Coarse-to-Fine Latent Diffusion Model for Glaucoma Forecast on Sequential Fundus Images

Yuhan Zhang[1,3], Kun Huang[4], Xikai Yang[1], Xiao Ma[4], Jian Wu[5,6], Ningli Wang[5,6], Xi Wang[1,*], and Pheng-Ann Heng[1,2]

[1] Department of Computer Science and Engineering,
The Chinese University of Hong Kong, Hong Kong, China
[2] Institute of Medical Intelligence and XR,
The Chinese University of Hong Kong, Hong Kong, China
[3] Shenzhen Research Institute,
The Chinese University of Hong Kong, Hong Kong, China
[4] Department of Computer Science and Engineering,
Nanjing University of Science and Technology, Nanjing, China
[5] Henan Academy of Innovations in Medical Science, Zhengzhou, China
[6] Beijing Tongren Eye Center,
Beijing Key Laboratory of Ophthalmology and Visual Sciences, Beijing, China
{xiwang@cse.cuhk.edu.hk}

Abstract. Glaucoma is one of the leading causes of irreversible blindness worldwide. Predicting the future status of glaucoma is essential for early detection and timely intervention of potential patients and avoiding the outcome of blindness. Based on historical fundus images from patients, existing glaucoma forecast methods directly predict the probability of developing glaucoma in the future. In this paper, we propose a novel glaucoma forecast method called Coarse-to-Fine Latent Diffusion Model (C2F-LDM) to generatively predict the possible features at any future time point in the latent space based on sequential fundus images. After obtaining the predicted features, we can detect the probability of developing glaucoma and reconstruct future fundus images for visualization. Since all fundus images in the sequence are sampled at irregular time points, we propose a time-adaptive sequence encoder that encodes the sequential fundus images with their irregular time intervals as the historical condition to guide the latent diffusion model, making the model capable of capturing the status changes of glaucoma over time. Furthermore, a coarse-to-fine diffusion strategy improves the quality of the predicted features. We verify C2F-LDM on the public glaucoma forecast dataset SIGF. C2F-LDM presents better quantitative results than other state-of-the-art forecast methods and provides visual results for qualitative evaluations.

Keywords: Latent Diffusion · Glaucoma Forecast · Fundus Image.

1 Introduction

Glaucoma is one of the leading causes of permanent vision loss on a global scale, affecting a substantial populace spanning diverse age cohorts and ethnicities.

Early detection and timely intervention play a pivotal role in mitigating vision deterioration and avoiding the consequences of complete blindness [22]. Therefore, forecasting the forthcoming degenerative trajectory of glaucoma has become an urgent demand. Glaucoma forecast shows its potential to identify individuals predisposed to glaucomatous affliction or to predict disease progression in those already bestowed with a diagnosis [11, 14].

Different from glaucoma detection [3, 7, 27] that diagnoses the current status of glaucoma based on the existing fundus images, glaucoma forecast predicts the future status of glaucoma by analyzing the historical fundus images. Several glaucoma forecast methods have been developed based on deep learning [13, 21, 20, 9]. For example, Lin *et al.* [16] proposed a multi-scale multi-structure siamese network (MMSNet) to predict the progress of glaucoma based on the current visit and the first visit. However, two fundus images lack enough ability to capture the glaucoma changes over time. Li *et al.* [14] established a glaucoma forecast dataset consisting of sequential fundus images and proposed a long short-term memory (LSTM)-based network DeepGF to learn spatial-temporal information from sequential fundus images of a patient. DeepGF outputs the probability of developing glaucoma at the next time step, but cannot specify when the next time is. Hu *et al.* [11] proposed a Transformer-based glaucoma forecast network GLIM-Net for irregularly sampled sequential fundus images, which introduced two time-related modules to control the prediction under a specific future time. All the above glaucoma forecast methods only predict the probability of developing glaucoma in the future, lacking visual results for further qualitative evaluation and interpretability.

Recently, the denoising diffusion model (DDM) [10], which as a type of generative model has achieved promising attention in image generation and image synthesis [1, 2]. DDM aims to denoise corrupted versions of the input images, which helps it learn the true distributions and capture the underlying data structure in the pixel space. To further improve the computational efficiency and flexibility, the latent diffusion model (LDM) [1, 18, 19] is developed by transforming the diffusion process from pixel space to latent space, making them attractive for various generative tasks. In the latent space, LDM abstracts away high-frequency and imperceptible details, enables more fine-grained control and is more scalable in terms of model size. By incorporating external conditional information into the diffusion process, controllable generation can be achieved by allowing for a more accurate representation and prediction of the variable's behavior. For example, Zbinden *et al.* [26] introduced categorical label maps and Yang *et al.* [24] designed a dual-granularity conditional guidance module as the conditional priors. Although existing works have tried to adopt the diffusion models for semantic segmentation and object classification [24, 17, 26], their potential for forecast tasks has yet to be fully explored.

In this paper, we propose a novel glaucoma forecast method called Coarse-to-Fine Latent Diffusion Model (C2F-LDM) based on irregularly sampled sequential fundus images. Our contributions are summarized as: (1) We predict the future possible features in the latent space by the latent diffusion model based on his-

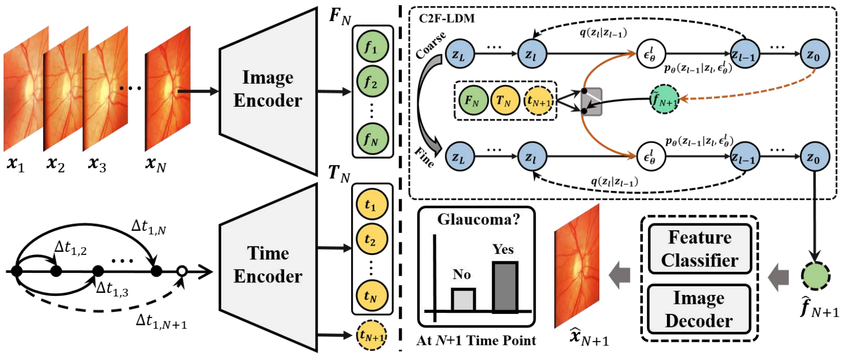


Fig. 1. Overview of our glaucoma forecast framework, which receives the historical sequential fundus images and corresponding time intervals and predicts the future possible features in the latent space for glaucoma forecast and fundus image reconstruction.

torical sequential fundus images, and the predicted features are used to detect the probability of developing glaucoma and reconstruct the future fundus images for visual evaluations; (2) We propose a time-adaptive sequence encoder to integrate irregular time intervals, making the model capable of capturing the status changes of glaucoma over time and making a flexible prediction by giving any future time point; (3) We propose a coarse-to-fine diffusion strategy to improve the quality of the predicted features; (4) The experimental results on the publicly available glaucoma forecast dataset SIGF show that C2F-LDM performs better than other state-of-the-art glaucoma forecast methods.

2 Methods

2.1 Forecast Framework

Fig. 1 presents our glaucoma forecast framework. Given N historical sequential fundus images $\boldsymbol{X}_N = \{\boldsymbol{x}_n \in \mathbb{R}^{3\times H\times W}\}_{n=1}^{N}$ and their corresponding irregular time intervals $\Delta\mathcal{T}_N = \{\Delta t_{(1,n)} \in \mathbb{R}^1\}_{n=1}^{N}$ with the first time point, we first encode them into features in the latent space by:

$$\begin{aligned}\boldsymbol{F}_N &= \mathcal{G}_{\mathrm{ie}}(\boldsymbol{X}_N), \quad \text{with } \boldsymbol{F}_N = \{\boldsymbol{f}_n \in \mathbb{R}^{d\times h\times w}\}_{n=1}^{N}\\ \boldsymbol{T}_N &= \mathcal{G}_{\mathrm{te}}(\Delta\mathcal{T}_N), \quad \text{with } \boldsymbol{T}_N = \{\boldsymbol{t}_n \in \mathbb{R}^{d}\}_{n=1}^{N}\end{aligned} \tag{1}$$

where $\mathcal{G}_{\mathrm{ie}}(\cdot)$ is the image encoder, $\mathcal{G}_{\mathrm{te}}(\cdot)$ is the time encoder, d is the feature dimension, $h \times w$ is the size of feature maps, and $H \times W$ is the size of fundus images. Then, our proposed C2F-LDM generatively predicts the future feature $\hat{\boldsymbol{f}}_{N+1}$ at $N{+}1$ time point by giving a specific future time $\Delta t_{(1,N+1)}$:

$$\boldsymbol{f}_{N+1} \sim \hat{\boldsymbol{f}}_{N+1} = \mathcal{P}(\boldsymbol{F}_N, \boldsymbol{T}_N, \boldsymbol{t}_{N+1}), \quad \text{with } \boldsymbol{t}_{N+1} = \mathcal{G}_{\mathrm{te}}(\Delta t_{(1,N+1)}) \tag{2}$$

where $\Delta t_{(1,N+1)} > \Delta t_{(1,N)}$, $\boldsymbol{f}_{N+1}$ is the true feature at $N{+}1$ time point. Based on the predicted feature $\hat{\boldsymbol{f}}_{N+1}$, it is possible to detect the probability $\hat{\boldsymbol{y}}_{N+1}$ of developing glaucoma by a feature classifier $\mathcal{G}_{\mathrm{c}}$, and reconstruct the future fundus image $\hat{\boldsymbol{x}}_{N+1}$ for visual evaluation by an image decoder $\mathcal{G}_{\mathrm{r}}$:

$$\hat{\boldsymbol{y}}_{N+1} \in \mathbb{R}^1 = \mathcal{G}_{\mathrm{c}}(\hat{\boldsymbol{f}}_{N+1}), \quad \hat{\boldsymbol{x}}_{N+1} \in \mathbb{R}^{3\times H\times W} = \mathcal{G}_{\mathrm{r}}(\hat{\boldsymbol{f}}_{N+1}) \tag{3}$$

2.2 Coarse-to-Fine Latent Diffusion Model (C2F-LDM)

C2F-LDM is built based on a conditional latent diffusion model, which contains a coarse latent diffusion (CLD) module and a fine latent diffusion (FLD) module. In CLD and FLD, the Gaussian noise is predicted by a conditional U-Net (CU-Net). Fig. 2 shows the architecture of C2F-LDM.

Coarse Latent Diffusion (CLD). During training, CLD consists of a forward diffusion process and a reverse denoising process. In the forward diffusion process, a random noisy variable $\boldsymbol{z}_l$ is sampled based on the true prior $\boldsymbol{z}_0$ across l steps:

$$\boldsymbol{z}_l = \sqrt{\overline{\alpha}_l}\boldsymbol{z}_0 + \sqrt{1-\overline{\alpha}_l}\epsilon, \quad \text{with } \overline{\alpha}_l = \prod\nolimits_l \alpha_l, \ \alpha_l = 1-\beta_l, \ l \in [1, L] \tag{4}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ is the Gaussian noise, $\{\beta_l \in (0,1)\}_{l=1}^{L}$ is a predefined noise schedule, L is the max step. After L steps, $\boldsymbol{z}_L \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ has a standard isotropic Gaussian distribution. In our work, $\boldsymbol{z}_0$ is initialized by the true feature $\boldsymbol{f}_{N+1}$ at $N+1$ time point. In the reverse denoising process, the true prior $\boldsymbol{z}_0$ is restored from $\boldsymbol{z}_L$ by multi-step employing CU-Net:

$$\begin{aligned} \boldsymbol{z}_{l-1} &= \frac{1}{\sqrt{\alpha_l}}(\boldsymbol{z}_l - \frac{1-\alpha_l}{\sqrt{1-\overline{\alpha}_l}}\epsilon_\theta(\boldsymbol{z}_l, \boldsymbol{E}_l, \mathcal{C})) + \sigma_l\epsilon \\ \text{with } \sigma_l &= \sqrt{\frac{1-\overline{\alpha}_{l-1}}{1-\overline{\alpha}_l}\beta_l}, \quad \mathcal{C} = \mathcal{F}_{\mathrm{hke}}(\boldsymbol{F}_N, \boldsymbol{T}_N, \boldsymbol{t}_{N+1}) \end{aligned} \tag{5}$$

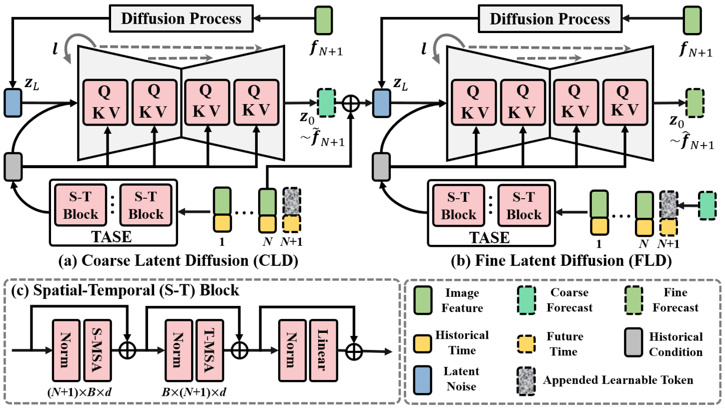
where $\boldsymbol{E}_l$ is the step embeddings, $\epsilon_\theta(\boldsymbol{z}_l, \boldsymbol{E}_l, \mathcal{C})$ is the predicted noise by joint sampling from $\boldsymbol{z}_l$ and historical condition $\mathcal{C}$ via CU-Net, $\mathcal{F}_{\mathrm{hke}}(\cdot)$ is the historical knowledge encoding module (section 2.3) for generating $\mathcal{C}$. In CU-Net, $\mathcal{C}$ is mapped to the intermediate layers via a cross-attention mechanism. Finally, the restored $\widetilde{\boldsymbol{f}}_{N+1} \sim \boldsymbol{z}_0 = p_\theta(\boldsymbol{z}_0|\boldsymbol{z}_1, \mathcal{C})$ follows the prior distributions of $\boldsymbol{f}_{N+1}$. We minimize the noise estimation loss for training CLD by:

$$\mathcal{L}(\theta) = \mathbb{E}_{\boldsymbol{z}_l,\epsilon,l}||\epsilon - \epsilon_\theta(\boldsymbol{z}_l, \boldsymbol{E}_l, \mathcal{C})||_2^2 \tag{6}$$

where θ is the learnable parameters in CLD.

Fine Latent Diffusion (FLD). To further improve the quality of predicted features, we continue to perform fine-grained diffusion by introducing $\widetilde{\boldsymbol{f}}_{N+1}$ from CLD and $\boldsymbol{f}_N$ from the last time point of historical sequence. In the forward diffusion process, $\boldsymbol{z}_l$ is jointly sampled based on $\boldsymbol{z}_0$, $\boldsymbol{f}_N$ and $\widetilde{\boldsymbol{f}}_{N+1}$ across l steps:

$$\boldsymbol{z}_l = \sqrt{\overline{\alpha}_l}\boldsymbol{z}_0 + \sqrt{1-\overline{\alpha}_l}\epsilon + (1-\sqrt{\overline{\alpha}_l})(\boldsymbol{f}_N + \widetilde{\boldsymbol{f}}_{N+1}) \tag{7}$$



(a) Coarse Latent Diffusion (CLD)

(b) Fine Latent Diffusion (FLD)

(c) Spatial-Temporal (S-T) Block

Fig. 2. Architecture of C2F-LDM for the feature forecast in the latent space.

Similarly, $\boldsymbol{z}_0$ is initialized by $\boldsymbol{f}_{N+1}$. In the reverse denoising process, the true prior $\boldsymbol{z}_0$ is restored from $\boldsymbol{z}_L$ by multi-step employing CU-Net:

$$\begin{aligned} \boldsymbol{z}_{l-1} &= \frac{1}{\sqrt{\alpha_l}}\left(\boldsymbol{z}_l - \frac{1-\alpha_l}{\sqrt{1-\bar{\alpha}_l}}\epsilon_\theta(\boldsymbol{z}_l, \boldsymbol{E}_l, \mathcal{C}_f)\right) + \sigma_l \epsilon \\ \text{with } \ \mathcal{C}_f &= \mathcal{F}_{\text{hke}}(\boldsymbol{F}_N, \boldsymbol{T}_N, \boldsymbol{t}_{N+1}, \widetilde{\boldsymbol{f}}_{N+1}) \end{aligned} \tag{8}$$

Finally, the restored $\hat{\boldsymbol{f}}_{N+1} \sim \boldsymbol{z}_0 = p_\theta(\boldsymbol{z}_0|\boldsymbol{z}_1, \mathcal{C}_f)$ can be used for the glaucoma detection by a feature classifier and the fundus image reconstruction by an image decoder. We optimize FLD by the same noise estimation loss as Eq. (6).

2.3 Historical Knowledge Encoding (HKE)

To encode sequential fundus images as a historical condition for the diffusion process, we propose a time-adaptive sequence encoder (TASE) that stacks K spatial-temporal (S-T) blocks, as shown in Fig. 2(c). Each S-T block is an extension of a ViT block [5] by introducing a spatial-temporal transformation [23]. Given the image feature sequence $\widetilde{\boldsymbol{F}}_N \in \mathbb{R}^{N\times d}$ by global average pooling $\boldsymbol{F}_N$, we append an extra learnable token $\boldsymbol{r}$ to represent the future feature at $N{+}1$ time point. The time features $\boldsymbol{T}_N$ and $\boldsymbol{t}_{N+1}$ are concatenated to the image feature sequence as the input of TASE. Finally, we use a linear head to map the feature of $N{+}1$ time point as the historical condition. The process is formulated as:

$$\begin{aligned} &[\boldsymbol{A}_{k+1}^{N}, \boldsymbol{A}_{k+1}^{N+1}] = \mathbb{B}_k([\boldsymbol{A}_k^N, \boldsymbol{A}_k^{N+1}]),\ k = 1, 2, ..., K \\ &\mathcal{C} = \text{Head}(\boldsymbol{A}_{K+1}^{N+1}) \\ &\text{with } \boldsymbol{A}_1^N = \text{Cat}(\widetilde{\boldsymbol{F}}_N, \boldsymbol{T}_N),\ \boldsymbol{A}_1^{N+1} = \text{Cat}(\boldsymbol{r}, \boldsymbol{t}_{N+1}) \end{aligned} \tag{9}$$

where $\mathbb{B}_k$ denotes k-th S-T block, $[\cdot,\cdot]$ indicates stacking on the sequence length dimension, $\text{Cat}(\cdot,\cdot)$ indicates concatenating on the feature length dimension. In FLD, $\boldsymbol{r}$ is initialized with $\widetilde{\boldsymbol{f}}_{N+1}$ from CLD. In our framework, HKE is integrated into CLD and FLD, and is optimized along with CU-Net.

2.4 Multi-Stage Training (MST)

As shown in Fig. 1, our proposed glaucoma forecast framework contains an image encoder, an image decoder, a time encoder, a feature classifier, a CLD module, and an FLD module. The time encoder requires no training because it follows the parameter-free time positional encoding method [11]. CLD and FLD have their own CU-Net, and their parameters are not shared. To optimize the model adequately, we train the different components in multiple stages. Firstly, we train the image encoder and the image decoder by VQGAN [6]. Next, we train the feature classifier by the glaucoma detection task while freezing the parameters of other modules. Finally, we train CLD and FLD separately by Eq. (6).

3 Experiments

3.1 Materials and Details

The efficacy of our proposed method is estimated on the publicly available glaucoma forecast dataset SIGF [14]. SIGF contains 405 sequences derived from distinct eyes. Each sequence is accompanied by no fewer than 6 fundus images, producing an average of 9 images per eye. Overall, SIGF comprises a total of 3671 fundus images. To establish a robust evaluation framework, the 405 sequences are randomly split for training (300), validation (35) and testing (70) at the patient level to ensure the avoidance of any potential biases. All fundus images are annotated with binary labels of glaucoma, *i.e.* positive or negative glaucoma. The 405 sequences are temporally segmented into 1146 clips, with each clip encompassing a continuous sequence of 6 fundus images.

The image encoder and image decoder follow the original U-Net encoder and decoder with three down-sampling operations, without skip connection. The feature classifier is a multi-layer perception with six layers. They were pre-trained based on two external glaucoma detection datasets LAG [15] and ACRIMA [4]. All fundus images are resized to 256×256 for consistent input. The feature dimension d is set to 768. K in HKE is set to 6. Experiments are built in NVIDIA TITAN Xp GPUs. More details can be found in our code[1].

3.2 Experimental Results

Quantitative Comparison with State-of-the-art Methods. We compare the proposed C2F-LDM with five other state-of-the-art glaucoma forecast methods, including CoG-Net [12], CABNet [8], MIL-VT [25], DeepGF [14] and GLIM-Net [11]. CoG-Net, CABNet and MIL-VT are the classification models for glaucoma detection on fundus images and we convert them for glaucoma forecast by supervising the models with the future status of glaucoma. DeepGF and GLIM-Net are the glaucoma forecast methods based on sequential fundus images. DeepGF learns the dynamic glaucoma transition based on the LSTM network,

[1] https://github.com/ZhangYH0502/C2F-LDM

Table 1. Quantitative comparison of glaucoma forecast with state-of-the-art methods over SIGF dataset based on the accuracy (ACC), sensitivity (SEN), specificity (SPE) and AUC metrics. For each metric, we show the mean and standard deviation.

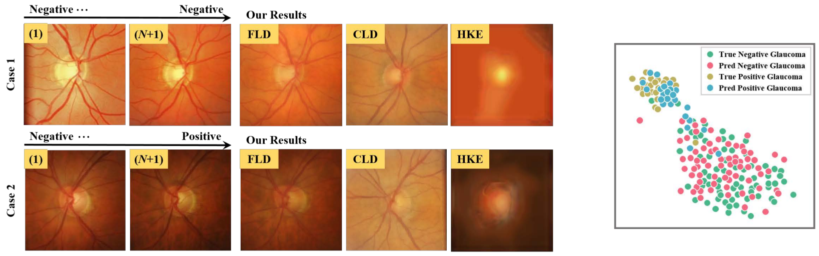
Methods	Core Technology	ACC (%)	SEN (%)	SPE (%)	AUC (%)
CoG-Net [12]	ImgPro+ConvNet	77.0±2.0	72.5±3.6	77.2±1.7	81.8±3.5
CABNet [8]	ConvNet+Atten.	73.9±1.6	74.4±1.6	73.9±1.6	78.7±2.6
MIL-VT [25]	Trans.+MIL	79.7±1.1	77.8±3.4	79.8±1.2	83.4±1.6
DeepGF [14]	ConvNet+LSTM	76.0±4.8	79.4±1.3	75.9±5.0	85.0±2.5
GLIM-Net [11]	Trans.+ConvNet	89.5±0.8	87.6±0.9	89.6±0.8	93.6±0.3
C2F-LDM	Trans.+Diffusion	**94.4±0.5**	**93.8±0.7**	**94.6±0.6**	**95.5±0.5**

but DeepGF has to make the prediction sequentially due to the unidirectional limitations of LSTM network. GLIM-Net models the fundus image sequence based on the Transformer architecture to better suit the irregularly sampled data by inserting the position encoding and time encoding. Table 2 presents the core technologies of all methods and compares the quantitative glaucoma forecast results by four metrics. Among all comparative methods, GLIM-Net significantly performs better than the other four methods, indicating the advantage of Transformer architecture in modeling sequence data and the importance of introducing the time factors. Furthermore, the proposed C2F-LDM performs an obvious quantitative improvement over GLIM-Net by achieving 94.4%, 93.8%, 94.6%, and 95.5% in accuracy, sensitivity, specificity and AUC, respectively.

Qualitative Evaluation via Visualization. We first evaluate the reconstructed fundus images by decoding the predicted features via the image decoder. Fig. 3(a) shows two cases of glaucoma sequences from different patients, wherein we only show the first time point in the historical sequence and the future $N+1$ time point we require to predict. In the first case, the status of glaucoma remains negative over time. In the second case, the status of glaucoma changes from negative to positive at the $N+1$ time point. In these two cases, the reconstructed fundus images own high enough image quality and keep the consistent glaucoma status with the true fundus images. We also utilize the t-SNE method to visualize the distributions of the true features and the predicted features from the feature classifier. As shown in Fig. 3(b), the predicted features are diacritical enough through the feature classifier, and show consistent distributions with the true features. Therefore, C2F-LDM has the potential to predict the future status of glaucoma interpretively by providing visual results as references.

3.3 Ablation Study

We first investigate the impact of latent diffusion by evaluating the results from FLD, CLD and HKE respectively. As shown in Table 2(b), the quantitative comparisons demonstrate that CLD and FLD effectively improves the quality



(a) Visualization of Reconstructed Images

(b) Visualization of Feature Distributions

Fig. 3. (a) Reconstructed fundus images of two cases by the image decoder. (b) Visualization of the true features and the predicted features from the feature classifier.

of predicted features based on historical conditions rather than directly using historical features for glaucoma forecast. From the visual comparisons in Fig. 3, we find that the quality of reconstructed fundus images from HKE is deficient, but CLD and FLD present powerful potential to improve image quality. In the results of CLD and FLD, pathological details keep high consistency with ground truth, while FLD shows more similar color distributions and pathological details with ground truth than CLD by introducing more controlled conditions.

Next, we verify the effect of TASE in the HKE module. As shown in Fig. 2(c), TASE contains two self-attention layers S-MSA and T-MSA for spatial and temporal modeling, and both of them are the reuse of multi-head self-attention layers. S-MSA first learns the spatial relationship among different sequences. Given the output of S-MSA with shape $(N+1) \times B \times d$, we reshape it into $B \times (N+1) \times d$ and then send it into T-MSA to learn the temporal relationship on the sequence dimension. When we replace TASE with the original ViT block, the total quantitative results drop by about 2% by comparing (a) and (c) in Table 2.

We lastly highlight the necessity of multi-stage training (MST) by removing FLD and training CLD with all other modules end-to-end. By comparing the first items of (b) and (d) in Table 2, mixed end-to-end training leads to obvious performance degradation because the image encoder fails to extract effective features from fundus images. Therefore, all other modules should be pre-trained before training these two latent diffusion modules.

4 Conclusion

This paper proposes a novel glaucoma forecast method C2F-LDM based on sequential fundus images. The main difference with existing forecast methods is that C2F-LDM generatively predicts the future possible features of glaucoma in the latent space rather than directly predicting the probabilities of developing glaucoma. Then the predicted features are used for glaucoma detection and image reconstruction. Besides, C2F-LDM considers irregular time intervals and can predict the glaucoma status at any future time point by artificial setup, showing higher flexibility and interpretability for clinical scenarios. C2F-LDM

Table 2. Ablation studies of different components based on the quantitative evaluation.

	HKE	CLD	FLD	TASE	MST	ACC (%)	SEN (%)	SPE (%)	AUC (%)
(a)	✓	✓	✓	✓	✓	**94.4±0.5**	**93.8±0.7**	**94.6±0.6**	**95.5±0.5**
(b)	✓	✓	×	✓	✓	91.1±0.9	87.5±1.1	91.3±0.9	93.4±0.8
	✓	×	×	✓	✓	83.2±2.1	81.3±2.0	83.3±1.5	84.9±1.4
(c)	✓	✓	✓	×	✓	91.7±0.7	87.5±0.9	91.9±0.8	92.8±0.6
(d)	✓	✓	×	✓	×	78.3±3.3	68.8±2.4	78.8±1.9	80.3±2.2

can also be extended for other forecast tasks on sequential medical images. C2F-LDM still has some limitations. Firstly, C2F-LDM cannot be trained end-to-end and multi-stage training is necessary for better performance. Secondly, C2F-LDM may be inability to generalize to unseen data.

Acknowledgments. This study was funded in part by the Research Grants Council of the Hong Kong Special Administrative Region, China, under Project T45-401/22-N; in part by Hong Kong Innovation and Technology Fund under Project MHP/085/21. This study was also funded by the National Natural Science Foundation of China under Grants (62202408).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22563–22575 (2023)
2. Croitoru, F.A., Hondru, V., Ionescu, R.T., Shah, M.: Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
3. De Vente, C., Vermeer, K.A., Jaccard, N., Wang, H., Sun, H., Khader, F., Truhn, D., Aimyshev, T., Zhanibekuly, Y., Le, T.D., et al.: Airogs: Artificial intelligence for robust glaucoma screening challenge. IEEE Transactions on Medical Imaging (2023)
4. Diaz-Pinto, A., Morales, S., Naranjo, V., Köhler, T., Mossi, J.M., Navea, A.: Cnns for automatic glaucoma assessment using fundus images: an extensive validation. Biomedical engineering online **18**, 1–19 (2019)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

6. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
7. Fan, R., Bowd, C., Brye, N., Christopher, M., Weinreb, R.N., Kriegman, D.J., Zangwill, L.M.: One-vote veto: Semi-supervised learning for low-shot glaucoma diagnosis. IEEE Transactions on Medical Imaging (2023)
8. He, A., Li, T., Li, N., Wang, K., Fu, H.: Cabnet: Category attention block for imbalanced diabetic retinopathy grading. IEEE Transactions on Medical Imaging **40**(1), 143–153 (2020)
9. Hemelings, R., Elen, B., Barbosa-Breda, J., Lemmens, S., Meire, M., Pourjavan, S., Vandewalle, E., Van de Veire, S., Blaschko, M.B., De Boever, P., et al.: Accurate prediction of glaucoma from colour fundus images with a convolutional neural network that relies on active and transfer learning. Acta ophthalmologica **98**(1), e94–e100 (2020)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
11. Hu, X., Zhang, L.X., Gao, L., Dai, W., Han, X., Lai, Y.K., Chen, Y.: Glim-net: chronic glaucoma forecast transformer for irregularly sampled sequential fundus images. IEEE Transactions on Medical Imaging (2023)
12. Juneja, M., Thakur, S., Uniyal, A., Wani, A., Thakur, N., Jindal, P.: Deep learning-based classification network for glaucoma in retinal images. Computers and Electrical Engineering **101**, 108009 (2022)
13. Kamal, M.S., Dey, N., Chowdhury, L., Hasan, S.I., Santosh, K.: Explainable ai for glaucoma prediction analysis to understand risk factors in treatment planning. IEEE Transactions on Instrumentation and Measurement **71**, 1–9 (2022)
14. Li, L., Wang, X., Xu, M., Liu, H., Chen, X.: Deepgf: Glaucoma forecast using the sequential fundus images. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23. pp. 626–635. Springer (2020)
15. Li, L., Xu, M., Wang, X., Jiang, L., Liu, H.: Attention based glaucoma detection: A large-scale database and cnn model. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10571–10580 (2019)
16. Lin, M., Liu, L., Gorden, M., Kass, M., Tassel, S.V., Wang, F., Peng, Y.: Multi-scale multi-structure siamese network (mmsnet) for primary open-angle glaucoma prediction. In: International Workshop on Machine Learning in Medical Imaging. pp. 436–445. Springer (2022)
17. Rahman, A., Valanarasu, J.M.J., Hacihaliloglu, I., Patel, V.M.: Ambiguous medical image segmentation using diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11536–11546 (2023)
18. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
19. Schramowski, P., Brack, M., Deiseroth, B., Kersting, K.: Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22522–22531 (2023)
20. Singh, L.K., Khanna, M., et al.: A novel multimodality based dual fusion integrated approach for efficient and early prediction of glaucoma. Biomedical Signal Processing and Control **73**, 103468 (2022)

21. Singh, L.K., Pooja, Garg, H., Khanna, M.: Deep learning system applicability for rapid glaucoma prediction from fundus images across various data sets. Evolving Systems **13**(6), 807–836 (2022)
22. Tham, Y.C., Li, X., Wong, T.Y., Quigley, H.A., Aung, T., Cheng, C.Y.: Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. Ophthalmology **121**(11), 2081–2090 (2014)
23. Yang, T., Zhu, Y., Xie, Y., Zhang, A., Chen, C., Li, M.: Aim: Adapting image models for efficient video action recognition. arXiv preprint arXiv:2302.03024 (2023)
24. Yang, Y., Fu, H., Aviles-Rivero, A.I., Schönlieb, C.B., Zhu, L.: Diffmic: Dual-guidance diffusion network for medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 95–105. Springer (2023)
25. Yu, S., Ma, K., Bi, Q., Bian, C., Ning, M., He, N., Li, Y., Liu, H., Zheng, Y.: Mil-vt: Multiple instance learning enhanced vision transformer for fundus image classification. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24. pp. 45–54. Springer (2021)
26. Zbinden, L., Doorenbos, L., Pissas, T., Huber, A.T., Sznitman, R., Márquez-Neila, P.: Stochastic segmentation with conditional categorical diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1119–1129 (2023)
27. Zhou, Y., Yang, G., Zhou, Y., Ding, D., Zhao, J.: Representation, alignment, fusion: A generic transformer-based framework for multi-modal glaucoma recognition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 704–713. Springer (2023)