GLIM-Net: Chronic Glaucoma Forecast Transformer for Irregularly Sampled Sequential Fundus Images

Xiaoyan Hu, Ling-Xiao Zhang, Lin Gao, *Member, IEEE*, Weiwei Dai, Xiaoguang Han, Yu-Kun Lai, *Member, IEEE*, and Yiqiang Chen, *Senior Member, IEEE*

***Abstract*— Chronic Glaucoma is an eye disease with progressive optic nerve damage. It is the second leading cause of blindness after cataract and the first leading cause of irreversible blindness. Glaucoma forecast can predict future eye state of a patient by analyzing the historical fundus images, which is helpful for early detection and intervention of potential patients and avoiding the outcome of blindness. In this paper, we propose a GLaucoma forecast transformer based on Irregularly saMpled fundus images named GLIM-Net to predict the probability of developing glaucoma in the future. The main challenge is that the existing fundus images are often sampled at irregular times, making it difficult to accurately capture the subtle progression of glaucoma over time. We therefore introduce two novel modules, namely time positional encoding and time-sensitive MSA (multi-head self-attention) modules, to address this challenge. Unlike many existing works that focus on prediction for an unspecified future time, we also propose an extended model which is further capable of prediction conditioned on a specific future time. The experimental results on the benchmark dataset SIGF show that the accuracy of our method outperforms the state-of-the-art models. In addition, the ablation experiments also confirm the effectiveness of the two modules we propose, which can provide a good reference for the optimization of Transformer models.**

***Index Terms*— Glaucoma forecast, transformer, attention mechanism, fundus image.**

I. Introduction

Glaucoma is the collective term for a group of diseases that results in progressive damage to the optic nerve and causes loss of vision, primarily associated with pathological intraocular pressure elevation. According to a 2014 global meta-analysis [1], the prevalence of glaucoma for population aged 40–80 years is 3.5%, or approximately 64.3 million people. The number of people with glaucoma is expected to increase to 112 million by 2040 due to population growth and aging.

Many early works [2]–[5] proposed measurement-based methods for automatic screening of glaucoma, which first segmented the pathological regions, such as optic disc and optic cup, and then calculated the relevant clinical values for glaucoma diagnosis. However, the accuracy of glaucoma screening can be seriously affected by the segmentation results, which can be easily influenced by the pathological regions and low image quality.

Recently, with the booming of deep learning, effort has been increasingly devoted to utilizing deep learning methods for automatic glaucoma diagnosis. For example, references [6], [7] applied convolutional neural networks (CNNs) to extract features and detect glaucoma from fundus images directly. However, different from diseases such as cataracts and myopia, the loss of vision caused by glaucoma is irreversible, but most of the loss of vision caused by glaucoma can be avoided by early detection and treatment. Therefore, it is vital to detect the potential deterioration ahead of time for earlier intervention. Li et al. [8] first established a dataset of sequential fundus images, called SIGF, and proposed a deep learning approach (DeepGF) for glaucoma forecast. DeepGF primarily consists of a novel long short-term memory (LSTM) network to learn the spatial and temporal information from sequential fundus images of a person. However, DeepGF outputs the probability of developing glaucoma at the next time step, but cannot specify when the next time is. Furthermore, LSTM may suffer from information loss when passing information from previous steps to the current one and cannot process data in parallel.

In this paper, we propose the first transformer-based glaucoma forecast network named GLIM-Net for irregularly sampled sequential fundus images. There are two main challenges that need to be addressed. First, the transformer architecture lacks inductive bias like convolutions and thus requires a large

This work was supported in part by the Science and Technology Service Network Initiative of the Chinese Academy of Sciences (No. KFJ-STS-QYZD-2021-11-001), in part by the Beijing Municipal Natural Science Foundation for Distinguished Young Scholars (No. JQ21013), in part by the Royal Society Newton Advanced Fellowship (No. NAF\R2\192151), in part by the Science and Technology Innovation Program of Hunan Province (No. 2021GK4015) and in part by the the Youth Innovation Promotion Association CAS. Xiaoyan Hu and Ling-Xiao Zhang contributed equally to this work.

Corresponding authors are Yiqiang Chen (yqchen@ict.ac.cn) and Lin Gao (gaolin@ict.ac.cn).

Xiaoyan Hu, Ling-Xiao Zhang, Lin Gao and Yiqiang Chen are with the Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences. Lin Gao and Yiqiang Chen are also with University of Chinese Academy of Sciences, Beijing, China (e-mail: yxhhxy1999@163.com, {zhanglingxiao, gaolin, yqchen}@ict.ac.cn).

Weiwei Dai is with the Changsha Aier Eye Hospital, Changsha, Hunan, China (e-mail: daiweiwei@aierchina.com).

Xiaoguang Han is with the SSE, The Chinese University of Hong Kong, Shenzhen, China (e-mail: hanxiaoguang@cuhk.edu.cn).

Yu-Kun Lai is with the School of Computer Science and Informatics, Cardiff University, Wales, UK (e-mail: LaiY4@cardiff.ac.uk).

amount of training data [9], but the scale of the SIGF dataset is relatively small. In addition, the images in SIGF dataset were captured at irregular times as it is impractical to request patients to take medical examinations on a regular basis. The time interval of fundus images in a sequence varies from the minimum of only one day to the maximum of 13 years, which will be a challenge for models to learn the status transition over time. To overcome the first challenge, we first replace the input embedding method of simple linear transformation, which was used in recent works of vision transformers [10], [11], with the polar convolutional neural network used in [8] to extract low-level features to enable the model to converge even if the scale of dataset is small. Transformer is designed on the hypothesis that the intervals between samples in a sequence are the same. So to overcome the second challenge, we further propose two novel modules i.e., time positional encoding module and time-sensitive MSA (multi-head self-attention) module. We redesign the positional encoding used in [12] and propose a new time positional encoding (TPE) to enable our model to effectively learn the time distribution of sequential fundus images, which are not sampled regularly. To make full use of the prior knowledge that the longer time the fundus image is captured from now, the less impact the fundus image will have on the current diagnosis, we propose a novel time-sensitive MSA to learn more effective temporal features. Forecasting the probability of developing glaucoma conditioned on a specific time is more meaningful and useful than that of developing glaucoma for an unspecified time, as it provides richer information to make more informed decisions for monitoring, early detection and further timely treatment to slow or halt the glaucoma progression. we therefore introduce an extended model capable of predicting glaucoma for a specific time by feeding the time condition into both input and output labels. Extensive experiments on SIGF dataset [8] demonstrate the effectiveness of our proposed model and our two modules.

Our main contributions can be summarized as follows:

- We propose the first transformer-based glaucoma forecast network named GLIM-Net with time positional encoding and time-sensitive MSA modules to better address the irregularly sampled data.
- We evaluate our model on SIGF dataset and experimental results demonstrate our GLIM-Net achieves better performance than other state-of-the-art methods with a remarkable margin.
- We extend our model to be able to predict glaucoma conditioned on a specific time, which addresses a problem that cannot be handled by existing works, and experiments show the effectiveness of the extended model.

II. Related Work

A. Early Prediction of Diseases

Early prediction of disease deterioration can help clinicians to better treat patients. It is estimated that 11% of patient deaths followed a failure of swift recognition and treatment [13]. Recently, many deep learning methods have been proposed for early prediction of diseases, such as predicting chronic lung disease, acute kidney injury, heart disease, Alzheimer's disease (AD), dementia and glaucoma.

Cheng et al. [14] used convolutional neural networks (CNNs) and temporal fusion mechanisms to analyze electronic health records (EHRs) and predict the probability of suffering from chronic lung disease in the future. Tomavsev et al. [15] proposed a recurrent neural network that processes sequential EHR of patients and outputs a probability of acute kidney injury occurring at any stage of severity within the next 48 hours. Ali et al. [16] first extracted features from both sensor data and EHR, and combined them using a feature fusion method. An ensemble deep learning model was then implemented to perform heart disease prediction.

The above methods are based on EHRs, but the high-dimensionality, sparsity and irregularity of EHR [14], [17] severely affect the accuracy of prediction. Li et al. [18] proposed a deep learning method to implicitly extract features from hippocampal magnetic resonance imaging (MRI) data and established a time-to-event prognostic model to predict the progression of subjects who meet criteria for mild cognitive impairment to Alzheimer's Disease (AD) dementia. Li et al. [8] used sequential fundus images that can better show the subtle pathological features of glaucoma to address the prediction of glaucoma. First, a CNN was applied to extract low-level features, which were then inputted into a long short-term memory (LSTM) network [19] that can better capture temporal features, and finally the probability of developing glaucoma in the future was outputted.

B. Transformers in Vision

Transformers [12] were first proposed for machine translation tasks and have made great success in many natural language processing (NLP) tasks, such as BERT [20], GPT [21], [22], and XLNet [23]. In recent years, self-attention mechanism has been extensively explored to build long-range dependencies. Reference [24] showed that attention modules can completely replace convolution operations and the study reported in [25] showed that self-attention layers deal with pixel-grid patterns analogously to CNN layers. Recently, many works have successfully applied self-attention to various vision tasks, such as image classification, object detection and image segmentation. DETR [26] regarded object detection as a direct set prediction task and reasoned about the relationships of the objects and the global image context to yield the final set of predictions straightforwardly in parallel. In ViT (Vision Transformer) [10], an image was split into patches and a pure transformer with MLP (Multi-Layer Perceptron) head was applied to perform image classification. On the basis of ViT, DeiT [11] proposed some training strategies and introduced the knowledge distillation (KD) [27] to make Transformer more efficient and achieved competitive results on small datasets. VisTR [28] regarded video instance segmentation as a direct coding/prediction task. With a video clip comprising a specified number of image frames as input, VisTR outputted a sequence of masks for individual instances in the video directly and in order.

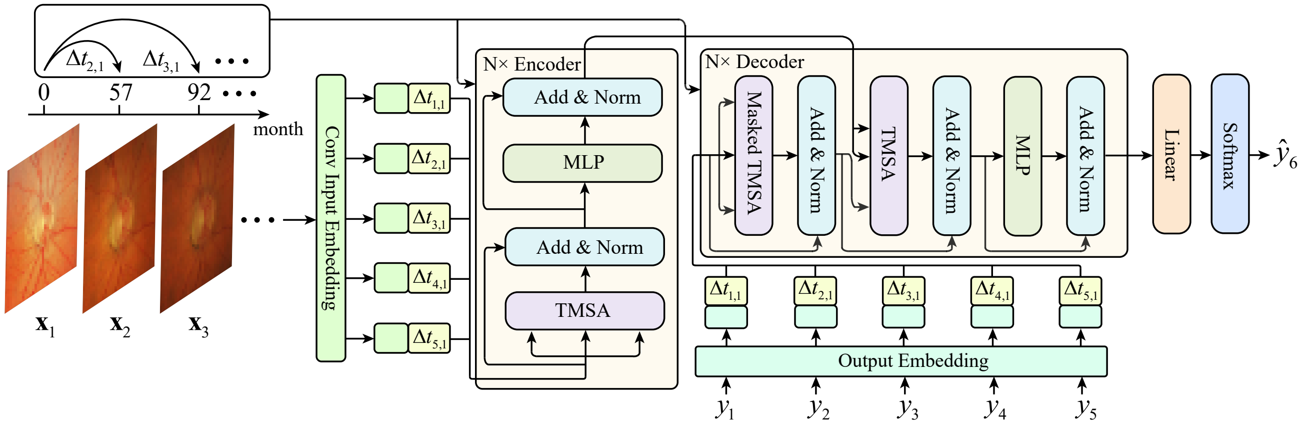

Fig. 1. The proposed network architecture is an encoder-decoder structure. The encoder inputs the embedding feature of a sequence of fundus images, which is then added with time positional encoding proposed to obtain temporal information. The decoder inputs the encoded feature and the embedding feature of the labels (0 for negative and 1 for positive) added with time positional encoding, and outputs the probability distribution of glaucoma. TMSA (Time-sensitive multi-head self-attention) is designed to make full use of the prior knowledge that the longer time the fundus image is captured from now, the less impact the fundus image will have on the current diagnosis. $\boldsymbol{\Delta t_{i,j}}$ denotes the time offset of the $\boldsymbol{i}$-th fundus image from the $\boldsymbol{j}$-th fundus image in the sequence. More details of the TMSA module are shown in Fig. 2. This illustration is based on taking the fundus images captured in the 5 visits as input and predict the probability of developing glaucoma at the time of the 6th visit, to match the setting of previous work [8]. Our method is general and can be applied to different numbers of images in the input sequence.

Some other studies [29]–[32] tried to apply self-attention mechanisms to the medical imaging field. Xiong et al. [29] proposed a reinforcement learning method, namely reinforced transformers for medical image captioning (RTMIC), to generate long and coherent medical imaging reports. Inspired by DETR, Prangemeier et al. [30] proposed a novel attention-based detection transformer called cell detection transformer (Cell-DETR) for faster cell instance segmentation. Valanarasu et al. [31] introduced a modified gated axial-attention by presenting an additional control mechanism and incorporated it as the building block of multi-head attention models for medical image segmentation. Cao et al. [32] introduced a Transformer-based U-shaped Encoder-Decoder architecture with skip-connections to extract local-global semantic features for medical image segmentation. However, most of there previous works cannot performance well on irregularly sampled data. In this paper, we propose two modules, namely time positional encoding and time-sensitive self-attention, to make full use of temporal information and enable the transformer to achieve good results on irregularly sampled data.

III. Method

In this section, we introduce our GLIM-Net in detail. We first discuss the vanilla transformer and our network architecture, and then introduce the two novel modules, time positional encoding and time-sensitive MSA, as well as the loss function. Finally we demonstrate the extended model with the capability of forecasting glaucoma conditioned on a specific time.

A. Revisiting Transformer Architecture

The vanilla Transformer [12] is an encoder-decoder structure with stacked encoder and decoder layers. Positional encodings are added to the embedded input prior to the first layer of encoder and decoder to make use of the order of the input sequence. Each encoder layer is composed of two sub-layers i.e., a multi-head self-attention mechanism followed by a fully connected feed-forward network. In addition to the two sub-layers of encoder layer, the decoder layer has an additional masked multi-head self-attention mechanism before the two sub-layers to ensure that the prediction of time t_i can only depend on the known data at positions before t_i. A residual connection [33] followed by layer normalization [34] is employed around each pair of sub-layers.

B. GLIM-Net Framework

The architecture of GLIM-Net is illustrated in Fig. 1, which is also an encoder-decoder structure. For a given sequence consisting of n fundus images, each image is denoted by $\mathbf{x}_i, i \in [1, ..., n]$, and has a corresponding time stamp t_i and label $y_i \in \{0, 1\}$, 0 for negative glaucoma and 1 for positive glaucoma. The output of GLIM-Net $\hat{y}_{i+1}$ is the probability of positive glaucoma of the image $\mathbf{x}_{i+1}$, We denote the GLIM-Net by $\mathcal{F}$, and $\hat{y}_{i+1} = \mathcal{F}_i = \mathcal{F}(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_i, y_1, y_2, ..., y_i)$. The final output is y_{n+1} corresponding to the the fundus image $\mathbf{x}_{n+1}$ captured in the next visit so unknown to the algorithm but used by clinicians to determine the ground truth label y_{n+1}. Specifically, we first apply a polar transformation [35] and a convolution network [8] on the fundus image $\mathbf{x}_i$ to extract low-level features $\mathbf{f}_i$. Then, the time positional encoding (see Sec. III-C) is added to $\mathbf{f}_i$ to form $\tilde{\mathbf{f}}_i \in \mathbb{R}^{d_m}$ as the input of the encoder. The encoder consists of N identical layers. We set $d_m = 512$ and $N = 6$ in our paper. Similar to the work [12], each layer is composed of two sub-layers, i.e., time-sensitive MSA and multilayer perceptron, and a residual connection [33] followed by layer normalization [34] is employed around each of the two sub-layers. The encoded feature will then be fed into the decoder, along with the labels' embedding feature $\tilde{\mathbf{y}}_i \in \{\mathbf{0}^{d_m}, \mathbf{1}^{d_m}\}$ added with time positional encoding, and the output is the probability of positive glaucoma $\hat{y}_{i+1}$. The structure of the decoder is similar to that of the encoder, and it is also composed of N identical layers. Each layer adds a masked time-sensitive MSA to ensure that the prediction of time t_i can only depend on the previous known data.

C. Time Positional Encoding

The self-attention module in the original Transformer [12] is permutation invariant, so it needs positional encoding to combine input embedding and position information to enable the model to have the ability to learn the order of the sequence. Positional encoding can either be fixed or learnable, and either absolute or relative. A fixed sinusoidal absolute positional encoding is proposed in [12]. Based on sinusoidal positional encoding, we propose a time positional encoding (TPE) that can make better use of the temporal information of irregularly sampled data. TPE can be expressed by the following

$$\mathrm{TPE}(\Delta t_{i,1}, q) = \begin{cases} \sin(\omega_q \cdot \Delta t_{i,1}), & q = 2k \\ \cos(\omega_q \cdot \Delta t_{i,1}), & q = 2k+1 \end{cases} \tag{1}$$

where $\Delta t_{i,1}$ is the time offset of the current fundus image $\mathbf{x}_i$ from the first fundus image $\mathbf{x}_1$ in the sequence and $\omega_q = 1/10000^{2q/d_m}, q \in [0, d_m]$. The wavelengths of functions increase from 2π to $10000 \cdot 2\pi$. To make encoding richer, sin and cos functions are used alternately. The modified function can not only learn the relative position relationship of sequential fundus images, but also learn the time distance relationship between fundus images in the sequence.

D. Time-Sensitive MSA

References [10], [11], [28] have proved that self-attention performs well in capturing long-range dependencies in the vision field. However, the data of these works is all regularly sampled, and thus models are designed based on the hypothesis that the intervals between samples in the sequence are equal, which makes self-attention sensitive to the position relationship of data but not sensitive to the time relationship of data. In this paper, we propose a novel time-sensitive self-attention module to enable the model to learn the time distance dependencies of sequential fundus images. Also we extend our time-sensitive self-attention to time-sensitive MSA to allow the model to attend to features from different representation subspaces.

1) Time-Sensitive Self-Attention: On the basis of the self-attention mechanism proposed by Vaswani et al. [12], we propose a time-sensitive self-attention mechanism to tackle the variable time interval problem of this task, as illustrated in Fig. 2 (a), which can be expressed by the following

$$\mathcal{G}(\mathbf{Q}, \mathbf{K}, \mathbf{T}, \mathbf{V}) = \mathrm{softmax}(\frac{\mathbf{Q}\mathbf{K}^T \circ \mathbf{T}}{\sqrt{d_m}})\mathbf{V} \tag{2}$$

$$\mathbf{Q} = \tilde{\mathbf{f}}\mathbf{W}^Q \tag{3}$$

$$\mathbf{K} = \tilde{\mathbf{f}}\mathbf{W}^K \tag{4}$$

$$\mathbf{V} = \tilde{\mathbf{f}}\mathbf{W}^V \tag{5}$$

where $\tilde{\mathbf{f}}$ is the result of the input embedding, $\tilde{\mathbf{f}} \in \mathbb{R}^{n \times d_m}$. We obtain the queries $\mathbf{Q} \in \mathbb{R}^{n \times d_q}$, keys $\mathbf{K} \in \mathbb{R}^{n \times d_k}$ and values $\mathbf{V} \in \mathbb{R}^{n \times d_v}$ by multiplying $\tilde{\mathbf{f}}$ with learnable weight matrices $\mathbf{W}^Q \in \mathbb{R}^{d_m \times d_q}$, $\mathbf{W}^K \in \mathbb{R}^{d_m \times d_k}$ and $\mathbf{W}^V \in \mathbb{R}^{d_m \times d_v}$, respectively. d_q, d_k and d_v are the dimensions of $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$. We set $d_q = d_k = d_v = d_m = 512$ in this paper. $\mathbf{T}$ is a weight matrix related to time interval of data and $\mathbf{T} \in \mathbb{R}^{d_m \times d_m}$. $\circ$ denotes the Hadamard product.

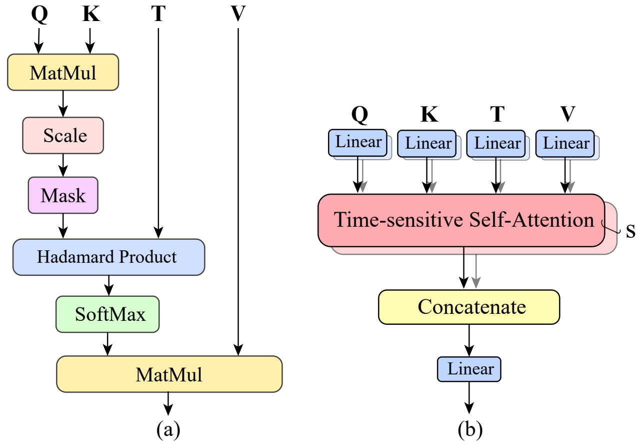

Fig. 2. (a) Time-Sensitive Self-Attention (b) Time-Sensitive MSA consisting of s heads.

Let $\mathbf{P} = \mathbf{Q}\mathbf{K}^T$, where $\mathbf{P} \in \mathbb{R}^{n \times n}$, and $\mathbf{K}^T$ is the transpose of $\mathbf{K}$. With the prior knowledge of vector multiplication, the value of $\mathbf{P}_{i,j}$ represents the influence of fundus image $\mathbf{x}_j$ on the fundus image $\mathbf{x}_i$. This is sufficient in the setting of machine translation, which represents the influence of the j-th word on the i-th word. However, in the context of glaucoma forecast task, the data of which is irregularly sampled, we propose to modify the self-attention to learn the temporal information of sequential fundus images. We have the prior knowledge that the longer time the fundus image is captured from now, the less impact the fundus image will have on the current diagnosis. Therefore, we impose an additional time-related matrix $\mathbf{T}$ on $\mathbf{P}$ through Hadamard product to enable the self-attention mechanism to better cope with the irregularly sampled data. The advantage of this design is to add a strong constraint to the model, so that the model pays more attention to the recent fundus images and less attention to the fundus images long time ago when processing a certain fundus image. The value of $\mathbf{T}_{i,j}$ is defined as

$$\mathbf{T}_{i,j} = \frac{1}{e^{A\Delta t_{i,j} + B}} \tag{6}$$

$$\Delta t_{i,j} = \max(|\Delta t_{i,j}|, \delta)/\delta \tag{7}$$

where A and B are learnable parameters. A is expected to be a positive number when the model converges, so that the function is a monotonically decreasing function, which conforms to our prior knowledge. $\Delta t_{i,j}$ is the time interval from the fundus image $\mathbf{x}_i$ to the image $\mathbf{x}_j$ in the sequence. Because the time stamp of the fundus images in the dataset is accurate to months, $\Delta t_{i,j}$ is counted in months. The maximum time interval between fundus images can reach 256 months in the dataset. However, the average time interval is only 39 months, which means that most of time interval in the dataset is much smaller than 256 months. Therefore, the time interval

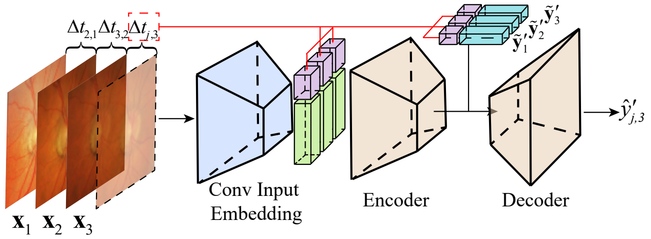

Fig. 3. Conditional Glaucoma Forecast Network. Time condition is incorporated in both the encoder and the decoder of GLIM-Net as additional input to enable the model to output the probability of developing glaucoma at a specified time. The illustration is based on taking the first 3 fundus images (i.e., $\boldsymbol{i} = \mathbf{3}$) to predict the probability of glaucoma at time $\boldsymbol{j}$, and our method works for general settings.

is clipped to a maximum value of δ and and then normalized by dividing it by δ using Eq. 7. The final experimental results prove the correctness of the conjecture.

The results of $\mathbf{P} \circ \mathbf{T}$ are then divided by $\sqrt{d_m}$ before softmax normalization to ensure the stability of the gradient descent during training. So far, the final weight matrix is obtained. We multiply the weight matrix by the value vector $\mathbf{V}$ and get the result of time-sensitive self-attention module $\mathcal{G}(\mathbf{Q}, \mathbf{K}, \mathbf{T}, \mathbf{V})$, which fuses all other data in the sequence according to the weight matrix on the basis of the current input.

2) Time-Sensitive MSA: According to the study reported in [12], instead of using the d_m-dimensional queries, keys and values to perform a single self-attention function, it is better to use the d_m/s-dimensional queries $\mathbf{Q}_i \in \mathbb{R}^{n \times d_q/s}$, keys $\mathbf{K}_i \in \mathbb{R}^{n \times d_k/s}$ and values $\mathbf{V}_i \in \mathbb{R}^{n \times d_v/s}$ to perform s times self-attention function, then concatenate the results of all the s times of operations together and project it to d_m by a linear projection. The multi-head self-attention (MSA) mechanism increases the representation subspaces that the model can attend to information. Therefore, we extend our time-sensitive self-attention to time-sensitive MSA, which is depicted in Fig. 2 (b) and can be expressed by the following

$$\mathcal{H}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{T}) = \text{Concat}(\mathbf{H}_1, ..., \mathbf{H}_\text{s})\mathbf{W}^\text{O} \tag{8}$$

where $\mathbf{H}_i = \mathcal{G}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i, \mathbf{T}_i)$. *Concat* function concatenates all the outputs of s time-sensitive self-attention modules together. $\mathbf{W}^O \in \mathbb{R}^{hd_v \times d_m}$ is a projection matrix used to project the result to d_m-dimension. In this paper, we set $s = 8$ heads and set $d_v = d_m/s = 64$ for each head.

E. Loss Function

As the glaucoma forecast task has only two classes, i.e., positive and negative glaucoma, we use the binary entropy loss as the loss function of our model, which can be expressed by the following

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n} y_{i+1}\log(\hat{y}_{i+1}) + (1 - y_{i+1})\log(1 - \hat{y}_{i+1}) \tag{9}$$

where $\hat{y}_{i+1} \in [0, 1]$ denotes the forecast probability of image $\mathbf{x}_{i+1}$.

F. Conditional Glaucoma Forecast Network

Forecasting the probability of developing glaucoma conditioned on a specific time is more meaningful and useful than that of developing glaucoma for an unspecified time, which provides richer information to help make more informed decisions for monitoring, early detection and further timely treatment to slow or halt the glaucoma progression. Therefore, we extend our GLIM-Net to conditional GLIM-Net to assist people who may develop glaucoma in a few years to take measures to prevent further vision loss. Specifically, the time condition is incorporated in both the encoder and the decoder as additional input. The architecture of conditional GLIM-Net is shown in Fig. 3. We denote C-GLIM-Net by $\hat{\mathcal{F}}$, and $\hat{y}'_{j,i} = \hat{\mathcal{F}}_{j,i} = \hat{\mathcal{F}}(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_i, y_1, y_2, ..., y_i|\Delta t_{j,i})$ means that the network inputs a sequence consisting of i fundus images and outputs the probability of glaucoma of $\mathbf{x}_j$ $(j > i)$. Different from GLIM-Net, the last layer of the convolution network [8] is changed to output a $(d_m - 1)$-dimension feature and it is concatenated with the time interval $\Delta t_{j,i}$ as the input of the encoder, and the embedding feature of the label is also changed to $(d_m - 1)$-dimension and is also concatenated with the time interval $\Delta t_{j,i}$ as the input of decoder. The decoder last outputs the probability of positive glaucoma $\hat{y}'_{j,i} \in [0, 1]$.

To train the C-GLIM-Net, we feed first i images of a sequence to predict the $\hat{y}'_{i+1,i}$ and $\hat{y}'_{i+2,i}$, under the conditions $\Delta t_{i+1,i}$ and $\Delta t_{i+2,i}$. The loss function is modified to

$$\begin{aligned}\mathcal{L}_\text{C} = -(&\frac{1}{n}\sum_{i=1}^{n}(y_{i+1}\log(\hat{y}'_{i+1,i}) + (1 - y_{i+1})\log(1 - \hat{y}'_{i+1,i})) \\ &+ \frac{1}{n-1}\sum_{i=1}^{n-1}(y_{i+2}\log(\hat{y}'_{i+2,i}) + (1 - y_{i+2})\log(1 - \hat{y}'_{i+2,i})))\end{aligned} \tag{10}$$

We conduct experiments of C-GLIM-Net model on the SIGF dataset by forecasting glaucoma conditioned on the time of the next two fundus images in the sequence simultaneously. Experimental results are detailed in the next section.

IV. Experiments and Results

A. Datasets

To evaluate the effectiveness of our proposed method, we first carry out experiments on a dataset consisting of sequential fundus images annotated with either positive or negative glaucoma named SIGF [8]. We further prove the effectiveness of our proposed method on a dataset named Tumor-CIFAR proposed by [36] that shares the same characteristics as the SIGF dataset.

1) SIGF: The SIGF contains 3671 fundus images in total and consists of 405 sequential fundus images from different eyes with an average of 9 images per eye ranging from 1986 to 2018. In the dataset, there are at least 6 fundus images for each eye. All the fundus images are annotated with positive glaucoma when they satisfy any of the three criteria, i.e., retinal nerve fiber layer defect, rim loss and optic disc hemorrhage. The sequences are divided into 2 types: time-variant and time-invariant as defined in [8]. Time-variant sequences are those that change from negative to positive

TABLE I
THE DISTRIBUTION OF TIME-VARIANT TYPE DATA AND TIME-INVARIANT TYPE DATA IN TRAINING, VALIDATION AND TEST SET.

	Training	Validation	Test
Time-variant	27	3	7
Time-invariant	273	32	63

glaucoma and time-invariant sequences are those that keep negative glaucoma. SIGF contains 37 time-variant sequences and 368 time-invariant sequences. In our experiments, we use the same training (300), validation (35) and test (70) sets used in [8], and the ratios of the time-variant and time-invariant sequences are roughly the same in the training, validation and test sets, detailed in Table I. There are 264 patients in total in the dataset, with 192 patients, 23 patients and 49 patients in the training set, validation set and test set, respectively. The dataset is randomly split at the patient level.

2) Tumor-CIFAR: The Tumor-CIFAR is a simulated dataset of nodules based on the CIFAR10 dataset [37]. Tumor-CIFAR contains 60,000 samples which are randomly divided into training (40,000), validation (10,000) and test (10,000) sets. Each sample consists of five sequential images that are extended from one image in the CIFAR10 dataset with two gradually growing nodules, the size of which is computed by the following

$$s_i = t_i \times g \tag{11}$$

where t_i is the time stamp from the beginning, g is the growth rate, i is the sequential index. The difference between malignant and benign nodules is the growth rate g. The growth rate of malignant pulmonary nodules is roughly three times as the benign one. The growth rate g of simulated nodules is expressed by the following

$$g = \frac{s_i}{t_i} \sim \begin{cases} N(3, 1.8), & \text{if malignant} \\ N(1, 0.2), & \text{if benign} \end{cases} \tag{12}$$

where $N(\mu, \sigma^2)$ denotes the Gaussian distribution whose mean and variance are μ and σ^2, respectively.

B. Experimental Setup

1) Implementation Details: We implement our GLIM-Net with Tensorflow [38] on a single NVIDIA TITAN X (Pascal) GPU with 12 GB memory. We employ the Adam algorithm [39] to update the parameters of our model with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1e - 8$. Considering that the data distribution of SIGF dataset is extremely imbalanced, we adopt the active convergence strategy used in [8] to self-update the distribution of the training set actively and adaptively, i.e., discarding the sequences the training loss of which ranks the lowest. Specifically, we set learning rate $l_r = \{3 \times 10^{-8}, 10^{-7}, 6 \times 10^{-6}, 6 \times 10^{-6}, 10^{-5}, 5 \times 10^{-7}\}$ each for 5 epochs. The parameters of time-sensitive MSA, A and B, are initialized as 0.9 and 2.0, respectively, according to our extensive experiments. The batch size is set to 4. Our model can gradually achieve the approximate global optimum after 400 epochs of iterations.

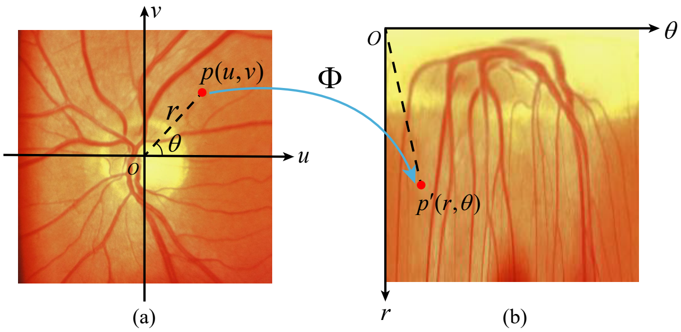

Fig. 4. The fundus image in the Cartesian coordinate system (a). The corresponding fundus image in polar coordinate system (b) through polar transformation in Eq. 13.

To augment the data set and avoid the overfitting problem, following the work [8], we segment the 405 sequences into 1146 clips, with each clip comprising T (=6) time steps and an overlap of 5 frames is allowed in segmenting the sequences into clips. Before being input into the model, the fundus images are reshaped to 224×224 and a polar transformation Φ is applied on them in order to enlarge the disc and cup structure, as illustrated in Fig. 4, which can be expressed by the following

$$\begin{cases} r = \sqrt{u^2 + v^2}, \\ \theta = \tan^{-1}(\frac{v}{u}), \end{cases} \tag{13}$$

where (u, v) denotes a point in the original Cartesian coordinate system, as shown in Fig. 4 (a). r and θ are the radius and directional angle in the polar coordinate system, as shown in Fig. 4 (b).

2) Evaluation Metrics: To compare our method with other state-of-the-art methods, we adopt four commonly-used metrics to assess our GLIM-Net and [8], [36], [40]–[42] on SIGF, including accuracy, sensitivity, specificity and AUC (Area Under Curve).

C. Comparison with State-of-the-art Methods

1) Comparison on the SIGF dataset: We use the vanilla transformer as our Baseline method, because transformer can build long-range dependencies and shows good performance in processing sequential data. We compare our model with the baseline method and 8 other state-of-the-art methods to verify the superiority of our proposed method, namely DeepGF [8], Deep CNN [40], AG-CNN [41], tLSTM [42], DLSTM [36], MIL-VT [43], CABNet [44] and CoG-Net [45]. Notice that in DeepGF, they convert image classification models to prediction models by supervising the models with the labels at the next time step, so we similarly try to use two sequential image classification models, tLSTM [42] and DLSTM [36], the setting of which is more similar to the setting of our task than that of the image classification models. The MIL-VT, CABNet and CoG-Net are designed for fundus image classification and we convert them to be prediction models by supervising the models with the labels at the next time step as DeepGF [8] does.

Table II tabulates the results of our model, baseline method and other state-of-the-art-methods. As shown, our baseline

TABLE II

COMPARISON WITH THE STATE-OF-THE-ART METHODS ON SIGF. ACC, SE AND SP DENOTE THE ACCURACY, SENSITIVITY AND SPECIFICITY, RESPECTIVELY. OUR METHOD OUTPERFORMS OTHER METHODS IN ALL FOUR METRICS WITH A CONSIDERABLE MARGIN. FOR EACH MEASURE, WE SHOW THE MEAN AND STANDARD DEVIATION.

Method	ACC (%)	SE (%)	SP (%)	AUC (%)
AG-CNN	47.5 ± 0.5	57.5 ± 4.6	47.1 ± 0.7	51.8 ± 2.5
Deep CNN	67.0 ± 5.5	62.1 ± 1.8	67.2 ± 5.7	63.0 ± 3.3
DeepGF	76.0 ± 4.8	79.4 ± 1.3	75.9 ± 5.0	85.0 ± 2.5
tLSTM	84.7 ± 1.4	72.5 ± 0.0	85.1 ± 1.4	83.6 ± 0.5
DLSTM	87.2 ± 1.1	84.3 ± 1.6	87.3 ± 1.2	92.4 ± 0.7
MIL-VT	79.7 ± 1.1	77.8 ± 3.4	79.8 ± 1.2	83.4 ± 1.6
CABNet	73.9 ± 1.6	74.4 ± 1.6	73.9 ± 1.6	78.7 ± 2.6
CoG-Net	77.0 ± 2.0	72.5 ± 3.6	77.2 ± 1.7	81.8 ± 3.5
Baseline	84.7 ± 0.2	77.8 ± 0.9	84.9 ± 0.1	86.1 ± 1.5
Ours	$\mathbf{89.5 \pm 0.8}$	$\mathbf{87.6 \pm 0.9}$	$\mathbf{89.6 \pm 0.8}$	$\mathbf{93.6 \pm 0.3}$

TABLE III

COMPARISON ON THE TUMOR-CIFAR DATASET.

Method	ACC (%)	SE (%)	SP (%)	AUC (%)
DLSTM	96.1 ± 0.3	96.0 ± 0.6	96.1 ± 0.7	99.3 ± 0.1
DeepGF	94.0 ± 0.1	94.1 ± 0.1	93.8 ± 0.2	98.6 ± 0.1
Ours	$\mathbf{97.3 \pm 0.6}$	$\mathbf{97.1 \pm 0.3}$	$\mathbf{97.4 \pm 0.8}$	$\mathbf{99.7 \pm 0.1}$

method performs better than 6 other state-of-the-art methods in terms of accuracy, sensitivity, specificity and AUC, and our modified model significantly improves the performance of baseline method. Our GLIM-Net model considerably outperforms 8 other state-of-the-art methods and the baseline method by achieving 89.5%, 87.6%, 89.6% and 93.6% in accuracy, sensitivity, specificity and AUC, respectively.

The loss, ACC and ROC (Receiving Operating Characteristic) curves of the training and validation phases are shown in Figure 5. (a) and (b) show the ROC curves of the training and validation phases. (c) presents the loss curves and (d) shows the ACC curves of the training and validation phases. The blue curve is the result of the training phase, and the red curve is the result of the validation phase.

DeepGF [8] and DLSTM [36] are based on the LSTM network, consisting of several recurrent layers, which need to process data sequentially. This enables the LSTM network to use more information of adjacent positions to make decisions. In contrast, the Transformer architecture is based entirely on the attention mechanism and the positional encoding. It leverages the information of all positions and allows for significantly more parallelization. And our proposed time positional encoding and time-sensitive multi-head self-attention make the transformer architecture better suit the irregularly sampled data.

2) Comparison on the Tumor-CIFAR dataset: To further verify the ability of handling irregularly sampled sequential data of our model, we also carry out experiments on the Tumor-CIFAR dataset [36]. As listed in Table III, our method achieves 97.3%, 97.1%, 97.4%, 99.7% in accuracy, sensitivity, specificity and AUC respectively, which illustrates that our method can deal with irregularly sampled sequential data well.

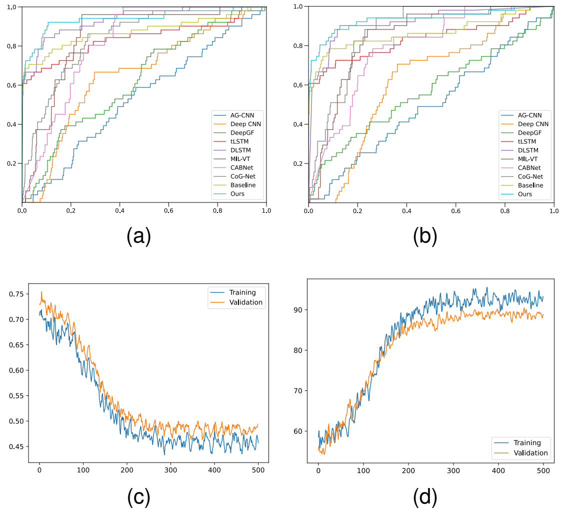

Fig. 5. The ROC curves of the (a) training and (b) validation phases. (c) the loss curves and (d) the ACC curves of the training and validation phases. The blue curve is the result of training phases, and orange curve is the result of validation phases.

TABLE IV

ABLATION STUDIES ON CONVOLUTIONAL INPUT EMBEDDING (CIE), TIME POSITIONAL ENCODING (TPE) AND TIME-SENSITIVE MSA (TMSA). ACCORDING TO THE RESULTS, THE TMSA MODULE IS THE MOST EFFECTIVE MODULE OF OUR METHOD.

Method	ACC (%)	SE (%)	SP (%)	AUC (%)
w/o CIE	85.3 ± 1.1	83.0 ± 0.9	85.4 ± 1.1	90.0 ± 0.4
w/o TPE	87.9 ± 0.6	81.1 ± 0.9	88.1 ± 0.6	88.6 ± 1.8
w/o TMSA	85.0 ± 1.5	81.1 ± 0.9	85.1 ± 1.5	88.5 ± 2.0
Ours	$\mathbf{89.5 \pm 0.8}$	$\mathbf{87.6 \pm 0.9}$	$\mathbf{89.6 \pm 0.8}$	$\mathbf{93.6 \pm 0.3}$

D. Ablation Study

To demonstrate the effectiveness of our each module, we conduct ablation studies on the convolutional input embedding, time positional encoding, time-sensitive MSA, the initialization of hyper-parameters A and B and the setting of threshold δ.

1) Convolutional Input Embedding (CIE): We first evaluate the impact of convolutional input embedding compared to the simple linear transformation by flattening the original fundus image $\mathbf{x}_{img} \in \mathbb{R}^{3\times H\times W}$ to a vector $\mathbf{x}'_{img} \in \mathbb{R}^{3HW}$ and then using linear transformation to embed it to d_m dimensions. As indicated in Table IV, the convolutional input embedding can improve accuracy and specificity by 4.2%. Also, sensitivity and AUC increase considerably from 83.0% and 90.0% to 87.6% and 93.6%, respectively. These results verify the rationality of replacing linear transformation with ConvNet as the input embedding method.

2) Time Positional Encoding (TPE): To demonstrate the superiority of our time positional encoding, we replace it with the vanilla positional encoding. As shown in Table IV, TPE improves the performance of glaucoma forecast in terms of all four evaluation metrics. Specifically, TPE improves the accuracy and specificity by 1.6% and 1.5%. Sensitivity and AUC reach 87.6% and 93.6% from 81.1% and 88.6%, respectively. These results verify that it is an effective module to make use of temporal information.

TABLE V

COMPARISON OF ACCURACY WITH DIFFERENT $\boldsymbol{A}$ IN EQ. 6. WE FIX $\boldsymbol{B = 2.0}$ AND THEN EVALUATE THE IMPACT OF $\boldsymbol{A}$.

A	0.3	0.6	0.9	1.2	1.5
ACC (%)	82.6±0.8	85.9±1.3	$\mathbf{89.5 \pm 0.8}$	87.1±1.4	82.6±1.2

TABLE VI

COMPARISON OF ACCURACY WITH DIFFERENT $\boldsymbol{B}$ IN EQ. 6. WE FIX $\boldsymbol{A = 0.9}$ AND THEN EVALUATE THE IMPACT OF $\boldsymbol{B}$.

B	0	1.0	2.0	3.0	4.0
ACC (%)	81.4±0.9	85.0±1.3	$\mathbf{89.5 \pm 0.8}$	86.5±0.7	82.0±1.0

3) Time-sensitive MSA (TMSA): We justify our time-sensitive MSA by simply removing the time related matrix ($\mathbf{T}$) from our model. As shown in Table IV, TMSA significantly improves the performance of the model, and this module improves the network performance most. Accuracy, sensitivity, specificity and AUC are all increased by at least 4%, to be specific, 4.5%, 6.5%, 4.5% and 5.1%, respectively.

4) Hyper-parameters A *and* B *in Eq. 6:* As aforementioned, we need the model to pay more attention to the most recent fundus images and less on the older ones when processing a particular fundus image. Eq. 6 is the key component to achieve this. A and B are two hyper-parameters controlling the overall trend of the function of Eq. 6. According to our prior knowledge, hyper-parameter A should be positive so that the function in Eq. 6 will be a monotonically decreasing function. Table V manifests that using a small and positive A, the model can perform better. And the model shows a performance drop when A is either smaller or bigger. Therefore, A is initialized as 0.9 in all experiments. Furthermore, we investigate the impact of different B on the performance of the model and find that model performs best when B is initialized as 2, as illustrated in Table VI. Setting B smaller or bigger will affect the performance of the model. Therefore, B is initialized as 2 in all experiments. Note that these settings are initializations and A and B are learnable parameters updated during training.

5) Hyper-parameter δ in Eq. 7: We evaluate the effect of different thresholds of time interval used in Eq. 7. As shown in Table VII, the model performs best when $\delta = 96$ (8 years) and shows a performance drop when δ is either smaller or bigger. This result is reasonable because there are only 4.6% of data in the dataset that have a bigger time interval than 96 months. Furthermore, the data of 96 months away from the current data may have little influence on the diagnosis. And if δ is too small, temporal information between fundus images in the sequence may be lost.

E. Conditional Glaucoma Forecast Network

With the help of time condition, C-GLIM-Net can output the probability of developing glaucoma of a patient conditioned on a specific time. In the following experiment, we feed in first $i = 4$ images of a sequence into C-GLIM-Net, under the condition $\Delta t_{i+1,i}$ and $\Delta t_{i+2,i}$ to predict the labels of the 5th and 6th images. We denote the results of the 5th and 6th predictions as C-GLIM$_1$ and C-GLIM$_2$, respectively. Because DeepGF is used to predict probability at the next time step and cannot predict the probability at a specified time, to compare with C-GLIM$_1$, we feed in first $i = 4$ images to train the DeepGF to only predict the label of the 5th image. As shown in Table VIII, our C-GLIM$_1$ outperforms the DeepGF by achieving 83.0%, 87.6%, 82.9% and 90.7% in accuracy, sensitivity, specificity and AUC respectively. The specificity is dropped slightly in C-GLIM$_2$ because it is harder to predict positives at a longer time. To further investigate this, we show the prediction of time-steps in-between the 5th and 6th visits in Fig. 6. Although the probability of positives are increased stably, the false negatives are increased due to the probabilities of some positive samples do not exceed the 50 % threshold at the 6th visit.

TABLE VII

COMPARISON OF ACCURACY WITH DIFFERENT $\boldsymbol{\delta}$ IN EQ. 7. THE NETWORK GETS BEST ACCURACY WHEN $\boldsymbol{\delta = 96}$. THIS RESULT IS REASONABLE BECAUSE THERE ARE ONLY 4.6% OF DATA IN THE DATASET THAT HAVE A BIGGER TIME INTERVAL THAN 96 MONTHS (8YEARS). FURTHERMORE, THE DATA OF 96 MONTHS AWAY FROM THE CURRENT DATA MAY HAVE LITTLE INFLUENCE ON THE DIAGNOSIS. AND IF $\boldsymbol{\delta}$ IS TOO SMALL, TEMPORAL INFORMATION BETWEEN FUNDUS IMAGES IN THE SEQUENCE MAY BE LOST.

δ	24	60	96	132	168
ACC (%)	81.5±1.2	84.3±0.5	$\mathbf{89.5 \pm 0.8}$	85.8±1.0	84.1±0.7

TABLE VIII

OUR CONDITIONAL GLIM-NET COMPARED WITH THE STATE-OF-THE-ART METHOD, DEEPGF, ON SIGF. C-GLIM$_\mathbf{1}$ AND C-GLIM$_\mathbf{2}$ DENOTE THE CONDITIONAL GLIM-NET CONDITIONED ON THE TIME OF NEXT FUNDUS IMAGE AND THE FUNDUS IMAGE AFTER THE NEXT ONE IN THE SEQUENCE, RESPECTIVELY.

Method	ACC (%)	SE (%)	SP (%)	AUC (%)
DeepGF	74.4 ± 5.3	74.1 ± 8.1	74.4 ± 5.4	79.8 ± 8.6
C-GLIM$_1$	$\mathbf{83.0 \pm 0.6}$	$\mathbf{87.6 \pm 0.9}$	82.9 ± 0.7	$\mathbf{90.7 \pm 0.3}$
C-GLIM$_2$	82.8 ± 0.6	74.7 ± 1.9	$\mathbf{83.1 \pm 0.6}$	81.8 ± 0.6

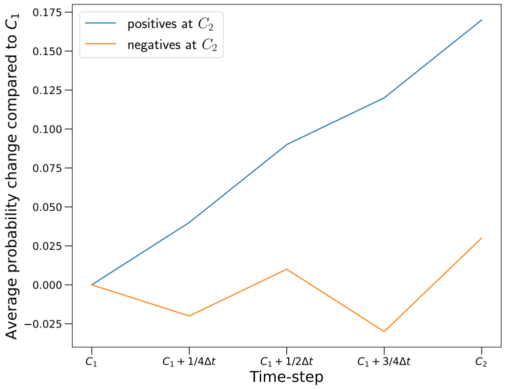

Fig. 6. Average probability changes compared to $\boldsymbol{C_1}$ of our C-GLIM-Net performing on the time-steps between the 5th and 6th visits. The blue curve denotes the average probability changes for samples that are annotated as positive at $\boldsymbol{C_2}$ while the orange one denotes the average probability changes for samples that are annotated as negative at $\boldsymbol{C_2}$. $\boldsymbol{C_1}$ denotes the time-step of the 5th visit, $\boldsymbol{C_2}$ denotes the time-step of the 6th visit, and $\boldsymbol{\Delta t = C_2 - C_1}$.

V. Conclusion

In this paper, we present the first transformer-based glaucoma forecast network for irregularly sampled sequential fundus images named GLIM-Net. Instead of using linear transformation as input embedding method, we utilize a convolutional neural network to embed the input into specific dimension to address the problem that transformer architecture needs large-scale training data. We propose time positional encoding and time-sensitive MSA to harness the challenge that the time interval of fundus images in a sequence varies from the minimum of only one day to the maximum of 13 years. To make our model more practical in real life, both input and output are concatenated with a time condition so as to output the probability of developing glaucoma conditioned on the time we feed to the model. Extensive experiments on the SIGF dataset demonstrate that our proposed GLIM-Net greatly outperforms other state-of-the-art methods in the glaucoma forecast task. Furthermore, we prove that the transformer architecture, which has achieved great success recently due to its effectiveness in traditional computer vision tasks, can also learn representative features of fundus images and can provide an idea of employing transformer architectures in the medical imaging area. Besides, extensive experiments prove that our two proposed modules can help to better capture temporal information of sequential data, which may provide a good reference for the problems with similar data structures as that of our data. A limitation of our current work is that the dataset used has a limited number of glaucoma cases for training. It would be better to collect more cases for training the model with better generalization ability in the future.

References

[1] Y.-C. Tham, X. Li, T. Y. Wong, H. A. Quigley, T. Aung, and C.-Y. Cheng, “Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis,” *Ophthalmology*, vol. 121, no. 11, pp. 2081–2090, 2014.

[2] G. D. Joshi, J. Sivaswamy, and S. Krishnadas, “Optic disk and cup segmentation from monocular color retinal images for glaucoma assessment,” *IEEE transactions on medical imaging*, vol. 30, no. 6, pp. 1192–1205, 2011.

[3] F. Yin, J. Liu, S. H. Ong, Y. Sun, D. W. Wong, N. M. Tan, C. Cheung, M. Baskaran, T. Aung, and T. Y. Wong, “Model-based optic nerve head segmentation on retinal fundus images,” in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2011, pp. 2626–2629.

[4] J. Cheng, J. Liu, Y. Xu, F. Yin, D. W. K. Wong, N.-M. Tan, D. Tao, C.-Y. Cheng, T. Aung, and T. Y. Wong, “Superpixel classification based optic disc and optic cup segmentation for glaucoma screening,” *IEEE transactions on medical imaging*, vol. 32, no. 6, pp. 1019–1032, 2013.

[5] J. Cheng, D. Tao, D. W. K. Wong, and J. Liu, “Quadratic divergence regularized SVM for optic disc segmentation,” *Biomedical optics express*, vol. 8, no. 5, pp. 2687–2696, 2017.

[6] X. Chen, Y. Xu, S. Yan, D. W. K. Wong, T. Y. Wong, and J. Liu, “Automatic feature learning for glaucoma detection based on deep learning,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 669–677.

[7] J. I. Orlando, E. Prokofyeva, M. del Fresno, and M. B. Blaschko, “Convolutional neural network transfer for automated glaucoma identification,” in *12th international symposium on medical information processing and analysis*, vol. 10160. International Society for Optics and Photonics, 2017, p. 101600U.

[8] L. Li, X. Wang, M. Xu, H. Liu, and X. Chen, “DeepGF: Glaucoma forecast using the sequential fundus images,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 626–635.

[9] Y. Xu, Q. ZHANG, J. Zhang, and D. Tao, “ViTAE: Vision transformer advanced by exploring intrinsic inductive bias,” in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 28 522–28 535.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations*.

[11] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[13] R. Thomson, D. Luettel, F. Healey, and S. Scobie, “Safer care for the acutely ill patient: learning from serious incidents,” *London: National Patient Safety Agency*, 2007.

[14] Y. Cheng, F. Wang, P. Zhang, and J. Hu, “Risk prediction with electronic health records: A deep learning approach,” in *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 2016, pp. 432–440.

[15] N. Tomašev, X. Glorot, J. W. Rae, M. Zielinski, H. Askham, A. Saraiva, A. Mottram, C. Meyer, S. Ravuri, I. Protsyuk *et al.*, “A clinically applicable approach to continuous prediction of future acute kidney injury,” *Nature*, vol. 572, no. 7767, pp. 116–119, 2019.

[16] F. Ali, S. El-Sappagh, S. R. Islam, D. Kwak, A. Ali, M. Imran, and K.-S. Kwak, “A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion,” *Information Fusion*, vol. 63, pp. 208–222, 2020.

[17] B. Jin, C. Che, Z. Liu, S. Zhang, X. Yin, and X. Wei, “Predicting the risk of heart failure with ehr sequential data modeling,” *Ieee Access*, vol. 6, pp. 9256–9261, 2018.

[18] H. Li, M. Habes, D. A. Wolk, Y. Fan, A. D. N. Initiative *et al.*, “A deep learning model for early prediction of Alzheimer’s disease dementia based on hippocampal magnetic resonance imaging data,” *Alzheimer’s & Dementia*, vol. 15, no. 8, pp. 1059–1070, 2019.

[19] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[21] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” 2018.

[22] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[23] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, “Xlnet: generalized autoregressive pretraining for language understanding,” in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 5753–5763.

[24] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, “Stand-alone self-attention in vision models,” in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 68–80.

[25] J.-B. Cordonnier, A. Loukas, and M. Jaggi, “On the relationship between self-attention and convolutional layers,” in *International Conference on Learning Representations*.

[26] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

[27] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.

[28] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng, H. Shen, and H. Xia, “End-to-end video instance segmentation with transformers,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8741–8750.

[29] Y. Xiong, B. Du, and P. Yan, “Reinforced transformer for medical image captioning,” in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2019, pp. 673–680.

[30] T. Prangemeier, C. Reich, and H. Koeppl, “Attention-based transformers for instance segmentation of cells in microstructures,” in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2020, pp. 700–707.

[31] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, “Medical transformer: Gated axial-attention for medical image segmentation,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, Eds. Cham: Springer International Publishing, 2021, pp. 36–46.

[32] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, “Swin-Unet: Unet-like pure transformer for medical image segmentation,” *arXiv preprint arXiv:2105.05537*, 2021.

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

[34] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” *arXiv preprint arXiv:1607.06450*, 2016.

[35] H. Fu, J. Cheng, Y. Xu, C. Zhang, D. W. K. Wong, J. Liu, and X. Cao, “Disc-aware ensemble network for glaucoma screening from fundus image,” *IEEE transactions on medical imaging*, vol. 37, no. 11, pp. 2493–2501, 2018.

[36] R. Gao, Y. Tang, K. Xu, Y. Huo, S. Bao, S. L. Antic, E. S. Epstein, S. Deppen, A. B. Paulson, K. L. Sandler *et al.*, “Time-distanced gates in long short-term memory networks,” *Medical image analysis*, vol. 65, p. 101785, 2020.

[37] A. Krizhevsky and G. Hinton, “Learning multiple layers of features from tiny images,” University of Toronto, Toronto, Ontario, Tech. Rep. 0, 2009.

[38] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, “TensorFlow: Large-scale machine learning on heterogeneous systems,” 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[39] D. P. Kingma and J. Ba, “ADAM: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[40] X. Chen, Y. Xu, D. W. K. Wong, T. Y. Wong, and J. Liu, “Glaucoma detection based on deep convolutional neural network,” in *2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 2015, pp. 715–718.

[41] L. Li, M. Xu, X. Wang, L. Jiang, and H. Liu, “Attention based glaucoma detection: A large-scale database and CNN model,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 571–10 580.

[42] R. Santeramo, S. Withey, and G. Montana, “Longitudinal detection of radiological abnormalities with time-modulated LSTM,” in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 326–333.

[43] S. Yu, K. Ma, Q. Bi, C. Bian, M. Ning, N. He, Y. Li, H. Liu, and Y. Zheng, “MIL-VT: Multiple instance learning enhanced vision transformer for fundus image classification,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 45–54.

[44] A. He, T. Li, N. Li, K. Wang, and H. Fu, “CABNet: Category attention block for imbalanced diabetic retinopathy grading,” *IEEE Transactions on Medical Imaging*, vol. 40, no. 1, pp. 143–153, 2021.

[45] M. Juneja, S. Thakur, A. Uniyal, A. Wani, N. Thakur, and P. Jindal, “Deep learning-based classification network for glaucoma in retinal images,” *Computers and Electrical Engineering*, vol. 101, p. 108009, 2022.