

# Robust 3D U-Net Segmentation of Macular Holes

Jonathan Frawley<sup>1,2</sup>[0000-0002-9437-7399], Chris  
G. Willcocks<sup>1</sup>[0000-0001-6821-3924], Maged Habib<sup>3,4</sup>[0000-0003-0931-3786], Caspar  
Geenen<sup>3</sup>[0000-0002-2778-6344], David H. Steel<sup>3,4</sup>[0000-0001-8734-3089], and  
Boguslaw Obara<sup>2,4,5</sup>[0000-0003-4084-7778]

<sup>1</sup> Department of Computer Science, Durham University, Durham, UK

<sup>2</sup> Gliff.ai, Durham, UK

<sup>3</sup> Sunderland Eye Infirmary, Sunderland, UK

<sup>4</sup> Bioscience Institute, Newcastle University, Newcastle Upon Tyne, UK

<sup>5</sup> School of Computing, Newcastle University, Newcastle Upon Tyne, UK

**Abstract.** Macular holes are a common eye condition which result in visual impairment. We look at the application of deep convolutional neural networks to the problem of macular hole segmentation. We use the 3D U-Net architecture as a basis and experiment with a number of design variants. Manually annotating and measuring macular holes is time consuming and error prone, taking dozens of minutes to annotate a single 3D scan. Previous automated approaches to macular hole segmentation take minutes to segment a single 3D scan. We found that, in less than one second, deep learning models generate significantly more accurate segmentations than previous automated approaches (Jaccard index boost of 0.08 – 0.09) and expert agreement (Jaccard index boost of 0.13 – 0.20). We also demonstrate that an approach of architectural simplification, by greatly simplifying the network capacity and depth, results in a model which is competitive with state-of-the-art models such as residual 3D U-Nets.

**Keywords:** Machine learning · image processing and computer vision · medicine · segmentation · neural nets · retina · macular holes.

## 1 Introduction

Idiopathic full thickness macular holes (iFTMH) are a common, and visually disabling condition, being bilateral in 10% of affected individuals. They occur at a prevalence of approximately 1 in 200 of the over 60-year-old population with an incidence of approximately 4000 per annum in the United Kingdom (UK)[1,13]. If left untreated they result in visual acuity below the definition of blindness and typically greater than 1.0 logMAR (logarithm of the minimum angle of resolution), where 0.1 logMAR is classed as normal.

3D high-resolution images of the retina can be created using optical coherence tomography (OCT) [9]. It is now the standard tool for diagnosing macular

holes [7]. Compared to previous imaging methods, OCT can more easily assist a clinician in differentiating a full-thickness macular hole from mimicking pathology, which is important in defining appropriate treatment [9]. An OCT scan of a macular hole is a 3D volume. Clinicians, however, typically view OCT scans as a series of 2D images, choose the central slice with maximum dimensions and perform measurements which are predictors of anatomical and visual success such as base diameter, macular hole inner opening and minimum linear diameter [12,2,15,19]. This approach is limited as it assumes that the macular hole base is circular, and would give incorrect results when it is elliptical [16], which is typically the case [2]. With the advent of automated 3D approaches, it is possible to begin to look at measurements in 3D and how they might be predictors of anatomical and visual success.

Neural networks are an interconnected group of artificial neurons, which can be reconfigured to solve a problem based on data. Convolutional neural networks (CNN) are a type of neural network inspired by how the brain processes visual information [11]. CNNs have been very successful in computer vision problems, such as automating the segmentation of medical images. For a CNN to learn to segment images in a supervised manner, it needs to have access to images with associated ground truth (GT) information which highlight the areas of the image for the task at hand. This is often done manually which is time consuming and requires expert knowledge.

The U-Net CNN architecture [18] is a highly utilized CNN architecture for biomedical image segmentation for use on 2D images. It has had success in segmentation to help diagnose other eye conditions such as macular edema, even when dataset sizes are limited [5]. We sought to examine the application of variants of the U-Net architecture to the problem of macular hole segmentation. Our proposed model is a smaller version of the model from the original 3D U-Net paper [3]. We also implemented and evaluated the proposed model with residual blocks added, similar to those described by He et al. [8]. In addition, we implemented a much more complex residual model, DeepMind’s OCT segmentation model [4], and ran the same tests with it.

Alternatives to U-Net have been created such as V-Net [14] which uses 3D convolutions and a Dice score-based loss. We use binary cross-entropy as our loss function, similar to the weighted cross entropy used in the original 3D U-Net paper. Early experiments showed that binary cross-entropy outperformed a Dice score-based loss for our problem. Additionally, a study that did a comparison between multiple model architectures on another biomedical image segmentation problem showed that V-Net-based models did not outperform U-Net-based models [6]. For these reasons, we chose 3D U-Net as the basis of our model rather than V-Net.

Our contribution can be summarized as developing an automated approach to macular hole segmentation based on deep learning which yields significantly improved results compared to prior methods. We present a comparison of the above-mentioned models against the current state-of-the-art automated approach [16]. The state-of-the-art method is a level set approach which does not use deep learn-

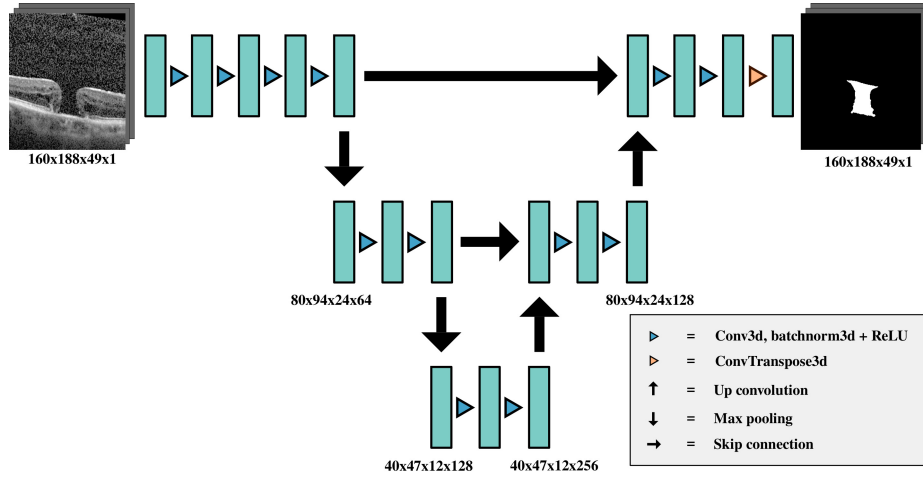


Fig. 1: Small 3D U-Net ( $M_1$ ). The proposed model is a cut-down version of 3D U-Net [3]. It has fewer levels and a carefully optimized capacity for our datasets.

ing. We show that simple low-capacity 3D U-Nets are capable of outperforming the state-of-the-art automated approach and that increasing the complexity of the architecture does not improve performance. The PyTorch-based code for this work has been released as an open-source project <sup>6</sup>.

## 2 Materials and Methods

### 2.1 Materials

All had undergone Spectral domain optical coherence tomography (SDOCT) imaging using the Heidelberg Spectralis (Heidelberg, Germany) as part of routine care, using the same imaging protocol. A high density central horizontal scanning protocol with 29-30 micron line spacing was used in the central 15 by 5 degrees. The individual OCT line scans were  $768 \times 496$  pixels with the scaling varying slightly between datasets but typically equating to 5.47 microns per pixel in the  $x$  (horizontal) axis and 3.87 microns per pixel in the  $y$  (vertical) axis. With 29-30 microns spacing between scans ( $z$  axis), there were 49 scans per dataset. All scans used a 16 automatic real time setting enabling multisampling and noise reduction over 16 images. All scans collected were from unique patients and were stored using the uncompressed TIFF file format.

All images were cropped to the same size and unnecessary information such as the fundus image were removed. Annotations were created by a mixture of clinicians and image experts using a 3D image annotation tool. Pixels on each slice of the OCT scan which represented macular hole were highlighted. There

<sup>6</sup> <https://github.com/gliff-ai/robust-3d-unet-macular-holes>

were 85 (image, annotation) pairs in the training dataset, 56 after combining annotations from multiple authors. There were 22 pairs in the validation dataset and 9 in the unseen test set.

Originally we had three annotations for each OCT image in the unseen test set. However, due to inconsistencies between authors, we combined all ground truths into a single ground truth per image. To do this, we used a voting system, where if  $\frac{2}{3}$  of the authors had annotated a voxel, that voxel was annotated in the resultant ground truth. All images and ground truths at full size had dimensions  $321 \times 376 \times 49$ . We did not augment our dataset as we found that augmentations did not improve the generalizability of our model. As we believe that our test and validation sets are large enough to be representative of the real-world problem, this was not deemed to be an issue.

## 2.2 Methods

Image segmentation involves the labelling of objects of interest in an image. For a 3D image, this is done by assigning voxels with shared characteristics to corresponding class labels. We wished to assign areas of the macular hole volume in an OCT image to white voxels and all other regions to black voxels.

We used binary cross-entropy as our loss function, which tells us how close our predicted macular hole regions are to those in the ground truth:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^N p_i \log q_i + (1 - p_i) \log(1 - q_i), \quad (1)$$

$N$  being the batch size,  $p_i$  being the ground truth and  $q_i$  being the output of our model. For images with multiple annotations in our training set, we trusted them with equal integrity and the target probabilities were averaged. The validation set had no samples with multiple annotations. As described in Section 2.1, for the unseen test set, we used a voting process to decide on the final target ground truth.

U-Net takes as input a 2D image and outputs a set of probabilities. Each entry in the output is the probability of each part of the image being a part of the segmented region. It is a U-shaped CNN architecture, consisting of a contracting path and an expansive path. The contracting path consists of 2D convolutions, ReLU activations and 2D max pooling at each level. The expansive path’s levels use *skip connections* to their contracting path equivalent, along with 2D convolutions, ReLU activations and 2D up convolutions. Skip connections allow for high-resolution information to be captured by the model while the contracting/expansive paths capture the abstract shape of the segmentation. The 3D U-Net architecture [3] is a version of U-Net designed for use with 3D images which uses 3D convolutions, up convolutions and max pooling layers. This allows for improved segmentation of 3D images as the context from multiple slices are used to decide whether an individual voxel is an object or not.

A number of models based on the 3D U-Net architecture were compared:

- $M_1$ : Small 3D U-Net (Proposal) [5,216,353 parameters],
- $M_2$ : Small residual 3D U-Net (Residual) [13,928,833 parameters],
- $M_3$ : Residual 3D U-Net for 2D slices (DeepMind) [4] [470,333,089 parameters].

A diagram of model  $M_1$  is shown in Fig. 1. Early on, versions of 2D U-Net were implemented as described in the original paper, however, performance was very poor for our dataset. The original 2D U-Net model has an input size of  $572 \times 572$  and an output size of  $388 \times 388$ . The poor performance we noticed is likely due to a lack of context from multiple slices. In addition, the input to the original 2D U-Net model is of a higher resolution than our image slices, which have a resolution of  $321 \times 376$ . Therefore, we needed to upscale our images which resulted in distortion and wasted memory usage. Similarly, we also implemented 3D U-Net as described in the original paper, however, this also performed poorly. Again, this is primarily due to the input and output sizes of the model being too dissimilar to our dataset's. The original 3D U-Net model has an input size of  $132 \times 132 \times 116$ . Our images only have 49 slices, which needed to be upsampled to 116 slices for this. This resulted in significant distortion of the input. The output of the original 3D U-Net model is also of a very low resolution:  $44 \times 44 \times 28$  which would have been very coarse when upsampled to our image dimensions of  $321 \times 376 \times 49$ . As our images were of resolution  $321 \times 376 \times 49$ , we aimed to keep the resolution of the input and output as close to this as possible. Different to the original U-Net and 3D U-Net papers, it was decided to keep the input and output dimensions equal to each other, to maximise the resolution of our output. We tweaked convolution sizes, padding and strides until we achieved this goal, while still fitting in available GPU memory.

Our experiments showed that using three levels for this model resulted in the best performance, rather than the four levels that the original 3D U-Net paper used. A scaled-down input image of  $160 \times 188 \times 49$  yielded the best results for models  $M_1$  and  $M_2$ . The output is of the same dimensions as the input.  $M_2$  is similar to  $M_1$  except that residual blocks have been added to each level.  $M_3$  is a very deep residual 3D U-Net architecture which takes nine slices of the OCT image as input and outputs a 2D probability map as output, representing the segmentation of a single slice of the OCT image. For  $M_3$ , the slice which we want to segment, along with 4 slices on either side is input to the model, which is a  $321 \times 376 \times 9$  image. This is based on a model architecture developed by DeepMind for segmenting OCT images [4]. For slices near the boundaries, we use mirroring to handle slices that are outside of the image. It outputs a set of  $321 \times 376$  probabilities, corresponding to one slice of the 3D OCT.  $M_3$ , therefore, requires 49 iterations to segment a whole 3D OCT image in our dataset. Model  $M_3$  has the most parameters of the models tested, with  $M_1$  having the fewest parameters.

The Jaccard index was used as the primary metric for measuring the performance of each method. This is one of the standard measures of the performance of image segmentation methods, especially in medical image segmentation [20]. The Dice similarity coefficient (DSC) is another commonly used metric and is closely related to the Jaccard index, with one being computable from the other.

For completeness and ease of comparison with other results, we also provide the DSC for our proposed model in Section 4.2.

### 3 Implementation

Our experiments were all conducted using the Python programming language and the PyTorch [17] deep learning framework on NVIDIA Turing GPUs with 24GB of memory. PyTorch is a state-of-the-art framework for building deep learning models which is highly optimized for modern GPU hardware. We trained each model for 500 epochs where each epoch ran over 10 3D images, which was enough for all models to stop substantially improving. This means that the models which output a 3D segmentation ( $M_1$  and  $M_2$ ) had 10 iterations per epoch, and the slice-based model ( $M_3$ ) had 490 iterations per epoch. As source code was not released for DeepMind’s model,  $M_3$  was implemented as closely as possible to the description provided in the original paper and slightly adapted to fit the binary classification problem.

In order to evaluate models  $M_1$  and  $M_2$ , we scaled up the output probability map to its original size using trilinear interpolation and thresholded it at 0.5 to generate a binary segmentation. For model  $M_3$ , we individually ran over all 49 slices of an image and recombined the 49 2D probability maps into a single 3D probability map. We then thresholded this combined map at 0.5 to generate a 3D binary segmentation. The Adam optimization algorithm [10] was used to optimize parameters of the models, with hyperparameters being found by experimentation. The *BCEWithLogitsLoss* function in PyTorch was used for loss calculation, which combines a sigmoid activation and binary cross entropy loss into one function. A similar number of experiments were conducted for each model. For model  $M_1$ , a learning rate of  $1e-4$  and weight decay of  $1e-6$  was used. For model  $M_2$ , a learning rate of  $1e-4$  and weight decay of  $1e-5$  was used. For model  $M_3$ , a learning rate of  $7.5e-5$  was used and weight decay was disabled. The 3D OCT images were normalized to the  $[0, 1]$  range prior to scaling or slicing.

Each model was trained and evaluated separately three times to assess the consistency of our results. We then calculated the Jaccard index, comparing each of the models’ predictions with the ground truth. Due to the fact that we only had a small number of images with multiple authors, we decided to keep the training, validation and unseen test sets static for all tests rather than using k-fold cross-validation. We reserved all images which had three annotations for the unseen test set, in order for us to be able to compare our results with expert agreement, which was a key goal of the research.

### 4 Results

In this section, we look at evaluating our models both qualitatively and quantitatively. For qualitative results, we primarily present results in 2D for ease of comparison with other methods. We also present a sample of segmented macular

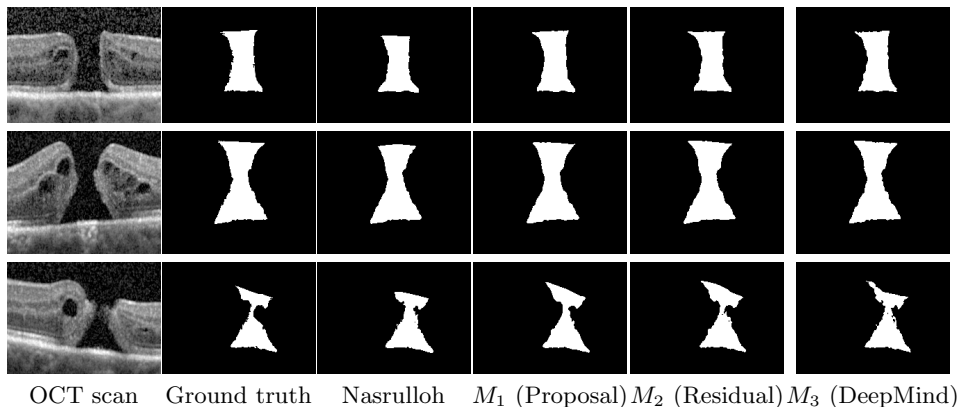


Fig. 2: Qualitative output on the unseen test set of our trained macular hole model ( $M_1$ ) compared with the ground truth, the state-of-the-art automated approach (Nasrulloh) [16], the residual model ( $M_2$ ) and DeepMind’s model ( $M_3$ ). For clarity, we zoomed in on the predicted regions.

hole volumes in 3D to demonstrate that our method captures the 3D shape of the volume. For quantitative results, we present an image-by-image comparison of each model’s performance using the Jaccard index against the state-of-the-art method. We then present a variety of other image segmentation metrics on the proposed model for ease of comparison with other methods.

#### 4.1 Qualitative Results

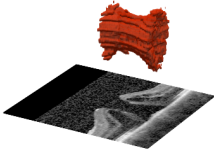
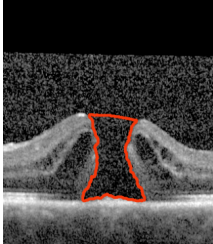
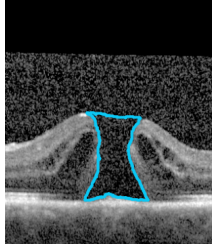
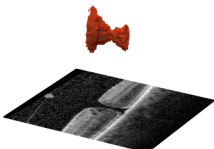
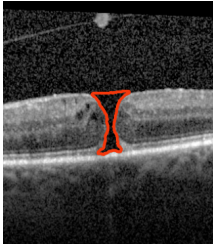
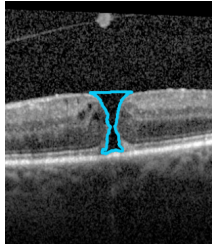
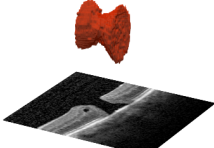
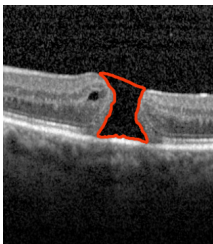
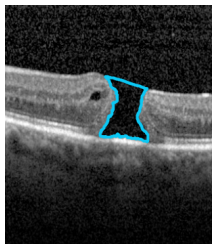
The qualitative results of running inference on the trained macular hole models are generally quite close to the ground truth, as seen in Fig. 2. In general, predictions from all of the models are closer to the ground truth than the state-of-the-art automated approach. We can see that the qualitative difference between the models tested is not hugely significant. This is surprising as  $M_3$  has significantly more capacity than  $M_1$ . This shows that adding more capacity to a model of a particular architecture does not necessarily yield an improvement in qualitative output.

3D visualizations of the output of our proposed model can be seen in Table 1. We can see how the 3D shape of the macular holes is preserved, and matches figures from similar works [16]. This type of view would allow the clinician to view the macular hole from every angle, rather than the 2D views which are currently widely used.

#### 4.2 Quantitative Results

Fig. 3 shows how the average Jaccard index on the unseen test set improved as  $M_1$  was trained and we can see that after 200 epochs it had surpassed the performance of the state-of-the-art automated approach and expert agreement.

Table 1: 3D and 2D segmentation output of model  $M_1$  (Proposal) on the unseen test along with ground truth.

3D segmentation	2D segmentation	2D ground truth
		
		
		

All of the trained macular hole models perform very well compared to the state-of-the-art automated approach [16] as we see in Table 2. Despite model  $M_1$  having by far the fewest parameters, it achieves performance which is similar to the highest-capacity model, and in some cases surpasses it. Further results in Table 3 show that  $M_1$  performs consistently well under other standard segmentation quality measures.

## 5 Discussion

The results show that previous automated approaches to this problem cannot compete with deep learning methods. All of the models tested performed significantly better than the level set method.

If we examine the model results in isolation, we can see that the results can be divided into two categories: the high-capacity 3D U-Net model ( $M_3$ )



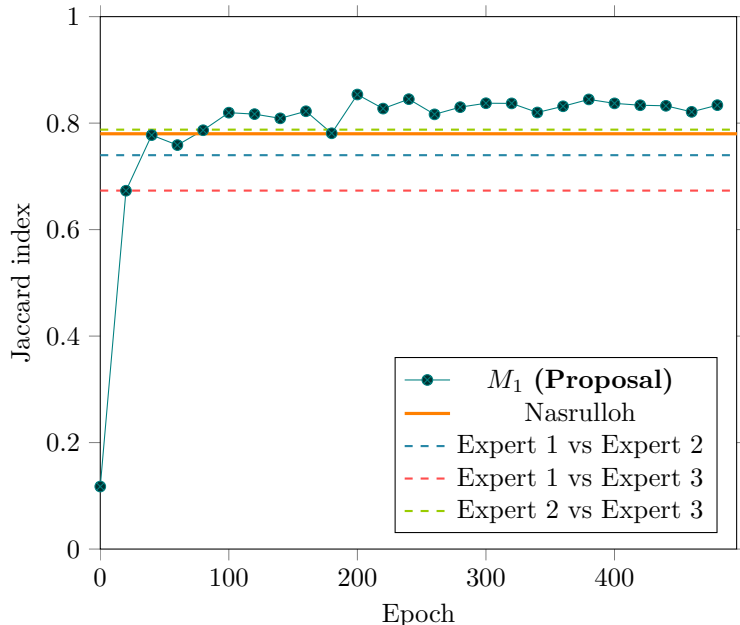


Fig. 3: Average Jaccard index of our proposed model ( $M_1$ ) over 3 runs on the unseen test set as the model was trained. We see that the model achieves significantly better results than the state-of-the-art automated approach (Nasrulloh) and expert agreement.  $M_1$  exceeded expert agreement by a Jaccard index of 0.13 – 0.20.

and the lower-capacity 3D U-Nets ( $M_1$  and  $M_2$ ). The low-capacity 3D U-Nets achieve the best results on the unseen test set. The high-capacity model, which has many times the number of parameters of the  $M_1$  model, does not have better generalizability. This is even more surprising given that the high-capacity model takes the full-resolution image as input and also outputs a full-resolution segmentation. Given that the low-capacity 3D U-Nets use a downsized 3D image as input and output, we would expect them to perform worse due to not having the same amount of information available. The fact that this does not occur implies that the chosen models do not need very high-resolution input and output to make an accurate segmentation of macular holes in OCT images.

It is a counterintuitive finding that we do not see an improvement in performance for a model which takes a full-resolution image and which has a significantly higher capacity. In a similar problem, a high-profile study used this high-capacity model for their segmentation [4]. Since that work did not present results from different architectures as we have done, it is difficult to know whether our results would be replicated there. Our work clearly shows that for some biomedical segmentation problems, it is important to consider lower-capacity models in addition to higher-capacity models.

Table 2: Jaccard index comparison between the state-of-the-art automated approach (Nasrulloh) and tested models on the unseen test set (mean and standard deviation over three runs except for state-of-the-art which is deterministic).

Image	Nasrulloh	$M_1$ (Proposal)	$M_2$ (Residual)	$M_3$ (DeepMind)
Image 1	0.714	$0.865 \pm 0.009$	$0.868 \pm 0.002$	$0.832 \pm 0.006$
Image 2	0.743	$0.891 \pm 0.02$	$0.887 \pm 0.014$	$0.893 \pm 0.012$
Image 3	0.772	$0.887 \pm 0.004$	$0.885 \pm 0.002$	$0.872 \pm 0.006$
Image 4	0.811	$0.895 \pm 0.012$	$0.884 \pm 0.001$	$0.894 \pm 0.006$
Image 5	0.787	$0.894 \pm 0.005$	$0.901 \pm 0.003$	$0.875 \pm 0.014$
Image 6	0.678	$0.804 \pm 0.008$	$0.815 \pm 0.007$	$0.765 \pm 0.006$
Image 7	0.845	$0.907 \pm 0.002$	$0.905 \pm 0.004$	$0.893 \pm 0.009$
Image 8	0.874	$0.874 \pm 0.012$	$0.862 \pm 0.002$	$0.893 \pm 0.006$
Image 9	0.787	$0.869 \pm 0.019$	$0.853 \pm 0.008$	$0.835 \pm 0.007$
Mean	0.779	$0.876 \pm 0.012$	$0.874 \pm 0.006$	$0.861 \pm 0.008$

Table 3: Other metrics for model  $M_1$  (Proposal) on the unseen test set (mean and standard deviation over three runs, DSC refers to the Dice similarity coefficient, AVD refers to absolute volume difference and AP refers to average precision).

Image	Precision	Recall	DSC	AVD	AP
Image 1	$0.93 \pm 0.009$	$0.926 \pm 0.012$	$0.928 \pm 0.005$	$1352 \pm 908.357$	$0.862 \pm 0.01$
Image 2	$0.954 \pm 0.008$	$0.931 \pm 0.014$	$0.942 \pm 0.011$	$1379 \pm 369.396$	$0.889 \pm 0.021$
Image 3	$0.949 \pm 0.003$	$0.931 \pm 0.003$	$0.94 \pm 0.002$	$2308 \pm 517.533$	$0.885 \pm 0.004$
Image 4	$0.974 \pm 0.005$	$0.917 \pm 0.016$	$0.945 \pm 0.007$	$5293 \pm 1817.849$	$0.895 \pm 0.011$
Image 5	$0.915 \pm 0.012$	$0.974 \pm 0.007$	$0.944 \pm 0.003$	$2320 \pm 763.08$	$0.892 \pm 0.005$
Image 6	$0.848 \pm 0.003$	$0.94 \pm 0.007$	$0.891 \pm 0.005$	$1911 \pm 80.168$	$0.797 \pm 0.009$
Image 7	$0.965 \pm 0.002$	$0.938 \pm 0.001$	$0.951 \pm 0.001$	$1564 \pm 157.11$	$0.906 \pm 0.002$
Image 8	$0.898 \pm 0.006$	$0.971 \pm 0.008$	$0.933 \pm 0.007$	$4544 \pm 155.656$	$0.872 \pm 0.013$
Image 9	$0.917 \pm 0.016$	$0.943 \pm 0.012$	$0.93 \pm 0.011$	$1186 \pm 906.832$	$0.865 \pm 0.02$

Our work concentrates on looking at OCT images from a particular type of device, from a single manufacturer. For future work, other models of OCT device should be tested and compared with our results. It has been shown that models trained on one device can be relatively easily trained to work with other devices [4]. Our data is from a particular population centre, namely North East England. For future work, it would be interesting to see if our results are replicated in other population centres, both nationally and internationally. As we have made our code available as an open-source project, it is hoped that this can be achieved.

## 6 Conclusions

All of the models tested exceeded the performance of the state-of-the-art automated approach which is a level set method. It is clear that deep learning methods allow for the generation of segmentations that are closer to what humans provide. Despite  $M_3$  having 90 times the parameters of  $M_1$ ,  $M_1$  gives excellent qualitative and quantitative results which are of a similar quality to  $M_3$ .  $M_1$ 's performance exceeded expert agreement by a Jaccard index of 0.13 – 0.20. As  $M_1$  is the smallest model, it requires the least amount of resources to run.  $M_1$  is also a quick model to run, requiring only one pass through the whole 3D image, whereas  $M_3$  requires one pass per slice. Once trained,  $M_1$  is capable of segmenting an OCT image in less than one second. In contrast, the state-of-the-art automated method requires minutes to run [16]. For these reasons,  $M_1$  is the best candidate to form the basis of future studies in a clinical setting. These findings show that careful tuning and in some cases architectural simplification can, for some simple task distributions, be as effective as very deep residual designs.

The code is provided as an open-source project in order for future researchers to replicate our results and build upon this research. Training and testing on different populations with different demographics will be crucial to determine that our trained models do not exhibit any bias. The lack of large-scale open data sets from different population centres for OCT imagery makes this a significant challenge that needs to be overcome.

## 7 Conflicts of Interest

In accordance with his ethical obligation as a researcher, Jonathan Frawley reports that he received funding for his PhD from Gliff.ai. Some of the work described was developed as part of his work as an employee at Gliff.ai. Gliff.ai also provided annotations created by non-clinicians. 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. All images were collected as part of routine care and anonymised.

## References

1. Ali, F.S., Stein, J.D., Blachley, T.S., Ackley, S., Stewart, J.M.: Incidence of and risk factors for developing idiopathic macular hole among a diverse group of patients throughout the United States. *JAMA Ophthalmology* **135**(4), 299–305 (2017)
2. Chen, Y., Nasrulloh, A.V., Wilson, I., Geenen, C., Habib, M., Obara, B., Steel, D.H.: Macular hole morphology and measurement using an automated three-dimensional image segmentation algorithm. *BMJ Open Ophthalmology* **5**(1), e000404 (2020)
3. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 424–432 (2016)

4. De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., et al.: Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine* **24**(9), 1342 (2018)
5. Frawley, J., Willcocks, C.G., Habib, M., Geenen, C., Steel, D.H., Obara, B.: Segmentation of macular edema datasets with small residual 3D U-Net architectures. In: 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE). pp. 582–587 (2020). <https://doi.org/10.1109/BIBE50027.2020.00100>
6. Ghavami, N., Hu, Y., Gibson, E., Bonmati, E., Emberton, M., Moore, C.M., Barratt, D.C.: Automatic segmentation of prostate MRI using convolutional neural networks: Investigating the impact of network architecture on the accuracy of volume measurement and MRI-ultrasound registration. *Medical Image Analysis* **58**, 101558 (2019)
7. Goldberg, R.A., Waheed, N.K., Duker, J.S.: Optical coherence tomography in the preoperative and postoperative management of macular hole and epiretinal membrane. *British Journal of Ophthalmology* **98**(Suppl 2), ii20–ii23 (2014)
8. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European Conference On Computer Vision. pp. 630–645 (2016)
9. Hee, M.R., Puliafito, C.A., Wong, C., Duker, J.S., Reichel, E., Schuman, J.S., Swanson, E.A., Fujimoto, J.G.: Optical coherence tomography of macular holes. *Ophthalmology* **102**(5), 748–756 (1995)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) International Conference on Learning Representations (2015)
11. Lindsay, G.W.: Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience* p. 1–15 (Feb 2020). [https://doi.org/10.1162/jocn\\_a.01544](https://doi.org/10.1162/jocn_a.01544)
12. Madi, H.A., Masri, I., Steel, D.H.: Optimal management of idiopathic macular holes. *Clinical Ophthalmology (Auckland, NZ)* **10**, 97 (2016)
13. McCannel, C.A., Ensminger, J.L., Diehl, N.N., Hodge, D.N.: Population-based incidence of macular holes. *Ophthalmology* **116**(7), 1366–1369 (2009)
14. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 565–571 (2016)
15. Murphy, D.C., Nasrulloh, A.V., Lendrem, C., Graziado, S., Alberti, M., la Cour, M., Obara, B., Steel, D.H.: Predicting postoperative vision for macular hole with automated image analysis. *Ophthalmology Retina* (2020). <https://doi.org/10.1016/j.oret.2020.06.005>, <http://www.sciencedirect.com/science/article/pii/S2468653020302311>
16. Nasrulloh, A., Willcocks, C., Jackson, P.T., Geenen, C., Habib, M.S., Steel, D.H., Obara, B.: Multi-scale segmentation and surface fitting for measuring 3D macular holes. *IEEE Transactions on Medical Imaging* **37**(2), 580–589 (2018)
17. Paszke, A., Gross, S., Massa, F., et al.: PyTorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32, pp. 8024–8035 (2019)
18. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention. vol. 9351, pp. 234–241 (2015)
19. Steel, D.H., Donachie, P.H., Aylward, G.W., Laidlaw, D.A., Williamson, T.H., Yorston, D.: Factors affecting anatomical and visual outcome after macular hole surgery: findings from a large prospective UK cohort. *Eye* pp. 1–10 (2020)
20. Taha, A.A., Hanbury, A.: Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging* **15**(1), 1–28 (2015)