SCSDM: Semi-supervised Cross-consistency Learning via Score-based Diffusion Model for lesion Segmentation in OCT images

Abstract

The lesion segmentation in Optical Coherence Tomography (OCT) images contributes to observing the disease status of impairment retina. Due to the limitation of the labeled data, the semi-supervised learning mechanism become a popular training strategy. However, most of the existing semi-supervised learning models are based on the collected labeled and unlabeled data, which show inferior ability of acquiring the relationship of data. To this end, we propose a Semi-supervised Cross-consistency Learning via Score-based Diffusion Model, called SCSDM, to effectively exploit the unlabeled data generated by score-based generative model for semi-supervised OCT lesion segmentation. Specifically, in the first stage, we utilize score-based generative model to synthesize unlabeled OCT images by excavating the relationship between the labeled and synthetic data. In the second stage, we filter out the low-quality images through some common image quality metrics. For the lesion segmentation framework, we incorporate boundary-transformer encoder jointly with CNN architecture to extract both local and global context for encoder. To acquire improved feature representation, the attention feature fusion module is devised to fuse local and global features with inconsistent semantic and scales. Further, we employ a mutual consistency constraint to align the output of main decoder with the masks generated by the varying levels of auxiliary decoders, enhancing the regularization of the model during model training. Experiments on one private dataset demonstrate that our proposed method is promising for the lesion segmentation using OCT images.

Keywords: OCT lesion segmentation, Score-based generative model, semi-supervised learning, boundary-transformer, consistency learning

1 Introduction

OPTICAL coherence tomography (OCT) a medical imaging modality for ophthalmologists to assess and diagnose diseases due to the properties of accessible, contactless and high resolution [1] [9] [19]. OCT has been employed as a practical tool for accurate segmentation and quantitative analysis of lesion areas [13] [16]. For instance, as a globally prevalent retinal degenerative disease, pathologic myopia is a main cause of blindness and irreversible [25], in which the pathologic myopia choroidal neovascularization (PMCNV) is the common complication. In addition, Macular Edema (ME) is the infiltration of fluid in the center of retina due to disruptions in blood retinal barrier [8] [15], which is related to retinal diseases such as age-related

macular degeneration (AMD) and diabetic macular edema (DME). The accurate segmentation of lesions in the retinal diseases above using OCT contributes to observing the progression of these diseases and assist the ophthalmologists to formulate the treatment strategies. For the lesion segmentation task in OCT, manual segmentation is time-consuming, laborious, and subjective. Hence, an automatic lesion segmentation method for OCT images is needed, which can help clinic doctors diagnose and assess the progression of retinal diseases, thereby reducing the workload of ophthalmologists.

In recent years, deep learning method has become a popular tool to execute various tasks (including segmentation , classification, Screening, etc.) using OCT images. However, the prerequisite for achieving good performance of deep learning models is sufficient labeled data. This is very difficult to obtain for clinical data, as the complex structure of clinical data makes annotation difficult, resulting in extremely high annotation costs. This issue is more prominent for OCT images, which require ophthalmologists examine each B-scans to make pixel-level annotation. To tackle this situation, some researchers investigated semi-supervised deep learning method to improve the performance [2] [24] . For example, Wang et al. proposed a semi-supervised segmentation method, called Semi-SGO, for joint segmentation of macular hole and cystoid macular edema in retinal OCT images by using unlabeled OCT images [20]. In order to make full use of the obtained information in the abundant unlabeled OCT images, Fazekas et al. presented a novel semi-supervised model for the segmentation task of the retinal layer [5]. These methods directly utilize the collected unlabeled data, however, there are not many available unlabeled data in clinical practice due to privacy or ethical reasons.

To this end, a novel semi-supervised mode attracts lots of attention, in which the produced data by generative models is used as the unlabeled data to build semi-supervised deep learning method for boosting performance. Wang et al. proposed a novel capsule conditional generative adversarial network (Caps-cGAN) to set up semi-supervised learning system for denoising the speckle noise in retinal OCT images [21]. Lyu et al. proposed a new semi-supervised model for Covid-19 pneumonia infection segmentation, in which the generative model produce synthetic images to match the pseudo-labels of unlabeled images, thereby enhancing the training data to improve performance [12]. Due to the powerful generative ability of the diffusion model [7] [4], it is also used as a generative model to synthesize unlabeled or pseudo labeled data to construct semi-supervised learning models. For example, Lim et al. presented Adaptive aggregation with Class-Attentive Diffusion (AdaCAD) to construct the semi-supervised learning model for the classification of different graph domains [10]. Tang et al. presented a Multi-level Global Context Cross-consistency (MGCC) framework, in which the synthetic images generated by a Latent Diffusion Model (LDM) as unlabeled images to build semi-supervised learning model for medical image segmentation [17]. However, the synthetic data generated by these generative models is confused, which cannot align the training data distribution, resulting in low efficiency.

To solve the issues above, we propose a novel semi-supervised cross-consistency learning via score-based diffusion model, called SCSDM, which consists of three-stage tasks (i.e. data generation, data selection, data segmentation.). In the first stage, we utilize score-based generative model to synthesize unlabeled OCT images by estimating the gradient of log-likelihood and applying the score-matching function to align the training data distribution. In the second stage, a quality-based filter is trained to remove the low-quality synthetic images by the generation stage. In the third stage, we leverage the labeled and synthetic high-quality images to build the semi-supervised segmentation network, in which the CNN and boundary-transformer are incorporated to extract both local and global contextual features and a mutual consistency constraint is employed to align

the output between main and auxiliary decoders. Extensive experiments on two public datasets demonstrate that our method is promising for the fluid segmentation in OCT images. Overall, our contributions can be summarized as below.

- We propose a novel semi-supervised fluid segmentation method in OCT images. A score-based generation model is utilized to synthetize unlabeled OCT images and a quality-based filter is trained to select the high-quality generation images.
- By integrating the labeled and synthetic unlabeled OCT images, we build the semi-supervised segmentation network. For encoder, we combine CNN and boundary-transformer to extract both local and global contextual features. And for decoder, a mutual consistency constraint is employed to align the output between main and auxiliary decoders.
- Extensive experiments on Myopia and UMN datasets demonstrate that our method can achieve superior segmentation performance for fluid in OCT images.

2 Related works

2.1 Lesion segmentation in OCT images

For retinal lesion segmentation using OCT images, several automated methods have been developed [16] [18] [6]. He et al. presented an intra-slice contrastive learning strategy to construct an inter-slice contrastive learning architecture, which can represent the similarity of adjacent OCT slices from one OCT volume and achieve promising fluid segmentation performance for OCT images [6]. Rashno et al. proposed a novel neutrosophic transformation and a graph-based shortest path method to segment fluid-associated and cyst regions in OCT images of subjects with diabetic macular edema. Based on the intensity of OCT images and retinal layer segmentations provided by a graph-cut algorithm, Lu et al. present a full convolution network to accomplish multiclass fluid segmentation and detection in the retinal OCT images [11]. Xing et al. integrated attention gate and a novel curvature regularization term in loss function to devise a FCN framework for the simultaneous segmentation of three types of pathological fluid lesions in OCT [22]. However, most of these methods are based on fully-supervised models designed with annotated data, which have a high degree of dependence on data, while clinical annotated data is difficult to obtain. To this end, we adopt semi-supervised learning mechanism to solve the limitation of scarce labeled data.

2.2 Semi-supervised learning for medical images segmentation

The semi-supervised learning strategy plays a significant role in the field of medical image analysis, especially for medical image segmentation [26]. For instance, Cheplygina et al. made an overview of semi-supervised, multiple instance, and transfer learning in medical imaging, both in diagnosis or segmentation tasks [3]. Luo et al. proposed a consistency regularization approach for semi-supervised medical image segmentation, where a set of segmentation predictions at different scales is obtained by pyramid-prediction network and integrated by multi-scale uncertainty rectification to boost the pyramid consistency regularization [12]. Zhou et al. presented a collaborative learning method to jointly improve the semi-supervised learning

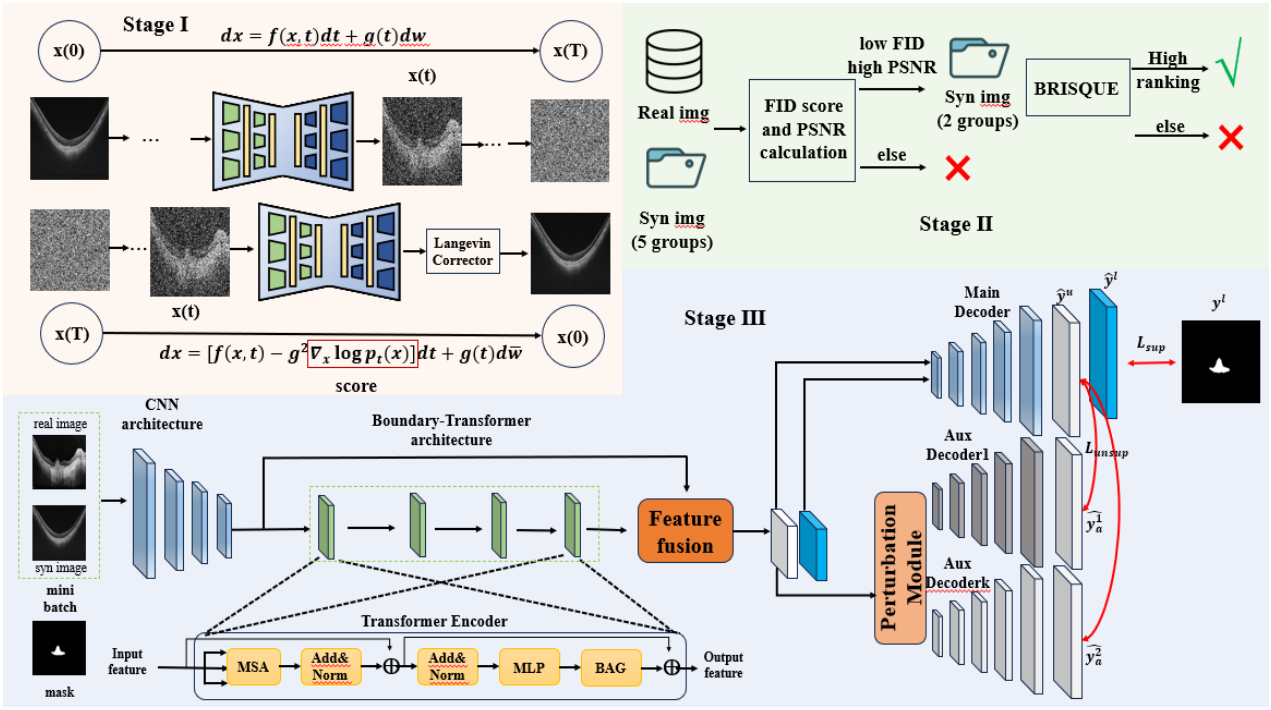

Figure 1. The flowchart of the proposed SCSDM. In the first stage, the score-based generative model is utilized to synthesize unlabeled OCT images by estimating the gradient of log-likelihood and applying the score-matching function to align the training data distribution. In the second stage, a quality-based filter is trained to remove the low-quality synthetic images by the generation stage. In the third stage, both the labeled and synthetic high-quality images are integrated to build the semi-supervised segmentation network, in which the CNN and boundary-transformer are incorporated to extract both local and global contextual features and a mutual consistency constraint is employed to align the output between main decoder and auxiliary decoders.

for diabetic retinopathy. Yang et al. proposed a semi-supervised retinal layer segmentation model, in which the lesion features and layer structure information are obtained by the labeled and unlabeled OCT images, respectively [23]. Most of the above semi-supervised methods are based on existing labeled and unlabeled data, while we cannot hold a large amount of unlabeled data in most cases. Hence, we need to synthetize unlabeled OCT images to build the semi-supervised learning model for improving the segmentation performance.

3 METHODOLOGY

3.1 Overview

This part integrates a score-based diffusion model with the semi-supervised framework to accomplish the lesion segmentation for OCT images, as shown in Fig. 1. The whole framework consists of three training stages: (1) Generate unlabeled OCT images. The score-based generative model is utilized to synthesize an extensive collection of OCT images. (2) Filter out high-quality generated OCT images. The filter module is adopted to screen out high-quality synthesis images for next step segmentation training. (3) Construct a semi-supervised segmentation network. We build global consistency network jointly with a boundary-aware transformer encoder to improve the representation capability for semi-supervised OCT lesion segmentation. The devised framework intends to explore both the scarce number of labeled images with pixel-wise annotations

and unlabeled synthetic images for accurate OCT lesion segmentation.

3.2 Stage I: Score-based generative model for unlabeled image synthesis

In the field of medical image segmentation, especially for OCT lesion segmentation task, it is not easily accessible to the large number of pixel-level annotation data, leading to the restricted segmentation performance. To solve this issue, many researchers developed generative models to synthetize data with labeled images to enrich the training dataset. As a type of generative models, the score-based generative model aims to generate complex and high-quality data samples by learning the underlying data distribution and estimating the gradient of log-likelihood, thereby utilizing the score-matching function to align the training data distribution. To alleviate scarcity problem of training data and acquire realistic results for OCT lesion segmentation, we attempt to synthesize target image examples using score-based generative model.

The principle of a score-based generative is that the score of the distribution $\bigtriangledown_x logp_t(x)$ can be approximated by training a score-based model on the sample data. There are two main processes: stochastic process and reverse process. For the stochastic process, the target is to transform the data into a prior distribution, which is a solution of stochastic differential equations (SDE) and can be formulated as

$$dx = f(x,t)dt + g(t)dw \tag{1}$$

where $f(,t)$ is called the drift coefficient of SDE and $g(t)$ represents the diffusion coefficient and w depicts the standard Brownian motion. In the diffusion process, we define $x(0) \sim p_0$ and $x(T) \sim p_T$. In this context, p_0 indcates the real data distribution and P_T represents the prior distribution that has a tractable form and is easy to sample.

The score-based generative model aims to reverse this process (from $t = T$ to $t = 0$) by generating new target samples from a prior distribution and solving the reversing SDE, which can be defined as:

$$dx = [f(x,t) - g^2(t) \bigtriangledown_x logp_t x]dt + g(t)d\overline{\omega} \tag{2}$$

where $\overline{\omega}$ is the Brownian motion in the reversing time with the gradient $\bigtriangledown_x logp_t(x)$. The $\bigtriangledown_x logp_t(x)$ is called score function, which can be approximately by a time-dependent score-based model $s_\theta(x,t)$.

$$s_\theta(x_t;t) \approx \bigtriangledown_x logp_t(x) \tag{3}$$

We use denoising score-matching to train our generative model. Instead of acquiring the score function of origin data, we seek to learn the score function of the modified data that have been perturbed by a specified perturbation function.

3.3 Stage II: Filter module for high-quality image selection

The low-quality medical images are synthesized inevitably by generative model, which can affect the segmentation performance when used as the training data in the segmentation task. To tackle this issue, a quality-based filter is trained to remove the low-quality synthetic OCT images for further segmentation training.

This section must be filled. If no related source codes are available, please indicate clearly. If there are any codes referenced in the process, please list them all and describe your usage in detail, highlighting your

own work, creative additions, noticeable improvements and/or new features. The differences and advantages must be dominant enough to demonstrate your contribution.

The image quality assessment plays a vital role in the filter process to ensure the fidelity and reliability of generated images. The combination of a lower FID value and a higher Inception Score(IS) signifies a higher level of quality for the group of images. Thus, for the synthetic OCT images, we utilize FID and IS to assess the quality between generated images and real images. We first divide the generation samples into different groups (5 groups in our work) to calculate the FID and IS between the sampled group and real group, which can be summarized by the formula:

$$FID = \|\mu_1 - \mu_2\|^2 + \text{Tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1\Sigma_2)^{1/2}) \tag{4}$$

$$IS = \exp\left(\mathbb{E}_x\left[D_{\text{KL}}(P(y|x) \parallel P(y))\right]\right) \tag{5}$$

where μ_1 and μ_2 are the mean vectors of real and generated data in the feature space in the Inception network, respectively, and Σ_1 and Σ_2 are their covariance matrices, $Tr(\cdot)$ denotes the trace operation. And x denotes the generated sample, $P(y|x)$ is the class distribution given sample x, $P(y)$ is the class distribution for all samples, and D_{KL} denotes the KL divergence.

Through the aforementioned steps, we can select the groups with low FID scores and high IS. Then, we handle the three greatest sample groups through the BRISQUE network [14], which can map the image to the quality space, producing a quantification of image quality. By utilizing this technique, the network can output the quality score of the input image to choose the high-ranking images. Thus, we can filter out the low-quality OCT images by the above steps for the subsequent semi-supervised segmentation training.

3.4 Stage III: The semi-supervised segmentation network

By stage I, we can obtain extensive unlabeled OCT images using score-based generative model. And the filter module is employed to select the high-quality synthetic OCT images for training segmentation network in stage II. With the combination of the generation and real OCT images, we can train a robust lesion segmentation network by semi-supervised strategy.

Although these generated samples do not have specific label information, they can be used as additional training samples to help the model learn image features and context information, which is beneficial for enhancing the OCT lesion segmentation performance. Since the OCT images show multiple levels of anatomical structures, boundary information can be used to localize and identify different ocular structures. In order to make full use of labeled samples and unlabeled samples generated by the score-based generative model, we build semi-supervised learning architecture based on CNN and boundary transformer network, which can extract rich local and global contextual features with boundary information. Subsequently, the extracted features above are fused by a feature fusion module.

1) Boundary-aware global information

Given an input image $I \in R^{H\times W\times 3}$ with height H and width W, the vanilla encoder-decoder architecture exploits standard CNN architecture (such as ResNet-50) and its variants to conduct effective medical image segmentation. However, the receptive field of CNN is limited by the intrinsic kernel size and pooling layers,

resulting in the lack of global contextual information. To this end, we integrate the features extracted by CNN into transformation network to enable the whole segmentation framework to express fine-grained features and enrich these features with comprehensive global contextual information.

In addition, the boundary information is an important component in medical images, whose well representation can boost the lesion segmentation performance. For OCT images, the boundary information can provide localize and distinguish the key areas within the OCT scans, which can improve the segmentation performance. Hence, we incorporate the boundary-aware mechanism into the transformer architecture. By this way, the global information and the boundary information can be obtained.

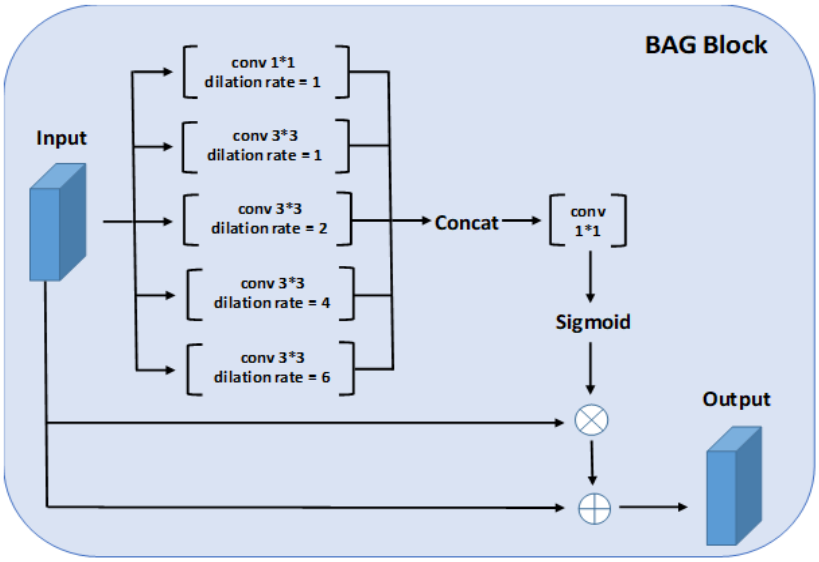

Figure 2. The illustration of BAG block.

2) Fusion of Local and Global context

Feature fusion is the procedure of merging distinct features of different sources to construct a more informative representation. Yet simply adding the features together or concatenating them cannot take full advantage of these fused features. In order to efficiently fuse the extracted global and local features, we deploy a multi-feature fusion method to integrate the local features from CNN module and the global features from the Transformer module to enhance the feature representation for lesion segmentation in OCT images.

Given two intermediate feature maps X and Y, where X denotes the local features from the CNN encoder and Y denotes the global features from the transformer encoder jointly with BAG block. Our feature fusion module mainly consists of two components: the local attention module and the global attention module. We first pass the Global Average Pooling(GAP) Module with the CNN Feature maps X and Transformer Feature maps Y, then we concat the features G_{CNN} and the origin Transformer features, and features G_{Trans} and the origin CNN features. Then the two mixture features will pass through a standard swin-transformer block to extract relevant spatial and temporal information. Those two features will be reshaped to 3d shape and then concated for the next decoder stage.

3) Cross-consistency regularization learning

To sufficiently utilize the synthetic OCT images D_u from stage I, in this part, we propose a global cross-consistency strategy to implement the consistency between the main decoder and those of auxiliary decoders, which can fully explore extra training information from the unlabeled synthetic set D_u.

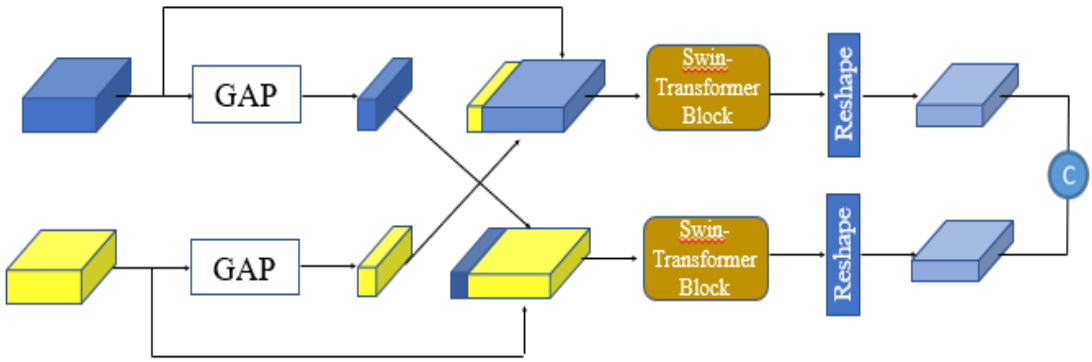

Figure 3. Fusion between CNN and Transformer features

We adopt a standard encoder-decoder architecture to build the segmentation network. Specifically, both the annotated images and synthetic images are imported into a shared encoder to extract the high-level features. To enhance the expression of the decoder, we devise a main decoder D_{main} to output the final segmentation prediction and a variety of auxiliary decoders $D^1_{aux}, \ldots, D^k_{aux}$ is incorporated to enforce the consistency between the main decoder and K auxiliary decoders. In each auxiliary decoder, we apply different perturbations to the encoder outputs.

For the supervised process, we define the OCT images x^l_i with the corresponding ground-truth y_i. And a binary cross-entropy loss is adopted to restrain the supervised network.

$$L_{sup} = \frac{1}{|D_L|} \sum y^i log\hat{y}^i + (1 - y^i) log(1 - \hat{y}^i) \tag{6}$$

where D_L represents the labeled dataset, y is the ground-truth of pixel x, and $\hat{y}_i$ is the prediction probability.

For the unsupervised process, we set up different types of perturbations to the intermediate features z_{ori} from the shared encoder to construct K different auxiliary decoders following []. In our work, we apply different compositions of these perturbations methods to select an optimal framework for global cross-consistency segmentation learning. We summarize six different perturbation methods in decoder architecture as indicated below, and the perturbation features from six perturbation methods are uniformly donated as $\tilde{z}$.

Feature-Noise Decoder: This module incorporates a feature-based generation module N, where a uniform distribution is sampled to create the noise features. Then the input features z_{ori} are multiplied with the noise features to generate the output features $\tilde{z} = (z_{ori} \cdot N) + z_{ori}$.

Feature Dropout Decoder: It takes the feature maps z_{ori} as input and the spatial attention operation is executed to generate the dropout mask M_{drop}. Subsequently, the features are modified by the mask to produce the output features $\tilde{z} = (z_{ori} \cdot M_{drop})$.

Dropout Decoder: This decoder apply the dropout regularization to the input tensor for random perturbation to gain the output feature.

Virtual Adversarial Decoder: This decoder defines a function to compute virtual adversarial perturbation and then perturbs the input features iteratively, which can enhance feature extraction based on the context and object presence in the output of main decoder.

GuidedMasking Decoder: Two version of context-based masking are employed to perturb both the de-

tected object and context to strengthen the capability of lesion understanding. Thus, then the object mask and context mask are generated to be applied to the input features.

CutoutMasking Decoder: This decoder executes on the output of the main decoder to find the detected object. A random crop is applied on each detected object from the features.

We use MSE distance to measure the probability distribution between the output of the main decoder and the auxiliary decoders. The total unsupervised loss can be summarized as:

$$L_{unsup} = \frac{1}{|D_U|}\frac{1}{K}\sum MSE(d_{main}(\tilde{z}), d^k_{aux}(\tilde{z})) \tag{7}$$

To sum up, we utilize the cross-entropy loss to restrain the supervised process. For the unsupervised process, a global cross-consistency approach is presented to maintain the consistency between the prediction of the main decoder and that of auxiliary decoders. The total loss L_{total} can be defined as:

$$L_{total} = L_{sup} + \omega L_{unsup} \tag{8}$$

where ω is an unsupervised loss weight.

4 Experiments

4.1 Datasets

Myopia dataset: The myopia dataset contains 2381 OCT images with the corresponding pixel-wise segmentation labels. In our experiment, we randomly split the myopia dataset into training set, validation set by the ratio of 8:2.

4.2 Implementation Details

All the methods are implemented using the PyTorch library and trained on NVIDIA 4090ti GPU to accelerate the process of training and testing. In stage I, we use a standard score-based generative model to synthesize OCT images for the next step. In the second stage, the synthesis images will be randomly split into different groups, then the image quality metrics including PSNR, PID are introduced to control the quality of generated images. Subsequently, we import the top three image sets with the highest quality into the BRISQUE to filter out the high-quality synthetic samples. For the semi-supervised learning framework, the semi-percentage is set to 0.5, which means we use half of the data with labels and half without labels to train the semi-supervised segmentation network. The Adam is selected to optimize the semi-supervised segmentation network with the scheduler of cosineAnnealingLR. The initial learning rate, the batch size, and the max epoch are set as 0.001, 8, and 100, respectively.

4.3 Evaluation Metrics

In our experiment, the common evaluation metrics include Intersection over Union (IoU), and Dice similarity coefficient (Dice) are emcployed to evaluate the segmentation performance, which can be expressed as follows:

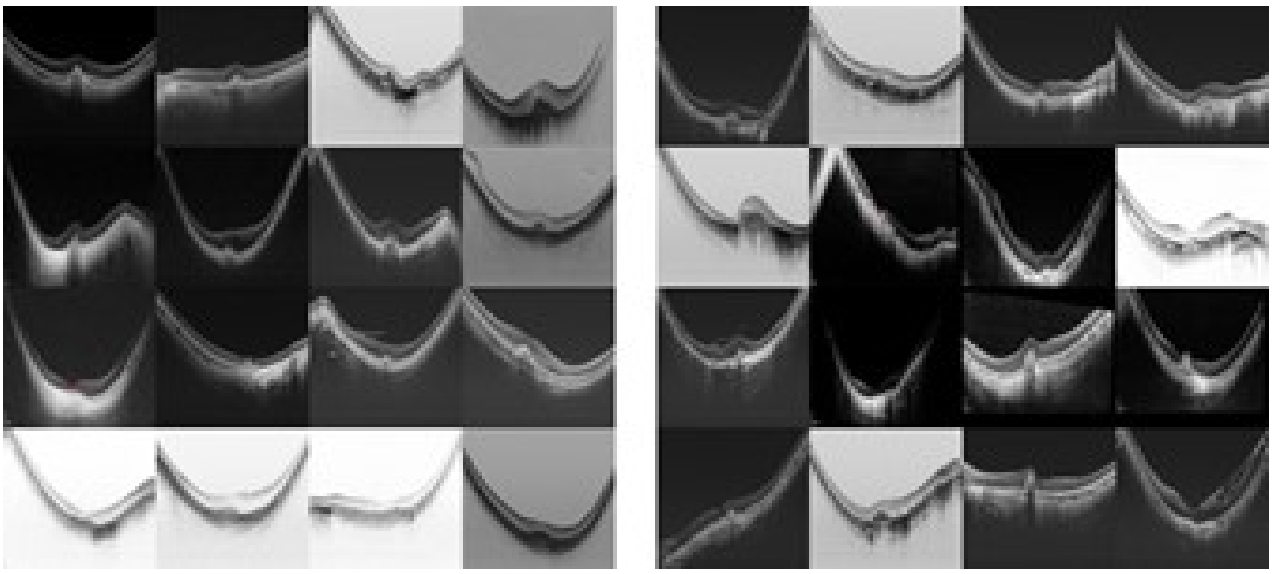

Figure 4. Visualization of the real images and the synthetic images by score-based generative model on Myopia dataset. The left side represents the synthesis samples, and the right side stands for the real samples, respectively.

$$IoU = \frac{TP}{TP + FP + FN} \tag{9}$$

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{10}$$

where TP, TN, FN, FP indicate the number of true positives, true negatives, false negatives, and false positives, respectively.

4.4 Visualization of synthetic OCT images

In our work, the score-based generative model is employed to synthesize unlabeled OCT data by solving the reverse-time SDE. The synthetic results of the score-based generative model are presented in Fig. 4 for Myopia datasets. We can observe that the generated OCT images by the score-based generative model are highly similar with the original OCT images in terms of shape and diversity. Furthermore, the lesion information in the synthetic OCT images can also be expressed well, which is beneficial for enriching the diversity of training data and boosting the segmentation performance of semi-supervised network.

4.5 Results of segmentation results compared to other methods

In this part, we utilize synthetic samples produced by the score-based generative model as an unlabeled dataset for the design of semi-supervised segmentation model. Due to the ratio of semi-supervised being set to 0.5, we employ the same number of the synthetic images as the real OCT images in Myopia dataset for the network training. Table I reports the results compared to other methods.

For fair comparison, we retrieve the articles with semi-supervised medical segmentation, transformer-based medical segmentation, and diffusion-based medical segmentation using same training settings in recent years.

Method	IoU	Dice
U-Net	76.02	85.62
U-Net++	76.37	86.09
MGCC	77.69	88.13
MC-Net	77.09	87.38
Med-Tran	78.58	87.94
Swin-Unet	79.08	89.15
MedsegDiff	79.20	89.23
Ours	**79.39**	**89.73**

Table 1. The quantified results of different methods on the Myopia dataset.

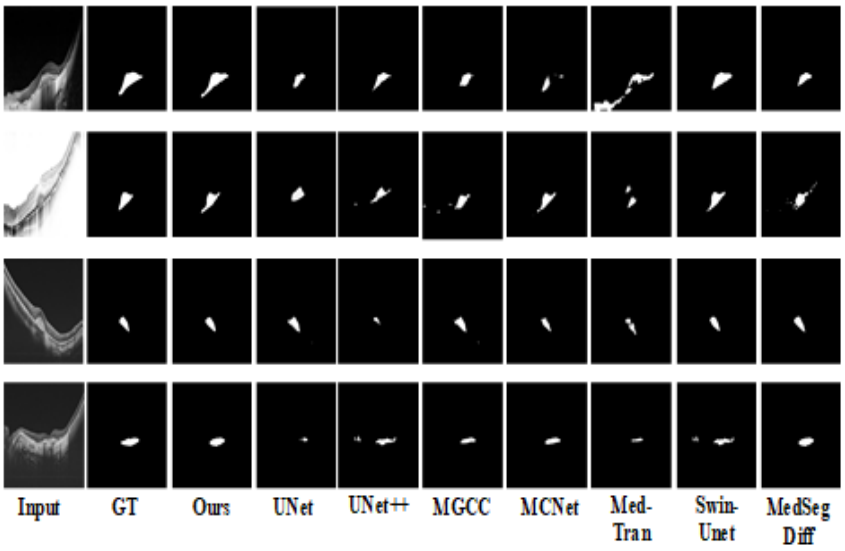

Figure 5. Qualitative results of the comparison methods on the Myopia dataset. Each column represents the segmentation result of a method.

From Table 1, we can observe that our proposed method achieves the best segmentation performance with the IoU of 79.39 and Dice score of 89.73, respectively. These results demonstrate that our semi-supervised segmentation method is effective for the prediction of lesion areas, which is benefited for the combination of the synthetic unlabeled samples, consistency learning, the fusion of local features and global features, and the boundary-aware information. By comparing the results of MGCC and Ours, we can see that our proposed method achieves a higher 1.7% than MGCC in terms of the metrics of IoU and Dice score for the both Myopia dataset. This results demonstrate that although the generative model with diffusion and semi-supervised learning mechanism are employed to conduct the lesion segmentation in OCT images, the score-based generative model can learn the underlying data distribution by estimating the gradient of log-likelihood and utilize the score-matching function to align the training data distribution, which improve the quality of the synthetic data to boost the segmentation performance.

4.6 Abalation study

In order to verify the effectiveness of individual component of our proposed method, we conduct multiple ablation studies on the Myopia dataset. In the experiments, we perform the training samples in the myopia dataset as a labeled set and the 1886 synthetic samples as an unlabeled set to train the semi-supervised segmentation network and the validation dataset is employed to evaluated the segmentation performance. And Figure 2 demonstrates that the four proposed components can effectively improve the IoU of fluid segmentation.

Table 2. The results of ablation studies on different components on the Myopia dataset.

Case	Quality-Filter	Trans.	Boundary Info	CNN-Trans Fusion	Iou(%)
case(1)					77.69%
case(2)	✓				78.27%
case(3)	✓	✓			78.85%
case(4)	✓	✓	✓		79.02%
case(5)	✓	✓	✓	✓	79.39%

5 Conclusion and future work

In this paper, we propose a novel semi-supervised segmentation framework based on score-based diffusion generative model for the lesion segmentation using OCT images. Specifically, we employ score-based generative model to synthesize unlabeled OCT mages, which are used to build the semi-supervised segmentation network using the labeled and unlabeled data. Considering that the score-based generative model may produce low-quality OCT images that may have a negative effect on improving segmentation, we devise a filter module by applying image quality metrics to filter out the inferior synthetic images. For the design of the semi-supervised segmentation network, we combine the CNN and boundary-aware transformer structures to extract rich local and global features. In addition, in order to fully explore the feature information from synthetic OCT images, six different auxiliary decoders are designed to preserve prediction consistency with that of the main decoder. Extension experimental results on two OCT datasets demonstrate our model achieves the superior segmentation performance over the comparison models, which can provide accurate lesion information for ophthalmologists to assess the progression of diseases.

In the future, we will explore the equilibrium between the computation cost and the quality of generative images by strengthening the efficacy of score-based diffusion sampler. For the semi-supervised segmentation framework, we will improve our network by replacing the boundary-aware transformer encoder with a light-weight transformer encoder to reduce the computational complexity. Additionally, we will integrate the light-weight transformer with CNN-architecture in the shallow layers to capture more global information during training to boost the segmentation performance.

References

[1] Voraporn Chaikitmongkol, Jun Kong, Preeyanuch Khunsongkiet, Direk Patikulsila, Mira Sachdeva, Pimploy Chavengsaksongkram, Chutikarn Dejkriengkraikul, Pawara Winaikosol, Janejit Choovuthayakorn, Nawat Watanachai, et al. Sensitivity and specificity of potential diagnostic features detected using fundus photography, optical coherence tomography, and fluorescein angiography for polypoidal choroidal vasculopathy. *JAMA ophthalmology*, 137(6):661–667, 2019.

[2] Shuai Chen, Gerda Bortsova, Antonio García-Uceda Juárez, Gijs Van Tulder, and Marleen De Bruijne. Multi-task attention-based semi-supervised learning for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*, pages 457–465. Springer, 2019.

[3] Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis*, 54:280–296, 2019.

[4] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[5] Botond Fazekas, Guilherme Aresta, Dmitrii Lachinov, Sophie Riedl, Julia Mai, Ursula Schmidt-Erfurth, and Hrvoje Bogunović. Sd-layernet: Semi-supervised retinal layer segmentation in oct using disentangled representation with anatomical priors. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 320–329. Springer, 2022.

[6] Xingxin He, Leyuan Fang, Mingkui Tan, and Xiangdong Chen. Intra-and inter-slice contrastive learning for point supervised oct fluid segmentation. *IEEE Transactions on Image Processing*, 31:1870–1881, 2022.

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[8] Junjie Hu, Yuanyuan Chen, and Zhang Yi. Automated segmentation of macular edema in oct using deep neural networks. *Medical image analysis*, 55:216–227, 2019.

[9] Thomas S Hwang, Simon S Gao, Liang Liu, Andreas K Lauer, Steven T Bailey, Christina J Flaxel, David J Wilson, David Huang, and Yali Jia. Automated quantification of capillary nonperfusion using optical coherence tomography angiography in diabetic retinopathy. *JAMA ophthalmology*, 134(4):367–373, 2016.

[10] Jongin Lim, Daeho Um, Hyung Jin Chang, Dae Ung Jo, and Jin Young Choi. Class-attentive diffusion network for semi-supervised classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8601–8609, 2021.

[11] Donghuan Lu, Morgan Heisler, Sieun Lee, Gavin Weiguang Ding, Eduardo Navajas, Marinko V Sarunic, and Mirza Faisal Beg. Deep-learning based multiclass retinal fluid segmentation and detection in optical coherence tomography images using a fully convolutional neural network. *Medical image analysis*, 54:100–110, 2019.

[12] Fei Lyu, Mang Ye, Jonathan Frederik Carlsen, Kenny Erleben, Sune Darkner, and Pong C Yuen. Pseudo-label guided image synthesis for semi-supervised covid-19 pneumonia infection segmentation. *IEEE Transactions on Medical Imaging*, 42(3):797–809, 2022.

[13] Yuhui Ma, Huaying Hao, Jianyang Xie, Huazhu Fu, Jiong Zhang, Jianlong Yang, Zhen Wang, Jiang Liu, Yalin Zheng, and Yitian Zhao. Rose: a retinal oct-angiography vessel segmentation dataset and new model. *IEEE transactions on medical imaging*, 40(3):928–939, 2020.

[14] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.

[15] Abdolreza Rashno, Dara D Koozekanani, Paul M Drayna, Behzad Nazari, Saeed Sadri, Hossein Rabbani, and Keshab K Parhi. Fully automated segmentation of fluid/cyst regions in optical coherence tomography images with diabetic macular edema using neutrosophic sets and graph algorithms. *IEEE Transactions on Biomedical Engineering*, 65(5):989–1001, 2017.

[16] Yuhe Shen, Jiang Li, Weifang Zhu, Kai Yu, Meng Wang, Yuanyuan Peng, Yi Zhou, Liling Guan, and Xinjian Chen. Graph attention u-net for retinal layer surface detection and choroid neovascularization segmentation in oct images. *IEEE Transactions on Medical Imaging*, 2023.

[17] Fenghe Tang, Jianrui Ding, Lingtao Wang, Min Xian, and Chunping Ning. Multi-level global context cross consistency model for semi-supervised ultrasound image segmentation with diffusion model. *arXiv preprint arXiv:2305.09447*, 2023.

[18] Yuhui Tao, Xiao Ma, Yizhe Zhang, Kun Huang, Zexuan Ji, Wen Fan, Songtao Yuan, and Qiang Chen. Lagan: Lesion-aware generative adversarial networks for edema area segmentation in sd-oct images. *IEEE Journal of Biomedical and Health Informatics*, 2023.

[19] Avinash V Varadarajan, Pinal Bavishi, Paisan Ruamviboonsuk, Peranut Chotcomwongse, Subhashini Venugopalan, Arunachalam Narayanaswamy, Jorge Cuadros, Kuniyoshi Kanai, George Bresnick, Mongkol Tadarati, et al. Predicting optical coherence tomography-derived diabetic macular edema grades from fundus photographs using deep learning. *Nature communications*, 11(1):130, 2020.

[20] Meng Wang, Tian Lin, Yuanyuan Peng, Weifang Zhu, Yi Zhou, Fei Shi, Kai Yu, Qingquan Meng, Yong Liu, Zhongyue Chen, et al. Self-guided optimization semi-supervised method for joint segmentation of macular hole and cystoid macular edema in retinal oct images. *IEEE Transactions on Biomedical Engineering*, 2023.

[21] Meng Wang, Weifang Zhu, Kai Yu, Zhongyue Chen, Fei Shi, Yi Zhou, Yuhui Ma, Yuanyuan Peng, Dengsen Bao, Shuanglang Feng, et al. Semi-supervised capsule cgan for speckle noise reduction in retinal oct images. *IEEE transactions on medical imaging*, 40(4):1168–1183, 2021.

[22] Gang Xing, Li Chen, Hualin Wang, Jiong Zhang, Dongke Sun, Feng Xu, Jianqin Lei, and Xiayu Xu. Multi-scale pathological fluid segmentation in oct with a novel curvature loss in convolutional neural network. *IEEE Transactions on Medical Imaging*, 41(6):1547–1559, 2022.

[23] Jiadong Yang, Yuhui Tao, Qiuzhuo Xu, Yuhan Zhang, Xiao Ma, Songtao Yuan, and Qiang Chen. Self-supervised sequence recovery for semi-supervised retinal layer segmentation. *IEEE Journal of Biomedical and Health Informatics*, 26(8):3872–3883, 2022.

[24] Yuhan Zhang, Mingchao Li, Zexuan Ji, Wen Fan, Songtao Yuan, Qinghuai Liu, and Qiang Chen. Twin self-supervision based semi-supervised learning (ts-ssl): Retinal anomaly classification in sd-oct images. *Neurocomputing*, 462:491–505, 2021.

[25] Feihui Zheng, Jacqueline Chua, Mengyuan Ke, Bingyao Tan, Marco Yu, Qinglan Hu, Chui Ming Gemmy Cheung, Marcus Ang, Shu Yen Lee, Tien Yin Wong, et al. Quantitative oct angiography of the retinal microvasculature and choriocapillaris in highly myopic eyes with myopic macular degeneration. *British Journal of Ophthalmology*, 106(5):681–688, 2022.

[26] Yi Zhou, Xiaodong He, Lei Huang, Li Liu, Fan Zhu, Shanshan Cui, and Ling Shao. Collaborative learning of semi-supervised segmentation and classification for medical images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2079–2088, 2019.