Label Refinement Network from Synthetic Error Augmentation for Medical Image Segmentation

Apélété Adodo SOSSOU and Raphaël Faure

Ecole Normale Supérieure de Paris-Saclay

Résumé

Les réseaux de neurones convolutifs (CNNs) représentent l'état de l'art pour de nombreuses tâches de segmentation d'images biomédicales. Parmi eux, le U-Net s'est imposé comme une référence grâce à ses connexions par sauts, qui améliorent précision et robustesse. Toutefois, ces modèles, y compris le U-Net, peinent souvent à capturer pleinement les informations structurelles des objets, ce qui entraîne des erreurs, notamment pour les structures tubulaires allongées comme les voies respiratoires.

Cet article propose une méthode pour intégrer implicitement les informations structurelles dans les CNN en ajoutant une étape de raffinement des labels. Elle repose sur la génération d'erreurs synthétiques dans les segmentations, utilisées pour entraîner un réseau de raffinement à corriger ces erreurs. Un réseau de simulation d'apparence réduit les écarts visuels entre erreurs synthétiques et segmentations réelles, renforçant ainsi la généralisation.

La méthode a été validée sur deux cas d'usage : la segmentation des voies respiratoires (CT thoracique) et des vaisseaux cérébraux (CTA 3D). Les résultats, comparés à des baselines comme U-Net, DoubleU-Net, SCAN et Post-DAE, montrent des performances supérieures. Une étude d'ablation a confirmé la pertinence des composants, et des expériences en apprentissage semi-supervisé ont démontré son efficacité même avec des données partiellement annotées.

Explications de l'article

Problèmes liés aux modèles actuels

Avant d'aborder les limites des modèles classiques tels que U-Net, il est important de rappeler ce qu'est le CT scan (tomodensitométrie) car c'est la méthode d'imagerie médicale utilisée ici. Cette technique exploite les rayons X pour produire des images 2D détaillées de l'intérieur du corps, lesquelles sont ensuite combinées pour créer un modèle volumétrique 3D. Elle permet de distinguer les différentes régions des zones examinées en fonction de leur niveau d'absorption des rayons X. Cela offre ainsi une visualisation précise des structures internes et est couramment utilisée pour analyser les poumons et les vaisseaux cérébraux, deux structures à morphologie arborescente étudiées dans ce travail.

Les arbres bronchiques et les vaisseaux

cérébraux présentent des caractéristiques similaires : des branches continues avec des bifurcations complexes et des terminaisons de faible volume. Ces spécificités posent deux défis majeurs pour les modèles comme U-Net. D'une part, la continuité des branches est une difficulté, les modèles ayant du mal à détecter des structures à la fois continues et hautement variables. D'autre part, les petites branches, souvent peu visibles en raison de leur faible volume et du contraste limité du CT scan, sont fréquemment ignorées ou mal segmentées.

Ces limitations soulignent le besoin de solutions avancées pour mieux traiter les particularités des structures arborescentes des voies respiratoires et des vaisseaux cérébraux.

Méthode de résolutions

Pour résoudre les limitations des modèles classiques de segmentation, les auteurs proposent d'ajouter un modèle de raffinement après le modèle de segmentation. Ce modèle corrige les erreurs classiques en s'appuyant sur les résidus et permet d'imposer des contraintes structurelles, telles que la continuité, qui ne sont pas explicitement encodables dans des modèles génératifs comme U-Net.

L'entraînement du modèle de raffinement repose sur deux types d'erreurs : les erreurs de segmentation initiales, générées par le modèle de segmentation, et des erreurs synthétiques, artificiellement injectées dans les segmentations de vérité terrain (g) pour produire des segmen- tations synthétiques (x_s). Les erreurs synthétiques incluent des branches terminales manquantes, contrôlées par le paramètre p_{a1}, et des discontinuités dans les branches, générées avec un taux p_{a2}. Ces erreurs imitent des défis usuels en segmentation médicale, comme les faibles résolutions ou contrastes, et permettent au modèle de raffinement d'apprendre à les corriger.

Pour rendre ces erreurs synthétiques plus réalistes, un apprentissage adversarial est utilisé via un simulateur d'apparence (LASN). Ce simulateur ajuste les segmentations synthétiques x_s pour produire des segmentations ajustées $\hat{x}_s$ proches des segmentations réelles x, tout en préservant les erreurs synthétiques. L'apprentissage adversarial combine deux composants : un simulateur f_a, qui ajuste l'apparence de x_s, et un discriminateur D, qui distingue les segmentations réelles x des segmentations ajustées $\hat{x}_s$. La perte adversariale est définie comme suit :

$$\mathcal{L}_{\text{adv}}(f_a, D) = \mathbb{E}_x[\log D(x)] + \mathbb{E}_{\hat{x}_s}[\log(1 - D(\hat{x}_s))].$$

f_a est optimisé pour minimiser L_{adv}, tandis que D la maximise. Une perte d'identité L_{dc} basée sur Dice est ajoutée pour garantir que les erreurs synthétiques soient préservées. La perte totale est donnée par :

$$f_a^* = \arg\min_{f_a} \left(\max_D \mathcal{L}_{\text{adv}}(f_a, D) + \lambda \mathcal{L}_{\text{dc}}(\hat{x}_s, x_s) \right),$$

où λ équilibre la préservation des erreurs synthétiques et l'ajustement d'apparence. Intuitivement, le générateur f_a est entraîné pour minimiser une combinaison de la perte adversariale $\mathcal{L}_{\text{adv}}$, maximisée par le discriminateur D, et d'une contrainte de cohérence $\mathcal{L}_{\text{dc}}$. Le min-max reflète la compétition entre f_a, qui cherche à tromper D, et D, qui vise à distinguer les exemples générés des données réelles. Les auteurs ici utilisent un U-Net pour le modèle de raffinement et optimisent avec la fonction de perte L_{dc}, garantissant des segmentations ajustées réalistes et utiles pour l'entraînement.

Pour évaluer les performances, plusieurs métriques sont utilisées. Le Dice coefficient mesure le recouvrement voxel par voxel entre la segmentation prédite Y et la vérité terrain G :

$$\text{Dice} = \frac{2|Y \cap G|}{|Y| + |G|}.$$

Les métriques spécifiques aux structures arborescentes permettent d'évaluer la qualité des segmentations en analysant des aspects essentiels à ces structures complexes.

La complétude des lignes médianes mesure la proportion de segments correctement détectés, en comparant les lignes médianes prédites à celles de la vérité terrain. La fuite des lignes médianes quantifie les fausses prédictions, en évaluant la proportion de segments incorrectement détectés dans l'arrière-plan. Enfin, le nombre de lacunes évalue les discontinuités dans les lignes médianes détectées en comptant les composantes connexes supplémentaires par rapport à la vérité terrain.

L'intérêt de ces métriques est qu'elles permettent une évaluation complète, mesurant à la fois la qualité globale et les caractéristiques spécifiques des structures arborescentes s'adaptant ainsi aux cas étudiés.

Données utilisées

Les expériences de segmentation ont été réalisées sur deux ensembles de données distincts : des scans CT thoraciques pour la segmentation des voies respiratoires et des scans CTA cérébraux pour la segmentation des vaisseaux sanguins.

1. **Données de scans CT thoraciques** : Les données proviennent d'une étude rétrospective réalisée sur des patients pédiatriques (âgés de 6 à 17 ans) atteints de fibrose kystique. Elles incluent 178 scans CT basse dose acquis en apnée à pleine inspiration. Les dimensions des coupes sont 512 × 512 pixels, avec un nombre de coupes variant entre 200 et 1000 par scan. La taille des voxels dans le plan est comprise entre 0,35 et 0,65 mm, et l'épaisseur des coupes entre 0,75 et 1,0 mm.

   Un sous-ensemble de 65 scans est annoté pour les lumières bronchiques à l'aide du logiciel LungQ, avec des corrections manuelles effectuées par des experts. Parmi ces scans annotés, 41 ont été utilisés pour le test, et les 24 restants ont été répartis en trois ensembles aléatoires pour l'entraînement (20 scans) et la validation (4 scans). Les 113 scans non annotés ont été utilisés comme données non labélisées pour l'apprentissage semi-supervisé.

2. **Données de scans CTA cérébraux** : Les scans CTA proviennent du registre MR CLEAN, qui collecte des données de patients traités pour des AVC ischémiques aigus dans 19 hôpitaux des Pays-Bas. Cet ensemble inclut 69 scans CTA, chacun ayant une taille de voxel dans le plan comprise entre 0,4 et 0,68 mm et une épaisseur de coupe entre 0,5 et 1,5 mm.

   Sur ces scans, 9 disposent d'annotations complètes des lignes médianes des vaisseaux, et 40 ont des annotations partielles sur un sous-volume de 140 × 140 × 140 voxels. Les 20 scans restants n'ont pas d'annotations et ont été utilisés pour l'apprentissage semi-supervisé. Pour le test, 2 scans annotés complets et 20 sous-volumes annotés ont été sélectionnés. Les données restantes ont été réparties en ensembles d'entraînement (7 scans complets et 14 sous-volumes) et de validation (6 sous-volumes).

Ces ensembles variés, comprenant à la fois des données annotées et non annotées, permettent d'entraîner et d'évaluer des modèles de segmentation robustes pour

des structures complexes comme les voies respiratoires et les vaisseaux cérébraux.

Résultats expérimentaux et conclusion

Une évaluation a été réalisée en comparant cette méthode avec le U-Net de base et plusieurs autres techniques de raffinement de segmentation, telles que DoubleU-Net, SCAN, Post-DAE, DVAE et le U-Net entraîné avec la clDice loss. De plus, une étude d'ablation a été menée pour analyser l'impact de chaque composant de la méthode. (cf. Figure 1 - Annexe).

En comparant la méthode proposée avec d'autres approches, intégrant le réseau de raffinement et le réseau d'ajustement d'apparence (LASN), des performances supérieures ont été observées en termes de précision de segmentation. Cette méthode a permis une meilleure continuité des structures segmentées, notamment pour les branches terminales des voies respiratoires et les petites branches des vaisseaux cérébraux. Elle a également réduit de manière significative les discontinuités et les erreurs de segmentation par rapport au U-Net de base, tout en corrigeant de manière plus robuste les erreurs structurelles, notamment face à SCAN et Post-DAE.

Par ailleurs, une étude d'ablation a mis en évidence l'importance des différents composants de la méthode. Trois variantes ont été testées. Sans l'ajout d'erreurs synthétiques (LR), les résultats montrent que la capacité du modèle à corriger les discontinuités est moindre. Avec des erreurs synthétiques mais sans LASN (LR+Syn), bien que les erreurs synthétiques aient un impact positif, l'absence du LASN limite la précision, surtout pour les structures plus petites. Enfin, la configuration complète (LR+Syn+LASN) a obtenu les meilleures performances, confirmant que l'ajustement d'apparence des erreurs synthétiques est essentiel pour maximiser la généralisation. L'utilisation d'apprentissage semi-supervisé a lui aussi porté ses fruits avec une précision au-dessus du modèle avec les données labélisées.

Concernant les résultats spécifiques, pour les voies respiratoires (CT thoraciques), la méthode complète a permis une augmentation notable du score Dice par rapport au U-Net de base, avec une réduction des faux négatifs dans les branches terminales. Pour les vaisseaux cérébraux (CTA), les résultats montrent une meilleure préservation de la connectivité des structures vasculaires, avec une réduction significative des discontinuités dans les longues branches.

Ces résultats confirment l'efficacité de la méthode proposée pour améliorer les segmentations de structures médicales complexes. En intégrant des mécanismes de raffinement robustes et des ajustements d'apparence pour les erreurs synthétiques, cette approche offre une solution performante pour les applications de segmentation médicale.

Limites de la méthode proposée

Bien que la méthode décrite dans cet article soit novatrice et présente des résultats prometteurs, elle n'est pas exempte de limites. Ces dernières ouvrent des perspectives intéressantes pour des améliorations futures et méritent une attention particulière. Voici quatre limitations principales :

1. Dépendance à l'expertise humaine pour la conception des erreurs

La méthodologie repose sur une analyse approfondie des erreurs typiques générées par les segmentations initiales. Ce

processus nécessite une intervention humaine pour identifier et concevoir les erreurs synthétiques pertinentes (comme les branches manquantes ou les discontinuités). Cette étape exige une expertise technique et médicale avancée, ce qui limite l'automatisation et l'évolutivité de la méthode. Par conséquent, son application à de nouveaux domaines ou ensembles de données nécessitera des ajustements manuels, rendant la méthode coûteuse et peu pratique pour des déploiements rapides ou à grande échelle.

2. Applicabilité limitée aux structures arborescentes

La méthode proposée a été spécifiquement conçue pour des structures arborescentes, comme les voies respiratoires et les vaisseaux sanguins, qui présentent des branches hiérarchiques et continues. Bien que ces résultats soient impressionnants, son application à d'autres formes anatomiques reste incertaine.

Pour des structures complexes mais non arborescentes, telles que les tumeurs ou les organes pleins, les erreurs synthétiques générées (comme les branches manquantes) ne reflètent pas les erreurs typiques de ces géométries. De plus, les métriques utilisées, comme la continuité et la complétude des branches, ne s'adaptent pas facilement à des formes volumétriques ou irrégulières.

Ainsi, cette limitation interroge la **versatilité** de la méthode, qui nécessiterait des ajustements importants pour s'appliquer à des tâches de segmentation impliquant des structures aux propriétés géométriques variées.

3. Prise en compte limitée des faux positifs

La méthode met l'accent sur la correction des faux négatifs, notamment les branches manquantes et les discontinuités. Cependant, elle aborde de manière marginale le problème des faux positifs, qui peuvent également affecter la qualité des segmentations. Les faux positifs apparaissent souvent sous forme de fragments isolés ou de « bruit » dans les données segmentées. Bien qu'ils soient moins fréquents dans les structures arborescentes, leur présence peut diminuer la précision globale et compliquer l'interprétation clinique. Une approche équilibrée qui traite à la fois les faux positifs et les faux négatifs serait préférable pour maximiser la robustesse de la méthode.

4. Fiabilité des erreurs synthétiques par rapport aux erreurs réelles

Le succès de la méthode repose fortement sur la capacité des erreurs synthétiques à imiter les erreurs réelles des segmentations initiales. Cependant, il existe toujours un risque que les erreurs générées ne capturent pas parfaitement la complexité des erreurs présentes dans des données réelles. Si les erreurs synthétiques ne sont pas suffisamment représentatives, le réseau de raffinement pourrait ne pas apprendre à corriger les erreurs les plus critiques. Cette limitation met en évidence la nécessité de valider soigneusement la correspondance entre les erreurs synthétiques et les erreurs observées en pratique, ainsi que d'explorer des moyens de rendre les erreurs simulées plus réalistes.

5. Sensibilité aux hyperparamètres

La méthode proposée repose sur plusieurs hyperparamètres critiques, tels que les taux d'injection des erreurs synthétiques (p_{a1} pour les branches manquantes et p_{a2} pour les discontinuités) et les poids associés aux différentes fonctions de perte (par exemple, la pondération entre la

perte Dice et la perte adversariale). Ces hyperparamètres jouent un rôle essentiel dans le succès de l'apprentissage du réseau de raffinement.

Toutefois, cette dépendance peut poser des problèmes. Une configuration inadéquate des hyperparamètres pourrait :

- Introduire des erreurs synthétiques non représentatives des erreurs réelles, limitant l'efficacité de l'apprentissage.
- Déséquilibrer les objectifs d'optimisation, par exemple en donnant trop d'importance à l'apparence des erreurs au détriment de leur correction structurelle.
- Rendre la méthode difficile à adapter à d'autres jeux de données, où les distributions des erreurs peuvent différer.

En pratique, l'optimisation des hyperparamètres nécessite des essais et des ajustements manuels approfondis, ce qui peut être coûteux en termes de temps et de ressources computationnelles. Cela limite l'accessibilité de la méthode pour des utilisateurs sans expertise avancée en apprentissage automatique ou des environnements où les ressources sont limitées.

Conclusion

Cet article propose une méthode innovante pour améliorer la segmentation des structures arborescentes telles que les voies respiratoires et les vaisseaux cérébraux. En utilisant des **erreurs synthétiques** et un **réseau de simulation d'apparence (LASN)**, la méthode corrige efficacement les erreurs structurelles et améliore la précision des segmentations.

Les résultats montrent une amélioration notable des scores Dice et des métriques structurelles par rapport aux méthodes existantes. La robustesse de la méthode a également été démontrée dans des configurations supervisées et semi-supervisées.

Cependant, cette approche pourrait être améliorée en automatisant la génération des erreurs synthétiques et en l'adaptant à d'autres types de structures anatomiques. Elle constitue une avancée prometteuse pour la segmentation médicale.

Répartition

— Résumé et explications de l'article : *Raphaël Faure*

— Limites de l'article et conclusion : *Apélété Adodo SOSSOU*

Référence

- *Label refinement network from synthetic error augmentation for medical image segmentation.* Chen, Shuai, Garcia-Uceda, Antonio, Su, Jiahang, Tulder, Gijs van, Wolff, Lennard, van Walsum, Theo, et de Bruijne, Marleen. *Elsevier*, 2025.

Annexe.

Table 1
Results for airway segmentation. Average performance (standard deviation) over the results obtained from three random data splits. LR: simple label refinement network. LR+Syn(init): label refinement method with synthetic errors on initial segmentations. LR+Syn: label refinement method with synthetic errors on ground truth segmentations. LR+Syn+LASN: label refinement method with label appearance simulation network. ↑: significantly better than the U-Net baseline ($p < 0.05$). ↓: significantly worse than the U-Net baseline ($p < 0.05$). P-values are calculated by the paired two-sided Student's T-test (on the average results from the three data splits). Boldface: best results, or not significantly different from the best results.

Method	Dice	Completeness	Leakage	Gaps
U-Net baseline (Garcia-Uceda et al., 2021)	0.76 (0.05)	0.74 (0.12)	0.23 (0.19)	95.73 (47.94)
DoubleU-Net (Jha et al., 2020)	0.77 (0.05)↑	0.73 (0.11)	0.21 (0.18)	99.93 (48.11)
SCAN (Dai et al., 2018)	0.77 (0.05)↑	0.75 (0.11)↑	0.31 (0.23)↓	98.83 (48.81)
Post-DAE (Larrazabal et al., 2020)	0.76 (0.06)	0.74 (0.12)	0.23 (0.19)	94.35 (48.17)
DVAE (Araújo et al., 2019)	0.75 (0.06)	0.72 (0.12)	0.18 (0.17)↑	**93.68** (49.69)↑
U-Net + clDice (Shit et al., 2021)	0.78 (0.05)↑	**0.75** (0.11)↑	0.25 (0.18)	95.96 (49.28)
LR	0.76 (0.05)	0.74 (0.11)↑	0.23 (0.17)	94.90 (47.66)
LR+Syn(init)	0.77 (0.06)	0.73 (0.12)	0.19 (0.17)	94.92 (50.14)
LR+Syn	**0.79** (0.05)↑	0.73 (0.12)	**0.17** (0.17)↑	**93.54** (50.83)↑
LR+Syn+LASN (proposed)	**0.79** (0.05)↑	**0.75** (0.11)↑	0.20 (0.16)↑	**91.63** (48.63)↑

FIGURE 1 – Résultats.