Graph Attention U-Net for Retinal Layer Surface Detection and Choroid Neovascularization Segmentation in OCT Images

Yuhe Shen, Jiang Li, Weifang Zhu, Kai Yu, Meng Wang, Yuanyuan Peng, Yi Zhou, Liling Guan, and Xinjian Chen, *Senior Member, IEEE.*

***Abstract*—Choroidal neovascularization (CNV) is a typical symptom of age-related macular degeneration (AMD) and is one of the leading causes for blindness. Accurate segmentation of CNV and detection of retinal layers are critical for eye disease diagnosis and monitoring. In this paper, we propose a novel graph attention U-Net (GA-UNet) for retinal layer surface detection and CNV segmentation in optical coherence tomography (OCT) images. Due to retinal layer deformation caused by CNV, it is challenging for existing models to segment CNV and detect retinal layer surfaces with the correct topological order. We propose two novel modules to address the challenge. The first module is a graph attention encoder (GAE) in a U-Net model that automatically integrates topological and pathological knowledge of retinal layers into the U-Net structure to achieve effective feature embedding. The second module is a graph decorrelation module (GDM) that takes reconstructed features by the decoder of the U-Net as inputs, it then decorrelates and removes information unrelated to retinal layer for improved retinal layer surface detection. In addition, we propose a new loss function to maintain the correct topological order of retinal layers and the continuity of their boundaries. The proposed model learns graph attention maps automatically during training and performs retinal layer surface detection and CNV segmentation simultaneously with the attention maps during inference. We evaluated the proposed model on our private AMD dataset and another public dataset. Experiment results show that the proposed model outperformed the competing methods for retinal layer surface detection and CNV segmentation and achieved new state of the arts on the datasets.**

***Index Terms*—Retinal layer detection, Choroidal neovascularization segmentation, Graph attention U-Net.**

I. INTRODUCTION

AGE-RELATED macular degeneration (AMD) is one of the leading causes for blindness among elderly. About 8.7% of the population age 60 and older are fully blind because of AMD [1]. Choroid neovascularization (CNV) is a typical symptom among advanced AMD patients, consisting of new abnormal blood vessels between choroid and retinal pigment epithelium (RPE). The occurrence of CNV is highly related to the high concentration of vascular endothelial growth factors (VEGF) [2]. Recent research demonstrated that the most effective medical treatment for CNV is intravitreal injection of anti-VEGF medicine. This treatment can suppress the growth of CNV and reduce the area of fluid below the RPE [3]. However, effectiveness of the treatment varies from patient to patient, and it needs to be assessed appropriately to provide better health care and patient management [4].

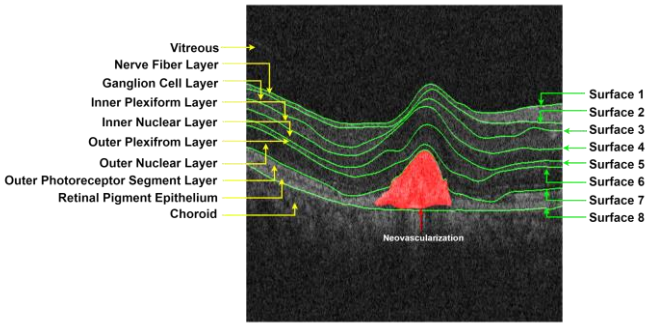

Fig. 1. An example OCT B-scan image with choroidal neovascularization

Optical coherence tomography (OCT) is a non-invasive, non-contact tomographic technique that visualizes cross-section of ultra-microscopic tissue structures for ophthalmic analysis [5]. It obtains a single one-dimensional scan (A-scan) of tissue by evaluating spectrum of the interference between reflected light and a fixed reference mirror. A two-dimensional image (B-scan) can be achieved by combining a series of A-scans horizontally, i.e., each column is an A-scan. The abnormality quantification and retinal layer thickness obtained in OCT images are often used for accurate diagnosis and monitoring of retinal disease including morphological changes in retinal layer structure due to CNV [6]. Fig. 1 shows a macular centered OCT B-scan image with CNV. The retinal layer surfaces are annotated as green lines and CNV is annotated as red. From top to bottom, there are eight layers including nerve fiber layer (NFL),

This study was supported in part by the National Key R&D Program of China (2018YFA0701700) and part by the National Nature Science Foundation of China (U20A20170, 61622114). Corresponding author: Xinjian Chen (xjchen@suda.edu.cn).

Yuhe Shen, Weifang Zhu, Yuanyuan Peng, Yi Zhou and Liling Guan are with the Medical Image Processing, Analysis and Visualization Lab, School of Electronics and Information Engineering, Soochow University, Jiangsu 215006, China (Email: yhshen@stu.suda.edu.cn).

Jiang Li is with Electrical and Computer Engineering, Old Dominion University, Norfolk, VA 23529.

Meng Wang and Kai Yu were with the Medical Image Processing, Analysis and Visualization Lab, School of Electronics and Information Engineering, Soochow University and now are with Institute of High Performance Computing, A*STAR, Singapore.

Xinjian Chen is with the School of Electronics and Information Engineering and the State Key Laboratory of Radiation Medicine and Protection, Soochow University, Jiangsu 215006, China.

ganglion cell layer (GCL), inner plexiform layer (IPL), inner nuclear layer (INL), outer plexiform layer (OPL), outer nuclear layer (ONL), external limiting membrane (ELM), outer photoreceptor segment layer (OPSL) and RPE.

Manually annotating retinal layers or CNV is time consuming and unrealistic in clinical situations. An automatic, reliable, and real-time segmentation method is desired for assessment of retinal layer thickness and CNV volume for AMD diagnosis [7, 8, 9]. Due to deformations among retinal layers caused by CNV, there are several challenges posed to traditional segmentation models. First, retinal layers have complex inner structures and usually are lack of clear boundary between layers in OCT images. Traditional pixel-wise segmentation methods including recent deep learning models have difficulties to reconstruct continuous and smooth layer surfaces [10]. Second, morphological deformation caused by CNV usually leads to failure of traditional graph-based methods for layer surface detection and CNV segmentation [11, 12].

In this paper, we propose a graph attention U-Net (GA-UNet) to address the challenges mentioned above. The proposed model consists of two novel modules. The first module is a graph attention encoder (GAE) that can automatically integrate topological and pathological knowledge of retinal layers to achieve effective feature embedding. The second one is a graph decorrelation module (GDM) that takes reconstructed features from decoder of the U-Net as inputs, it then decorrelates and removes information unrelated to retinal layer for improved retinal layer surface detection. A weighted mean square error (MSE) loss function is designed to maintain both the retinal layer topological order and its continuity during detection. The proposed model learns graph attention maps automatically during training and performs retinal layer surface detection and CNV segmentation simultaneously with the attention maps during inference. Our main contributions are:

1) A novel graph attention module is proposed in GA-UNet to learn topological information of retinal structure in OCT images that can achieve effective feature embedding for CNV segmentation and retinal layer surface detection.

2) A new graph decorrelation module is designed to maintain retinal layer interface information in the graph for improved retinal layer surface detection.

3) A new loss function is crafted for retinal layer surface detection by incorporating topological constraints of retinal layers into the loss function.

4) The proposed model achieves superb performances in retinal layer surface detection and CNV segmentation, establishing new state of the arts on our dataset.

II. Related Work

Retinal layer surface detection in OCT images has been extensively studied in the past decades, and a large number of automatic methods have been proposed and validated on patients with different retinal diseases. These methods can be categorized into two groups: traditional rule-based approaches led by graph search algorithms [13]-[20], and deep learning methods including pixel-wise classification and boundary regression [21]-[33], [40], [41].

A. Traditional Ruled-based Methods

Graph search and level-set. Methods based on graph search and level-set typically used an initial retinal layer surface segmentation as constraint for the final detection. The "Iowa Reference Algorithms" developed by Garvin et al. combined unary terms derived from filter responses with hard and soft constraints on different retinal layers to construct a graph for segmentation of retinal layers [13]. Song et al. proposed a 3-D graph-theoretic framework that incorporated both shape and context prior knowledge to penalize local shape and surface-distance changes for segmentation of retinal layers [14]. Dufour et al. developed a graph based multi-surface segmentation method for retinal layer segmentation [15]. The algorithm used soft constraints to add prior information from a learned model and achieved good performances in both normal and drusen OCT images. Novosel et al. designed a loosely coupled level-set method for segmentation of retinal layers and fluids in OCT images with central serous retinopathy [16]. Attenuation coefficients and thickness of different layers derived from OCT images with anatomical prior knowledge were used to constraint the algorithm.

Graph search with machine learning. Another subset of rule-based methods combined graph search with traditional machining learning algorithms. Lang et al. introduced a graph-cut based solution to infer retinal layers in OCT images and trained a random forest classifier to compute the unary term of an energy function to increase performance of the method [17]. Liu et al. utilized a random forest model to generate a probability map for retinal layer boundaries and optimized the algorithm by a fast level-set method to avoid disorder of layers for segmentation of retinal layers in macular-centered OCT images [18]. Xiang et al. trained a neural network model to generate initial retinal layer boundaries by twenty-four selected features and proposed an advanced graph search method to enhance the constraints between retinal layers and overcome morphological changes due to the occurrence of CNV. Finally, retinal layer surfaces and neovascularization were detected simultaneously [19]. One of the major limitations of these approaches is that these methods were built upon manually selected features or application specific graph parameters so that a finetuning step was almost always needed for new applications [20,21], which is time consuming and difficult especially for cases with pathology.

Traditional rule-based methods typically rely heavily on parameter tuning, which is prone to overfitting, leading to good performances on the data the models were tuned on but poor performances on unseen data. These methods are also computationally expensive.

B. Deep Learning Models

Retinal layer surface detection by pixel-wise classification. This category of algorithms treated the pixels belong to each retinal layer as a unique class and performed a pixel-wise classification to detect different layers. Most methods utilized fully convolutional network (FCN) [22] or U-Net [23] with the encoder-decoder structure as backbone for their models. For example, the ReLayNet model proposed by Roy et al. modified

the basic U-Net structure by replacing the deconvolution decoder branch with an unpooling layer [24]. It used indices from the max pooling layer in encoder to upscale feature maps at these positions in decoder while filling the remaining positions with zeros. Retinal layers and fluid were segmented with diabetic macular edema. The BRU-Net model, another U-Net variant, used a feature pyramid at each level in encoder to provide image information at each scale [25]. A soft constraint on anatomical retinal layer order was added to maintain topology of the layer structure. This method showed a much better performance in highly pathological scans with AMD than these by graph-based methods and U-Net.

Retinal layer surface detection by boundary regression. Methods identifying boundaries between layers without recognizing their classes had also been explored for retinal layer surface detection. Fang et al. successfully combined deep convolutional neural networks with a graph-based optimization algorithm to predict retinal layer surfaces in OCT images with non-exudative AMD but without CNV and fluid [26]. Gopinath et al. used an additional FCN model to identify edges and reproduced a final consistent segmentation with correct topological order [27]. He et al. proposed the SR-Net model by combining two cascaded deep networks for both retinal layer surface detection and microcystic macular edema (MME) segmentation [28], where a S-Net performed pixel-wise classification and a R-Net regressed retinal layer boundaries based on outputs from the S-Net. They improved their model by proposing a two-branch encoder-decoder framework for retinal layer surface detection in [29]. The first branch output a pixel-wise segmentation map for both retinal layers and lesions, and the second branch modeled the distribution of retinal layer surface positions and output positions of retinal layers in each A-scan.

The above methods improved performances of retinal layer surface detection and CNV segmentation by adopting graph optimization with deep learning models [26, 30]. We believe that if the shape and topology prior knowledge of retinal layer structure is implicitly incorporated, performances of the models can be further improved. In addition, few existing methods confused shape or intensity features with pathological structure features in the forward propagation [31, 32], which may lead to degraded performances.

III. Method

A. Overview Architecture

In this paper, we propose an attention graph convolutional neural network, GA-UNet, to segment both CNV and retinal layer surfaces in OCT images and the overall diagram is shown in Fig. 2. The backbone of our model is a graph attention U-Net, consisting of a graph attention encoder (GAE) and a decoder. The basic U-Net structure was developed in 2015 and has been widely used in many computer vision applications [23]. The proposed model also contains a graph decorrelation module (GDM) and a semantic topological constraint graph operator (STCGO). After pre-processing, an OCT image is fed into GA-UNet to achieve CNV segmentation. For retinal layer surface detection, the proposed model utilizes GDM to decorrelate the reconstructed feature maps of different resolutions from the decoder to generate topological constraints for retinal layer surface detection. Topological constraints derived from prior knowledge of retinal layer structure are also integrated by STCGO to learn features for retinal layer surface detection. In addition, a new loss function is crafted based on the outputs of GDM to maintain topological order and continuity of retinal layer surfaces. We will detail each module in the proposed model in the following subsections.

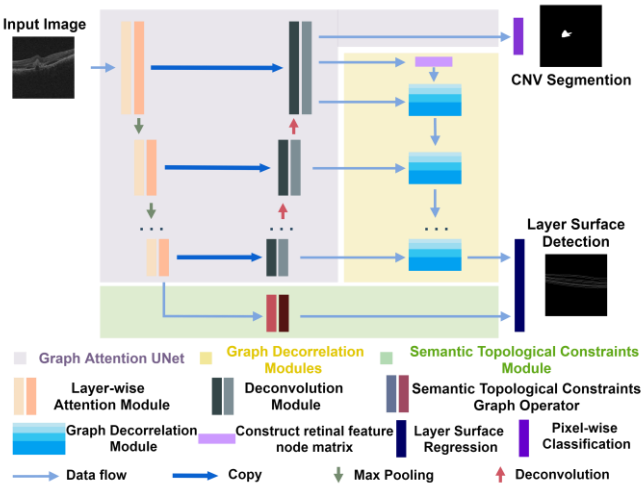

Fig. 2. Diagram of the proposed model. The proposed GA-UNet model takes an OCT image and produces retinal layer surface detection and CNV segmentation simultaneously. Topological constraints derived from prior knowledge of retinal layer structures are integrated by the semantic topological constraints graph operator (STCGO) to learn features for retinal layer surface detection through regression. The graph decorrelation module (GDM) decorrelates the reconstructed feature maps to generate constraints, and a novel loss function is designed to benefit from the constraints for better layer surface detection.

B. Preprocessing

OCT images contain artifacts caused by uncontrollable factors such as eye movement [11]. We apply the retinal boundary flattening and intensity normalization techniques to eliminate those artifacts and make spatial coordinates and intensities consistent among OCT images [28].

1) Retinal Boundary Flattening

Image flattening has been widely used to correct irregular displacements of retinal interfaces in OCT images [36]. We first identify Bruch's membrane (BM), the bottom interface between retina and vitreous, as reference. We then locate the bottom pixel as the reference in each column and shift down all other columns in the image so that the reference interface is flattened, i.e., to be a straight horizontal line in the image. BM is chosen as the reference interface because it is the surface under RPE in OCT images and it is often undamaged and clean.

2) Intensity Normalization

OCT scanners typically perform real-time intensity rescaling and averaging to enhance image contrast [34]. However, this process still cannot resolve the challenge of insignificant grayscale differences among different retinal tissues, which may lead to inefficient model training and incorrect retinal layer segmentation. We utilize the following equation to further enhance image contrast:

$$I_{N(i,j)} = \begin{cases} 1; & I_{i,j} = I_{max} \\ \frac{I_{i,j} - I_{min}}{I_{max} - I_{min}}; & I_{min} < I_{i,j} < I_{max} \\ 0; & I_{i,j} = I_{min} \end{cases} \quad (1)$$

where i and j denote pixel coordinates, $I_{i,j}$ and $I_{N(i,j)}$ are original and enhanced pixel intensities, I_{max} and I_{min} are maximal and minimal intensities in the image, respectively. After normalization, the intensity value range is $[0, 1]$.

C. Graph Attention Encoder

The proposed GA-UNet model consisting of an encoder, GAE, and a decoder as shown in Fig. 2. Unlike pure graph convolution based neural network such as GAU-Net [48] that adopted graph convolution layer to completely replace traditional CNN for feature extraction, GAE contains six layer-wise attention modules (LAM) of different spatial resolutions to extract features for retinal layer detection and CNV segmentation. Each LAM consists of three components as shown in Fig. 3 including 1) Feature decomposition, 2) Squeeze-and-Excitation (SE) [35], and 3) Layer-wise graph operator (LGO). Outputs of LAM are the composition of the outputs from both SE and LGO. The decoder of GA-UNet is a typical deconvolutional network and is briefly shown in Fig. 7 to save space. Here we describe LAM in detail as it is a building block of GA-UNet.

1) Feature Decomposition

Retinal layers are alternating bright and dark bands in OCT images while CNV tissue appears as groups of bright pixels between RPE and choroid layers. We decompose the input feature map into boundary and region components for effective feature representations of retinal layers and CNV tissue as follows (top left in Fig. 3).

We first learn a $1 \times 1 \times c$ convolution layer to squeeze the c-channel input feature map to a single-channel image. We then apply the SoftMax function to the image to obtain a positive valued boundary mask. By multiplying the boundary mask with each channel in the original input feature map, we obtain a boundary enhanced feature map for LGO to extract topological features for both retinal layer detection and CNV segmentation.

To enhance region features for CNV segmentation, we first generate a boundary mask using the same way as described above. We then subtract the boundary mask from a same-sized all-one matrix and multiply the resulted matrix with each channel in the original input feature map to enhance region features embraced by layer boundaries. The enhanced feature map is then input to SE for further enhancement.

2) Squeeze-and-Excitation

The SE module is a channel attention strategy that allows the network to selectively enhance or suppress feature channels [35]. We adopt three convolution layers, a global average pooling layer and two fully connected layers in SE (bottom in Fig. 3) to selectively enhance channels in the input feature map.

3) Layer-wise Graph Operator

Traditional convolution layers operate on a regular grid and are isotropic in responding to patterns from all directions. However, retinal layer surfaces are highly anisotropic having strong horizontal structures in OCT images. We define a layer-wise graph operator (LGO) to leverage this prior knowledge as shown at top right in Fig. 3.

3.1 Graph Definition

We formulate the graph as $G(N, E)$, where N represents the node set of the graph, which is a $w \times h \times c$ feature map volume and each row image is a node of size $w \times 1 \times c$. E represents edges connecting nodes and it is expressed as a trainable adjacency matrix A of size $h \times h$. The graph has $|h|$ nodes in total. LGO takes a feature map from the *Feature Decomposition* block as input and generates a new feature map through graph attention convolution as shown in Fig. 3. The trainable adjacency matrix A can be treated as a global attention matrix such that a node can be enhanced by its neighbors if they are similar. Note that neighbors are not limited to those that are physically adjacent. A node could be a neighbor of any node in N since A is a dense matrix. In addition, traditional graph-based methods for image segmentation typically treat each pixel as a node, leading to a huge graph with much higher complexity. Our design significantly reduces the complexity.

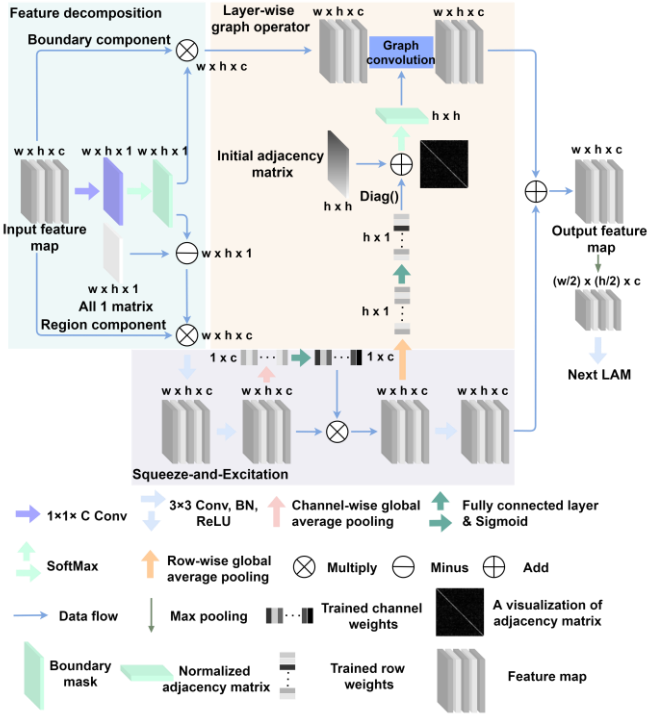

Fig. 3. Structure of layer-wise attention module (LAM). The encoder, GAE, of the proposed GA-UNet model consists of six LAMs. Each LAM has three components including feature decomposition, layer-wise graph operator (LGO) and Squeeze-and-Excitation (SE).

3.2 Attention Adjacency Matrix

We build attention into the adjacency matrix A and the attention consists of two parts: attention from all other nodes (reflected by the off-diagonal components in A) and attention from the node itself (diagonal components in A). These adjacency matrices are randomly initialized with 0.1 or 1, and the diagonal components are set to zero, representing attention from all other nodes excluding itself (Fig. 4). Attention from nodes themselves are learned through a separate path from the outputs of the SE block (Fig. 3). We first apply row-wise global pooling to the feature map output by SE of size $w \times h \times c$ to obtain a vector of size $h \times 1$, and then utilize a two-layer fully

connected network to achieve attention for each node. Finally, we diagonalize the attention vector to make its size as $h \times h$ and add it to the initial adjacency matrix A to obtain the final attention adjacency matrix for graph convolution.

The attention adjacency matrices in LGOs are trainable and each value in the matrix represents the similarity between two nodes in the input feature map. Fig. 4 shows a trained adjacency matrix with size of 128 × 128 from an LGO module. Most elements have values close to 0 (black) meaning no similarity. Diagonal components have large values, representing high self-similarities.

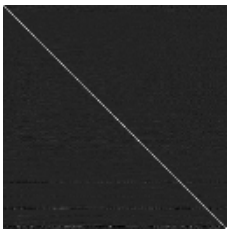

Fig. 4. Example of a trained adjacency matrix with size of 128 × 128 from LGO.

3.3 Graph Convolution

Assume the input feature map is denoted as X, the graph convolution is performed as,

$$\widetilde{A_{ij}} = \frac{\exp(A_{ij})}{\sum_{j=1}^{w} \exp(A_{ij})}, X' = \sigma(\tilde{A}XW) \quad (2)$$

$\widetilde{A_{ij}}$ represents the attention coefficient between the ith node and the jth node in the graph. We adopt ReLU [38] for the activation function $\sigma()$, W is a trainable weight matrix. Different from standard graph convolution, we choose the *SoftMax* function to perform normalization on each row in the matrix so that the sum of all edge weights connecting to each node is 1.

The matrix multiplication operation in (2) performs a weighted summation over the nodes in the feature map X using the weights given by the normalized adjacency matrix $\tilde{A}$ and the learnable matrix W. If a node is highly similar to another node, the component in the adjacency matrix corresponding to the two nodes have a larger value. Therefore, a node will be enhanced by similar nodes in the graph convolution step. Both the adjacency matrix $\tilde{A}$ and W are trainable so similar nodes can be found through training.

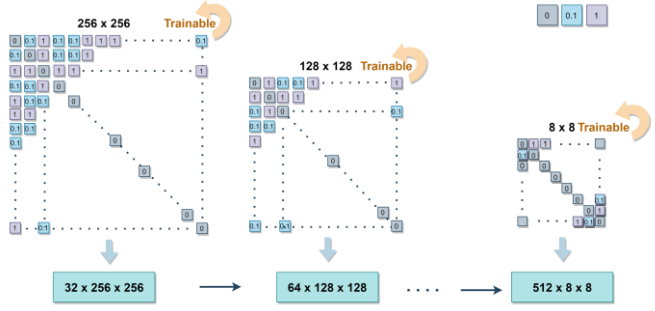

Fig. 5. Example initial adjacency matrices in the six LAMs of sizes 256 × 256, 128 × 128, 64 × 64, 32 × 32, 16 × 16 and 8 × 8. The corresponding nodes in the six LAMs have sizes of 256 × 256 × 32, 128 × 128 × 64, 64 × 64 × 128, 32 × 32 × 256, 16 × 16 × 512 and 8 × 8 × 512, respectively.

Outputs of LAM are the composition of feature maps from graph convolution and outputs from SE as shown in Fig. 3. The feature map composition is then max pooled with a factor of 2 × 2 and passed to the next LAM. There are skip connections to copy features from GAEs directly to the graph decoder at the same resolution level as shown in Fig. 2. Example heatmaps generated based on enhanced feature maps by graph attention convolution are shown in Fig. 6.

D. Semantic Topological Constraint Graph Operator

The last LAM in the GAE module outputs a feature map volume of size 8 × 8 × 512 having the lowest spatial resolution. There are 8 retinal layers in OCT images, and they follow certain rules. For example, NFL is adjacent to GCL and vitreous only. INL can only be adjacent to IPL and OPL. We design the STCGO (Fig. 7) module to integrate the topological knowledge into the adjacency matrix A as shown at bottom of Fig. 2 to learn feature representations for retinal layer regression. STCGO has the same structure as LGO (Fig. 3) except that the adjacency matrix A in STCGO is fixed and the self-attention matrix is the identity matrix, while the adjacent matrices in LGO are trainable. The highly regulated features from STCGO will be used for subsequent retinal layer surface detection through regression.

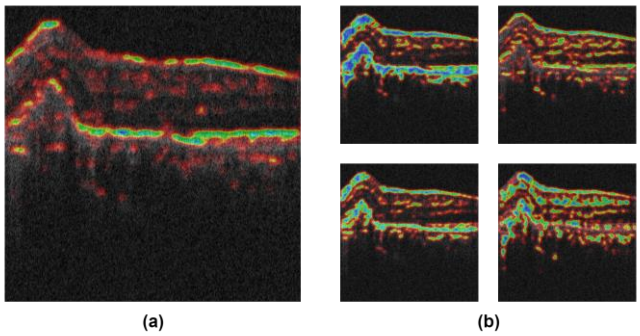

Fig. 6. Examples of enhanced feature maps by LAMs shown as heatmaps. Retinal layer surface structures are enhanced. (a) One enhanced feature map of size 256 × 256. (b) Four feature maps of size 128 × 128.

E. Graph Decorrelation Module

1) Architecture of GDM

GDM takes outputs from the decoder in GA-UNet as inputs (left in Fig. 8). The decoder has six levels of different resolutions, and each level receives two inputs: one is from the below level through regular deconvolution, and another is a direct copy from the LAM encoder at the same level. GDM (right in Fig. 8) is used to learn topological constraints for layer surface detection, and each resolution level in GA-UNet has a GDM module. All the six GDMs are connected as a hierarchy and output a three-dimensional array of size 8 × 512 × 8, which is used in the loss function as topological constrains described in Section III G.2. The objective of GDM is to remove non retinal layer boundary pixels in the feature map volume and to use the remaining pixels in the feature volume as constraints for retinal layer surface regression.

2) Decorrelation Pooling

Using the highest resolution level as example shown in Fig. 8, the inputs of GDM consist of two components. One is the reconstructed feature map by GA-UNet of size 512 × 512 × 16, another is generated by the 1 × 1 × 8 convolution layer denoted

as G_p and can be seen as a coarse prediction map for retinal layer tissues. There are 512 × 512 pixels and 16 learned features for each pixel in the reconstructed feature map. Among these 512^2 pixels, some pixels are located on retinal layer interfaces and are our target pixels to keep. Others are located between retinal layers' regions and will be eliminated by GDM.

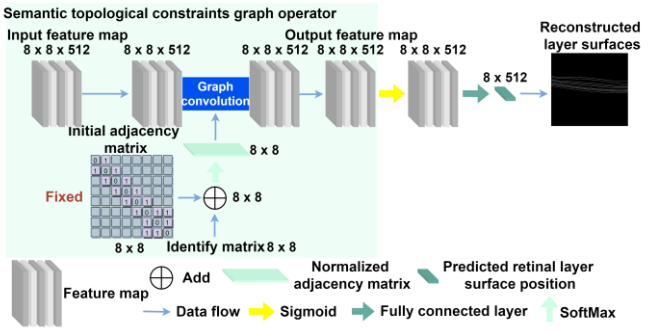

Fig. 7. Architecture of STCGO. STCGO has the same structure as LGO shown in Fig. 3 except that it has a fixed adjacency matrix. Under the constraints of the fixed adjacent matrix, features of a retinal layer will be enhanced only by those features from the specified nearby layers during graph convolution, and the highly regulated features from STCGO will be used for subsequent layer surface detection.

GDM divides the reconstructed feature map element-wisely by a same-sized volume obtained by a 3 × 3 convolution layer. It is then combined as a 512 × 512 image by a 1 × 1 × 16 convolution kernel, and pixels belong to retinal layers have larger values in the image. After sigmoid function being applied to each pixel, we rank pixels by the values in each column and only keep the half top ranked pixels making the size of the image as 256 × 512. Then, we normalize the image so that each row sums to 1. Finally, we multiply this image of size 256 × 512 with G_p of size 512 × 512 × 8 to generate a tensor product of size 256 × 512 × 8, which is the decorrelated feature volume. After six GDMs, the original retinal feature node set G_p of size 512 × 512 × 8 is transformed into a retinal surface node set G_{layer} of size 8 × 512 × 8 only containing nodes fall on the 8 retinal layer surfaces, which is sent to the novel loss function as constraints for retinal layer surface detection.

F. CNV Segmentation and Layer Surface Detection

1) CNV Segmentation

CNV segmentation is performed by the GAE module and the decoder in GA-UNet. First, we preprocess OCT images by flattening retinal layers with BM as reference and normalizing pixel values to the range of [0, 1]. Then, the preprocessed OCT images are sent to GA-UNet for CNV segmentation. Note that the proposed GA-UNet performs semantic segmentation for CNV in the same way as the traditional U-Net except that the encoder of GA-UNet utilizes the attention mechanism to achieve effective feature embedding.

2) Layer Surface Detection

Retinal layer surface detection is performed by GA-UNet, GDM and STCGO. Same as CNV segmentation, OCT images are first preprocessed and are sent to GA-UNet for processing. GDM then takes the feature maps reconstructed by GA-UNet at different resolution levels and produces decorrelated feature maps, which are learned topological constraints from the retinal layers. Finally, STCGO takes the lowest resolution feature map extracted by LAM, and results produced by STCGO are used to regress layer surfaces with the guidance provided by the constraints from GDM through the proposed loss function described below.

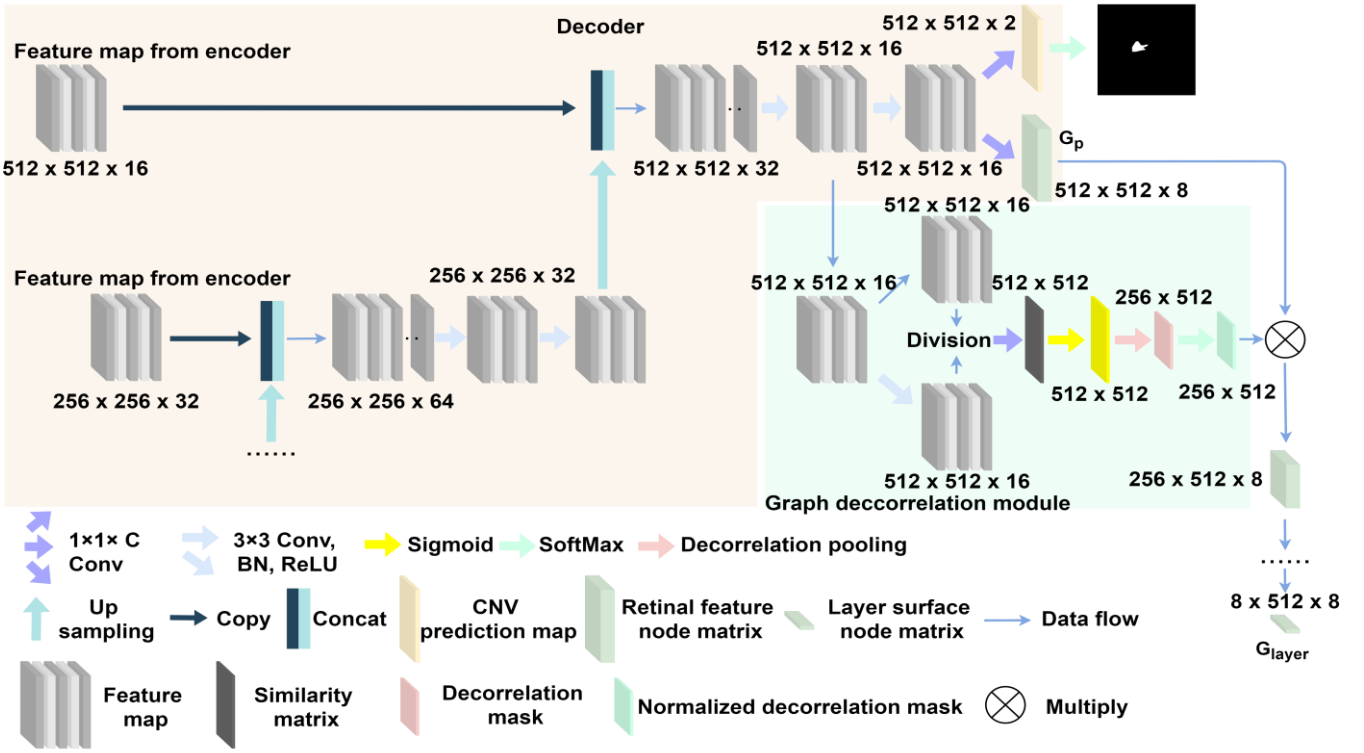

Fig. 8. Architecture of the GDM and the decoder of the GA-UNet. One GDM consists of a 3 × 3 convolution layer, a 1 × 1 convolution layer, a sigmoid layer, a decorrelation pooling layer and a SoftMax layer. The green area shows one example GDM at the highest resolution level. There are six GDMs in the proposed model.

G. Loss Functions

The proposed model is trained by jointly optimizing the following loss functions:

1) Loss Function for CNV Segmentation

CNV segmentation is a pixel-wise binary classification task, and we utilize the binary cross-entropy loss for training,

$$\mathcal{L}_{CE} = -(\sum_{x\in X} p(x)log(q(x)) + (1-p(x))log(1-q(x))) \quad (3)$$

where $p(x)$ represents the class label for pixel x, which is 1 for CNV otherwise 0, $q(x)$ is the estimated probability of pixel x belong to CNV. The *Dice* loss is used to evaluate spatial overlap between ground truth and predicted CNV area. The total loss of CNV segmentation is the combination of $\mathcal{L}_{CE}$ and $\mathcal{L}_{dice}$ [24] as,

$$\mathcal{L}_{dice} = 1 - \frac{2\sum_{x\in X} p(x)q(x)}{\sum_{x\in X} p(x)+\sum_{x\in X} q(x)} \quad (4)$$

$$\mathcal{L}_{CNV} = \mathcal{L}_{CE} + \mathcal{L}_{dice} \quad (5)$$

2) Loss Function for Retinal Layer Surface Detection

We utilize a weighted mean square error loss for retinal layer surface detection,

$$\mathcal{L}_{surface} = \frac{1}{h\times w}\sum_{i=1}^{h=8}\sum_{j=1}^{w=512}\omega_{i,j}(s_i(j)-r_i(j))^2 \quad (6)$$

where $s_i(j)$ and $r_i(j)$ represent ground truth and predicted surface position of surface i in the jth A-scan, $\omega_{i,j}$ is the weight assigned to pixel $x_{i,j}$ and is defined as,

$$\omega_{i,j} = 2 - (\omega_{vertical} + \omega_{horizontal}) + \omega_{min} \quad (7)$$

$$\omega_{vertical} = 1 - \overrightarrow{G_{layer}(\imath,\jmath)}^T \cdot \overrightarrow{G_{layer}(\imath+1,\jmath)} \quad (8)$$

$$\omega_{horizontal} = \overrightarrow{G_{layer}(\imath,\jmath)}^T \cdot \overrightarrow{G_{layer}(\imath,\jmath+1)} \quad (9)$$

where G_{layer} is the decorrelated feature map from GDM of size $8 \times 512 \times 8$. $\overrightarrow{G_{layer}(\imath,\jmath)}$ is a learned feature vector of 8 dimensions and there are 8×512 such feature vectors. Ideally, if two feature vectors belong to the same layer, the correlation between them should be one and otherwise 0. $\omega_{vertical}$ and $\omega_{horizontal}$ are the vertical and horizontal correlation (dot product) of the decorrelated map at $G_{layer}(i,j)$, and ω_{min} is a minimal correlation which is set as 0.0001. After convergence, $\omega_{vertical}$ and $\omega_{horizontal}$ are maximized, forcing each row in G_{layer} to represent a separate retinal layer for ideal case. The final training loss is the combination of the CNV segmentation loss and the retinal layer surface detection loss and both tasks share the same weight [29],

$$L_{GA-UNet} = L_{CNV} + L_{surface} \quad (10)$$

IV. Experiment Setup

A. Datasets

The study was approved by the Institutional Review Board of Soochow University, and informed consents were obtained from all subjects. The proposed method was validated on two datasets. The first one is a private dataset containing 51 macular OCT scans from AMD patients acquired by a Cirrus HD-OCT 4000 scanner at the Joint Shantou International Eye Center. Each scan consists of 128 B-scans of size 512 × 1024. The lateral and axial resolutions and B-scan separation are 11.74 μm, 1.96μm, and 47.24μm, respectively. Eight retinal layer surfaces and CNV in each OCT image were manually annotated by two retinal specialists. Each expert annotated half of the dataset and reviewed the annotated data by another expert. Any annotation disagreement was resolved by discussion till a consensus was reached. We randomly divided the 51 OCT scans consisting of 6528 B-scans into 5 folds and conducted 5-fold cross-validation (CV) for model comparison. This dataset was used for both retinal layer surface detection and CNV segmentation.

The second OCT dataset is publicly available consisting of OCT scans from 35 subjects (14 healthy controls (HC) and 21 subjects with multiple sclerosis (MS)) [42]. Each OCT scan consists of 49 B-scans of size 496 × 1024. The depth resolution is 3.9μm. Nine retinal layer surfaces were manually annotated. We divided 30 of the 35 OCT scans (12 HC + 18 MS) with 1470 B-scans into 5 folds and conducted 5-fold cross-validation to fine-tune parameters for each model in comparison. Once the parameters were fine-tuned, the model was retrained with all the 30 OCT scans and the trained model was applied to the remaining 5 OCT scans with 245 B-scans for testing. This dataset was used for retinal layer surface detection but not for CNV segmentation since CNV annotation is not available.

B. Implementation Details

The proposed method was implemented with PyTorch and was trained with one NVIDIA 3070 GPU. For the AMD dataset, the initial learning rate was set to 0.00025 and was reduced to 0.0001 after 50 epochs, and further down to 0.000025 after another 100 epochs. The number of training epochs was set to 200. Momentum and weight decay coefficients were set to 0.999 and 0.0001, respectively. For the public dataset, the initial learning rate was also set to 0.00025 and was reduced to 0.0001 after 100 epochs, and after another 100 epochs the learning rate was reduced to 0.00025. The number of training epochs was set to 300. Momentum and weight decay coefficients were also set to 0.999 and 0.0001, respectively. The SGD optimizer was used in training for both datasets and the mini batch size was set to 2.

C. Ablation Study

The proposed model consists of four components and one customized loss function including the basic U-Net, GAE, GDM, STCGO and the weighted MSE loss. However, the weighted MSE loss is defined on the outputs of GDM, so it is not independent. Therefore, we combined GDM and the weighted MSE loss as one component and investigated the contribution of each of the four components in the ablation study. Eight experiments were conducted for retinal surface detection and CNV segmentation, respectively, including 1) U-Net, 2) GAE+U-Net, 3) STCGO+U-Net, 4) GAE+U-Net+STCGO, 5) GDM+U-Net, 6) GAE+GDM+U-Net, 7) GDM+U-Net+STCGO, and 8) GAE+GDM+U-Net+STCGO, where the last one is the proposed model named as GA-UNet. As the weighted MSE loss was combined with GDM, if GDM was included in the ablution study, it means that the weighted MSE loss was utilized. Otherwise, the regular MSE loss was employed in the experiment. We also conducted experiments with different weights in the overall loss function for both tasks.

TABLE I

AVERAGE USPES OF RETINAL LAYER DETECTION ON THE AMD DATASET IN ABLATION STUDY (MEAN(STD) μm),'*' INDICATES THAT THE RESULT IS SIGNIFICANTLY WORSE THAN THAT BY GA-UNET

Method/Surface	1	2	3	4	5	6	7	8	Overall
U-Net	3.86(3.41)	3.37(3.93)	4.32(3.88)	4.27(3.35)	4.21(3.25)	3.99(3.13)	3.83(3.18)	3.78(3.92)	3.95(3.50)
GAE+U-Net	2.49(1.80)	2.39(2.13)	3.56(1.35)	4.05(0.97)	3.69(0.80)	3.17(1.24)	3.35(1.37)	2.65(1.05)	3.17(1.34)
STCGO+U-Net	2.85(2.34)	3.00(2.14)	3.70(1.85)	4.10(1.29)	3.84(1.12)	3.58(1.28)	3.45(1.44)	2.84(1.36)	3.42(1.60)
GAE+STCGO+U-Net	2.39(1.71)	2.23(2.01)	3.41(1.26)	3.92*(0.83)	3.47*(1.03)	3.02(1.12)	3.15(1.25)	2.44*(0.85)	3.00(1.26)
GDM+U-Net	2.45(1.77)	2.85(1.91)	3.37(1.63)	4.13(0.98)	4.01(0.94)	4.15(0.83)	3.83(1.54)	2.85(0.85)	3.45(1.31)
GAE+GDM+U-Net	2.27(1.46)	2.16(1.75)	3.15(1.16)	3.81*(0.92)	3.25*(0.92)	3.99(0.99)	3.06(1.32)	2.39*(0.88)	2.89(1.17)
GDM+STCGO+U-Net	2.34(1.34)	2.27(1.66)	3.24*(0.96)	3.44*(0.78)	3.41*(0.71)	3.03*(0.82)	3.11(1.20)	2.46*(0.93)	2.92(1.05)
GA-UNet (Proposed)	2.16(1.15)	1.89(1.16)	2.40(0.96)	3.08(0.81)	2.39(0.58)	2.79(0.62)	2.71(0.97)	1.98(0.65)	2.42(0.86)

TABLE II

AVERAGE DSCS OF CNV SEGMENTATION ON THE AMD DATASET IN ABLATION STUDY (MEAN(STD)%) ,'*' INDICATES THAT THE RESULT IS SIGNIFICANTLY WORSE THAN THAT BY GA-UNET

Method	DSC	Accuracy	Precision	Sensitivity	Specificity
U-Net	89.90(1.95)	99.13(0.07)	88.10 (0.57)	89.97(1.81)	99.27(0.27)
GAE+U-Net	93.84(0.66)	99.42(0.04)	92.62(0.58)	94.13(0.69)	99.62(0.27)
STCGO+U-Net	90.39(1.76)	99.18(0.05)	88.64(0.64)	90.42(0.95)	99.32(0.32)
GAE+STCGO+U-Net	93.91(0.63)	99.47 (0.06)	92.73(0.60)	94.33(0.76)	99.67(0.91)
GDM+U-Net	90.66(0.67)	99.36(0.07)	90.27(0.56)	90.88(0.70)	99.44(0.28)
GAE+GDM+U-Net	94.07(0.47)	99.46(0.05)	92.95(0.60)	94.52(0.49)	99.80(0.23)
GDM+STCGO+U-Net	91.60*(0.71)	99.40*(0.06)	91.34*(0.63)	92.04*(0.53)	99.49*(0.35)
GA-UNet (Proposed)	94.14(0.51)	99.54(0.04)	93.34(0.45)	94.99 (0.56)	99.87(0.21)

D. Comparison Study

For retinal surface detection, we compared our model with seven state-of-the-art methods including 1) AUtomated Retinal Analysis tools (AURA) [17], a graph search and random forest-based model for retinal surface detection. 2) SR-Net [28] or Topology Guaranteed F-CNN (TG-FCNN) [29]. SR-Net is a deep neural network consisting of a segmentation network and a regression network and TG-FCNN is the updated version of SR-Net, we reported one of the two whichever was better. 3) RelayNet [24], a unique encoder-decoder architecture to preserve spatial information in feature reconstruction. 4) NNCGS [19], a constrained graph search algorithm with a neural network feature extractor. 5) Deep Learning-Shortest Path (DL-SP) [39], a U-shape network for pre-segmentation with shortest path for layer surface estimation. 6) CNN-S [30], a CNN based regression model for multiple layer surface detection and 7) Multi-scare CNN combined graph searching (multi-CNN GS) [32], a multi-scare deep learning architecture which learned graph-edge weight and optimization with graph searching.

For CNV segmentation, eight state-of-the-art deep learning segmentation methods and an advanced graph search method were compared with our model including 1) CE-Net [43], an encoder-decoder model with a context extractor for medical image segmentation. 2) Attention U-Net [44], an encoder-decoder architecture with an attention gate module for medical image segmentation. 3) IA-Net [45], an attention deep learning model for CNV segmentation. 4) Trans-UNet [46], a full transformer-based U-shape architecture for medical image segmentation. 5) Swin-UNet [47], a Swin-transformer based U-Net for medical image segmentation. 6) SR-Net [28] or TG-FCNN [29] whichever was better, 7) RelayNet [24] and 8) NNCGS [19]. All the studies utilized the same retinal boundary flatten strategy and the intensity normalization step.

E. Performance Metrics

We used the unsigned surface position error (USPE), the Euclidean distance in the z-axis between the detected surface and its ground truth [20], and the unsigned maximum surface position error (UMSPE) to evaluate retinal surface detection. We utilized four performance metrices to evaluate CNV segmentation including Accuracy, Sensitivity, Specificity, Precision and Dice similarity coefficient (DSC) [20]. Standard deviation for each performance metric was computed across data samples. Paired t-test was conducted for both CNV segmentation and retinal surface detection to test if performance differences among different models are significant, i.e., p-value is less than 0.05.

V. RESULTS

A. Results of Ablation Study

Ablation study was conducted on the AMD dataset through 5-fold CV and results of surface detection are shown in Table I. We performed paired t-test between results by the proposed method (GA-UNet) against those obtained in experiments 4), 6), 7) where only one module was removed from GA-UNet in each of the experiments, and '*' indicates that the difference was significant. The backbone U-Net alone did not perform well achieving a mean USPE of 3.95μm. GAE reduced the mean USPE to 3.17μm if combined with U-Net. Adding STCGO further reduced USPE to 3.00μm. GDM also reduced the mean USPE to 3.45μm if combined with U-Net. With this

configuration, using GAE to replace the encoder of U-Net decreased USPE to 2.89μm. Meanwhile, adding STCGO to the combination of GDM and U-Net reduced USPE to 2.92μm. Overall, the final proposed model GA-UNet utilized all the components achieved the best mean USPE of 2.42μm. In summary, all components are necessary for the final proposed model, but GAE contributed the most. The weighted MSE loss also played an important role, making GA-UNet significantly better.

The 5-fold CV ablation study results on the AMD dataset for CNV segmentation are shown in Table II. An '*' indicates results in experiments 4), 6) and 7) are significantly worse than these achieved by GA-UNet. Only GAE is necessary and it enhanced the backbone U-Net by large margins in terms of all the five metrics. GDM and STCGO also contributed to CNV segmentation but not significantly. For example, in terms of DSC, GAE improved upon U-Net by 3.94% from 89.90% to 93.84% while STCGO and GDM only improved by 0.49% and 0.76%, respectively. Putting them together, the proposed model achieved the final DSC of 94.14%. Similar observations are observed for the other metrics as listed in Table II.

The ablation study on weights in the overall loss function shows that equal weight configuration achieved the best performance. If we change the weight of CNV segmentation task to 0.7and weight of layer surface detection to 0.3, the mean USPE decreased to 2.53μm and the DSC decreased 0.08%. Meanwhile, if we change the weight of CNV segmentation task to 0.3 and weight of layer surface detection to 0.7, the mean USPE decreased to 2.49μm and the DSC decreased 0.12%. Both are slightly worse than the result by the loss function with equal weights.

B. Results of Retinal Layer Surface Detection

1) AMD dataset

Fig. 9 visualizes surface detection 5-fold CV results of two OCT B-scan images with CNV and Table III lists means and standard deviations of USPEs by different models. Note that TG-FCNN is the updated version of SR-Net but TG-FCNN did not perform well on this dataset and its results are not reported here. We also performed the paired *t*-test for the proposed method (GA-UNet) versus all other competing methods and '*' indicates that the differences are significant. If we look at these eight retinal layer surfaces individually, the proposed model won 55 cases out of the 56 head-to-head comparisons. The only lost case by the proposed model was on surface 1 where AURA won but the difference was not significant ($p > 0.05$). The 43 out of 55 winning cases had significant margins ($p < 0.05$). The UMSPEs are listed in Table IV where '/' represents a failure case during layer surface detection and UMSPE is too large and meaningless. All the peak errors of the proposed method are smaller than these by competing methods with a minimum difference of 0.31μm, except surface 8 where NNCGS achieved the smallest UMSPE of 5.34μm for surface 8.

Among these competing methods, SR-Net achieved the second-best overall performance and all others performed similarly with USPEs larger than 3.30μm. On these eight individual layer surfaces, AURA performed the best on surface 1. Layer surface 3-7 were more difficult to detect by most of the models including the proposed model. Detection of surfaces 1, 2 and 8 were less challenging for most methods and achieved smaller USPEs.

2) Public Dataset

The public dataset annotated 9 retinal layer surfaces. First, we selected the same 8 surfaces as those in the AMD dataset and trained the proposed model without changing the setting of these adjacency matrices. Table V lists means and standard deviations of USPEs for different models. Again, TG-FCNN performed better than its old version SR-Net and only TG-FCNN is reported. We performed paired *t*-test for the proposed method (GA-UNet) versus all other competing methods. Our proposed model won all the 56 head-to-head comparisons and 52 of them had significant margins ($p < 0.05$).

Second, we changed the settings of the adjacency matrices based on the prior knowledge of all the 9 retinal layer structures in the public dataset and conducted experiment to detect all the nine retinal layer surfaces. Means and standard deviations of USPEs are shown in Table VI. Still, our method outperformed the competing methods in all the 63 cases. However, 16 of these winning cases were not significant ($p < 0.05$). Fig. 10 shows examples of the detection results.

In the above two experiments, our proposed model, TG-FCNN and Multi-scale CNN+GS were always ranked among top three. Our method ranked the first twice, TG-FCNN [29] won the second twice and Multi-scale CNN+GS were always the third. Other methods all had USPEs larger than 2.90μm. In detection of the 9 retinal layer surfaces, our method was not significantly better than other methods especially on surface 4 and surface 7, with *p*-values larger than 0.05 in 8 out of 14 head-to-head comparisons. The USPEs on the public dataset were generally smaller than the USPEs on the AMD dataset, indicating that the public dataset was less challenging.

C. Results of CNV Segmentation

Since the public dataset does not have CNV annotation, CNV segmentation was only performed on the AMD dataset. Fig. 11 shows three examples of CNV segmentation, where ground truth CNV tissues were colored as red, segmentation regions as blue, and their overlaps as white. Table VII lists the means and standard deviations of DSC, Accuracy, Precision, Sensitivity and Specificity for all competing models. Again, SR-Net performed better than TG-FCNN and is reported here. Paired *t*-test of GA-UNet versus competing methods were performed. Though there are not much visual differences among these segmentation results, our proposed method achieved the best results in terms of all the five metrics and all the margins were statistically significant.

SR-Net and RelayNet are two deep learning frameworks for both retinal layer detection and CNV segmentation. These two multitask model achieved upstream performances in all the competing methods. IA-Net is an attention based deep learning model customized for CNV segmentation and it achieved third best performance. NNCGS is a traditional graph-based method and achieved the worst performance. Trans-UNet, Swin-UNet, CE-Net and Attention U-Net were much better than NNCGS

but were still significantly worse than the proposed model.

D. Computational Complexity

Table VIII lists number of parameters, number of floating-point operations (FLOPs) and the time required during inference to process one B-scan image by all the multi-task competing methods. As NNCGS is not a deep neural network, we only list its running time. Our proposed model has almost twice and triple number of parameters as these of TG-FCNN and SR-Net, respectively. However, our model has much less FLOPs, which represents the direct computational complexity, than the other methods. Our model only needs 0.04s to process one B-scan, the most efficient method among the multi-task competitors.

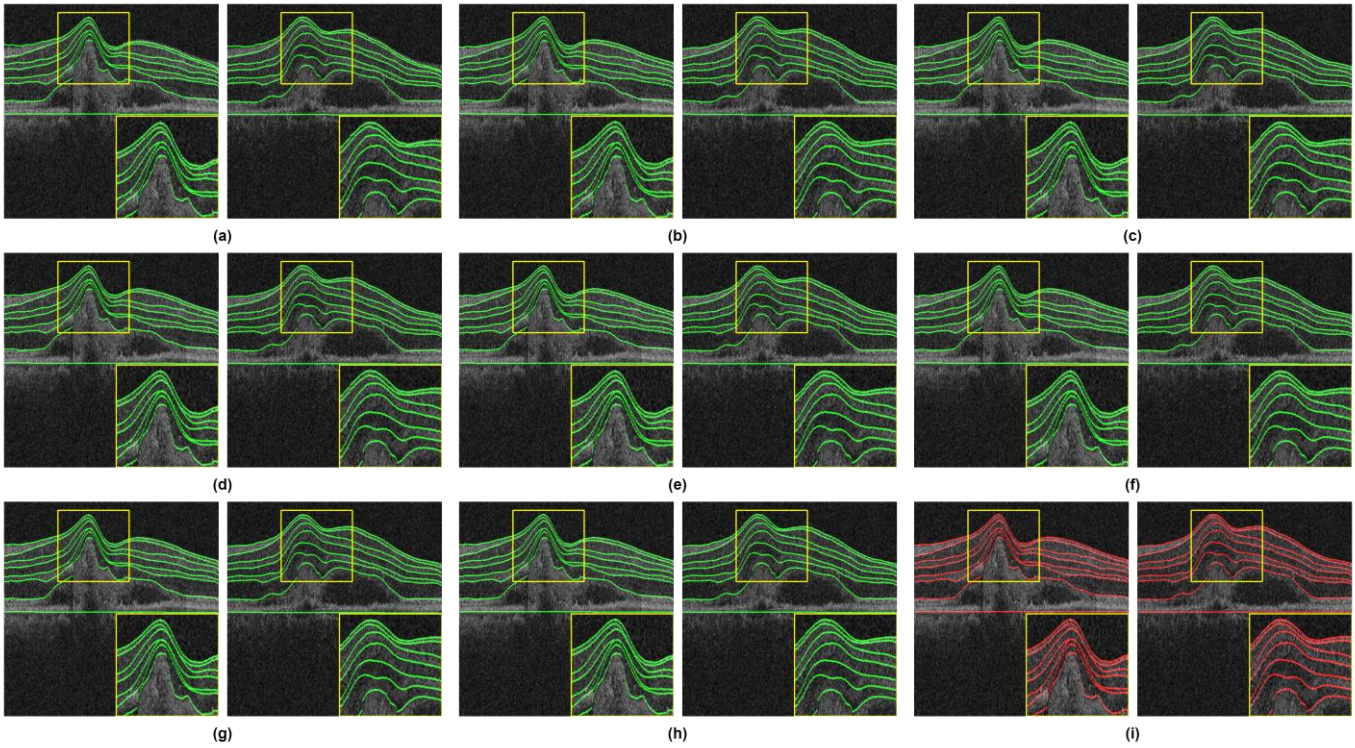

Fig. 9. Results of retinal layer surface detection on two OCT images with CNV. (a) NNCGS, (b) AURA, (c) SR-Net, (d) RelayNet, (e) DL-SP, (f) CNN-S, (g) Multi-scare CNN + GS, (h) Proposed method, (i) Ground truth.

TABLE III

AVERAGE USPES BY DIFFERENT METHODS IN RETINAL LAYER SURFACE DETECTION ON THE AMD DATASET (MEAN(STD) μm),'*' INDICATES THAT THE COMPARED MODEL IS SIGNIFICANTLY WORSE THAN GA-UNET

Method/Surface	1	2	3	4	5	6	7	8	Overall
NNCGS [19]	7.85*(7.98)	6.37*(4.85)	5.45*(3.66)	4.21*(3.78)	4.31*(2.38)	3.84*(1.55)	2.89(0.41)	2.11(0.25)	4.63(3.07)
AURA [17]	1.97(0.79)	3.41*(1.37)	3.44*(1.92)	4.97*(2.31)	5.29*(2.08)	4.88*(2.91)	5.82*(3.79)	3.39*(2.02)	4.15(2.15)
SR-Net [28]	2.95(1.65)	3.05*(2.05)	2.77(1.29)	4.08*(0.96)	4.11*(0.66)	3.31*(1.01)	2.82(1.46)	3.25*(0.79)	3.29(1.23)
RelayNet [24]	2.83(2.45)	3.07*(1.85)	3.58*(1.99)	4.02*(1.81)	4.51*(1.01)	3.42*(1.01)	3.63(2.40)	3.19*(1.13)	3.53(1.71)
DL-SP [39]	3.07(1.49)	3.02*(1.68)	3.01*(1.50)	3.75*(0.93)	4.46*(0.91)	4.39*(0.98)	4.05*(1.83)	3.16*(0.82)	3.61(1.27)
CNN-S [30]	3.06*(1.13)	3.28*(1.09)	3.81*(1.23)	3.94*(1.41)	4.45*(0.67)	3.72*(0.93)	3.63*(1.23)	3.20*(1.25)	3.64(1.12)
Multi-scare CNN +GS [32]	2.74(1.34)	2.89(1.41)	2.95*(1.11)	3.86(1.45)	4.15*(1.01)	3.23(1.16)	3.80*(1.49)	3.17*(1.05)	3.35(1.25)
GA-UNet (proposed)	2.16(1.15)	1.89(1.16)	2.40(0.96)	3.08(0.81)	2.39(0.58)	2.79(0.62)	2.71(0.97)	1.98(0.65)	2.42(0.86)

TABLE IV

UMSPES BY DIFFERENT METHODS IN RETINAL LAYER SURFACE DETECTION ON THE AMD DATASET (μm)

Method/Surface	1	2	3	4	5	6	7	8
NNCGS [19]	33.17	27.68	19.54	21.76	14.32	/	9.73	5.34
AURA [17]	7.62	10.45	11.37	14.56	17.39	17.91	20.21	15.62
SR-Net [28]	8.32	9.17	8.29	8.99	9.93	10.72	8.91	11.07
RelayNet [24]	9.11	10.34	8.72	10.25	/	12.33	11.94	8.42
DL-SP [39]	7.39	10.22	12.13	10.37	9.65	8.42	7.81	8.23
CNN-S [30]	10.44	9.83	11.83	9.49	10.32	11.31	9.78	11.65
Multi-scare CNN +GS [32]	9.55	9.19	8.22	9.01	9.77	10.28	11.96	9.03
GA-UNet (proposed)	5.73	6.45	6.08	5.67	6.76	8.11	7.98	7.91

VI. Discussion

A. Summary of Our Results

The proposed model achieved a 2.42μm mean USPE on the AMD dataset which was the best as compared with the seven state-of-the-art methods [17, 19, 24, 28, 30, 32, 39]. The proposed model was significantly better than the other competing methods in 43 out of the 56 head-to-head comparisons and achieved insignificant margins in the other 12 cases. For the one lost case, the difference was insignificant. Results of CNV segmentation on the AMD dataset by the proposed model also outperformed the eight state-of-the-art methods [19, 24, 28, 43, 44, 45, 46, 47] with a DSC of 94.14%, at least 1.25% of improvement, a Precision of 93.34%, 1.95% higher than the second highest, which were all significant in the 40 head-to-head comparisons. These competing methods consisted of traditional graph-based methods and deep learning models.

The AMD dataset was collected from AMD patients and the retinal layers were distorted making retinal layer detection and CNV segmentation challenging for these traditional models. The proposed model has customized modules to handle the distortions, i.e., GAE improved the mean USPE by 0.78μm, while STCGO improved the mean USPE by 0.53μm and GDM together with the weighted MSE loss decreased it by 0.50μm. In total, the mean USPE was decreased from 3.95μm to 2.42μm. The improvements of DSC by GAE, STCGO and GDM were 3.94%, 0.49% and 0.76%, respectively. With these improvements, DSC was improved from 89.90% to 94.14% which was the best among all the competing methods.

The proposed model achieved a 2.06μm mean USPE on the public dataset in the 8-layer surface detection experiment and 2.17μm in the 9-layer surface detection experiment, both were the best as compared with the seven state-of-the-art methods. In the 8-layer surface detection experiment, the proposed model was significantly better than the other competing methods in 52 out of the 56 head-to-head comparisons and achieved insignificant margins in the other 4 cases. In the 9-layer surface detection. The proposed model was significantly better than the other competing methods in 47 cases out of the 63 comparisons and was insignificantly better in the remaining 16 cases.

B. OCT image with AMD

In an OCT image with AMD, CNV is typically located between choroid and RPE, and causes large and irregular morphological deformations on surfaces 7 and 8 [4]. Existing graph methods or deep learning models performed well in normal OCT images, but performances degraded significantly in OCT images with AMD. Detection of retinal layers in diseased OCT images often started with a pre-processing step to flatten the retinal layers. Next, initialization of retinal layers was performed. Classifiers such as random forest [17] or a simple neural network [19] could be trained on number of features, computed from OCT images, to generate initial retinal surface boundaries. Finally, the initialized boundaries were optimized to generate final results. Due to the appearance of CNV in AMD patients, it posts severe challenges to the existing graph-based methods and deep learning models.

C. Limitation of Traditional Graph-based Methods

Challenges posed to traditional graph-based methods: In these methods, 1) retinal layer surfaces to be detected were typically initialized with outputs of classifiers trained with

TABLE V

AVERAGE USPES BY DIFFERENT ALGORITHMS IN 8-LAYER SURFACE DETECTION ON THE PUBLIC DATASET [42] (MEAN(STD) μm),'*' INDICATES THAT THE COMPARED MODEL IS SIGNIFICANTLY WORSE THAN GA-UNET

Method/Surface	1	2	3	4	5	6	7	8	Overall
NNCGS [19]	2.89*(0.77)	3.63*(1.13)	3.67*(1.37)	4.01*(1.67)	4.72*(1.99)	4.33*(2.13)	2.99*(1.56)	3.79*(2.33)	3.75(1.44)
AURA [17]	2.37*(0.36)	3.09*(0.64)	3.43*(0.53)	3.25*(0.48)	2.96*(0.55)	2.69*(0.44)	2.07(0.81)	3.77*(0.94)	2.95(0.90)
TG-FCNN [29]	2.35*(0.41)	2.83*(0.67)	2.79*(0.50)	3.15*(0.56)	2.71*(0.63)	2.63(0.89)	2.09*(0.42)	3.43*(1.11)	2.75(0.65)
RelayNet [24]	3.19*(0.70)	3.69*(0.79)	3.47*(0.41)	3.48*(0.38)	3.37*(0.71)	2.99*(0.42)	2.82*(0.46)	3.85*(1.56)	3.36(0.68)
DL-SP [39]	2.50*(0.43)	3.21*(0.71)	3.07*(0.47)	3.23*(0.70)	2.91*(0.57)	2.61*(0.45)	2.17*(0.63)	3.67*(0.72)	2.92(0.59)
CNN-S [30]	3.20*(0.65)	3.58*(0.72)	3.62*(0.43)	3.55*(0.44)	3.32*(0.68)	3.00*(0.45)	2.79*(0.53)	3.77*(0.98)	3.35(0.61)
Multi-scare CNN+GS [32]	2.52*(0.39)	2.97*(0.65)	2.92*(0.49)	3.17(0.62)	3.01*(0.55)	2.72*(0.67)	2.13(0.51)	3.46*(0.89)	2.86(0.60)
GA-UNet (proposed)	1.57(0.33)	1.85(0.41)	2.23(0.42)	2.54(0.30)	1.91(0.39)	2.33(0.50)	1.68(0.63)	2.39(0.71)	2.06(0.46)

TABLE VI

AVERAGE USPES BY DIFFERENT ALGORITHMS IN 9-LAYER SURFACE DETECTION ON THE PUBLIC DATASET [42] (MEAN(STD) μm),'*' INDICATES THAT THE COMPARED MODEL IS SIGNIFICANTLY WORSE THAN GA-UNET

Method/Surface	1	2	3	4	5	6	7	8	9	Overall
NNCGS [19]	2.93*(0.73)	3.70*(1.22)	3.79*(1.55)	4.07*(1.37)	4.53*(2.10)	4.27*(1.73)	3.11*(1.54)	3.94(2.65)	4.23*(3.17)	3.84(1.78)
AURA [17]	2.35*(0.39)	3.13*(0.61)	3.44*(0.60)	3.21(0.52)	2.94*(0.51)	2.74*(0.43)	2.11(0.89)	3.81*(0.97)	2.99(2.32)	2.97(0.80)
TG-FCNN [29]	2.50*(0.40)	3.03(0.74)	2.85*(0.53)	3.14(0.44)	2.77*(0.70)	2.63(0.62)	2.09(0.59)	3.59(1.03)	3.11*(1.92)	2.86(0.77)
RelayNet [24]	3.12*(0.59)	3.81*(0.87)	3.41*(0.46)	3.69*(0.45)	3.31*(0.61)	3.02*(0.43)	2.75*(0.52)	4.24*(1.53)	3.09*(1.55)	3.38(0.78)
DL-SP [39]	2.49*(0.42)	3.23*(0.74)	3.00*(0.43)	3.19(0.72)	2.90*(0.66)	3.05*(0.49)	2.19(0.58)	3.73*(0.66)	3.20(1.12)	3.00(0.65)
CNN-S [30]	3.18*(0.67)	3.61*(0.73)	3.63*(0.39)	3.49*(0.44)	3.37*(0.69)	3.03*(0.41)	2.75*(0.50)	3.72(0.82)	3.13*(0.89)	3.32(0.62)
Multi-scare CNN +GS [32]	2.50*(0.37)	3.03*(0.64)	2.95(0.44)	3.22(0.63)	2.98*(0.55)	2.65*(0.63)	2.17(0.48)	3.42(0.87)	3.15*(1.01)	2.90(0.62)
GA-UNet (proposed)	1.58(0.33)	2.18(0.47)	2.40(0.57)	2.72(0.31)	2.22(0.47)	2.15(0.49)	1.53(0.40)	2.64(0.79)	2.11(0.99)	2.17(0.54)

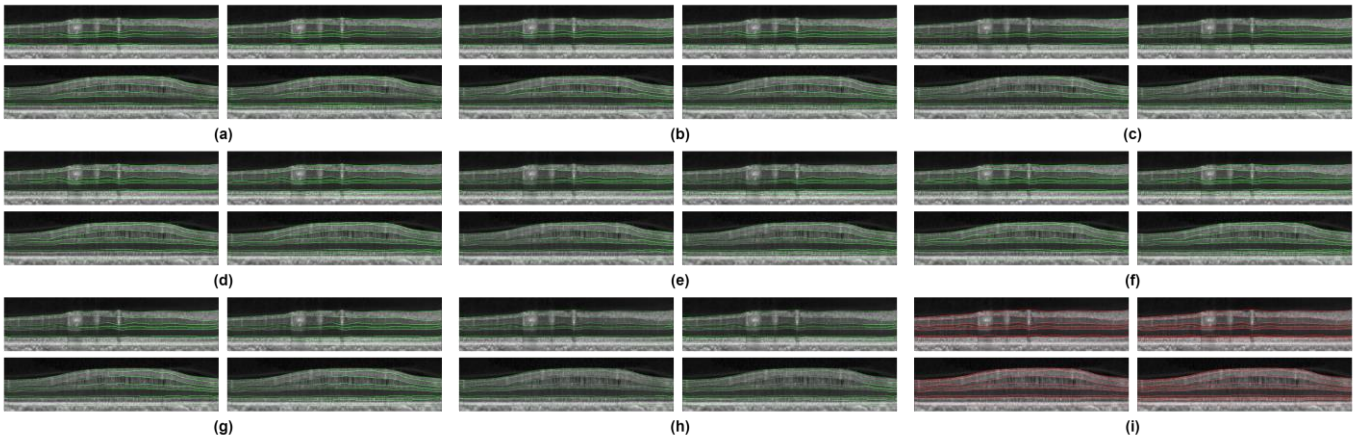

Fig. 10. Results of retinal layer surface detection on two OCT B-scan images with MS. (a) NNCGS, (b) AURA, (c) TG-FCNN, (d) RelayNet, (e) DL-SP, (f) CNN-S, (g) Multi-scare CNN + GS, (h) Proposed method, (i) Ground truth.

TABLE VII
AVERAGE DSCS IN CNV SEGMENTATION ON THE AMD DATASET BY DIFFERENT METHODS (MEAN(STD)%),'*' INDICATES THAT THE COMPARED MODEL IS SIGNIFICANTLY WORSE THAN GA-UNET

Method	DSC	Accuracy	Precision	Sensitivity	Specificity
NNCGS [19]	84.91*(7.65)	94.37*(1.10)	78.93*(5.85)	86.17*(7.43)	97.01*(0.88)
CE-Net [43]	92.33*(0.75)	99.15*(0.06)	90.11*(0.81)	93.03*(0.65)	99.28*(0.31)
Att-UNet [44]	91.25*(0.63)	99.13*(0.05)	90.00*(0.88)	92.51*(0.79)	99.11*(0.36)
SR-Net [28]	92.38*(0.65)	99.34*(0.08)	91.35*(0.40)	93.19*(0.61)	99.42*(0.25)
RelayNet [24]	92.89*(0.77)	99.35*(0.04)	91.39*(0.62)	93.55*(0.69)	99.44*(0.19)
IA-Net [45]	92.71*(0.83)	99.31*(0.07)	91.35*(0.59)	93.42*(0.57)	99.41*(0.21)
Trans-UNet [46]	91.45*(0.55)	99.17*(0.06)	90.01*(0.61)	92.81*(0.77)	99.21*(0.27)
Swin-UNet [47]	91.68*(0.75)	99.15*(0.05)	90.12*(0.67)	92.95*(0.68)	99.25*(0.39)
GA-UNet (Proposed)	94.14(0.51)	99.54(0.04)	93.34(0.45)	94.99 (0.56)	99.87(0.21)

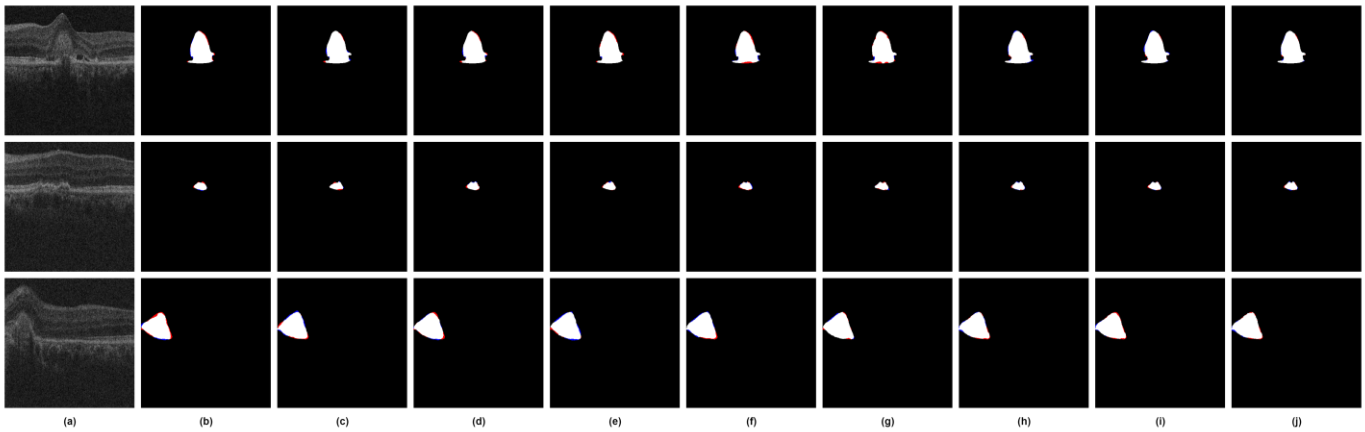

Fig. 11. Results of CNV segmentation by different models. (a) Original OCT B-scan images, (b) NNCGS, (c) CE-Net, (d) Attention U-Net, (e) SR-Net, (f) RelayNet, (g) IA-Net, (h Trans-UNet, (i) Swin-UNet, (j) Proposed method.

TABLE VIII
COMPUTATIONAL COMPLEXITY OF THE COMPETING METHODS

Method	Parameters (M)	FLOPs (G)	Time (s)
NNCGS [19]	/	/	19.13
SR-Net [28]	37.42	42.23	0.05
TG-FCNN [29]	52.91	242.10	0.10
GA-UNet (Proposed)	92.26	28.71	0.04

engineering features on limited data, deformations caused by CNV will directly impact graph construction and graph search since those engineered features were hard to fully represent the deformations. 2) CNV has blurred boundaries and low layer surface contrasts, making graph-based methods easily to be trapped in local minima and less generalizable [37]. For example, NNCGS, a graph-based algorithm, had the largest USPEs and achieved the lowest DSC on the AMD dataset as shown in Table III and Table VII.

Our solutions: In the proposed method, 1) we developed an end-to-end deep learning system rather than manual feature engineering for feature extraction. This automatic feature learning capability empowered convolutional neural network in many applications including computer vision. We showed in this paper that it could also improve graph-based methods. 2) We constructed a layer-wise graph structure in which each row

of the image was considered as a node in the graph. Retinal layer surfaces had very strong horizontal edge patterns in the image. The adjacent matrices representing correlations among nodes in the image were learned in an end-to-end manner to search similar nodes for node enhancement so that these layer surface edges were enhanced. With these innovative modules, our proposed model achieved significantly better results as shown in Table III and Table VII.

D. Limitation of Traditional CNN Models

Challenges posed to deep learning models: CNN-based models such as RelayNet [24], SR-Net [28] and DL-SP [39] resolved the manual feature engineering challenge and achieved better results as compared with these graph-based competing methods as shown in Table V and Table XI. However, the special topological structures in OCT images still limited the potential of CNN models. 1) Retinal layer surfaces in OCT images appear as very strong global edges. Though convolution kernels in traditional CNN can detect edges after training, there was no specialized design to leverage the large strong global edge structures. In addition, receptive fields in both horizontal and vertical directions increased equally in CNN models, which were not in favor for the retinal layer structures. 2) It is very challenging to build prior topological knowledge into CNN models. Several attempts have been made to resolve these challenges in the competing CNN based models [30]. These attempts did improve performances but not specifically designed to address the two main challenges.

Our solutions: In the proposed model, 1) we developed a layer-wise graph operator with global receptive fields in GAE to enhance layer surfaces. Graph construction by treating feature map rows as nodes allowed nodes to learn to find similar nodes for enhancement through these trainable adjacency matrices. In addition, the graph provided global receptive fields at different resolution levels, making the features sensitive to the large-scale retinal layer surfaces. It led to enhanced layer-like features in feature map as shown in Fig. 6. 2) We utilized the topological prior knowledge of retinal layer as hard constraints in STCGO to regularize outputs of GAE for layer surface detection, which eliminated redundancy and reduced computational complexity, and 3) we implemented the GDM module to decorrelate feature map and generate topological constraints for retinal layer surface detection, which maintained the continuity and smoothness of layer surfaces. Our proposed method outperformed all the competing CNN deep learning-based models.

E. Limitation of Our Work

There are several limitations in the proposed model. First, the accuracy of retinal layer surface detection was constrained by image contrast between layers in OCT images. If the boundaries are invisible due to the occurrence of CNV, the proposed model will fail. As shown in the first row of Fig. 9, the CNV changed the morphological structure due to the RPE layer bulge caused by blood vessel, making surface 7 to overlap with surfaces 6 and 5. Despite the constraints posed by the prior topological knowledge, the overlapped regions gave wrong information and led to incorrect surface detection results. The eighth row of Table V shows that surface detection errors of surfaces 4-7 were larger than these of other surfaces.

Second, we treated each row in the OCT image at input and in feature maps produced by subsequent convolutional layers as a node in the graph model. Though we found that these strip layers share similarity and layer surfaces were enhanced by the weighed summation of similar layers from other feature maps through the learned attention maps or adjacency matrices, we expect that using each pixel in the feature map as a node to construct a graph should lead to improved retinal layer surface detection. However, the number of parameters is too large for that case. Taking an input feature size (w, h) = (256, 256) as example, the size of Adjacency matrix will be (256×256) × (256×256), hundreds of times larger than that in our proposed model, and the graph will be much difficult to be managed.

Third, the proposed model had more parameters than these CNN based competing methods, due to the adoption of dynamic adjacency and weight matrices in convolutional layers. These matrices had the same sizes as their corresponding feature maps, making the number of parameters at high resolution layers much larger than these in CNN models. For example, in a convolutional layer with a kernel size of 3 × 3 and the input feature map with a size of 256 × 256 × 16, a CNN convolution layer contains 16 × 3 × 3 parameters, while the graph operator in the proposed model had 256 × 256 parameters. This limitation makes the proposed model as twice large as these traditional encoder-decoder multi-task structures [28, 29]. However, the inference time of the proposed model was the fastest among those multi-task methods including NNCGS [19], SR-Net [28] and TG-FCNN [29] as listed in Table VIII.

VII. Conclusion

In this paper, we proposed a graph based neural network for both retinal layer surface detection and CNV segmentation in OCT images with AMD. Our model used U-Net as backbone and consisted of two major novel components: GAE focusing on layer-wise feature embedding in the images and GDM learning to generate topological constraints for retinal layer surfaces to maintain correct topological order for better layer surface detection. In addition, we proposed a novel loss function that combined the learned topological constraints with MSE loss to improve the detection of retinal layer surface. Our proposed framework achieved the best results for CNV segmentation and retinal layer surface detection and established new state of the arts for our dataset.

References

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

[2] N. Kwak et al., "VEGF is major stimulator in model of choroidal neovascularization," *Investigative ophthalmology & visual science*, vol. 41, no. 10, pp. 3158–3164, 2000.

[3] A. Kubicka-Trzaska et al., "Circulating antiretinal antibodies predict the outcome of anti-VEGF therapy in patients with exudative age-related macular degeneration," *Acta Ophthalmologica*, vol. 90, no. 1, pp. 21–24, 2012.

[4] D. S. Friedman et al. "Prevalence of age-related macular degeneration in the United States." *Arch Ophthalmol*, vol. 122, no. 4, pp. 564-572, 2004.

[5] M. Wojtkowski, R. Leitgeb, A. Kowalczyk, T. Bajraszewski, A. F. Fercher et al., "In vivo human retinal imaging by fourier domain optical coherence tomography," *Journal of biomedical optics*, vol. 7, no. 3, pp. 457–463, 2002.

[6] Jia, Y., et al. "Quantitative Optical Coherence Tomography Angiography of Choroidal Neovascularization in Age-Related Macular Degeneration." *Ophthalmology,* vol. 121, no. 7, pp.1435-1444, 2014.

[7] X. Chen et al., "Quantification of external limiting membrane disruption caused by diabetic macular edema from SD-OCT," *Investigative ophthalmology & visual science,* vol. 53, no. 13, pp. 8042-8048, 2012.

[8] Q. Zhang et al., "Automated quantitation of choroidal neovascularization: A comparison study between spectral-domain and sweptsource OCT angiograms," *Investigative ophthalmology & visual science,* vol. 58, no. 3, pp. 1506–1513, 2017.

[9] H. Chen, H. Xia, Z. Qiu, W. Chen, and X. Chen, "Correlation of optical intensity on optical coherence tomography and visual outcome in central retinal artery occlusion," *Retina*, vol. 36, no. 10, pp. 1964–1970, 2016.

[10] D. Xiang et al., "Automatic Retinal Layer Segmentation of OCT Images With Central Serous Retinopathy," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 1, pp. 283-295, Jan. 2019

[11] F. Shi et al., "Automated 3-D retinal layer segmentation of macular optical coherence tomography images with serous pigment epithelial detachments," *IEEE transactions on medical imaging*, vol. 34, no. 2, pp. 441–452, Feb. 2015.

[12] M. K. Garvin, M. D. Abramoff, R. Kardon, S. R. Russell, X. Wu, and M. Sonka, "Intraretinal layer segmentation of macular optical coherence tomography images using optimal 3-D graph search," *IEEE transactions on medical imaging*, vol. 27, no. 10, pp. 1495–1505, Oct. 2008.

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

[14] Q. Song, J. Bai, M. K. Garvin, M. Sonka, J. M. Buatti, and X. Wu, "Optimal multiple surface segmentation with shape and context priors," *IEEE transactions on medical imaging*, vol. 32, no. 2, pp. 376–386, Feb. 2013.

[15] P. A. Dufour et al., "Graph-based multi-surface segmentation of OCT data using trained hard and soft constraints," *IEEE transactions on medical imaging*, vol. 32, no. 3, pp. 531–543, Mar. 2013.

[16] J. Novosel, et al. "Locally-adaptive loosely-coupled level sets for retinal layer and fluid segmentation in subjects with central serous retinopathy," *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2016, pp. 702-705.

[17] A. Lang et al., "Retinal layer segmentation of macular OCT images using boundary classification," *Biomedical Optics Express*, vol. 4, no. 7, pp. 1133–1152, 2013.

[18] Y. Liu, et al., "Multi-layer fast level set segmentation for macular OCT," *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI)*, IEEE, 2018, pp. 1445-1448,

[19] D. Xiang *et al*., "Automatic Segmentation of Retinal Layer in OCT Images With Choroidal Neovascularization," in *IEEE Transactions on Image Processing*, vol. 27, no. 12, pp. 5880-5891, Dec. 2018

[20] K. Yu, et al. "Shared-hole graph search with adaptive constraints for 3D optic nerve head optical coherence tomography image segmentation." *Biomedical Optics Express,* vol. 9, no. 3, pp. 962-983, 2018.

[21] S. Lee, et al., "Atlas-based shape analysis and classification of retinal optical coherence tomography images using the functional shape (fshape) framework," *Medical Image Analysis,* vol. 35, pp. 570-581, 2017.

[22] E. Shelhamer, J. Long and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640-651, April 2017.

[23] Ronneberger, O., P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation." *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, Cham, 2015, pp 234-241.

[24] A. G. Roy, et al., "Relaynet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks," *Biomedical Optics Express,* vol. 8, no. 8, pp. 3627–3642, 2017.

[25] S. Apostolopoulos et al., "Pathological OCT retinal layer segmentation using branch residual U-shape networks," *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, Cham, 2017, pp. 294-301.

[26] L. Fang, D. Cunefare, C. Wang, R. H. Guymer, S. Li, and S. Farsiu, "Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search," *Biomedical Optics Express,* vol. 8, no. 5, pp. 2732-2744, 2017.

[27] K. Gopinath, S. B. Rangrej and J. Sivaswamy, "A Deep Learning Framework for Segmentation of Retinal Layers from OCT Images," *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*, IEEE, 2017, pp. 888-893.

[28] He, Y., et al. "Deep learning based topology guaranteed surface and MME segmentation of multiple sclerosis subjects from retinal OCT." *Biomedical Optics Express,* vol. 10, no. 10, pp. 5042-5058, 2019.

[29] Y. He et al. "Fully convolutional boundary regression for retina OCT segmentation." *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, Cham, 2019, pp. 120-128.

[30] Abhay Shah, Leixin Zhou, Michael D. Abrámoff, and Xiaodong Wu, "Multiple surface segmentation using convolution neural nets: application to retinal layer segmentation in OCT images," *Biomedical Optics Express,* vol. 9, no. 9, pp. 4509-4526, 2018.

[31] S. Masood, et al., "Automatic choroid layer segmentation from optical coherence tomography images using deep learning," *Scientific Reports*, vol. 9, no. 1, pp. 1-18, 2019.

[32] X. Sui, et al. "Choroid segmentation from optical coherence tomography with graph-edge weights learned from deep convolutional neural networks," *Neurocomputing*, vol. 237, no. C, pp. 332-341, 2017.

[33] Liu, Y., et al. "Layer boundary evolution method for macular OCT layer segmentation." *Biomedical Optics Express,* vol. 10, no. 3, pp. 1064-1080, 2019.

[34] X. Xu, K. Lee, L. Zhang, M. Sonka, and M. D. Abràmoff, "Stratified sampling Voxel classification for segmentation of intraretinal and subretinal fluid in longitudinal clinical OCT data," *IEEE transactions on medical imaging*, vol. 34, no. 7, pp. 1616-1623, Jul. 2015.

[35] J. Hu, L. Shen, S. Albanie, G. Sun and E. Wu, "Squeeze-and-Excitation Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011-2023, 2020.

[36] Y, Liu, et al. "Layer boundary evolution method for macular OCT layer segmentation." *Biomedical Optics Express,* vol. 10, no. 3, pp. 1064-1080, 2019.

[37] M Wang, et al., "Semi-Supervised Capsule cGAN for Speckle Noise Reduction in Retinal OCT Images." *IEEE transactions on medical imaging*, vol. 40. No. 4, pp. 1168-1183, 2021.

[38] Kip F, T. N., and M. Welling. "Semi-Supervised Classification with Graph Convolutional Networks." *arXiv preprint arXiv:1609.02907*, 2016.

[39] Mishra, Z., et al. "Automated Retinal Layer Segmentation Using Graph-based Algorithm Incorporating Deep-learning-derived Information." *Scientific Reports*, vol. 10, no. 1, pp. 9541-9550, 2020.

[40] Raja, H., et al. "Extraction of Retinal Layers Through Convolution Neural Network (CNN) in an OCT Image for Glaucoma Diagnosis." *Journal of Digital Imaging*, vol. 33, pp. 1428–1442, 2020.

[41] B. Wang, W. Wei, S. Qiu, S. Wang, D. Li, and H. He, "Boundary aware U-Net for retinal layers segmentation in optical coherence tomography images," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 8, pp. 3029 -3040, Aug. 2021.

[42] He Y, et al., "Retinal layer parcellation of optical coherence tomography images: data resource for multiple sclerosis and healthy controls." *Data in Brief,* vol. 22, pp. 601–604, 2018.

[43] Z. Gu *et al*., "CE-Net: Context Encoder Network for 2D Medical Image Segmentation," *IEEE Transactions on Medical Imaging*, vol. 38, no. 10, pp. 2281-2292, Oct. 2019.

[44] Oktay, O., et al. "Attention U-Net: Learning Where to Look for the Pancreas." *arXiv preprint arXiv:1804.03999*, 2018.

[45] Xi, X., Meng, X., Qin, Z., Nie, X., Yin, Y., & Chen, X. "IA-net: informative attention convolutional neural network for choroidal neovascularization segmentation in OCT images." *Biomedical Optics Express*, vol. 11, no. 11, pp. 6122-6136, 2020.

[46] Chen J, et al. "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation." *arXiv preprint arXiv: 2102.04306*, 2021.

[47] Cao H, et al. "Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation." *arXiv preprint arXiv: 2105.05537*, 2021.

[48] Haoran Zhao and Chengwei Zhang. "GAU-Nets: Graph Attention U-Nets for Image Classification." *The 5th International Workshop on Advanced Algorithms and Control Engineering*, Journal of Physics: Conference Series, Vol. 1861, February 2021.