Original Article

Loss-Modified Transformer-Based U-Net for Accurate Segmentation of Fluids in Optical Coherence Tomography Images of Retinal Diseases

Abstract

Background: Optical coherence tomography (OCT) imaging significantly contributes to ophthalmology in the diagnosis of retinal disorders such as age-related macular degeneration and diabetic macular edema. Both diseases involve the abnormal accumulation of fluids, location, and volume, which is vitally informative in detecting the severity of the diseases. Automated and accurate fluid segmentation in OCT images could potentially improve the current clinical diagnosis. This becomes more important by considering the limitations of manual fluid segmentation as a time-consuming and subjective to error method. **Methods:** Deep learning techniques have been applied to various image processing tasks, and their performance has already been explored in the segmentation of fluids in OCTs. This article suggests a novel automated deep learning method utilizing the U-Net structure as the basis. The modifications consist of the application of transformers in the encoder path of the U-Net with the purpose of more concentrated feature extraction. Furthermore, a custom loss function is empirically tailored to efficiently incorporate proper loss functions to deal with the imbalance and noisy images. A weighted combination of Dice loss, focal Tversky loss, and weighted binary cross-entropy is employed. **Results:** Different metrics are calculated. The results show high accuracy (Dice coefficient of 95.52) and robustness of the proposed method in comparison to different methods after adding extra noise to the images (Dice coefficient of 92.79). **Conclusions:** The segmentation of fluid regions in retinal OCT images is critical because it assists clinicians in diagnosing macular edema and executing therapeutic operations more quickly. This study suggests a deep learning framework and novel loss function for automated fluid segmentation of retinal OCT images with excellent accuracy and rapid convergence result.

Keywords: *Customized loss function, deep learning, fluid accumulation, optical coherence tomography, semantic segmentation*

Reza Darooei[1,2], Milad Nazari[3,4], Rahle Kafieh[1,5], Hossein Rabbani[1,2]

[1]Medical Image and Signal Processing Research Center, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, [2]Department of Bioelectrics and Biomedical Engineering, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, [3]Department of Molecular Biology and Genetics, Aarhus University, [4]DANDRITE – The Danish Research Institute of Translational Neuroscience, Aarhus University, Aarhus, Denmark, [5]Department of Engineering, Durham University, Durham, United Kingdom

Submitted: 20-Aug-2022 **Revised:** 29-Jan-2023 **Accepted:** 06-Apr-2023 **Published:** 31-Aug-2023

Introduction

Optical coherence tomography (OCT) is a noninvasive modality that can reconstruct high-resolution cross-sectional images containing structural and molecular information of biological tissues.[1,2] For example, ocular OCT is widely used to diagnose different retinal pathologies, such as hyperreflective retinal foci and cystoid areas.[3]

Macular edema (ME) refers to swelling of the macula in the area of the central vision. It is an abnormal blister of fluid due to bleeding of the retinal barrier. ME can cause different retinal diseases, such as age-related macular degeneration (AMD) and diabetic ME (DME). DME happens when excess fluid builds up and is a complication of diabetes. Early and accurate diagnosis of DME has many benefits for the public health system patient due to the on-time cure and follow-up of the severity level of DME.[4]

The detection and identification of fluid-filled regions in ME disorders are medically essential. Because of the nature of OCT images, diagnosis of fluid-filled areas is time-consuming, and it needs the experience to figure out fluids correctly.[5]

Semantic segmentation is a fundamental and challenging task in image analysis for fluid or cyst segmentation which is to label each pixel of an image with a corresponding class. Over the past few years, unlike many conventional machine learning methods, deep learning models have achieved

For reprints contact: WKHLRPMedknow_reprints@wolterskluwer.com

How to cite this article: Darooei R, Nazari M, Kafieh R, Rabbani H. Loss-modified transformer-based U-Net for accurate segmentation of fluids in optical coherence tomography images of retinal diseases. J Med Sign Sens 2023;13:253-60.

Address for correspondence:
Dr. Hossein Rabbani,
Medical Image and Signal Processing Research Center, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, Iran.
E-mail: h_rabbani@med.mui.ac.ir

Access this article online

Website: www.jmssjournal.net

DOI: 10.4103/jmss.jmss_52_22

Quick Response Code:

remarkable performances and the highest accuracy rates on popular benchmarks, resulting in a paradigm shift in the field.[6]

One of the first deep learning methods for semantic segmentation is convolutional neural networks (CNNs) and fully convolutional networks.[7] A fully CNN called U-Net was created for the segmentation of medical images and is widely used in OCT segmentation for different tasks such as layers and cyst segmentation.[8] U-Net structure consists of two parts, namely the encoder and decoder, with skip connections to connect these two parts. Since U-Net is a CNN-based method, it has some drawbacks, such as the intrinsic locality of convolution operations, and it is not easy to model long-range relations.[9] Different methods, such as attention U-Net,[10] U-Net3^{++},[11] and Res-U-Net,[12] have been developed that try to remedy intrinsic locality problems. To choose the best model to serve as the base for evaluating various loss functions in the next stage, we used a variety of cutting-edge semantic segmentation techniques. The different types of U-Net,[7] such as Attention U-Net,[10] U-Net^{+++},[11] R2 U-Net,[13] Trans-U-Net,[14] and Swin-U-Net,[15] are implemented to get the best performance for OCT fluid semantic segmentation. On the other hand, OCT cyst segments are imbalanced areas in the B-scans,[16] so learning the U-Net using a custom loss function is needed.

To overcome the mentioned problems, we propose an automated fluid semantic segmentation method for new OCT images. The main contributions of this article are as follows:

1. Different state-of-arts semantic segmentation methods were applied to get the best performance on the OCT semantic segmentation dataset
2. We suggested a fully automated approach based on Trans-U-Net
3. A customized loss function was implemented to work in the imbalanced datasets
4. We added noise to show the robustness of the proposed U-Net method compared to other loss functions.

Methods

Different state-of-art U-Nets are utilized to analyze the effect of each network on the performance of OCT semantic segmentation in the deep learning framework. Various deep learning architectures, such as Trans-U-Net, Swin-U-Net, U-Net, attention U-Net, and R2 U-Net, are applied with similar and optimal parameter numbers. Because the Trans-U-Net in the results section has better performance than the others, we only focus on the Trans-U-Net in this section. We extracted different data augmentation subimages of size 512 × 512 pixels for preventing overfitting from the dataset, resulting in a stratified training set with around 2000 subimages. These data augmentation processes were coded in the Python programming language. In this section, we first introduce the Trans-U-Net duo to show that Trans-U-Net has better performance than other networks, then, in the last part, we provide the dataset and discuss the theoretical foundations of the suggested loss function.

Trans-U-Net

Encoder part

Transformers were first applied in natural language processing, and it used in the state-of-the-art in many NLP tasks.[17] The input of a Trans-U-Net sequence[14] is an image (*H,W, and C*) with a spatial resolution of *H* × *W* and *C* number of color channels. Our goal is to predict the corresponding pixel-wise mask. Network architecture is illustrated in Figure 1. In the first stage, five convolutional layers with stride two are applied to extract low-resolution feature maps. In the next step, input images are tokenized and reshaped to the two-dimensional flatten patches. The transformer's input is the length of *L* and the sequence, and the size of the patches is (*P* × *P*).

$$L = \frac{H \times W}{(\text{pach\_size})^2} \tag{1}$$

Patches are projected into the vector, and it maps into D-dimensional. Position embedding data are used to encrypt patch spatial information. The patch spatial information is then encoded using a trainable positional embedding layer to store the spatial information that the transformer encoder layer can accurately model. Eq. 2 displays the encoder's starting value, where *E* is a patch embedding projection,

$E_{POSITION}$ is position embedding and x_p^i shows *i*th vectorized patch.

$$z_0 = \left[x_p^1 E; x_p^2 E\, x_p^1 E\, ; . . . ; x_p^N E\right] + E_{\text{POSITION}} \tag{2}$$

Each layer of the transformer encoder has *L* layers of multi-head self-attention (MSA) blocks and multilayer perceptron (MLP) blocks shown as:

$$\begin{aligned} z_l^{'} &= \text{Multihead Self - Attention}\left(LN\left(z_{l\text{-}1}\right)\right) + z_{l\text{-}1} \\ &= MSA\left(LN\left(z_{l\text{-}1}\right)\right) + z_{l\text{-}1} \end{aligned} \tag{3}$$

$$\begin{aligned} z_l &= \text{Multilayer Perceptron}\left(LN\left(z_{l\text{-}1}\right)\right) \\ &= z_{l\text{-}1}^{'} MLP\left(LN\left(z_{l\text{-}1}\right)\right) + z_{l\text{-}1}^{'} \end{aligned} \tag{4}$$

LN(.) denotes the layer normalization operator, a residual connection that bypasses each block to form an identity mapping and a layer normalization operator. The encoded image is finally obtained after iterative (3)-(4) calculations.

Decoder part

Similar to the U-Net concatenation component, the CNN decoder is used to take the abstract representation. However, there is a slight difference between absorbing feature maps from the transformer encoder and predicting the image. The decoder block starts with upsampling. Next, regular convolution operations concatenate feature maps from the previous layer. It resembles the U-Net decoder

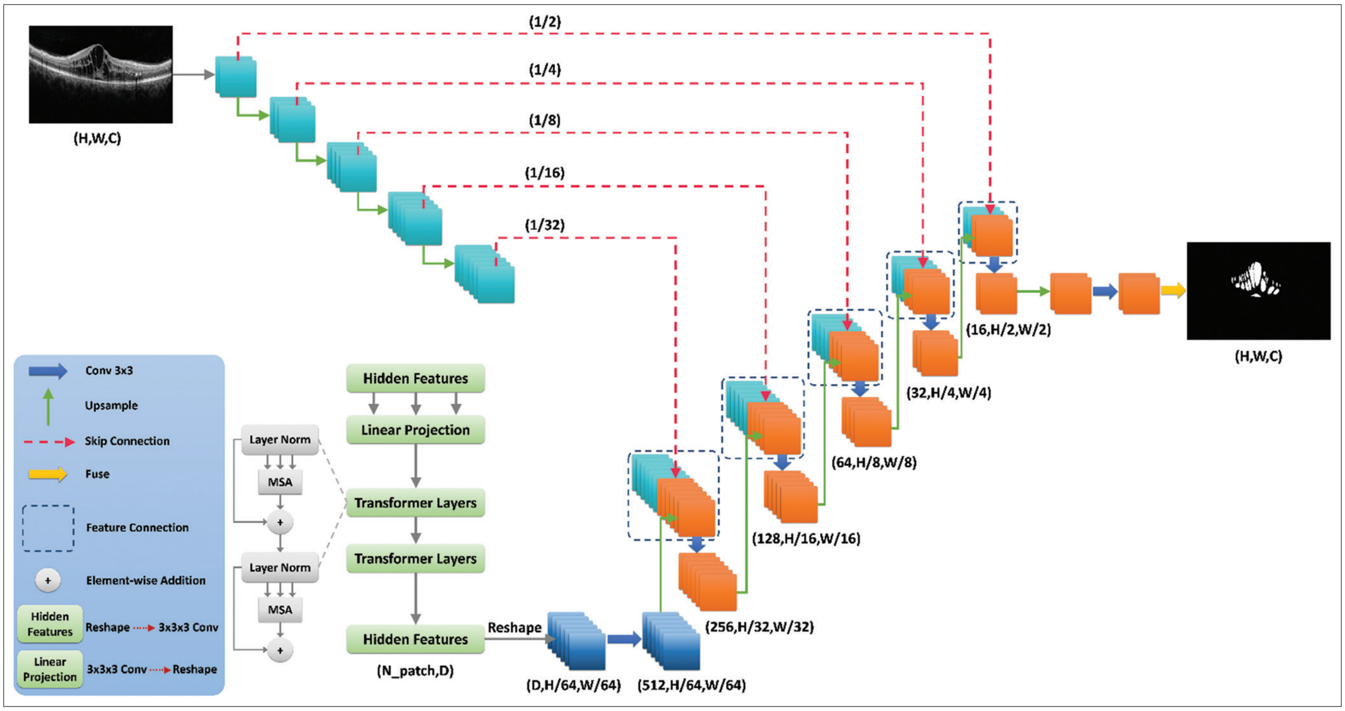

Figure 1: The network architecture that is suggested for the optical coherence tomography semantic segmentation. MSA: Multi-head self-attention

up to this point. It is now combined with the extracted transformer feature maps to create the final tensor. The fusion of the last layer results in the final step will provide a one-dimensional predicted mask.

Dataset

The dataset used in this study comes from two disparate datasets (normal and DME) introduced in.[18] A raw dataset, including fluid images, was used. The acquired data contains 194 scans with fluid from the cirrus OCT device. In addition, we added 185 normal images to balance the proposed dataset. To increase the size of our dataset and evaluate the agreement and repeatability of the suggested approach network, we applied data augmentation, including rotation, shift, and crop. The dataset was split into 80% and 20%, four quarters of the dataset were used as a training set, and the remaining were used as the test set. We applied k-fold cross-validation. The number of groups into which a given data sample is to be divided is specified by a single parameter called k. The performance of the k-fold cross-validation method for loss function was examined. We used five-fold cross-validation and drew graphs based on the mean of folds. The image sizes of the dataset are different. Dataset and masks are reshaped into 512 × 512.

To test the robustness of the suggested technique, Gaussian noise is applied to the images in the pixel domain. This results in a new dataset.

Loss function

Handling unbalanced data is one of the main issues in the OCT fluid segmentation task.[16,19] The OCT mask foreground region is too small in the context of the background image. As mentioned before, the Dice score coefficient is an overlap index that treats false negatives and false positives equally and has good performance in the segmentation. Tversky loss is based on linear Dice loss.[20] To reduce the imbalance effect, focal Dice loss, which is a generalization of the Tversky loss, is also used.[9] Weighted binary cross-entropy (WBCE) loss encourages the proposed segmentation network to predict near the ground truth mask and help loss converge faster.[21-23] These loss functions are combined by weights *α, β* and *γ*, respectively. Hence, our custom loss function using a combination of WBCE, focal Tversky, and Tversky to handle the imbalanced dataset is defined as follows.

$$L_{\text{Custom}} = \alpha\text{Loss}_{\text{Dice}} + \beta\text{Loss}_{\text{Focal\_Dice}} + \gamma\text{Loss}_{\text{WBCE}} \quad (5)$$

Metrics

Dice coefficient

We use the Dice coefficient and Jaccard coefficient as the proper basis for comparing the located fluid objects. The Dice coefficient, also known as the f1-score, reflects how comparable the predicted mask and the grand truth mask of images are, and it depends on the overlap between the results. The formula is defined as follows:

$$\text{Dice}(A, B) = \frac{2\,|\,\text{A} \cap \text{B}\,|}{|\,\text{A}\,| + |\,\text{B}\,|} \quad (6)$$

Jaccard index

In semantic segmentation, one of the metrics that is used most often is called the Intersection over Union, which is also known as the Jaccard Index. It measures similarities between sets, and the below formula defines it:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{7}$$

Results and Discussion

In the problem of OCT fluid segmentation, the fluid region is only a small portion of the image. This image's mask suffers from class imbalance. We employed different methods to overcome this problem and improve results. Different U-Nets and loss functions were proposed to handle this problem. To segment fluid regions automatically, we designed an improved Trans-U-Net-based novel loss function.

Network selection

The results are collected and implemented in Python 3.9 using two GeForce GTX 1080Ti GPUs. In this work segmentation, the model was implemented with four different loss functions, two sets of data, and also six deep learning networks with the optimal parameters. The proposed datasets are simple and noisy. We used our method based on different U-Nets to segment OCT fluids. The proposed method was cross-validated by splitting 80% as train and 20% as the test data. In the first step, we choose the best network for deep learning semantic segmentation. Loss functions of all U-Nets are Dice loss as the base loss function and 100 epochs for all U-Nets are applied.

We evaluated the effectiveness of many cutting-edge U-Nets for fluid segmentation approaches for fluid localization in DME and AMD patients. In the beginning, we do a quantitative analysis of the segmentation results on images from our dataset. Figure 2 illustrates the segmentation results of each network.

The quantitative results of the segmentation are shown in Figure 2, which was created by applying hyperparameter optimization to the proposed networks and then applying those networks to images from six different networks. In addition, Table 1 presents the findings from several models utilizing Jaccard and Dice outcomes.

Figure 2's first and second columns display the results of multiple networks' Dice and Jaccard. The third column displays the results of the loss function with fixed Dice loss. The Trans-U-Net approach outperforms other U-Nets and cutting-edge networks, as can be shown in the diagram. When using the settings from Table 1, Trans-U-Net outperforms other networks using the same filter sizes in terms of validation results. Figure 2 and Table 1 show that Trans-U-Net has better performance than other networks.

The number of the encoder CNN layers was five, the patch size was (4, 4), and the other model parameters are shown in Table 2.

Figure 3 displays the segmentation outcomes for each network in additional cells, the gold-standard annotations (second column), and the qualitative findings of networks with fluid. The Trans-U-Net method, as shown in the image, performs well in both closed and large fluid environments.

Table 1: The results of the 5-fold cross-validation score of cutting-edge U-Nets models after performing hyperparameter tuning

	Dice (validation)	Jaccard (validation)
U-Net	85.90 (78.70)	78.70 (55.73)
Attention U-Net	88.30 (79.30)	79.30 (53.80)
Trans-U-Net	89.54 (80.40)	80.60 (66.80)
R^2 U-Net	88.40 (78.90)	78.90 (63.16)
U-Net+++	76.50 (63.90)	63.90 (45.70)
Swin-U-Net	77.60 (65.60)	71.63 (42.30)

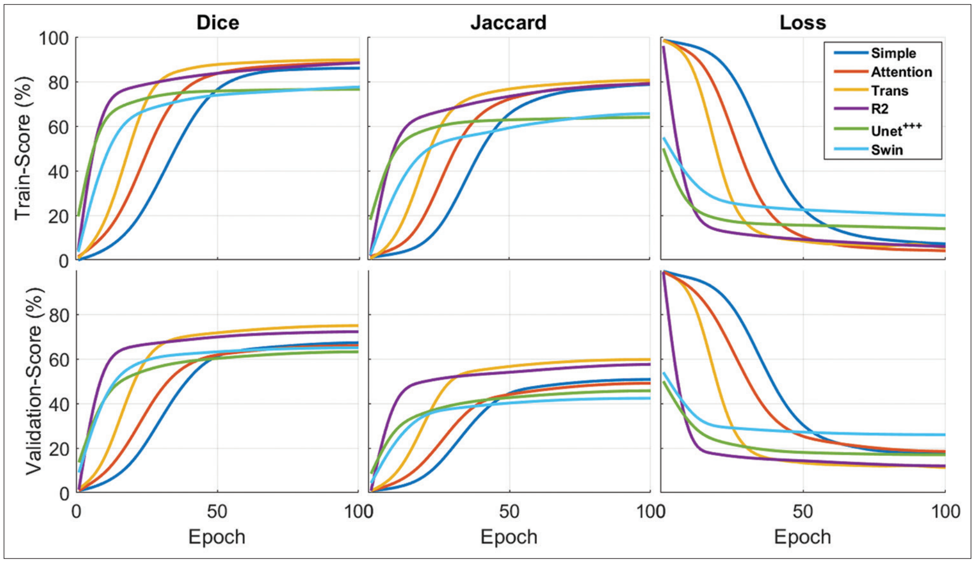

Figure 2: Compares the various models using the Dice, Jaccard, and Loss curves. In the metrics comparison, all of the models perform well. Trans-U-Net models, with 89.54 and 77.2 validation, provide the best Dice and Jaccard outcomes, whereas Swin-U-Net produces the poorest

These findings led us to choose Trans-U-Net with the best settings as the main network for loss function analysis.

Table 2: Network parameters

	Modified Trans-Unet parameters
Filter number	5
Transformer blocks	2
Convolutional layers per down-sampling level	2
Convolutional layers (after concatenation) per up-sampling level	2
Attention heads	2
MLP nodes per vision transformer	384
Embedding dimensions	96
Activation for transformer MLPs	Gaussian error linear unit

MLP: Multi-layer perceptrons

Loss function analysis

In the last part of the network section part, Trans-U-Net with the optimal parameters is selected. In this section, we are trying to improve the network using a modified loss function.

The plot of the assessment of the quantitative results (Dice, Jaccard, and Loss) versus the number of epochs in the proposed dataset is shown in Figures 4-6.

Figures depict the convergence speed and result in the proposed loss function during the training and validation phase having better performance than the cutting-edge loss functions for the imbalance dataset.

Table 3 displays the Dice and Jaccard findings for normal images, whereas Table 4 shows the results for noisy

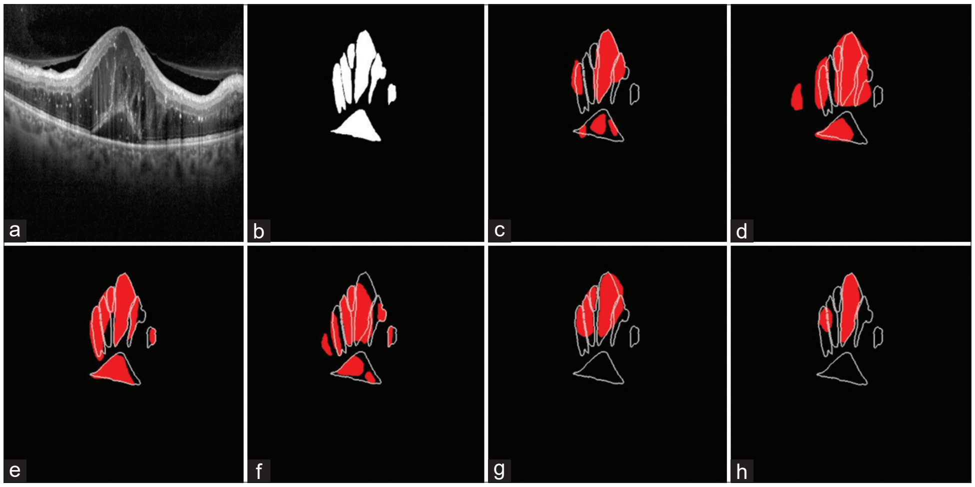

Figure 3: Comparing segmentation outcomes from several U-Net experiments on OCT B-scan images. (a) The original image, (b) True mask, (c) Simple U-Net, (d) Attention U-Net, (e) Trans-U-Net, (f) R2 U-Net, (g) U-Net^{+++}, U-Net, and Swin-U-Net, (h) Swin-U-Net

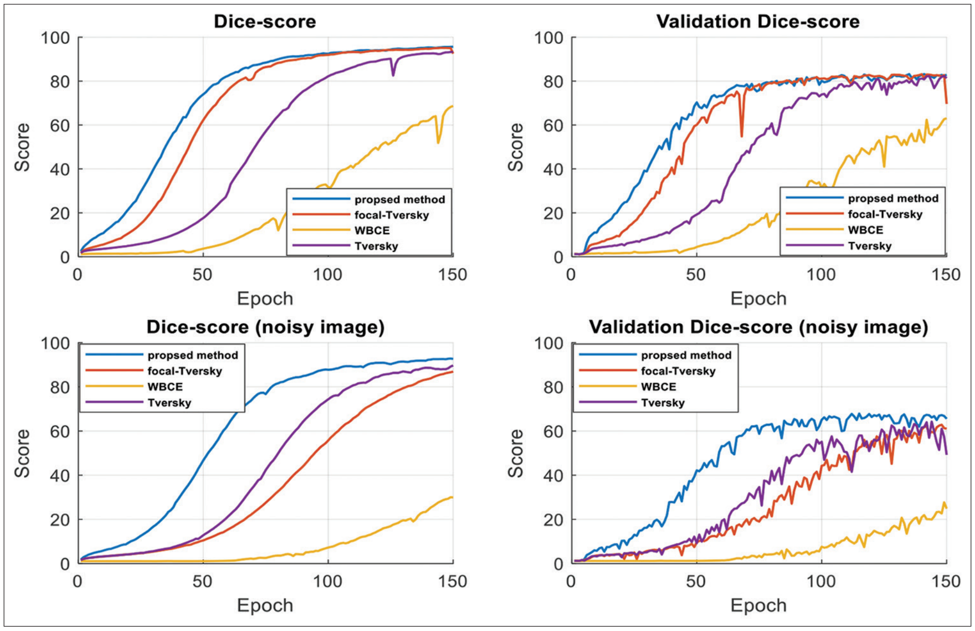

Figure 4: Dice score for initial and noisy image datasets. WBCE: Weighted binary cross-entropy

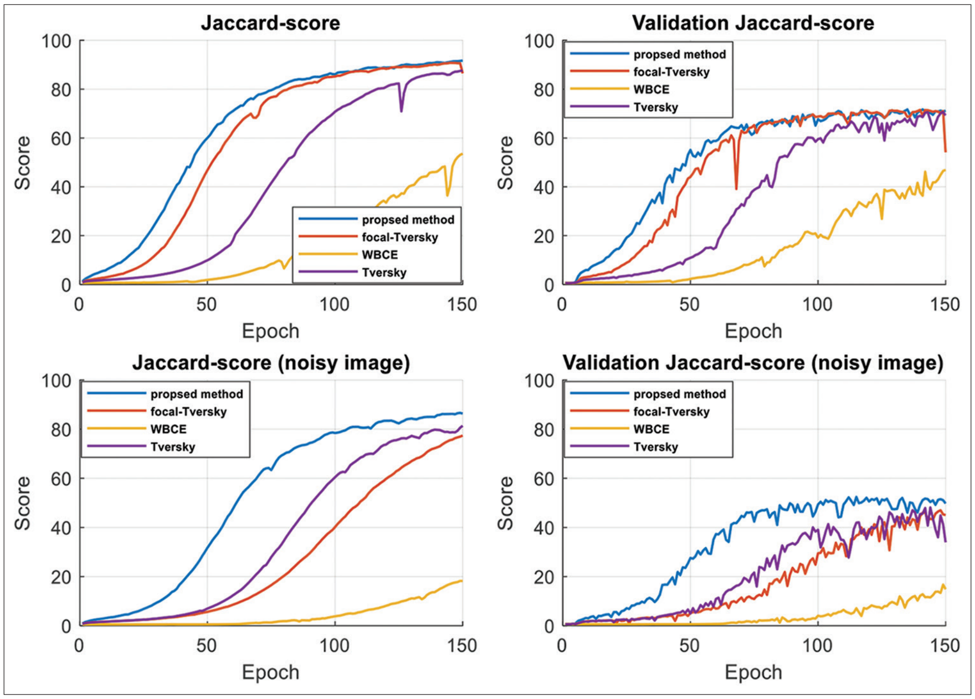

Figure 5: Jaccard score of initial and noisy image datasets. WBCE: Weighted binary cross-entropy

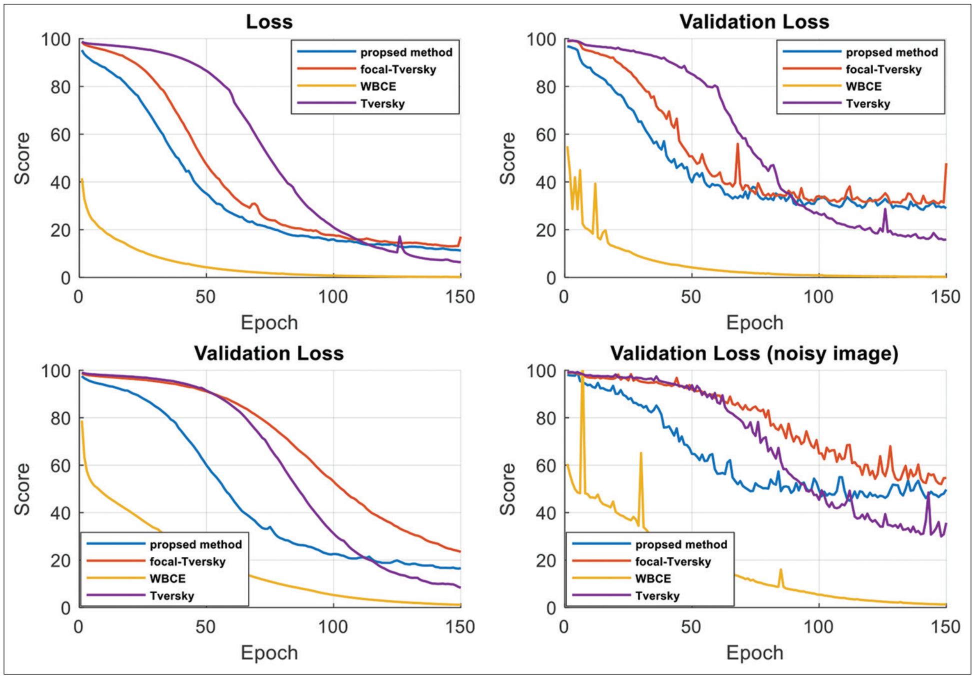

Figure 6: Loss and validation loss of initial and noisy images. WBCE: Weighted binary cross-entropy

images. According to the results, the proposed method achieved higher performance in quantitative analysis. The numerical results express great superiority in Dice, Jaccard, and Loss metrics and in speed for segmenting fluid regions for the proposed initial and noisy datasets.

Figure 7 compares the visual results of different loss functions and the proposed method. Figure 7a is the results for the initial images, and Figure 7b is the results for noisy images. The proposed method with hybrid loss functions achieves the best performance in segmentation results than others.

Conclusions

Accurate fluid detection in OCT images is crucial for following up on ocular diseases such as DME and AMD. However, the cyst segmentation process is

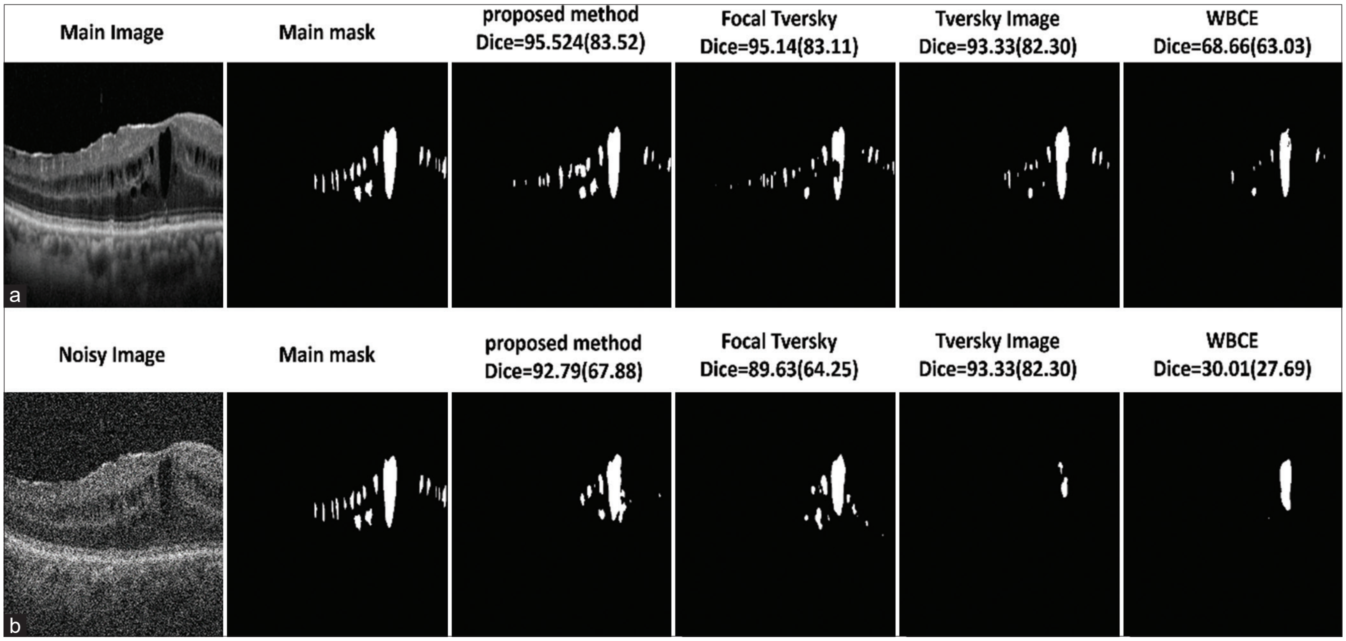
Main Image
Main mask
proposed method Dice=95.524(83.52)
Focal Tversky Dice=95.14(83.11)
Tversky Image Dice=93.33(82.30)
WBCE Dice=68.66(63.03)
a
Noisy Image
Main mask
proposed method Dice=92.79(67.88)
Focal Tversky Dice=89.63(64.25)
Tversky Image Dice=93.33(82.30)
WBCE Dice=30.01(27.69)
b

Figure 7: The visual segmentation results for different loss function methods for a sample B-scan. The number inside the bracket shows dice validation results for each method. (a) Initial images. (b) Noisy images. WBCE: Weighted binary cross-entropy

Table 3: The quantitative results of the initial image dataset

	Dice (validation)	Jaccard (validation)
Proposed method	95.524 (83.52)	90.88 (72.21)
Focal dice	95.14 (83.11)	90.75 (71.52)
WBCE	68.66 (63.03)	53.56 (46.94)
Tversky	93.33 (82.30)	87.57 (70.63)

WBCE: Weighted binary cross-entropy

Table 4: The quantitative results of the noisy image dataset

	Dice (validation)	Jaccard (validation)
Proposed method	92.79 (67.88)	86.61 (56.20)
Focal dice	86.88 (62.82)	77.37 (48.14)
WBCE	30.01 (27.69)	18.24 (12.21)
Tversky	89.63 (64.25)	81.34 (47.34)

WBCE: Weighted binary cross-entropy

highly time-consuming and challenging. Therefore, automatic semantic segmentation is a valuable solution to this problem. In this article, we introduce a novel loss function for using a deep learning network on an imbalanced OCT dataset and implement it in our dataset.

The resulting architecture of the proposed method inherits the advantages of other methods. It improves network performance and learning convergence speed, and the results were robust in semantic segmentation compared to loss functions. In future work, we will explore the genetics algorithm for automatic parameter selection and improve Trans-U-Net architecture by adding residual layers.

Financial support and sponsorship

None.

Conflicts of interest

There are no conflicts of interest.

References

1. Aumann S, Donner S, Fischer J, Müller F. Optical coherence tomography (OCT): principle and technical realization. High resolution imaging in microscopy and ophthalmology: New frontiers in biomedical optics 2019. p. 59-85.
2. Podoleanu AG. Optical coherence tomography. Br J Radiol 2005;78:976-88.
3. Mokhtari M, Kamasi ZG, Rabbani H. Automatic Detection of Hyperreflective Foci in Optical Coherence Tomography B-scans Using Morphological Component Analysis. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE; 2017.
4. Marmor MF. Mechanisms of fluid accumulation in retinal edema. InMacular Edema: Conference Proceedings of the 2nd International Symposium on Macular Edema, Lausanne, Springer Netherlands. 23–25 April 1998 2000 (pp. 35-45).
5. Ouchi M, West K, Crabb JW, Kinoshita S, Kamei M. Proteomic analysis of vitreous from diabetic macular edema. Exp Eye Res 2005;81:176-82.
6. Lateef F, Ruichek Y. Survey on semantic segmentation using deep learning techniques. Neurocomputing 2019;338:321-48.
7. Long J, Shelhamer E, Darrell T. Fully Convolutional Networks for Semantic Segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015.
8. Ronneberger O, Fischer P, Brox T. U-net: Convolutional Networks for Biomedical Image Segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2015.
9. Abraham N, Khan NM. A Novel Focal Tversky Loss Function with Improved Attention U-Net for Lesion Segmentation.

In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE; 2019.

10. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, *et al*. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999. 2018.
11. Huang H, Lin L, Tong R, Hu H, Zhang Q, Iwamoto Y, *et al*. Unet 3+: A full-scale connected unet for medical image segmentation. InICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2020 (pp. 1055-1059). IEEE.
12. Diakogiannis FI, Waldner F, Caccetta P, Wu C. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journal of Photogrammetry and Remote Sensing 2020;162:94-114.
13. Alom MZ, Hasan M, Yakopcic C, Taha TM, Asari VK. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. arXiv preprint arXiv:1802.06955. 2018.
14. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, *et al*. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306. 2021.
15. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, *et al*. Swin-unet: unet-like pure transformer for medical image segmentation. CoRR. arXiv preprint arXiv:2105.05537. 2021.
16. Oguz I, Zhang L, Abràmoff MD, Sonka M. Optimal retinal cyst segmentation from OCT images. InMedical Imaging 2016: Image Processing SPIE. 2016;9784:375-81.
17. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, *et al*. Transformers: State-of-the-art natural language processing. InProceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations 2020 p. 38-45.
18. Montazerin M, Sajjadifar Z, Khalili Pour E, Riazi-Esfahani H, Mahmoudi T, Rabbani H, *et al*. Livelayer: A semi-automatic software program for segmentation of layers and diabetic macular edema in optical coherence tomography images. Sci Rep 2021;11:13794.
19. Liu W, Sun Y, Ji Q. Mdan-Unet: multi-scale and dual attention enhanced nested u-net architecture for segmentation of optical coherence tomography images. Algorithms 2020;13:60.
20. Hashemi SR, Salehi SS, Erdogmus D, Prabhu SP, Warfield SK, Gholipour A. Asymmetric loss functions and deep densely connected networks for highly imbalanced medical image segmentation: Application to multiple sclerosis lesion detection. IEEE Access 2019;7:721-1735.
21. Ho Y, Wookey S. The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. IEEE Access 2019;8:4806-13.
22. Aurelio YS, De Almeida GM, de Castro CL, Braga AP. Learning from imbalanced data sets with weighted cross-entropy function. Neural processing letters 2019;50:1937-49.
23. Rezaei-Dastjerdehei MR, Mijani A, Fatemizadeh E. Addressing Imbalance in Multi-Label Classification Using Weighted Cross Entropy Loss Function. In: 2020 27th National and 5th International Iranian Conference on Biomedical Engineering (ICBME). IEEE; 2020.