Attention U-Net pour la Segmentation des Pores de la Lame Criblée

Nan DING[1], Hélène URIEN[1], Florence ROSSANT[1], Jérémie SUBLIME[1], Paul BASTELICA[2], Michel PAQUES[3] *

[1]ISEP - Institut Supérieur d'Électronique de Paris, 10 rue de Vanves, 92130 Issy les Moulineaux, France

[2]Centre Hospitalier National d'ophtalmologie, IHU FOReSiGHT, Hôpital des Quinze-Vingts, 75012 Paris, France

[3]Centre d'investigation Clinique 1423, INSERM & Direction de l'Hospitalisation et des Soins, Hôpital des Quinze-Vingts, Sorbonne Université, 75012 Paris, France

nding@ext.isep.fr, (hurien, frossant, jsublime)@isep.fr, (pbastelica,mpaques)@15-20.fr

Résumé – Le glaucome est une neuropathie optique, dont la physiopathologie est encore mal connue. La lame criblée, structure poreuse à travers laquelle les fibres nerveuses passent pour sortir du globe oculaire afin de rejoindre le cerveau, a été identifiée comme site principal des dommages neuronaux retrouvés dans le glaucome. Nous présentons dans cet article une méthode de segmentation des pores de la lame criblée dans des images OCT, par apprentissage profond. Les difficultés sont dues à la faible résolution des images par rapport aux dimensions des zones à segmenter, et au faible rapport signal-à-bruit, ce qui rend la détection des pores, et a fortiori leur segmentation, très difficile. Ainsi nous proposons une architecture de type U-net, avec des optimisations permettant un apprentissage à partir d'annotations incomplètes et la segmentation de petits objets mal contrastés. Les résultats expérimentaux montrent que 71,8% des pores annotés sont segmentés avec succès.

Abstract – Glaucoma is the second leading cause of blindness in the world. Although its physiopathology remains unclear, the lamina cribrosa, a 3D mesh-like structure consisting of pores, that allow the axons passing through to join the brain, has been identified as the primary site of damage. In this work we present an extended version of U-Net for pore segmentation in OCT images with partial points annotation, i.e. having only a small portion of pore locations in each image labeled. Our method combines the attention gate and the context information to address the difficulties caused by small object segmentation in low signal to noise ratio images. Experimental results show that 71.8% of the annotated pores are successfully segmented.

1 Introduction

Le glaucome est une pathologie cécitante qui affecte le nerf optique. La lame criblée (LC), structure poreuse à travers laquelle les axones rétiniens passent pour rejoindre le cerveau, a été identifiée comme site principal des dommages axonaux. L'observation in vivo des pores (i.e. voies axonales) en 3D est désormais possible grâce aux progrès de la technologie de tomographie par cohérence optique (OCT) (Fig. 1). Des modifications de la surface et de la longueur des pores, ont été observées chez des patients atteints de glaucome [1]. Cependant, la physiopathologie de cette maladie est encore très mal connue. Ainsi, notre projet de recherche vise à caractériser les pores de la LC dans le glaucome, ainsi que les modifications morphologiques survenant au cours de la maladie, par l'analyse automatique des images OCT 3D de la LC.

Les études antérieures ont principalement porté sur l'épaisseur de la LC [2], et seulement deux approches ont été proposées pour étudier plus en détails les pores. Les auteurs de [3] ont utilisé un filtre médian et un seuillage local pour segmenter les pores dans les plans 2D *en-face*. Cette méthode s'applique sur chaque image 2D séparément, sans prise en compte de la continuité de l'information dans les coupes consécutives. Dans nos travaux précédents [4], les pores candidats ont été sélectionnés comme les minima locaux d'intensité les plus contrastés dans les images *en-face*, et les voies axonales ont été reconstruites à partir de ces points grâce à un algorithme de suivi. Cependant, l'étape de détection des pores candidats n'est pas fiable, car la forme des pores n'est pas suffisamment prise en compte. De plus, les deux méthodes citées nécessitent une délimitation manuelle des masques périphériques dans les images *en-face*, afin de ne traiter que les régions avec des pores identifiables, étape coûteuse lorsqu'on traite de larges bases de données.

Dans cet article, nous proposons une méthode de segmen-

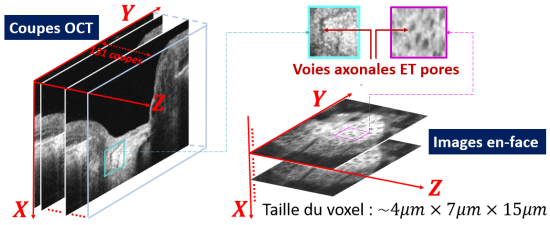

FIGURE 1 – Données OCT 3D. Les images *en-face* (à droite) sont construites à partir des coupes OCT (à gauche) acquises. Axe X : profondeur dans la papille. Les taches sombres dans les images *en-face* correspondent aux pores de la LC.

*Cette étude est co-financiée par l'Institut de la Vision et l'ISEP (IHU M20JRAS004).

tation automatique des pores de la LC dans les images OCT *en-face* par apprentissage profond, ce qui est, à notre connaissance, la première fois que ce problème est abordé avec ce type d'approche. Les méthodes de segmentation faiblement supervisées ne permettent pas d'atteindre la fiabilité requise pour les études cliniques, car les difficultés sont trop importantes : très mauvais rapport signal à bruit, faible résolution des images par rapport à la taille des pores, grande variabilité de forme et d'intensité de ces derniers. Ainsi, dans notre approche précédente [4], un post-traitement manuel était nécessaire pour sélectionner les pores pertinents et éliminer les fausses détections. En revanche, les réseaux de neurones de type U-net [6] ont démontré leur capacité à modéliser des problèmes complexes de segmentation par les données, y compris pour le traitement de structures rétiniennes [7], même de petite taille [8]. Notre objectif est d'obtenir, grâce à un réseau U-net adapté et optimisé, une segmentation précise des pores dans toutes les images *en-face*, suffisamment fiable pour permettre une future reconstruction automatique en 3D des voies axonales, et ceci malgré les difficultés d'annotation manuelles des images pour l'apprentissage.

2 Segmentation des pores avec l'annotation partielle des points

Jeu de données. Notre base de données contient 41 volumes OCT provenant de 17 sujets, sélectionnés par un expert médical, qui a vérifié, pour la faisabilité de l'étude, que les pores sont bien visibles dans les images acquises. Plusieurs volumes peuvent correspondre au même œil pour le suivi longitudinal. Toutes les images sont acquises avec le même appareil Spectralis de Heidelberg Engineering. Chaque acquisition est centrée sur la papille et contient 131 coupes OCT 2D (496 768 pixels, Fig. 1), dont les résolutions transversale et axiale théoriques sont de respectivement 7 m et 4 m. Le pas d'échantillonnage entre deux coupes consécutives est 15 m.

Les images *en-face* sont extraites du volume OCT 3D (Fig. 1), et les pores sont partiellement annotés par 2 experts. Les pores sélectionnés sont marqués par un point proche du centroïde (Fig. 2a). L'objectif de cette annotation est d'identifier les pores les plus larges et qui présentent une continuité dans le volume (i.e. d'un plan en-face aux plans adjacents). Chercher à segmenter exhaustivement tous les pores serait irréaliste : ils sont trop nombreux, beaucoup sont ambigus à cause du faible contraste et des artéfacts, surtout dans les régions dans l'ombre des vaisseaux. Au final, notre jeu de données de 41 volumes est composé de 2101 images *en-face*, avec une moyenne de 8:78 4:75 pores manuellement identifiées, donc une faible proportion de la totalité des pores existants.

Pré-traitements. Les niveaux de gris des images *en-face* (Fig. 2a) sont normalisés sur $[0; 1]$, et chacune est filtrée par deux filtres morphologiques [4] afin de rehausser les pores. Notons I_o l'image prétraitée (Fig. 2b). Les cartes binaires de segmentation de vérité terrain (VT, Fig. 2c), dont on a besoin pour l'apprentissage de notre réseau, sont obtenues en appliquant un algorithme de croissance de région sur I_o, en prenant comme graine les points annotés par les experts. Le critère de similarité est la distance L1 (DST) entre l'intensité d'un pixel non alloué et l'intensité moyenne de la région courante. Considérant le voisinage de la région courante (connexité 8), le pixel avec le plus petit DST sera intégré à la région si DST est inférieur à un seuil $DST_{th} = 0:04$; la croissance s'arrête lorsque DST devient supérieur à DST_{th} pour tous les voisins. Cette valeur seuil a été fixée expérimentalement, afin d'éviter toute sur-segmentation, certains contours étant à peine discernables.

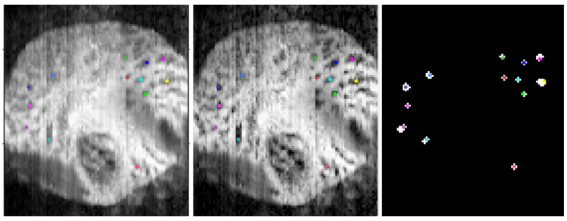

FIGURE 2 – (a) Annotation incomplète des pores. Un pore identifié est marqué d'un point sur l'image *en-face*, en vérifiant la continuité dans les images adjacentes. (b) Image pré-traitée avec des filtres morphologiques pour rehausser les pores. (c) Croissance des régions sur (b) pour générer la vérité terrain.

Méthode proposée. L'U-Net [6] est une architecture neuronale pour la segmentation dont les performances ont été démontrées en imagerie médicale pour de nombreuses modalités. Les *skip connexions* du U-Net aident à récupérer les informations spatiales perdues dans l'encodeur, permettant ainsi une localisation précise des objets d'intérêt. Pour notre application, les pores ont généralement une taille inférieure à 5 5 pixels pour des images d'environ 150 130 pixels (après recadrage pour ne conserver que la région de la papille). Afin d'améliorer la détection des petits objets, les *features* produites peuvent être améliorées en intégrant des mécanismes d'attention dans l'U-Net pour aider à capturer les régions d'intérêt (ROI). Une approche populaire proposée dans [9] repose sur un module *Attention Gate (AG)* (Fig. 4) ajouté à un U-Net. L'AG permet d'estimer les zones potentielles où les pores sont les plus susceptibles d'apparaître en supprimant l'activation des *features* dans les régions non pertinentes, le tout sans avoir besoin de supervision externe. D'autre part, les voies axonales étant régulières, les intensités de pores sont similaires entre les images *en-face* adjacentes, tandis que le centroïde et la forme varient également peu. Par conséquent, une application naïve de U-Net prend le risque de passer à côté de ces propriétés de régularité. De ce fait, nous proposons un réseau en utilisant 3 images adjacentes à l'entrée, et nous produisons une seule carte de segmentation pour l'image du milieu.

La méthode proposée avec l'attention U-Net est illustrée sur la Fig. 3. 3 images *en-face* adjacentes d'entrée sont redimensionnées à 160 160 pixels, puis elles sont progressivement filtrées par (2) blocs de convolution et sous-échantillonnées dans l'encodeur. Pour le décodeur : chaque couche a une *AG*

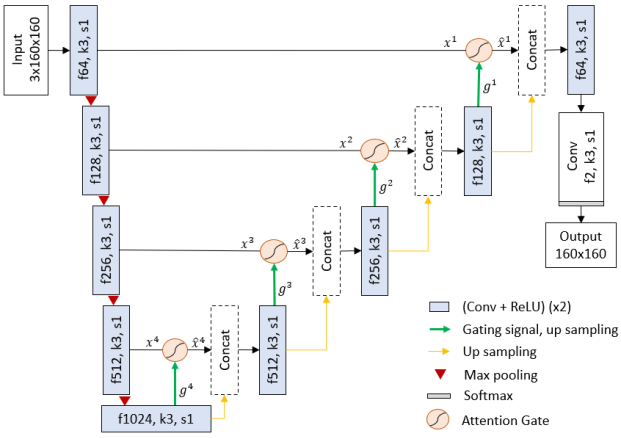

FIGURE 3 – L'architecture d'attention U-Net adaptée de [9]. 'f64,k3,s1' indique le nombre de *feature maps* (64), la taille du noyau (3) et le *stride* (1).

à travers laquelle les caractéristiques de la couche l de l'encodeur doivent passer avant d'être concaténées avec les caractéristiques venant de la couche supérieure $(l+1)$ dans le décodeur. Enfin, une activation *softmax* est utilisée pour générer des cartes de probabilité.

La Fig. 4 montre l'architecture AG adaptée de [9] pour la segmentation des pores. Nous définissons la *feature map* x au pixel $i \in \{1,\ldots,N\}$ dans la couche $l \in \{1,\ldots,L\}$ comme $x_i^l \in \mathbb{R}^{F_l}$, où F_l est la nombre de *feature map* dans la couche l. Un coefficient d'attention $\alpha_i^l \in [0,1]$ est calculé par l'AG pour identifier les ROI. La sortie de AG est une multiplication élément par élément $\hat{x}^l = \alpha_i^l x_i^l$.

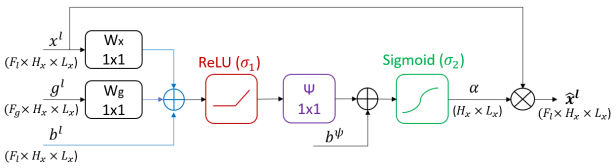

FIGURE 4 – AG adapté de [9]. H_x, L_x sont la hauteur et la largeur du *feature map* x.

Les *feature maps* sont progressivement sous-échantillonnées dans l'encodeur pour capturer un large champ récepteur. Les *features* de la couche $(l+1)$ identifient l'emplacement grossier des objets cibles, et ces emplacements approximatifs peuvent servir de base $g^l \in \mathbb{R}^{F_g}$ pour fournir l'information globale au x_i^l afin de ne se concentrer que sur des ROI. Les coefficients d'attention additifs α_i^l sont calculés comme suit :

$$\alpha_i^l = \sigma_2(q_{att}^l(x_i^l, g_i^l; \Theta_{att})) \quad (1)$$

$$q_{att}^l = \psi^T(\sigma_1(W_x^T x_i^l + W_g^T g_i^l + b_i^l)) + b_i \quad (2)$$

Les transformations linéaires $W_x \in \mathbb{R}^{F_l \times F_l}$, $W_g \in \mathbb{R}^{F_g \times F_l}$, $\psi \in \mathbb{R}^{F_l \times 1}$, et les biais $b_i^l \in \mathbb{R}^{F_l}$, $b_i \in \mathbb{R}$ forment le jeu de paramètres. W_g et W_x garantissent l'addition de x^l et g^l pour apprendre les ROI. σ_1 est la fonction *ReLU* pour la non-linéarité et σ_2 est l'activation *sigmoïde* pour la normalisation. $\{W_x, W_g, \psi\}$ sont implémentées comme 1×1 convolution.

La fonction de coût *Dice généralisé* (GDL) [10] est utilisé pour résoudre le problème de déséquilibre du au fait que seules certaines petites régions de l'image sont annotées.

$$GDL = 1 - 2\frac{\sum_{l=0}^{1} w_l \sum_{n=1}^{N} p_n g_n}{\sum_{l=0}^{1} w_l \sum_{n=1}^{N} p_n + g_n}; w_l = \frac{1}{\sum_{n=1}^{N} g_n} \quad (3)$$

où $VT = \{v_n | v_n \in \{0,1\}\}$ est la VT pour une image de N pixels, et $P = \{p_n | p_n \in [0,1]\}$ est la carte de probabilité de la segmentation automatique. Le poids w_l permet une invariance aux différentes propriétés du jeu d'annotation.

3 Etudes expérimentales

Métriques d'évaluation. Le Dice (DSC) est utilisé pour évaluer la segmentation au niveau pixel, tandis que le rappel (Rec_p) et la précision (Pre_p) portent sur la détection des pores :

$$Rec_p = VP_p/RP; \quad Pre_p = VP_p/PP$$

où un pore dans la VT est correctement détecté (VP_p) s'il a une intersection non nulle avec un pore segmenté automatiquement. Nous appelons RP le nombre de pores dans la VT, et PP le nombre de pores dans la segmentation automatique.

Implémentation. Les expériences ont été menées sur le jeu de données décrit dans la Sec. 2. Les opérations suivantes ont été aléatoirement appliquées pour l'augmentation de données : retournement horizontal/vertical, rotation, changement de luminosité et de contraste, ajout de bruit gaussien et déformation élastique. Nous avons utilisé une *validation croisée à 4 blocs* avec 13 sujets pour l'entraînement, 2 pour la validation, étant donné que les 2 sujets pour le test ne servent pas pour l'apprentissage. Avec les 2 modèles avec les meilleurs DSC et Rec_p, nous prenons la moyenne des cartes de probabilités, et la seuillons à 0,4 pour générer la carte binaire de segmentation automatique. Le taux d'apprentissage est initialisé à 3×10^{-4}, puis diminue selon une stratégie adaptative. L'optimiseur Adam est utilisé et la taille de batch de 32 pour 500 époques.

Résultats expérimentaux. Nous comparons la méthode proposée avec la méthode variationnelle développée dans [11], avec les réseaux de neurones supervisés [6, 9] ou non [12]. Nous avons délimité manuellement un masque périphérique pour ne conserver que la région de la papille pour [11] et [12]. Nous observons (Fig. 5) que les approches variationnelle et non supervisée sont sensibles aux bruits et aux artefacts, tandis que celles supervisées sont plus robustes. En outre, la méthode proposée est capable de prédire plus de pores candidats dans les régions à faible contraste et dans les régions voisines. L'analyse visuelle des coefficients d'attention (Fig. 6) illustre que notre réseau offre des performances particulièrement bonnes grâce aux AGs dans les couches les plus profondes.

La méthode proposée est robuste pour identifier les vrais pores avec une valeur Rec_p élevée (Table 1). En revanche, les