Segmentation of Macular Edema Datasets with Small Residual 3D U-Net Architectures

Jonathan Frawley*†, Chris G. Willcocks*, Maged Habib‡, Caspar Geenen‡, David H. Steel‡§ and Boguslaw Obara*†
*Department of Computer Science, Durham University, Durham, UK
†Intogral Limited, Durham, UK
‡Sunderland Eye Infirmary, Sunderland, UK
§Newcastle University, Newcastle Upon Tyne, UK

***Abstract*—This paper investigates the application of deep convolutional neural networks with prohibitively small datasets to the problem of macular edema segmentation. In particular, we investigate several different heavily regularized architectures. We find that, contrary to popular belief, neural architectures within this application setting are able to achieve close to human-level performance on unseen test images without requiring large numbers of training examples. Annotating these 3D datasets is difficult, with multiple criteria required. It takes an experienced clinician two days to annotate a single 3D image, whereas our trained model achieves similar performance in less than a second. We found that an approach which uses targeted dataset augmentation, alongside architectural simplification with an emphasis on residual design, has acceptable generalization performance - despite relying on fewer than 15 training examples.**

***Index Terms*—Machine learning, image processing and computer vision, medicine, segmentation, neural nets**

I. Introduction

The number of adults with diabetes worldwide has increased from 108 million to 422 million in the period 1980-2014 [1]. The number of affected adults worldwide is expected to rise to 592 million by 2035 [2]. About 25% of people with diabetes have some form of diabetic retinopathy [3]. This is one of the leading causes of blindness for working-aged adults in the United Kingdom [4] [5]. Diabetic macular edema is the accumulation of extracellular fluid in the retina secondary to inner retinal blood barrier breakdown associated with diabetes. It results in retinal thickening in the important central retina and causes impaired vision. It is the leading cause of decreased vision caused by diabetic retinopathy [6]. Recent research from the United Kingdom suggests that, with effective screening, the number of cases that can be caught and treated early rises significantly [4] [7]. Automated detection of diabetic retinopathy has been shown to reduce the burden on screening services [8].

Optical coherence tomography (OCT) is a non-invasive, high-resolution imaging technique that uses infrared light to provide 3D imaging of the retina [9]. OCT is capable of generating high-resolution, 3D images of the retina [10]. It is now the de facto standard tool for diagnosing multiple retinal and macular diseases, including macular edema.

Ophthalmologists currently use OCT scans to analyse the progression of macular edema both qualitatively and quantitatively. Although quantitatively the thickness of the retina can be measured relatively easily, the extent and location of intraretinal edema relative to the remaining neuro-retinal tissue is of key importance in assessing prognosis and monitoring response to treatment. This is a complex problem. Quantification of the intraretinal fluid (IRF) is something that can be done manually, but it is a slow and error prone process. Classical image techniques have failed to yield an automated solution to this problem.

Convolutional neural networks (CNN) are a deep learning based technique for solving many image-based segmentation problems. Most CNNs today are applied in areas where a lot of data is available to train on. In the case of medical images, there is often a data availability problem. Data is of a much more highly sensitive nature than is typical for many domains and has to be anonymized, which is a non-trivial process [11]. The annotation of this data with ground truth (GT) information is also a difficult, time-consuming task. It takes an experienced clinician between 30-45 minutes to annotate each slice of the OCT image due to ambiguity, shadowing, and often there being no clear edges in intensity to follow. With the typical scan consisting of up to one hundred individual slices at 30-120 microns separation, it can take two days for each image to be annotated. These challenges together limit the amount of data that can be gathered. The requirement for deep learning processes to work with small datasets is therefore of great importance in the field of medical imaging.

The U-Net CNN architecture [12] represented a step forward for the accuracy of deep learning-based biological image segmentation. It takes as input a 2D medical image and outputs a segmentation probability map. This represents a set of probabilities $p \in [0, 1]$ of each pixel being a part of the segmented region. It comprises a series of downsampling convolutions followed by a series of upsampling mixed with 2D convolutions. The key contribution of U-Net is the addition of *skip-connections* which connect the downsampling layers with their upsampling equivalent. This allows the model to capture fine details in the result, while the lower layers of the model will capture the general shape of the segmentation. The combination of the two approaches has yielded very good results in a wide range of biomedical image segmentation problems [13] [14] [15]. The 3D U-Net architecture [16] ex-

tends U-Net for use with 3D images by using 3D convolutions in place of 2D convolutions. Using 3D images allows for improved segmentation as context from multiple slices aids the decision about whether an individual voxel is an object or not.

The majority of medical imaging deep learning research has involved developing segmentations for different forms of cancerous tumors and brain disease. There has been relatively little research done on ophthalmic segmentation using deep learning [17]. This is starting to change, with recent research [18] [19], but it still lags behind other areas of medical imaging.

We have based one of our models on 3D U-Net with added residual blocks similar to He et al. [20]. We also present the result of combining the above models with a Wasserstein Generative Adversarial Network (WGAN) [21], acting as a regularization approach. Past work on liver segmentation shows improved results from combining U-Net with WGANs [22].

We propose that using the above techniques for macular edema segmentation on a small, carefully augmented dataset yields results comparable to human performance. While some prior work has looked at using deep learning on problems with small training datasets [23] [24], there have been none which the authors are aware of that specifically look at biomedical image segmentation. Existing works on macular edema segmentation using deep learning have used an order of magnitude more training images than our model [25] [26] [27].

Our contribution can be summarised as an automated approach to macular edema segmentation based on deep learning with fewer training images than any known prior work. While this paper specifically looks at the problem of macular edema segmentation, we believe that the results should be generalizable to other biomedical image segmentation problems.

II. Method

Segmentation involves labelling objects in an image, by assigning pixels with shared characteristics to corresponding class labels. In our case, we wish to assign areas of IRF in an OCT image to white pixels, and non-IRF regions to black pixels.

This means we have two classes, IRF or non-IRF, which is an example of binary image segmentation:

$$S(x,y,z) = \begin{cases} 1 & S(x,y,z) \in D \\ 0 & S(x,y,z) \notin D \end{cases} \tag{1}$$

where D is the set of voxels which correspond to disease in the original image, and x, y and z represent the coordinates of that voxel [28].

Therefore we estimate the probability of each voxel either being IRF or not, where we minimize the binary cross-entropy:

$$\begin{aligned} \mathcal{L}_{\text{BCE}} &= -\sum_{i}^{n=2} p_i \log q_i \\ &= -(p_i \log q_i + (1-p_i)\log(1-q_i)) \end{aligned} \tag{2}$$

where p_i are the target probabilities, and q_i is the output of our model.

In cases of multiple annotations per image, which we trust with equal integrity, the target probabilities p_i are averaged, although we do not have any such cases except in our test set.

A. Adaption of U-Net

We investigated and designed a number of models for comparison:

M_1: Original 3D U-Net
M_2: Small 3D U-Net
M_3: Small residual 3D U-Net
M_4: M_2 with WGAN
M_5: M_3 with WGAN

A diagram of model M_3 is shown in Fig. 1. Two residual blocks have been added to each layer. We experimented with different model depths and found that having three layers gave the best results, while still fitting in available GPU memory. Similarly, by experimentation, we found that the best input to all of our models is a $128 \times 128 \times 49$ image, the output is similarly a set of $128 \times 128 \times 49$ probabilities.

The WGAN models M_4 and M_5 adversarially train the U-Net against the discriminator network, such as to regularize the output to look like the same distribution as the annotations.

M_1 is a close replica of the original 3D U-Net [16] with batch normalization. The input to this is a $132 \times 132 \times 116$ image and the output is a set of $44 \times 44 \times 28$ probabilities. As the input to this model has more slices than our source images, this model is not very well suited to our dataset.

We optimize our network parameters using the Adam optimization algorithm, which is shown to give state-of-the-art performance in a number of settings [29]. We considered stochastic gradient descent (SGD) as it requires less memory, but found the improvements offered by Adam to outweigh the additional memory.

B. Data Augmentation in 3D

Data augmentation is the process of expanding the training dataset by adding transformations to the inputs, artificially simulating variations that may otherwise occur naturally. It is important that the generated data is representative of real world data.

At first, no data augmentation was performed on our dataset. The model performed poorly on images which were at different scales to the training data. To counteract this, we used the following transforms, all performed in 3D:

- Scaling up of images
- Cropping of images (equivalent to zooming in images)
- Elastic deformation of images

For the first two transforms, a random size is chosen to either crop or scale to. Random samples are drawn from the uniform number distribution. For the scaled up case, the randomly chosen size is limited to between 1x and 4x the original dimensions of the image. The x:y and x:z aspect ratios of the 3D image are preserved with scaling. For the cropped

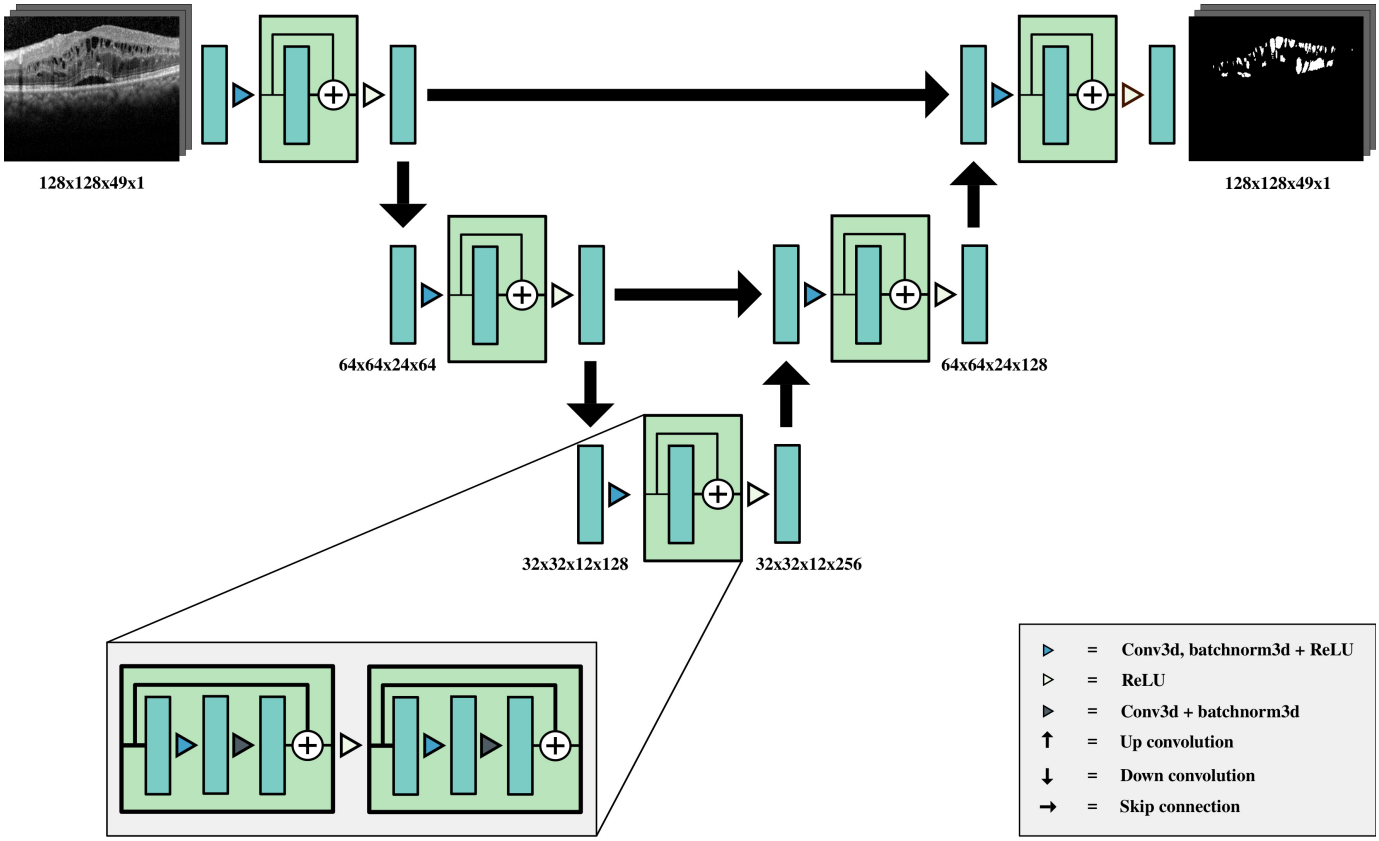


Fig. 1. Small residual 3D U-Net (M_3). This model yielded the best generalization performance in our analysis. This architecture has one fewer layer than the original 3D U-Net by Çiçek et al. [16] and the input and output sizes have been modified to better suit our data.

augmentation case, a random size is chosen between 1/4 of the size and the full size of the image, again maintaining aspect ratios. For areas outside of the boundary of the image, reflection of the original image is used. Using a combination of these forms of augmentation, the trained model is able to cope with data at a variety of resolutions and scales.

Performing elastic deformation augmentation improves generalization performance by creating new images which are still biologically realistic [16]. Parameters used for elastic deformation augmentation are $\sigma = 10$ and $points = 6$. We uniformly randomly choose between these augmentation methods and performing no augmentation when creating our augmented dataset. All of our models were trained on the result of running this augmentation on each image in our training set multiple times.

C. Training

The model described in Section II-A was implemented and trained using our training set. The images were scaled down to $128 \times 128 \times 49$ in order for the model to fit in GPU memory. In order to evaluate the model quantitatively, we scale up the image to the original size using trilinear interpolation and threshold the output probability map at 0.5 to generate a binary image. An example of this can be seen in the rightmost column of Fig. 2. For models M_2 to M_5, a learning rate of $1\mathrm{e}{-4}$ was used. For regularization, we experimented with different values of weight decay for our optimizer and found that $1\mathrm{e}{-4}$ consistently resulted in the best performance on our validation set for models M_2 to M_4. For model M_1, disabling weight decay and using a learning rate of $1\mathrm{e}{-5}$ produced the best results on our validation set. We trained and evaluated each model separately three times to assess the consistency of our results.

III. Results

A. Qualitative Results

The qualitative results of running the trained model are generally quite close to the ground truth, as seen in Fig. 2. The *3D Segmentation* column of TABLE I and TABLE II shows what the output segmentation of our model looks like in 3D across the training, validation and unseen test sets. Additionally, the *Our Prediction* and *GT* columns of TABLE I and TABLE II correspond to the output segmentation of our model and the ground truth respectively. In general, the system performs well at capturing the frequency and general shape of the edema. Fig. 3 is a heatmap showing the areas of greatest difference between our model's output and the clinician's annotation. The system tends to make most mistakes in areas around the edges of edema. This is likely due to the small size of the training set as well as the training set containing ground truths from authors with different skill levels and thresholds

Fig. 2. A single slice of one of our unseen test OCT images alongside annotations by a clinician (A_1) and non-clinician (A_2). Comparing the manual annotations to the output of our model (on the right), it can be seen that our model captures the general structure of the IRF but fails to capture finer details from the original annotations.

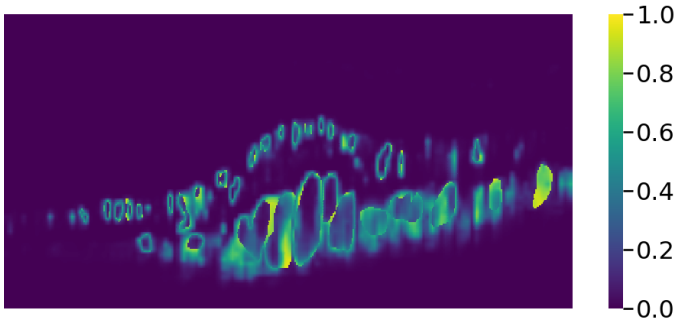


Fig. 3. Heatmap of the example in Fig. 2, visualising the delta between A_1's annotation and the unthresholded model output. Brighter colours indicate areas of greater difference between the annotation and our prediction. We can see that our model tends to make most prediction errors around the edges of edema.

for delineating small areas of IRF. The model's reduced input and output size relative to the true size of the images possibly means that it misses out on finer features in the input image.

A web-based viewer was developed to make it easier for clinicians to test and visualise the output of our model.

B. Quantitative Results

As we have only one OCT image with annotations from different authors, our comparison to human performance is limited. For this one image, the best model achieves within 4% of human performance as can be seen in TABLE III. Higher values are best for these, with M_3 achieving the best result over three runs. M_2 has the smallest standard deviation over three runs.

Performance of the best performing model using a variety of standard image segmentation metrics is shown in TABLE IV. For all metrics except absolute volume difference, higher values are best.

Fig. 4 shows how the Jaccard index improves as the models are trained, where each epoch is 10 iterations long. Data points are averaged every 200 epochs. M_3 reaches the highest peak while M_2 has the smoothest curve. M_4 and M_5 both perform well below the non-GAN models, and these also take longer to train. M_1 reaches the lowest peak performance of all models. This could be partially explained by the significantly

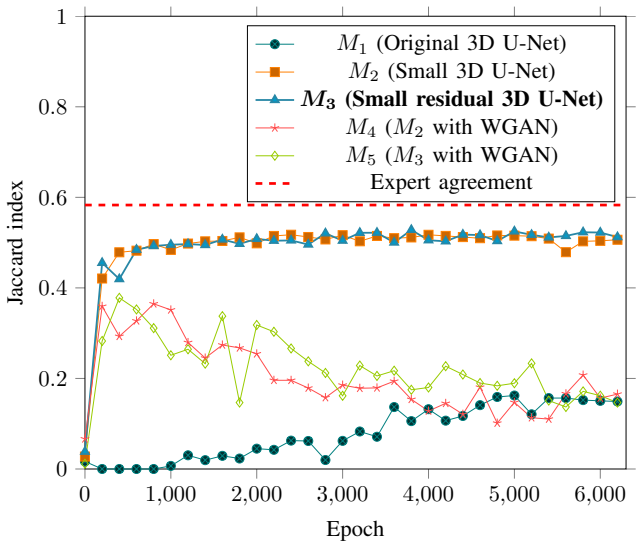


Fig. 4. Average Jaccard index over 3 runs on unseen test set as model is trained (higher is better). M_3 achieves the highest peak performance, reaching within 4% of expert agreement.

lower resolution output compared to the other models tested. The residual and non-residual models' average performance is broadly similar, but the residual model has slightly better peak performance.

Implementation

The models were implemented using PyTorch [30] and were trained on 11GB − 24GB NVIDIA Pascal and Turing architecture GPUs. Each model was trained for 6400 epochs, which was enough for models to stop substantially improving test performance as can be seen in Fig. 4.

Inference takes less than 1s per input image using the GPU-accelerated version of our model. PyTorch was used as it enables quick prototyping of deep learning models while also having good performance.

The Jaccard index was primarily used for evaluating the performance of our algorithm (where *Pred* is our prediction

TABLE I
3D SEGMENTATION VOLUMES AND 2D CROSS-SECTIONS OF TRAINING DATASET

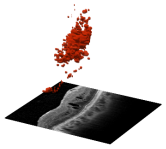
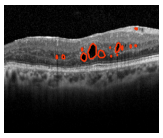
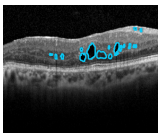
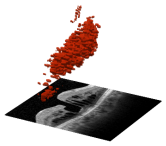
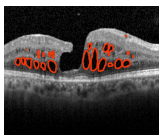
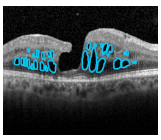
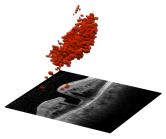
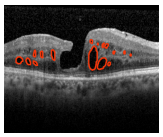
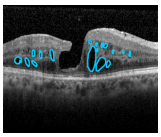
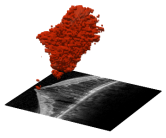
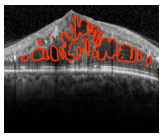
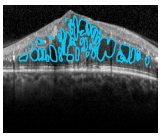
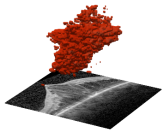
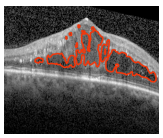
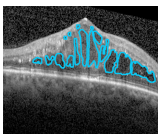
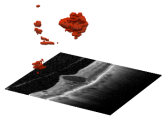
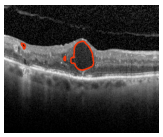
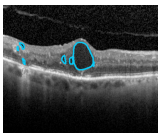
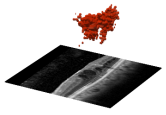
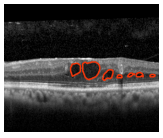
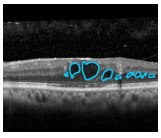
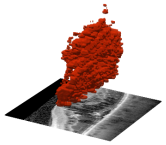
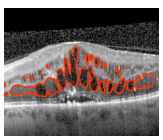
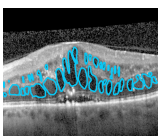
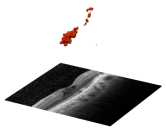
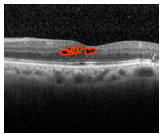
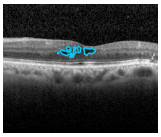
	3D Segmentation	2D Cross-Section	
		Our Prediction	GT
1			
2			
3			
4			
5			
6			
7			
8			
9			

TABLE II
3D SEGMENTATION VOLUMES AND 2D CROSS-SECTIONS OF TRAINING (10-14), VALIDATION (15 AND 16) AND TEST (17) DATASETS

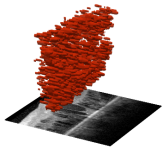
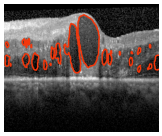
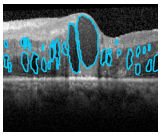
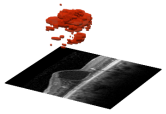
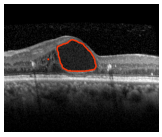
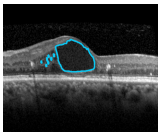
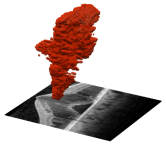
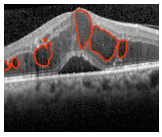
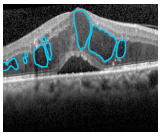
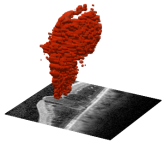
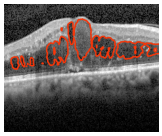
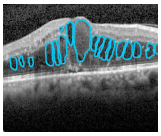
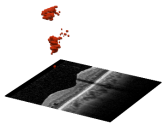
	3D Segmentation	2D Cross-Section	
		Our Prediction	GT
10			
11			
12			
13			
14			
15			
16			
17			

TABLE III
PEAK JACCARD INDEX OF TESTED MODELS AGAINST EXPERT PERFORMANCE ON THE UNSEEN TEST DATASET (MEANS AND STANDARD DEVIATIONS OVER THREE RUNS)

	Models					
Author	M_1	M_2	M_3	M_4	M_5	Expert agreement
A_1	0.206 (0.006)	0.545 (0.006)	**0.552 (0.011)**	0.437 (0.029)	0.441 (0.034)	0.583
A_2	0.225 (0.0)	0.521 (0.003)	**0.527 (0.012)**	0.452 (0.016)	0.467 (0.012)	0.583

TABLE IV
DETAILED STATISTICS OF BEST PERFORMING MODEL M_3 (MEANS AND STANDARD DEVIATIONS OVER THREE RUNS)

Author	Precision	Recall	Dice similarity coefficient	Absolute volume difference	Average precision
A_1	0.675 (0.013)	0.751 (0.007)	0.711 (0.01)	13783 (2099.839)	0.511 (0.013)
A_2	0.712 (0.006)	0.669 (0.015)	0.69 (0.01)	8731 (2099.839)	0.482 (0.014)

TABLE V
NUMBER OF ANNOTATIONS PER DATASET AND AUTHOR

	Author		
Dataset	A_1	A_2	A_3
Train	8	0	6
Validation	0	0	2
Test	1	1	0

and *GT* is the ground truth):

$$J = \frac{|Pred \cap GT|}{|Pred \cup GT|} \tag{3}$$

Also known as Intersection over Union (IoU), this is a commonly used metric for comparing the similarity of two sets. In this case, the Jaccard index represents the intersection of the model's prediction and the ground truth divided by the union of the model's prediction and the ground truth. It was used as a key metric when evaluating the performance of our prediction, thresholded at 0.5, as it is a robust indicator of how close the resultant segmentation is to the ground truth.

The open source software, Scikit-learn, was used to compute all metrics [31]. The Elasticdeform Python package was used to perform elastic deformation dataset augmentation [32].

MATERIALS

OCT images were exported from a Heidelberg SPECTRALIS HRA+OCT machine with software version 1.10.4.0. These images were cropped to remove unnecessary information and the fundus image. Annotations were created manually using a 3D image annotation tool, slice by slice in the z-dimension, by highlighting the pixels on the OCT scan which are of IRF. TABLE V shows how (image, annotation) pairs were divided into training, validation and test sets based on who authored each ground truth. Due to the many hours it takes to annotate a single image, our dataset sizes are small. Also note the imbalance of annotations for each author. A_1 is a clinician, A_2 and A_3 are non-clinician image and data experts. We use the image with multiple annotations as our test set in order to be able to compare our model against expert agreement. All images and ground truths at full size have dimensions width $= 461$, height $= 381$, slices $= 49$.

IV. CONCLUSIONS AND FUTURE WORK

It is hypothesised that our model performs so well due to its simplicity. As we are capable of learning a working solution with fewer layers than the original 3D U-Net, we think that macular edema segmentation is well suited to a small residual U-Net architecture.

It takes days to create a single annotated 3D OCT image by hand. This makes it infeasible for clinicians to manually create these annotations for every patient. Our solution can generate a similar quality result automatically in less than a second. Future clinical research will be able to assess correlation of these metrics with disease progression and treatment outcomes.

As it takes so long to create annotations, increasing the size of the dataset is difficult. We do think, however, that if more data from expert clinicians were trained on, results would continue to improve.

The model primarily makes mistakes around the edges of edema. We believe this is partially due to the fact that the images are scaled down prior to being input to the model. It would be useful for the model to work on the full resolution image. This could be done using techniques similar to those used to create super-resolution images as described by Dong et al. [33], or by using GPUs with larger amounts of memory available.

The imaging device used also provides fundus image output. An interesting extension of this project would be to use the fundus output along with the OCT image and see if that improves the prediction. The fundus image could help with prediction by utilising features for maculopathy such as exudates and red lesions (microaneurysms). The role of vessel width and geometry analysis in maculopathy prediction

could be added to the OCT biomarkers to improve prediction accuracy for disease progression and response to treatment. Recent work on deep learning models for diagnosing age-related macular degeneration (AMD) has shown that combining OCT and fundus image output can yield improved results [34].

The use of another medical imaging technique known as OCT angiography (OCT-A) has shown promise in helping to diagnose diabetic retinopathy and macular edema [35] [36]. Recent work has suggested combining OCT-A, OCT and fundus images to improve the accuracy of models to diagnose AMD [37]. Applying such an approach to the automated diagnosis of macular edema may help to improve prediction accuracy.

We have trained and tested images from a single device type (Heidelberg SPECTRALIS HRA+OCT). In order to create a real-world diagnostic solution, it would be required to train and test our model on images from a variety of OCT device manufacturers. Recent work by De Fauw et al. [18] has shown that it is possible to train a model on a single OCT device and refine it to work for another OCT device.

Conflict of Interest & Attribution

In accordance with his ethical obligation as a researcher, Jonathan Frawley reports that he receives funding for his PhD from Intogral Ltd. Some of the work described was developed as part of his work as an employee at Intogral Ltd. Data and annotations by the clinician for this project were kindly provided by Maged Habib, Caspar Geenen and David H. Steel of the Sunderland Eye Infirmary, South Tyneside and Sunderland NHS Foundation Trust, UK. Intogral Ltd also provided annotations created by non-clinicians.

References

[1] B. Zhou, Y. Lu, K. Hajifathalian, J. Bentham, M. Di Cesare, G. Danaei, H. Bixby, M. J. Cowan, M. K. Ali, C. Taddei *et al.*, "Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4· 4 million participants," *The Lancet*, vol. 387, no. 10027, pp. 1513–1530, 2016.

[2] L. Guariguata, D. R. Whiting, I. Hambleton, J. Beagley, U. Linnenkamp, and J. E. Shaw, "Global estimates of diabetes prevalence for 2013 and projections for 2035," *Diabetes Research and Clinical Practice*, vol. 103, no. 2, pp. 137–149, 2014.

[3] G. Virgili, F. Menchini, G. Casazza, R. Hogg, R. R. Das, X. Wang, and M. Michelessi, "Optical coherence tomography (OCT) for detection of macular oedema in patients with diabetic retinopathy," *Cochrane Database of Systematic Reviews*, no. 1, 2015.

[4] C. Arun, A. Al-Bermani, K. Stannard, and R. Taylor, "Long-term impact of retinal screening on significant diabetes-related visual impairment in the working age population," *Diabetic Medicine*, vol. 26, no. 5, pp. 489–492, 2009.

[5] C. Bunce and R. Wormald, "Leading causes of certification for blindness and partial sight in England & Wales," *BMC Public Health*, vol. 6, p. 58, 2006.

[6] F. L. Ferris III and A. Patz, "Macular edema. a complication of diabetic retinopathy," *Survey of Ophthalmology*, vol. 28, pp. 452–461, 1984.

[7] G. Liew, M. Michaelides, and C. Bunce, "A comparison of the causes of blindness certifications in england and wales in working age adults (16–64 years), 1999–2000 with 2009–2010," *BMJ Open*, vol. 4, no. 2, p. e004015, 2014.

[8] D. Usher, M. Dumskyj, M. Himaga, T. H. Williamson, S. Nussey, and J. Boyce, "Automated detection of diabetic retinopathy in digital retinal images: a tool for diabetic retinopathy screening," *Diabetic Medicine*, vol. 21, no. 1, pp. 84–90, 2004.

[9] G. Trichonas and P. K. Kaiser, "Optical coherence tomography imaging of macular oedema," *British Journal of Ophthalmology*, vol. 98, pp. 24–29, 2014.

[10] M. R. Hee, C. A. Puliafito, C. Wong, J. S. Duker, E. Reichel, J. S. Schuman, E. A. Swanson, and J. G. Fujimoto, "Optical coherence tomography of macular holes," *Ophthalmology*, vol. 102, no. 5, pp. 748–756, 1995.

[11] K. El Emam, S. Rodgers, and B. Malin, "Anonymising and sharing individual patient data," *The BMJ*, vol. 350, p. 1139, 2015.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, vol. 9351, 2015, pp. 234–241.

[13] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: learning where to look for the pancreas," in *Medical Imaging with Deep Learning*, 2018.

[14] T. Falk, D. Mai, R. Bensch, Ö. Çiçek, A. Abdulkadir, Y. Marrakchi, A. Böhm, J. Deubner, Z. Jäckel, K. Seiwald *et al.*, "U-Net: deep learning for cell counting, detection, and morphometry," *Nature Methods*, vol. 16, no. 1, pp. 67–70, 2019.

[15] H. Dong, G. Yang, F. Liu, Y. Mo, and Y. Guo, "Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks," in *Medical Image Understanding and Analysis*, 2017, pp. 506–517.

[16] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2016, pp. 424–432.

[17] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[18] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature Medicine*, vol. 24, no. 9, p. 1342, 2018.

[19] T. Schlegl, S. M. Waldstein, H. Bogunovic, F. Endstraßer, A. Sadeghipour, A.-M. Philip, D. Podkowinski, B. S. Gerendas, G. Langs, and U. Schmidt-Erfurth, "Fully automated detection and quantification of macular fluid in OCT using deep learning," *Ophthalmology*, vol. 125, no. 4, 2017.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European Conference On Computer Vision*, 2016, pp. 630–645.

[21] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," *arXiv:1701.07875 [cs, stat]*, 2017.

[22] Y. Enokiya, Y. Iwamoto, Y.-W. Chen, and X.-H. Han, "Automatic liver segmentation using u-net with wasserstein gans," *Journal of Image and Graphics*, vol. 6, no. 2, 2018.

[23] A. Milan, T. Pham, K. Vijay, D. Morrison, A. W. Tow, L. Liu, J. Erskine, R. Grinover, A. Gurman, T. Hunn *et al.*, "Semantic segmentation from limited training data," in *IEEE International Conference on Robotics and Automation*, 2018, pp. 1908–1915.

[24] A. Davari, E. Aptoula, B. Yanikoglu, A. Maier, and C. Riess, "Gmm-based synthetic samples for classification of hyperspectral images with limited training data," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 6, pp. 942–946, 2018.

[25] M.-X. Li, S.-Q. Yu, W. Zhang, H. Zhou, X. Xu, T.-W. Qian, and Y.-J. Wan, "Segmentation of retinal fluid based on deep learning: application of three-dimensional fully convolutional neural networks in optical coherence tomography images," *International Journal of Ophthalmology*, vol. 12, no. 6, p. 1012, 2019.

[26] C. S. Lee, A. J. Tyring, N. P. Deruyter, Y. Wu, A. Rokem, and A. Y. Lee, "Deep-learning based, automated segmentation of macular edema in optical coherence tomography," *Biomedical Optics Express*, vol. 8, no. 7, pp. 3440–3448, 2017.

[27] D. Lu, M. Heisler, S. Lee, G. W. Ding, E. Navajas, M. V. Sarunic, and M. F. Beg, "Deep-learning based multiclass retinal fluid segmentation and detection in optical coherence tomography images using a fully convolutional neural network," *Medical Image Analysis*, vol. 54, pp. 100–110, 2019.

[28] J. Chen, H. Shao, and C. Hu, "Image segmentation based on mathematical morphological operator," in *Colorimetry and Image Processing*, 2017.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, Y. Bengio and Y. LeCun, Eds., 2015.

[30] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: an imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, 2019, pp. 8024–8035.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[32] "WWW: Web page of the elasticdeform project," https://pypi.org/project/elasticdeform/, [Online; accessed 11-March-2020].

[33] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[34] T. K. Yoo, J. Y. Choi, J. G. Seo, B. Ramasubramanian, S. Selvaperumal, and D. W. Kim, "The possibility of the combination of oct and fundus images for improving the diagnostic accuracy of deep learning for age-related macular degeneration: a preliminary experiment," *Medical & Biological Engineering & Computing*, vol. 57, no. 3, pp. 677–687, 2019.

[35] A. Y. Kim, Z. Chu, A. Shahidzadeh, R. K. Wang, C. A. Puliafito, and A. H. Kashani, "Quantifying microvascular density and morphology in diabetic retinopathy using spectral-domain optical coherence tomography angiography," *Investigative Ophthalmology & Visual Science*, vol. 57, no. 9, pp. 362–370, 2016.

[36] L. Mao, S.-s. Weng, Y.-y. Gong, and S.-q. Yu, "Optical coherence tomography angiography of macular telangiectasia type 1: Comparison with mild diabetic macular edema," *Lasers in Surgery and Medicine*, vol. 49, no. 3, pp. 225–232, 2017.

[37] E. Vaghefi, S. Hill, H. M. Kersten, and D. Squirrell, "Multimodal retinal image analysis via deep learning for the diagnosis of intermediate dry age-related macular degeneration: a feasibility study," *Journal of Ophthalmology*, pp. 1–7, 2020.