its magnitude to the *amount* of the morphing transformation. The magnitude increases with ρ while the direction remains the same.

This property may be explored in the future for different applications. Balanced-Mixup [49] generates artificial training

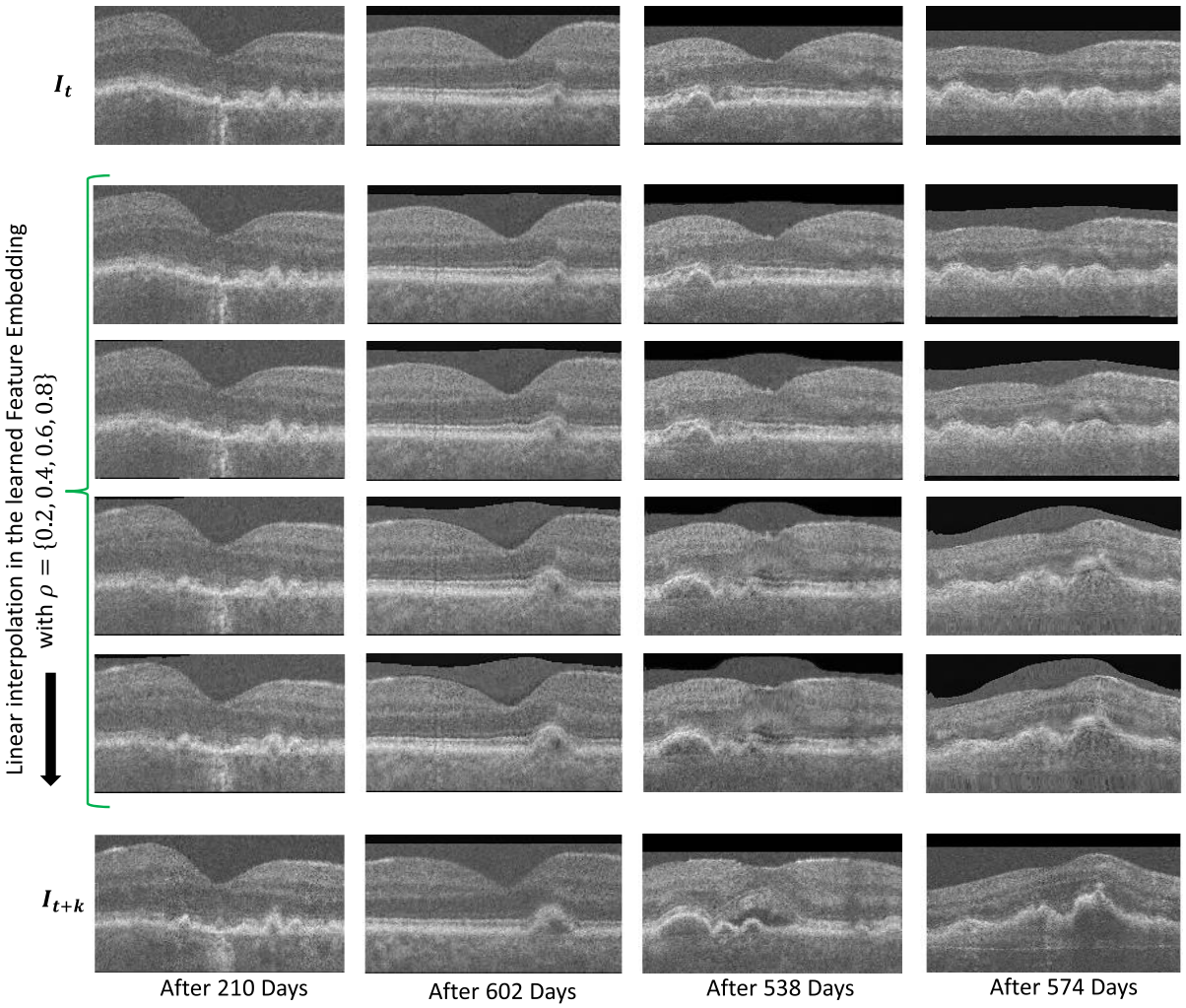

Fig. 5. Qualitative visualization of the linear interpolation between the features extracted from two OCT volumes $\mathbf{I}_t$ (first row) and $\mathbf{I}_{t+k}$ of the same eye (last row). A single B-scan from the 3D volume has been depicted for a different eye in each column. The smooth transition in the generated intermediate images demonstrates Morph-SSL's ability to learn a feature representation with a meaningful notion of distance and direction that correspond to specific morphological changes in the input scan.

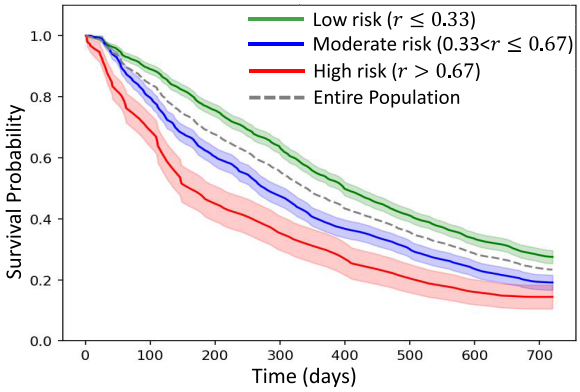

Fig. 6. Kaplan-Meier curves for different risk groups.

samples by directly interpolating the voxels between two training images, which may produce blurry images. Through small interpolations in our feature embedding instead, better training samples may be generated. Another potential application could be to generate approximations of future OCT scans to visualize disease progression. A Recurrent Neural Network to predict the features for future visits may be explored for this task.

VI. Conclusion

A vast amount of unlabelled longitudinal OCT scans are generated in clinics to monitor AMD. To leverage this data, we have proposed Morph-SSL, a novel SSL method designed to capture the temporal changes caused by disease progression. It ensures that the displacement in features between two OCT scans captures the morphological changes in the retina between them. With the Morph-SSL trained Encoder, we have developed a prognostic model for TTC estimation that predicts the future risk of conversion from iAMD to nAMD from the current OCT scan. The lack of reliable biomarkers and wide variability in the rate of AMD progression makes it a challenging task. We modelled the CDF of TTC with a sigmoidal function over time. The Morph-SSL features were found to perform well on the TTC task even without fine-tuning and showed significant improvements over training from scratch or fine-tuning standard 3D-CNNs with pre-trained weights. It also outperformed popular SSL methods

TABLE VI

SCAN-LEVEL EVALUATION (MEAN±STD. DEVIATION) OF MORPH-SSL PRE-TRAINING AFTER FREEZING WEIGHTS (FR) AND END-TO-END FINE-TUNING (FN), COMPARED WITH RANDOM NETWORK INITIALIZATION (RI). THE PROPOSED METHOD OF MODELING THE TIME-TO-CONVERSION IS COMPARED AGAINST REGRESSION IN ROWS 4-5. THE AUC VALUES HIGHLIGHTED WITH * ARE **NOT** STATISTICALLY DIFFERENT FROM PROPOSED-FR (ROW 1) WITH $p > 0.05$

SL. No.		0 month		6 month		12 month		18 month	
		AUC	Bal Acc.	AUC	Bal Acc.	AUC	Bal Acc.	AUC	Bal Acc.
1	Proposed-FR	$\mathbf{0.876 \pm 0.02}$	0.810 ± 0.02	0.767 ± 0.02	0.721 ± 0.01	0.718 ± 0.04	0.686 ± 0.02	0.693 ± 0.06	0.673 ± 0.04
2	Proposed-FN	$0.871 \pm 0.02^*$	$\mathbf{0.812 \pm 0.02}$	$\mathbf{0.768 \pm 0.02}^*$	$\mathbf{0.727 \pm 0.01}$	$\mathbf{0.721 \pm 0.05}^*$	$\mathbf{0.691 \pm 0.03}$	$\mathbf{0.700 \pm 0.06}$	$\mathbf{0.674 \pm 0.05}$
3	Proposed-RI	0.799 ± 0.04	0.742 ± 0.04	0.704 ± 0.03	0.673 ± 0.02	0.668 ± 0.06	0.650 ± 0.04	0.657 ± 0.08	0.638 ± 0.06
4	Regression-FR	—	0.686 ± 0.02	—	0.644 ± 0.02	—	0.618 ± 0.03	—	0.566 ± 0.05
5	Regression-FN	—	0.714 ± 0.02	—	0.675 ± 0.04	—	0.636 ± 0.03	—	0.592 ± 0.05

TABLE VII

SCAN-LEVEL PERFORMANCE (MEAN ± STD.DEV) FOR ABLATION ON THE ENCODER-DECODER ARCHITECTURE. EACH NETWORK IS PRE-TRAINED WITH MORPH-SSL AND EVALUATED BY EITHER FREEZING (FR) WEIGHTS OR END-TO-END FINETUNING (FN) ON THE DOWNSTREAM TASK. THE BEST VALUE IN EACH COLUMN IS HIGHLIGHTED IN BOLD. THE STATISTICAL SIGNIFICANCE OF THE AUC VALUES IN ROWS 2-4 IS COMPARED WITH ROW 1 AND ROWS 6-8 WITH ROW 5 WITH THE DELONG TEST. THE VALUES HIGHLIGHTED WITH * ARE **NOT** STATISTICALLY DIFFERENT WITH $p > 0.05$

SL. No.		0 month		6 month		12 month		18 month	
		AUC	Bal Acc.	AUC	Bal Acc.	AUC	Bal Acc.	AUC	Bal Acc.
1	Proposed-FR	$\mathbf{0.876 \pm 0.02}$	0.810 ± 0.02	0.767 ± 0.02	0.721 ± 0.01	0.718 ± 0.04	0.686 ± 0.02	0.693 ± 0.06	0.673 ± 0.04
2	3D Convolutions-FR	0.859 ± 0.03	0.801 ± 0.02	0.744 ± 0.02	0.695 ± 0.01	0.695 ± 0.04	0.663 ± 0.03	0.679 ± 0.06	0.653 ± 0.04
3	BatchNorm-FR	0.850 ± 0.03	0.803 ± 0.02	0.743 ± 0.02	0.705 ± 0.02	0.699 ± 0.04	0.668 ± 0.03	$0.690 \pm 0.06^*$	0.667 ± 0.04
4	Additive skip conn-FR	0.845 ± 0.03	0.789 ± 0.03	0.743 ± 0.02	0.694 ± 0.02	0.686 ± 0.03	0.657 ± 0.02	0.661 ± 0.04	0.643 ± 0.03
5	Proposed-FN	0.871 ± 0.02	$\mathbf{0.812 \pm 0.02}$	$\mathbf{0.768 \pm 0.02}$	$\mathbf{0.727 \pm 0.01}$	$\mathbf{0.721 \pm 0.05}$	$\mathbf{0.691 \pm 0.03}$	0.700 ± 0.06	$\mathbf{0.674 \pm 0.05}$
6	3D Convolutions-FN	0.858 ± 0.02	0.795 ± 0.03	0.748 ± 0.01	0.706 ± 0.02	0.698 ± 0.04	0.671 ± 0.02	0.684 ± 0.06	0.663 ± 0.04
7	BatchNorm-FN	0.846 ± 0.03	0.793 ± 0.02	0.750 ± 0.02	0.704 ± 0.02	0.710 ± 0.03	0.680 ± 0.01	$\mathbf{0.701 \pm 0.04}^*$	0.673 ± 0.01
8	Additive skip conn-FN	0.841 ± 0.03	0.777 ± 0.02	0.740 ± 0.02	0.697 ± 0.01	0.687 ± 0.04	0.658 ± 0.02	0.653 ± 0.05	0.639 ± 0.04

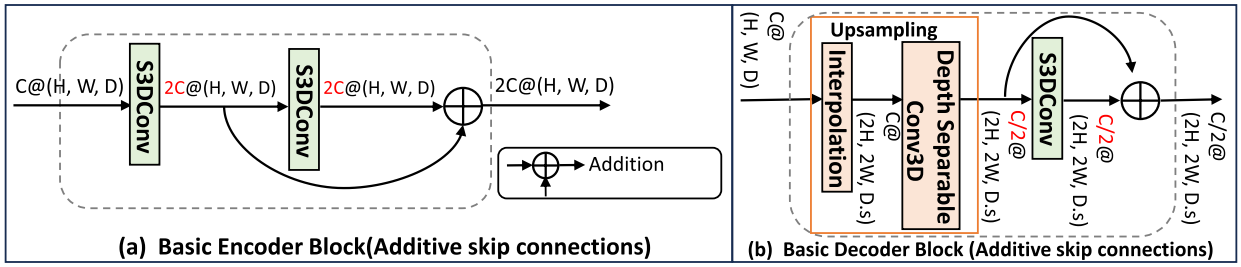

Fig. 7. The alternative architectures with additive skip connections used to replace the proposed Basic Encoder and Decoder Blocks (see Fig. 2(b), (d)) in the ablation experiments presented in Table II, VII.

TABLE VIII

SCAN-LEVEL PERFORMANCE (MEAN ± STD. DEVIATION) FOR ABLATION ON THE DOWNSTREAM TTC CLASSIFICATION LOSS (ROWS 2-3) AND THE CLASSIFIER ARCHITECTURE (ROWS 4-5). THE MORPH-SSL PRE-TRAINED ENCODER WEIGHTS ARE FROZEN AND ONLY THE CLASSIFIER IS TRAINED IN EACH EXPERIMENT. THE BEST VALUE IN EACH COLUMN IS HIGHLIGHTED IN BOLD. THE AUC VALUES IN ROWS 2-5 ARE COMPARED WITH ROW 1 USING THE DELONG TEST AND THE VALUES WHICH ARE **NOT** STATISTICALLY SIGNIFICANT ($p > 0.05$) ARE HIGHLIGHTED WITH *

SL. No.		0 month		6 month		12 month		18 month	
		AUC	Bal Acc.	AUC	Bal Acc.	AUC	Bal Acc.	AUC	Bal Acc.
1	Proposed	0.876 ± 0.02	0.810 ± 0.02	$\mathbf{0.767 \pm 0.02}$	$\mathbf{0.721 \pm 0.01}$	$\mathbf{0.718 \pm 0.04}$	$\mathbf{0.686 \pm 0.02}$	$\mathbf{0.693 \pm 0.06}$	$\mathbf{0.673 \pm 0.04}$
2	no $\|\|a\|\|_2^2$	$0.869 \pm 0.02^*$	0.809 ± 0.02	0.753 ± 0.02	0.706 ± 0.01	0.705 ± 0.04	0.666 ± 0.02	0.679 ± 0.06	0.653 ± 0.04
3	no $\|\|\mathbf{M} \odot (1 - \widehat{\mathbf{R}})\|\|_1$	0.865 ± 0.03	0.808 ± 0.02	0.755 ± 0.02	0.711 ± 0.01	$0.714 \pm 0.05^*$	0.685 ± 0.02	$0.692 \pm 0.06^*$	0.671 ± 0.03
4	Multilabel Classifier	$\mathbf{0.885 \pm 0.01}$	$\mathbf{0.841 \pm 0.01}$	$0.763 \pm 0.02^*$	0.720 ± 0.02	0.709 ± 0.05	0.681 ± 0.02	0.686 ± 0.06	0.667 ± 0.03
5	Separate a prediction	$0.869 \pm 0.03^*$	0.812 ± 0.03	$0.766 \pm 0.02^*$	0.715 ± 0.02	$0.717 \pm 0.04^*$	0.684 ± 0.02	$0.689 \pm 0.05^*$	0.659 ± 0.03

with significant gains in scenarios where SSL features are used off-the-shelf or fine-tuned on limited labeled data. We also derived a risk score that could be used to stratify eyes into low or high risk categories. Identifying iAMD patients with a high risk of progressing to nAMD can enable ophthalmologists to prioritize these cases for closer monitoring. Initiating treatment at the earliest sign of nAMD onset is crucial to prevent irreversible vision loss. Thus, our method to forecast the risk of future AMD progression can play a critical role in enabling patient-specific disease management and also aid in enriching

TABLE IX

SCAN-LEVEL PERFORMANCE (MEAN±STD. DEVIATION) TO BENCHMARK THE PROPOSED METHOD (FINE-TUNED FROM MORPH-SSL PRE-TRAINED WEIGHTS) AGAINST STANDARD 3D NETWORKS (FINE-TUNED FROM THEIR AVAILABLE WEIGHTS PRE-TRAINED ON THE KINETICS DATASET) AND HANDCRAFTED BIOMARKERS WITH A RANDOM FOREST CLASSIFIER. THE BEST PERFORMANCE IN EACH COLUMN IS HIGHLIGHTED IN BOLD. THE AUC DIFFERENCE OF ROWS 2-4 COMPARED TO ROW 1 WAS FOUND TO BE STATISTICALLY SIGNIFICANT ($p < 0.05$) FOR ALL TIME-POINTS USING THE DELONG TEST

SL. No.		0 month		6 month		12 month		18 month	
		AUC	Bal Acc.	AUC	Bal Acc.	AUC	Bal Acc.	AUC	Bal Acc.
1	Proposed-Finetune	$\mathbf{0.871 \pm 0.02}$	$\mathbf{0.812 \pm 0.02}$	$\mathbf{0.768 \pm 0.02}$	$\mathbf{0.727 \pm 0.01}$	$\mathbf{0.721 \pm 0.05}$	$\mathbf{0.691 \pm 0.03}$	$\mathbf{0.700 \pm 0.06}$	$\mathbf{0.674 \pm 0.05}$
2	I3D [45]	0.796 ± 0.05	0.752 ± 0.04	0.701 ± 0.02	0.677 ± 0.01	0.657 ± 0.03	0.640 ± 0.01	0.654 ± 0.03	0.625 ± 0.02
3	X3D [46]	0.792 ± 0.01	0.756 ± 0.02	0.710 ± 0.01	0.683 ± 0.01	0.673 ± 0.02	0.650 ± 0.02	0.666 ± 0.02	0.645 ± 0.01
4	Biomarkers+Random Forest	0.753 ± 0.03	0.708 ± 0.02	0.673 ± 0.04	0.653 ± 0.04	0.628 ± 0.05	0.619 ± 0.04	0.609 ± 0.06	0.607 ± 0.04

TABLE X

SCAN-LEVEL AREA UNDER THE ROC CURVE (MEAN±STD. DEVIATION) TO COMPARE SSL METHODS UNDER DIFFERENT TRAINING CONFIGURATIONS BY: EITHER TRAINING ON ONE-THIRD OR THE ENTIRE TRAINING DATASET; EITHER FREEZING SSL-TRAINED WEIGHTS OR FINETUNING END-TO-END. THE VALUES HIGHLIGHTED WITH * IN EACH COLUMN ARE **NOT** STATISTICALLY DIFFERENT ($p > 0.05$) COMPARED TO THE PROPOSED METHOD TRAINED WITH IDENTICAL DATA (EITHER ONE-THIRD OR THE ENTIRE DATASET) AND PROTOCOL (EITHER FREEZE SSL WEIGHTS OR FINETUNE). THE BEST PERFORMANCE IN EACH COLUMN IS HIGHLIGHTED IN **BOLD**

	One-third Training data				Entire Training data			
	0 month	6 month	12 month	18 month	0 month	6 month	12 month	18 month
Proposed-Freeze	0.842 ± 0.04	0.745 ± 0.03	0.697 ± 0.05	0.679 ± 0.06	$\mathbf{0.876 \pm 0.02}$	0.767 ± 0.02	0.718 ± 0.04	0.693 ± 0.06
Proposed-Finetune	$\mathbf{0.864 \pm 0.02}$	$\mathbf{0.764 \pm 0.03}$	$\mathbf{0.715 \pm 0.05}$	$\mathbf{0.690 \pm 0.06}$	0.871 ± 0.02	$\mathbf{0.768 \pm 0.02}$	$\mathbf{0.721 \pm 0.05}$	$\mathbf{0.700 \pm 0.06}$
Model Genesis-Freeze [3]	0.801 ± 0.02	0.704 ± 0.01	0.659 ± 0.02	0.648 ± 0.03	0.820 ± 0.02	0.717 ± 0.01	0.670 ± 0.03	0.658 ± 0.05
Model Genesis-Finetune [3]	0.783 ± 0.03	0.698 ± 0.04	0.653 ± 0.06	0.626 ± 0.06	0.846 ± 0.02	0.744 ± 0.02	$\mathbf{0.721 \pm 0.04}^*$	$0.699 \pm 0.06^*$
Time prediction-Freeze [14]	0.600 ± 0.04	0.542 ± 0.04	0.521 ± 0.06	0.522 ± 0.05	0.715 ± 0.06	0.590 ± 0.07	0.540 ± 0.07	0.536 ± 0.07
Time prediction-Finetune [14]	0.646 ± 0.12	0.610 ± 0.09	0.580 ± 0.07	0.565 ± 0.06	0.759 ± 0.05	0.655 ± 0.03	0.615 ± 0.02	0.583 ± 0.03
Barlow Twins-Freeze [10]	0.776 ± 0.03	0.669 ± 0.03	0.598 ± 0.03	0.559 ± 0.06	0.751 ± 0.03	0.678 ± 0.03	0.635 ± 0.03	0.612 ± 0.04
Barlow Twins-Finetune [10]	0.775 ± 0.05	0.670 ± 0.04	0.628 ± 0.05	0.610 ± 0.07	0.774 ± 0.03	0.679 ± 0.04	0.641 ± 0.04	0.614 ± 0.04
VICReg-Freeze [9]	0.783 ± 0.04	0.702 ± 0.02	0.668 ± 0.03	0.658 ± 0.03	0.842 ± 0.02	0.733 ± 0.01	0.684 ± 0.02	0.659 ± 0.03
VICReg-Finetune [9]	0.822 ± 0.03	0.738 ± 0.03	0.693 ± 0.04	$0.684 \pm 0.05^*$	0.852 ± 0.02	0.755 ± 0.01	0.700 ± 0.02	0.680 ± 0.03

clinical trial populations through the recruitment of patients at risk.

A. Limitations and Future Directions

The large amount of time required for training multiple 3D CNN networks prevented an exhaustive search for the optimal network architecture, the size of Encoder's output feature map, and the tunable weights of the loss terms used during Morph-SSL training, which remains a limitation of this work. Currently, the classification network for forecasting the conversion risk treated each scan acquired at different time-points of the same eye as independent training samples. Although the current model showed a moderate amount of consistency between predictions from different visits for the same future time-point (eCI = 0.73), an alternate approach for the supervised downstream task training may be explored in the future to explicitly enforce this consistency constraint. Finally, although Morph-SSL was primarily developed to pre-train the Encoder in an unsupervised manner, the learned Encoder-Decoder network can additionally smoothly interpolate between the scans from two visits. This offers promising future research directions for using the interpolated scans as a data augmentation or to visualize the expected future morphological changes if a Recurrent Neural Network could be trained to predict the feature representations of future visits. Adapting Morph-SSL to other prognostic tasks in the medical domain, such as forecasting cancer progression from ultrasound images or predicting the future onset of dementia from MRI scans, offers important directions for future work.

APPENDIX

See Tables VI–X and Fig. 7.

REFERENCES

[1] W. L. Wong et al., "Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: A systematic review and meta-analysis," *Lancet Global Health*, vol. 2, no. 2, pp. e106–e116, Feb. 2014.

[2] J. A. Hallak, L. de Sisternes, A. Osborne, B. Yaspan, D. L. Rubin, and T. Leng, "Imaging, genetic, and demographic factors associated with conversion to neovascular age-related macular degeneration: Secondary analysis of a randomized clinical trial," *JAMA Ophthalmol.*, vol. 137, no. 7, pp. 738–744, 2019.

[3] Z. Zhou, V. Sodha, J. Pang, M. B. Gotway, and J. Liang, "Models genesis," *Med. Image Anal.*, vol. 67, Jan. 2021, Art. no. 101840.

[4] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised learning for medical image analysis using image context restoration," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101539.

[5] S. Azizi et al., "Big self-supervised models advance medical image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3478–3488.

[6] O. Ciga, T. Xu, and A. L. Martel, "Self supervised contrastive learning for digital histopathology," *Mach. Learn. With Appl.*, vol. 7, Mar. 2022, Art. no. 100198.

[7] D. Zeng et al., "Positional contrastive learning for volumetric medical image segmentation," in *Proc. MICCAI*, 2021, pp. 221–230.

[8] Y. Chen et al., "USCL: Pretraining deep ultrasound image diagnosis model through video contrastive representation learning," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 627–637.

[9] A. Bardes, J. Ponce, and Y. LeCun, "VICReg: Variance-invariance-covariance regularization for self-supervised learning," in *Proc. ICLR*, 2022.

[10] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 12310–12320.

[11] J.-B. Grill et al., "Bootstrap your own latent—A new approach to self-supervised learning," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 21271–21284.

[12] R. Holland et al., "Metadata-enhanced contrastive learning from retinal optical coherence tomography images," 2022, *arXiv:2208.02529*.

[13] T. Emre et al., "TINC: Temporally informed non-contrastive learning for disease progression modeling in retinal OCT volumes," in *Proc. MICCAI*, 2022, pp. 625–634.

[14] A. Rivail et al., "Modeling disease progression in retinal OCTs with longitudinal self-supervised learning," in *Proc. PRIME, MICCAI Workshop*, 2019, pp. 44–52.

[15] M. D. Davis, "The age-related eye disease study severity scale for age-related macular degeneration: AREDS report no. 17," *Arch. Ophthalmol.*, vol. 123, no. 11, pp. 1484–1498, 2005.

[16] X. Hu et al., "Morphological and functional characteristics at the onset of exudative conversion in age-related macular degeneration," *Retina*, vol. 40, no. 6, pp. 1070–1078, 2020.

[17] S. M. Waldstein, W.-D. Vogl, H. Bogunovic, A. Sadeghipour, S. Riedl, and U. Schmidt-Erfurth, "Characterization of drusen and hyperreflective foci as biomarkers for disease progression in age-related macular degeneration using artificial intelligence in optical coherence tomography," *JAMA Ophthalmol.*, vol. 138, no. 7, p. 740, Jul. 2020.

[18] Y. Wakatsuki et al., "Optical coherence tomography biomarkers for conversion to exudative neovascular age-related macular degeneration," *Amer. J. Ophthalmol.*, vol. 247, pp. 137–144, Mar. 2023.

[19] J. Sutton et al., "Developing and validating a multivariable prediction model which predicts progression of intermediate to late age-related macular degeneration—The PINNACLE trial protocol," *Eye*, vol. 37, pp. 1275–1283, May 2022.

[20] W.-D. Vogl, H. Bogunović, S. M. Waldstein, S. Riedl, and U. Schmidt-Erfurth, "Spatio-temporal alterations in retinal and choroidal layers in the progression of age-related macular degeneration (AMD) in optical coherence tomography," *Sci. Rep.*, vol. 11, no. 1, pp. 1–11, Mar. 2021.

[21] P. M. Burlina, N. Joshi, K. D. Pacheco, D. E. Freund, J. Kong, and N. M. Bressler, "Use of deep learning for detailed severity characterization and estimation of 5-year risk among patients with age-related macular degeneration," *JAMA Ophthalmol.*, vol. 136, no. 12, pp. 1359–1366, 2018.

[22] A. Bhuiyan, T. Y. Wong, D. S. W. Ting, A. Govindaiah, E. H. Souied, and R. T. Smith, "Artificial intelligence to stratify severity of age-related macular degeneration (AMD) and predict risk of progression to late AMD," *Transl. Vis. Sci. Technol.*, vol. 9, no. 2, p. 25, Apr. 2020.

[23] J. Bridge, S. Harding, and Y. Zheng, "Development and validation of a novel prognostic model for predicting AMD progression using longitudinal fundus images," *BMJ Open Ophthalmol.*, vol. 5, no. 1, Oct. 2020, Art. no. e000569.

[24] C. Yin, S. E. Moroi, and P. Zhang, "Predicting age-related macular degeneration progression with contrastive attention and time-aware LSTM," in *Proc. ACM KDD*, 2022, pp. 4402–4412.

[25] A. Ganjdanesh, J. Zhang, E. Y. Chew, Y. Ding, H. Huang, and W. Chen, "LONGL-Net: Temporal correlation structure guided deep learning model to predict longitudinal age-related macular degeneration severity," *PNAS Nexus*, vol. 1, no. 1, Mar. 2022, Art. no. pgab003.

[26] Q. Yan et al., "Deep-learning-based prediction of late age-related macular degeneration progression," *Nat. Mach. Intell.*, vol. 2, no. 2, pp. 141–150, 2020.

[27] I. Banerjee et al., "Prediction of age-related macular degeneration disease using a sequential deep learning approach on longitudinal SD-OCT imaging biomarkers," *Sci. Rep.*, vol. 10, no. 1, pp. 1–16, Sep. 2020.

[28] U. Schmidt-Erfurth et al., "Prediction of individual disease conversion in early AMD using artificial intelligence," *Investigative Opthalmol. Vis. Sci.*, vol. 59, no. 8, p. 3199, Jul. 2018.

[29] L. de Sisternes, N. Simon, R. Tibshirani, T. Leng, and D. L. Rubin, "Quantitative SD-OCT imaging biomarkers as indicators of age-related macular degeneration progression," *Investigative Opthalmol. Vis. Sci.*, vol. 55, no. 11, p. 7093, Nov. 2014.

[30] D. B. Russakoff, A. Lamin, J. D. Oakley, A. M. Dubis, and S. Sivaprasad, "Deep learning for prediction of AMD progression: A pilot study," *Investigative Opthalmol. Vis. Sci.*, vol. 60, no. 2, p. 712, Feb. 2019.

[31] J. Yim, "Predicting conversion to wet age-related macular degeneration using deep learning," *Nature Med.*, vol. 26, no. 6, pp. 892–899, 2020.

[32] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "VoxelMorph: A learning framework for deformable medical image registration," *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1788–1800, Aug. 2019.

[33] J. Zhang, "Inverse-consistent deep networks for unsupervised deformable image registration," 2018, *arXiv:1809.03443*.

[34] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. NeuRIPS*, 2015, pp. 2017–2025.

[35] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *Proc. NeurIPS*, 2016, pp. 658–666.

[36] Q. Yang et al., "Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1348–1357, Jun. 2018.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 630–645.

[38] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

[39] M. K. Garvin, M. D. Abramoff, X. Wu, S. R. Russell, T. L. Burns, and M. Sonka, "Automated 3-D intraretinal layer segmentation of macular spectral-domain optical coherence tomography images," *IEEE Trans. Med. Imag.*, vol. 28, no. 9, pp. 1436–1447, Sep. 2009.

[40] A. Johnson, L. Bulgarelli, and T. Pollard, "alistairewj/pyroc: pyroc v0.2.0," Zenodo, v0.2.0, Jul. 2022, doi: 10.5281/zenodo.6819206.

[41] D. P. Kingma and J. Ba, "ADAM: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[42] L. N. Smith, "Cyclical learning rates for training neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 464–472.

[43] S. Zhang, L. Yang, M. B. Mi, X. Zheng, and A. Yao, "Improving deep regression with ordinal entropy," in *Proc. ICLR*, 2023.

[44] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4700–4708.

[45] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.

[46] C. Feichtenhofer, "X3D: Expanding architectures for efficient video recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 203–213.

[47] H. Bogunovic et al., "Machine learning of the progression of intermediate age-related macular degeneration based on OCT imaging," *Investigative Opthalmol. Vis. Sci.*, vol. 58, no. 6, Jun. 2017, Art. no. BIO141.

[48] T. Schlegl et al., "Fully automated segmentation of hyperreflective foci in optical coherence tomography images," 2018, *arXiv:1805.03278*.

[49] A. Galdran, G. Carneiro, and M. A. G. Ballester, "Balanced-mixup for highly imbalanced medical image classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 12905. Cham, Switzerland: Springer, 2021, pp. 323–333.