Semantic segmentation of retinal pathological fluid from tomographic images with neural models

Ana Ezquerro
ana.ezquerro@udc.es

July 12, 2023

Abstract

Modern Deep Learning architectures have been applied lately in biomedical field in order to provide an non-invasive and automatic diagnostic from high quality images. Convolutional networks have demonstrated their potential to extract deep features from images in order to support and extend the human expert labeling process in anomaly detection from biomedical images [25]. In this work we propose neural-based approaches for semantic segmentation of the retinal pathological fluid from a small set of tomographic images. We tackle the problem of extracting semantic knowledge from less than one hundred labeled images and demonstrate the power of fine-tuning and data augmentation techniques.

1 Introduction

Semantic segmentation is a core task in Computer Vision where a system classifies each pixel of a given image in an specific class [22]. In the last decade, the most successful approaches in semantic segmentation have been neural-based models. Specifically, fully-convolutional neural networks [20, 7, 30] and *Transformer* blocks [29, 31].

The development of these systems in the biomedical field is agreed to be of vital relevance since a good performance might support an automatic non-invasive detection of anomalies and diseases from a set of patients, reducing costs by automating the manual labeling. Despite the most popular datasets in semantic segmentation, ADE20k [33], Cityscapes [8] and PASCAL VOC [11], do not belong to the medical field, several studies have proposed fully-convolutional neural networks to achieve great performances in biomedical datasets. In this context, the Optical Coherence Tomography (OCT) is a popular technique used in semantic segmentation to detect anomalies in high quality images of the retina [1].

In this work we evaluate different neural approaches in semantic segmentation to detect pathological fluid from retinal tomography images. The most popular models in the state of the art have demonstrated a great performance in other settings, trained and evaluated over massive datasets, so we now explore the behavior of these neural approaches limited to extract knowledge from a smaller set of labeled images from different environments.

2 Image dataset

Our dataset consists of 50 retinal images with their corresponding masks which label the presence of pathological fluid at each position of the image. Figure 1 shows an example of the *dataset*, the tomography image and its respective mask.

Due to the small size of the dataset, in order to obtain a robust evaluation of the tested models, k-fold cross-validation ($k = 10$) was used to validate our results, thus, 5 images were reserved for

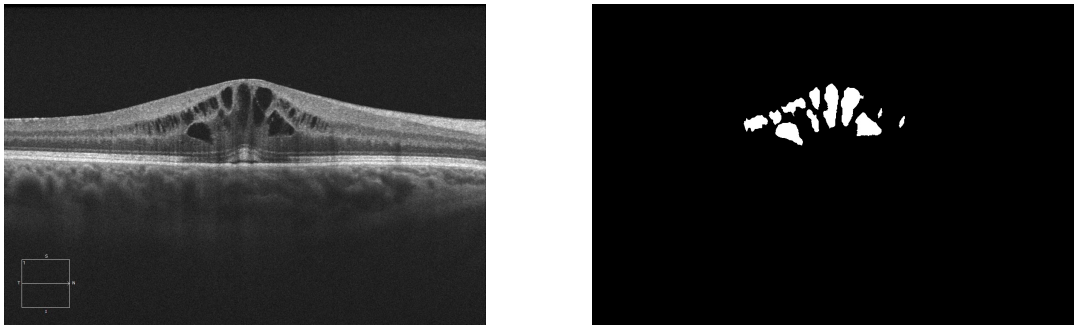

Figure 1: On the left, the captured image of the retina. On the left, the binary mask that shows the presence of pathological fluid.

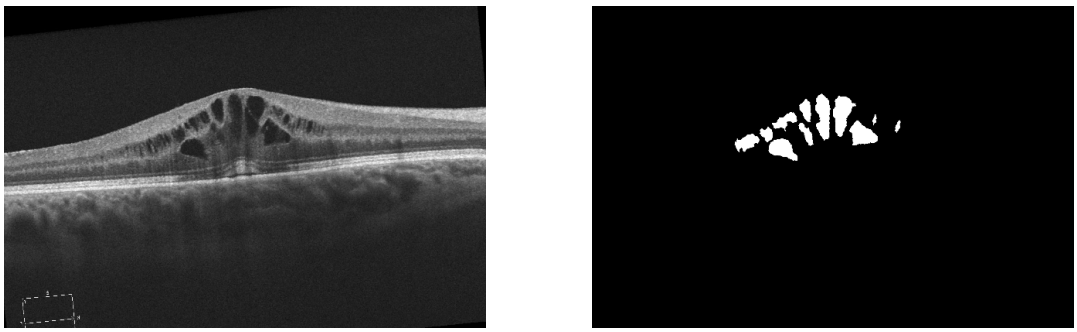

Figure 2: A la izquierda, la imagen con las transformaciones aplicadas. A la derecha, la máscara binaria resultante.

validation and 45 for training in each iteration. Image resolution was adapted to the model, but it was always kept around 400×600 pixels, maintaining a good quality to detect all the details in the image.

We trained each model twice to test the impact of data augmentation. At each epoch, the image is randomly transformed by applying horizontal splits, rotations (of maximum 7 degrees), translations (of maximum 10% of the image), zooms (of maximum 120% of the image) and elastic deformations. In several studies the effectiveness of data augmentation has been proven in order to avoid the overfitting and improve the generalization of the model in semantic segmentation tasks [21, 17, 18]. In Figure 2 we show an example of the final result when this data augmentation operations are applied in a given image. The new image maintains some general features regarding the original image but introduces some variation and noise to the training process in order to avoid the overfitting.

3 Approaches

3.1 Baseline

We took a similar architecture to the popular model U-Net [25] as a baseline for our experiments. U-Net was one of the first convolutional architectures to be successfully applied in semantic segmentation of biomedical images. The success of U-Net comes from the data augmentation technique based on elastic deformations and the connections between the encoder and decoder blocks of its architecture, which help the model to preserve the input information. In Figure 3 we see that the encoder uses two convolutional layers followed by a max pooling layer to reduce the tensor dimension (contraction). The decoder expands the size of the tensor using two convolutional layers and an up-sampling layer to recover the original input size. At each expansion, the features at the decider

are concatenated to the features of the encoder in the same level.

The implementation of this approach was obtained in the GitHub repository pytorch-unet and the image resolution employed was 412×624.

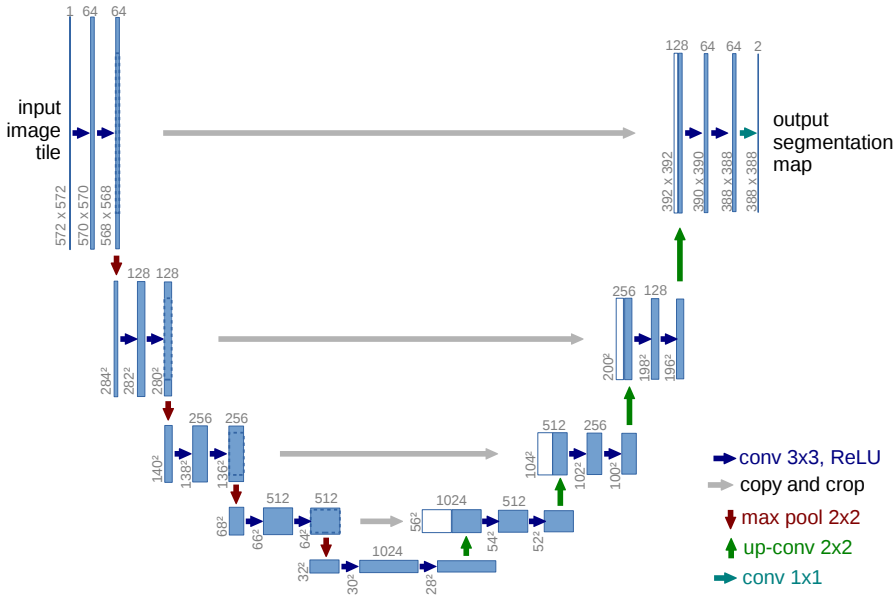

Figure 3: U-Net architecture

We selected Adam [16] as the optimizer and 100 epochs per fold training. The validation set was used for early stopping and save the weights with the best F-score.

3.2 Convolutional models pretrained with ImageNet

In the next approach we used the implementation of the library `segmentation_models_pytorch` to finetune three popular architectures based on convolutional networks with pretrained weights in the encoder: U-Net [25], LinkNet [4] y PSPNet [32], pretrained with ImageNet [10]. The selected *encoder* follows the architecture of ResNet-50 [14] and the same training procedure described in Section 3.1 was followed to finetune the networks.

PSPNet [32] (Pyramid Scene Parsing Network) introduces the pyramid pooling module (Figure 5) which consists of dividing the encoder output in different scales (originally, 4 scales).

1. The encoder consists of a fully convolutional network which maps an input image to its corresponding neural feature representation, an eight times smaller tensor than the original image with m filters ($m \times H/8 \times H/8$).

2. With the encoded image, PSPNet applies pyramid pooling in 4 scales and passes each result to a 1×1 convolution with an up-sampling operation to obtain the original input dimension.

3. All up-sampled images of the pyramid pooling module are concatenated and passed through a new convolutional network to obtain the final mask.

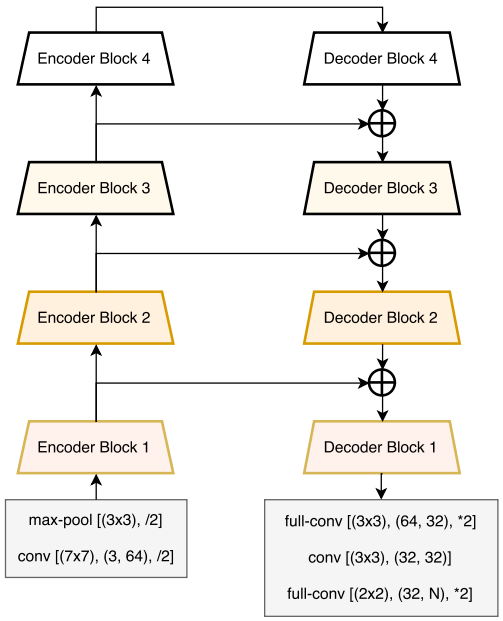

Figure 4: LinkNet architecture

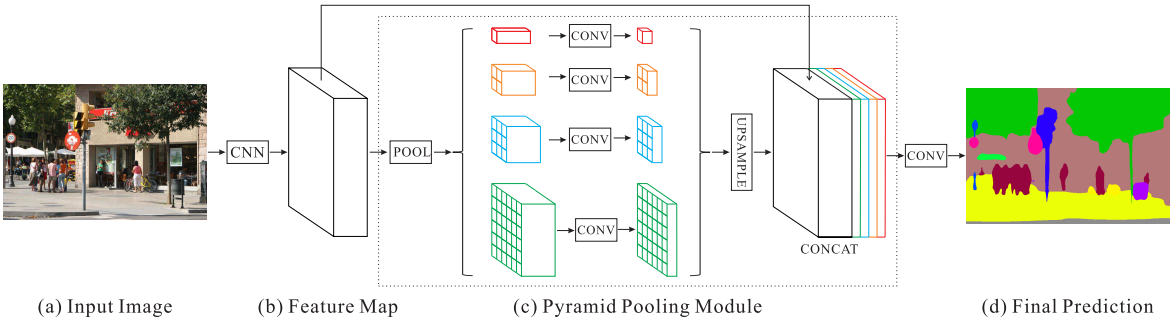

Figure 5: PSPNet architecture [32]

PSPNet introduced successfully the pyramid pooling in semantic segmentation. The pyramid pooling allows the network to learn separately image features from the global environment to local levels.

U-Net, LinkNet and PSPNet were finetuned and compared with the baseline in order to study if the previously learned knowledge during pretraining phase could be transferred to a different dataset.

3.3 Transformers for segmentation tasks

Since the integration the Transformer block [28] and the Self-Attention mechanism in neural network, the state-of-the-art models in the vast majority of Deep Learning problems have been adapted to these architectures. In semantic segmentation, combining Transformers with convolutional layers have demonstrated being extremely effective to robustly extract local features for each pixel to produce a classification [19, 23, 12, 5, 3, 13].

In this approach we recollected semantic segmentation Transformer-based models with pretrained and non pretrained weights:

- PAN [19]: To integrate the [28] mechanism, [19] proposes a new architecture which emulates the Self-Attention module, replacing the pyramid pooling of PSPNet [32] with the Feature Pyramid

Attention and a Global Attention Upsampling module (Figure 6).

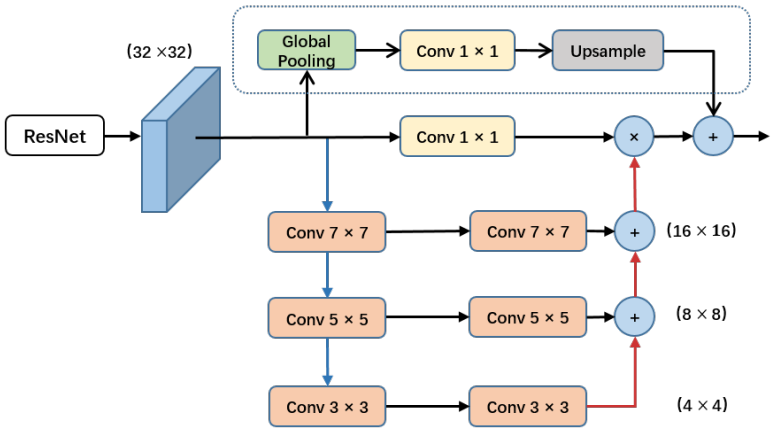

Figure 6: PAN [19] architecture

- Attention U-Net [23]: Inspired on U-Net [25], adds the attention mechanism in the encoder-decoder connections of the original architecture. The objective is fit decoder weights with the information of the contraction process in the encoder (Figure 7).

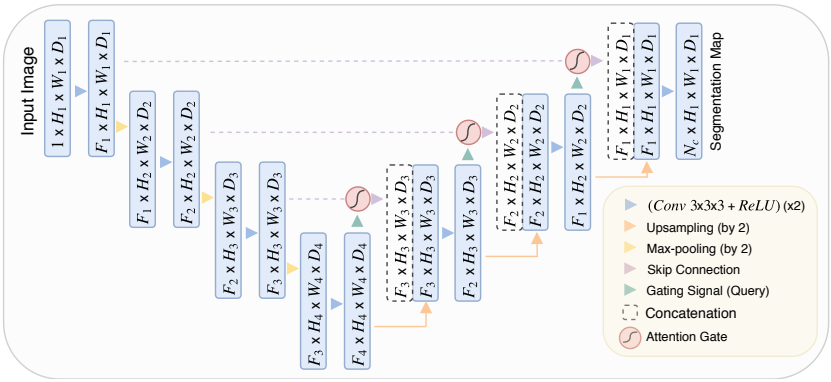

Figure 7: Attention U-Net [23] architecture

Although these approaches have been out-performed by other proposals with more complex architectures and expensive training procedures [30, 6], their release obtained an excellent performance in semantic segmentation and they are still feasible in some environments with low resources.

3.4 Deformable convolutions

One of the most important advances in Computer Vision on the last decade is the development of the deformable convolutions [9, 34, 30]. From 2017 until now, three different architectures based on the deformable convolution have been proposed and, at the time of their publication, they out-performed the state of the art in the majority of Computer Vision tasks. Deformable convolutions essentially consists of adapting the dilation level of the original convolution system by learning the kernel offsets of the receptive field.

In this study we adapted the implementation of PyTorch-Deformable-Convolution-v2 and the static convolution of the baseline (Section 3.1) was substituted by this approach.

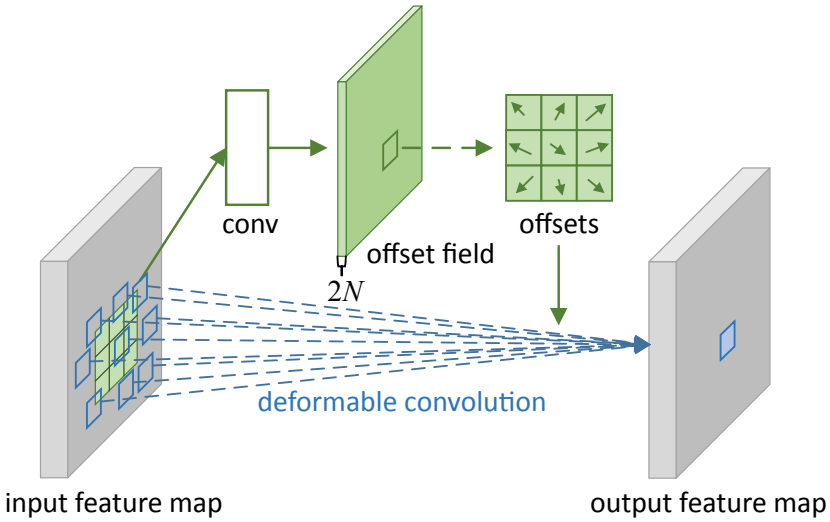

Figure 8: 3×3 deformable convolution [9]

3.5 Adversarial learning

In our last approach, we followed [15, 26] proposals to extend and complete our experiments and trained the approaches introduced in Section 3.2. In this approach, the generative network G consists of one of the previously mentioned models of Section 3.2, while the discriminative network D is a classification convolutional network which accepts an image as input and outputs a unique real number between 0 and 1.

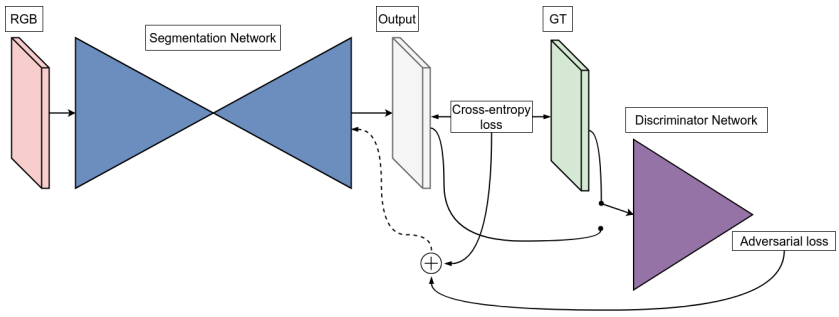

Figure 9: Adversarial learning approach [26]

The adversarial training consists of optimizing at the same time (Figure 9):

- The target of the generative network, the segmentation of the tomographic image to obtain the detection mask of the pathological fluid in the retina.
- The target of the discriminative network, which takes as input the output of the generative network and tries to distinguish which image comes from the ground-truth or form the generative network.

In our experiments the generative network is selected from the models introduced in Section 3.2 and the discriminative network is constructed with 3 convolutional blocks of 32 filters and kernel size $k = 7$, stride of $d = 4$ pixels and dilation of 4, 2 and 2 offsets, respectively. The result is flattened and concatenated to pass it through a feed-forward network with a sigmoid activation function.

4 Experiments: training and evaluation

All of these approaches have been trained twice to evaluate the impact of data augmentation and the robustness of training was validated via k-fold cross validation (where $k = 10$).

The evaluation metrics are the accuracy, precision, recall, F-score and IoU (Intersection over Union). Since the number of prediction classes is imbalanced (approximately the 98% of the pixels corresponds to background and the 2% is pathological fluid), we took as reference the F-score (f_s) and the IoU ($\mathcal{IU}$).

During training, the optimal threshold is determined via IoU with a grid of 20 equidistant values between the minimum and maximum output of the network in the training set. Formally, the optimal threshold τ^* is defined as:

$$\tau^* = \arg\max_{\tau} \Big\{ \mathcal{IU}\Big(\mathbb{I}\{\mathcal{M}(\mathbf{X}) > \tau\}\Big) \mid \tau \in \ell_{20}\Big(\min\{\mathcal{M}(\mathbf{X})\}, \max\{\mathcal{M}(\mathbf{X})\}\Big)\Big\}$$

where $\mathbf{X}$ is the training input set, $\mathcal{M}$ is the evaluated model, $\mathbb{I}(\cdot)$ is the indicator variable of $\cdot$ and $\ell_\xi(y_0, y_1)$ is a function which returns the set of ξ equidistant values between y_0 and y_1.

In order to train the models we used as loss function the binary cross-entropy with logits [24], which facilitates a more numerically stable optimization of the network. Thus, the output of all models was adapted to substitute the final activation function with a linear function (excluding Attention U-Net and Deform U-Net since it was manually tested the improvement of using the sigmoid function as the final activation function of the network).

In Table 1 we see a comparative of the computational costs of each approach with adversarial learning. In Table 2 we see a summary of the training configuration of each approach.

	Default benchmark	*Adversarial benchmark*
Baseline	7.78M	-
U-Net	32.5M	32.6M
LinkNet	31.1M	31.27M
PSPNet	24.9M	24.39M
PAN	24.2M	24.35M
Attention U-Net	42.7M	-
Deform U-Net	28.2M	-

Table 1: Number of trainable parameters of our models

	Loss	*Batch-size*	Tasa aprendizaje	Tamaño imagen
Baseline	*logit*	10	10^{-3}	416×614
U-Net	*logit*	10	10^{-3}	416×640
LinkNet	*logit*	10	10^{-3}	416×640
PSPNet	*logit*	10	10^{-3}	416×640
PAN	*logit*	10	10^{-3}	416×640
Attention U-Net	BCE	5	10^{-4}	416×624
Deform U-Net	BCE	10	10^{-4}	416×624

Table 2: Training configuration of our models

5 Results

In Table 3 we provide a summary of the results with 10-fold cross-validation. The pretrained models in ImageNet [10] are marked with the symbol †.

	Default benchmark				*Adversarial benchmark*			
	No aug.		*Data aug.*		*No aug.*		*Data aug.*	
	f_s	$\mathcal{IU}$	f_s	$\mathcal{IU}$	f_s	$\mathcal{IU}$	f_s	$\mathcal{IU}$
Baseline	$0.6_{0.15}$	$0.44_{0.13}$	$0.65_{0.16}$	$0.56_{0.11}$	-	-	-	-
U-Net †	$0.7_{0.05}$	$0.54_{0.06}$	$0.6_{0.09}$	$0.48_{0.1}$	$0.87_{0.03}$	$0.78_{0.04}$	$0.86_{0.06}$	$0.76_{0.08}$
LinkNet †	$0.6_{0.1}$	$0.43_{0.11}$	$0.57_{0.09}$	$0.4_{0.09}$	$0.77_{0.25}$	$0.67_{0.22}$	$0.83_{0.12}$	$0.72_{0.14}$
PSPNet †	$0.67_{0.04}$	$0.5_{0.04}$	$0.67_{0.05}$	$0.51_{0.06}$	$0.82_{0.03}$	$0.69_{0.04}$	$0.84_{0.04}$	$0.73_{0.05}$
PAN †	$0.7_{0.05}$	$0.54_{0.05}$	$0.66_{0.09}$	$0.5_{0.03}$	-	-	-	-
Attention U-Net	$0.74_{0.06}$	$0.59_{0.07}$	$0.81_{0.08}$	$0.72_{0.08}$	-	-	-	-
Deform U-Net	$0.71_{0.06}$	$0.55_{0.06}$	$0.73_{0.09}$	$0.58_{0.07}$	-	-	-	-

Table 3: Mean and standard deviation (in subscripts) of F-Score and IoU in 10-fold cross-validation

Impact of data augmentation and pre-training Our baseline attains 0.6 F-score and 0.44 IoU without data augmentation. In this summary we see the impact of adding data augmentation and using pretrained weights:

- Using pretrained weights (U-Net, LinkNet y PSPNet) evidences the influence of the data resources. When no data augmentation is applied, the non pretrained model has less potential than the pretrained approaches which have learned to extract image features from ImageNet.
- Applying data augmentation with the baseline model outperforms in 5 points in F-score and 10 points in IoU the metrics of the baseline approach without this technique.

All pretrained models attain approximately 0.7 in F-score and 0.5 in IoU, and maintain their performance when adding data augmentation. We might conclude that the effect of the pretrained weights is similar to the impact of data augmentation.

Transformer-based models The models based on the attention mechanism (Attention U-Net and PAN) reach a similar performance, slightly superior with data augmentation. This effect might be due to the number of parameters (Attention U-Net has six times more parameters than the baseline) and the effectiveness of the attention mechanism, as demonstrated in [27, 5, 12].

Deformable convolutional models Deform U-Net attains an intermediate performance between the attention-based models and fully-convolutional models, despite using non pretrained weights. Deformable convolutions provide flexibility to the model to learn the dilation of the receptive field to optimize the loss function in the backward pass. This allows a higher adaptability of the model to the features that should extract to obtain an accurate representation for the classification problem. Thus, Deform U-Net outperforms the classical fully-convolutional networks.

Adversarial learning Although there is no in-depth analysis to interpret the explainability of this benchmark in semantic segmentation, [15, 26] substantiated that segmentation can be improved

using an adversarial loss function in high resolution satelital images and small image datasets. By reproducing this experiment, we obtained some performance improvements with respect to the classic benchmark.

6 Conclusions

Through this study we have explored different approaches to extract the pathological retinal fluid in tomographic images. Despite we have obtained competitive results, our proposals outperformed by the state of the art (InternImage [30], BEiT [2] and DeepLab v3+ [7]) in popular segmentation datasets, where the IoU is quite superior. However in this work we propose a segmentation problem with limited resources and demonstrate that competitive results can be achieved with simpler segmentation models and a reasonable number of parameters.

References

[1] Deep learning in retinal optical coherence tomography (OCT): A comprehensive survey. *Neurocomputing*, 2022.

[2] H. Bao, L. Dong, S. Piao, and F. Wei. BEiT: BERT Pre-Training of Image Transformers, 2022.

[3] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation, 2021.

[4] A. Chaurasia and E. Culurciello. LinkNet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2017.

[5] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation, 2021.

[6] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking Atrous Convolution for Semantic Image Segmentation, 2017.

[7] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, 2018.

[8] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding, 2016.

[9] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable Convolutional Networks, 2017.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database, 2009.

[11] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vision*, 2010.

[12] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. Dual Attention Network for Scene Segmentation, 2019.

[13] Y. Gao, M. Zhou, D. Liu, Z. Yan, S. Zhang, and D. N. Metaxas. A Data-scalable Transformer for Medical Image Segmentation: Architecture, Model Efficiency, and Benchmark, 2023.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[15] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang. Adversarial Learning for Semi-Supervised Semantic Segmentation, 2018.

[16] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization, 2017.

[17] J. Kugelman, D. Alonso-Caneiro, S. A. Read, J. Hamwood, S. J. Vincent, F. K. Chen, and M. J. Collins. Automatic choroidal segmentation in OCT images using supervised deep learning methods. *Scientific Reports*, 2019.

[18] J. Lee, J. N. Kim, L. Gomez-Perez, Y. Gharaibeh, I. Motairek, G. briel T. R. Pereira, V. N. Zimin, L. A. P. Dallan, A. Hoori, S. Al-Kindi, G. Guagliumi, H. G. Bezerra, and D. L. Wilson. Automated segmentation of microvessels in intravascular OCT images using deep learning, 2022.

[19] H. Li, P. Xiong, J. An, and L. Wang. Pyramid Attention Network for Semantic Segmentation, 2018.

[20] J. Long, E. Shelhamer, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation, 2015.

[21] D. Mahapatra, B. Bozorgtabar, and L. Shao. Pathological retinal region segmentation from oct images using geometric relation based augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[22] Y. Mo, Y. Wu, X. Yang, F. Liu, and Y. Liao. Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing*, 2022.

[23] 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, 2018.

[24] Pytorch. PyTorch: Binary Crossentropy Loss with Sigmoid, 2023.

[25] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation, 2015.

[26] C. Sebastian, R. Imbriaco, E. Bondarev, and P. H. N. de With. Adversarial Loss for Semantic Segmentation of Aerial Imagery, 2020.

[27] R. Strudel, R. Garcia, I. Laptev, and C. Schmid. Segmenter: Transformer for Semantic Segmentation, 2021.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is All You Need, 2017.

[29] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som, and F. Wei. Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks. 2022.

[30] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li, X. Wang, and Y. Qiao. InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions, 2023.

[31] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen. CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation with Transformers, 2023.

[32] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid Scene Parsing Network, 2017.

[33] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene Parsing through ADE20K Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[34] X. Zhu, H. Hu, S. Lin, and J. Dai. Deformable ConvNets v2: More Deformable, Better Results, 2018.