Etude expérimentale de SegFormer, un concurrent sérieux à U-Net pour la segmentation d'images médicales

Théo SOURGET[1] Nouman Syed HASANY[1] Fabrice MÉRIAUDEAU[2] Caroline PETITJEAN[1]

[1]Univ Rouen Normandie, Université Le Havre Normandie, INSA Rouen Normandie, Normandie Univ, LITIS UR 4108, 76000 Rouen, France

[2]ICMUB UMR CNRS 6302, Université Bourgogne, 21000 Dijon, France

Résumé – Le modèle U-Net, introduit en 2015, s'est imposé comme l'état de l'art pour la segmentation des images médicales, avec ses variantes UNet++, nnU-Net, V-Net, etc. En 2021, les "vision transformers" se sont imposés en vision par ordinateur. Depuis lors, de nombreuses architectures basées sur des transformers ou des architectures hybrides (combinant des blocs convolutifs et des blocs de transformer) ont été proposées pour la segmentation d'images, remettant en cause la suprématie de U-Net. Les transformers pourraient-ils arriver à détrôner U-Net pour la segmentation d'images médicales ? Dans cet article, nous prenons l'une des architectures de transformers les plus populaires pour effectuer la segmentation, SegFormer, et nous la comparons à U-Net dans trois jeux de données d'images médicales publiques, englobant diverses modalités et organes : la segmentation de structures cardiaques dans des images ultrasons issues du challenge CAMUS, la segmentation de polype dans des images d'endoscopie et la segmentation d'instruments dans des images de coloscopie issues du challenge MedAI. Nous les comparons à la lumière de plusieurs critères (performances de segmentation, temps d'apprentissage), et nous montrons que SegFormer peut être un concurrent sérieux d'U-Net dans le domaine de la segmentation des images médicales.

Abstract – The U-Net model, which was introduced in 2015, has been widely recognized as the state-of-the-art architecture for medical image segmentation, and its variants such as UNet++, nnU-Net, and V-Net have also gained popularity. However, in 2021, vision transformers made a breakthrough in the computer vision field. As a result, many transformer-based or hybrid architectures (combining convolutional and transformer blocks) have been proposed for image segmentation, challenging the dominance of U-Net. In this paper, we investigate whether transformers could surpass U-Net in medical image segmentation. We compare SegFormer, one of the most popular transformer architectures for segmentation, with U-Net using three publicly available medical image datasets that include various modalities and organs. These datasets include the segmentation of cardiac structures in ultrasound images from the CAMUS challenge, segmentation of polyps in endoscopy images, and segmentation of instruments in colonoscopy images from the MedAI challenge. We compare the two models based on various metrics such as segmentation performance and training time, and we demonstrate that SegFormer can be a true competitor to U-Net. Thus, we believe SegFormer should be carefully considered for future medical image segmentation tasks.

1 Introduction

Depuis 2015, U-Net [12] est l'état de l'art en segmentation d'images médicales, avec ses variantes UNet++, nnU-Net [7], V-Net, etc. Cette domination a été remise en question par l'arrivée des transformers en 2021 [2]. En effet, s'inspirant des excellents résultats des transformers sur des problèmes de traitement du langage naturel, des architectures transformer ont été proposées pour des tâches de traitements d'images, en commençant par la classification d'images avec le Vision Transformer ou ViT[3]. Les transformers traitent les images comme une séquence de patchs (généralement de taille 16×16 pixels) et semblent particulièrement efficaces grâce au méchanisme d'attention qui permet de modéliser des interactions distantes entre les patchs, à la différence du champ réceptif réduit des noyaux convolutionnels.

La transition du ViT à une architecture transformer pour la segmentation d'images n'est pas évidente. Plusieurs architectures ont été proposées, comme "Segmenter Transformer" SETR [15], PVT [13] ou SegFormer [14], dont l'encodeur a une structure pyramidale afin d'imiter l'encodage d'un CNN. L'outil récent "Segment Anything" [9], qui a obtenu des résultats impressionnants sur les images naturelles, est également basé sur une architecture transformer. Concernant les images médicales, des architectures hybrides combinant convolution et transformer comme TransUnet [2], CATS [11] ou UNETR [5] ont été proposées [1, 6]. Ces modèles sont cependant généralement très complexes avec plusieurs dizaines de millions de paramètres et demandent ainsi beaucoup de temps pour être entraînés.

La question est maintenant de savoir si une architecture transformer peut être un vrai concurrent à U-Net pour la segmentation d'images médicales. Dans cette étude, nous choisissons SegFormer [14], une architecture légère pour la segmentation d'images, conçue afin d'éviter l'utilisation d'une décodeur complexe. Son efficacité, sa précision et sa fiabilité ont été démontrées sur plusieurs jeux de données comme ADE20K, Cityscapes et COCO-Stuff. En particulier, SegFormer obtient de meilleurs résultats que SETR [15] sur ADE20K qui était jusqu'alors le meilleur modèle sur ce jeu de données. Dans cet article, notre objectif est de comparer SegFormer, pré-entraîné ou non, par rapport à U-Net, à la fois en termes de précision et de temps de calcul afin de savoir si SegFormer est une alternative viable à U-Net pour la segmentation d'images médicales.

Dans la suite de l'article, nous détaillons l'architecture Seg-

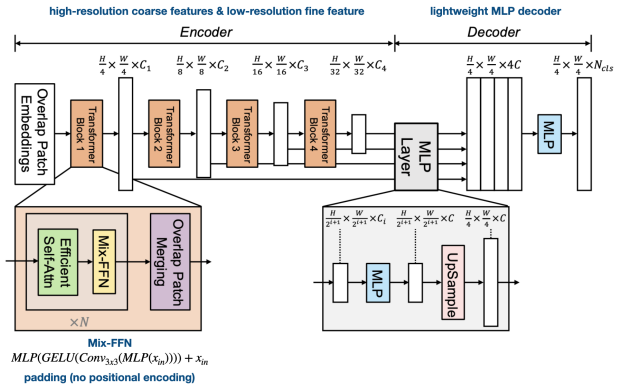

FIGURE 1 : Architecture SegFormer. Les deux principales parties sont le transformer hiérarchique et le décodeur MLP combinant les caractéristiques multi-résolution de l'encodeur.

Former. Puis nous présentons les trois jeux de données présentant différentes tâches de segmentation d'images médicales ainsi que le protocole expérimental. Enfin nous présentons des résultats à la fois quantitatifs et qualitatifs de segmentation.

2 SegFormer

SegFormer est composé de deux parties principales : un encodeur hiérarchique basé sur les transformers et un decodeur entièrement basé sur un perceptron multicouche (MLP) (figure 1). C'est le décodeur qui est un point clé de SegFormer, et qui permet d'obtenir une architecture plus légère en comparaison avec les autres architectures transformer pour la segmentation comme [15]. Ces aspects de SegFormer ont plusieurs avantages. Tout d'abord, l'encodeur génère des cartes de caractéristiques à plusieurs niveaux de résolution. Par ailleurs, l'encodeur se base sur l'opération "mix-FFN" qui consiste à appliquer une convolution 3×3 avec un "0-padding" directement sur la sortie d'un MLP et comme démontré par [8], l'utilisation d'opérations de convolution avec 0-padding permet au modèle d'apprendre une information positionnelle dans les images. Cette opération permet donc de remplacer l'encodage positionel utilisé habituellement par les autres architectures. Cela permet à SegFormer d'être robuste au résolutions d'images différant entre l'apprentissage et le test. De plus, comme le modèle est moins complexe que les autres architectures transformer, il demande moins de données pour être entraîné et peut être utilisé dans des applications temps-réel.

Enfin, la partie encodeur de l'architecture peut être mise à l'échelle de B0 à B5 en augmentant le nombre de filtres où la dimension de chaque bloc d'encodage. Dans cet article, nous utilisons la configuration B0 qui a 3.1M de paramètres.

3 Expérimentations

3.1 Datasets

Nous avons évalué les deux architectures sur trois jeux de données différents : CAMUS, Polyp et Instruments. Le jeu de données issu du challenge "Cardiac Acquisitions for Multi-structure Ultrasound Segmentation" (CAMUS)[10] contient les images 2D d'échographies cardiaques de 500 patients avec

TABLE 1 : Nombre d'images par jeu de données

Jeu de données	Apprentissage	Validation	Test
Camus	800	100	100
Polyp	640	160	200
Instrument	377	95	118

une segmentation manuelle en 4 classes (endocarde, épicarde, atrium et fond), chaque examen incluant des images 4-cavités et 2-cavités. Dans nos expérimentations, nous n'avons utilisé que les images de diastole pour l'apprentissage et le test. Polyp et Instruments sont deux jeux de données issus de MedAI: Transparency in Medical Image Segmentation. Polyp consiste à segmenter des polypes dans 1000 images d'endoscopie du jeu de données Kvasir-SEG. Instrument consiste à segmenter (de manière binaire) des outils dans des images de coloscopie issus du jeu de données Kvasir-Instrument qui contient 590 images.

3.2 Protocole

Nous comparons le modèle U-Net original (31M de paramètres) avec deux versions de SegFormer-B0 : un pré-entraîné sur ImageNet-1k et l'autre sans pré-entraînement. Nous utilisons également une version "allégée" de U-Net (U-Net Lite) où nous avons diminué le nombre de filtres par couche passant de [64,128,256,512,1024] dans [12] à [22,44,88,176,352] afin d'obtenir le même nombre de paramètres que SegFormer-B0 (3.7M). Pour le jeu de données CAMUS, la fonction de coût est l'entropie croisée, l'algorithme d'utilisation est Adam avec un taux d'apprentissage de 10^{-3} pour U-Net et 10^{-4} pour SegFormer. Pour Polyp et Instrument, la fonction de coût est une moyenne de l'entropie croisée et du Dice avec comme algorithme d'optimisation AdamW avec un taux d'apprentisage fixe de 10^{-4} pour les deux modèles.

Nous utilisons le modèle SegFormer-B0 issu de HuggingFace et les poids pré-entraînés sur Imagenet-1k lors de apprentissage par transfert. Dans un souci de reproductibilité, les codes utilisés par CAMUS pour U-Net et SegFormer sont disponibles en ligne : https://github.com/TheoSourget/UNet_CAMUS, https://github.com/TheoSourget/SegFormer_CAMUS.

Le tableau 1 montre le découpage des jeux de données pour l'apprentissage, la validation et le test. Les images sont redimensionnées en 256×256 pixels pour CAMUS et 224×224 pour Polyp et Instrument. Une augmentation de données est faite en appliquant des transformations géométriques aux images (miroir, rotation) à chaque epoch.

4 Résultats et Discussion

4.1 Précision de la segmentation

Les scores de Dice moyens pour chaque jeu de données sont présentés dans le tableau 2 et la Figure 4. Tout d'abord, concernant U-Net Lite, il est intéressant de noter que même si le nombre de paramètres est quasiment divisé par 10 par rapport à U-Net, seule les performances sur Polyp et Instrument sont significativement impactées. Cela corrobore les observations faites par [4, 10] montrant que des modèles plus simples

TABLE 2 : Scores de Dice moyens de U-Net et SegFormer sur 3 jeux de données : CAMUS, Polyp et Instrument. * indique que le score est significativement différent de celui de UNet (p<0.05). Représentation graphique en figure 4

		U-Net	U-Net Lite	SegFormer	SegFormer pré-entraîné
pré-entraîné ?		Non	Non	Non	Oui
# param		31M	3.7M	3.7M	3.7M
CAMUS	Endo	0.90	0.90	0.89	**0.91***
	Epi	0.80	0.79	0.81	**0.83***
	Atrium	0.83	0.84	0.81	0.85
Polyp		0.74	0.67	0.60	**0.83***
Instrum		0.79	0.75	0.82	**0.92***

(a) SegFormer (b) SegFormer pré-entraîné

FIGURE 2 : Évolution de la fonction de coût pendant l'apprentissage pour SegFormer avec et sans pré-entraînement sur ImageNet-1K

peuvent obtenir des résultats similaires à ceux de modèles plus complexes.

Sans surprise, SegFormer avec pré-entraînement est plus performant que SegFormer non pré-entraîné, parfois avec une marge importante. Le modèle pré-entraîné semble aussi converger plus rapidement, comme le montre l'évolution de la fonction de coût sur le jeu de données CAMUS (figure 2).

Enfin, le SegFormer pré-entraîné obtient des performances significativement supérieures à celles de U-Net pour toutes les régions segmentées, i.e. endocarde et épicarde de CAMUS, Polyp et Instrument, sauf pour l'atrium de CAMUS. La significativité statistique a été réalisée grâce à un test bilatéral de Wilcoxon (p<0.05).

On peut également observer une différence sur le comportement lors de l'apprentissage : alors que SegFormer est capable de prédire toutes les classes dès la fin de la première epoch, U-Net ne prédit que la classe "fond" pendant les deux premières epochs avant d'être capable de prédire les classes objet. Ce phénomène est illustré par la figure 3 avec les segmentations de U-Net, SegFormer et SegFormer avec pré-entraînement, après la première et cinquième epoch. Cette figure montre également le fort impact de l'utilisation de poids pré-entraînés pour SegFormer.

4.2 Temps d'apprentissage

Le tableau 3 montre la variation du temps d'apprentissage de chacun des modèles, en comparaison avec le temps d'apprentissage du U-Net. On peut observer que SegFormer est toujours

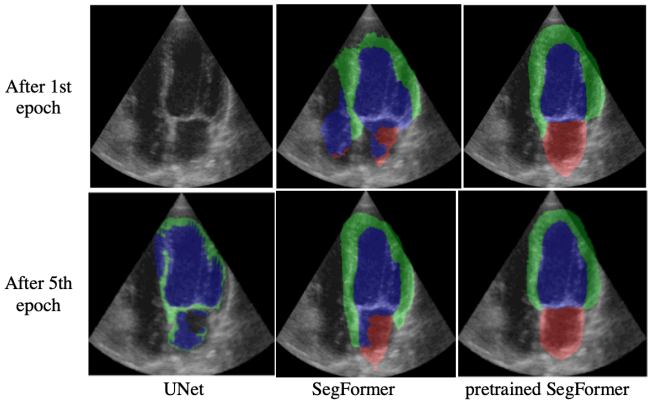

FIGURE 3 : Comparaison des résultats de segmentation entre U-Net, SegFormer et SegFormer pré-formé après la 1ère et la 5ème epoch

TABLE 3 : Diminution du temps d'apprentissage pour U-Net Lite, SegFormer et SegFormer pré-entraîné par rapport à celui de U-Net

modèle	U-Net Lite	SegFormer	SegFormer pré-entraîné	Epochs
CAMUS	-57.5%	-49.7%	-53.0%	50
Polyp	-51.2%	-62.3%	-65.1%	80
Instrum	-40.4%	-46.4%	-46.9%	80

plus rapide à entraîner que le U-Net originel et souvent plus rapide que la version "Lite", même sans pré-entraînement, pour le nombre d'epoch donné.

5 Conclusion

Dans cette étude, notre objectif était de comparer le modèle U-Net à une nouvelle architecture transformer nommée SegFormer. L'objectif sous-jacent est de commencer à évaluer si l'ascension récente des "vision transformer" va bouleverser l'état de l'art en segmentation d'images médicales, tant en termes de précision que de temps de calcul. Dans cette étude, nous nous sommes intéressés à la segmentation de structures cardiaques dans des images échographiques, de polypes en endoscopie et d'instruments dans des coloscopies. Sur chacune des tâches, SegFormer-B0 avec pré-entraînement a obtenu des résultats globalement meilleurs, que le modèle U-Net originel (sauf pour une région segmentée), et ce avec un temps d'entraînement moindre grâce à son architecture légère. Avec ces résultats, nous avons montré que même si les transformers ont généralement besoin de plus de données pour être entraînés, ils peuvent tout de même être utilisés pour de la segmentation d'images médicales, et fournir résultats compétitifs sur des jeux de données de taille limitée, grâce à l'apprentissage par transfert. Les résultats de cette étude préliminaire demandent cependant à être validés dans un cadre plus large, sur d'autres jeux de données.

Remerciements

Les auteurs remercient l'Agence Nationale de la Recherche pour le financement du projet MediSEG ANR-21-CE23-0013.

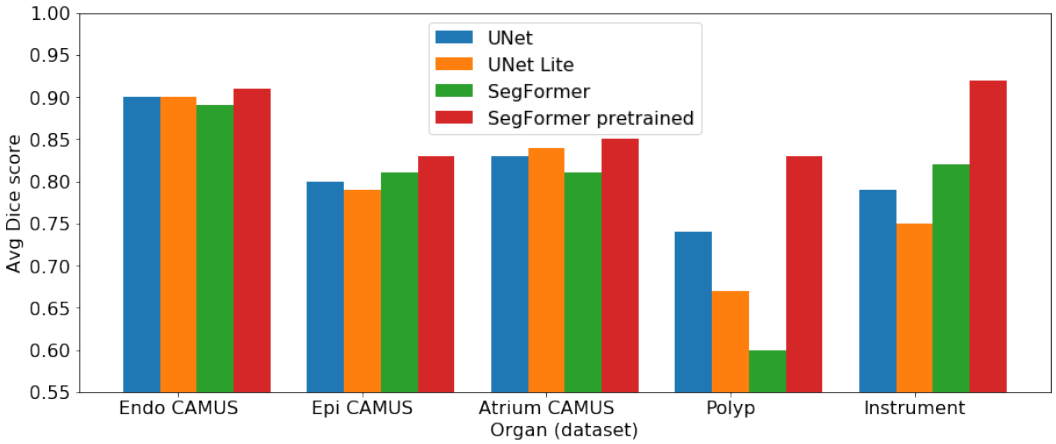

FIGURE 4 : Scores de Dice moyens de U-Net et SegFormer en test de 3 jeux de données : CAMUS, Polyp and Instrument. Correspond aux résultats du tableau 2.

Références

[1] Reza AZAD, Amirhossein KAZEROUNI, Moein HEIDARI, Ehsan Khodapanah AGHDAM, Amirali MOLAEI, Yiwei JIA, Abin JOSE, Rijo ROY et Dorit MERHOF : Advances in Medical Image Analysis with Vision Transformers : A Comprehensive Review, janvier 2023. arXiv :2301.03505 [cs].

[2] Jieneng CHEN, Yongyi LU, Qihang YU, Xiangde LUO, Ehsan ADELI, Yan WANG, Le LU, L. Alan YUILLE et Yuyin ZHOU : Transunet : Transformers make strong encoders for medical image segmentation. 2021.

[3] Alexey DOSOVITSKIY, Lucas BEYER, Alexander KOLESNIKOV, Dirk WEISSENBORN, Xiaohua ZHAI, Thomas UNTERTHINER, Mostafa DEHGHANI, Matthias MINDERER, Georg HEIGOLD, Sylvain GELLY, Jakob USZKOREIT et Neil HOULSBY : An image is worth 16x16 words : Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020.

[4] Adrian GALDRAN, André ANJOS, José DOLZ, Hadi CHAKOR, Hervé LOMBAERT et Ismail Ben AYED : The little w-net that could : State-of-the-art retinal vessel segmentation with minimalistic models, 2020.

[5] Ali HATAMIZADEH, Yucheng TANG, Vishwesh NATH, Dong YANG, Andriy MYRONENKO, A. Bennett LANDMAN, R. Holger ROTH et Daguang XU : Unetr - transformers for 3d medical image segmentation. *WACV*, pages 1748–1758, 2022.

[6] Kelei HE, Chen GAN, Zhuoyuan LI, Islem REKIK, Zihao YIN, Wen JI, Yang GAO, Qian WANG, Junfeng ZHANG et Dinggang SHEN : Transformers in medical image analysis. *Intelligent Medicine*, 3(1):59–78, février 2023.

[7] Fabian ISENSEE, Jens PETERSEN, André KLEIN, David ZIMMERER, Paul F. JAEGER, Simon KOHL, Jakob WASSERTHAL, Gregor KÖHLER, Tobias NORAJITRA, Sebastian J. WIRKERT et Klaus H. MAIER-HEIN : nnu-net : Self-adapting framework for u-net-based medical image segmentation. *CoRR*, abs/1809.10486, 2018.

[8] Md. Amirul ISLAM, Sen JIA et Neil D. B. BRUCE : How much position information do convolutional neural networks encode ? *CoRR*, abs/2001.08248, 2020.

[9] Alexander KIRILLOV, Eric MINTUN, Nikhila RAVI, Hanzi MAO, Chloe ROLLAND, Laura GUSTAFSON, Tete XIAO, Spencer WHITEHEAD, Alexander C. BERG, Wan-Yen LO, Piotr DOLLÁR et Ross GIRSHICK : Segment anything, 2023.

[10] Sarah LECLERC, Erik SMISTAD, Joao PEDROSA, Andreas OSTVIK, Frederic CERVENANSKY, Florian ESPINOSA, Torvald ESPELAND, Erik Andreas Rye BERG, Pierre-Marc JODOIN, Thomas GRENIER, Carole LARTIZIEN, Jan DHOOGE, Lasse LOVSTAKKEN et Olivier BERNARD : Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE Transactions on Medical Imaging*, 38(9):2198–2210, sep 2019.

[11] Hao LI, Dewei HU, Han LIU, Jiacheng WANG et Ipek OGUZ : Cats : Complementary cnn and transformer encoders for segmentation. *In IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2022.

[12] Olaf RONNEBERGER, Philipp FISCHER et Thomas BROX : U-net : Convolutional networks for biomedical image segmentation. *MICCAI*, page 234–241, 2015.

[13] Wenhai WANG, Enze XIE, Xiang LI, Deng-Ping FAN, Kaitao SONG, Ding LIANG, Tong LU, Ping LUO et Ling SHAO : Pyramid vision transformer : A versatile backbone for dense prediction without convolutions. *CoRR*, abs/2102.12122, 2021.

[14] Enze XIE, Wenhai WANG, Zhiding YU, Anima ANANDKUMAR, Jose M. ALVAREZ et Ping LUO : Segformer : Simple and efficient design for semantic segmentation with transformers. *CoRR*, abs/2105.15203, 2021.

[15] Sixiao ZHENG, Jiachen LU, Hengshuang ZHAO, Xiatian ZHU, Zekun LUO, Yabiao WANG, Yanwei FU, Jianfeng FENG, Tao XIANG, Philip H. S. TORR et Li ZHANG : Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. *CoRR*, abs/2012.15840, 2020.