Learn Single-horizon Disease Evolution for Predictive Generation of Post-therapeutic Neovascular Age-related Macular Degeneration

Yuhan Zhang^a, Kun Huang^a, Mingchao Li^a, Songtao Yuan^b, Qiang Chen^{a,*}

^aSchool of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210094, China

^bDepartment of Ophthalmology, The First Affiliated Hospital with Nanjing Medical University, Nanjing, 210094, China

Abstract

Background and Objective: Most of the existing disease prediction methods in the field of medical image processing fall into two classes, namely image-to-category predictions and image-to-parameter predictions. Few works have focused on image-to-image predictions. Different from multi-horizon predictions in other fields, ophthalmologists prefer to show more confidence in single-horizon predictions due to the low tolerance of predictive risk.

Methods: We propose a single-horizon disease evolution network (SHENet) to predictively generate post-therapeutic SD-OCT images by inputting pretherapeutic SD-OCT images with neovascular age-related macular degeneration (nAMD). In SHENet, a feature encoder converts the input SD-OCT images to deep features, then a graph evolution module predicts the process of disease evolution in high-dimensional latent space and outputs the predicted deep features, and lastly, feature decoder recovers the predicted deep features to SD-OCT images. We further propose an evolution reinforcement module to ensure the effectiveness of disease evolution learning and obtain realistic SD-OCT images by adversarial training.

Results: SHENet is validated on 383 SD-OCT cubes of 22 nAMD patients based on three well-designed schemes (P-0, P-1 and P-M) based on the

^{*}Qiang Chen is the corresponding author of this work.

Email addresses: zhangyuhan@njust.edu.cn (Yuhan Zhang),

huangkun@njust.edu.cn (Kun Huang), chaosli@njust.edu.cn (Mingchao Li),

yuansongtao@vip.sina.com (Songtao Yuan), chen2qiang@njust.edu.cn (Qiang Chen)

quantitative and qualitative evaluations. Three metrics (PSNR, SSIM, 1-LPIPS) are used here for quantitative evaluations. Compared with other generative methods, the generative SD-OCT images of SHENet have the highest image quality (P-0: 23.659, P-1: 23.875, P-M: 24.198) by PSNR. Besides, SHENet achieves the best structure protection (P-0: 0.326, P-1: 0.337, P-M: 0.349) by SSIM and content prediction (P-0: 0.609, P-1: 0.626, P-M: 0.642) by 1-LPIPS. Qualitative evaluations also demonstrate that SHENet has a better visual effect than other methods.

Conclusions: SHENet can generate post-therapeutic SD-OCT images with both high prediction performance and good image quality, which has great potential to help ophthalmologists forecast the therapeutic effect of nAMD.

Keywords:

nAMD, Generative Adversarial Network, Graph Neural Network, Predictive Generation

1. Introduction

1.1. Application Background

Neovascular age-related macular degeneration (nAMD) is a main subtype of AMD. As the intravitreal vascular endothelial growth factor (VEGF) level elevates, choroidal neovasculars invade the avascular outer retina and severely damage photoreceptors, resulting in rapid vision loss [1]. At present, anti-VEGF injection is considered the preferred nAMD therapy [2]. Although ophthalmologists always give anti-VEGF injections after nAMD diagnosis, nAMD does not always respond satisfactorily to treatment. Besides, due to the lack of uniform guidelines, it is difficult for ophthalmologists to predict the short-term therapeutic response after anti-VEGF injection according to their subjective experiences [3]. This results in huge economic pressures and waste of resources [4]. Therefore, based on the known pre-therapeutic status of nAMD at time point t_1 , predicting the post-therapeutic status of nAMD at time point $t_2 = t_1 + \Delta t$ can effectively forecast the efficacy of anti-VEGF injection for each patient. This can promote better clinical decision making.

1.2. SD-OCT Imaging Brief

Spectral-domain optical coherence tomography (SD-OCT) is a noninvasive, depth-resolved, high-resolution, and volumetric imaging technique. SD-OCT



Figure 1: Clinical Application of Our Proposed Approach. (a) shows the workflow of neovascular age-related macular degeneration (nAMD) therapy. Ophthalmologists diagnose nAMD by SD-OCT imaging at time point t_1 and then give anti-VEGF injection for treatment. Generally one month later ($t_2 = t_1 + \Delta t$), ophthalmologists can estimate the therapy response by SD-OCT imaging again. Red line and green line indicate two retinal layer surfaces, which will be used for data pre-processing (flatten and alignment). (b) shows the clinical application of our proposed approach. Based on pre-processed SD-OCT images at time point t_1 , our proposed approach predictively generates SD-OCT images at time point t_2 to help ophthalmologists forecast the therapy response of anti-VEGF injection, and consequently make more reasonable treatment decisions.

has become a pivotal diagnostic tool to visualize and quantitatively evaluate retinal morphological changes [5], including the diagnosis and tracing of nAMD [6, 7]. Each SD-OCT imaging can produce 3D volumetric images, also known as a **cube**. As shown in Fig. 1(a), we take Cirrus SD-OCT device as the example, each SD-OCT cube contains $1024 \times 512 \times 128$ voxels with a corresponding trim size of $2mm \times 6mm \times 6mm$ on the retina. In a cube, each slice along the vertical direction, with the size of 1024×512 , is known as a **B-scan**. A complete SD-OCT cube contains 128 continuous B-scans in space.

1.3. Dilemma of Multi-horizon Predictions on Medical Images

Diseases-associated predictions are more restrictive than general predictions. In the field of pattern recognition, multi-horizon predictions have been widely applied for natural language predictions, action predictions, video predictions, traffic predictions and etc. Given $X_{t_1:t_N} = [\mathbf{X}_{t_1}, \mathbf{X}_{t_2}, \cdots, \mathbf{X}_{t_N}] \in \mathbb{R}$ as the historical N observations, each observation is obtained at the different time point and the time interval between any two adjacent time point is uniform. The actual future M observations are formally expressed as $X_{t_{N+1}:t_{N+M}} = [\mathbf{X}_{t_{N+1}}, \mathbf{X}_{t_{N+2}}, \cdots, \mathbf{X}_{t_{N+M}}] \in \mathbb{R}$. We expect multi-horizon predictions to learn a mapping function $\mathcal{F} : \mathbb{X}_{t_1:t_N} \to \widetilde{\mathbb{X}}_{t_{N+1}:t_{N+M}}$ to obtain the prediction result $\widetilde{\mathbb{X}}_{t_{N+1}:t_{N+M}}$ as close as $\mathbb{X}_{t_{N+1}:t_{N+M}}$, as shown in Fig. 2(a). However, when applying multi-horizon predictions on medical images, several actual challenges raise:

- Obtaining long-series observations from the same patient is intractable in practical clinical scenes, and it also is necessary to make effective predictions for a new patient with only one observation at the current time point.
- For serial medical data, given any two different time points $t_i, t_j \in (t_2, t_N]$, time intervals from adjacent time points may be different, namely $t_i t_{i-1} \not\equiv t_j t_{j-1}$.
- Most medical observations generate 3D data and it is expensive for GPUs to learn from serial 3D data.
- Therapeutic intervention dependent on medicine injection at a random time point is a key factor that cannot be ignored for diseases-associated predictions, and most medicine injection treatments only work for a short period of time.
- It is well-known that the prediction accuracy decreases over time and risk tolerance on medical predictions is lower than other predictions, so clinicians always pay more confidence on the short horizon than the longer horizon.

These difficulties make it difficult for multi-horizon predictions to be really applied to medical images, thus learning a single-horizon prediction for medical images is more practical and realistic. Given one historical observation $\mathbf{X}_{t_i} \in \mathbb{R}$ at time point t_i , single-horizon prediction learns a mapping function $\mathcal{F} : \mathbf{X}_{t_i} \to \widetilde{\mathbf{X}}_{t_{i+1}}$ to obtain the prediction result at time point t_{i+1} , as shown in Fig. 2(b).

1.4. Target of Our Work

Nowadays, most existing disease prediction methods in the field of medical image processing fall into two classes, namely classification-based image-tocategory (I2C) predictions, and regression-based image-to-parameter (I2P) predictions. Few works have focused on generation-based image-to-image (I2I)



Figure 2: **Prediction Visualization.** (a) is multi-horizon prediction, (b) is single-horizon prediction.

predictions, even though generative adversarial networks (GANs) have been widely used for modality transformation between medical images from different imaging devices. The aim of post-therapeutic prediction is to generate an image that explains how the anatomical appearance changes after treatment. We consider that the predictive post-therapeutic SD-OCT images would enable a better understanding of nAMD and clinical decision-making by presenting a visual post-therapeutic status of nAMD. According to the above, the target of our work is defined as:

Given a serial SD-OCT cubes with fixed time interval Δt , writing as $\mathbb{X} = [\mathbf{X}_{t_1}, \mathbf{X}_{t_2}, \cdots, \mathbf{X}_{t_{N+M}}] \in \mathbb{R}$, and any time point t_i can be represented as $t_i = t_1 + (i-1)\Delta t$, $i \in [1, N+M]$. Anti-VEGF injection is given at each time point. For any SD-OCT cube \mathbf{X}_{t_i} , we hope to learn a mapping function $\mathcal{F} : \mathbf{X}_{t_i} \to \widetilde{\mathbf{X}}_{t_{i+1}}$ to predictively generate $\widetilde{\mathbf{X}}_{t_{i+1}}$ as close as $\mathbf{X}_{t_{i+1}}$.

However, learning the mapping function \mathcal{F} directly based on a 3D SD-OCT cube is difficult and cost-consuming. Thus, for the SD-OCT cube $\mathbf{X}_{t_i} = [\mathbf{x}_{t_i}^1, \mathbf{x}_{t_i}^2, \cdots, \mathbf{x}_{t_i}^{128}]$, where $\mathbf{x}_{t_i}^j$ is *j*-th B-scan in \mathbf{X}_{t_i} , we learn the 2D mapping function $\mathcal{F}_{2D} : \mathbf{x}_{t_i}^j \to \widetilde{\mathbf{x}}_{t_{i+1}}^j$. \mathcal{F}_{2D} is carried out 128 times until the SD-OCT cube is predicted completely.

1.5. Contributions

In this paper, we present a Single-Horizon disease Evolution Network (SHENet) to predictively generate the post-therapeutic SD-OCT images based on pre-therapeutic SD-OCT images with nAMD. SHENet can help forecast the short-term response of anti-VEGF injection for individual nAMD patients. The main contributions in this paper are summarized as:

• We explore the possibility of predictively generating post-therapeutic

SD-OCT images based on pre-therapeutic SD-OCT images with nAMD, and further propose SHENet to solve this problem.

- Graph evolution module (GEM) is proposed to imprison the process of disease evolution in the high-dimensional latent space by graph representation learning.
- Evaluation reinforcement module (ERM) is proposed to reinforce the disease evolution process by combining an additional reconstruction generator and contrastive learning.
- We design targeted experimental schemes according to actual clinical realities, and the results demonstrate that SHENet has great potential to generate post-therapeutic SD-OCT images with both high prediction performance and good image quality.

2. Related Work

2.1. Predictions on AMD

In clinical scenarios, several prediction requirements have been raised around AMD by ophthalmologists. In present, there are no uniform guidelines to be helpful for making predictions and ophthalmologists rely only on their own rich clinical experiences. Thus, ophthalmologists warrant the need for objective outcomes of subjective predictions. AMD typically develops from an early to an advanced form and advanced AMD is difficult to be cured effectively. When AMD is in its early stage, ophthalmologists predict the risk of progression to advanced AMD within the future short term in order to adapt therapies, recommendations, and follow-up frequency [8, 9, 10, 11, 12] For advanced nAMD, ophthalmologists predict best-corrected visual acuity (BCVA) outcomes in patients receiving standard therapy [13, 14, 15]. For geographic atrophy (GA), as non-neovascular advanced AMD, there is a lack of effective treatments for retinal areas that have progressed to GA. But predicting the GA progression [16, 17, 18, 19, 20, 21] could allow for a better understanding of the pathogenesis and forewarn preventive treatment to normal retinal areas that are at high risk of developing GA in the future. Besides, several other predictions also reveal clinical needs. Bogunovic et al. [22] predicted low and high anti-VEGF injection requirements based on sets of SD-OCT images acquired during the initiation phase in nAMD. Liu et al. [23] and Lee et al. [24] exploringly generated individualized post-therapeutic

SD-OCT images that could predict the short-term response of anti-VEGF injection for nAMD based on pre-therapeutic images using Pix2Pix. Forshaw et al. [25] predicted the visual gain from cataract surgery when the main cause of vision loss is nAMD. Pham et al. [26] generated future fundus images for early age-related macular degeneration based on generative adversarial networks.

2.2. Generative Adversarial Networks (GANs)

Generative adversarial networks (GANs) [27] have become one of the widely leveraged techniques for generating images that look like the real thing. Traditional L1 or L2 supervision often results in blurred images [28], but GANs introduce an additional discriminator to play a min-max game with the generator that enforces the generator to output more realistic images. The discriminator is mainly used to estimate the divergence difference between generated fake images and real images. Different adversarial losses estimate different divergences between Wasserstein divergence distribution [29] and f-divergence family distribution [30]. If target domains own multiple distributions, conditional GAN (cGAN) [31] uses conditional labels to guide the generator to produce the images by fitting the specified distribution. For example, Yoo et al. [32] proposed a postoperative appearance prediction model for orbital decompression surgery for thyroid ophthalmopathy using a conditional GAN. GANs are also used for image-to-image (I2I) translation, in which pixel-level losses or feature-level losses are embedded to ensure the quality and stability of the generated image [26, 33, 34]. Besides, literatures [35, 36] explored the impact of discriminator architecture on GANs.

3. Methods

This study was approved by the ethic committee of the First Affiliated Hospital with Nanjing Medical University.

3.1. Pre-processing: Voxel-wise Serial Alignment

The random deviation of rotation angle and displacement among SD-OCT cubes at different time points caused by human operations cannot be learned, which further limits the I2I prediction. In other words, for two original SD-OCT cubes from the same patient captured at different time points, the B-scan with the same index in two cubes may not correspond to the same anatomical tissue. Thus, we perform voxel-wise serial alignment, including image flattening in vertical direction and image alignment in horizontal-axial directions, for all SD-OCT cubes before running SHENet. In this way, all SD-OCT cubes at different time points are aligned in the 3D space. The aligned SD-OCT cube can be seen in Fig. 1(b).

Image Flattening in Vertical Direction. We first obtain the locations of Bruch's membrance (BM) of all B-scans by layer segmentation approach [37], and then all B-scans are flattened based on BM. Finally, we crop the B-scan restricted to the region from 0.75mm above BM to 0.25mm below BM. In this processing, negligible vitreum regions and sclera regions are removed as much as possible. Image flattening also reduces the size of the input SD-OCT images to alleviate the memory pressure of hardware.

Image Alignment on Horizontal-Axial Directions. We first generate the 2D vessel fundus images of each SD-OCT cube by restricting the projection region between inner segments/outer segments (IS/OS) and BM. The vessel fundus images are aligned using a scale-invariant feature transform (SIFT) flow method to obtain the transformation matrixes. Lastly, the transformation matrixes are applied to horizontal-axial directions of SD-OCT cubes to obtain the aligned SD-OCT cubes.

3.2. Overview of SHENet

Pix2Pix [38] has been a great success to apply conditional GAN (cGAN) to supervised I2I translation. It regards input images as additional conditions to learn an I2I mapping, consequently producing specified output images. SHENet further extends Pix2Pix for our nAMD prediction task and its overview is illustrated in Fig. 3. In terms of model architecture, SHENet consists of:

(1) Prediction Generator \mathcal{G}^p is the core of SHENet, including a feature encoder, a graph evolution module (GEM), and a feature decoder, which can predictively generate post-therapeutic SD-OCT images by inputting pre-therapeutic SD-OCT images with nAMD.

(2) Reconstruction Generator \mathcal{G}^r is an auxiliary generator in the training process and will be removed in the model inference stage. \mathcal{G}^r removes GEM from \mathcal{G}^p and utilizes a reconstruction task to help \mathcal{G}^p imprison the process of disease evolution in the high-dimensional latent space, and further to distill the function of feature encoder and decoder.

(3) Evolution Reinforcement Module (ERM) reinforces the process of disease evolution by working with \mathcal{G}^r based on contrastive learning.



Figure 3: **Overview of SHENet in Training Stage.** SHENet consists of a prediction generator \mathcal{G}^p , a reconstruction generator \mathcal{G}^r , an evolution reinforcement module (ERM), a quality discriminator \mathcal{D}^q and a pair discriminator \mathcal{D}^p . A little piece of B-scans $\mathbf{X}_{t_1}^{[j;\Delta s]} \in \mathbf{X}_{t_1}$ as multi-channel inputs are sent to \mathcal{G}^p to predictively generate $\tilde{\mathbf{x}}_{t_2}^j$, meanwhile a little piece of B-scans $\mathbf{X}_{t_2}^{[j;\Delta s]} \in \mathbf{X}_{t_2}$ as multi-channel inputs are sent to \mathcal{G}^r to reconstruct $\overline{\mathbf{x}}_{t_2}^j$. ERM reinforces the disease evolution in the high-dimensional latent space. \mathcal{D}^q ensures the image quality of $\tilde{\mathbf{x}}_{t_2}^j$ and $\overline{\mathbf{x}}_{t_2}^j$ is realistic, and \mathcal{G}^p decides whether $\tilde{\mathbf{x}}_{t_2}^j$ and $\overline{\mathbf{x}}_{t_2}^j$ are paired with $\mathbf{x}_{t_1}^j$ in terms of pathological characterization that reflects therapy response. In the inference stage of SHENet, only prediction generator \mathcal{G}^p remains and other all components are removed.

(4) Quality Discriminator \mathcal{D}^q ensures that the predicted and reconstructed images look realistic.

(5) Pair Discriminator \mathcal{D}^p determines whether predicted and reconstructed images are paired with input images in terms of pathological characterization that reflects therapy response.

3.2.1. Motivation: Why learn the disease evolution in the highdimensional latent space?

Liu et al. [23] and Lee et al. [24] have preliminarily used Pix2Pix to generate post-therapeutic SD-OCT images and show the feasibility, but learning a pixel-to-pixel prediction on SD-OCT images seems not to be a delicate work. Severe speckle noise in SD-OCT images and the imbalanced proportion of foreground pixels (*i.e.* nAMD) relative to the background pixels (*i.e.* non-nAMD) degenerate the pixel-to-pixel prediction performance. Compared with Pix2Pix, SHENet imprisons the process of disease evolution in the high-dimensional latent space:

$$\begin{array}{l} Pix2Pix: \boldsymbol{x}_{t_1}^j \stackrel{Generator}{\longrightarrow} \widetilde{\boldsymbol{x}}_{t_2}^j \\ SHENet: \boldsymbol{x}_{t_1}^j \stackrel{Enc}{\longrightarrow} \boldsymbol{f}_{t_1}^j \stackrel{Pred}{\longrightarrow} \widetilde{\boldsymbol{f}}_{t_2}^j \stackrel{Dec}{\longrightarrow} \widetilde{\boldsymbol{x}}_{t_2}^j \end{array} \tag{1}$$

where $f_{t_1}^j$ is the encoding features of $x_{t_1}^j$ and $\tilde{f}_{t_2}^j$ is the predicted features of $x_{t_1}^j$. The high-dimensional features are more condensed and effective, because the invalid background and noise information are removed, and distinct disease information is retained. Therefore, we consider that learning the process of disease evolution after treatment in the high-dimensional latent space is more reasonable than performing a pixel-to-pixel prediction.

3.2.2. Multiple B-scans As Model Input.

Given an SD-OCT cube at time point t_1 as $\mathbf{X}_{t_1} = [\mathbf{x}_{t_1}^1, \mathbf{x}_{t_1}^2, \cdots, \mathbf{x}_{t_1}^{128}]$ and aligned SD-OCT cube at time point $t_2 = t_1 + \Delta t$ as $\mathbf{X}_{t_2} = [\mathbf{x}_{t_2}^1, \mathbf{x}_{t_2}^2, \cdots, \mathbf{x}_{t_2}^{128}]$. In general, we should train a 2D mapping \mathcal{F}_{2D} to predictively generate $\widetilde{\mathbf{x}}_{t_2}^j$ as close as real $\mathbf{x}_{t_2}^j$:

$$\mathcal{F}_{2D}: \boldsymbol{x}_{t_1}^j \to \widetilde{\boldsymbol{x}}_{t_2}^j \tag{2}$$

However, in SD-OCT images, $\boldsymbol{x}_{t_1}^j$ with similar pathological characterization may develop to $\boldsymbol{x}_{t_2}^j$ with different pathological characterization, resulting in difficult model convergence and random prediction results. For example, for patient-1, a healthy SD-OCT B-scan at time point t_1 evolves to an SD-OCT B-scan with nAMD at time point t_2 . However, for patient-2, a healthy SD-OCT B-scan at time point t_1 may remain its healthy status at time point t_2 . Thus, to speed up the model convergence and improve the model robustness, we choose to stack a little piece of B-scans of \mathbf{X}_{t_1} as $\mathbf{X}_{t_1}^{[j;\Delta s]} = [\boldsymbol{x}_{t_1}^{j-\Delta s}, \cdots, \boldsymbol{x}_{t_1}^{j}, \cdots, \boldsymbol{x}_{t_1}^{j+\Delta s}] \in \mathbf{X}_{t_1}$ as multi-channel inputs to predict $\tilde{\boldsymbol{x}}_{t_2}^j$:

$$\mathcal{F}_{2D}: \mathbf{X}_{t_1}^{[j;\Delta s]} \to \widetilde{\mathbf{x}}_{t_2}^j \tag{3}$$

 \mathcal{F}_{2D} slides on each B-scan of \mathbf{X}_{t_1} until $\widetilde{\mathbf{X}}_{t_2}$ is predicted completely.

3.3. Prediction Generator \mathcal{G}^p of SHENet

The prediction generator \mathcal{G}^p of SHENet consists of three modules: feature encoder (\mathcal{G}_E), graph evolution module (GEM), and feature decoder (\mathcal{G}_D). The pre-therapeutic SD-OCT images are mapped to high-dimensional latent space



Figure 4: **Details of Prediction Generator** \mathcal{G}^p . (a) is feature encoder \mathcal{G}_E consisting of several customized encoding blocks. (b) is channel attention block (CAB), where GAP indicates global average pooling. (c) is graph evolution module (GEM) that stacks three multi-head GAT modules. (d) is feature decoder \mathcal{G}_D consisting of several customized decoding blocks.

by \mathcal{G}_E , then GEM predicts the process of disease evolution after treatment, and finally, \mathcal{G}_D recovers the predicted features to post-therapeutic SD-OCT images.

Feature Encoder (\mathcal{G}_E). Fig. 4(a) illustrates the architecture of feature encoder \mathcal{G}_E that stacks several encoding blocks (EncBlock) and a mapping layer. Each EncBlock consists of two 3×3 convolutional layers with ReLU activation, one channel attention block (CAB) [39], and one max-pooling

layer for down-sampling. CAB (Fig. 4(b)) improves the quality of features by explicitly modeling the interdependencies between the channels of its convolutional features. Through CAB, informative features are selectively emphasized and less useful ones are suppressed.

Graph Evolution Module (GEM). Given the encoding features $f_{t_1}^j \in \mathbb{R}^{h' \times w' \times d'}$ by \mathcal{G}_E , where $P = h' \times w'$ is feature numbers and d' is feature dimension. We consider the evolution of each feature should be related to other features. Intuitively, a series of convolution layers or a linear regression is easier to implement to predict the change in the feature level. However, a series of convolution layers or a linear regression defaults that all other features contribute equally to the targeted feature. In fact, the contributions of all features are unequal due to their different semantic information. Graph neural networks can model the contributions of all features automatically, which is more reasonable than a series of convolution layers or a linear regression. Let each feature be the vertex of the graph, and we can represent $f_{t_1}^j$ as a fully-connected undirected graph $\mathcal{R} = (\mathcal{V}, \mathcal{E})$, where \mathcal{V} is vertex set and \mathcal{E} is edge set. Further given the adjacency matrix A, the diagonal degree matrix D and the identity matrix I, the relation of features can be extracted by the following formula:

$$\boldsymbol{H}^{l+1} = \sigma(\widetilde{\boldsymbol{A}}\boldsymbol{H}^{l}\boldsymbol{W}^{l}), \widetilde{\boldsymbol{A}} = \boldsymbol{D}^{-\frac{1}{2}}(\boldsymbol{A} + \boldsymbol{I})\boldsymbol{D}^{-\frac{1}{2}}$$
(4)

where $\mathbf{W}^{l} \in \mathbb{R}^{d_{in} \times d_{out}}$ is the learnable weight, and σ is the *Mish* function [40]. $\mathbf{H}^{l} \in \mathbb{R}^{P \times d_{in}}, \mathbf{H}^{l+1} \in \mathbb{R}^{P \times d_{out}}$ are the input features and the updated features at *l*-th layer. Intuitively, for *p*-th vertex \mathcal{V}_{p} , its all neighbors contribute unequally to the evolution of \mathcal{V}_{p} . To model this, we introduce the multihead GAT [41] to explicitly consider the importance of the neighbors. Take independent GAT as an example, we calculate the attentive score γ_{pq} of the vertex pair $(\mathbf{h}_{p}, \mathbf{h}_{q})$:

$$\gamma_{pq} = \frac{exp(LReLU(\boldsymbol{\alpha}^{T}[\boldsymbol{W}\boldsymbol{h}_{p}, \boldsymbol{W}\boldsymbol{h}_{q}]))}{\sum_{k \in \mathcal{N}_{p}} exp(LReLU(\boldsymbol{\alpha}^{T}[\boldsymbol{W}\boldsymbol{h}_{p}, \boldsymbol{W}\boldsymbol{h}_{k}]))}$$
(5)

where \mathcal{N}_p is the neighbors of *p*-th vertex in the graph, and $[\cdot]$ is a concatenation operation. $\mathbf{W} \in \mathbb{R}^{d_{out} \times d_{in}}$ indicates a linear mapping and $\boldsymbol{\alpha} \in \mathbb{R}^{2d_{out}}$ indicates a single-layer fully-connected layer. *LReLU* is the nonlinear activation. Then, we compute the average of multi-head GAT for the output of vertex \mathcal{V}_p :

$$\widetilde{\boldsymbol{h}}_{p} = \sigma\left(\frac{1}{G}\sum_{g=1}^{G}\sum_{\sum_{q\in\mathcal{N}_{i}}}\gamma_{pq}^{g}\boldsymbol{W}^{g}\boldsymbol{h}_{p}\right)$$
(6)

where G=5 is the number of heads. Similarly, we repeat the GAT computation on each vertex to obtain the complete output. Considering the powerful ability of graph neural networks for information inference, GEM is only composed of 3 multi-head GATs with the channel numbers 1024, 1024, 1024. The overview of GEM is shown in Fig. 4(c).

Feature Decoder (\mathcal{G}_D) . Fig. 4(d) illustrates the architecture of feature decoder \mathcal{G}_D that stacks several decoding blocks (DecBlock) and a mapping layer. Each DecBlock consists of one de-convolution layer with ReLU activation, and two convolutional layers with ReLU activation.

3.4. Evolution Reinforcement Module (ERM)

To reinforce the process of disease evolution, we introduce reconstruction generator $\mathcal{G}^{\mathbf{r}}$ that removes GEM from $\mathcal{G}^{\mathbf{p}}$ to reconstruct $\overline{\boldsymbol{x}}_{t_2}^j$ by inputting $\mathbf{X}_{t_2}^{[j;\Delta s]}$:

$$\begin{aligned}
\mathcal{G}^{p} : \mathbf{X}_{t_{1}}^{[j;\Delta s]} \xrightarrow{\mathcal{G}_{E}} \boldsymbol{f}_{t_{1}}^{j} \xrightarrow{GEM} \widetilde{\boldsymbol{f}}_{t_{2}}^{j} \xrightarrow{\mathcal{G}_{D}} \widetilde{\boldsymbol{x}}_{t_{2}}^{j} \\
\mathcal{G}^{r} : \mathbf{X}_{t_{2}}^{[j;\Delta s]} \xrightarrow{\mathcal{G}_{E}} \boldsymbol{f}_{t_{2}}^{j} \xrightarrow{\mathcal{G}_{D}} \overline{\boldsymbol{x}}_{t_{2}}^{j}
\end{aligned} \tag{7}$$

Prediction generator $\mathcal{G}^{\mathbf{p}}$ and reconstruction generator $\mathcal{G}^{\mathbf{r}}$ share feature encoder \mathcal{G}_{E} and feature decoder \mathcal{G}_{D} . We use the project head $h(\cdot)$ [42] to map $\boldsymbol{f}_{t_{1}}^{j}$, $\boldsymbol{\tilde{f}}_{t_{2}}^{j}$, $\boldsymbol{f}_{t_{2}}^{j}$ to $\boldsymbol{z}_{t_{1}}^{j}$, $\boldsymbol{\tilde{z}}_{t_{2}}^{j}$, $\boldsymbol{z}_{t_{2}}^{j}$ for evolution reinforcement learning:

$$\boldsymbol{z} = h(\boldsymbol{f}) = \boldsymbol{W}^{(2)}(ReLU(\boldsymbol{W}^{(1)} \cdot GAP(\boldsymbol{f})))$$
(8)

where $\boldsymbol{W}^{(1)}$, $\boldsymbol{W}^{(2)}$ are weight matrixes, $GAP(\cdot)$ indicates global average pooling. We hope predicted $\tilde{\boldsymbol{z}}_{t_2}^j$ and real $\boldsymbol{z}_{t_2}^j$ should be as similar as possible, while predicted $\tilde{\boldsymbol{z}}_{t_2}^j$ and real $\boldsymbol{z}_{t_1}^j$ should be different. Thus, we build the evolution reinforcement learning based on contrastive loss:

$$\mathcal{L}_{ERM}(\mathcal{G}^p, \mathcal{G}^r) = -log \frac{exp(Sim(\widetilde{\boldsymbol{z}}_{t_2}^j, \boldsymbol{z}_{t_2}^j)/\tau)}{exp(Sim(\widetilde{\boldsymbol{z}}_{t_2}^j, \boldsymbol{z}_{t_1}^j)/\tau)}$$
(9)

where $\tau = 1$ is the temperature factor, $Sim(\cdot)$ is the cosine similarity metric. ERM also further distills the function of the feature encoder and feature decoder.

3.5. Discriminators of SHENet

For the predictively generated $\widetilde{x}_{t_2}^j$ and the reconstructed $\overline{x}_{t_2}^j$, we firstly use a quality discriminator \mathcal{D}^q to ensure their image quality is realistic:

$$\mathcal{D}^{q}(\widetilde{\boldsymbol{x}}_{t_{2}}^{j}) \to T/F, \quad \mathcal{D}^{q}(\overline{\boldsymbol{x}}_{t_{2}}^{j}) \to T/F$$
 (10)

Furthermore, we consider that pathological manifestation of $\tilde{x}_{t_2}^j$ and $\bar{x}_{t_2}^j$ should show the response of $x_{t_1}^j$ after anti-VEGF injection. In other words, $\tilde{x}_{t_2}^j$ and $\bar{x}_{t_2}^j$ should be paired with $x_{t_1}^j$. Thus we also use a pair discriminator \mathcal{D}^p to make the decision:

$$\mathcal{D}^{p}(\widetilde{\boldsymbol{x}}_{t_{2}}^{j}, \boldsymbol{x}_{t_{1}}^{j}) \to T/F, \quad \mathcal{D}^{p}(\overline{\boldsymbol{x}}_{t_{2}}^{j}, \boldsymbol{x}_{t_{1}}^{j}) \to T/F$$
(11)

In SHENet, both \mathcal{D}^q and \mathcal{D}^p follow the popular PatchGAN [38].

3.6. Adversarial Training

In the training process, we use \mathcal{L}_{L1} measures the L1 distance between output images and ground truth:

$$\mathcal{L}_{L1}(\mathcal{G}^p) = \mathbb{E}[||\boldsymbol{x}_{t_2}^j - \mathcal{G}^p(\mathbf{X}_{t_1}^{[j;\Delta s]})||_1]$$

$$\mathcal{L}_{L1}(\mathcal{G}^r) = \mathbb{E}[||\boldsymbol{x}_{t_2}^j - \mathcal{G}^r(\mathbf{X}_{t_2}^{[j;\Delta s]})||_1]$$
(12)

A pair discriminator \mathcal{D}^p makes the decision whether $(\tilde{\boldsymbol{x}}_{t_2}^j, \boldsymbol{x}_{t_1}^j)$ and $(\bar{\boldsymbol{x}}_{t_2}^j, \boldsymbol{x}_{t_1}^j)$ are paired:

$$\mathcal{L}_{GANp}(\mathcal{G}^{p}, \mathcal{D}^{p}) = \mathbb{E}[log\mathcal{D}^{p}(\boldsymbol{x}_{t_{1}}^{j}, \boldsymbol{x}_{t_{2}}^{j})] + \mathbb{E}[log(1 - \mathcal{D}^{p}(\boldsymbol{x}_{t_{1}}^{j}, \mathcal{G}^{p}(\mathbf{X}_{t_{1}}^{[j;\Delta s]})))]$$

$$\mathcal{L}_{GANr}(\mathcal{G}^{r}, \mathcal{D}^{p}) = \mathbb{E}[log\mathcal{D}^{p}(\boldsymbol{x}_{t_{1}}^{j}, \boldsymbol{x}_{t_{2}}^{j})] + \mathbb{E}[log(1 - \mathcal{D}^{p}(\boldsymbol{x}_{t_{1}}^{j}, \mathcal{G}^{r}(\mathbf{X}_{t_{2}}^{[j;\Delta s]})))]$$
(13)

A quality discriminator \mathcal{D}^q ensures $\widetilde{x}_{t_2}^j$ and $\overline{x}_{t_2}^j$ are realistic:

$$\mathcal{L}_{GANq}(\mathcal{G}^{p}, \mathcal{D}^{q}) = \mathbb{E}[log\mathcal{D}^{q}(\boldsymbol{x}_{t_{2}}^{j})] + \mathbb{E}[log(1 - \mathcal{D}^{q}(\mathcal{G}^{p}(\mathbf{X}_{t_{1}}^{[j;\Delta s]})))]$$

$$\mathcal{L}_{GANq}(\mathcal{G}^{r}, \mathcal{D}^{q}) = \mathbb{E}[log\mathcal{D}^{q}(\boldsymbol{x}_{t_{2}}^{j})] + \mathbb{E}[log(1 - \mathcal{D}^{q}(\mathcal{G}^{r}(\mathbf{X}_{t_{2}}^{[j;\Delta s]})))]$$
(14)

We combine all losses together and the complete optimization can be represented as:

$$\mathcal{G}^{p} = \arg \min_{\mathcal{G}^{p}, \mathcal{G}^{r}} \max_{\mathcal{D}^{p}, \mathcal{D}^{q}} [\mathcal{L}_{GANp}(\mathcal{G}^{p}, \mathcal{D}^{p}) + \mathcal{L}_{GANp}(\mathcal{G}^{r}, \mathcal{D}^{p}) + \mathcal{L}_{GANq}(\mathcal{G}^{p}, \mathcal{D}^{q}) + \mathcal{L}_{GANq}(\mathcal{G}^{r}, \mathcal{D}^{q}) + \lambda (\mathcal{L}_{L1}(\mathcal{G}^{p}) + \mathcal{L}_{L1}(\mathcal{G}^{r})) + \mu \mathcal{L}_{ERM}(\mathcal{G}^{p}, \mathcal{G}^{r})]$$
(15)

4. Results

4.1. Data Acquisition

According to clinical research, the therapeutic effect of anti-VEGF injection for nAMD lasts about one month, so nAMD patients treated should be followed up one month later to evaluate the therapeutic response of nAMD. Then, ophthalmologists can make the decision on whether further treatment is needed. However, to the best of our knowledge, there are no large-scale, public, and annotated SD-OCT datasets that can be acquired for our model validation due to the differences in imaging protocols, privacy problems, and lack of medical integration. This is also a common problem in the field of medical image processing.

The experimental data includes 46208 paired SD-OCT images obtained from 383 SD-OCT cubes of 22 nAMD patients. Only one eye per nAMD patient is included in the dataset, namely 22 eyes are included. Each patient contains about 17 serial SD-OCT cubes captured at different time points. During the SD-OCT imaging performed once a month, ophthalmologists gave anti-VEGF injections for nAMD treatment. The time interval between any adjacent time points is about one month, that is to say, the predictive single horizon is fixed to one month. These SD-OCT cubes were captured by Cirrus SD-OCT device *Carl Zeiss Meditec, Inc., Dublin, CA*. The size of an SD-OCT cube is $1024 \times 512 \times 128$. Detailed attribute descriptions are presented in Table 1.

Attributes	Values
Cube Number	383
Patient Number	22
Time Range	Jan 2012-Dec 2015
Time Interval	$1 \mathrm{month}$
Patient Age	Avg: 70 years (range: 45-84 years)
Patient Gender	17 males, 5 females
Type	\mathbf{nAMD}
Volume	Avg: $0.8024 \text{ mm}^3 (0.0133 \text{-} 7.1147 \text{ mm}^3)$
SD-OCT Device	Carl Zeiss Meditec, Inc., Dublin, CA
Cube Size	$1024{\times}512{\times}128$
Trim Size	$2\text{mm} \times 6\text{mm} \times 6\text{mm}$
Injection	$\operatorname{anti-VEGF}$

Table 1: Detailed attribute descriptions of materials.

4.2. Evaluation & Comparison

For the predictively generated post-therapeutic SD-OCT images, qualitative evaluation using human subjective judgment is the most straightforward and effective way to verify the prediction performance of SHENet. Besides, we also use three metrics, peak signal-to-noise ratio (PSNR), structural similarity (SSIM) and learned perceptual image patch similarity (LPIPS), to evaluate the prediction performance quantitatively. These three metrics have been previously used for the evaluation of image generation [43]. PSNR measures the image quality and higher PSNR means less distortion. SSIM measures the structure similarity and higher SSIM means higher structure similarity. LPIPS measures the similarity of deep visual features and lower LPIPS means higher feature similarity. To make LPIPS keep the consistency with PSNR and SSIM, namely higher value indicates better performance, here we use 1-LPIPLS to replace LPIPS.

To highlight the advantages of SHENet in terms of I2I prediction performance, we choose four GANs-based methods for competitive comparison, namely Pix2Pix [38], Pix2PixHD [44], Fundus2Angio [45], Att2Angio [46]. Pix2Pix and Pix2PixHD have become popular baselines and are widely used for image generation. Fundus2Angio and Att2Angio are the latest GANsbased methods for modality transformation in the field of medical image processing.



Figure 5: **Experimental Designs** on P-0 evaluation, P-1 evaluation and P-M evaluation. Each row denotes the serial SD-OCT observations of a nAMD patient and the time interval is one month.

4.3. Experimental Designs

In order to conform to real clinical application scenes, we design our experiments from three sub-evaluations: **P-0 Evaluation.** For the new nAMD patients that only have one SD-OCT imaging at time point t_1 , we want to predict post-therapeutic SD-OCT images at time point t_2 . Thus, the SD-OCT images from all other patients are used for training SHENet. We use five-fold cross-validation until all patients are tested, as shown in Fig. 5(a).

P-1 Evaluation. For the nAMD patients that have two SD-OCT imaging at time points t_1 and t_2 , we would like to predict post-therapeutic SD-OCT images at time point t_3 . Thus, we transfer the model parameters from P-0 evaluation and fine-tune SHENet only using $(\mathbf{X}_{t_1}, \mathbf{X}_{t_2})$. In the model inference stage, we take \mathbf{X}_{t_2} as model input to predictively generate \mathbf{X}_{t_3} . We use five-fold cross-validation until all patients are tested, as shown in Fig. 5(b).

P-M Evaluation. For the nAMD patients that have owned multiple regular SD-OCT imaging, we would like to predict consequent post-therapeutic SD-OCT images. Thus, we remain the last two SD-OCT cubes of all patients for model validation and the remaining SD-OCT cubes are used for training model, as shown in Fig. 5(c).

4.4. Implementation Details

The experiment is constructed in a hardware condition with Intel Xeon CPU, one GeForce RTX 3090 GPU and 128 GB RAM, and a software condition with Python3.5 and Pytorch.

For the input multiple B-scans, we choose $\Delta s = 3$ and zero-padding operation is used if $j - \Delta s < 0$ or $j + \Delta s > 128$, thus the model input is a three-channel SD-OCT image. The number of encoding blocks and decoding blocks is 5. The output dimensions of 5 encoding blocks are {128, 256, 512, 1024, 2048} respectively. The output feature size of the feature encoder is $16 \times 16 \times 2048$, which means the vertex number of a fully-connected graph in GEM is 16×16 . In the loss function, $\lambda = 100$ and $\mu = 10$ balance the weights among GAN losses, L1 losses and contrastive loss. Flip and rotation operations are used here for data augmentation.

Adam optimizer with initial learning rate of 0.0001 and weight decay of 0.1 is chosen for model optimization. The batch-size is set to 2. SHENet was trained for 100 epochs and the training was properly completed.

4.5. Qualitative Evaluation.

We qualitatively compare our SHENet with competing methods by visualizing the examples from two different nAMD patients based on P-0, P-1 and



Figure 6: Qualitative Comparison by visualizing the examples from two different nAMD patients. We respectively input a little piece of B-scans $\mathbf{X}_{t_1}^{[j;\Delta s]}$ to Pix2Pix [38], Pix2PixHD [44], Fundus2Angio [45], Att2Angio [46] and the proposed SHENet to predictively generate $\tilde{\boldsymbol{x}}_{t_2}^j$ based on three experimental designs (P-0, P-1 and P-M), and further evaluate them with ground truth $\boldsymbol{x}_{t_2}^j$.

P-M evaluations, as shown in Fig. 6. In terms of image quality, benefiting from adversarial training, all methods can produce clear and unblurred SD-OCT images. Besides, SHENet can stably maintain the structural integrity of the retina to achieve the better visual effects and actually predict the status of nAMD one month later after anti-VEGF injections.

M	P-0				P-1		P-M		
Methods	PSNR	SSIM	1-LPIPS	PSNR	SSIM	1-LPIPS	PSNR	SSIM	1-LPIPS
Pix2Pix [38]	21.114	0.273	0.552	21.272	0.276	0.559	21.481	0.284	0.568
Pix2PixHD [44]	21.415	0.286	0.581	21.525	0.290	0.585	21.764	0.299	0.591
Fundus2Angio [45]	21.333	0.288	0.569	21.471	0.293	0.568	21.679	0.301	0.576
Att2Angio [46]	21.466	0.294	0.589	21.716	0.311	0.591	22.149	0.321	0.607
SHENet (Ours)	23.659	0.326	0.609	23.875	0.337	0.626	24.198	0.349	0.642

Table 2: Quantitative Comparison with Other Competing Methods, based on three experimental designs.

4.6. Quantitative Evaluation.

We quantitatively evaluate the prediction performance based on P-0, P-1 and P-M evaluations using PSNR, SSIM and 1-LPIPS metrics, as shown in Table 2. For P-0 and P-1 evaluations, we calculate the average of five-fold crossvalidation as the final results. Overall, SHENet obtains consistently superior results than competing methods on three experimental designs. Among three experimental designs, SHENet achieves the best prediction performance on P-M evaluation (PSNR: 24.198, SSIM: 0.349, 1-LPIPS: 0.642), followed by P-1 evaluation (PSNR: 23.875, SSIM: 0.337, 1-LPIPS: 0.626), and P-0 evaluation (PSNR: 23.659, SSIM: 0.326, 1-LPIPS: 0.609) has the worst results, which shows the consistency with qualitative evaluation in Fig. 6. That is because the difference among patients is significantly greater than that among serial SD-OCT observations from the same patient. We consider that the gap will narrow by further collecting more samples from more patients.

4.7. Ablation Analysis

Evaluation of Hyper-parameter Values. We analyze the influence of hyper-parameters λ and μ in loss function with the experiment. We first fix μ to be 1 and vary λ from 10 to 190 with interval 10. As shown in Fig. 7(a), 1-LPIPS metrics of three evaluations are consistently increasing when rising the of λ from 10 to 100 and consistently decreasing when further rising λ . We then fix the value of λ to 100 and vary μ from 1 to 19 with interval 1. As shown in Fig. 7(b), the performance of SHENet achieves the highest values when $\mu = 10$.

Evaluation of Model Inputs. In terms of model input, we investigate the difference between single B-scan and multiple B-scans, and the quantitative



Figure 7: Hyper-parameter Tuning. Disease evolution prediction 1-LPIPS values with different hyper-parameters λ and μ based on three experimental designs.

Table 3: Results Comparison with Different Inputs (single B-scan and multiple B-scans), based on three experimental designs.

	Model Inputs	PSNR	SSIM	1-LPIPS
P-0	Single B-scan	17.542	0.242	0.513
	Multiple B-scans	23.659	0.326	0.609
P-1	Single B-scan	17.567	0.244	0.515
	Multiple B-scans	23.875	0.337	0.626
P-M	Single B-scan	17.585	0.250	0.521
	Multiple B-scans	24.198	0.349	0.642

comparisons are recorded in Table 3. We find that the prediction results using multiple B-scans as model input demonstrate significant improvements to those using a single B-scan as model input on three experimental designs.

Table 4: Results Comparison by Stacking GEM, \mathcal{G}^r and ERM One by One, based on three experimental designs.

GEM \mathcal{G}^r	Qr	ERM	P-0				P-1		P-M		
	G.		PSNR	SSIM	1-LPIPS	PSNR	SSIM	1-LPIPS	PSNR	SSIM	1-LPIPS
×	×	×	21.188	0.276	0.559	21.277	0.281	0.563	21.485	0.288	0.574
\checkmark	\times	\times	21.842	0.285	0.571	21.943	0.289	0.577	22.197	0.295	0.588
\checkmark		×	22.648	0.292	0.588	22.791	0.298	0.594	23.016	0.309	0.611
		\checkmark	23.659	0.326	0.609	23.875	0.337	0.626	24.198	0.349	0.642

Evaluation of Model Architecture. In the training process, we build our single-horizon disease evolution in the high-dimensional latent space based on the cooperation of GEM, \mathcal{G}^r and ERM. We investigate the impact of three components by stacking them one by one. As shown in Table 4, SHENet achieves better results than others when considering these three components jointly. This evidences that SHENet effectively imprisons the process of disease evolution in GEM and further reinforces the process by \mathcal{G}^r +ERM. Therefore, adopting all three can improve prediction performance.

Evaluation of Discriminator. To verify the relevance of two discriminators $(\mathcal{D}^q, \mathcal{D}^p)$, we separately analyze the prediction results of SHENet when one of them is reserved. When only using one of the two, quantitative metrics in Table 5 show performance degradation. Besides, we also observe that single \mathcal{D}^p performs better than single \mathcal{D}^q , as \mathcal{D}^p can also control the image quality to a certain extent. Thus, it shows the necessity to use two independent discriminators to respectively control the image quality and the pathological characterization.

0										
Da	\mathcal{D}^p	P-0				P-1		P-M		
D^{I}		PSNR	SSIM	1-LPIPS	PSNR	SSIM	1-LPIPS	PSNR	SSIM	1-LPIPS
	×	23.574	0.320	0.602	23.783	0.325	0.618	24.088	0.337	0.633
×		23.613	0.322	0.607	23.829	0.331	0.622	24.165	0.346	0.639
		23.659	0.326	0.609	23.875	0.337	0.626	24.198	0.349	0.642

Table 5: Results Comparison using \mathcal{D}^p , \mathcal{D}^q or the both, based on three experimental designs.

5. Discussion

In this paper, according to the actual clinical requirement, we explore the possibility of predictively generating post-therapeutic SD-OCT images based on pre-therapeutic SD-OCT images with nAMD, and propose a singlehorizon disease evolution network (SHENet) to solve it. SHENet learns the process of disease evolution in the high-dimensional latent space, rather than performing pixel-to-pixel prediction. This has the advantage of eliminating the influence of speckle noise and redundant background context on the prediction. Considering several inherent characteristics of medical images different from other modal data, we choose single-horizon prediction rather than multi-horizon prediction to simplify the problem. In other words, we only predictively generate post-therapeutic SD-OCT images with a one-month time interval, and would not continue to predict the nAMD status of the third month.

From Fig. 6, we can observe structural damage of the retina from competing methods, but SHENet can stably maintain the structural integrity of the retina to achieve a better visual effect. This benefits from two improvements to the proposed model: 1) SHENet uses two discriminators to respectively manage image quality and pathological characterization, but other competing methods only use one discriminator; 2) SHENet imprisons the process of disease evolution in the high-dimensional latent space, which makes the feature encoder and feature decoder concentrate on the restoration of image details.

Furthermore, we should pay more attention to the correctness of predictive generation of nAMD, namely whether SHENet can actually predict the status of nAMD one month later after anti-VEGF injection. Generally speaking, reduction of nAMD volume is the best indicator of therapeutic response, but texture change and improvement of additional diseases also need to be taken into account. In example-1 in Fig. 6, by comparing the pre-therapeutic SD-OCT image with the post-therapeutic ground truth, nAMD is significantly reduced in size, which shows an excellent therapeutic response. In example-2 in Fig. 6, although the nAMD volume does not change significantly, the thickness of the retina becomes thinner, which still indicates an effective treatment as additional diseases are improved. By qualitative comparisons, the predicted visual nAMD of SHENet is closer to ground truth than that of competing methods, demonstrating that SHENet owns a more powerful ability to learn the disease evolution. We also visualize the latent space of Pix2Pix and our SHENet from sequential observations for further comparison. As observed in Fig. 8, the image features generated by our SHENet are more high-density than those generated by Pix2Pix. That is because our designed GEM and ERM can filter out disease-unrelated information effectively. Highquality image features can provide a good basis for disease evolution learning in the latent space.

From Tables 2-5, we also observe lower PSNR & SSIM metrics than 1-LPIPS metric, and we consider this is caused by inherent severe speckle noise of SD-OCT images. Although denoising techniques can reduce the impact of speckle noise, the image quality would also become more blurry and the image details would be lost. Considering that the feature encoder owns the denoising ability, we do not apply any denoising technique during



Figure 8: Visualization of latent space of Pix2Pix method and our SHENet from sequential observations.

the pre-processing, avoiding an overly prediction model. In SHENet, to make the predicted images look realistic, the feature decoder adds random speckle noise to the generated SD-OCT images. PSNR & SSIM metrics are evaluated based on the whole SD-OCT images, thus speckle noise can result in low quantitative values, but 1-LPIPS metric is evaluated based on the deep visual features that show higher values than PSNR & SSIM.

To explore the influence of the number of historical observations on predictive results, we conducted P-0 evaluation, P-1 evaluation and P-M evaluation in our experiments. These three evaluations used zero, one and multiple historical observations respectively from the testing patients for model training. From Table 2, we can find that all methods achieve the best results on P-M evaluation and have the worst results on P-0 evaluation, demonstrating that more historical observations of the same patients can improve their own prediction performance. Although our dataset contains only 22 nAMD patients, each patient contains about 17 sequential SD-OCT cubes captured at different time points with regular medicine injections. A total of 46208 SD-OCT image pairs are used for our model validation. Besides. common data augmentation operations (rotation and horizontal flipping) are used. Therefore, our method does not occur the overfitting phenomenon, which could be observed from the results of P-0 evaluation. In the future, we will validate our method on other medical images, such as fundus photos for AMD and OCT for DME.

Fig. 9 shows more cases of predicted post-therapeutic SD-OCT images from P-0 evaluation, where green dotted box denotes the best results and crimson dotted box denotes the worst results. Overall, SHENet has the



Figure 9: Cases of predicted post-therapeutic SD-OCT images from P-0 evaluation, where (1-4) show the best results and (5-8) show the worst results.

ability to predict the change trend of nAMD status with high image quality, but several flaws still appear. First, SHENet fails to predict the dramatic texture change of nAMD precisely, as shown in Fig. 9(5). Second, due to individual differences, SHENet is hard to model the personalized change rate of nAMD after medicine injection. For example, SHENet overestimates the effect of medicine injection in Figs. 9(6-7) and underestimates the effect of medicine injection in Fig. 9(8). We consider the main reason is the limited number of patients, leading to SHENet cannot learn comprehensive information due to the complexity of nAMD. We believe that SHENet can be improved significantly after learning from more nAMD patients, and we will also continually validate SHENet by collecting more data in the future.

Although SHENet can produce high-quality SD-OCT images to visually reflect the status of nAMD one month later after anti-VEGF injection, it also has limitations. First, SD-OCT cubes at all time points need to be aligned by manual or automated alignment methods, since SHENet cannot learn the random deviations caused by man-made operations. Second, the time interval between the model input and ground truth must be consistent in the training process. Thus, SHENet can only predict the nAMD status after a single horizon and is incapable to predict longer horizons. Third, each time point for model training must ensure the same treatment intervention. SHENet can also be extended to train on the serial SD-OCT cubes without treatment intervention for predicting the single-horizon progression of nAMD.

6. Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (62172223, 61671242), and the Fundamental Research Funds for the Central Universities (30921013105).

References

- A. Foss, T. Rotsos, T. Empeslidis, V. Chong, Development of macular atrophy in patients with wet age-related macular degeneration receiving anti-vegf treatment, Ophthalmologica 245 (3) (2022) 204–217.
- [2] R. Tadayoni, L. Sararols, G. Weissgerber, R. Verma, A. Clemens, F. G. Holz, Brolucizumab: a newly developed anti-vegf molecule for the treatment of neovascular age-related macular degeneration, Ophthalmologica 244 (2) (2021) 93–101.
- [3] P. S. Mettu, M. J. Allingham, S. W. Cousins, Incomplete response to anti-vegf therapy in neovascular amd: Exploring disease mechanisms and therapeutic opportunities, Progress in Retinal and Eye Research 82 (2021) 100906.
- [4] M. G. Maguire, D. F. Martin, G.-s. Ying, G. J. Jaffe, E. Daniel, J. E. Grunwald, C. A. Toth, F. L. Ferris III, S. L. Fine, C. of Age-related Macular Degeneration Treatments Trials (CATT) Research Group, et al., Five-year outcomes with anti-vascular endothelial growth factor treatment of neovascular age-related macular degeneration: the comparison of age-related macular degeneration treatments trials, Ophthalmology 123 (8) (2016) 1751–1761.

- [5] G. Lan, J. Xu, Z. Hu, Y. Huang, Y. Wei, X. Yuan, H. Liu, J. Qin, Y. Wang, Q. Shi, et al., Design of 1300 nm spectral domain optical coherence tomography angiography system for iris microvascular imaging, Journal of Physics D: Applied Physics 54 (26) (2021) 264002.
- [6] M. Gharbiya, R. Giustolisi, J. Marchiori, A. Bruscolini, F. Mallone, V. Fameli, M. Nebbioso, S. Abdolrahimzadeh, Comparison of short-term choroidal thickness and retinal morphological changes after intravitreal anti-vegf therapy with ranibizumab or affibercept in treatment-naive eyes, Current eye research 43 (3) (2018) 391–396.
- [7] M. Saito, M. Kano, K. Itagaki, T. Sekiryu, Efficacy of intravitreal affibercept in japanese patients with exudative age-related macular degeneration, Japanese journal of ophthalmology 61 (1) (2017) 74–83.
- [8] J. Yim, R. Chopra, T. Spitz, J. Winkens, A. Obika, C. Kelly, H. Askham, M. Lukic, J. Huemer, K. Fasler, et al., Predicting conversion to wet age-related macular degeneration using deep learning, Nature Medicine 26 (6) (2020) 892–899.
- [9] S. Ajana, A. Cougnard-Grégoire, J. M. Colijn, B. M. Merle, T. Verzijden, P. T. de Jong, A. Hofman, J. R. Vingerling, B. P. Hejblum, J.-F. Korobelnik, et al., Predicting progression to advanced age-related macular degeneration from clinical, genetic, and lifestyle factors using machine learning, Ophthalmology 128 (4) (2021) 587–597.
- [10] I. Banerjee, L. de Sisternes, J. A. Hallak, T. Leng, A. Osborne, P. J. Rosenfeld, G. Gregori, M. Durbin, D. Rubin, Prediction of age-related macular degeneration disease using a sequential deep learning approach on longitudinal sd-oct imaging biomarkers, Scientific reports 10 (1) (2020) 1–16.
- [11] Q. Yan, Y. Jiang, H. Huang, A. Swaroop, E. Y. Chew, D. E. Weeks, W. Chen, Y. Ding, Genome-wide association studies-based machine learning for prediction of age-related macular degeneration risk, Translational vision science & technology 10 (2) (2021) 29–29.
- [12] A. Bhuiyan, T. Y. Wong, D. S. W. Ting, A. Govindaiah, E. H. Souied, R. T. Smith, Artificial intelligence to stratify severity of age-related

macular degeneration (amd) and predict risk of progression to late amd, Translational vision science & technology 9 (2) (2020) 25–25.

- [13] U. Schmidt-Erfurth, H. Bogunovic, A. Sadeghipour, T. Schlegl, G. Langs, B. S. Gerendas, A. Osborne, S. M. Waldstein, Machine learning to analyze the prognostic value of current imaging biomarkers in neovascular agerelated macular degeneration, Ophthalmology Retina 2 (1) (2018) 24–30.
- [14] M. Rohm, V. Tresp, M. Müller, C. Kern, I. Manakov, M. Weiss, D. A. Sim, S. Priglinger, P. A. Keane, K. Kortuem, Predicting visual acuity by using machine learning in patients treated for neovascular age-related macular degeneration, Ophthalmology 125 (7) (2018) 1028–1036.
- [15] C. Diack, D. Schwab, V. Cosson, V. Buchheit, N. Mazer, N. Frey, A baseline score to predict response to ranibizumab treatment in neovascular age-related macular degeneration, Translational Vision Science & Technology 10 (6) (2021) 11–11.
- [16] F. Rossant, M. Paques, Normalization of series of fundus images to monitor the geographic atrophy growth in dry age-related macular degeneration, Computer Methods and Programs in Biomedicine 208 (2021) 106234.
- [17] Y. Zhang, X. Zhang, Z. Ji, S. Niu, T. Leng, D. L. Rubin, S. Yuan, Q. Chen, An integrated time adaptive geographic atrophy prediction model for sd-oct images, Medical Image Analysis 68 (2021) 101893.
- [18] G. S. Reiter, R. Told, L. Baumann, S. Sacu, U. Schmidt-Erfurth, A. Pollreisz, Investigating a growth prediction model in advanced age-related macular degeneration with solitary geographic atrophy using quantitative autofluorescence, Retina 40 (9) (2020) 1657–1664.
- [19] K. Nattagh, H. Zhou, N. Rinella, Q. Zhang, Y. Dai, K. G. Foote, C. Keiner, M. Deiner, J. L. Duncan, T. C. Porco, et al., Oct angiography to predict geographic atrophy progression using choriocapillaris flow void as a biomarker, Translational vision science & technology 9 (7) (2020) 6–6.
- [20] Y. Zhang, Z. Ji, S. Niu, T. Leng, D. L. Rubin, Q. Chen, A multi-scale deep convolutional neural network for joint segmentation and prediction

of geographic atrophy in sd-oct images, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE, 2019, pp. 565–568.

- [21] Q. Yang, N. Anegondi, V. Steffen, C. Rabe, D. Ferrara, S. S. Gao, Multi-modal geographic atrophy lesion growth rate prediction using deep learning, Investigative Ophthalmology & Visual Science 62 (8) (2021) 235–235.
- [22] H. Bogunović, S. M. Waldstein, T. Schlegl, G. Langs, A. Sadeghipour, X. Liu, B. S. Gerendas, A. Osborne, U. Schmidt-Erfurth, Prediction of anti-vegf treatment requirements in neovascular amd using a machine learning approach, Investigative ophthalmology & visual science 58 (7) (2017) 3240–3248.
- [23] Y. Liu, J. Yang, Y. Zhou, W. Wang, J. Zhao, W. Yu, D. Zhang, D. Ding, X. Li, Y. Chen, Prediction of oct images of short-term response to anti-vegf treatment for neovascular age-related macular degeneration using generative adversarial network, British Journal of Ophthalmology 104 (12) (2020) 1735–1740.
- [24] H. Lee, S. Kim, M. A. Kim, H. Chung, H. C. Kim, Post-treatment prediction of optical coherence tomography using a conditional generative adversarial network in age-related macular degeneration, Retina 41 (3) (2021) 572–580.
- [25] T. R. J. Forshaw, H. J. Ahmed, T. W. Kjær, S. Andréasson, T. L. Sørensen, Full-field electroretinography in age-related macular degeneration: can retinal electrophysiology predict the subjective visual outcome of cataract surgery?, Acta ophthalmologica 98 (7) (2020) 693–700.
- [26] Q. T. Pham, S. Ahn, J. Shin, S. J. Song, Generating future fundus images for early age-related macular degeneration based on generative adversarial networks, Computer Methods and Programs in Biomedicine 216 (2022) 106648.
- [27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Advances in neural information processing systems 27 (2014).

- [28] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: European conference on computer vision, Springer, 2016, pp. 694–711.
- [29] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: International conference on machine learning, PMLR, 2017, pp. 214–223.
- [30] S. Nowozin, B. Cseke, R. Tomioka, f-gan: Training generative neural samplers using variational divergence minimization, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016, pp. 271–279.
- [31] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv preprint arXiv:1411.1784 (2014).
- [32] T. K. Yoo, J. Y. Choi, H. K. Kim, A generative adversarial network approach to predicting postoperative appearance after orbital decompression surgery for thyroid eye disease, Computers in biology and medicine 118 (2020) 103628.
- [33] D. Qiu, Y. Cheng, X. Wang, Improved generative adversarial network for retinal image super-resolution, Computer Methods and Programs in Biomedicine 225 (2022) 106995.
- [34] J. Zhang, X. He, L. Qing, F. Gao, B. Wang, Bpgan: Brain pet synthesis from mri using generative adversarial network for multi-modal alzheimer's disease diagnosis, Computer Methods and Programs in Biomedicine 217 (2022) 106676.
- [35] E. Schonfeld, B. Schiele, A. Khoreva, A u-net based discriminator for generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8207–8216.
- [36] T. Park, M.-Y. Liu, T.-C. Wang, J.-Y. Zhu, Semantic image synthesis with spatially-adaptive normalization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2337–2346.

- [37] Y. Zhang, M. Li, S. Yuan, Q. Liu, Q. Chen, Robust region encoding and layer attribute protection for the segmentation of retina with multifarious abnormalities, Medical Physics 48 (12) (2021) 7773–7789.
- [38] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125–1134.
- [39] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
- [40] D. Misra, Mish: A self regularized non-monotonic neural activation function, arXiv preprint arXiv:1908.08681 4 (2019) 2.
- [41] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, arXiv preprint arXiv:1710.10903 (2017).
- [42] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR, 2020, pp. 1597–1607.
- [43] N. Wang, Y. Zhang, L. Zhang, Dynamic selection network for image inpainting, IEEE Transactions on Image Processing 30 (2021) 1784–1798.
- [44] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, B. Catanzaro, Highresolution image synthesis and semantic manipulation with conditional gans, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8798–8807.
- [45] S. A. Kamran, K. F. Hossain, A. Tavakkoli, S. Zuckerbrod, S. A. Baker, K. M. Sanders, Fundus2angio: A conditional gan architecture for generating fluorescein angiography images from retinal fundus photography, in: International Symposium on Visual Computing, Springer, 2020, pp. 125–138.
- [46] S. A. Kamran, K. F. Hossain, A. Tavakkoli, S. L. Zuckerbrod, Attention2angiogan: Synthesizing fluorescein angiography from retinal fundus images using generative adversarial networks, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 9122–9129.