Overview of some segmentation models

Réalisé par : OUGNI Imane

Table des matières

1. Introduction

La segmentation d'images est une tâche fondamentale en vision par ordinateur, visant à identifier et isoler différentes régions d'intérêt dans une image. Elle trouve des applications dans de nombreux domaines, tels que la médecine, l'urbanisme, l'agriculture et l'environnement.

Le projet Augmented Forest s'inscrit dans ce contexte et se concentre sur l'analyse automatique des images forestières pour mieux comprendre la couverture végétale, détecter les changements environnementaux et faciliter la gestion durable des ressources naturelles.
Dans ce rapport, nous présentons un aperçu de plusieurs modèles de segmentation d'images basés sur l'apprentissage profond, en expliquant leur architecture, leurs principes et leurs avantages, afin de fournir une base pour le benchmarking sur le dataset Augmented Forest.

2. La segmentation des images

La segmentation d'images est une étape fondamentale en vision par ordinateur. Elle consiste à diviser une image en plusieurs régions homogènes afin d'isoler les objets ou les zones d'intérêt qu'elle contient. L'objectif est de regrouper les pixels présentant des caractéristiques similaires (couleur, texture, intensité, position, etc.) pour obtenir une représentation plus structurée et interprétable de la scène.

En apprentissage profond, la segmentation sémantique permet d'attribuer à chaque pixel de l'image une classe. Ce type de traitement est particulièrement utile pour des environnements complexes, tels que les zones urbaines ou forestières, où la compréhension précise du contexte visuel est essentielle.

On distingue plusieurs formes de segmentation :

- Segmentation binaire : distingue un objet du fond (par exemple, détecter la forêt ou la chaussée).
- Segmentation sémantique : classe chaque pixel selon la catégorie d'objet à laquelle il appartient.
- Segmentation d'instance : différencie les objets appartenant à une même classe (par exemple, plusieurs arbres distincts).
- Segmentation panoptique : combine segmentation sémantique et segmentation d'instance pour offrir une compréhension globale de la scène.

Grâce aux réseaux de neurones convolutifs (CNN) et à leurs variantes plus avancées, les performances des modèles de segmentation se sont considérablement améliorées, permettant des résultats précis même sur des images complexes et réalistes.

Dans le contexte de l'étude des environnements forestiers, ces techniques permettent d'analyser efficacement la végétation et les différentes composantes du paysage, comme le montre le dataset Augmented Forest.

3. Augmented forest

Le projet Augmented Forest s'inscrit dans le domaine de l'analyse environnementale et de la vision par ordinateur appliquée à la foresterie. Il consiste à utiliser des techniques avancées de segmentation d'images pour identifier et classifier automatiquement différentes zones dans les images de forêt.

L'intérêt de ce type de projet est multiple :

- Suivi de la couverture forestière : Il permet de détecter les changements dans les forêts, tels que la déforestation, la régénération naturelle ou l'impact des activités humaines.
- Gestion durable des ressources : Les informations extraites peuvent aider les autorités et les chercheurs à planifier la conservation et la gestion des écosystèmes forestiers.
- Développement de modèles intelligents : Ce projet est un terrain idéal pour tester et comparer différents modèles de segmentation d'images, contribuant à l'avancement des techniques de vision par ordinateur.
- Applications réelles et sociales : En plus de l'aspect scientifique, ce type de projet peut avoir un impact direct sur l'environnement et la société, en améliorant la surveillance des forêts et la prise de décision écologique.

Ainsi, le projet Augmented Forest illustre parfaitement l'intérêt de combiner l'intelligence artificielle et la vision par ordinateur pour résoudre des problèmes complexes du monde réel, en particulier dans le domaine de la protection et de l'analyse des forêts.

4. Aperçu des modèles de segmentation

Au cours des dernières années, de nombreux modèles de segmentation d'images basés sur l'apprentissage profond ont été développés afin d'améliorer la compréhension des scènes visuelles complexes. Ces modèles diffèrent par leur architecture, leur capacité à extraire des caractéristiques multi-échelles et leur performance en termes de précision et de vitesse d'exécution.

4.1. U-Net

Le U-Net est un réseau de neurones convolutifs conçu initialement pour la segmentation d'images biomédicales, mais il est aujourd'hui largement utilisé dans d'autres domaines, notamment la segmentation d'images urbaines. Son nom provient de la forme en "U" de son architecture, qui relie une partie d'encodage (compression) à une partie de décodage (reconstruction).

L'architecture du U-Net est symétrique et se compose de trois grandes parties :

- **Chemin contractant (Encoder) :**

Cette partie a pour rôle d'extraire les caractéristiques importantes de l'image. Elle utilise des convolutions 3×3 suivies d'une fonction d'activation ReLU pour introduire la non-linéarité, puis des opérations de max pooling 2×2 pour réduire la taille spatiale de l'image tout en conservant les informations essentielles.

- **Bottleneck (Couche intermédiaire) :**

C'est la zone centrale du réseau, où les représentations de l'image sont les plus compressées et les plus abstraites. Elle relie directement l'encodeur au décodeur et capture les caractéristiques les plus importantes.

- **Chemin expansif (Decoder) :**

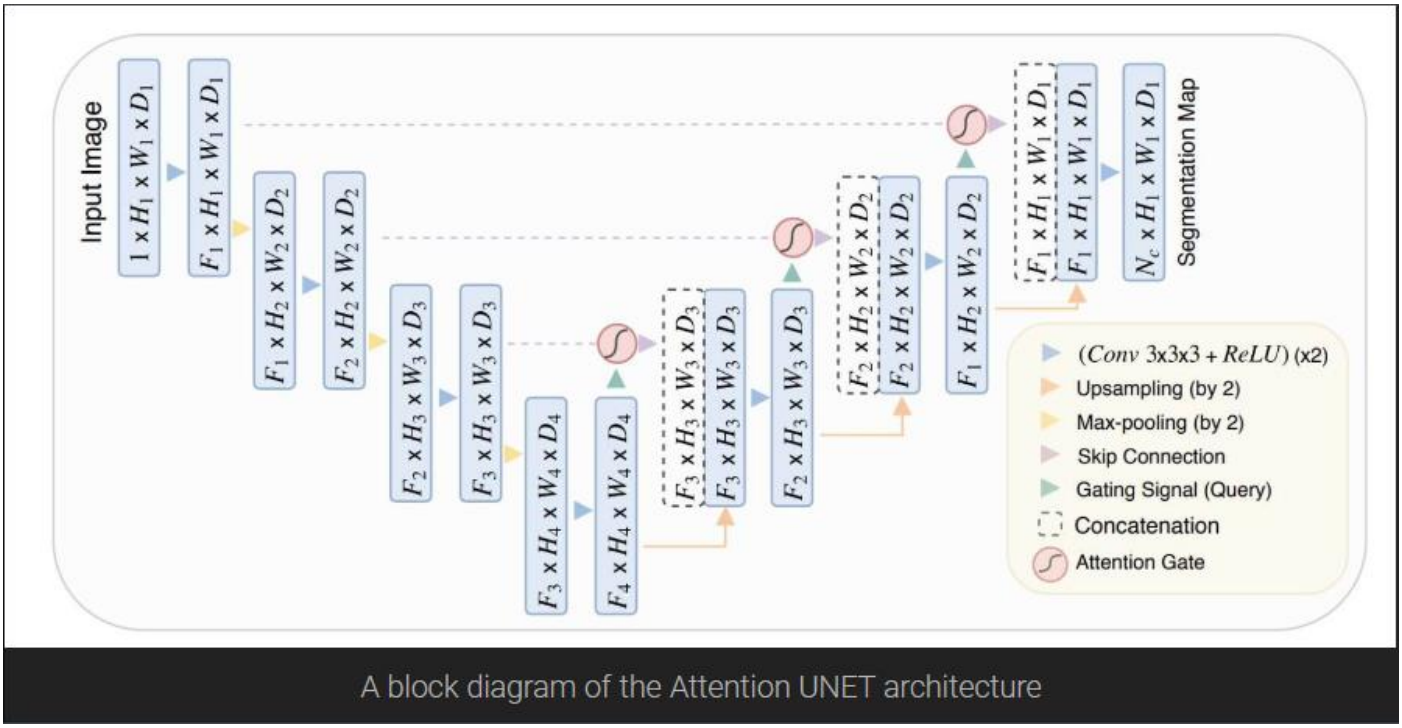
Cette partie reconstruit l'image segmentée en effectuant des opérations d'upsampling (augmentation de la taille de l'image) et de convolution.
Des "skip connections" relient les couches du décodeur aux couches correspondantes de l'encodeur, permettant de récupérer des détails spatiaux perdus pendant la phase de compression.

Le principe clé du U-Net est donc de combiner les informations globales (contenu sémantique) apprises par l'encodeur avec les informations locales (détails spatiaux) transmises par les connexions de saut.
Ainsi, le modèle parvient à segmenter finement chaque pixel, même avec un nombre limité d'images annotées.

Grâce à sa simplicité, sa robustesse et son efficacité, le U-Net est devenu une référence incontournable dans la segmentation d'images, qu'il s'agisse de domaines médicaux, agricoles ou urbains.

4.2. Attention U-Net

Attention U-Net est une extension du modèle classique U-Net qui intègre des mécanismes d'attention pour améliorer la segmentation des objets d'intérêt, en particulier lorsque ces objets représentent une petite portion de l'image ou sont entourés de zones complexes.



A block diagram of the Attention UNET architecture

L'architecture repose sur le principe du U-Net classique, avec un encodeur, un bottleneck et un décodeur, mais elle ajoute des modules d'attention sur les skip connections.

- **Modules d'attention :**

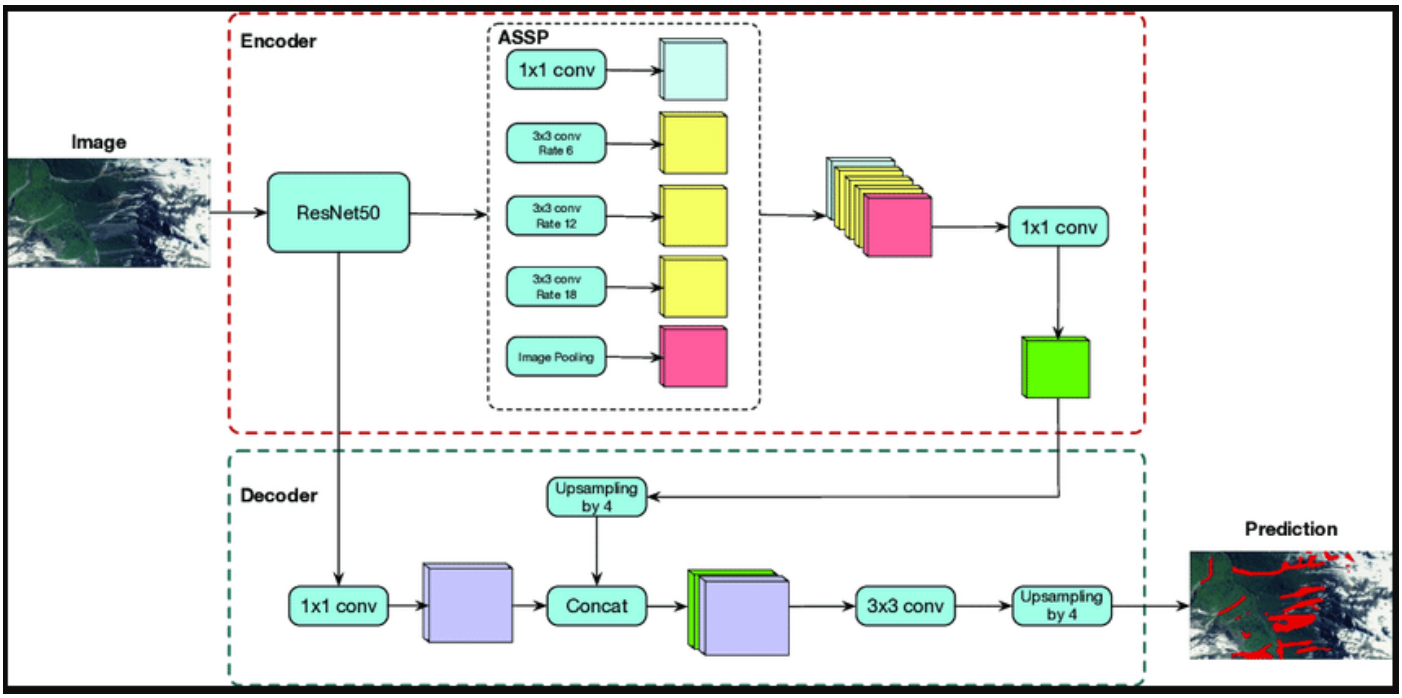
Ces modules permettent au réseau de pondérer les informations transmises du décodeur, en mettant l'accent sur les régions pertinentes de l'image et en filtrant les informations non importantes ou bruitées. Cela améliore la précision de la segmentation, en particulier pour les objets difficiles à détecter.

- **Combinaison informations locales et globales :**

Comme dans le U-Net, les skip connections transmettent les détails spatiaux de l'encodeur vers le décodeur. L'attention permet de sélectionner intelligemment quelles informations sont utiles, renforçant ainsi la qualité des contours et des détails fins.

4.3. DeepLabV3+

DeepLabV3+ est une architecture avancée de segmentation d'images qui améliore la précision sur des scènes complexes grâce à l'extraction de caractéristiques multi-échelles. Il est largement utilisé dans les domaines urbains, naturels et environnementaux, où les objets à segmenter peuvent varier fortement en taille et en forme.



L’architecture de DeepLabV3+ repose sur deux composants principaux :

- **Encoder (Extraction des caractéristiques multi-échelles) :**

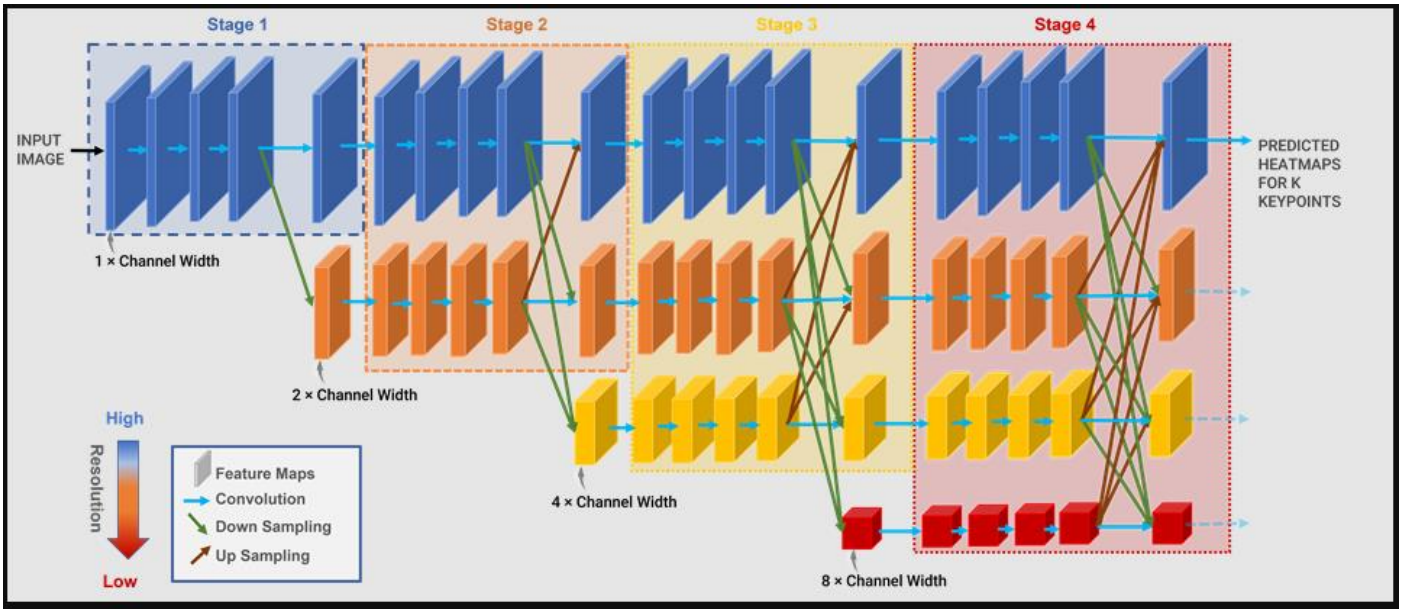
DeepLabV3+ utilise un backbone CNN (souvent ResNet ou Xception) pour extraire les caractéristiques profondes de l’image. Il intègre également le module Atrous Spatial Pyramid Pooling (ASPP), qui applique des convolutions dilatées à différentes échelles pour capturer à la fois les détails fins et le contexte global de l’image. Cette approche permet de segmenter efficacement des objets de tailles variées.

- **Decoder (Reconstruction de la segmentation) :**

Contrairement à DeepLabV3, DeepLabV3+ inclut un décodeur qui combine les caractéristiques extraites par l’encodeur avec les informations spatiales des couches précédentes pour produire des contours plus précis et des segmentations détaillées.

4.4. HRNet

High-Resolution Network (HRNet) est une architecture de segmentation d’images conçue pour maintenir des résolutions élevées tout au long du réseau, contrairement à la majorité des architectures classiques qui réduisent progressivement la taille spatiale pour extraire des caractéristiques profondes. Cette particularité permet à HRNet de conserver les détails fins et les structures spatiales précises de l’image, ce qui est essentiel pour des applications comme la segmentation forestière.



L’architecture de HRNet repose sur les principes suivants :

- **Maintien de multiples résolutions :**

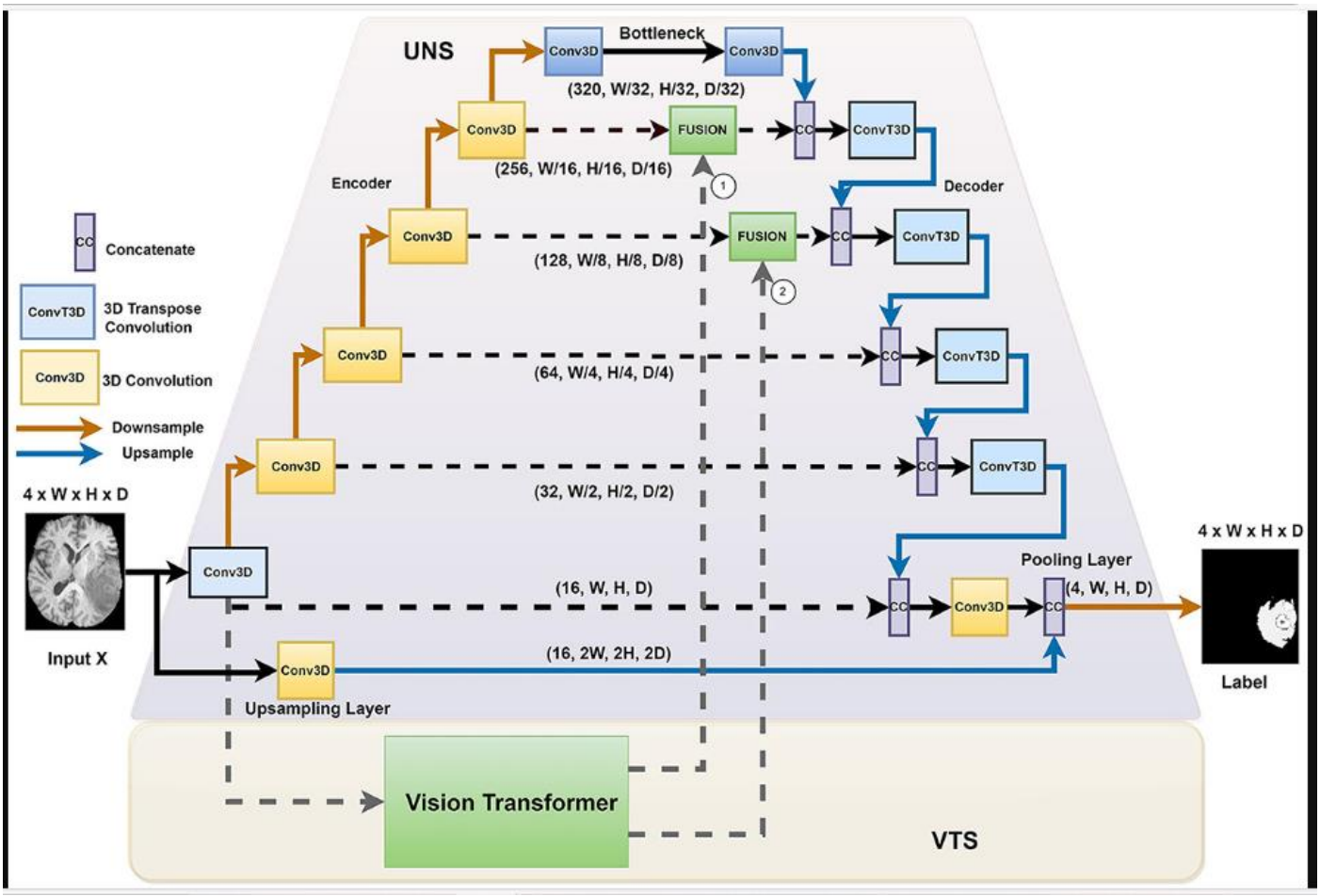
HRNet conserve simultanément plusieurs flux de résolution, du plus élevé au plus bas. Chaque flux capture différents niveaux de détails et de contexte, permettant au réseau de comprendre à la fois les petites structures et les informations globales.

- **Fusion continue des flux :**

Les informations provenant des différents flux de résolution sont régulièrement fusionnées à travers le réseau, ce qui permet de combiner les caractéristiques globales et locales à chaque étape.

4.5. UNETR

UNETR (U-Net with Transformers) est une architecture récente qui combine les avantages des Transformers avec la structure classique des réseaux de type U-Net pour la segmentation d’images. Elle a été conçue pour exploiter la capacité des Transformers à capturer des dépendances globales dans l’image, tout en conservant la reconstruction fine des détails spatiaux grâce au décodeur de type U-Net.



L'architecture de UNETR se compose de deux parties principales :

- **Encodeur basé sur Transformers :**

L'image est découpée en patches, qui sont ensuite transformés en vecteurs de caractéristiques et traités par un Transformer encoder. Cette approche permet de capturer les relations globales et contextuelles à longue portée, ce qui est particulièrement utile pour segmenter des structures complexes et variées.

- **Décodeur de type U-Net :**

Les représentations extraites par le Transformer sont transmises au décodeur via des skip connections. Le décodeur reconstruit ensuite l'image segmentée en combinant les informations globales du Transformer avec les détails locaux, assurant une segmentation précise des contours et des structures fines.

5. Conclusion

Dans ce rapport, nous avons présenté les principes fondamentaux de la segmentation d'images et l'intérêt des projets appliqués à l'analyse forestière, illustré par le dataset **Augmented Forest**. Nous avons ensuite détaillé plusieurs modèles de segmentation avancés, tels que **U-Net, Attention U-Net, DeepLabV3+, HRNet et UNETR**, en soulignant leurs architectures et leurs points forts. Ces modèles offrent des approches complémentaires pour segmenter des images complexes, chacun ayant ses avantages en termes de précision, de préservation des détails spatiaux ou de capacité à capturer le contexte global.