

# Evaluation of OCT Image Synthesis for Choroidal and Retinal Layer Segmentation Using Denoising Diffusion Probabilistic Models

Yudai Yamauchi<sup>1,2</sup><sup>a</sup>, Yuli Wu<sup>2</sup><sup>b</sup> and Eiji Okada<sup>1</sup><sup>c</sup>

<sup>1</sup>Keio University, Department of Electronics and Electrical Engineering, Japan

<sup>2</sup>Institute of Imaging and Computer Vision, RWTH Aachen University, Germany

**Keywords:** Retinal OCT Image, Denoising Diffusion Probabilistic Models, Image Synthesis.

**Abstract:** Machine learning can automatically conduct the layer segmentation task of retinal optical coherence tomography (OCT) image, but annotated data is required to train these models. Synthetic retinal OCT images are generated using denoising diffusion probabilistic models (DDPMs), which can be used to train segmentation models effectively and automatically create annotated data. However, the extent to which these synthetic images contribute to segmentation accuracy compared to real data has not been investigated. In this study, we synthesized retinal OCT images from sketch images using DDPMs, trained a segmentation model using synthetic and real images, and evaluated how the use of synthetic images influenced the accuracy of choroidal and retinal layer segmentation compared to results using only real images. Through a comparison of the Dice score, we confirmed that training with both synthetic and real OCT images led to higher Dice scores than training with only real OCT images. These findings suggest that using synthetic images can enhance segmentation accuracy, offering a promising approach to improving model performance in situations with limited annotated real data.


## 1 INTRODUCTION


The human retina is a collection of thin layers that line the inner wall of the eye, and it is a vital organ that receives light information from outside and converts it into visual information. It is widely recognized that the thickness of retinal layers is associated with various diseases. For instance, the thickness of the choroid, which underlies the retina, is also associated with certain diseases and increases the likelihood of early-atrophic age-related macular degeneration (Sigler et al., 2014). Tomographic images essential for diagnosing retinal diseases are obtained through optical coherence tomography (OCT), a technology widely applied in ophthalmology and other fields. OCT is a non-invasive technique used to obtain information about the refractive index structure within biological tissues. OCT can capture high-resolution internal tissue structures without the need for complex algorithms (Schmitt, 1999).


Usually, the retinal layers in an OCT image are

manually segmented by an ophthalmologist to measure the thickness of the layers. However, manual segmentation is highly complex and time-consuming (Ye et al., 2023). Therefore, automatic segmentation using machine learning has been widely studied in recent years (He, 2021). Machine learning models for segmentation are usually supervised learning, and the supervised learning model requires an annotated dataset. For this reason, annotated datasets such as the COCO dataset (Lin et al., 2014), which contains over 200k labelled images for segmentation tasks, and the Open Images v4 Dataset (Kuznetsova et al., 2020), which contains 9.2M images, are publicly available and are widely used for training models. Therefore, automatic segmentation using machine learning requires a sufficiently annotated dataset for training, but annotated medical image datasets are usually very limited (Wang et al., 2021). Hence, research has been carried out to generate synthetic images corresponding to the ground-truth label using generative models.

Generative adversarial networks (GANs) (Goodfellow et al., 2014) have shown remarkable results in various generative tasks and have been used in various settings. It is used for various medical imaging tasks, such as synthesizing retinal OCT images (Zheng

<sup>a</sup> <https://orcid.org/0009-0008-5579-8116>

<sup>b</sup> <https://orcid.org/0000-0002-6216-4911>

<sup>c</sup> <https://orcid.org/0000-0002-7846-7677>

et al., 2020) and brain tumor MR images (Mukherjee et al., 2022). The flow-based model (Rezende and Mohamed, 2015) can directly learn the data distribution of the training data, enabling the rapid generation of many images. It has been used in applications such as the reconstruction of CT and MR images (Denker et al., 2021), as well as synthesizing of chest X-ray images (Hajij et al., 2022). While GAN can synthesize high-quality images, it is known to suffer from "mode collapse," in which the model becomes unstable during training and synthesizes images only similar to those in the input data. The flow-based model demonstrates stable training, but the quality of the generated images is low (Xiao et al., 2021).

Most recently, denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020) have emerged as one of the most promising generative models, known for their stability during training and their ability to produce high-quality data. Unlike GAN and the flow-based model, DDPMs demonstrate stable learning and consistently generate higher-quality images, while also being capable of generating diverse data without causing "mode collapse" (Dhariwal and Nichol, 2021; Müller-Franzes et al., 2023; Xiao et al., 2021). DDPMs have successfully synthesised high-quality microscopy images from simple structural sketches for cell tracking tasks (Eschweiler et al., 2024; Yilmaz et al., 2024). Previous studies have also used DDPMs to generate annotated datasets to train segmentation models, improving the accuracy of automatic layer segmentation in retinal OCT images (Wu et al., 2024). As suggested in (Eschweiler et al., 2024; Yilmaz et al., 2024; Wu et al., 2024), it is important to evaluate the generated biomedical images directly with the downstream task, such as segmentation and tracking. However, since synthetic data is not real data but artificially generated, it is necessary to evaluate the extent to which synthetic data can substitute for real data. In this study, we assess the quality of DDPM-generated images by evaluating the performance of a layer segmentation model trained on synthetic retinal OCT images and comparing it with a model trained on real retinal OCT images.

## 2 METHODS

### 2.1 Denoising Diffusion Probabilistic Models

There are two processes in the DDPMs training phase. The first is the diffusion process, which progressively adds noise to the original image—the image we want to train—until it becomes "pure noise," meaning it no

longer retains any information about the original image. The second is the reverse process, which reverses the diffusion process by removing noise to reconstruct the original image. Figure 1a illustrates the training pipeline, where  $X_0$  represents the retinal OCT image, and  $T$  denotes the total number of time steps during which noise accumulates. The noise is determined at each time according to the following equation 1.

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t \mathbf{I}) \quad (1)$$

where  $t$  is  $1 < t < T$ . Moreover,  $\beta_t$  is a variable that determines the amount of noise added. At each time step,  $\beta_t$  is incrementally adjusted, being set to  $\{\beta_1, \beta_2, \dots, \beta_T\}$ .

Next, the reverse process is described. In this process, a neural network is used to denoise the original OCT image from pure noise.  $p_\theta$ , the diffusion process that generates the data  $x_{t-1}$  from  $x_t$  to the previous time step can be expressed as follows. The learning process starts with the pure noise distribution  $p(x_T) = \mathcal{N}(x_T; 0, I)$ , and proceeds through Gaussian transitions as described in Equation 2:

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \quad (2)$$

After the training phase, we move to the synthesis phase. Figure 1b illustrates the synthetic pipeline. Retinal OCT image is input in the training phase. However, in the synthesis phase, we use an image that has a rough texture and structure of the retinas: the sketch image. Assuming the total time step is trained at 1000, the number of time steps can be changed to obtain images with different modalities. The upper part of Figure 1b shows an example with time step 100 and the lower part with time step 900. The synthetic image not only retains many of the characteristics of the sketch image but also fails to capture the characteristic qualities inherent to the OCT image. In contrast, in the case of time step 900, the amount of noise added is large, so the synthetic image captures the characteristic qualities inherent to the OCT image. However, due to too much noise being added, the synthetic image doesn't retain the shape of the sketch image. Therefore, optimization is needed to identify the ideal time step with the OCT image's features and preserve the sketch image's shape.

### 2.2 Choroidal and Retinal Layer Segmentation

Segmentation was performed using U-Net (Ronneberger et al., 2015), widely used in medical image segmentation tasks. U-Net is constructed by encoder

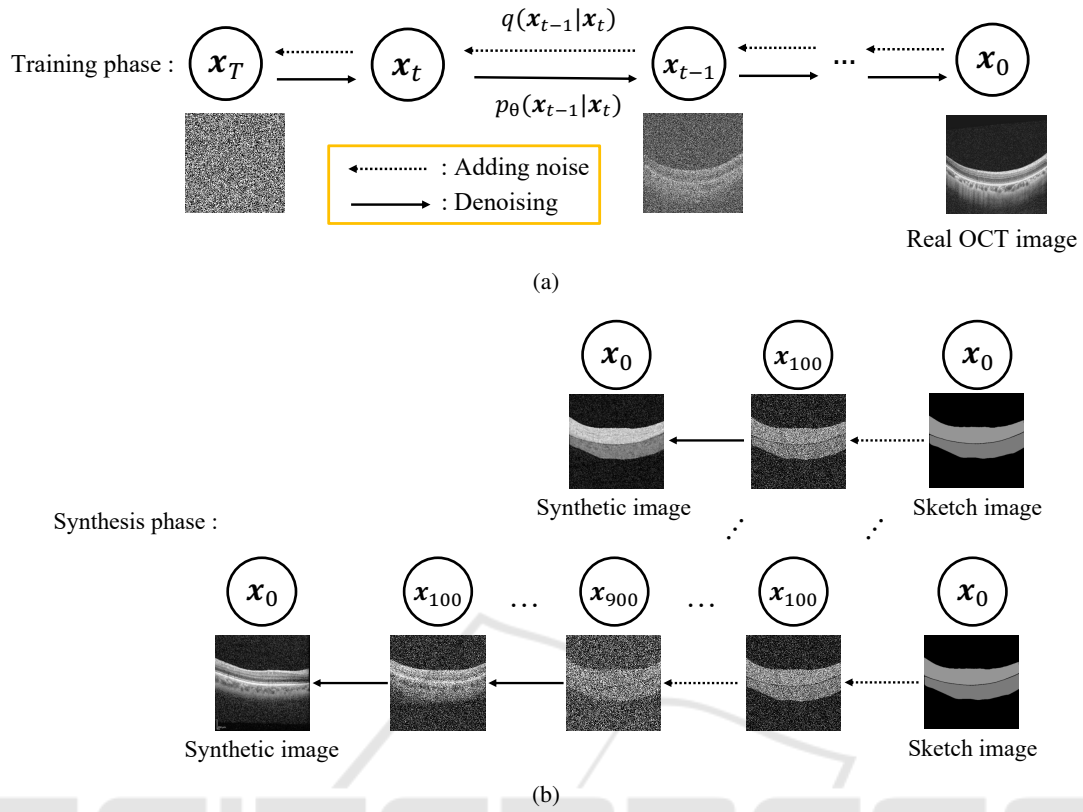


Figure 1: (a) Training pipeline. This workflow illustrates the DDPMs training using real OCT Image. (b) Synthesis pipeline. The upper workflow illustrates image synthesis with time step 100. The lower workflow illustrates image synthesis with time step 900.

and decoder. The encoder converts the input image into a low-dimensional feature representation, which extracts local and global features of the image by progressively reducing the image's resolution. The decoder uses the output of the encoder to restore the low-dimensional feature map to the resolution of the original input image. Dice score is often used to quantify the result of segmentation and is defined as follows Equation 3.

$$\text{Dice}(c) = \frac{2|A_c \cap B_c|}{|A_c| + |B_c|} \quad (3)$$

Here,  $|A_c|$  is the number of pixels of class  $c$  in the ground truth for the input image,  $|B_c|$  is the number of pixels of class  $c$  in the predicted image for output, and  $|A_c \cap B_c|$  denotes the number of overlapping pixels of class  $c$  between the ground truth and the predicted image.

### 3 EXPERIMENTS AND RESULTS

#### 3.1 Dataset

We used the OIMHS dataset, an open-source retinal OCT image dataset. This dataset was used by Ye's research group (Ye et al., 2023) using the Spectral-domain OCT (SD-OCT) system (Spectralis HRA OCT, Heidelberg Engineering, Heidelberg, Germany) to obtain retinal OCT images of patients with macular holes. This dataset contains 3859 retinal OCT images of 119 patients with macular holes and a set of four segmentation labels provided by a skilled ophthalmologist: retinal layer, macular hole, intraretinal cysts, and choroidal layer. The image set also contains an image quality assessment based on an objective assessment (low signal strength) and two subjective perspectives (i.e. signal shield and image blur). This study targets pure retina layer segmentation. Thus, we used OCT images containing only the retinal and choroidal layers. Additionally, to avoid differences in the images used for training, we employed a set of 1,179 images that were not classified as having any

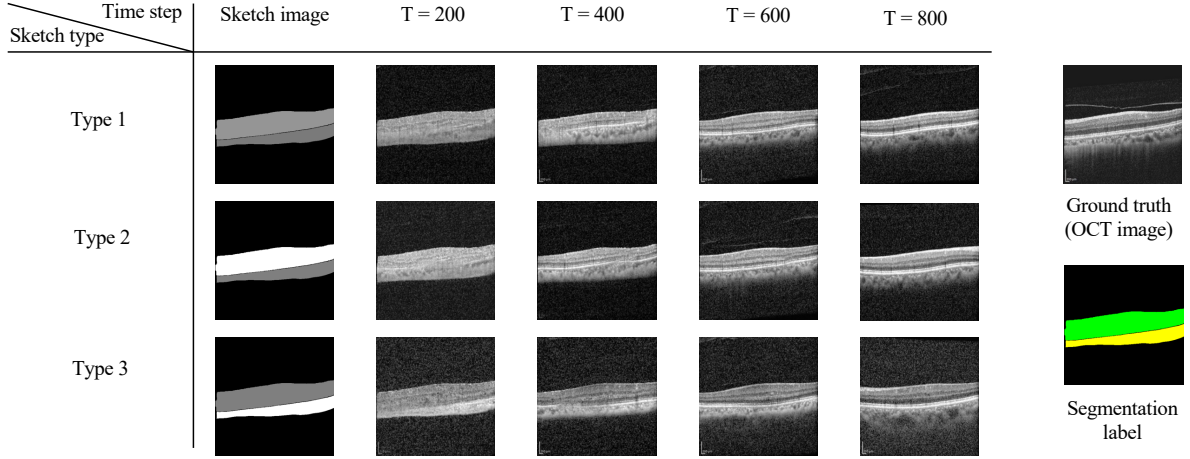


Figure 2: Example of synthetic OCT images generated using three types sketch image. The row shows the type of sketch. The column shows the input sketch image and each time step. sketch image represents the segmentation label(right below the image) in grey scale. Type 1: The sketch image uses the same pixel values as the OCT image in the dataset. Types 2 and 3: sketch images with the highest contrast between layers. Specifically, the pixel values in each type of sketch image are as follows: in Type 1, the retinal layer is set to 149, and the choroidal layer is set to 122; in Type 2, the retinal layer is 255, and the choroidal layer is 127; in Type 3, the retinal layer is 127, and the choroidal layer is 255.

problems based on quality assessment. The original image size of  $512 \times 512$  was downsampled to  $256 \times 256$  to reduce computer memory. 79 images were also allocated for segmentation testing and 1100 images for training the segmentation model. Finally, 550 images from the segmentation training images were allocated for training the DDPMs and 550 images for synthesising images using the DDPMs.

### 3.2 Image Synthesis

We set the total time step:  $T = 1000$  for DDPMs training and cosine-based scheduling of the variance  $\beta_t$  ( $\beta_1 = 10^{-4}$  to  $\beta_T = 0.02$ ). The sketch image was created by grayscaling the segmentation label of the dataset. Three types of sketches were created because there are three possible pixel values for grayscaling. The average pixel value in the retinal layer is 149, and the choroidal layer is 122, which are pixel values of the OCT image in the dataset. Therefore, we set the same pixel value to the sketch image, the input image to DDPMs in the synthesis phase. This sketch image was defined as Type 1.

On the other hand, as the synthetic image is used for segmentation training, it is considered more effective during segmentation if the contrast between the layers is emphasized. Therefore, the OCT pixel values are ignored, and the sketch image with the highest contrast between layers is used. In this case, the pixel value range is from 0 to 255, so the background is set to 0, and 127 and 255 are allocated to the two layers.

The sketch image with 255 allocated to the reti-

nal layer and 127 to the choroidal layer is defined as sketch image Type 2, whereas a sketch image with 255 allocated to the choroidal layer and 127 to the retinal layer is defined as sketch image Type 3.

We trained the DDPMs on 550 retinal OCT images and synthesized 550 images of three different sketch images with different time steps  $t$  starting from  $\{100 \text{ to } 900, \text{ interval } 100\}$ . An example of this image synthesis is shown in Figure 2.

Moreover, sketch images as input to DDPMs have unnatural changes in pixel values between layers. Gaussian blur was added to create natural boundaries and perturb the pixel intensity of the image. It has been reported that applying this makes the composite image more closely resemble the original OCT image (Wu et al., 2024).

We evaluated the image accuracy of the synthetic image using the quantitative evaluation methods Peak signal-to-noise ratio (PSNR), Structural similarity index measure (SSIM) (Wang and Bovik, 2009) and Fréchet inception distance (FID) (Heusel et al., 2017). The PSNR represents the ratio between the maximum possible power of a signal and its noise. A higher PSNR value indicates less distortion or error, meaning the signal retains more of its original quality; SSIM evaluates the similarity between two images by considering changes in structural information, luminance, and contrast. A higher SSIM value indicates more similarity, meaning the compared images retain similar structural and perceptual qualities; FID evaluates how close the generated image is to the real image by comparing the mean and variance, with lower



Table 1: Evaluate image quality. Image quality metrics are Peak signal-to-noise ratio (PSNR), Structural similarity index measure (SSIM) and Fréchet inception distance (FID).

time step	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$
<b>Type 1</b>			
100	$16.59 \pm 0.66$	$0.29 \pm 0.06$	268.32
200	$16.71 \pm 0.62$	$0.30 \pm 0.06$	189.20
300	$16.85 \pm 0.61$	$0.30 \pm 0.05$	148.04
400	$17.06 \pm 0.64$	$0.31 \pm 0.05$	115.54
500	$17.20 \pm 0.66$	$0.31 \pm 0.05$	98.90
600	$17.40 \pm 0.72$	$0.31 \pm 0.05$	86.09
700	$17.60 \pm 0.80$	$0.31 \pm 0.06$	75.21
800	$17.55 \pm 0.97$	$0.31 \pm 0.06$	72.36
900	$16.77 \pm 1.11$	$0.31 \pm 0.06$	74.89
<b>Type 2</b>			
100	$16.56 \pm 0.66$	$0.29 \pm 0.06$	266.24
200	$16.70 \pm 0.62$	$0.30 \pm 0.06$	190.01
300	$16.83 \pm 0.60$	$0.30 \pm 0.05$	145.95
400	$16.96 \pm 0.61$	$0.31 \pm 0.05$	120.05
500	$17.18 \pm 0.69$	$0.31 \pm 0.05$	96.56
600	$17.43 \pm 0.73$	$0.31 \pm 0.05$	86.35
700	$17.57 \pm 0.75$	$0.31 \pm 0.06$	76.28
800	$17.53 \pm 0.95$	$0.31 \pm 0.06$	71.02
900	$16.83 \pm 1.06$	$0.31 \pm 0.06$	77.53
<b>Type 3</b>			
100	$16.47 \pm 0.82$	$0.23 \pm 0.06$	284.65
200	$16.71 \pm 0.78$	$0.24 \pm 0.05$	214.99
300	$16.91 \pm 0.79$	$0.25 \pm 0.05$	163.79
400	$17.09 \pm 0.81$	$0.26 \pm 0.06$	132.98
500	$17.21 \pm 0.91$	$0.26 \pm 0.06$	115.62
600	$17.25 \pm 0.97$	$0.26 \pm 0.06$	110.35
700	$17.18 \pm 1.06$	$0.25 \pm 0.06$	108.37
800	$16.88 \pm 1.14$	$0.26 \pm 0.06$	106.87
900	$15.98 \pm 1.09$	$0.27 \pm 0.05$	109.95

FID indicating more significant similarity to the real image. For this work, we adopted the sentence informing about using Clean-FID (Parmar et al., 2022) that improved the reliability and consistency compared with a usual FID. Table 1 compares the original images and synthesized images at time steps 100 to 900 and sketch image Types 1 to 3, using PSNR, SSIM, and FID as evaluation metrics. The results show that the quality of both images improves as time step increases. However, it can be seen that the accuracy of PSNR and SSIM decreases after time step 700. A possible explanation for this may be the mismatching of composite images, as mentioned in Section 3.1; it is known that if too much noise is added, the original image shape cannot be preserved. Therefore, when the total time step is 1000, it can be seen that image mismatching occurs after time step 700.

In addition, the FID of Type 1 and Type 2 decrease almost the same way, but only for Type 3 the FID does

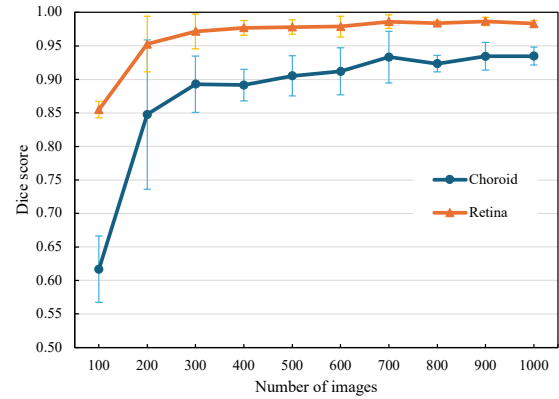


Figure 3: Mean dice score (y axis) trained with retinal OCT images among a number of training image (x axis), orange line shows result of retinal layer, dark blue line shows result of choroidal layer.

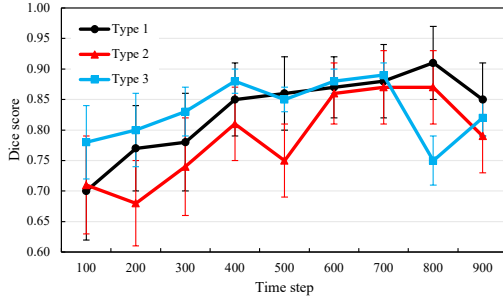
not fall below 100 even if the time step is increased. In OCT images, the average pixel intensity is higher in the retinal layer and lower in the choroidal layer. The results in Table 1 indicate that in both cases—when the pixel values in the sketch image are close to those in the OCT image and when the contrast between layers in the sketch image is clear—aligning the pixel value order with the original image leads to a more accurate synthesized image.

The subsequent step involves evaluating the types of synthetic images optimal for training a segmentation model. For the segmentation evaluation, 500 of the 550 synthetic images were used for training, 50 for validation and 79 test OCT images were used for segmentation during testing. Figure 4 shows the Dice score when synthetic images at each time step were used for training. The results of Dice score show that Type 2 has a better Dice score than Types 1 and 3 when the time step is below 600. However, for time steps higher than Time 700, the Dice score is reversed between Types 1, 3, and 2. Type 2 was created to emphasize the contrast between layers in sketch images, which may have facilitated segmentation when the time step was low; as the time step improved, the pixel values in Type 1 converged and were closer to those in the original OCT images. Therefore, it is considered that the best Dice score was produced.

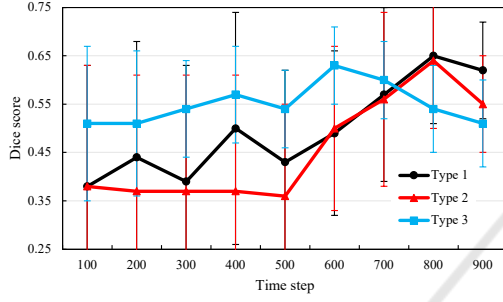
### 3.3 Evaluating Synthesis with Layer Segmentation

The number of training samples significantly influences segmentation performance when training a segmentation model solely with OCT images.

Figure 3 illustrates the model's performance as the



(a) Retinal Layer



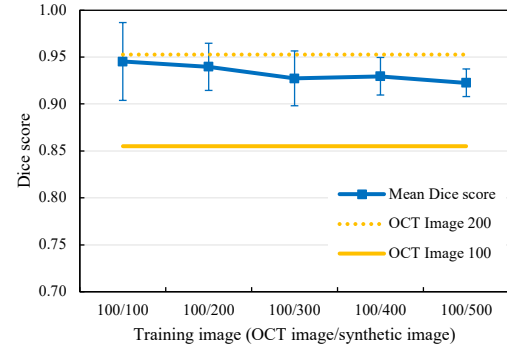
(b) Choroidal Layer

Figure 4: Mean Dice score (y axis) trained with 500 synthesized image among each time step (x axis), types of 3 sketch images that are input to DDPMs synthesis. (a) Results of the retinal layer. (b) Results of the choroidal Layer.

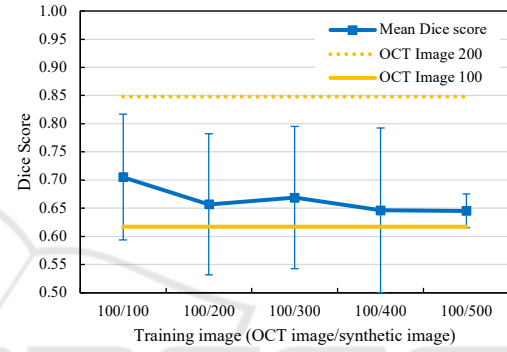
number of OCT images used for training increases incrementally from 100 to 1000 images in steps of 100 images. As shown in Figure 3, the Dice score remains notably lower when the model is trained on only 100 images, compared to the performance achieved with 200 or more training images. This observation suggests that a dataset of less than 100 images may need to be sufficient for the model to adequately learn segmentation features, indicating a threshold in sample size necessary to achieve stable segmentation performance.

Secondly, we evaluate how much effect the Segmentation predicted is using synthetic images. We calculate a Dice score from a segmentation model that is trained with 100 OCT images and synthetic images. Figure 5 shows the Dice score of layer segmentation when training with 100-500 synthetic images for 100 OCT images.

Based on the findings in Section 3.2, we selected synthetic images of Sketch Type 1 as input for the DDPM, specifically choosing those generated at time step 800 as training data for the segmentation model. Our results show a clear improvement in Dice score accuracy when these synthetic images are added to



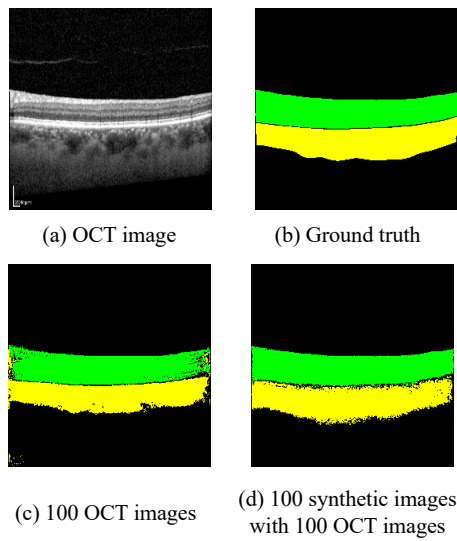
(a) Retinal Layer



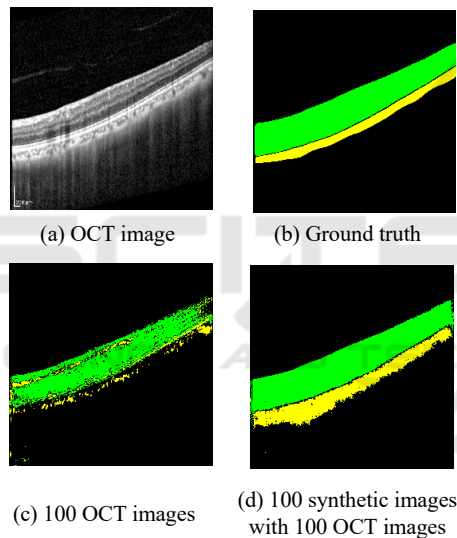
(b) Choroidal Layer

Figure 5: Mean Dice score (y axis) trained with 100 OCT and 100 to 500 synthesized image. The blue line is the average Dice score; the yellow line is the average Dice score trained on 100 OCT images; the yellow dotted line is the average Dice score trained on 200 OCT images. (a) Results of the retinal layer. (b) Results of the choroidal layer.

a training set of 100 OCT images. This suggests that synthetic images contribute positively to model performance, helping to alleviate some of the limitations posed by a smaller dataset of real OCT images. However, it is important to note that while the addition of synthetic images enhances segmentation performance, it does not fully match the accuracy achieved when training with 200 OCT images alone for both the retinal and choroidal layers. A likely reason for this is a structural mismatch within the synthetic images. Specifically, although synthetic images generated at time step 800 share properties with the OCT modality on which the DDPM was trained, they do not entirely replicate the detailed layer structures present in real OCT images or those outlined in the input sketch images. As a result, when the segmentation model is trained using a combination of synthetic and real OCT images, its performance remains slightly lower than when trained with an equivalent



I. Best examples.



II. Worst examples.

Figure 6: Input OCT images predicted each layer when the segmentation model is trained with 100 OCT images and the best Dice score (I) and the worst Dice score (II). The green region shows the layer, and The yellow region shows the choroidal layer. (a) Input OCT image. (b) Ground truth. (c) Segmentation predicted from the segmentation model trained on 100 OCT images. (d) Segmentation predicted from the segmentation model trained on synthetic 100 images and 100 OCT images.

number of only real OCT images. Nevertheless, the addition of synthetic images led to a notable improvement in the Dice score for retinal layer segmentation, with an increase of nearly 0.1. This result is comparable to the performance achieved when training on 200 real OCT images, highlighting synthetic images' potential to effectively augment training datasets when

real data is limited. However, it was observed that the segmentation accuracy did not improve when more than 200 synthetic images were added. This is likely because the contribution of real images diminishes as the absolute quantity of synthetic images increases.

We also discuss the results of individual segmentations. Figure 6-I show the segmentation results with the highest Dice score, achieved by training with only 100 OCT images, Figure 6-II shows the segmentation results with the lowest Dice score. In Figure 6-I, training with 100 OCT images, shows that the model misclassified a region that should be the retinal layer (as per the ground truth) as the choroidal layer. However, when trained with an additional 100 synthetic images, the model correctly identified this region as the retinal layer. Similarly, Figure 6-II highlights a significant improvement in the accuracy of choroidal layer segmentation when synthetic images were added. Examining Dice scores, we see that training with only 100 OCT images yields scores of 0.88 for the retinal layer and 0.75 for the choroidal layer in Figure 6-II. Adding 100 synthetic images improves these scores to 0.92 and 0.83, respectively. Likewise, in Figure 6-I, training with only 100 OCT images produces Dice scores of 0.78 for the retinal layer and 0.21 for the choroidal layer, which increases to 0.93 and 0.68, respectively, when synthetic images are included. Notably, in Figure 6-II, the Dice score for the choroidal layer improves by 0.47. In supervised learning, limited training data often hinders model performance on new data. This limitation is evident in Figure 6-II, where segmentation performance is suboptimal with only OCT images. By supplementing the dataset with synthetic images, we effectively increased the training data, resulting in a marked improvement in Dice scores.

## 4 CONCLUSIONS

In this study, we synthesized retinal OCT images and compared the predicted segmentation results obtained by using both the synthetic and real OCT images against those obtained by using only the real OCT images. The results showed that in the retinal layer, the segmentation predicted were comparable to those obtained when only OCT images were used for training. It was also found that the pixel values of the sketch image used as input to DDPMs during image synthesis should be based on the pixel values of the training images of DDPMs to achieve higher-quality synthesized images. This study demonstrates that incorporating images synthesized through DDPMs can effectively enhance segmentation model training, par-

ticularly in cases with limited real images. Notably, segmentation accuracy improved markedly in cases where initial segmentation accuracy was lower, underscoring the value of synthetic images for segmentation model training in achieving robust model performance.

## REFERENCES

- Denker, A., Schmidt, M., Leuschner, J., and Maass, P. (2021). Conditional invertible neural networks for medical imaging.
- Dhariwal, P. and Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794.
- Eschweiler, D., Yilmaz, R., Baumann, M., Laube, I., Roy, R., Jose, A., Brückner, D., and Stegmaier, J. (2024). Denoising diffusion probabilistic models for generation of realistic fully-annotated microscopy image datasets. *PLOS Computational Biology*, 20(2):e1011890.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.
- Hajij, M., Zamzmi, G., Paul, R., and Thukar, L. (2022). Normalizing flow for synthetic medical images generation. In *2022 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT)*, pages 46–49.
- He, Y. (2021). *Retinal OCT image analysis using deep learning*. PhD thesis, Johns Hopkins University.
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.
- Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.
- Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Mallocci, M., Kolesnikov, A., et al. (2020). The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.
- Mukherjee, D., Saha, P., Kaplun, D., Sinitca, A., and Sarkar, R. (2022). Brain tumor image generation using an aggregation of gan models with style transfer. *Scientific reports*, 12(1):9141.
- Müller-Franzes, G., Niehues, J. M., Khader, F., Arasteh, S. T., Haarbuerger, C., Kuhl, C., Wang, T., Han, T., Nolte, T., Nebelung, S., et al. (2023). A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Scientific Reports*, 13(1):12098.
- Parmar, G., Zhang, R., and Zhu, J.-Y. (2022). On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11410–11420.
- Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR.
- Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18*, pages 234–241. Springer.
- Schmitt, J. M. (1999). Optical coherence tomography (oct): a review. *IEEE Journal of selected topics in quantum electronics*, 5(4):1205–1215.
- Sigler, E., Randolph, J., Calzada, J., and Charles, S. (2014). Smoking and choroidal thickness in patients over 65 with early-atrophic age-related macular degeneration and normals. *Eye*, 28(7):838–846.
- Wang, S., Li, C., Wang, R., Liu, Z., Wang, M., Tan, H., Wu, Y., Liu, X., Sun, H., Yang, R., et al. (2021). Annotation-efficient deep learning for automatic medical image segmentation. *Nature communications*, 12(1):5915.
- Wang, Z. and Bovik, A. C. (2009). Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, 26(1):98–117.
- Wu, Y., He, W., Eschweiler, D., Dou, N., Fan, Z., Mi, S., Walter, P., and Stegmaier, J. (2024). Retinal oct synthesis with denoising diffusion probabilistic models for layer segmentation. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE.
- Xiao, Z., Kreis, K., and Vahdat, A. (2021). Tackling the generative learning trilemma with denoising diffusion gans. *arXiv preprint arXiv:2112.07804*.
- Ye, X., He, S., Zhong, X., Yu, J., Yang, S., Shen, Y., Chen, Y., Wang, Y., Huang, X., and Shen, L. (2023). Oimhs: An optical coherence tomography image dataset based on macular hole manual segmentation. *Scientific Data*, 10(1):769.
- Yilmaz, R., Eschweiler, D., and Stegmaier, J. (2024). Annotated biomedical video generation using denoising diffusion probabilistic models and flow fields. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 197–207.
- Zheng, C., Xie, X., Zhou, K., Chen, B., Chen, J., Ye, H., Li, W., Qiao, T., Gao, S., Yang, J., et al. (2020). Assessment of generative adversarial networks model for synthetic optical coherence tomography images of retinal disorders. *Translational Vision Science & Technology*, 9(2):29–29.