Contrastive Learning for Generating Optical Coherence Tomography Images of the Retina

Kaplan Sinan, Lensu Lasse

DOI: 10.1007/978-3-031-16980-9_11

Contrastive Learning for Generating Optical Coherence Tomography Images of the Retina

Sinan Kaplan$^{1[0000-0002-0849-934X]}$ and Lasse Lensu$^{1[0000-0002-7691-121X]}$

[1]Department of Computational Engineering, Lappeenranta-Lahti University of Technology LUT, P.O. Box 20, 53850 Lappeenranta, Finland
sinan.kaplan@student.lut.fi, lasse.lensu@lut.fi

Abstract. As a self-supervised learning technique, contrastive learning is an effective way to learn rich and discriminative representations from data. In this study, we propose a variational autoencoder (VAE) based approach to apply contrastive learning for the generation of optical coherence tomography (OCT) images of the retina. The approach first learns embedding representation from data by contrastive learning. Secondly, the learnt embeddings are used to synthesize disease-specific OCT images using VAEs. Our results reveal that the diseases are separated well in the embedding space and the proposed approach is able to generate high-quality images with fine-grained spatial details. The source code of the experiments in this paper can be found on Github[1].

Keywords: Optical coherence tomography · contrastive learning · variational autoencoder · deep generative model · deep learning · artificial intelligence.

1 Introduction

An increasing amount of effort has been put to the research and development of deep learning (DL) and its applications [6]. This methodology has shown its effectiveness as the state-of-the-art solution for many tasks [5] including medical image analysis. DL has also leveraged the potential for early detection and recognition of abnormalities, such as diabetic retinopathy and age-related macula degeneration, from retinal images [1]. One of the retinal imaging techniques that has benefited from DL is OCT [9]. This imaging modality sheds light on pathological structures of the retina in 3D, through which it is possible to reliably diagnose diseases such as choroidal neovascularization (CNV) and diabetic macular edema (DME). As a result of progress in both fields, DL based solutions are studied for the detection and recognition of such abnormalities from OCT images.

Despite its potential and successful applications in a variety of tasks, the practical utilisation of deep neural network (DNN) in safety-critical tasks like medical diagnosis systems is limited [11]. One of the underlying reasons of this

[1] https://github.com/kaplansinan/OCTRetImageGen_CLcVAE

is the black-box nature of the DL models [6]. The term black-box refers to the inability of understanding how the DL algorithm makes a particular decision [33]. This arises from the inherent structure of DL models, which are complex, non-linear, challenging to interpret, and the amount of data needed to train such models is typically large. To address the issues arising from the nature of DL models, there exists several solutions under the name of an emerging field called explainable artificial intelligence (XAI) [24, 2, 33]. XAI represents the techniques and methods to understand why artificial intelligence (AI) makes a particular decision. It is considered as key methodology to understand model decisions and build trust between the users and AI solutions [26, 3].

In XAI literature, the methods are split into two categories [2]: global and local. The global methods try to explain a particular model and data set used for training the model, whereas the local methods are utilized for understanding post-hoc decisions at instance level [24]. For example, considering a DL model trained to recognize certain diseases in OCT images [29], post-hoc XAI methods help in explaining which features in a target image are relevant to the model while performing inference. Such explanation is achieved by highlighting relevant regions in the target image [36]. Post-hoc explanations are often used for sensitivity analysis [25, 19], where the aim is to understand how the behavior of a model changes while manipulating the input image. There are different instance manipulation techniques [24] such as applying specific image transformations and augmentations like cropping, deleting certain part of the image [32], color transformation, and copy-pasting a part of the image.

An ideal solution for sensitivity analysis would be to avoid the limitations of existing data with synthetic data generated from an underlying distribution of real data used for training a DL model. To do so, one may apply deep generative models [16] like generative adversarial networks (GANs) [8] and VAEs [15] for image generation. In this paper, we propose a framework to synthesize OCT images using conditional variational autoencoder (cVAE) and contrastive learning [17]. The goal is to generate high-quality OCT images, which can be further used for the sensitivity analysis of OCT image classification tasks [13, 36]. In addition, the framework can be used to synthesize images for augmenting data sets, or constructing a benchmark OCT set.

The rest of the paper is organized as follows: Section 2 reviews the studies in OCT imaging, how DL is used for OCT image analysis and OCT image generation tasks. Section 3 introduces the proposed solution and Section 4 presents the experimental results. Finally, we present the conclusions that can be made based on the current experiment and give possible future directions in Section 5.

2 Related Work

OCT imaging and deep learning. OCT is a technique to acquire high resolution images of cross-sections of the retina [9]. It enables diagnosis of retinal disorders. For instance, CNV, DME, and DRUSEN are such disease, which can be diagnosed through OCT imaging [13, 23]. To detect such diseases from OCT images, techniques based on DL have received substantial attention in the medical image analysis field. For instance, classification [13] and segmentation tasks [9] are implemented to support clinicians while diagnosing specific abnormalities using OCT imaging.

OCT image generation by deep generative models. The potential of deep generative models for synthesizing high quality images have been proposed for OCT image generation tasks [20]. As part of deep generative models, GANs are studied widely compared to VAEs. For instance, Zha et al. proposed conditional generative adversarial networks (cGANs) to solve the class imbalance problem exhibit in OCT data sets [37]. In addition, by considering the difficulty of finding rare disease examples in OCT data sets, Xiao et al. adapted a GAN based method to generate and create an open OCT data set [34]. Furthermore, in another study [35], the authors aimed to improve the performance of OCT image classification tasks by using cGANs to generate images, thereby increasing the number of data samples.

OCT image generation by variational autoencoder. Although it has not received that much attention in OCT image generation tasks [22, 7], VAEs are an option for deep image generation. Compared to GANs, VAEs are easy to train and do not suffer from mode collapse [18]. In addition, another important advantage of VAEs is that they learn the characteristics of input data samples by mapping them into a latent space [16]. After training, new data samples are generated by sampling from this latent space. This way VAEs introduce a controlled way for the generation.

Our work. Since VAEs enable us to alter and explore the variations over the data, we choose VAEs to generate OCT images. To do so, generation is conditioned on learnt embeddings via the contrastive learning approach. Thus, our work combines contrastive learning and VAEs to synthesize disease-specific OCT images with appropriate visual details. To best of our knowledge, no other study has proposed this solution for the synthesis of OCT images of the retina.

3 Methods

The proposed solution consists of two stages. In the first stage, we use contrastive learning [17] to have class-wise discrimination in the learned embedding space. The embedding as a discriminative data representation enables class-specific image generation. In the second stage, we train cVAEs to generate the disease-specific OCT images. As the conditioning is done using the embeddings from the first stage, we are able to control the disease-wise data generation. Fig. 1 illustrates the model architectures used in the two stages.

Contrastive learning is a self-supervised learning technique widely applied for image retrieval tasks [17, 4]. The goal is to learn an embedding space in which the distances between similar samples are minimized while the distances between dissimilar samples are maximized. To learn such an embedding space, contrastive learning models are trained with specific loss functions, such as SimCLR [4], triplet-loss [27], and n-pair loss [30].

In our work, the contrastive learning model is inspired by the work in [14]. As being an effective batch construction method and reducing convergence time of the model, We use n-pair loss [30] to train the contrastive learning part. In addition, instead of using Resnet34 as the encoder part of the contrastive model, we use Resnet50 to increase the learning capacity of the model, thus having more discriminative embeddings for each class.

To generate new samples, we use a variational autoencoder representing one of the deep generative models. VAE maps the input data into a latent space in a probabilistic way by an encoder module. Afterwards, a decoder module is used to synthesize new data samples by sampling a latent vector from the latent space [16]. In a VAE, the goal is to minimize the distance between distribution of the input data and the distribution of the latent space using Kullback–Leibler (KL) divergence loss [15]. In this paper, we apply a cVAE, which conditions the synthesis of new samples on a given extra information, such as labels [16]. This contributes to the generation of the data in a desired way.

Conditional variational autoencoder is optimized with a weighted set of loss functions. VAEs often generate blurry images due to pixel-wise reconstruction loss. To avoid this issue, we replace the reconstruction loss by perceptual loss [10] and deep feature loss [21]. Hence, the objective of our cVAEs is to minimize the following weighted loss function:

$$L_{\text{CVAE}} = w_1 * L_{\text{perceptual}} + w_2 * L_{\text{feature}} + w_3 * L_{\text{KL}} \tag{1}$$

where $L_{\text{perceptual}}$ is perceptual loss, L_{feature} is feature loss and L_{KL} is KL divergence loss.

It is important to note that whilst designing the architecture of cVAE, in the decoder part we use sub-pixel convolutional layer [28] to increase the quality of generated images. This layer basically learns an array of image upscaling filters described in the original paper.

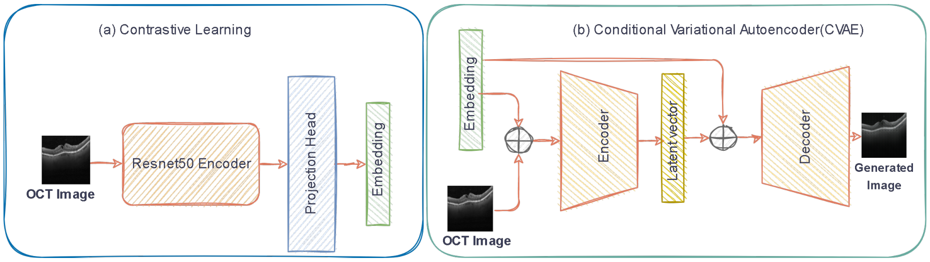

Fig. 1. The architectures of the proposed solution: a) contrastive learning model trained in Stage 1 and b) cVAE model trained in Stage 2. CVAE is conditioned on the embedding learnt by the contrastive learning part.

4 Experiments and Results

In this section, we cover characteristics of the training data, training procedure and results from contrastive learning and cVAE. [2]

4.1 Dataset

We use OCT data from the study in [12]. It consists of 84 495 labeled images split into 4 categories as follows: 37 200 CNV, 8 618 DRUSEN, 11300 DME and 26 300 NORMAL (healthy) images. The data has the issue of class imbalance. As a remedy, we apply a representative sampling approach given in the supplementary material. After the representative sampling, we reduce the data size to 19 980 samples equally distributed across each category. A few representative samples from the training set are presented in Fig. 2.

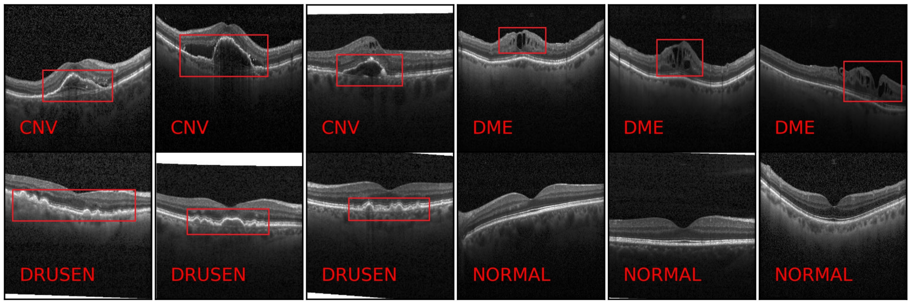

Fig. 2. Representative samples of OCT data for each category in the set. The red rectangle highlights the characteristics of each disease in an image.

4.2 Model training

We performed training in two stages. First the contrastive learning is trained to learn embeddings, which is used as conditioning information in the next stage of cVAE training. In the second stage, we train a cVAE model for each class to synthesize new OCT images. More details regarding the training hyperparameters, model input/output size, and training environment is given in the supplementary material.

[2] The in-depth details regarding the data set, training hyperparameters, and randomly generated OCT images by the trained model and the supplementary material can be found in the Github repo.

4.3 Results

Contrastive Learning The goal of contrastive learning is to learn discriminative embedding from each class. To verify this after the model training, we visualize the embedding space by principal-component analysis (PCA). Based on the visualization in Fig. 3, the contrastive learning provides good discrimination between each class in the embedding space. This is important for accurately mapping the input data into the latent space in the next stage.

It is also important to notice the representation of the NORMAL (healthy) cases in the embedding space. Based on the visualization, they are at the intersection of each disease and this strengthens the idea that the generation of new disease-specific OCT images representing different levels of severity of the condition is possible. While studying the existing images individually, we observed that the further an instance is located from the center of NORMAL cases, the more severe the disease is.

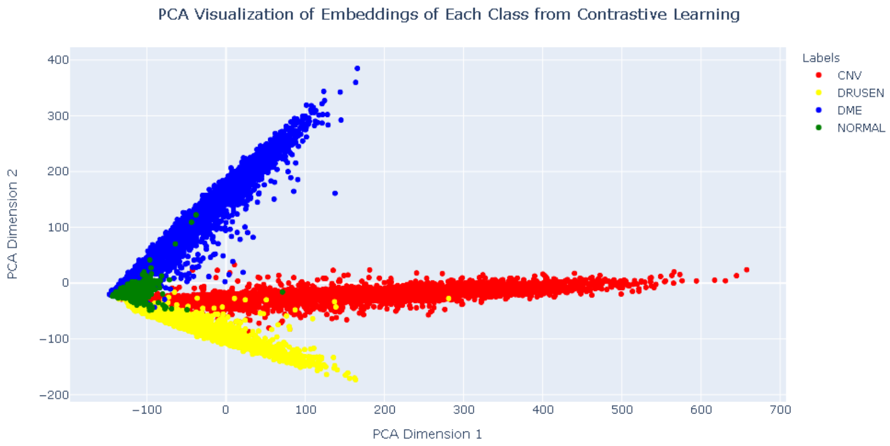

Fig. 3. Principal-component analysis projection of embedding of training set from contrastive learning.

CVAE In the second stage, by training a cVAE model for each disease, we generate high quality images that capture the characteristics of each disease successfully. We demonstrate pairwise visual comparison of generated images for each disease in Fig. 4, Fig. 5 and Fig. 6, respectively[3].

Pathological structures in the OCT images are captured well with disease specific details. However, we observe that the quality of generated images are

[3] Randomly generated samples from each class are presented in the supplementary material, which can be found in the Github repo.

better if there is less variation within a class. For instance, not all the fine-grained details are captured in the CNV class. The underlying reason behind this is that the class contains far more variability of images compared to DME and DRUSEN classes. The variation in CNV also exhibits in the embedding space of contrastive learning model (see Fig. 3).

In some of the generated images (see the supplementary material), we encounter checkerboard artefacts, which is due to the upsampling layers used in the decoder module of cVAE [31].

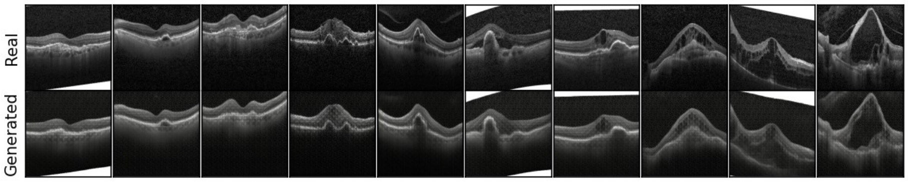

Fig. 4. The pair-wise visualization of generated choroidal neovascularization (CNV) samples: First row - real images; Second row - corresponding generated images.

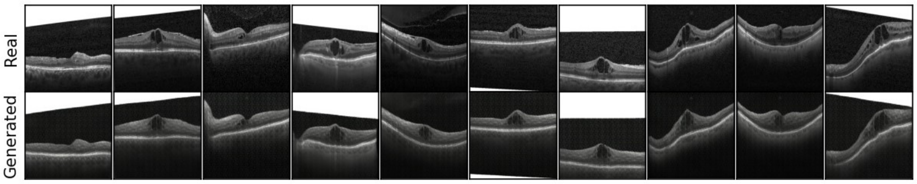

Fig. 5. The pair-wise visualization of generated diabetic macular edema (DME) samples: First row - real images; Second row - corresponding generated images.

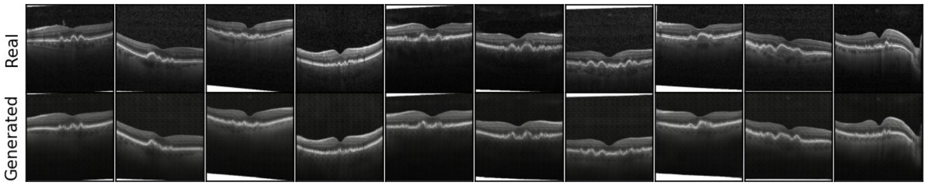

Fig. 6. The pair-wise visualization of generated DRUSEN samples: First row - real images; Second row - corresponding generated images.

5 Conclusions

In this paper, we study a contrastive learning based approach for synthesizing OCT images using cVAE. The contrastive learning is applied to extract rich representations from the data, which is further used by cVAE to generate new samples. Based on the presented results, the proposed method enables successful synthesis of visually quality OCT images representing CNV, DME, DRUSEN in fine-grained details. Among the aforementioned diseases, DME, DRUSEN cases are synthesized better than CNV.

Although, our main motivation is to generate images to be used in sensitivity analysis tasks, the image generation can be used variety of other tasks such as augmenting existing sets, counterfactual image generation and disease progression simulation. In the future work, we plan to combine OCT image classification done on the same set and use the proposed cVAE model to conduct sensitivity analysis and simulate disease progression. Also, in the extended study of this work, we plan to incorporate expert opinions to validate our observation about the different levels of severity of the diseases revealed in the contrastive learning part. We believe this can be helpful for both automated grading of the diseases from OCT images and simulating the progression of a certain disease.

References

1. Badar, M., Haris, M., Fatima, A.: Application of deep learning for retinal image analysis: A review. Computer Science Review **35**, 100203 (2020)
2. Bai, X., Wang, X., Liu, X., Liu, Q., Song, J., Sebe, N., Kim, B.: Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments. Pattern Recognition **120**, 108102 (2021)
3. Bruckert, S., Finzel, B., Schmid, U.: The next generation of medical decision support: A roadmap toward transparent expert companions. Frontiers in artificial intelligence p. 75 (2020)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
5. Dargan, S., Kumar, M., Ayyagari, M.R., Kumar, G.: A survey of deep learning and its applications: a new paradigm to machine learning. Archives of Computational Methods in Engineering **27**(4), 1071–1092 (2020)
6. Dong, S., Wang, P., Abbas, K.: A survey on deep learning and its applications. Computer Science Review **40**, 100379 (2021)
7. Gan, M., Wang, C.: Esophageal optical coherence tomography image synthesis using an adversarially learned variational autoencoder. Biomedical Optics Express **13**(3), 1188–1201 (2022)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
9. Islam, M.S., Wang, J.K., Johnson, S.S., Thurtell, M.J., Kardon, R.H., Garvin, M.K.: A deep-learning approach for automated oct en-face retinal vessel segmentation in cases of optic disc swelling using multiple en-face

images as input. Translational vision science technology **9**, 1–15 (2020). https://doi.org/10.1167/TVST.9.2.17

10. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016)
11. Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G., King, D.: Key challenges for delivering clinical impact with artificial intelligence. BMC medicine **17**(1), 1–9 (2019)
12. Kermany, D., Zhang, K., Goldbaum, M., et al.: Labeled optical coherence tomography (oct) and chest x-ray images for classification. Mendeley data **2**(2) (2018)
13. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell **172**(5), 1122–1131 (2018)
14. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in Neural Information Processing Systems **33**, 18661–18673 (2020)
15. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
16. Kingma, D.P., Welling, M.: An introduction to variational autoencoders. arXiv preprint arXiv:1906.02691 (2019)
17. Le-Khac, P.H., Healy, G., Smeaton, A.F.: Contrastive representation learning: A framework and review. IEEE Access **8**, 193907–193934 (2020)
18. Liu, Z.S., Siu, W.C., Chan, Y.L.: Photo-realistic image super-resolution via variational autoencoders. IEEE Transactions on Circuits and Systems for Video Technology **31**, 1351–1365 (4 2021). https://doi.org/10.1109/TCSVT.2020.3003832
19. Markus, A.F., Kors, J.A., Rijnbeek, P.R.: The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. Journal of Biomedical Informatics **113**, 103655 (2021)
20. Pan, H., Yang, D.I., Yuan, Z., Liang, Y.: More realistic low-resolution oct image generation approach for training deep neural networks. OSA Continuum, Vol. 3, Issue 11, pp. 3197-3205 **3**, 3197–3205 (11 2020). https://doi.org/10.1364/OSAC.408712
21. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2337–2346 (2019)
22. Pesteie, M., Abolmaesumi, P., Rohling, R.N.: Adaptive augmentation of medical data using independently conditional variational autoencoders. IEEE Transactions on Medical Imaging **38**, 2807–2820 (12 2019). https://doi.org/10.1109/TMI.2019.2914656
23. Ran, A., Cheung, C.Y.: Deep learning-based optical coherence tomography and optical coherence tomography angiography image analysis: An updated summary. Asia-Pacific journal of ophthalmology (Philadelphia, Pa.) **10**, 253–260 (5 2021). https://doi.org/10.1097/APO.0000000000000405
24. Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.R.: Explainable AI: interpreting, explaining and visualizing deep learning, vol. 11700. Springer Nature (2019)
25. Samek, W., Wiegand, T., Müller, K.R.: Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296 (2017)

26. Schoonderwoerd, T.A., Jorritsma, W., Neerincx, M.A., Van Den Bosch, K.: Human-centered xai: Developing design patterns for explanations of clinical decision support systems. International Journal of Human-Computer Studies **154**, 102684 (2021)
27. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
28. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1874–1883 (2016)
29. Singh, A., Mohammed, A.R., Zelek, J., Lakshminarayanan, V.: Interpretation of deep learning using attributions: application to ophthalmic diagnosis. https://doi.org/10.1117/12.2568631 **11511**, 39–49 (8 2020). https://doi.org/10.1117/12.2568631
30. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. Advances in neural information processing systems **29** (2016)
31. Sugawara, Y., Shiota, S., Kiya, H.: Super-resolution using convolutional neural networks without any checkerboard artifacts. In: 2018 25th IEEE International Conference on Image Processing (ICIP). pp. 66–70. IEEE (2018)
32. Uzunova, H., Ehrhardt, J., Kepp, T., Handels, H.: Interpretable explanations of black box classifiers applied on medical images by meaningful perturbations using variational autoencoders (2019). https://doi.org/10.1117/12.2511964, `https://doi.org/10.1117/12.2511964`
33. van der Velden, B.H., Kuijf, H.J., Gilhuijs, K.G., Viergever, M.A.: Explainable artificial intelligence (xai) in deep learning-based medical image analysis. Medical Image Analysis p. 102470 (2022)
34. Xiao, Y., Gao, S., Chai, Z., Zhou, K., Zhang, T., Zhao, Y., Cheng, J., Liu, J.: Open-set oct image recognition with synthetic learning. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). pp. 1788–1792. IEEE (2020)
35. Yoo, T.K., Choi, J.Y., Kim, H.K.: Feasibility study to improve deep learning in oct diagnosis of rare retinal diseases with few-shot classification. Medical and Biological Engineering and Computing **59**, 401–415 (2 2021). https://doi.org/10.1007/S11517-021-02321-1/FIGURES/12, `https://link.springer.com/article/10.1007/s11517-021-02321-1`
36. Yoon, J., Han, J., Park, J.I., Hwang, J.S., Han, J.M., Sohn, J., Park, K.H., Hwang, D.D.J.: Optical coherence tomography-based deep-learning model for detecting central serous chorioretinopathy. Scientific Reports **10**(1), 1–9 (2020)
37. Zha, X., Shi, F., Ma, Y., Zhu, W., Chen, X.: Generation of retinal oct images with diseases based on cgan. https://doi.org/10.1117/12.2510967 **10949**, 544–549 (3 2019). https://doi.org/10.1117/12.2510967