OPENCLASSROOMS

Segmentation sémantique d'images sous-marines avec SegFormer

Project 7 : Développer une preuve de concept

Mestapha Oumouni

Mentor : Samir Tanfous

OpenClassrooms,

Ingénierie Machine Learning

14 octobre 2022

Table des matières

1 Introduction

L'objectif de ce rapport est d'élaborer une preuve de concept (poc) en réalisant une veille thématique via la recherche de sources pertinentes. Nous cherchons à résoudre un problème concret en testant de nouvelles approches conduisant à de meilleures solutions en termes de temps de calcul et de précision. Nous cherchons à résoudre un problème concret en testant de nouvelles approches conduisant à de meilleures solutions en termes de temps de calcul et de précision.

Le projet 6 du parcours concerne le problème de classification d'images avec CNN, un thème du domaine de la vision par ordinateur. Nous nous plaçons ici dans le cadre d'un autre sujet de ce domaine de la vision. Il s'agit de la segmentation sémantique d'images, un sujet clé avec neuf machines d'application, y compris la compréhension des scènes, l'imagerie médicale, la perception de la robotique, la vidéosurveillance...

Plusieurs modèles inspirés de ceux du problème de la classification CNN ont été développés pour performer la tache de la segmentation sémantique [7]. Notamment le modèle "SegFormer", un framework de transformer de pointe qui considère conjointement l'efficacité, la précision et la robustesse [6].

L'objectif est de comparer, de vérifier l'efficacité de ce modèle sur des images sous-marines. Ce type d'images représentent un défi pour la tâche de la segmentation. En effet, leur contenu visuel est entièrement différent en raison des catégories d'objets spécifiques au domaine, des motifs d'arrière-plan. Des proplèmes liés à la mauvaise apparence visuelle et des artefacts de distorsion et de la diffusion de la lumière. En outre, les classes des images contiennent des étiquettes avec des contours complexes (rugosité et variabilité spatiale fortes).

2 Segmentation d'images

La segmentation sémantique des images est un sous-domaine fondamental de la vision par ordinateur. Elle est liée à la classification des images car elle produit une prédiction de catégorie par pixel au lieu d'une prédiction au niveau de l'image.

La partition de l'image (ou image vidéo) consiste à étiqueter des régions spécifiques de celle-ci avec des classes correspondantes en fonction de ce qui est représenté. Cette tâche est parfois appelée prédiction dense, puisqu'on déduit des étiquettes pour chaque pixel, de sorte que chaque pixel est étiqueté avec la classe de sa région.

Il existe deux types de segmentation d'images :

— Segmentation sémantique : les pixel sont classés avec une étiquette.
— Segmentation d'instance : les pixels sont classés avec une différentation de chaque instance d'objet.

La segmentation d'image joue un rôle central dans un large éventail domaine d'application, par exemple :

— Analyse d'images médicales (par exemple : extraction des limites tumorales)
— Véhicules autonomes : comprendre l'environnement et repérer la surface navigable et piétonnière.
— La vidéosurveillance pour comprendre les scènes, la réalité augmentée et d'autre applications.

2.1 Résumé de l'état de l'art

Le modèle FCN est l'un des premier modèle fondamental de la segmentation sémantique [1], c'est est un réseau entièrement convolutif . Il effectue une classification pixel à pixel en se basant sur réseau d'encodeurs suivi d'un réseau de décodeurs (Figure1). Sous différents aspects, des améliorations du FCN ont été proposées (l'élargissement du champ réceptif) pour améliorer l'efficacité du réseau FCN. D'autres modèles initialement développés pour segmentation d'images biomédicales, qui s'inspirent FCN et modèles, par exemple U-Net [7] et V-Net [9].

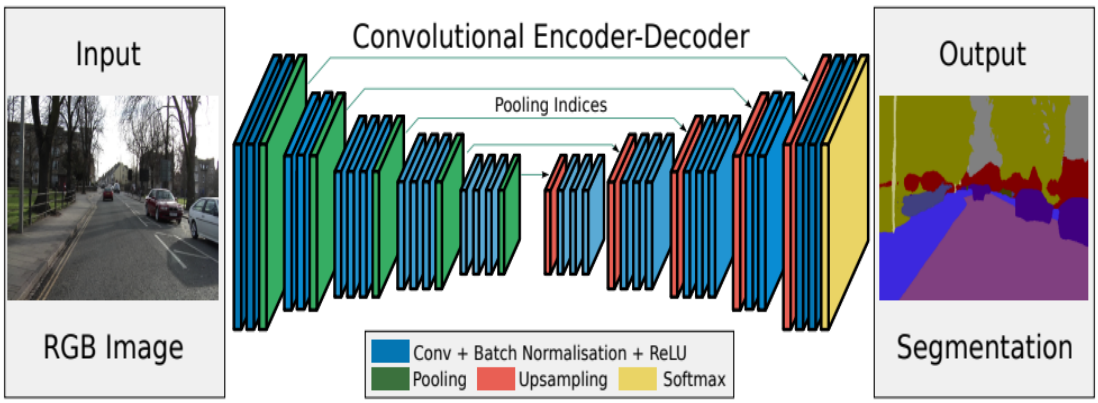


FIGURE 1 – Architecture de segmentation sémantique générale : un réseau d'encodeurs suivi d'un réseau de décodeurs.

Récemment, suite au succès des transformateurs aux sujet du traitement du langage naturel, ces derniers ont été introduis dans les sujets de la vision par ordinateur. Par exemple , le modèle VIT vision pour la classification des images [8]. Les auteurs dans [12] ont proposé SETR pour démontrer la faisabilité de l'utilisation de transformateurspour la tâche de la segmentation d'image. Malgré les bonnes performances, ViT a quelques limites, notamment l'absence d'une performance à plusieurs échelles et un coût exorbitant pour des grandes résolutions. Une extension naturelle de ViT avec des structures pyramidales (PVT) pour une prédiction dense à été proposé dans [10], afin de pallier ces limites avec des amélioration considérable pour une prédiction dense. Cependant, avec d'autres méthodes émergentes considèrent principalement la conception du codeur, et elles négligent la contribution du décodeur pour plus d'améliorations. Dans l'article [6] SegFormer, un cadre de transformateur de pointe pour la segmentation sémantique qui redéfinit à la fois l'encodeur et le décodeur (Figure. 1). Avec une conception d'encodeur

de transformateur hiérarchique et un décodeur All-MLP léger, le modèle fournit une représentation puissante avec moins de complexité informatique (Figure. 5).

2.2 Données d'images considérées

Les données considérées sont des images sous-marines qui ont été récupérées depuis la plate-forme Kaggle [2], plus d'informations sur les moyens d'acquisition et de notation dans [4]. L'ensemble des données contient 1635 images avec des annotations en pixels pour huit catégories d'objets et partagé sous deux ensembles. Un groupe d'images d'entrainement et de validation contenant 1525 images et un autre groupe d'images de test de nombre 110. Les images contiennent diverses configurations homme-robot et un ensemble

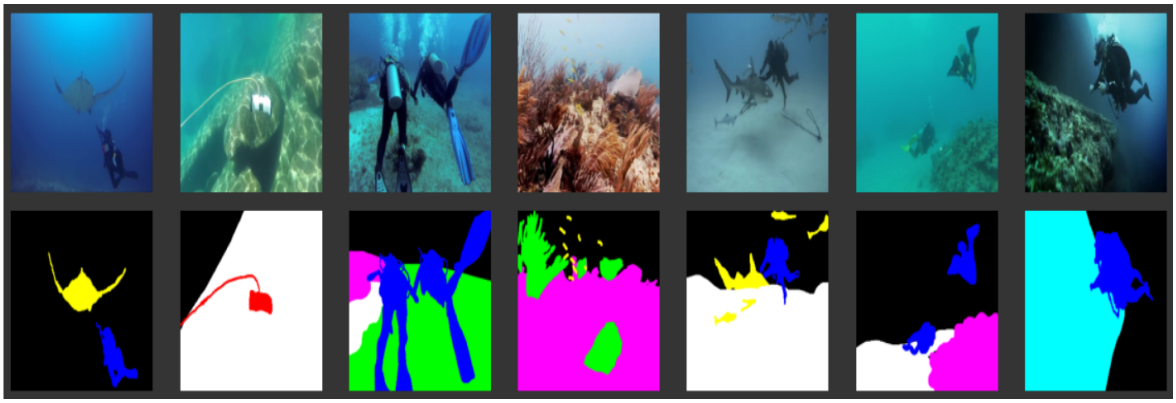


FIGURE 2 – Exemple d'images de dataset sous-marrines et leurs masques

diversifié de scènes sous-marines naturelles. La figure. 2 montre un exemple d'images ainsi que le masque correspondant. Le tableau 1 résume la population de chaque catégorie d'objets contenate dans l'ensemble des images sous-marine considérees, le nombre d'image contenant chaque catégorie. On peut remarquer cette distribution d'image par classe et un peu déséquilibrées.

Catégorie d'objet	RGB couleur	Code (couleur)	# d'image/catégorie
Fond (plan d'eau)	000	BW(noir)	1288
Plongeurs (humain)	001	HD(bleu)	405
Plantes, herbiers marins	010	PF(vert)	239
Épaves ou ruines	011	WR (ciel)	275
Robots (instruments)	100	RO (rouge)	101
Récifs, invertébrés	101	RI(rose)	1028
Poissons, vertébrés	110	FV(jaune)	1030
plancher, rochers	111	SR(blanc)	635

TABLE 1 – Catégories d'objets et couleur codes pour les annotations de pixels et nombre d'images par classe.

3 Modèle de segmentation : Unet et SegFormer

Nous considérons deux modèles de réseaux encodeur-décodeur de segmentation d'images pour les tester sur les images sous-marines. Le modèle base line U-net que nous allons comparer avec le modèle SegFormer.

3.1 Architecture U-net

U-Net est un réseau entièrement convolutif (FCN) où il ne contient que des couches convolutives et ne contient aucune couche dense. L'architecture d'un U-Net contient deux blocs : le premier est le chemin de contraction (aussi connu sous le nom d'encodeur), le deuxième chemain expansif, connu sous le nom décodeur. Les deux chemins sont reliés par un pont. Dans l'article original [7], le U-Net est décrit comme suit : (Figure 3). Le

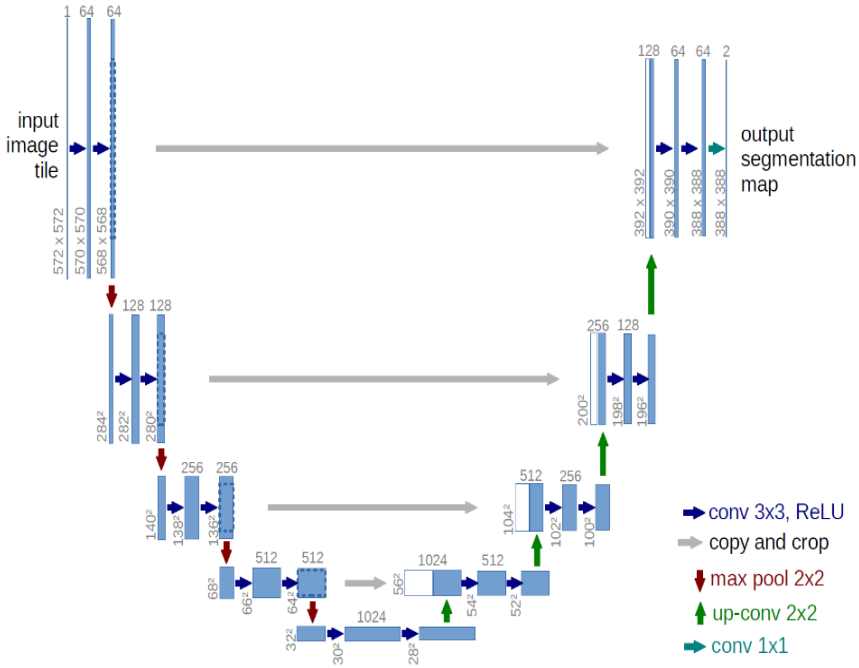


FIGURE 3 – U-Net architecture (image source [7]).

bloc d'encodeur sert à capturer les caractéristiques (features) de l'image et de réduire sa dimension afin diminuer le nombre de paramètres du réseau. Il s'agit de la répétition de deux couches de convolution 3x3, chaque couche est suivie d'une fonction d'activation Relu et de normalisation des lots (batch normalisation). Puis, une opération de max pooling 2x2 est appliquée pour réduire les dimensions spatiales

Le pont relie les deux réseaux d'encodeurs et de décodeurs et complète le flux d'informations. Il se compose de deux convolutions 3x3, où chaque convolution est suivie d'une fonction d'activation ReLU.

Le réseau décodeur est utilisé pour prendre la représentation abstraite et générer un masque de segmentation sémantique. Le bloc décodeur commence par une convolution

de transposition 2x2. Ensuite, il est concaténé avec la carte de caractéristiques de saut de connexion correspondante du bloc d'encodeur. Ces connexions de saut fournissent des fonctionnalités des couches précédentes qui sont parfois perdues en raison de la profondeur du réseau. Après cela, deux convolutions 3x3 sont utilisées, où chaque convolution est suivie d'une fonction d'activation ReLU.

La sortie du dernier décodeur passe par une convolution 1x1 avec activation sigmoïde. La fonction d'activation sigmoïde donne le masque de segmentation représentant la classification pixel par pixel.

3.2 Architecture de SegFormer

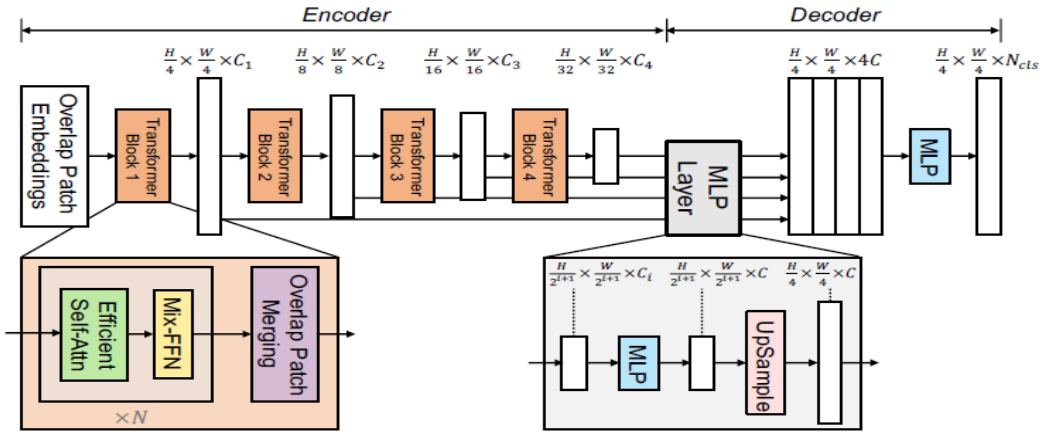
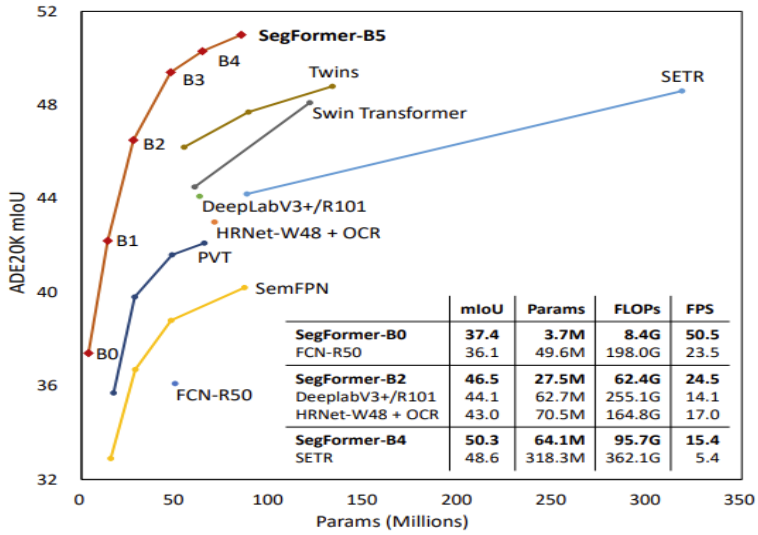


FIGURE 4 – SegFormer proposé se compose de deux modules principaux.

L'architecture du modèle SegFormer [6] est basée sur une architecture de transformateur inspirée du modèle ViT [8] avec deux modules d'encodeur-décodeur comme le montre la Figure. 4. Le premier est un encodeur de transformateur hiérarchique pour générer des features grossières à haute résolution. Le deuxième bloc est un décodeur All-MLP léger pour fusionner ces features multiniveaux afin de produire le masque final de la segmentation sémantique.

Contrairement à ViT qui ne peut générer qu'une carte d'entités à résolution unique, l'architecture du codeur SegFormer est de nature hiérarchique, il produit des caractéristiques multi-échelles de type CNN. Plus précisément, étant donné une image d'entrée avec une résolution de $H \times W \times 3$, le modèle effectue une fusion de patchs pour obtenir une carte de caractéristiques hiérarchique F_i avec un résolution de $\frac{H}{2^{i+1}} \times \frac{H}{2^{i+1}} \times C_i$, avec $i = 1, 2, 3, 4$. et $C_i \leq C_{i+1}$. Contrairement à d'autres décodeurs complexes, SegFormer applique un décodeur MLP simple qui agrège les informations de différentes couches et combine ainsi à la fois l'attention locale et l'attention globale pour rendre des représentations puissantes.

SegFormer est présenté dans [6] sous forme d'une série d'encodeurs Mix Transformer (MiT), MiT-B0 à MiT-B5, avec la même architecture mais des tailles différentes. MiT-B0 est le modèle léger pour une inférence rapide, tandis que MiT-B5 est le plus grand modèle pour des meilleures performances comme le montre la Figure 5.



	mIoU	Params	FLOPs	FPS
SegFormer-B0	**37.4**	**3.7M**	**8.4G**	**50.5**
FCN-R50	36.1	49.6M	198.0G	23.5
SegFormer-B2	**46.5**	**27.5M**	**62.4G**	**24.5**
DeeplabV3+/R101	44.1	62.7M	255.1G	14.1
HRNet-W48 + OCR	43.0	70.5M	164.8G	17.0
SegFormer-B4	**50.3**	**64.1M**	**95.7G**	**15.4**
SETR	48.6	318.3M	362.1G	5.4

FIGURE 5 – Performance vs Efficacité sur les données ADE20K [6].

4 Résultats et comparaison de Modèles

4.1 Résultats de Unet

Nous commençons par présenter les résultats du modèle (baseline) U-Net. Nous préparons les images, encoder les masques puis charger les couples (images, masques) par lots en utilisant des fonctions du module de Keras tensorflow.keras.utils. Le modèle Unet construit prend des images de taille (256,256,3) en entré.

On considère les métriques : accuracy et IoU appelée aussi coefficient de Jaccard. On entraine le modèle avec l'optimiseur Adam sur 30 epochs.

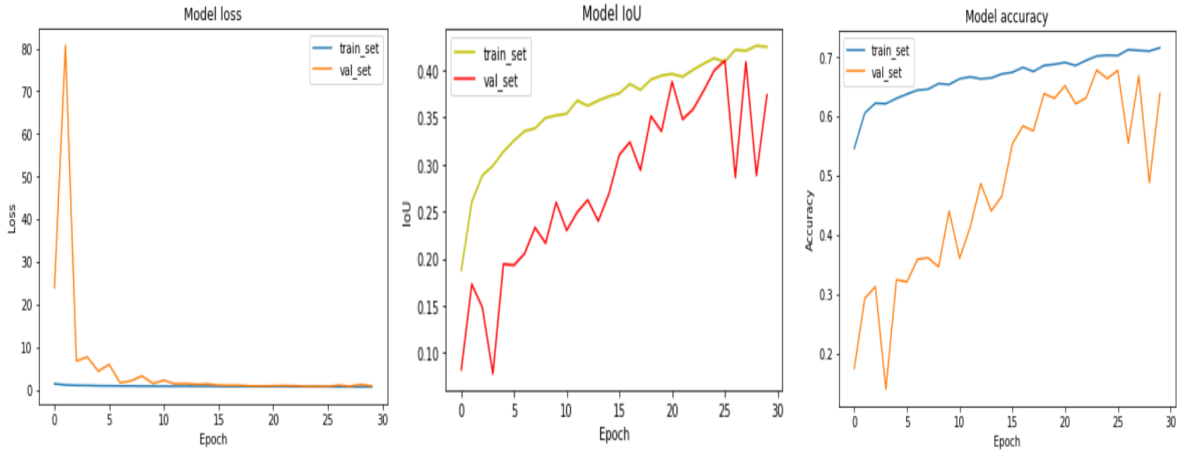


FIGURE 6 – Courbes des scores et d'erreurs du modèle U-Net.

La Figure 6 plot les courbes des scores de l'entrainement. Le tableau 2 résume le temps de d'ajustement et les scores sur les données de test. La Figure 7 montre un exemple de prédiction des masques sur des données de test.

temps d'entraînement (min)	43.35
mIoU (test)	0.24
Accuracy (test)	0.55

TABLE 2 – Scores et temps de fit sur les données test de Unet

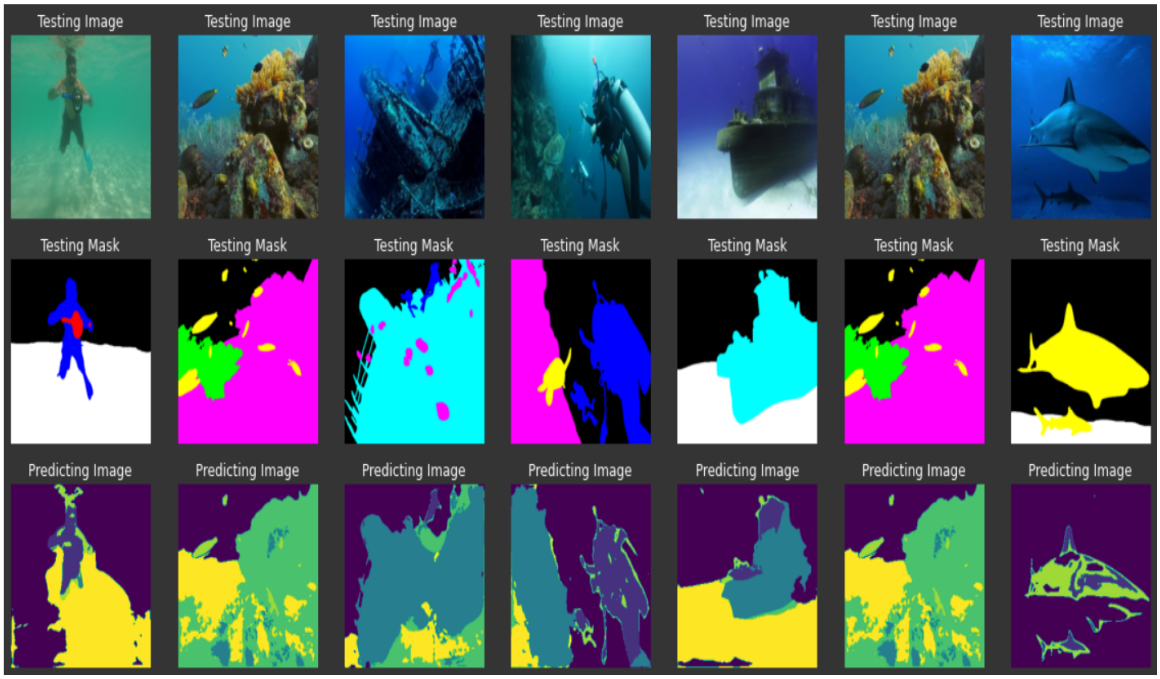


FIGURE 7 – Exemple de prédiction des masques par U-Net.

4.2 Résultas de SegFormer

Le modèle provient de la plateforme HuggingFace [3] pre-entrainé sur des données ADE20k et la version "mit-b0" :

```
model = SegformerForSemanticSegmentation.from_pretrained("nvidia/mit-b0",
                                                         num_labels=8,
                                                         id2label=id_to_label,
                                                         label2id=label_to_id,)
```

Malheureusement, l'extraction des features avec la fonction "feature_extractor" du modèle envois des mauvais codes pour les huits classes du dataset. Ainsi, on construit une classes qui charge les données où encode les masques avec des bonnes codes dans l'ensemble $\{0, 1, \ldots, 7\}$. Ensuite, on charge les données (images, masques) avec la primitive de DataLoader de pytorch.

On choisit la métrique d'évaluation, l'accuracy, IoU et la fonction de perte (loss), charger le modèle en précisant le nombre de classes souhaitées afin de procéder au fine-tuning avec des valeurs par défaut sur 20epochs. Le modèle utilse l'optimiseur AdamW, où nous utilisons le même taux d'apprentissage que celui rapporté dans [6].

Figure 8 plot les courbe des scores de l'entrainement, il montre une amélioration du modèle finetuné aussi que les scores sont meilleurs de ceux du modèle U-net. Cependant, le temps de l'entrainement de SegForm est important en comparaison avec Unet comme le montre le Tableau. 3. La Figure 9 montre un exemple de prédiction avec SegFormer des masques sur des données de test.

Le tableau de la figure 10 montre les scores IoU pour chaque classe sur les données de test. On peut remarquer que les scores sont bon except pour la classe box qui remplace la classe robot. qui ne figure pas dans la liste des classes des données ADE20k.

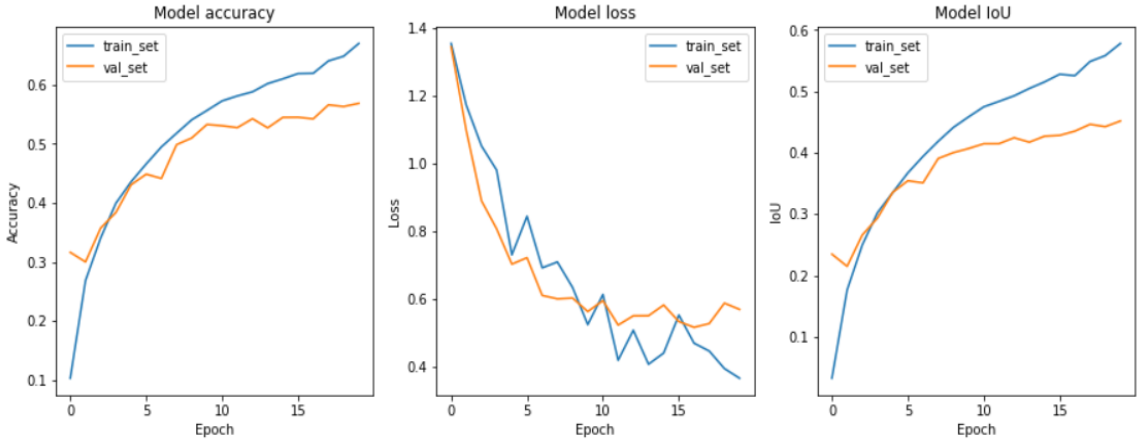


FIGURE 8 – Courbes des scores et de perte du modèle SegFormer.

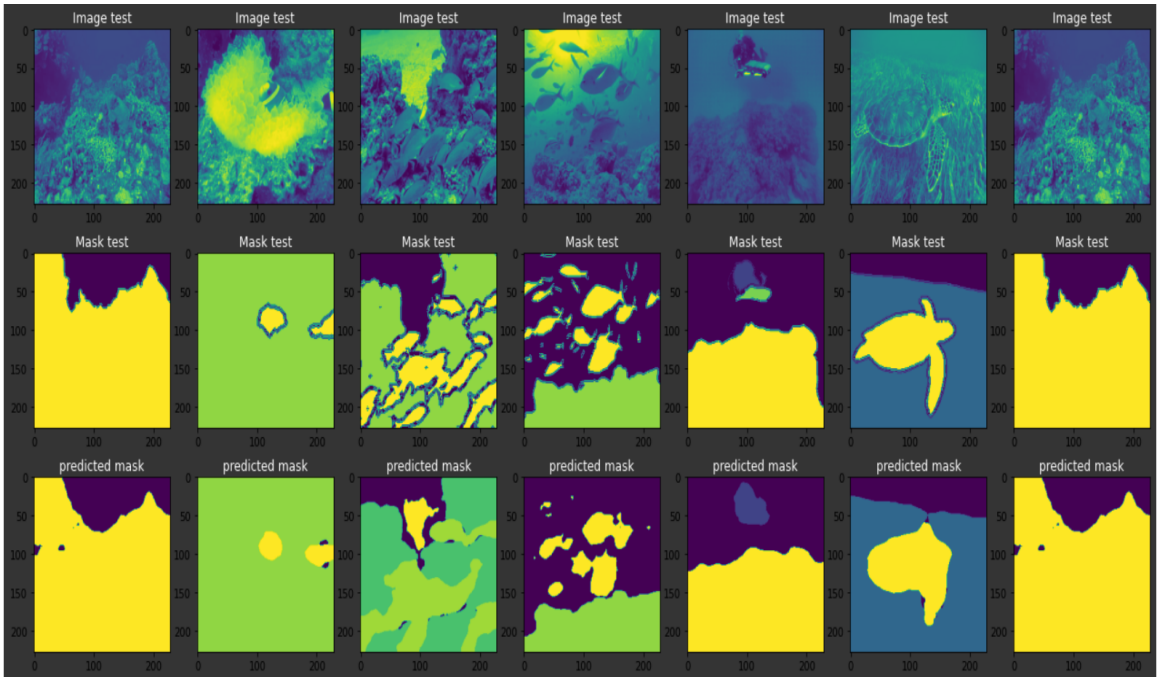


FIGURE 9 – Exemple de prédiction des masques par SegFormer.

temps d'ajustement (min)	126.98
mIoU test	0.60
Accuracy test	0.71

TABLE 3 – Scores et temps de fit sur les données test de SegFormer

per-category metrics:

	IoU
water	0.760330
person	0.518212
plants-grass	0.507251
wrecks-ruins	0.768283
box	0.093362
reefs-invertebrates	0.824704
animal	0.625970
sea-floor-rocks	0.746131

FIGURE 10 – IoU score par catégorie avec SegFormer.

5 Conclusion

SegFormer est un modèle de segmentation sémantique simple et puissante. L'architecture du modèle (codeur transformer hiérarchique plus et un décodeur All-MLP léger) lui permet à la fois une efficacité et une performances élevée. Cepdant, avec des images contenant des artéfacts et une faible visibilité, le modèle peut connaître une baisse de performance et d'efficacité.

Dans ce travail nous avons expérimenté SegFormer pour segmenter des images sous-marines, généralement avec des artefects de la lumière, des classes spécifiques et des bords complexes. Le modèle préentrainé sur des données ADE20k [13] et finetuné montre des scores meilleurs que le modèle U-net considéré comme méthode de base. Toutefois le temps d'ajustement est important pour le modèle SegFormer avec 3.7 millions de paramètres à apprendre en comparaiosn avec Unet.

Les scores du modèle SegFormer seront améliorés avec ces points suivant :

— pre-prossing spécifique d'images afin d'améliorer leur qualité
— exploiter l'augmentation des données
— enrichir la base de donnée par des images synthétiques
— prendre en compte l'asymétrie des catégories.

Références

[1] J. Long, E. Shelhamer, and T.Darrell. *Fully convolutional networks for semantic segmentation.* In CVPR, 2015.

[2] `https://www.kaggle.com/datasets/ashish2001/semantic-segmentation-of-`underwater-imagery-suim.

[3] `https://www.huggingface.co/nvidia/segformer-b0-finetuned-ade-512-512`.

[4] M. Islam, Ch. Edge, Y. Xiao, P. Luo, M. Mehtaz, Ch. Morse, Sadman. Enan, J. Sattar *Semantic Segmentation of Underwater Imagery : Dataset and Benchmark*. arXiv :2004.01241.

[5] S. Minaee, Y. Boykov, Fa. Porikli, A. Plaza, N. Kehtarnavaz, D. Terzopoulos . *Image Segmentation Using Deep Learning : A Survey* , arxiv : 2001.05566.

[6] E. Xie, W. Wang, Z. Yu, A. Anandkumar, Jose M. Alvarez, P. Luo *SegFormer : Simple and Efficient Design for Semantic Segmentation with Transformers*. arXiv :2105.15203.

[7] O. Ronneberger, P. Fischer, and T. Brox, *U-net : Convolutional networks for biomedical image segmentation,* in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.

[8] W. Wang, E. Xie, X. Li, De. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. *Pyramid vision transformer : A versatile backbone for dense prediction without convolutions.* arXiv, 2021.

[9] F. Milletari, N. Navab, and S.-A. Ahmadi, *V-net : Fully convolutional neural networks for volumetric medical image segmentation*, in 2016 Fourth International Conference on 3D Vision (3DV). IEEE, 2016, pp. 565–571.

[10] W. Wang, E. Xie, X. Li, D.Fan, K. Song, D.Liang, T. Lu, P.Luo, and .Shao ; *Pyramid vision transformer : A versatile backbone for dense prediction without convolutions.* arXiv, 2021.

[11] S. Zheng, J.Lu, H. Zhao, X. Zhu, Z.Luo, Y. Wang, Y.Fu, J. Feng, T. Xiang, P. Torr, et al.*Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers.* CVPR, 2021.

[12] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos *Image Segmentation Using Deep Learning : A Survey*; arXiv :2001.05566.

[13] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba.*Scene parsing through ade20k dataset.* In CVPR, 2017