Applied Soft Computing Journal 155 (2024) 111431

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

# A spatiotemporal convolution recurrent neural network for pixel-level peripapillary atrophy prediction using sequential fundus images

Mengxuan Li [a], Weihang Zhang [a], He Zhao [a], Yubin Xu [a], Jie Xu [b], Huiqi Li [a,*]

[a] Beijing Institute of Technology, Beijing, 100081, China
[b] Beijing Institute of Ophthalmology, Beijing Tongren Hospital, Capital Medical University, Beijing, 100005, China

ARTICLE INFO

*Keywords:*
Peripapillary atrophy prediction
Sequential fundus images
Temporal memory
Spatiotemporal prediction
Scheduled sampling

ABSTRACT

The progression of peripapillary atrophy (PPA) is closely associated with the development of retinal diseases such as myopia and glaucoma. PPA prediction employing longitudinal images to obtain its progress trend can facilitate personalized treatment. Although existing studies have attempted to predict the persistence of PPA, such studies cannot provide quantitative measurement for personalized treatment. In this paper, we propose a spatiotemporal framework for pixel-level PPA prediction using sequential fundus images, including feature extractor, temporal memory, and spatiotemporal prediction modules. To take advantage of historical information, a temporal memory module is used, integrating current and prior features to build sequential data of features. To further enhance the prediction performance, the recurrent neural network states in a spatiotemporal prediction module transmit between different layers, enabling high-level states to guide the learning of low-level states. To handle missing data in clinical follow-up data, we use the predicted output of the spatiotemporal prediction module to substitute the missing data, and the scheduled-sampling strategy is employed in training. Extensive experiments conducted using a clinical dataset demonstrate that our proposed method achieves a satisfactory performance compared with the start-of-the-art models. The proposed approach can be applied using clinical data to obtain various quantitative indicators for personalized treatment and prevention of retinal disease.

## 1. Introduction

Peripapillary atrophy (PPA) is associated with chorioretinal thinning and the disruption of retinal pigment epithelium, which can be divided into $\alpha$ and $\beta$ zones [1,2]. $\alpha$-PPA is defined as irregular pigmentation that is located circumferentially away from the optic disc (OD) [3]. $\beta$-PPA is adjacent to OD and is characterized by the thinning of the chorioretinal tissue with the visible sclera and choroidal vessels [4,5]. Several studies [5–9] have found that the morphology change of $\beta$-PPA is positively correlated to the progression of myopia and glaucoma, both of which may cause visual impairment or even blindness. By analyzing the $\beta$-PPA regions in fundus images, ophthalmologists can assess and track related fundus diseases accurately. Currently, the clinical diagnosis of $\beta$-PPA regions still relies on manual annotation by experienced ophthalmologists, whereas the prediction of $\beta$-PPA progression requires ophthalmologists to analyze the labeled regions in sequential data according to their experience. If the progression of the $\beta$-PPA region can be predicted automatically in a computer-assisted manner, ophthalmologists can detect the related diseases at an early stage for timely treatment, which helps prevent visual impairment and minimize the socioeconomic cost caused by the related diseases [10].

The progression of $\beta$-PPA is a gradual process; thus, the temporal information for its accurate prediction should be utilized. Only sequential data inherently present temporal information, and signal-moment images only convey transient information. Therefore, we propose to use the sequential fundus images for $\beta$-PPA (abbreviated as PPA in this paper) prediction.

In recent years, a few methods have been proposed for PPA prediction. Li et al. [11] proposed a causal hidden Markov model for PPA prediction, in which the hidden variables that propagate to generate medical observation are introduced. Wu et al. [12] exploited the irreversibility of prior and progression learning to predict PPA. These two methods only predict whether there will be PPA regions at the image level, lacking the capability to provide pixel-level information crucial for analyzing disease progression. For other ocular disease prediction, Pham et al. [13] proposed a generative adversarial network-based method to determine the progression of age-related macular degeneration by comparing the predicted drusen mask of the generated

* Corresponding author.
*E-mail address:* huiqili@bit.edu.cn (H. Li).

https://doi.org/10.1016/j.asoc.2024.111431
Received 9 August 2023; Received in revised form 4 February 2024; Accepted 19 February 2024
Available online 2 March 2024

future fundus image with the ground truth. Although the dataset used in this method is longitudinal, the temporal information is not as fully explored as the methods in previous studies [11,12]. To address this problem and provide comprehensive information for predicting disease progression using longitudinal data, we explore to use of the spatiotemporal prediction model.

Because the input and output are both spatiotemporal sequential data, PPA-region prediction can be regarded as a spatiotemporal prediction task, in which temporal and spatial features must be considered. The algorithms used for spatiotemporal prediction can be divided into two categories: convolution recurrent neural network (ConvRNN)-based methods [14–20] and transformer-based methods [21,22]. ConvRNN-based methods sequentially process data by combining the capabilities of the convolutional neural network and recurrent neural network (RNN) with a hybrid structure, allowing the predicted results to supplement missing data in the input sequence. Transformer-based methods handle historical input in parallel through transformer blocks for self-attention learning. However, they fail to effectively handle missing data. Although transformer-based methods capture the long-range dependencies better than the ConvRNN-based methods, they involve higher computational complexity and cost. Because missing data are a common phenomenon in clinical scenarios, we anticipate utilizing the recurrent input structure of the ConvRNN-based methods to address the challenge of incomplete input data in PPA-region prediction.

Currently, most studies have focused on improving the ConvRNN unit in spatiotemporal sequence prediction, such as ConvLSTM [14], ConvGRU [15], Trajectory GRU [15], PredRNN [16], PredRNN++ [17], SA-JSTN [18], CSA-ConvLSTM [19], and PredRANN [20]. In terms of the overall prediction framework, two categories can be obtained: one involves stacking multiple layers of ConvRNN units [14,20,23–25] and the other encodes all input through stacked ConvRNN units and reverses the connection order of RNN states during predicting (which can guide the learning of low-level states with high-level states) [15,19,26]. In spatiotemporal prediction, the distributions of input and output data exhibit strong spatial similarity and highly correlated underlying changes may exist in the temporal domain. Therefore, increasing the dependence of predictions on memory representation learned by recurrent modules across different layers becomes important. However, only spatiotemporal memory can be transmitted between different layers in a zigzag direction in the existing frameworks. Other memory states, such as the cell memory in ConvLSTM, are typically limited to transmitting within the same layer. We facilitate the transmission of the RNN memory states between different layers in our method to enhance the performance of spatiotemporal prediction.

In this study, we propose a spatiotemporal framework that can use sequential fundus images for pixel-level PPA prediction to provide information for clinical treatment even in the presence of missing data. The main contributions of this study can be summarized as follows:

1. We introduce a novel spatiotemporal framework that can predict future PPA regions at a pixel level using longitudinal fundus images, which can assist ophthalmologists in formulating personalized treatment plans for patients with related fundus diseases. A temporal memory module (TMM) is proposed to fully utilize historical information. Moreover, a spatiotemporal prediction module based on ConvRNN is proposed, where the RNN states transmit between different layers to improve the prediction performance.
2. A scheduled-sampling strategy is introduced to network training, overcoming the challenge of missing input data and improving the model performance.
3. Comprehensive experiments are conducted to evaluate our approach. Results show that our method can achieve good performance for pixel-level PPA prediction, especially when some input data are missing.

## 2. Method

### 2.1. Overview

We propose a spatiotemporal prediction framework that utilizes longitudinal data to predict the future PPA region at a pixel level. PPA primarily appears in the vicinity of OD, and not all fundus images contain this region. Furthermore, the PPA region grows progressively, and its size is small at the early stage, posing challenges for network training based on the PPA region only. By contrast, OD exhibits a regular shape and size, which makes it easy for model learning. To enhance the performance of PPA prediction, OD information is incorporated into the prediction task for a stable performance.

Fig. 1 illustrates the framework of the proposed method. The training process is as follows. The input images are first encoded using the feature extractor module, which learns the segmentation features of PPA and OD. This operation helps mitigate the influence of inconsistency on imaging across different timelines and ensure the robustness of network training. Because missing data exist in actual clinical scenarios, we involve a scheduled-sampling strategy to simulate the missing data by hiding some of the actual features randomly during training. Specifically, when a hidden operation is performed on the data at a certain moment, the hidden feature is replaced by the output of the spatiotemporal prediction module at the corresponding moment. As the proposed framework is based on ConvRNN, data are sequentially input into the network. To fully utilize the long-term dependencies in the temporal dimension, the output features processed by a scheduled-sampling operation from the initial moment until the current input state are all inputted into the TMM. Then, the output features are fed into the spatiotemporal prediction module, which is the core component of the proposed framework. The spatiotemporal prediction module predicts future features based on historical data. Specifically, this module takes the input sequence $\{F_1, F_2, \dots, F_T\}$ and produces the output sequence $\{\hat{F}_2, \hat{F}_3, \dots, \hat{F}_{T+1}\}$, where $\hat{F}_{T+1}$ represents the future output feature. Because we aim to predict the PPA region, it is crucial to process the predicted features via a segmentation head (a convolution cascaded after the concatenating features) to obtain the final PPA segmentation. The loss function comprises two components: the feature loss and segmentation loss. The feature loss is calculated based on the difference between input features and prediction features. The segmentation loss is calculated based on the difference between the ground truth and the output of the segmentation head. Model training is accomplished by optimizing this loss function. During testing in real applications, scheduled-sampling is excluded from the model, while the other parts remain unchanged. When data are missing, the predicted result of the spatiotemporal prediction module at the corresponding moment is used to substitute the missing input.

### 2.2. Network architecture

#### 2.2.1. Feature extractor module

Instead of training directly using images, we use the pretraining strategy to obtain the feature maps first and then use them as the input for the subsequent modules. Directly using original images as input can lead to various issues. First, owing to different conditions of data collection, the style of fundus images can be inconsistent, which may adversely affect the performance of the network. Second, it is difficult for the prediction network to focus on the features of PPA and OD. Fundus images are complicated and contain various structures, increasing the difficulty of network training, resulting in poor prediction of PPA and OD. Moreover, owing to different distributions of input and output, when some input data are missing, the predicted PPA segmentation cannot be used directly to compensate for the missing input retinal image. To handle the abovementioned issues, a feature extractor module is used. Specifically, we train a segmentation network using $(X_i, Y_i)$, where $X_i$ and $Y_i$ are the $i$th image and annotation labeled

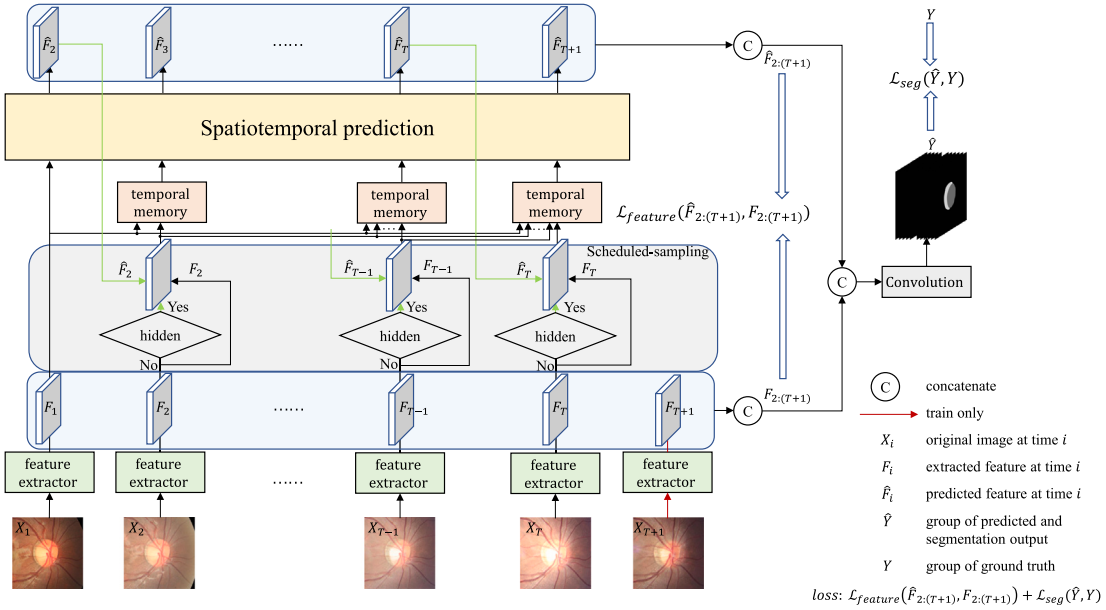


**Fig. 1.** Diagram of the proposed pixel-level PPA region-prediction framework.

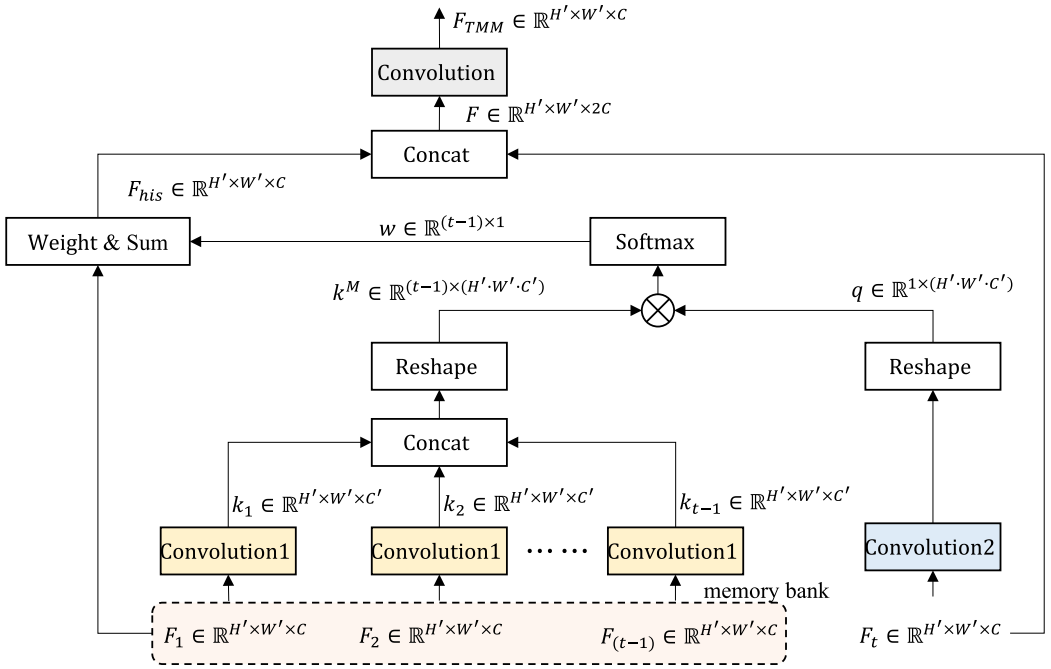


**Fig. 2.** Structure of the temporal memory module. The features in the memory bank are historical features, and $F_i$ is the input feature at the current moment.

by ophthalmologists, respectively. Notably, the annotation contains two objects: the PPA region and the OD region. We choose Unet [27] as the segmentation structure. After pretraining, the feature extractor module is constructed by removing the last convolutional layer, which is originally used to project features to the number of categories for each pixel. If the last convolution layer is not removed, the channel of features output by the feature extractor requires expansion using convolution operations to enrich the discriminative information, which is equivalent to adding two unnecessary convolution operations into the overall network. The parameters of this module are fixed during subsequent usage to reduce the risk of forgetting previously learned knowledge and reduce the requirement for computational resources.

### 2.2.2. Temporal memory module

Because the progression of PPA is a gradual process, fully leveraging the temporal information can promote PPA prediction. To effectively utilize long-term dependencies along the temporal dimension while following the causality in sequential data, we introduce the TMM. This module inputs the features processed by the scheduled-sampling strategy from the initial until the current time $(F_1, \ldots, F_t)$, enabling the network to retain information on past time steps. The TMM is used starting from the second input because the historical information is available only at the beginning of the second input. The specific operation of the TMM is shown in Fig. 2. When $F_t$ is input, the previous $t-1$ features are sent to the memory bank. Each moment in the

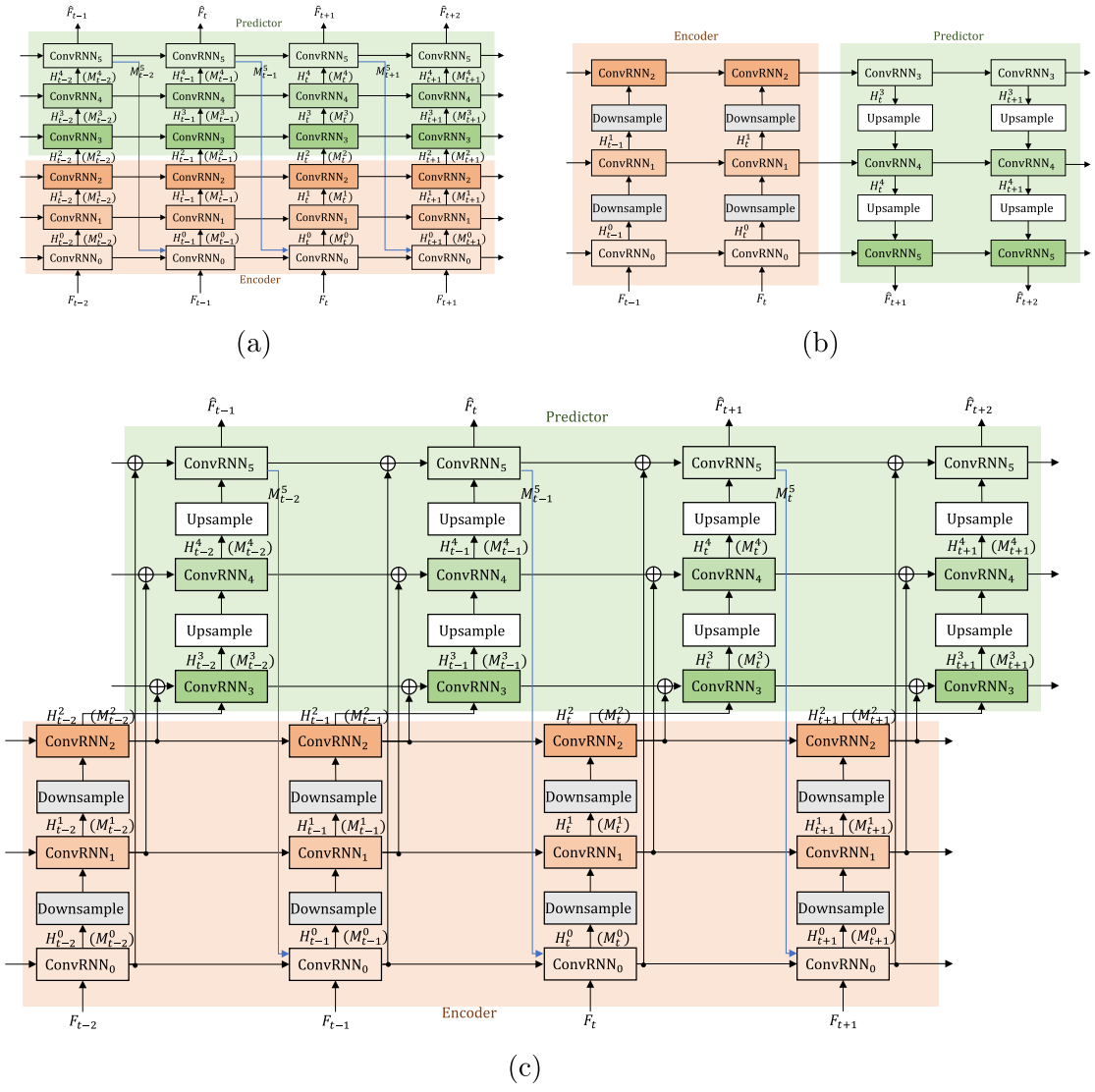


**Fig. 3.** Spatiotemporal prediction structure. (a) The structure proposed in Ref. [14]. (b) The structure proposed in Ref. [15]. (c) Our proposed spatiotemporal prediction module. $H$ represents the hidden state, and $M$ indicates the spatiotemporal memory state.

memory bank is operated by the same convolution layer to generate the key $k_i \in \mathbb{R}^{H' \times W' \times C'}$, which is used for addressing. $H'$, $W'$, and $C'$, representing the height, width, and number of channels of the feature map, respectively. Then, the keys $(k_1, \ldots, k_{t-1})$ calculated at different moments are concatenated together and reshaped, affording $k^M \in \mathbb{R}^{(t-1)\times(H' \cdot W' \cdot C')}$, where $C' = C/4$ owing to the dimensionality reduction during convolutional operation to reduce the amount of calculation and $C$ is the channel number of original features. The current input $F_t$ is used as the query feature, which is performed by another convolution layer and reshaped to generate the query $q \in \mathbb{R}^{1\times(H' \cdot W' \cdot C')}$.

Because each historical moment has a distinct impact on the present input, it is crucial to assess the significance of every previous input by evaluating the similarity between the query and each key. By calculating and normalizing the similarity scores using Formula (1), we can obtain the corresponding weight assigned to each key. A higher weight signifies a more pronounced impact of the historical moment on the current input.

$$w = softmax\left(\frac{k^M q^T}{\sqrt{t-1}}\right) \tag{1}$$

The obtained weights can be used to weigh and sum the features present in the original memory bank for producing the integrated historical feature $F_{his} \in \mathbb{R}^{H' \times W' \times C}$. Subsequently, the obtained feature is concatenated with the query feature to get $F \in \mathbb{R}^{H' \times W' \times 2C}$. Finally, the dimensionality reduction operation is performed to obtain the final output feature $F_{TMM} \in \mathbb{R}^{H' \times W' \times C}$, which is sequentially inputted to the spatiotemporal prediction module.

### 2.2.3. Spatiotemporal prediction module

In spatiotemporal prediction tasks, a common method to improve model representation is to stack multiple layers of the ConvRNN units to expand the convolutional domain. ConvLSTM [14] is a fundamental technique for spatiotemporal sequence prediction, which incorporates convolution operations into long short-term memory (LSTM) [28] to capture spatial context information. Some similar network structures have been proposed by replacing LSTM with different RNN structures [29,30], such as ConvGRU and Trajectory GRU proposed in Ref. [15]. However, the transmission of spatial information across different layers may be inadequate when using the abovementioned ConvRNN units, thus hindering the effective propagation of information. To address this issue, PredRNN [16] and PredRNN++ [17] introduce an additional global spatial memory to preserve the spatial features of each layer. Although these methods can capture spatial and temporal information simultaneously, they rely on modeling local context information obtained through convolutional operations for capturing spatial information. To capture local and global spatial features

simultaneously, a self-attention mechanism has been incorporated into the ConvRNN unit in previous studies [18,19].

Existing methods primarily focus on optimizing the ConvRNN unit. When considering information interaction between different layers, current research mainly emphasizes spatial information and ignores memory information. In spatiotemporal learning, the distributions of input and output data are very similar in the spatial domain, so there may be highly correlated underlying changes in the temporal domain. Thus, enabling the transmission of memory information between different layers is very important. The overall framework of current spatiotemporal prediction research can be summarized into two categories, which are shown in Fig. 3(a) and (b). The method of directly stacking multiple layers of the ConvRNN units (stack-ConvRNN) [14] (Fig. 3(a)) can output predicted results for each timestamp but cannot fully utilize the memory information across different layers, in which only spatiotemporal memory $M^5$ can be transmitted between different layers along a zigzag direction (blue arrow in Fig. 3), while other memories can only be transmitted within the same layer. The structure shown in Fig. 3(b) is the enhanced version [15] of Fig. 3(a) where the order of memory information in the predictor is reversed to enable the high-level states to guide the learning of low-level states (reverse-ConvRNN). Meanwhile, downsampling and upsampling operations are added to reduce the amount of calculation and memory usage. However, the structure presented in Fig. 3(b) also has the following limitations: (1) the structure first encodes all inputs and then performs predicting operations, resulting in the incapability to obtain predicted outputs for all timestamps. Therefore, this structure fails to handle missing data, as the common solution is to substitute the missing data with the predicted outputs. (2) The structure fails to fully utilize memory information between different layers at all timestamps, as it only achieves the transmission of all memory states between different layers when predicting the timestamp $t+1$.

To address the abovementioned limitations, we propose a spatiotemporal prediction module inspired by the structures presented in Fig. 3(a) and (b). The proposed module is shown in Fig. 3(c), in which the advantages of the predicted results for each timestamp and the memory information transmitted between different layers are combined. Specifically, the first three layers are encoders that encode the feature of the current moment, and the last three layers are the predictors that predict the feature of the next moment. This structure enables predictions for all timestamps, making the structure suitable for scenarios with missing input data. Moreover, for the ConvRNN units in the encoder, the RNN memory states of the horizontal input are the output of the previous moment, while for the ConvRNN units in the predictor, the horizontal input is the sum of the output at the previous moment and the output of the corresponding encoder layer at the current moment. The connection of the predictor enables the horizontally transmitted memory information to be fully utilized between different layers.

### 2.3. Scheduled-sampling

To address the challenge of missing input data in actual clinical scenarios, commonly, the missing input is substituted with the predicted output of the spatiotemporal prediction module during inference. However, if missing data are ignored in the training phase, the error of the previous output might rapidly accumulate, affecting the prediction of follow-up sequences. To address this issue, a sampling strategy is introduced during training, which can simulate the clinical situation and force the model to learn more about long-term information to improve its generalization. The specific operations are as follows, which are inspired by Refs. [31,32]. First, some real inputs are randomly hidden for the sequence data except the initial input. Then, the output of the spatiotemporal prediction module is used to replace the hidden data as the input of the TMM; the replacement is indicated by the green arrows marked in Fig. 1. We assumed that the probability of the hidden operation for the sequence data is $p$. At the beginning of training, $p$ is set to 0. When the network learning tends to be stable, $p$ is gradually increased to simulate the clinical situation of missing data.

### 2.4. Loss function

The loss function for optimizing the entire network can be expressed using Eq. (2), which comprises two parts: feature consistency loss $\mathcal{L}_{feature}$ and segmentation loss $\mathcal{L}_{seg}$.

$$\mathcal{L} = \mathcal{L}_{feature} + \mathcal{L}_{seg} \tag{2}$$

**Feature consistency loss:** This loss imposes constraints between features output by the feature extractor module and the predicted features [33]. In practical applications, it is common to encounter missing data at certain timestamps. When performing consistency constraints on the features of each timestamp, the output of the spatiotemporal prediction module can replace the missing input, addressing the issue of missing data. The feature consistency loss can be formulated as follows:

$$\mathcal{L}_{feature} = \frac{1}{N}\sum_{i=1}^{N}(F_{i,2:(T+1)} - \hat{F}_{i,2:(T+1)})^2 \tag{3}$$

where $F_{i,2:(T+1)}$ represents the features output by the feature extractor from time 2 to $T+1$ of the data group $i$ and $\hat{F}_{i,2:(T+1)}$ indicates the features output by the spatiotemporal prediction module from time 2 to $T+1$ of the data group $i$. $N$ is the batch size.

**Segmentation loss:** This loss is calculated based on the difference between the ground truth and the output of the segmentation head, which is a combination of cross-entropy loss and dice loss [34,35], as described in Eq. (4). Combining these two loss functions allows the model to focus on classification accuracy and segmentation accuracy. The segmentation task is equivalent to performing multiple classifications on each pixel. Cross-entropy loss is a standard loss function used for multiclass classification problems, ensuring accurate pixel-level classification. The PPA region occupies a small proportion compared with the background within the image, leading to a class imbalance issue. The utilization of the dice loss helps handle this imbalance, thus improving the model performance.

$$\mathcal{L}_{seg} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{Cls} Y_{i,j} log(\hat{Y}_{i,j}) + \frac{1}{Cls-1}\sum_{j=2}^{Cls}\left(1 - \frac{2\times\sum_{i=1}^{N}|Y_{i,j}\cap\hat{Y}_{i,j}|}{\sum_{i=1}^{N}|Y_{i,j}| + |\hat{Y}_{i,j}|}\right) \tag{4}$$

where $Cls$ represents the number of categories, which is set to 3 in our application, including background, PPA, and OD. Only the foreground is calculated to determine the dice loss. Because there is no PPA region in some images, the dice loss is computed on a per-batch basis. According to the scheduled-sampling strategy, $\hat{Y}_i$ contains two results for the data group $i$: the segmentation result obtained by decoding the output of the feature extractor and the prediction result obtained by decoding the output of the spatiotemporal prediction module. $Y_i$ represents the group truth of the data group $i$, in which moments correspond to those in $\hat{Y}_i$.

The pseudo-code of our method for real applications is shown in Algorithm 1.

## 3. Experiments

### 3.1. Dataset description

The dataset used in following experiments is provided by the Beijing Institute of Ophthalmology, Beijing Tongren Hospital; the data were collected from the same group of primary school students from grade one (age $6.3\pm0.4$ years, 2011) to six (2016) in Dongcheng and Huairou Districts, Beijing, China. All the fundus images were acquired by a 45°, CR-DGI camera, Canon Inc, Tokyo, Japan, in 2011–2013, and a 45°, CR-II camera, Canon Inc, Tokyo, Japan, in 2014–2016. The collected data were grouped based on individual patients with a label indicating left or right eyes. Only the groups with complete data for 6 years were

**Algorithm 1** The practical application of our method.

**Input:**
Input sequential images: $\{X_1, \cdots, X_T\}$ If $X_i (i \in [2, T])$ is missing, it is set to None.
Initialization parameters:
$H$ : Hidden state
$C$ : Memory state
Variable definitions:
$\mathcal{M}_{Enc}$ : Feature extractor module
$\mathcal{M}_{TMM}$ : Temporal memory module
$\mathcal{M}_{ConvRNN}$ : Spatiotemporal prediction module
$\mathcal{M}_{Seg}$ : Segmentation head
**Output:**
Predicted PPA regions: $\{\hat{Y}_2, \cdots, \hat{Y}_{T+1}\}$
**for** $i$ in $[1, T]$ **do**
  **if** $i == 1$ **then**
    $F_i = \mathcal{M}_{Enc}(X_i)$
    $\hat{F}_{i+1}, H, C = \mathcal{M}_{ConvRNN}(F_i, H, C)$
  **else**
    **if** $X_i$ is None **then**
      $F_i = \hat{F}_i$
    **else**
      $F_i = \mathcal{M}_{Enc}(X_i)$
    **end if**
    $F_{TMM} = \mathcal{M}_{TMM}([F_1, \cdots, F_{i-1}], F_i)$
    $\hat{F}_{i+1}, H, C = \mathcal{M}_{ConvRNN}(F_{TMM}, H, C)$
  **end if**
  $\hat{Y}_{i+1} = \mathcal{M}_{Seg}(\hat{F}_{i+1})$
**end for**

retained, amounting to a total of 342 groups. The resolutions of the original fundus images include 2592 × 3888, 1696 × 2544, 1728 × 2592, and 1556 × 1924. We selected 250 groups as the training set, 42 groups as the validation set, and the remaining 50 groups as the test set. When evaluating the situation of missing input data in the follow-up sequences, we randomly pruned data from the test set, where one image was missing from 25 groups and two images were missing from the other 25 groups. The missing data were selected from grades 2 to 5. In the following experiments, we refer to the test set without missing data as Dataset 1 and the test set with missing data as Dataset 2. The imaging condition of fundus images collected at different moments is not exactly the same.

To eliminate this effect and facilitate the intuitive observation of changes in the PPA region, registration was performed for each group. We treated the first-year image as the standard and registered other images using a partial intensity invariant feature descriptor [36]. As PPA primarily appears in the vicinity of OD, we used the region of interest (ROI) area that focused on the region surrounding OD for the following experiments to reduce the complexity of model learning. The ROI area was extracted according to the pretrained segmentation model followed by a resize operation to 512 × 512. One example of the original input is shown in Fig. 4(a), and the extracted ROI region after registration is shown in Fig. 4(b).

### 3.2. Implementation details

Because each group in the dataset contains 6 years of clinical data, the fundus images of the first 5 years were used in the experiment to predict the region of PPA in the sixth year. The code of our proposed method was implemented using PyTorch. The models were trained on an AMD Ryzen Threadripper 3960X 24-Core Processor and a NVIDIA RTX 3090 24G GPU. For all experiments, 100 epochs were trained. The initial learning rate was set to 0.0003, and the cosine annealing strategy was used. The epoch number and learning rate were determined via experiments. An excessively large learning rate could lead to training instability, while a small learning rate might result in a slow training process. The cosine annealing strategy could prevent model overfitting and enhance model stability. We followed existing spatiotemporal prediction methods [15–20] to use the Adam optimizer. The batch size comprised four groups of data that could fully utilize the hardware resources of GPU. When using the scheduled-sampling strategy, we started with $p = 0$ for the first 30 epochs, increased $p$ to 0.5 linearly from epoch 30 to 60, and then kept $p$ at 0.5 after epoch 60. The parameter setting was determined based on the validation loss obtained during training without scheduled-sampling. When the epoch number was 30, the decline rate of loss tended to be gentle and the model gradually became stable. When the epoch number exceeded 90, the validation loss showed minimal fluctuation. The epoch range 30–90 contained two phases: linearly increasing $p$ and maintaining $p$ at a constant value. We distributed these 60 epochs equally to accommodate both the phases. For data augmentation, we applied random left–right flipping, random Gaussian noise, random motion blur, and random brightness and contrast change of the input image. In the test phase, the model with the lowest loss on the validation set was selected.

The $F1\ score$, $Precision$, $Recall$, and $Accuracy$ were used to evaluate the performance of prediction results. The performance evaluation metrics were calculated using the following formulas:

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

$$F1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{8}$$

where $TP$ refers to the number of pixels of the PPA area labeled as PPA, $FP$ denotes the number of pixels of background labeled as PPA, $FN$ represents the number of pixels of PPA masked as background, and $TN$ indicates the number of pixels in background labeled as background. Because $Precision$ and $Recall$ are often mutually restrictive, the $F1\ score$ can comprehensively consider the abovementioned two metrics. Therefore, the $F1\ score$ is the main metric for evaluating the quality of models. We also ran a statistical test using the Mann–Whitney U test [37], and we consider that the results are significantly different when the $p$-value is less than 0.05 (i.e., $p < 0.05$).

### 3.3. Comparison with other methods

Currently, only the methods proposed in Refs. [11,12] are reported for PPA prediction; however, these two methods are used to predict the existence of PPA in a future stage and cannot be refined to pixel-level prediction. Moreover, because the data used in these methods contain image data as well as clinical measurements, our method cannot be compared fairly with them. As pixel-level PPA prediction is a spatiotemporal prediction method in essence, we compare our method with the ConvRNN-based methods including stack-ConvLSTM [14] and reverse-ConvLSTM [15], as well as transformer-based methods, including encoder–decoder transformer [21] and encoder–decoder Swin transformer [22].

In this comparison study, ConvLSTM is selected as the ConvRNN unit in our method. Table 1 presents the comparison of the results of different spatiotemporal prediction methods using Dataset 1, in which no missing data exist. To compare the quantitative results of different methods intuitively, we draw a multicriterion graph of each metric presented in Table 1, as shown in Fig. 5(a). Because the $F1\ score$ is a comprehensive metric, it is considered to be the highest priority metric when evaluating the model performance. The ConvRNN-based methods are shown to generally perform better than the transformer-based

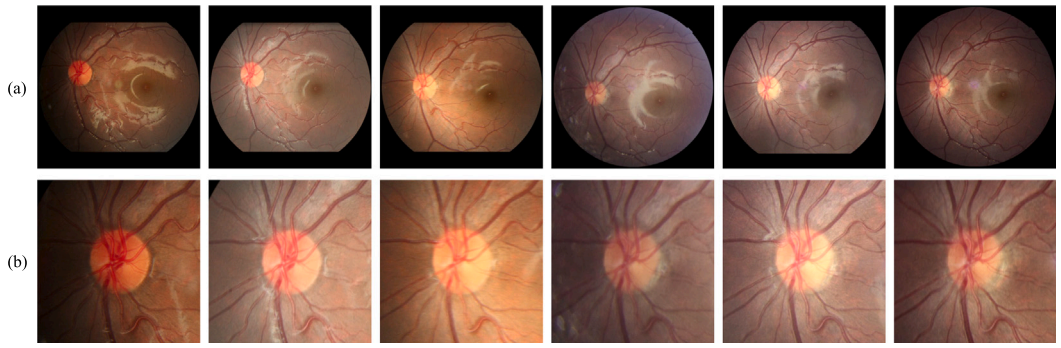


**Fig. 4.** Data preprocessing. (a) Original sequential images of an individual patient. (b) The corresponding extracted ROI after registration.

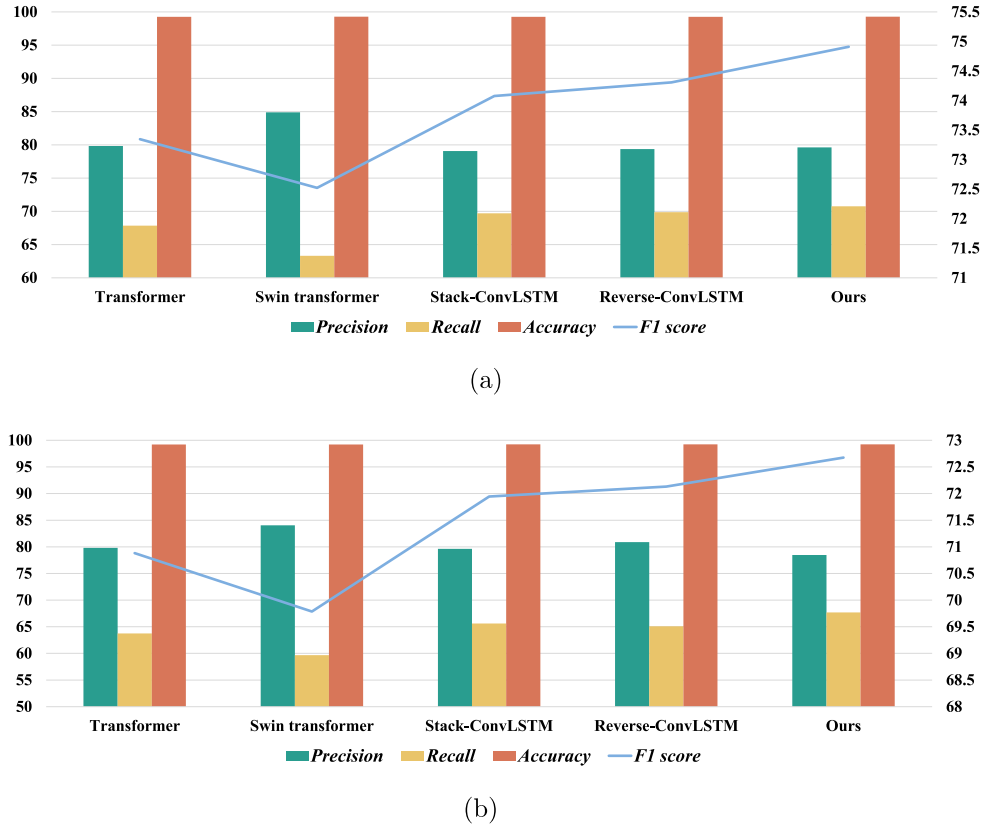


**Fig. 5.** Multicriterion graphs obtained for a comparison study. The highest priority metric *F1 score* is represented by the line chart, with the corresponding scale (%) shown on the right axis. Other metrics are represented by the bar chart, with the corresponding scale (%) shown on the left axis. (a) Results obtained using Dataset 1. (b) Results obtained using Dataset 2, in which the first four methods use interpolation to simulate missing data.

methods. Compared with stack-ConvLSTM and reverse-ConvLSTM, our method not only transmits memory information between different layers but also completely utilizes temporal information, leading to the best performance among these methods. Fig. 6 shows the visual comparison between our method and others using Dataset 1. The predictions obtained by our method are closer to the ground truth.

In actual clinical applications, follow-up data often have missing data. Therefore, we also compare our method with different spatiotemporal prediction methods using Dataset 2 comprising missing data. The issue of missing input data has not yet been considered in the comparison methods presented in Table 1. In clinical research, this situation is commonly addressed by replacing the missing data via means of interpolation based on adjacent data. In our proposed method, missing data are simulated during training instead of interpolation of input data. Our method uses all training data but incorporates the scheduled-sampling strategy during training to simulate missing data by randomly hiding a part of actual features. In testing, the missing

**Table 1**
Prediction results of different spatiotemporal prediction methods using Dataset 1. The value after ± indicates the standard deviation.

| | *Precision* | *Recall* | *Accuracy* | *F1 score* |
|---|---|---|---|---|
| Transformer | 79.829 | 67.839 | 99.254 | 73.346 (±0.223)* |
| Swin transformer | **84.889** | 63.308 | 99.274 | 72.521 (±0.158)* |
| Stack-ConvLSTM | 79.058 | 69.689 | 99.262 | 74.077 (±0.393)* |
| Reverse-ConvLSTM | 79.353 | 69.864 | 99.269 | 74.307 (±0.281)* |
| Ours | 79.605 | **70.744** | **99.283** | **74.910** (±0.189) |

* Indicates that the difference between our method and the other methods is significant ($p < 0.05$).

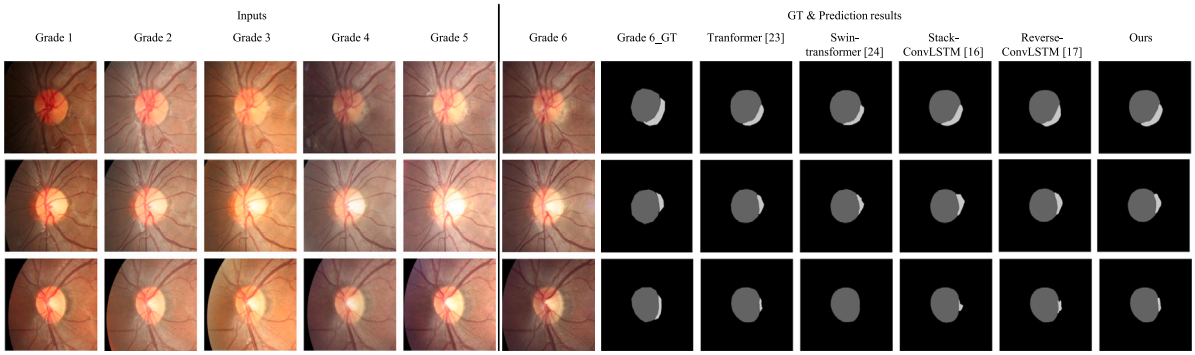


**Fig. 6.** Visual prediction results of different spatiotemporal prediction methods using Dataset 1.

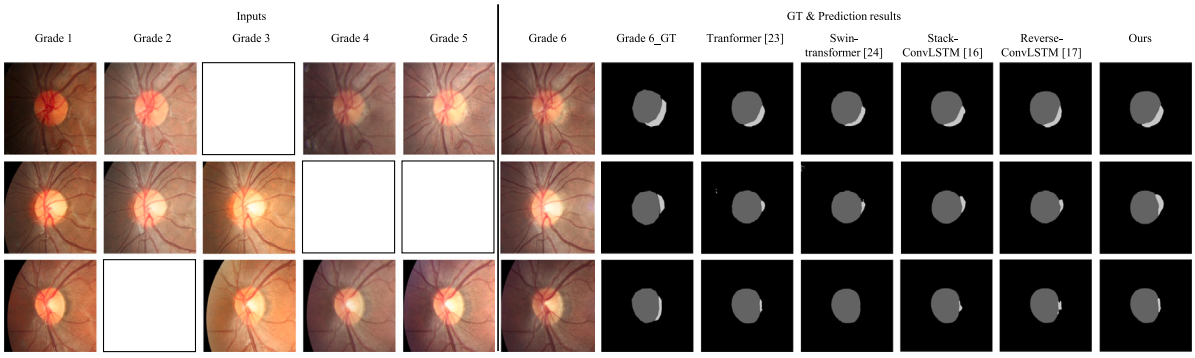


**Fig. 7.** Visual prediction results of different spatiotemporal prediction methods using Dataset 2. Blank areas indicate missing data.

**Table 2**
Prediction results of different spatiotemporal prediction methods using Dataset 2. The value after ± indicates the standard deviation.

| | *Precision* | *Recall* | *Accuracy* | *F1 score* |
|---|---|---|---|---|
| Transformer+interpolation | 79.801 | 63.757 | 99.207 | 70.881 (±0.346)* |
| Swin transformer+interpolation | **84.049** | 59.671 | 99.218 | 69.786 (±0.245)* |
| Stack-ConvLSTM+interpolation | 79.623 | 65.624 | 99.226 | 71.945 (±0.361)* |
| Reverse-ConvLSTM+interpolation | 80.890 | 65.085 | **99.239** | 72.131 (±0.194)* |
| Ours | 78.474 | **67.693** | 99.230 | **72.676** (±0.250) |

* Indicates that the difference between our method and the other methods is significant ($p < 0.05$).

data are estimated in advance using interpolation for other comparative methods. The quantitative results are shown in Table 2, and the multi-criterion graph is presented in Fig. 5(b). Compared with interpolation methods, our method achieves better results using other metrics except for its weaker performance when using *Precision*. For the highest priority metric *F1 score*, our method achieves the best performance. Fig. 7 shows the visual comparison between our method and others applying interpolation using Dataset 2. The predictions obtained using our method are closer to the ground truth. Because our method with scheduled-sampling takes advantage of the characteristics of sequential input of the ConvRNN structure by directly compensating the missing input with the predicted output of the spatiotemporal prediction module, our method is more convenient to operate compared with other methods using interpolation to address missing data and can better learn the spatiotemporal changes through the input data. Irrespective of the presence of missing data in the test set, our method can achieve the best prediction performance quantitatively and visually. Moreover, our method significantly outperforms ($p < 0.05$) the other methods for the *F1 score*.

### 3.4. Generalizability verification

The proposed method sequentially processes data and can output the result of the next moment after each input. As missing data are also considered during training, our method can predict future timestamps for any given years of data. To assess the generalizability of our method, we randomly select several years of data from Dataset 1 to predict the PPA region of the next 1 and 2 years, with quantitative results shown in Table 3. Our method achieves good prediction results under every listed condition, and its *F1 score* can reach more than 69%. In addition, as the input years increase, the prediction results become more accurate, exhibiting the advantage of using longitudinal images for future prediction.

Figs. 8 and 9 show the visual results. Fig. 8 presents the results of predicting the PPA regions in grades 4 and 5 by inputting the data of grades 2 and 3. Fig. 9 shows the results of predicting the PPA regions in grades 5 and 6 by inputting the data of grades 1, 3, and 4. Figs. 8 and 9 show that the prediction results obtained by our method are close to the ground truth. In addition, the change of PPA over time shows that our model can learn the progression of PPA, especially as shown by Example 1 in Fig. 9. Our method can also successfully predict when PPA will appear even though there is no PPA in the input images. Both quantitative and visual results show that our method can effectively predict the PPA regions with remarkable generalizability.

### 3.5. Ablaton study

To demonstrate the generality of the proposed spatiotemporal framework, different ConvRNN units are compared, including ConvLSTM [14], ConvGRU [15], and spatiotemporal LSTM (ST-LSTM) [16]. The experimental results are shown in Table 4. We regard the model without TMM and the scheduled-sampling strategy as the baseline.

**Table 3**
Prediction results under various conditions.

| | Ours w/SS | | | | Ours w/o SS + interpolation | | | |
|---|---|---|---|---|---|---|---|---|
| | *Precision* | *Recall* | *Accuracy* | *F1 score* | *Precision* | *Recall* | *Accuracy* | *F1 score* |
| Two years data→next year result | 78.341 | 65.266 | 99.391 | 71.205 | 79.417 | 63.613 | 99.390 | 70.640 |
| Three years data→next year result | 79.421 | 67.465 | 99.348 | 72.955 | 79.670 | 66.567 | 99.343 | 72.530 |
| Four years data→next year result | 80.766 | 68.433 | 99.324 | 74.089 | 81.298 | 67.776 | 99.325 | 73.923 |
| Two years data→two-year later result | 76.289 | 63.326 | 99.297 | 69.201 | 81.821 | 54.582 | 99.282 | 65.479 |
| Three years data→two-year later result | 78.831 | 64.266 | 99.265 | 70.805 | 81.500 | 59.516 | 99.252 | 68.793 |
| Four years data→two-year later result | 79.020 | 66.035 | 99.220 | 71.929 | 78.871 | 65.850 | 99.216 | 71.774 |

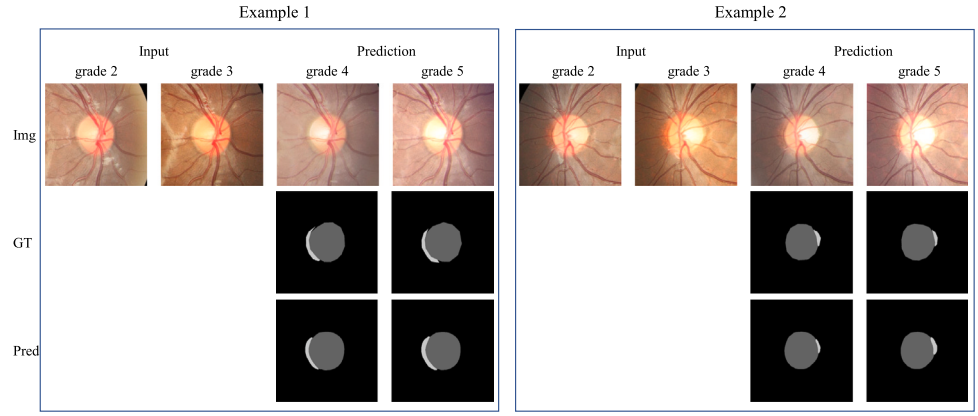


**Fig. 8.** Predicted PPA results of age in grades 4 and 5 by inputting the data of grades 2 and 3.

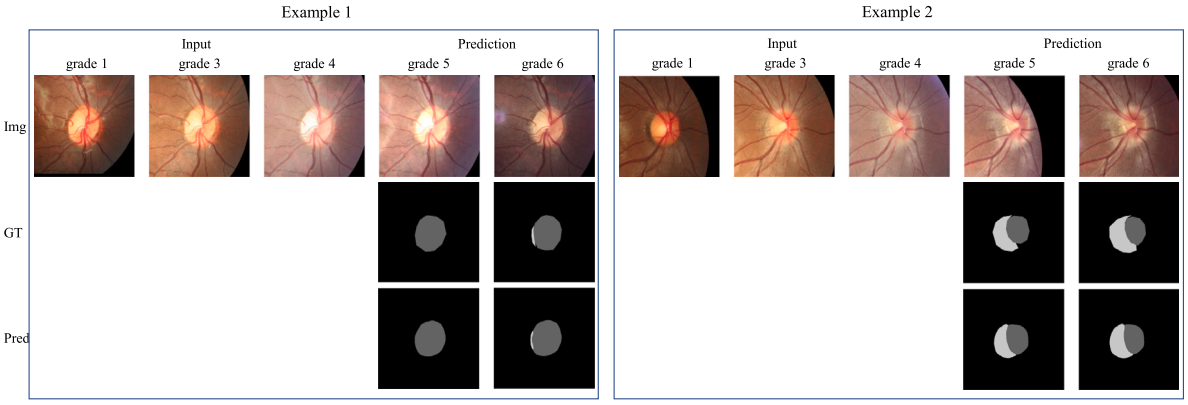


**Fig. 9.** Predicted PPA results of age in grades 5 and 6 by inputting the data of grades 1, 3, and 4.

In Tables 1–4, the term 'Ours' indicates the proposed spatiotemporal framework. When using the baseline to test Dataset 2, the interpolation method is applied to estimate missing data. Table 4 shows that no matter which ConvRNN units are adopted, our method can achieve better results using Datasets 1 and 2 when compared with the baseline. The results show that the proposed spatiotemporal framework is generalizable and can be extended to other ConvRNN units. In the following experiments, ConvLSTM is selected as the backbone of the ConvRNN unit.

We first verify the effectiveness of the feature extractor. The experimental results are shown in Table 5, which are performed using Dataset 1. At this stage, we only discuss the baseline model. Table 5 shows that the prediction performance of the network is considerably improved after incorporating the feature extractor module.

Then, we compare the baseline result of the ConvLSTM unit presented in Table 4 with the results of the stack-ConvLSTM and reverse-ConvLSTM units presented in Tables 1 and 2. Only the spatiotemporal prediction module is different in the above three methods. We can see that the use of the spatiotemporal prediction module can afford a higher *F1 score*, which shows that the memory information transmitted in different layers can improve the prediction performance.

Furthermore, we verify the effectiveness of TMM and scheduled-sampling strategy. Table 6 displays the results of the ablation study. When analyzing Dataset 2, except for the last row that uses the predicted result to substitute missing data, the other rows use the interpolation method to solve the problem of missing data. The result shows that the TMM can improve the performance compared with the baseline model. Because the input data of the spatiotemporal prediction module are sequential, the previous information will be gradually forgotten with time. The TMM enables the input at each moment to make full use of the historical information, resulting in more efficient use of long-term dependencies in the temporal dimension. When the scheduled-sampling strategy is further added to network training, the performances when using Datasets 1 and 2 are improved. The scheduled-sampling strategy simulates the lack of data during training but still uses all the training data. Hiding a part of the actual data

**Table 4**
Ablation test conducted on different ConvRNN units.

| | | Dataset 1 | | | | Dataset 2 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *Precision* | *Recall* | *Accuracy* | *F1 score* | *Precision* | *Recall* | *Accuracy* | *F1 score* |
| ConvLSTM | Baseline | 78.976 | 70.636 | 99.271 | 74.573 | **80.152** | 66.230 | **99.241** | 72.527 |
| | Ours | **79.605** | **70.744** | **99.283** | **74.910** | 78.474 | **67.693** | 99.230 | **72.676** |
| ConvGRU | Baseline | **79.145** | 69.728 | 99.263 | 74.119 | **80.111** | 65.957 | **99.236** | 72.330 |
| | Ours | 78.590 | **70.569** | **99.264** | **74.363** | 77.525 | **68.463** | 99.222 | **72.708** |
| ST-LSTM | Baseline | **79.229** | 70.185 | 99.270 | 74.433 | **79.851** | 66.598 | **99.240** | 72.625 |
| | Ours | 78.945 | **70.771** | **99.272** | **74.635** | 77.162 | **68.803** | 99.220 | **72.743** |

**Table 5**
Ablation study of feature extractor module using Dataset 1.

| Feature extractor | *Precision* | *Recall* | *Accuracy* | *F1 score* |
|---|---|---|---|---|
| | 64.485 | 64.546 | 98.923 | 64.483 |
| ✓ | 78.976 | 70.636 | 99.271 | 74.573 |

can force the network to learn long-term information to afford better performance. We also verify the effectiveness of scheduled sampling in more general scenarios. Table 3 shows the experimental results. The table shows that our method outperforms the interpolation method, especially in predicting long-term future moments. The interpolation method heavily relies on temporal smoothness and may be biased if the subject's condition deteriorates, while our method uses the predicted results to fill in the missing data, which can better learn the change in PPA through the input data to obtain more accurate prediction results.

To explore the impact of the proposed modules on the computational complexity of the network, we conduct an ablation study for complexity analysis. The results are shown in Table 7. After adding the TMM, 1.745 giga (G) of floating point operations (FLOPs) and 0.002 mega (M) parameters (Params) are increased. After further adding the scheduled-sampling strategy, no change is observed in the number of Params and FLOPs. The experiments demonstrate that incorporating the proposed modules enables the network to capture long-term dependencies more effectively in the temporal dimension, resulting in improved performance when compared with the baseline with a limited increase in computational resource consumption.

## 4. Discussion

The above experiments have shown the effectiveness of our framework. We first utilize the feature extractor module to encode the segmentation features of sequential fundus images. Then, the scheduled-sampling strategy is used to simulate the situation of missing data in practical application. To fully utilize the long-term dependencies in the temporal domain, the TMM is used. Finally, the spatiotemporal prediction module combined with a segmentation head is used to obtain the final prediction results. We have verified the generality of our proposed framework in experiments using different ConvRNN units, including ConvLSTM, ConvGRU, and ST-LSTM. For a general extension, any ConvRNN unit can be used in our framework to further improve the PPA prediction performance.

In the field of ocular disease prediction, most methods use sequential data to predict diseases at the image level, which cannot provide sufficient information for disease progression [11,12]. The method reported in Ref. [13] can be used to obtain the progression trend of age-related macular degeneration by generating fundus images at future moments. However, this method only takes a single moment's image as input; therefore, it lacks temporal information in comparison with sequential data. Disease progression is gradual and patient-specific, making temporal information essential for accurate prediction. Our method can not only effectively utilize the temporal information in sequential data but also refine the prediction to the pixel level.

The objective of this study is to predict the PPA region. Various structures, textures, and details of fundus images will impact the model's learning of PPA. Although the pretrained feature extractor module can extract accurate PPA segmentation features, PPA is not obvious at the early stage of development, making precise segmentation a challenge during this stage. The inaccurate PPA segmentation feature will influence the subsequent prediction stage, leading to false predictions. In our future study, we will focus on how to obtain precise segmentation features.

The scheduled-sampling strategy simulates missing data by randomly hiding real input features during training. In practical applications, the features predicted by the spatiotemporal prediction module can be used to substitute the missing data. Although the scheduled-sampling strategy can obtain good performance when dealing with missing data, differences between the filled predicted features and the real features remain, which will also lead to false predictions. Furthermore, in the current way of estimating missing data during training, although the input data are reduced, the number of iterations required for the model is not decreased. In the future study, we will explore a method that only uses sequential data and time intervals to achieve PPA prediction, especially in scenarios where data may be missing.

Our proposed framework is designed for PPA prediction, but it is also a spatiotemporal prediction framework that can provide inspiration for other spatiotemporal prediction problems such as drusen prediction [13] and precipitation nowcasting [25,26].

## 5. Conclusion

In this study, we propose a spatiotemporal framework for pixel-level PPA prediction using sequential fundus images. Our approach integrates a TMM as well as a spatiotemporal prediction module based on ConvRNN and employs a scheduled-sampling strategy during training. Our method enables the accurate prediction of the PPA region even in the presence of missing data, thereby enhancing the ability to gain quantitative insights into disease progression.

Our framework is trained on 250 groups of data and tested on 50 groups of data. The experimental results showed that our method achieves the highest *F1 score* regardless of the completeness of the input follow-up data, i.e., whether tested using Dataset 1 without missing data (74.910%) or Dataset 2 with missing data (72.676%). Our method is designed to transmit memory information between different layers in a spatiotemporal prediction module, and the integration of the TMM can make full use of the memory information across diverse inputs and different layers of the same input, thereby enhancing the *F1 score* by 0.826%.

Compared with the interpolation method, our method with the scheduled-sampling strategy during training can effectively handle the challenge of incomplete follow-up data, achieving better prediction performance without requiring additional processing of the input data. Because the follow-up data are often incomplete in actual clinical scenarios, our method has considerable practical value for clinical application. Our findings highlight the potential of our approach in predicting pixel-level PPA regions, representing a notable advancement in personalized patient diagnosis and treatment. Moreover, our method can inspire spatiotemporal prediction in other applications.

**Table 6**
Ablation study using the TMM and scheduled-sampling (SS).

| | Dataset 1 | | | | Dataset 2 | | | |
|---|---|---|---|---|---|---|---|---|
| | *Precision* | *Recall* | *Accuracy* | *F1 score* | *Precision* | *Recall* | *Accuracy* | *F1 score* |
| Baseline | 78.976 | 70.636 | 99.271 | 74.573 | 80.152 | 66.230 | 99.241 | 72.527 |
| Baseline+TMM | **79.679** | 70.669 | **99.283** | 74.903 | **80.270** | 66.348 | **99.244** | 72.647 |
| Baseline+TMM+SS | 79.605 | **70.744** | **99.283** | **74.910** | 78.474 | **67.693** | 99.230 | **72.676** |

**Table 7**
Complexity analysis of the ablation study.

| Baseline | TMM | SS | FLOPs (G) | Total params (M) | Trainable params (M) |
|---|---|---|---|---|---|
| ✓ | | | 433.754 | 32.829 | 0.308 |
| ✓ | ✓ | | 435.499 | 32.831 | 0.310 |
| ✓ | ✓ | ✓ | 435.499 | 32.831 | 0.310 |

Although the present investigation is a pioneer study using sequential fundus images for future-stage pixel-level PPA-region prediction, our method has some limitations. In future research, we will further improve the performance of the feature extractor module to obtain more precise segmentation features. Moreover, we will improve the architecture to only use sequential data and time intervals to achieve PPA prediction.

## CRediT authorship contribution statement

**Mengxuan Li:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation. **Weihang Zhang:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **He Zhao:** Supervision, Methodology, Conceptualization. **Yubin Xu:** Validation, Software. **Jie Xu:** Writing – review & editing, Resources, Formal analysis, Data curation. **Huiqi Li:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Formal analysis, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

The authors do not have permission to share data.

## Acknowledgments

This research is supported by the National Natural Science Foundation of China (NSFC) (Grant No. 82072007), Beijing Natural Science Foundation (Grant No. IS23112), and Beijing Institute of Technology Research Fund Program for Young Scholars.

## References

[1] V. Manjunath, H. Shah, J. Fujimoto, J. Duker, Analysis of peripapillary atrophy using spectral domain optical coherence tomography, Ophthalmology 118 (3) (2011) 531–536.
[2] K. Lee, A. Tomidokoro, R. Sakata, et al., Cross-sectional anatomic configurations of peripapillary atrophy evaluated with spectral domain-optical coherence tomography, Investig. Pphthalmol. Visual Sci. 51 (2) (2010) 666–671.
[3] J. Jonas, S. Jonas, R. Jonas, et al., Parapapillary atrophy: histological gamma zone and delta zone, PLoS One 7 (10) (2012) e47237.
[4] H. Uchida, S. Ugurlu, J. Caprioli, Increasing peripapillary atrophy is associated with progressive glaucoma, Ophthalmology 105 (8) (1998) 1541–1545.
[5] H. Park, S. Jeon, C. Park, Features of the choroidal microvasculature in peripapillary atrophy are associated with visual field damage in myopic patients, Amer. J. Ophthalmol. 192 (2018) 206–216.
[6] H. Li, H. Li, J. Kang, Y. Feng, J. Xu, Automatic detection of parapapillary atrophy and its association with children myopia, Comput. Methods Programs Biomed. 183 (2020) 105090.
[7] J. Zhang, J. Li, J. Wang, et al., The association of myopia progression with the morphological changes of optic disc and $\beta$-peripapillary atrophy in primary school, Graefes Arch. Clin. Exp. Ophthalmol. 260 (2) (2022) 677–687.
[8] Y. Moon, H. Lim, Relationship between peripapillary atrophy and myopia progression in the eyes of young school children, Eye 35 (2) (2021) 665–671.
[9] Y. Guo, Y. Wang, L. Xu, et al., Five-year follow-up of parapapillary atrophy: the Beijing eye study, PLoS One 7 (5) (2012) e32005.
[10] J. Xu, L. Xu, Six dimensional evaluation for myopia prevention and control, Chin. J. Optom. Ophthalmol. Vis. Sci. 20 (3) (2018) 129–132.
[11] J. Li, B. Wu, X. Sun, Y. Wang, Causal hidden markov model for time series disease forecasting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 12105–12114.
[12] B. Wu, S. Ren, J. Li, et al., Forecasting irreversible disease via progression learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 8117–8125.
[13] Q. Pham, S. Ahn, J. Shin, S. Song, Generating future fundus images for early age-related macular degeneration based on generative adversarial networks, Comput. Methods Programs Biomed. 216 (2022) 106648.
[14] X. Shi, Z. Chen, H. Wang, et al., Convolutional LSTM network: a machine learning approach for precipitation nowcasting, in: Advances in Neural Information Processing Systems, NIPS, Vol. 28, 2015, pp. 802–810.
[15] X. Shi, Z. Gao, L. Lausen, et al., Deep learning for precipitation nowcasting: a benchmark and a new model, in: Advances in Neural Information Processing Systems, NIPS, Vol. 30, 2017, pp. 5618–5628.
[16] Y. Wang, M. Long, J. Wang, Z. Gao, P. Yu, Predrnn: recurrent neural networks for predictive learning using spatiotemporal LSTMs, in: Advances in Neural Information Processing Systems, NIPS, Vol. 30, 2017, pp. 880–889.
[17] Y. Wang, Z. Gao, M. Long, J. Wang, P. Yu, PredRNN++: towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning, in: International Conference on Machine Learning, ICML, 2018, pp. 5123–5132.
[18] L. Shi, N. Liang, X. Xu, T. Li, Z. Zhang, SA-JSTN: self-attention joint spatiotemporal network for temperature forecasting, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14 (2021) 9475–9485.
[19] T. Xiong, J. He, H. Wang, et al., Contextual Sa-attention convolutional LSTM for precipitation nowcasting: a spatiotemporal sequence forecasting view, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14 (2021) 12479–12491.
[20] C. Luo, X. Zhao, Y. Sun, X. Li, Y. Ye, PredRANN: the spatiotemporal attention convolution recurrent neural network for precipitation nowcasting, Knowl.-Based Syst. 239 (2022) 107900.
[21] X. Zhang, Q. Jin, T. Yu, et al., Multi-modal spatio-temporal meteorological forecasting with deep neural network, ISPRS J. Photogramm. Remote Sens. 188 (2022) 380–393.
[22] A. Bojesomo, H. Al-Marzouqi, P. Liatsis, Spatiotemporal swin-transformer network for short time weather forecasting, in: International Conference on Information and Knowledge Management (CIKM) Workshops, Vol. 3052, 2021, pp. 1–5.
[23] Y. Xue, Y. Cao, M. Zhou, et al., Rock mass fracture maps prediction based on spatiotemporal image sequence modeling, Comput.-Aided Civ. Infrastruct. Eng. 38 (4) (2023) 470–488.
[24] H. Geng, L. Geng, MCCS-LSTM: extracting full-image contextual information and multi-scale spatiotemporal feature for radar echo extrapolation, Atmosphere 13 (2) (2022) 192.
[25] Z. Zhang, C. Luo, S. Feng, et al., RAP-Net: region attention predictive network for precipitation nowcasting, Geosci. Model Dev. 15 (13) (2022) 5407–5419.
[26] C. Luo, X. Li, Y. Ye, PFST-LSTM: a spatiotemporal LSTM model with pseudoflow prediction for precipitation nowcasting, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14 (2020) 843–857.
[27] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI, Vol. 9351, 2015, pp. 234–241.
[28] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[29] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014, arXiv preprint arXiv:1412.3555.
[30] K. Cho, B. Merrienboer, C. Gulcehre, et al., Learning phrase representations using RNN encoder–decoder for statistical machine translation, 2014, arXiv preprint arXiv:1406.1078.

[31] S. Bengio, O. Vinyals, N. Jaitly, N. Shazzer, Scheduled sampling for sequence prediction with recurrent neural networks, in: Advances in Neural Information Processing Systems, NIPS, Vol. 28, 2015, pp. 1171–1179.
[32] Y. Wang, H. Wu, J. Zhang, et al., PredRNN: a recurrent neural network for spatiotemporal predictive learning, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2) (2022) 2208–2225.
[33] Z. Zhao, F. Zhou, Z. Zeng, et al., Meta-hallucinator: towards few-shot cross-modality cardiac image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI, Vol. 13435, 2022, pp. 128–139.
[34] Z. Xing, L. Yu, L. Wan, et al., NestedFormer: nested modality-aware transformer for brain tumor segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI, Vol. 13435, 2022, pp. 140–150.
[35] X. Sun, L. Cheng, S. Plein, et al., Transformer based feature fusion for left ventricle segmentation in 4D flow MRI, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI, Vol. 13435, 2022, pp. 370–379.
[36] J. Chen, J. Tian, N. Lee, et al., A partial intensity invariant feature descriptor for multimodal retinal image registration, IEEE Trans. Biomed. Eng. 57 (7) (2010) 1707–1718.
[37] M. Fay, M. Proschan, Wilcoxon-Mann–Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules, Stat. Surv. 4 (2010) 1–39.