Projet de fin d'études

Évaluation du potentiel des réseaux de neurones avec apprentissage profond pour la discrimination par télédétection spatiale de plantations arborées tropicales

PRÉPARÉ PAR :

- Abir ZAHI
Elève ingénieure en troisième année
à l'IMT d'Alès

ENCADRANTS DE STAGE :

- Frédéric BORNE, CIRAD
- Camille LELONG-RICHAUD, CIRAD
- Marc CHAUMONT, LIRMM

Soutenance le 12 septembre 2023

Remerciements

Je tiens à exprimer ma plus profonde gratitude à mes encadrants de stage, Madame Camille LELONG-RICHAUD, Monsieur Frédéric BORNE et Monsieur Marc CHAUMONT, pour m'avoir offert l'opportunité de réaliser mon stage de Projet de Fin d'Études au sein de l'UMR AMAP et en bénéficiant de l'appui de l'UMR Tetis et du LIRMM. Leurs soutien, conseils et expertise ont joué un rôle inestimable dans mon développement professionnel.

Je tiens également à adresser un chaleureux remerciement à toute l'équipe de l'UMR AMAP pour son accueil et son encadrement bienveillant. Chacun d'entre eux a contribué de manière significative à rendre mon expérience de stage à la fois enrichissante et mémorable.

Je souhaite également exprimer ma reconnaissance envers mon institution académique, l'Institut des Mines et Télécom d'Alès, pour avoir fourni les connaissances et les compétences qui ont été essentielles à la réussite de ce stage. Leur enseignement a constitué les fondations de cette réalisation importante.

Je tiens aussi à exprimer ma profonde gratitude à toutes les personnes qui ont contribué de manière directe ou indirecte à la réussite de mon stage.

Ce travail a bénéficié d'une aide de l'État français gérée par l'Agence Nationale de la Recherche au titre du programme d'Investissements d'Avenir portant la référence ANR-16-CONV-0004.
This work was supported by the French National Research Agency under the Investments for the Future Program, referred as ANR-16-CONV-0004.

1. Introduction

1.1. Organisme d'accueil

Le présent projet de fin d'étude a été mis en œuvre au sein du **CIRAD**, et plus précisément des deux unités de recherche **AMAP** et **TETIS**, et en collaboration avec le LIRMM. Le **CIRAD** est un EPIC (Établissement Public à Caractère Industriel et Commercial) français dédié à la recherche agronomique et à la coopération internationale pour le développement durable des régions tropicales et méditerranéennes. Le **CIRAD** a pour but d'inventer des systèmes d'agriculture durables dans le cadre des transitions agroécologiques, et de protéger la biodiversité et la durabilité des systèmes alimentaires.

L'UMR TETIS "Territoire, Environnement, Télédétection et Information Spatiale" est une unité mixte de recherches interdisciplinaires regroupant une centaine de personnes (Inrae / CIRAD / CNRS / AgroParisTech / INRIA) ayant la mission de développer la maîtrise et l'usage de l'information spatiale pour la compréhension de la complexité territoriale, des agro-éco systèmes et l'accompagnement des acteurs. Son projet scientifique s'articule autour de recherches conceptuelles, méthodologiques et thématiques, en particulier pour l'extraction et l'analyse de l'information spatio- temporelle à partir de données de télédétection.

L'UMR AMAP est une unité mixte de recherche interdisciplinaire qui travaille à l'acquisition de connaissances fondamentales sur les plantes dans le but de prévoir la réponse des écosystèmes aux forçages environnementaux, en termes de distribution/conservation des espèces et de la biodiversité, production des cultures agronomiques, stockage du carbone dans la biomasse végétale, protection de l'environnement et des services écosystémiques.

Le LIRMM laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM) est une Unité Mixte de Recherche de l'Université de Montpellier (UM) et du Centre National de la Recherche Scientifique (CNRS). Le laboratoire a un partenariat avec Inria, l'Université Paul Valéry de Montpellier (UPV) et l'Université de Perpignan (UPVD),il est constitué de 3 départements scientifiques et de 22 équipes de recherche et concernent l'Informatique (algorithmique et complexité, intelligence artificielle, science des données, génie logiciel, recherche opérationnelle, etc.), la Robotique (robotique pour l'industrie du futur, robotique humanoïde et interactions homme-robot, robotique d'exploration sous-marine et robotique chirurgicale et dispositifs médicaux) et la Microélectronique (modélisation, conception, test, fiabilité, sécurité et efficacité énergétique de systèmes électroniques intégrés complexes).

1.2. Cadre du projet

Plus particulièrement, mon travail s'insère dans le **projet européen**[1] **OBSYDYA** (Observatoire pilote des dynamiques agricoles et des paysages ruraux du Bénin (zones Nord et Centre)), dont l'objectif scientifique est de construire des indicateurs de structures et de dynamiques des systèmes agraires et des paysages à partir d'images satellitaires. Ces indicateurs concernent en particulier la taille des exploitations agricoles, les types de culture et la dynamique des zones agricoles, la dynamique des vergers et plantations arborées, ainsi que la dégradation forestière. Dans ce contexte général, l'un des objectifs techniques du projet, par exemple, est de développer un outil opérationnel permettant la cartographie de l'occupation du sol et sa mise à jour régulière à l'échelle des grandes régions agricoles du Bénin à partir de séries temporelles d'images à moyenne résolution spatiale (Sentinel-2). Un autre de ces objectifs concerne le développement d'un outil

[1] initiative de la Communauté Européenne DeSIRA (Development Smart Innovation through Research in Agriculture)

permettant l'évaluation des surfaces plantées de diverses essences fruitières (anacardiers et manguiers en particulier) et des autres plantations arborées (teck, eucalyptus, mais aussi forêts) ainsi que leur suivi spatio-temporel, plutôt à partir d'images à très haute résolution spatiale (Pléiades, WorldView…).
Mon étude a donc consisté, dans ce cadre, à explorer le potentiel des méthodes issues de l'Intelligence Artificielle et en particulier de l'apprentissage profond pour la reconnaissance, la discrimination, ainsi que la délimitation de différents vergers afin de les cartographier à partir d'images satellites acquises par le capteur Pléiades-Néo.

1.3. Contexte

L'utilisation des algorithmes d'apprentissage profond, notamment des réseaux de neurones convolutifs, a véritablement développé le traitement de l'imagerie satellitaire. Ces avancées technologiques ont ouvert de nouvelles perspectives dans de nombreux domaines, offrant de nouvelles approches pour améliorer la cartographie, la surveillance environnementale et la gestion des ressources naturelles.

Grâce à l'IA, les chercheurs et les professionnels disposent des outils puissants pour extraire des informations riches et précises à partir d'images satellitaires. Cette capacité à analyser en profondeur les données visuelles permet une compréhension plus fine des territoires, des écosystèmes et des phénomènes environnementaux [1].

Un des avantages majeurs de l'utilisation de l'IA dans le traitement de l'imagerie satellitaire est censé être plus rapide et plus facile à utiliser après la finalisation de l'algorithme et du modèle. Les algorithmes d'apprentissage profond peuvent en particulier traiter de grandes quantités d'images rapidement, réduisant les délais de traitement et facilitant la prise de décisions.

Dans le contexte spécifique de notre projet de fin d'études, axé sur la cartographie des anacardiers à des fins d'inventaire des surfaces plantées, nous nous appuierons sur cette approche par apprentissage profond. Notre méthodologie de reconnaissance et de discrimination des différentes parcelles contenant des anacardiers, des manguiers et d'autres types d'arbres, tels que le teck ou les forêts, repose sur la combinaison de ces avancées en traitement d'image et de l'expertise en télédétection.

2. Etat de l'art

2.1. La segmentation d'images

La segmentation d'images joue un rôle crucial dans de nombreuses tâches de vision par ordinateur, telles que la détection d'objets et la classification d'images. Cette méthode consiste à diviser une image en zones, généralement en fonction des contours ou des limites visibles des objets présents, dans le but de simplifier la complexité de l'image.

Un des aspects de la segmentation est l'attribution d'étiquettes à chaque pixel de l'image, permettant ainsi de définir les éléments importants présents dans celle-ci. Cette approche peut être appliquée dans divers domaines tels que les véhicules autonomes, l'analyse d'images médicales, les images satellites, la vidéosurveillance, ainsi que d'autres tâches de reconnaissance et de détection.
Dans le domaine des images satellites, la segmentation aide à identifier et à cartographier différentes caractéristiques du terrain, telles que les cours d'eau, les forêts ou les zones urbaines, ce qui peut être utile pour la planification urbaine et la gestion des ressources.
Segmentation sémantique : Elle permet d'associer chaque pixel d'une image à une étiquette de catégorie, comme une voiture, un arbre, un fruit ou une personne. Elle traite plusieurs objets de la même catégorie comme une seule entité.
Segmentation d'instances : Elle ne permet pas d'associer chaque pixel d'une image à une étiquette de catégorie. Elle traite plusieurs objets de la même catégorie en tant qu'instances individuelles distinctes, sans nécessairement reconnaître les instances individuelles. Par exemple, deux voitures 1 et 2 seront représentées par des couleurs différentes dans une image car elles représentent deux objets différents.

L'intégration de l'IA a joué un rôle essentiel dans l'amélioration de la segmentation sémantique pour l'analyse d'images, aboutissant à une segmentation plus précise dans une grande variété d'applications [2]. Les deux approches qui ont marqué une révolution dans ce contexte sont les réseaux de neurones convolutifs (CNN) et les architectures de réseaux à base de Transformers dédiés à la vision, les Vision Transformers (ViTs)

2.2. Les CNNs

Les CNNs sont un type de réseau neuronal artificiel conçu pour traiter et analyser les données visuelles, ce qui les rend idéaux pour les tâches de segmentation d'images. Ces réseaux se composent de plusieurs couches, chaque couche étant responsable de l'extraction de différentes caractéristiques de l'image d'entrée. Au fur et à mesure que l'image traverse le réseau, les couches apprennent progressivement des caractéristiques plus complexes et abstraites, conduisant finalement à l'identification et à la segmentation des objets dans l'image. L'utilisation des CNN dans la segmentation sémantique a conduit à des améliorations significatives en termes de précision et d'efficacité [2]. Les CNN peuvent apprendre et s'adapter automatiquement aux différentes caractéristiques des données d'entrée, ce qui leur permet de segmenter les images avec plus de précision et de robustesse. De plus, les CNN peuvent être formés sur de grands ensembles de données, ce qui leur permet de bien généraliser aux images nouvelles et inédites, améliorant encore leurs performances dans les applications du monde réel.

2.2.1 ResNet 34 et ResNet 50

ResNet, abréviation de Réseau Résiduel, a été introduit en 2015 par Kaiming He et al. [4] pour résoudre des problèmes liés à la formation de réseaux neuronaux profonds. Les réseaux neuronaux convolutifs profonds ont en effet montré des résultats impressionnants dans diverses tâches de vision par ordinateur, mais l'ajout de couches supplémentaires peut entraîner une dégradation des performances : alors que le nombre de couches empilées peut enrichir les fonctionnalités du modèle, un réseau plus profond peut révéler le problème de la dégradation. En d'autres termes, à mesure que le nombre de couches du réseau neuronal augmente, les niveaux de précision peuvent devenir saturés et se dégrader lentement après un certain point. En conséquence, les performances du modèle se détériorent à la fois sur les données de d'apprentissage et de test.

La technique clé des ResNets est l'utilisation de connexions résiduelles, ou "sauts", qui permettent aux informations de contourner certaines couches et ainsi faciliter l'apprentissage de fonctions complexes. Cela résout le problème de la dégradation des performances en permettant aux couches supérieures de ne pas devenir moins performantes que les couches inférieures.

Les variantes de ResNet, telles que ResNet-34 et ResNet-50, diffèrent par le nombre de couches utilisées dans leur architecture. ResNet-34 a été le premier modèle à introduire les connexions résiduelles, tandis que ResNet-50 a présenté des blocs de goulot d'étranglement à 3 couches pour accélérer la formation.

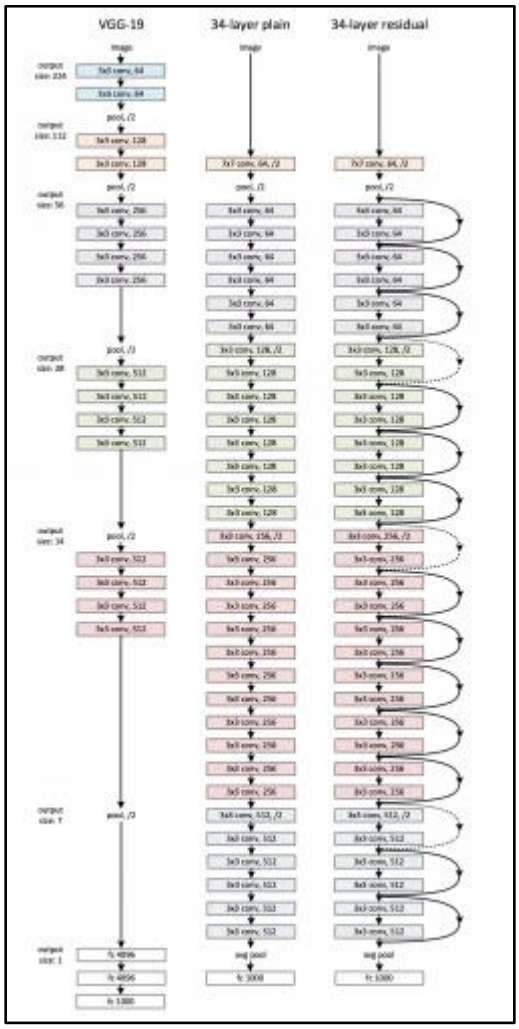


Fig. 1 : Exemples d'architectures réseau pour ImageNet[2]. A gauche : le modèle VGG-19 (19,6 milliards de FLOP) comme référence. Au milieu : un réseau simple avec 34 couches (3,6 milliards de FLOP). À droite : ResNet avec 34 couches (3,6 milliards de FLOP). Les raccourcis en pointillés augmentent les dimensions[5]

Sur la base du réseau simple ci-dessus, une connexion raccourcie est insérée (Fig.1, à droite) qui transforme le réseau en sa version résiduelle homologue.

[2] une base de données d'images annotées utilisée pour la recherche en vision par ordinateur. Lancée par l'organisation du même nom, ImageNet contient des images annotées avec des objets et des boîtes englobantes. Elle contient environ 1,5 Millions d'images.

2.2.1 U-net

U-Net, issu du réseau de neurones CNN traditionnel, a été conçu et appliqué pour la première fois en 2015 pour traiter des images biomédicales [6]. Il est capable de localiser et de distinguer les frontières des éléments composant une image en faisant la classification sur chaque pixel. Détaillons l'architecture de ce modèle afin de comprendre son fonctionnement et ce qui fait qu'il soit aussi précis et adapté à des situations complexes.

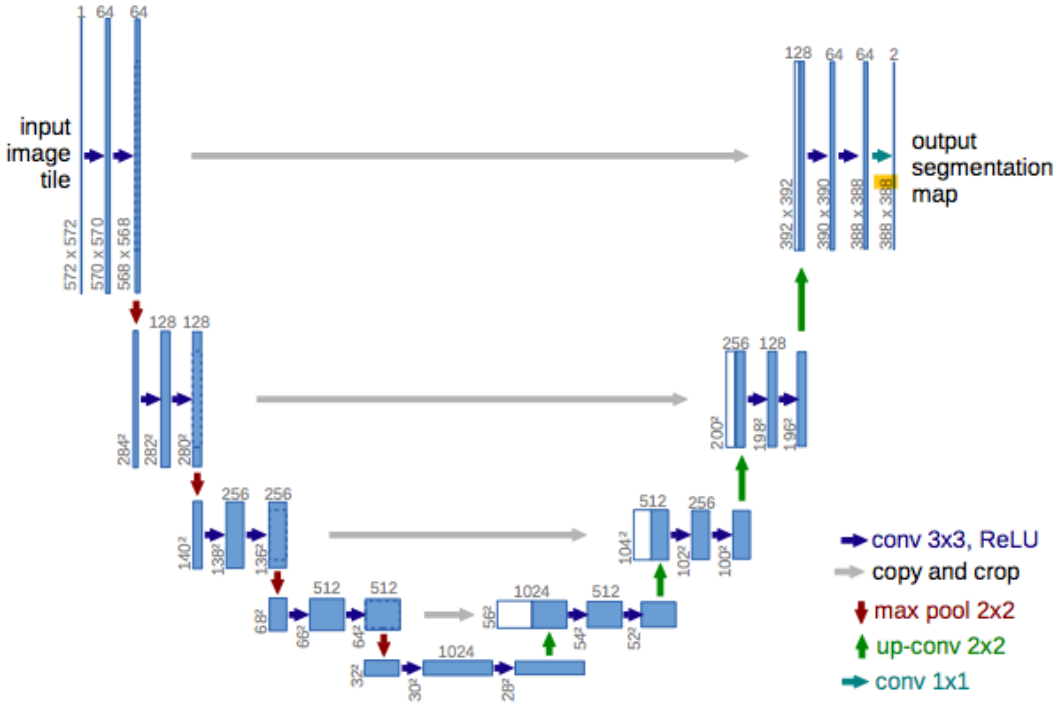


Fig. 2 : Architecture d'U-NET.

Visuellement, il a une forme en « U ». L'architecture est symétrique et se compose de trois sections : La contraction, le goulot d'étranglement et la section d'expansion.

Le premier bloc (partie gauche du réseau), aussi appelé **encodeur,** est utilisé pour récupérer le contexte d'une image. Ce bloc consiste en un assemblage de couches de convolution et de couches de max pooling permettant de capturer les caractéristiques d'une image et de réduire sa taille pour diminuer le nombre de paramètres du réseau. Cela consiste en l'application répétée de deux couches de convolution 3x3. Chaque couche est suivie d'une fonction d'activation ReLU et d'une normalisation par lots (batch normalization). Ensuite, une opération de max pooling 2x2 est appliquée pour réduire les dimensions spatiales.

Le pont, appelé aussi goulot d'étranglement, relie l'encodeur et le réseau de **décodeurs**, et complète le flux d'informations. Il se compose de deux couches de convolutions 3x3, où chaque couche est suivie d'une fonction d'activation ReLU. Le second bloc est celui du décodeur. Il permet la localisation précise grâce à la convolution transposée et permet également de retrouver la taille initiale de l'image. Le bloc décodeur commence par un sur-échantillonnage (upsampling) de la carte des caractéristiques suivie d'une couche de convolution 2x2 transposée. Après, deux couches de convolutions 3x3 sont utilisées, où chaque convolution

est suivie d’une fonction d’activation ReLU. La sortie du dernier décodeur passe par une couche de convolution 1x1 avec une fonction d'activation sigmoïde.

U-Net utilise une fonction de perte pour chaque pixel de l’image. La fonction Softmax est appliquée à chaque pixel suivi d’une fonction de perte. Ceci convertit le problème de segmentation en un problème de classification où on doit classer chaque pixel dans l’une des classes.

2.3. LinkNet

LinkNet a été conçu pour surmonter certaines limitations des architectures précédentes, tout en conservant les avantages clés de ces dernières.

La principale différence entre LinkNet et U-Net réside dans la manière dont ils gèrent les informations à différentes échelles. U-Net, une architecture utilisée fréquemment pour la segmentation sémantique, fonctionne en utilisant une structure d'encodeur-décodeur, où l'encodeur capture les caractéristiques de l'image à différentes résolutions, tandis que le décodeur restaure la segmentation à la résolution d'origine en utilisant ces caractéristiques. Cependant, U-Net peut rencontrer des problèmes de perte d'information fine lors de la transition entre les échelles.

LinkNet a été conçu pour résoudre ce problème en utilisant une approche différente. Il utilise des blocs de résidus et des connexions de sauts (skip connections) pour relier directement les caractéristiques d'une échelle à une autre. Cette stratégie permet de préserver les informations à haute résolution tout au long du réseau, améliorant ainsi la précision de la segmentation, notamment pour les objets de petite taille ou les détails fins.

En outre, LinkNet emploie des couches de convolutions fractionnaires, qui réduisent le nombre de paramètres et accélèrent le processus d’apprentissage, tout en maintenant des performances de segmentation compétitives. Cela en fait une option attrayante pour les applications en temps réel ou les situations où l'efficacité de calcul est cruciale.

2.4. PSPNet

PSPNet (Pyramid Scene Parsing Network) est une architecture de réseau neuronal convolutif (CNN) avancée, qui est particulièrement puissant pour la compréhension des scènes complexes et la segmentation d'objets de grande échelle.
L'élément clé de PSPNet est le module de pyramide spatiale (PSP), qui permet de capturer des informations contextuelles à différentes résolutions spatiales. Ce module divise l'image en plusieurs résolutions et agrège des informations contextuelles à chaque échelle. Cela signifie que PSPNet peut non seulement détecter des caractéristiques locales, mais aussi comprendre le contexte global de l'image, ce qui améliore considérablement la précision de la segmentation.

PSPNet peut être pré-entraîné sur de grandes bases de données d'images pour apprendre des caractéristiques générales, puis « fine-tuné » sur des données spécifiques à la tâche.

3. Problématique

3.1. Zone d'étude

La zone étudiée couvre environ 100 km2 et se situe au Nord Bénin dans la périphérie Est de la ville de Parakou (9°21' de latitude Nord, 2°36' de longitude Est), qui est la troisième ville la plus importante du pays. Cette région essentiellement rurale est située à une altitude moyenne de 350 m et bénéficie d'un climat tropical de type soudanien, caractérisé par une saison sèche de cinq mois, d'octobre à avril, et une saison pluvieuse le reste de l'année. Elle offre un paysage varié et diversifié, dont le couvert végétal est dominé par une savane arborée comprenant des essences à valeur ajoutée telles que, par exemple, le néré, le faux acajou, le bois d'ébène, ou encore le karité. L'agriculture y est l'activité prédominante, les principales cultures vivrières étant l'igname, le maïs, le manioc, le sorgho et le niébé, suivies du soja et du riz. En parallèle, les cultures de rente comme le coton, l'anacarde, l'arachide et le teck contribuent également à l'économie locale. Cependant, en raison des problèmes écologiques et financiers, la culture du coton est en déclin, laissant place à l'essor des plantations de tecks, d'anacardiers et de manguiers qui offrent une double opportunité : protéger les sols tout en générant des revenus importants pour les agriculteurs. Cette récente dynamique vers un développement des exploitations arborées est encore mal connue et l'opportunité de suivre les surfaces plantées par télédétection apporterait donc aux gestionnaires du territoire un outil permettant d'en appréhender la réelle étendue et d'en assurer le suivi.

Fig. 4 : Localisation du Bénin en Afrique

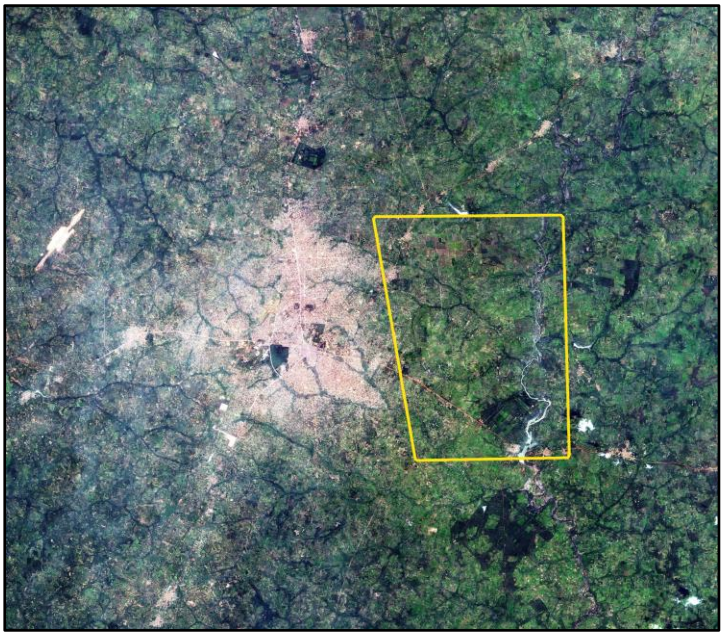
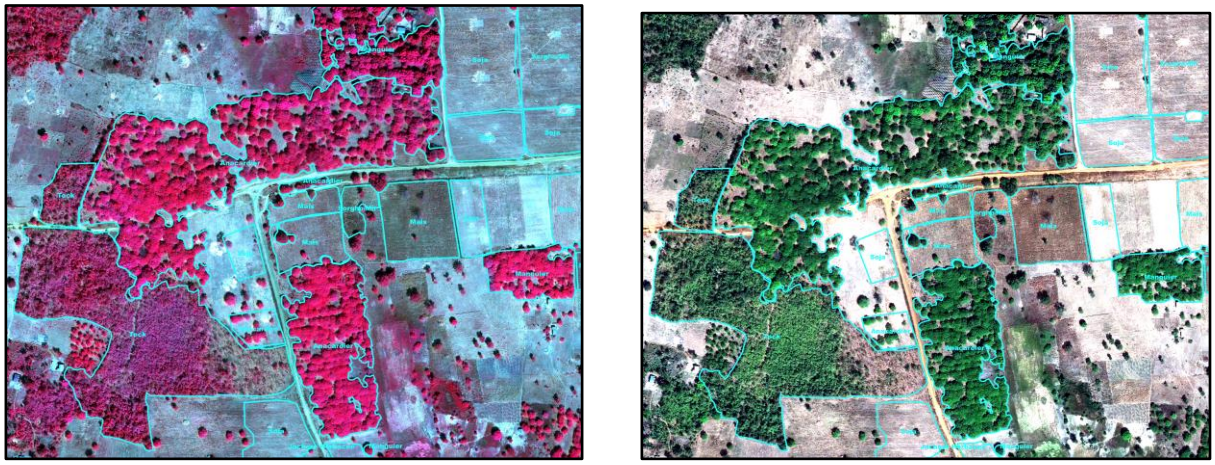


Fig. 5 : Emprise de l'image Pléiades-NEO totale (cadre jaune) dans son contexte. On voit nettement la ville de Parakou juste à l'ouest de la zone.

3.2. La donnée d'imagerie en télédétection satellitaire, Pléiades-NEO

Les données traitées consistent en une image multispectrale de la zone d'étude de Parakou, que nous appellerons image de référence. Cette image a été acquise par le capteur satellitaire Pléiades-NE0[8], dans 6 bandes spectrales. Elle se présente sous la forme d'un fichier multicouches au format TIFF, géoréférencé, couvrant une superficie de 10 km de large par 13 km de haut avec une résolution spatiale de 1,20m par pixel, accompagné d'un fichier monobande, dit panchromatique, couvrant une large gamme de longueurs d'ondes

du visible au proche infrarouge, acquise à 30 cm de résolution spatiale. Cette spécification offre une vue détaillée des divers types d'occupation et d'usage du terrain, qu'ils soient boisés, cultivés ou non végétalisés.



Figures 6 et 7 : extrait de l'image source fusionnée à 30cm de résolution spatiale, représenté en mode RGB avec les bandes spectrales infrarouge, rouge et verte à gauche et avec les bandes spectrales rouge, verte et bleue à droite. Les lignes bleues représentent les délimitations des parcelles annotées sous forme de polygones.

Dans la tâche de cartographie des anacardiers et de gestion des surfaces plantées, les bandes spectrales jouent un rôle crucial dans la caractérisation des parcelles et l'identification des espèces d'arbres. Pour cette mission, le capteur utilisé est le Pléiades-NEO Imager [8], embarqué à bord de deux satellites, qui acquiert des images optiques multispectrales avec une résolution spatiale de 1,20 m et une image panchromatique à 30cm de résolution spatiale. Les pixels de chacune de ces différentes images ont pour valeur la quantité de lumière réfléchie par la surface observée (réflectance) dans la bande de longueur d'onde d'acquisition et apporte donc une information sur la composition et la structure de cette surface.

Les différentes bandes spectrales mesurées sont le Deep Blue, Blue, Green, Red, Red Edge et Near-infrared, couvrant une large gamme de longueurs d'onde permettant de détecter des caractéristiques spécifiques de la surface terrestre. Chaque bande apporte une information sur l'absorption de la lumière par la surface dans cette gamme de longueur d'onde, selon sa composition, son épaisseur optique, sa transparence, etc.... Ce qui peut être mis, dans le cas des végétaux par exemple, en relation avec des propriétés biophysiques.

En particulier, la bande rouge (Red) est corrélée à la vitalité et l'activité chlorophyllienne de la végétation car elle correspond à la bande d'absorption de la chlorophylle. La bande Red Edge, par exemple, est particulièrement utile pour évaluer la santé de la végétation, car elle permet de détecter les variations subtiles dans la chlorophylle et la teneur en eau des plantes. Elle est aussi corrélée à la phénologie des plantes. Enfin, la bande acquise dans l'infrarouge (Near-infrared) est directement corrélée à la densité et la vigueur de la végétation car elle correspond à la longueur d'onde de la réflexion multiple et diffusion de la lumière au sein des différentes strates de feuilles.

Afin de bénéficier à la fois de la plus grande richesse spectrale et de la plus forte résolution spatiale des données brutes, on applique un modèle de fusion entre l'image multispectrale et l'image panchromatique (appelé pansharpening en anglais) afin de produire un seul fichier multibande à 6 couches et à 30 cm de résolution spatiale.

Enfin, pour introduire les aspects texturaux dans le voisinage de chaque pixel, et donc intégrer l'information de structure spatiale en supplément aux informations radiométriques, nous avons calculé 3 indices de texture issus de la statistique du second ordre de la matrice de cooccurrence calculée sur l'image panchromatique dans un voisinage de 35 pixels soit 10 mètres : moyenne, contraste et entropie (cf. Haralick et al., 1973) [9]

3.3. Description de la base de données de référence

Afin de fournir une référence fiable sur la réalité des différentes surfaces de la zone d'étude, une base de données appelée "vérité-terrain" a été constituée à partir d'informations collectées sur le terrain. Cette base se présente sous la forme d'un fichier "shapefile" : un format de stockage de données vectorielles développé par Esri[3], qui permet d'enregistrer l'emplacement géographique, la forme et les attributs des entités présentes sur la carte. En d'autres termes, il contient des annotations pour les différentes classes d'occupation et d'usage du sol, dont les arbres tels que les manguiers, les anacardiers, ou les tecks, mais aussi les cultures annuelles comme le maïs ou le soja, ou encore les zones artificialisées telles que les routes, les bâtiments et les sols nus.

Cette combinaison de l'image Pléiades-NEO géoréférencée et de données vectorielles permet d'effectuer des analyses approfondies sur la couverture terrestre et de mieux comprendre la répartition des différentes caractéristiques environnementales dans la région nord du Bénin.

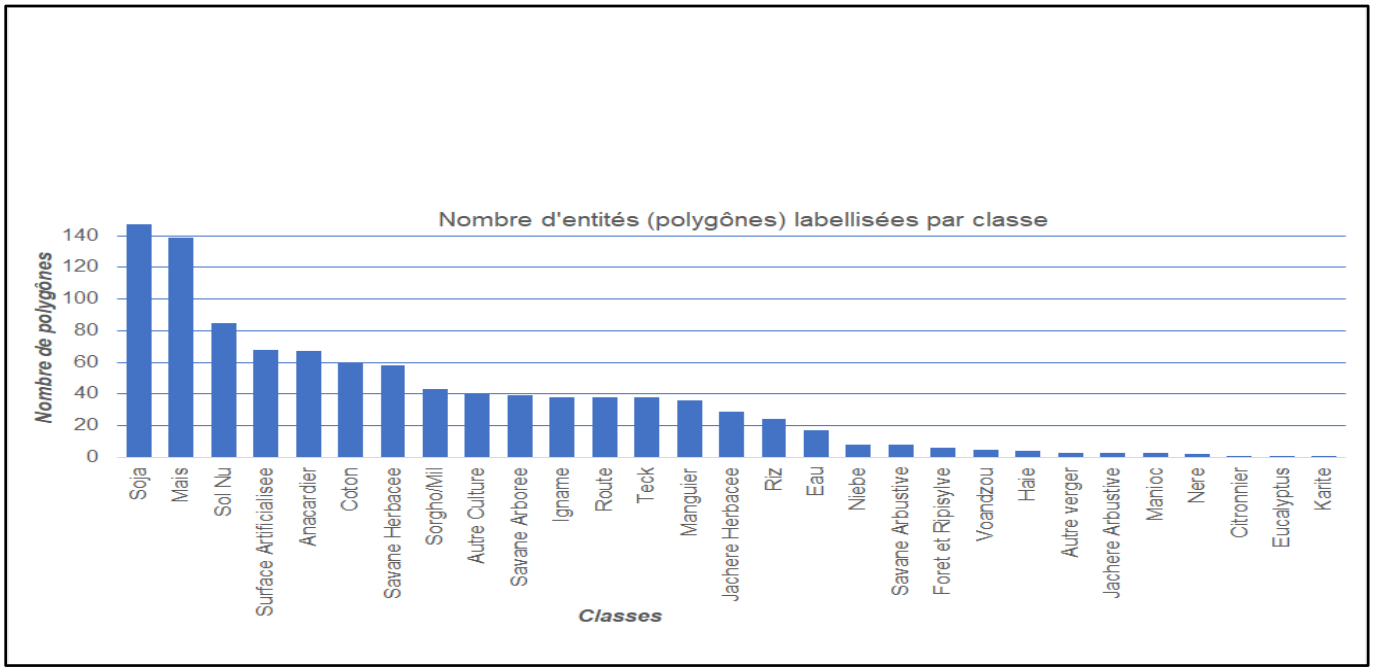


Fig. 8 : Nombres d'entités labellisées -fournies dans le shapefile sous forme de polygones- par classe.

[3] une plateforme spatiale accessible à tous, partout et tout le temps.

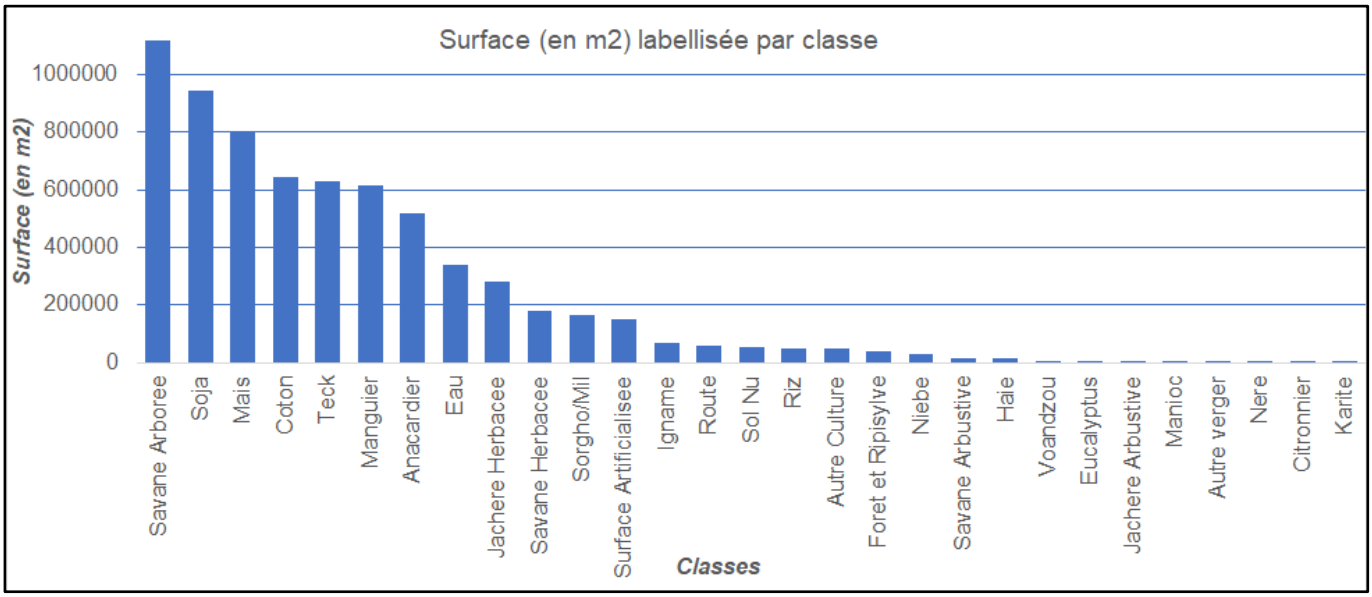


Fig. 9 : Surfaces totales couvertes par classe.

Notre première difficulté se présente à cette étape. Elle a trait au fait que les annotations ne couvrent pas l'intégralité de l'image, ce qui pose un problème : cela ne garantit pas l'absence des classes concernées dans les zones non annotées. En d'autres termes, nous ne pouvons pas être certains que certaines caractéristiques ou objets importants ne se trouvent pas dans les zones non étiquetées.

Pour résoudre ce problème, différentes propositions ont été discutées au sein de notre équipe. L'une d'entre elles consistait à utiliser Papri ou un modèle d'apprentissage profond sur QGIS pour enrichir les annotations existantes. Cependant, cette approche présente un inconvénient majeur : elle pourrait entraîner la création d'une base d'apprentissage inexacte et sujette à des imprécisions. En effet, les prédictions basées sur des modèles pourraient ne pas être fiables et affecteraient donc la qualité globale des annotations.

Une autre proposition envisagée était d'adopter une méthode de "copier-coller", où l'on extrait toutes les zones déjà annotées pour reconstruire des imagettes à partir de celles-ci. Cette approche présente toutefois un inconvénient important : elle néglige les dépendances locales entre les différentes parcelles ou objets. En supposant que ces dépendances n'existent pas, nous pourrions manquer des informations cruciales sur la distribution spatiale des classes d'intérêt, ce qui pourrait limiter la précision de notre segmentation.

Enfin, une dernière proposition consistait à masquer les zones qui ne sont pas labellisées, c'est-à-dire les zones non annotées. En agissant ainsi, nous pourrions éviter de fausses interprétations des données non étiquetées, tout en nous concentrant uniquement sur les zones d'intérêt définies par les annotations existantes. Cependant, cette approche pourrait également conduire à la perte potentielle d'informations précieuses présentes dans les zones masquées, ce qui limiterait notre compréhension globale du paysage.

En conclusion, nous devons trouver un compromis entre ces différentes approches pour résoudre le problème de la couverture partielle des annotations. Il est essentiel d'utiliser des méthodes qui préservent au mieux les informations tout en évitant les erreurs et les biais importants dans notre base de données.

3.4. Choix de la tâche

La question posée est une estimation des surfaces arborées cultivées et non la connaissance précise et individuelle de chacun des polygones constituant ces surfaces. La segmentation sémantique est donc la tâche la plus adaptée à notre projet de cartographie des anacardiers et de gestion des surfaces plantées en raison de la nécessité de labelliser tous les pixels non annotés. Contrairement à la segmentation d'instances, qui se

concentre sur la détection et l'identification d'objets individuels, la segmentation sémantique attribue une étiquette sémantique à chaque pixel de l'image, permettant ainsi une cartographie précise et détaillée de la couverture terrestre.
De plus, dans la segmentation sémantique, l'information est propagée dans tout le réseau pour déterminer la classe sémantique de chaque pixel. Cela signifie que le réseau est capable de prendre en compte le contexte global de l'image, ce qui est crucial pour la compréhension complète de la scène. La segmentation d'objet, en revanche, se concentre sur l'identification et la délimitation précises d'objets individuels, mais elle peut perdre de vue le contexte global.
Ainsi, la segmentation sémantique peut servir de base pour d'autres tâches de vision par ordinateur, telles que la détection d'objets. En utilisant une carte de segmentation sémantique, il est plus facile de localiser et de classifier des objets spécifiques dans une image, car on connaît déjà les classes sémantiques de tous les pixels.
En segmentation d'instances, l'objectif est d'attribuer un identifiant unique à chaque instance d'objet, même s'il y a plusieurs instances de la même classe dans une image. Lorsque les objets sont coupés en plusieurs tuiles (cela se produit généralement lorsque les objets s'étendent au-delà des limites d'une seule tuile d'image) cela peut entraîner une subdivision des objets en plusieurs parties distinctes, chacune se voyant attribuer un identifiant d'instance différent ce qui compliquera la tâche aussi, chaque objet détecté est encadré par une boîte englobante qui définit sa limite. Cette boîte englobante est souvent utilisée pour isoler et extraire chaque objet en tant qu'instance indépendante de l'image. Cependant, cela peut poser des défis spécifiques dans des scénarios où les objets se chevauchent ou sont très proches les uns des autres.

3.5. Choix du réseau

Nous avons pris la décision de travailler avec U-Net pour plusieurs raisons importantes qui correspondent parfaitement à nos objectifs de projet.
La communauté de R&D en vision par ordinateur et en segmentation d'images aériennes a largement adopté U-Net, ce qui signifie qu'il existe de nombreux tutoriels, exemples de code et forums en ligne qui nous ont guidés tout au long de notre projet [7].
Un autre facteur important était la limitation de nos données d'entraînement. Nous disposions d'un ensemble de données relativement petit, ce qui aurait pu poser des problèmes pour certains modèles complexes nécessitant une grande quantité de données pour bien généraliser. Cependant, U-Net a montré une très bonne efficacité dans la segmentation d'images avec des ensembles de données restreints [6]. Sa capacité à extraire des informations riches et à capturer les caractéristiques importantes de l'image nous a rassurés quant à sa capacité à produire de bons résultats malgré le manque de données d'entraînement.
De plus, notre projet nécessitait de tester plusieurs combinaisons de bandes spectrales et de classes pour identifier la configuration la plus appropriée pour notre tâche de segmentation. U-Net offre la flexibilité nécessaire pour modifier facilement le nombre de classes et les canaux d'entrée, ce qui nous permet de réaliser rapidement des expériences comparatives. Sa structure avec des opérations de fusion facilite également la gestion des différentes combinaisons sans ajouter une complexité excessive à notre réseau.
Enfin, notre choix pour un réseau léger et peu complexe se justifie par le besoin d'obtenir des résultats rapidement et de manière efficace. U-Net, avec son architecture simple mais efficace, répond parfaitement à ces exigences et nous permet de mener des expériences avec une bonne vitesse de traitement sans sacrifier la précision de la segmentation.

De plus, sur le plan pratique, la facilité d'implémentation et la disponibilité des ressources sur internet ont été un des facteurs supplémentaires dans notre choix de modèle.

4. Méthodologie

4.1. Librairies et ressources mises en œuvre

Dans ce projet, j'ai utilisé les bibliothèques TorchGeo, Tifffile, TensorFlow et Keras, adaptées au traitement d'images géospatiales et l'apprentissage profond. TorchGeo offre des outils spécifiques pour la manipulation de données géospatiales, tandis que Tifffile facilite la gestion d'images au format TIFF, pour l'écriture et la lecture. TensorFlow et Keras sont des cadres de développement populaires pour créer des modèles de réseaux de neurones puissants, permettant de gérer les différentes étapes de construction, d'entraînement et d'évaluation des modèles.

Une autre bibliothèque utilisée dans ce projet : "Segmentation Models" , représente un outil puissant pour l'implémentation de modèles de segmentation sémantique en apprentissage profond. Elle fournit une collection de modèles de pointe, tels que U-Net, FPN, PSPNet, etc., qui sont pré-implémentés et pré-entraînés sur de vastes ensembles de données.[10]

L'ordinateur qui a servi aux expérimentations a la configuration suivante :

- processeur Intel Xeon W-2235 (6 cœurs HT, 3.8Ghz, 4.6GHzTurbo, 8.25M Cache, 130W),
- deux cartes graphiques NVIDIA RTX A5000 24Go (8192 cœurs CUDA, 256 cœurs Tensor gen3, 64 cœurs RT 2ème gen, 4xDP),
- une RAM de 64Go 3200 MHz DDR4 - 4x16Go ECC
- 4 To d'espace disque de stockage

4.2. Métriques d'évaluation utilisées

4.2.1. Matrice de confusion

D'une manière générale, l'évaluation des performances d'un réseau en télédétection est compliquée. La vérité terrain est souvent faible en quantité d'échantillons et la qualité est aussi un problème pour un pourcentage significatif des échantillons. Il est en effet difficile de définir le contour de certaines zones, notamment dans des milieux difficiles d'accès comme les forêts, et le contour tracé n'est pas toujours en correspondance avec l'image (aérienne / drone / satellite).
Pour évaluer les performances de notre modèle, on évalue ses prédictions en utilisant la matrice de confusion, aussi connue sous le nom de tableau de contingences [11]. Une matrice qui permet d'analyser les résultats corrects et incorrects du modèle, en mettant en évidence les types d'erreurs commises. Chaque colonne de la matrice correspond à une classe prédite par l'algorithme, et chaque ligne correspond aux classes réelles.
Les résultats sont divisés en quatre catégories :
Vrai Positif (TP) : L'algorithme prédit correctement une valeur positive.
Vrai Négatif (TN) : L'algorithme prédit correctement une valeur négative.
Faux Positif (FP) : L'algorithme prédit à tort une valeur positive alors qu'elle est négative en réalité.
Faux Négatif (FN) : L'algorithme prédit à tort une valeur négative alors qu'elle est positive en réalité.

Les matrices qu'on visualisera tout au long de nos résultats sont des matrices normalisées sur les lignes, en d'autres termes sur les classes réelles, ce qu'on appelle en géomatique précision.
En statistique ce terme correspond à la sensibilité, mesurant la proportion des prédictions positives correctement identifiées.

$$\text{Précision} = \frac{TP}{TP+FN}$$

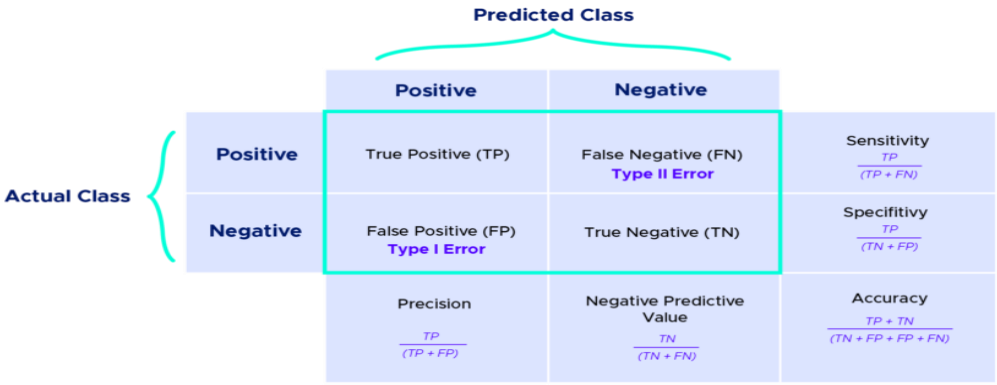


Fig. 10 : Descriptif de la matrice de confusion.[12]

4.2.2. Intersection over Union

L'IoU (Intersection over Union), également connu sous le nom de Jaccard Index, est une métrique couramment utilisée pour évaluer la précision des modèles de détection d'objets et de segmentation d'images, en particulier dans le domaine de la vision par ordinateur et de l'apprentissage profond.
L'IoU mesure le degré de chevauchement entre deux ensembles, généralement des ensembles de pixels ou de régions d'intérêt. Dans le contexte de la détection d'objets ou de la segmentation, l'IoU compare la zone d'intersection entre la prédiction du modèle et la vérité terrain (la véritable position de l'objet ou de la région) par rapport à l'union des deux ensembles.
Tout au long des résultats, on calcule l'IoU pour chaque classe pour une évaluation fine.

4.3. Préparation des données

Pour préparer notre base de données d'apprentissage, nous avons pris la décision de masquer les zones non labellisées sur les images satellitaires. Cette approche vise à exclure les régions de l'image qui ne sont pas annotées, afin de ne pas les prendre en compte ni lors de l'entraînement du modèle ni lors de son évaluation.

Nous avons procédé par la suite au découpage de l'image en imagettes de taille 512x512, en adéquation avec l'entrée des modèles que nous utilisons. Il est important de noter que nous avons choisi de ne pas avoir de chevauchement (overlap) entre les imagettes. Cette décision a été prise en vue de l'évaluation ultérieure

du modèle. Nous avons par la suite séparé notre jeu de données en deux répertoires, constitués en sélectionnant des imagettes d'une manière aléatoire : 80% des imagettes servent pour l'entraînement et les 20% restants pour le test qui servira à l'évaluation des performances du modèle sur ces données qui n'ont pas servi à l'entraînement. Nous nous assurons donc que le modèle apprend à faire des prédictions sur des régions isolées des imagettes d'entraînement sans utiliser des informations superposées.

Enfin, il nous a fallu définir une nomenclature (une liste des classes) de travail, contenant uniquement des classes pertinentes, tant du point de vue thématique (objectifs du projet) que technique (séparabilité des classes dans l'espace des données). Cette sélection des classes est très importante pour aboutir à une classification correcte. Par exemple, lorsqu'il s'agit des parcelles de cultures qui, à la date d'acquisition de l'image, sont en fin de saison culturale, la végétation peut être sèche, peu couvrante, voire en partie ou totalement récoltée, elle peut donc ressembler à un sol nu. Ces parcelles seront donc plutôt classées en sol nu, contrairement aux autres parcelles de végétation plus dense. Ainsi, plusieurs tests effectués sur les classes thématiques originellement définies sur le terrain montrent qu'il y a une importante confusion entre surfaces végétales non arborées et surfaces non végétales. Outre l'état des cultures cité précédemment, cette confusion vient aussi du fait que des parcelles appartenant à la classe des surfaces non végétalisées (les zones artificialisées, notamment) peuvent contenir de petits éléments végétaux (arbre isolé, par exemple, petit patch herbeux, jardin…). Cette complexité et cette hétérogénéité dans les caractéristiques de terrain engendrent une très grande variabilité des pixels au sein d'une même "parcelle", avec des recouvrements de classes. Cela rend difficile la distinction entre différentes catégories définies comme des objets lorsqu'on travaille à l'échelle du pixel, d'où la décision de finalement combiner les classes de cultures annuelles, autres zones herbacées, mais aussi sols nus et zones artificialisées. Dans la mesure où les objectifs de ce travail concernent la discrimination de différentes classes arborées, d'une part, et que ces classes fusionnées sont plus facilement classifiables avec d'autres méthodes traditionnelles orientées objet (SVM, Random Forest…), d'autre part, cela n'impacte pas le résultat attendu. D'autre part les étendues d'eau se distinguent nettement des autres zones dépourvues de végétation. Par ailleurs, la catégorie « autres arbres cultivés » ne contient qu'un nombre restreint d'exemples (moins d'une dizaine) et ne représente que des arbres isolés. Nous avons donc jugé plus prudent de supprimer cette classe en masquant les pixels labellisés sous cette classe et en les considérant comme s'ils étaient inconnus

.

Finalement, **six classes pertinentes** ont été retenues pour notre étude :

- **l'anacardier,**
- **le manguier**
- **le teck**
- **les forêts et savanes arborées**
- **les surfaces végétales non arborées (cultures annuelles, jachères herbacées, etc...) et les surfaces non végétales (sol nu, bâtiments et zones "artificialisées", routes),**
- **l'eau**

Dans la suite de notre processus, nous avons opté pour une approche à deux répertoires distincts. Nous avons filtré les images, écartant celles ne présentant pas d'annotations. Par la suite, pour équilibrer de manière optimale le nombre d'images et la préservation des annotations cruciales, nous avons introduit deux seuils de pixels noirs distincts : l'un fixé à 0,3 et l'autre à 0,7. Nous cherchons à travers ces deux répertoires à trouver un compromis entre la quantité d'informations annotées et la gestion des zones non labellisées au sein de chaque image. Suite à nos expérimentations, nous avons également testé le seuil 0,75 pour augmenter un peu plus le nombre des imagettes par rapport à 0.7 et avoir plus des exemples de nos classes

d’intérêt. Il est intéressant de noter que travailler avec un seuil plus élevé s'est avéré plus favorable, car cela a entraîné une amélioration dans la reconnaissance des pixels noirs. Finalement, on présente ci-dessous des figures décrivant notre jeu de données final.

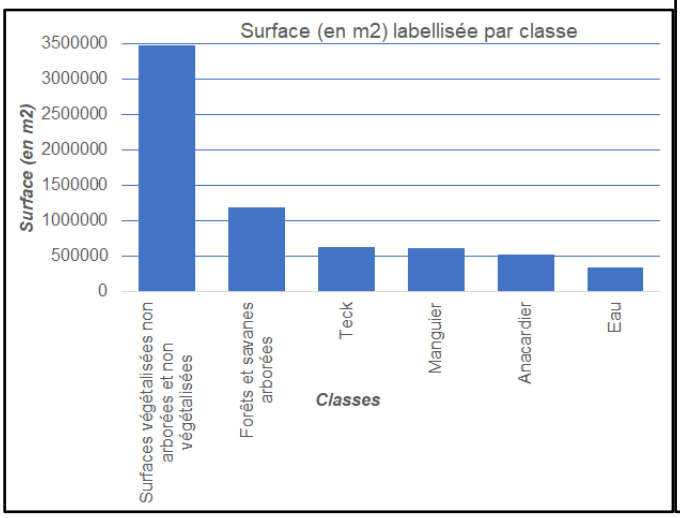


Fig. 11 : Surfaces totales labellisées par classe

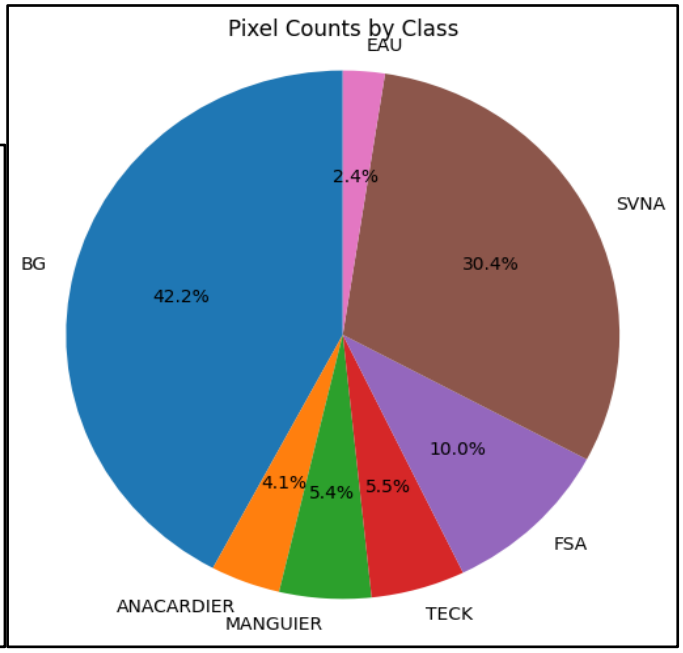


Fig. 12 : Proportions des pixels labellisés par classe

Nous avons mis en œuvre une approche de prétraitement des données pour améliorer la robustesse de notre modèle. Nous avons utilisé le générateur de données fourni par la bibliothèque Keras pour appliquer des transformations aux images de notre ensemble d'entraînement. Ces transformations incluent une rotation aléatoire dans une plage de 0 à 180 degrés, ainsi que des retournements horizontaux et verticaux aléatoires. Tout d'abord, cela nous permettra d’augmenter la diversité des données d'entraînement en créant des variations artificielles des images originales. Cette variabilité a permis à notre modèle d'apprendre à gérer des images sous différents angles et orientations, ce qui est particulièrement pertinent pour les images aériennes soumises à des conditions d'acquisition variées.

De plus, le générateur de données avec ces transformations permet d'augmenter artificiellement la taille de l'ensemble de données d'entraînement en créant des versions modifiées des images existantes. Cela peut aider le modèle à s'entraîner sur une gamme plus large de scénarios possibles et à mieux généraliser aux données de test.

4.4. Expérimentations et tests

Après avoir préparé notre jeu de données qui restera fixé pour les prochaines étapes, nous allons entreprendre une série de tests méthodiques visant à explorer et analyser divers paramètres de notre réseau. Cette approche implique la variation ciblée de quelques paramètres clés, dans le but de comprendre leur impact sur les performances du réseau. Une fois ces expérimentations réalisées, nous évaluerons les performances de chaque configuration résultante. Notre objectif ultime est de déterminer le compromis optimal entre ces paramètres, en tenant compte des objectifs spécifiques que nous avons fixés pour ce projet.

4.4.1. Focal Loss

La Focal Loss est une fonction de perte (loss function) spécialement conçue pour traiter le problème de la classification de données déséquilibrées, en particulier dans le contexte de l'apprentissage en profondeur (deep learning) et de la détection d'objets dans l'apprentissage supervisé. Elle a été introduite pour la première fois par Lin et al. dans leur article "Focal Loss for Dense Object Detection" en 2017.

La Focal Loss a pour objectif de donner plus de poids aux exemples difficiles (c'est-à-dire ceux qui sont mal classés avec une probabilité très faible) pendant l'entraînement du modèle, tout en atténuant le problème du déséquilibre de classe. Elle fonctionne en introduisant deux paramètres clés : l'indice de focalisation (gamma) et le paramètre d'équilibre (alpha).

La Focal Loss a pour objectif de donner plus de poids aux exemples difficiles (c'est-à-dire ceux qui sont mal classés avec une probabilité très faible) pendant l'entraînement du modèle, tout en atténuant le problème du déséquilibre de classe. Elle fonctionne en introduisant deux paramètres clés : l'indice de focalisation (gamma) et le paramètre d'équilibre (alpha).

Voici une explication plus détaillée du fonctionnement de la Focal Loss et de ses paramètres :

L'indice de focalisation (gamma) : L'indice de focalisation est un paramètre positif qui contrôle à quel point la Focal Loss donne plus de poids aux exemples difficiles. Plus gamma est élevé, plus la focalisation est forte sur les exemples mal classés. En d'autres termes, la Focal Loss accentue la perte pour les exemples qui sont difficiles à classer correctement par le modèle.

Le paramètre d'équilibre (alpha) : Le paramètre d'équilibre est un autre paramètre de la Focal Loss qui permet de gérer le déséquilibre de classe. Il est généralement utilisé lorsque les classes sont inégalement réparties dans les données d'entraînement. Si une classe est sous-représentée par rapport aux autres, on peut ajuster alpha pour donner plus de poids à cette classe, ce qui aidera le modèle à mieux apprendre à la distinguer.

La formule de la Focal Loss est la suivante :

$$\mathrm{FL}(p_\mathrm{t}) = -\alpha_\mathrm{t}(1-p_\mathrm{t})^\gamma \log(p_\mathrm{t}).$$

pt : est la probabilité prédite par le modèle pour la classe correcte.
αt : est le paramètre d'équilibre pour la classe cible.
γ est l'indice de focalisation.

En utilisant cette fonction de perte, les exemples faciles (ceux avec une probabilité élevée) ne contribuent pas de manière significative à la perte globale, tandis que les exemples difficiles (ceux avec une probabilité faible) reçoivent une focalisation accrue grâce à l'indice de focalisation.

La Focal Loss a été au cœur de notre exploration approfondie. Nous avons cherché à déterminer les deux meilleurs paramètres alpha et gamma en appliquant le même modèle (U-Net non pré-entraîné) à différentes combinaisons du couple (alpha, gamma) pour des valeurs comprises entre [0-1] pour alpha et 0,5 et 5 pour gamma [13]. Ces essais nous ont permis d'explorer un large éventail de possibilités, avec pour objectif de cerner les valeurs qui apportent les meilleurs résultats.

4.4.2. Résolution spatiale

Dans la mesure où nous nous intéressons à classifier des parcelles, c'est à dire des étendues spatiales homogènes du point de vue d'un type d'occupation du sol, et non à discerner des objets plus fins comme les arbres ou d'autres éléments isolés, et vu l'approche pixel choisie pour cette étude, nous nous sommes interrogés sur l'influence de la résolution spatiale sur le résultat de la segmentation sémantique. Potentiellement, une plus faible résolution pourrait autoriser une plus grande homogénéité des classes, et donc améliorer leur discrimination. Nous avons donc voulu tester la démarche sur des images de plus faible résolution spatiale, en considérant les images multispectrales originales non fusionnées avec le panchromatique, c'est-à-dire à 1,20m/pixel au lieu de 0,30 m/pixel.
Cependant, pour garantir le même cadrage spatial, c'est-à-dire conserver les mêmes zones couvertes dans nos nouvelles imagettes, nous avons dû réaliser un découpage de 128 x 128 pixels au lieu de 512x512 (un pixel de 1,20m de côté couvre en effet la même surface que 16 pixels de 30 cm de côté).
Nous avons donc appliqué à ce nouveau jeu d'imagettes les deux modèles considérés les plus performants sur le jeu initial, c'est à dire deux configurations différentes d'alpha et gamma de la focal loss : (gamma=2, alpha=0,25) et (gamma=2, alpha=0,75).

4.4.3. Pré-entraînement sur ImageNet

À cette étape de notre démarche, nous avons cherché à explorer l'impact du pré-entraînement sur l'ensemble de données ImageNet sur nos modèles U-Net. Pour cela, nous avons comparé les performances de deux modèles U-Net distincts : l'un ayant été pré-entraîné sur le vaste ensemble de données ImageNet[6], et l'autre ne bénéficiant pas de ce pré-entraînement préalable. Cette comparaison a permis de déterminer si le transfert d'apprentissage à partir d'ImageNet apportait des améliorations significatives à notre tâche spécifique de segmentation sémantique.

Post-traitement :

Enfin, nous avons abordé spécifiquement le problème de la surestimation des pixels noirs au sein des imagettes. Pour résoudre ce problème, nous avons mis en place un processus de post-traitement. Cette approche de post-traitement visait à corriger les prédictions générées par les modèles afin d'améliorer la qualité de la segmentation de la classe noire et à atténuer les prédictions excessives de pixels noirs dans les imagettes. Il est important de noter que nous n'avons pas cherché à améliorer spécifiquement la segmentation du fond dans ce contexte, et ce choix était basé sur des considérations pratiques. Bien que cela puisse avoir entraîné un biais dans les performances, nous avons opté pour cette approche pour des raisons d'efficacité au niveau des classes d'intérêt.

Pour ce faire, nous avons adopté une démarche en deux étapes. Tout d'abord, nous avons mis en place un algorithme pour contraindre le modèle à prédire la classe noire lorsque le pixel d'entrée est noir. Cette étape visait à réguler les prédictions erronées associées aux pixels noirs, contribuant ainsi à une meilleure délimitation des parcelles des autres classes.

Pour les points d'entrée qui n'étaient pas noirs, mais pour lesquels le modèle prédisait la classe noire, nous avons adopté une approche alternative. Au lieu de prendre en compte la probabilité associée à la classe du noire (la probabilité maximale), nous avons opté pour la deuxième probabilité la plus élevée comme prédiction finale. Cette démarche avait pour objectif de rectifier les prédictions incorrectes liées aux points d'entrée non noirs identifiés à tort comme faisant partie de la classe noire.
Les résultats des tests menés à cette phase ont été évalués en utilisant un modèle U-Net pré-entraîné et pas. Les performances ont été comparées dans deux scénarios distincts : avec l'application d'une étape de post-traitement (postprocess) et sans cette étape.

4.4.4. Comparaison entre U-Net et LinkNet

Tout en préservant les paramètres qui ont été identifiés précédemment, nous avons entrepris une nouvelle phase d'exploration en nous concentrant sur des architectures de modèles différentes. Cette étape visait à évaluer l'impact de l'architecture de l'encodeur sur les performances générales de notre système de détection. On utilise deux architectures différentes U-net et LinkNet, chacune avec deux encodeurs différents, resnet34 et resnet50, avec la focal loss avec la meilleure configuration obtenue (gamma=2, alpha=0.25)
U-Net avec ResNet34 : U-Net est une architecture de réseau de convolution qui est déjà relativement légère par nature. Lorsqu'elle est associée à l'encodeur ResNet34, elle conserve cette légèreté.

U-Net avec ResNet50 : En comparaison, l'utilisation de ResNet50 avec U-Net introduit une profondeur de réseau significativement plus grande. ResNet50 est capable de capturer des caractéristiques plus complexes dans les images, ce qui peut être crucial lorsque les données d'entrée contiennent des détails fins ou des variations subtiles. Cette configuration est plus adaptée aux scénarios où une précision plus élevée est requise, mais qui nécessite davantage des ressources de calcul.

LinkNet avec ResNet34 : LinkNet est une autre architecture qui, lorsqu'elle est couplée à ResNet34, se caractérise par sa capacité à rétablir des connexions entre les différentes couches du réseau (Ces connexions permettent de rétablir des liens directs entre des couches situées à des niveaux différents de la hiérarchie du réseau. Concrètement, cela signifie que des informations peuvent être transmises non seulement de manière séquentielle de l'entrée à la sortie, mais aussi directement entre des couches plus profondes et des couches moins profondes du réseau). Cette capacité de reconnexion est bénéfique pour maintenir des informations spatiales dans les données d'entrée, et elle fonctionne bien lorsque des détails spatiaux sont importants dans la tâche de segmentation.

LinkNet avec ResNet50 : L'utilisation de ResNet50 avec LinkNet apporte la même profondeur accrue que mentionnée précédemment. Cela signifie que cette configuration peut gérer des données d'entrée complexes avec des détails fins tout en maintenant une bonne compréhension des caractéristiques spatiales.

4.4.5. Comparaison entre U-Net et PSPNet

A cette étape nous comparons les performances de PSPNet avec U-Net, la spécificité de PSPNet réside dans sa capacité à traiter des images à plusieurs échelles, à agréger des informations contextuelles à partir de différentes résolutions et à maintenir des détails spatiaux importants. Ces caractéristiques en font un choix puissant pour la segmentation sémantique, en particulier lorsque les images contiennent une grande variabilité de taille d'objet et de contexte.

Similairement à ce qu'on a fait précédemment, on teste U-Net et PSPNet sans préentraînement avec la focal loss et la meilleure configuration obtenue.

4.4.6 Richesse des données en entrée (nombre de canaux)

Nous avons étendu notre exploration en incorporant différents canaux d'information (bandes spectrales, indices de texture) dans le processus de modélisation. Pour ce faire, nous avons évalué les performances d'U-Net en comparant plusieurs configurations de données, puis sélectionné le modèle correspondant à la configuration la plus performante. L'objectif était de vérifier notre hypothèse selon laquelle plus on utilise de canaux d'information en entrée, meilleurs sont les résultats.

Lors de ces expérimentations, nous avons comparé les résultats des modèles avec trois bandes spectrales d'une part et six bandes spectrales d'autre part, avec un accent particulier sur les bandes correspondant au vert, au rouge et au proche infrarouge, correspondant aux bandes les plus significatives par rapport aux caractéristiques de la végétation. Nous avons également examiné les performances du modèle utilisant neuf bandes : les six bandes spectrales plus les trois indices de texture, pour évaluer l'apport des informations contextuelles sur l'agencement des pixels voisins.

5. Résultats et discussion

5.1. Evaluation des performances pour la sélection d'un modèle

Nous sommes sur le point d'introduire les résultats et les conclusions des nombreuses expérimentations évoquées précédemment. Ces expériences ont été menées avec rigueur et méthodologie afin de répondre aux questions de recherche cruciales qui sous-tendent notre projet. Les analyses effectuées nous permettront de tirer des conclusions significatives et d'identifier des tendances importantes dans nos observations

5.1.1. Focal Loss

On montre les résultats des différentes combinaisons (alpha, gamma). On note que pour l'instant et jusqu'à ce qu'on indique une autre configuration, on travaille avec un U-Net non pré-entraîné.

	alpha=0,25	**alpha=0,5**	**alpha=0,75**	**alpha=1**
gamma=0,5	Model 1	Model 2	Model 3	Model 4
gamma=1	Model 5	Model 6	Model 7	Model 8
gamma=2	Model 9	Model10	Model 11	Model 12

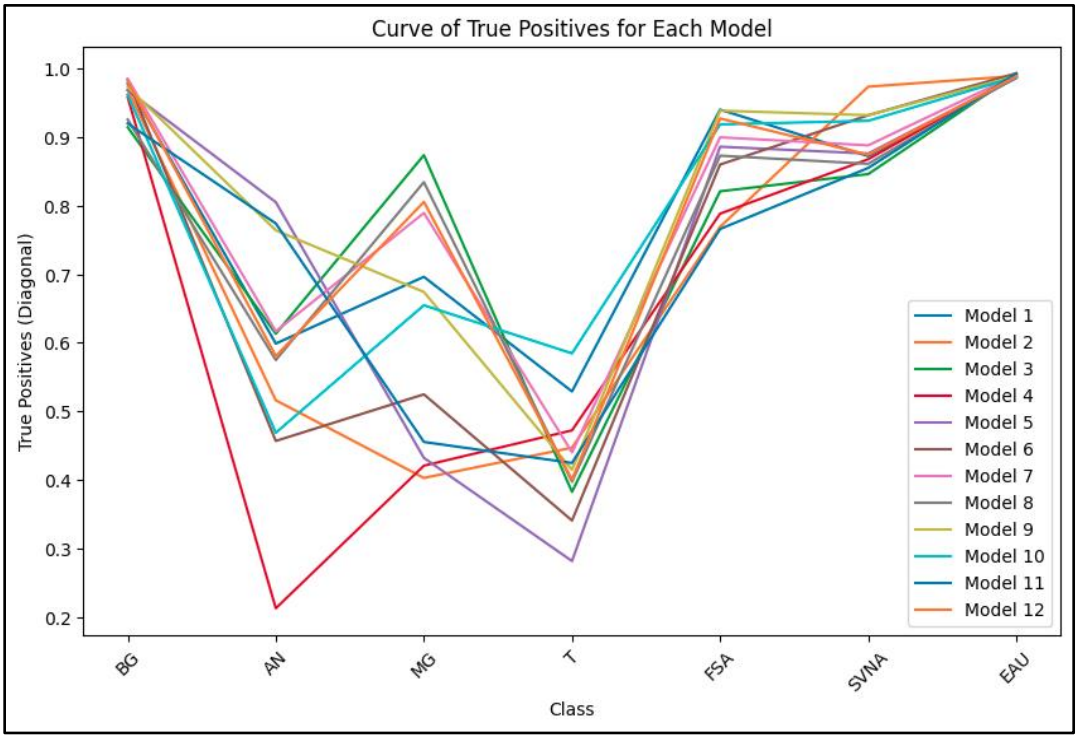


Fig. 13 : Valeurs normalisées issues des 12 modèles avec les 12 configurations différentes.

Renvoyer ici aux figures à analyser

Les résultats varient considérablement d'une configuration à l'autre, principalement au niveau de nos classes d'intérêt. Parmi ces classes, la moins bien identifiée est celle du "Teck" avec une précision globale moyenne de tous les essais de **0.425** une précision de **0.58**. Elle est souvent confondue avec la classe "Forêts et savanes arborées". Cette confusion est compréhensible étant donnée la forte similitude entre ces deux classes.

En ce qui concerne les classes "Anacardier" et "Manguier", de très bonnes performances sont atteintes avec certaines configurations : une précision maximale de **0,81** pour la classe des anacardiers avec la configuration (gamma=1, alpha=0,25 correspondant au modèle 5) et de **0,87** pour la classe des manguiers avec la configuration (gamma=0,5, alpha=0,75, correspondant au modèle 3). Mais malheureusement, pour le premier modèle la classe Manguier n'est reconnue qu'à 0,42% et pour le deuxième modèle la classe Anacardier n'atteint que 0,61% de précision. Il est donc très difficile de sélectionner la meilleure configuration et il va falloir faire des compromis. Aussi, on observe beaucoup de confusion pour ces classes avec les pixels noirs. Il est remarquable que dans les matrices de confusion, les premières lignes indiquent qu'il y a peu ou pas d'erreurs pour la détection des pixels noirs par les modèles. En revanche, dans les premières colonnes, de nombreuses erreurs sont observées, indiquant que les modèles ont tendance à classer des pixels non noirs comme étant des pixels noirs, en particulier pour les classes plus difficiles à identifier.

Suite à cette analyse approfondie, nous avons identifié le modèle qui a affiché les meilleures performances avec la Focal Loss. Les valeurs optimales des paramètres alpha et gamma de la Focal Loss, qui ont conduit aux meilleurs résultats, ont été identifiées comme étant **(gamma=2, alpha=0,25)**. Ce choix a été guidé par une amélioration significative de la précision globale, atteignant une valeur maximale de **0,81**. De plus, les

précisions de classification des classes d'intérêt, à savoir les anacardiers, les manguiers et les tecks, ont atteint un meilleur **compromis** avec **0,76, 0,67** et **0,41**.
Il est important de noter que dans les configurations où les anacardiers ont été mal classés, ils sont confondus avec les manguiers et vice versa. Cette confusion est attribuable à la grande similitude visuelle entre les deux types d'arbres sur nos images. Les caractéristiques distinctives qui permettent de les différencier, comme les feuilles ou la taille des couronnes ou autres, sont fines et peuvent échapper à la capacité discriminative de notre modèle, ce qui souligne une complexité fondamentale de notre tâche.

Nous avons ensuite procédé à une comparaison directe entre ce modèle et les résultats obtenus sans utiliser la Focal Loss, en utilisant la perte d'entropie croisée catégorique (categorical cross entropy). Les résultats de cette comparaison se sont avérés remarquablement éclairants. En effet, il était clair que l'application de la Focal Loss avait sensiblement amélioré les performances globales du modèle.
Cette démarche de test et de comparaison approfondie a validé l'efficacité de la Focal Loss pour notre tâche spécifique. Elle a également renforcé notre confiance dans les avantages de cette approche en termes d'optimisation de la fonction de perte.

5.1.2. Résolution spatiale

Renvoyer ici aux figures à analyser
Les résultats obtenus ont été particulièrement instructifs, mettant en évidence le rôle important de la résolution spatiale des imagettes sur les performances du modèle.

Les performances obtenues avec une résolution de 1,20m ont été tout aussi bonnes, voire meilleures, au niveau des trois dernières classes : forêts et savanes arborées, surfaces végétales non arborées et eau. En revanche, pour les classes spécifiques des objets forestiers, les résultats ont varié en fonction de la résolution. Par exemple, pour la classe des manguiers, les performances étaient très similaires avec une configuration (gamma=2, alpha=0,25), atteignant une valeur normalisée des vrais positifs de 0,57 avec une résolution de 1,20m et de 0,67 avec une résolution de 30 cm. En revanche, avec une configuration de (gamma=2, alpha=0,75), les performances étaient meilleures avec une résolution de 1,20m (0,60) par rapport à une résolution de 30 cm (0,42). Cependant, une grande disparité a été observée au niveau des classes de l'anacardier et du teck, où les performances avec une résolution de 1,20m étaient insuffisantes (0 pour les deux classes et les deux configurations). La résolution de 30 cm s'est avérée nettement plus performante dans ces cas.
En résumé, une résolution de 1,20m s'est avérée efficace pour les objets contrastés et facilement identifiables à l'œil nu, comme les zones forestières par rapport au sol nu ou végétalisé, ainsi que par rapport à l'eau. En revanche, pour les objets présentant moins de similarités, tels que différentes espèces d'arbres, l'identification s'est avérée nettement plus difficile avec une résolution plus élevée.
Le teck, de son côté, a connu une forte confusion avec la classe correspondant aux forêts et aux savanes arborées.

En poursuivant notre démarche, nous avons choisi d'ajouter un autre volet à nos tests en explorant l'effet de la résolution spatiale sur des patchs de dimensions plus grandes. Nous avons ainsi effectué des tests en utilisant des patchs de 256 x 256 pixels à la résolution de 1,20m. Cependant ces tests n'ont pas produit les

résultats escomptés. Les performances obtenues avec les patchs plus grands n'étaient pas satisfaisantes, principalement en raison de la limitation du nombre d'imagettes disponibles pour les entraîner et les évaluer. Cette limitation de données a eu un impact négatif sur la généralisation et la précision du modèle, ce qui a finalement rendu les résultats moins convaincants que ceux obtenus avec les patchs plus petits. Ces essais infructueux avec les patchs de plus grande taille ont souligné l'importance cruciale de la taille de l'ensemble de données dans le processus de modélisation. Bien que l'idée d'une plus grande résolution spatiale à travers des patchs plus grands semblait prometteuse, les contraintes inhérentes aux données disponibles ont finalement restreint la faisabilité et l'efficacité de cette approche alternative.

5.1.3. Préentraînement sur ImageNet

Renvoyer ici aux figures à analyser

Pour l'analyse des résultats issus du pré-entraînement du modèle U-Net, on analyse les matrices de confusion des deux expérimentations (avec préentraînement et sans préentraînement) avant postprocess, on constate que les résultats de l'entraînement from scratch (sans préentraînement) ont été meilleurs.

Sur les classes "forêts et savanes arborées", "surfaces végétales non arborées" on atteint respectivement les précisions 0,94 et 0,93 avec le modèle from scratch et 0,75 et 0,87 avec préentraînement, cette différence de performances a été aussi constaté au niveau de nos trois classes d'intérêt "anacardier", "manguier" et "teck", sur ces trois seulement, la précision moyenne (des trois classes) s'élève à 0,62 dans le cas de l'entraînement from scratch et 0,5 dans le cas du préentraînement.

La différence notable de performances entre le modèle entraîné "from scratch" (sans préentraînement) et le modèle préentraîné peut être en grande partie attribuée à la divergence fondamentale entre la base de données ImageNet et notre domaine d'images aériennes. En effet, ImageNet est principalement axé sur la reconnaissance d'objets tels que des espèces animales ou des objets de la vie quotidienne, comme des bureaux ou des ballons, qui présentent des caractéristiques visuelles très différentes de celles que nous rencontrons dans notre ensemble de données. Dans le contexte des images aériennes, les structures, les motifs visuels, et les éléments à reconnaître diffèrent considérablement de ceux présents dans les images d'ImageNet. Cette disparité rend les transferts de connaissances depuis un modèle préentraîné sur ImageNet moins efficaces pour notre tâche de classification spécifique. En conséquence, le modèle "from scratch" a eu l'opportunité d'apprendre des caractéristiques mieux adaptées à notre domaine d'intérêt, expliquant ainsi les performances supérieures observées dans certaines classes.

5.1.4. Post-processing

Renvoyer ici aux figures à analyser

En reconsidérant les points noirs au sein des imagettes, les résultats ont révélé une amélioration significative des vrais positifs pour chaque classe. Les résultats ont mis en lumière une amélioration significative des vrais positifs pour chaque classe. Cette tendance révèle que la deuxième probabilité prédite par le modèle, lorsque la prédiction initiale était la classe noire, est en grande partie cohérente avec la classe correcte. En effet, par exemple au niveau du deuxième exemple celui qui est à droite sur la figure, presque la moitié des 42 % des pixels des anacardiers qui ont été prédit comme des pixels noirs ont été remis à leurs classe correcte et une autre partie a été remise à la classe des manguiers ce qui est raisonnable vu la similarité entre ces deux classes avec une petite partie restante distribuée entre les autres classes. De la même manière, pour la

classe du teck, les 57 % des pixels du teck qui ont été prédit comme classe noire ont été remis en grande partie à leur classe correcte "teck", augmentant la précision de cette classe de 0,20 à 0,60 et le reste qui a été classé comme "forêts et savanes arborés", ce qui reste raisonnable.
Cette évolution reflète une amélioration globale de la précision des prédictions.
Ainsi, sur le plan pratique, nous ne nous intéressons pas à la classe noire car, en réalité, elle n'existe pas. Nous l'avons introduite pour résoudre le problème de la dispersion des annotations. Par cette approche, on exclut cette classe de nos évaluations des performances et on obtient des résultats plus raisonnables et plus corrects.
Ces conclusions soulignent la valeur de l'approche de post-traitement que nous avons adoptée.

5.1.5. Comparaison avec d'autres architectures

Renvoyer ici aux figures à analyser.

Il est essentiel de noter que les résultats présentés ici sont basés sur la configuration spécifique des paramètres alpha et gamma de la focal loss 2 pour gamma et 0,75 pour alpha.
Il convient de mentionner qu'avec le post-traitement que nous avons appliqué, nous avons constaté que cette configuration spécifique a donné de bien meilleurs résultats par rapport à la configuration avec les valeurs de 2 pour gamma et 0,25 pour alpha.

Lorsque nous comparons les performances d'U-Net avec ResNet34 et U-Net avec ResNet50, nous observons une nette amélioration des résultats. Initialement, avec U-Net et ResNet34, nous avions des précisions de 0,62 en anacardier, 0,63 en manguier et 0,71 en teck. Cependant, en utilisant U-Net avec ResNet50, ces performances ont considérablement augmenté, atteignant respectivement 0,79, 0,83 et 0,75. Ces améliorations témoignent de l'impact positif de l'utilisation d'un encodeur plus profond et complexe pour extraire des caractéristiques plus riches des images aériennes.
On procède à comparer maintenant, les modèles U-Net avec ResNet 50, LinkNet avec ResNet 50 et PSPNet avec ResNet 50. U-Net atteint une précision globale de 0,90, LinkNet de 0,89 et PSPNet de 0,91.
Au niveau des trois classes d'intérêt, PSPNet l'emporte sur les deux réseaux avec une précision de 0,91 pour les anacardiers, 0,81 pour les manguiers et 0,82 pour les tecks.
Même si au niveau des précisions les valeurs sont élevées, on constate au niveau des prédictions du modèle que les performances ne sont pas aussi améliorées, le modèle manque de précision contextuelle locale et généralise d'une manière globale au niveau des prédiction de chaque imagette, en effet, les grandes parcelles dominantes sont bien classées mais au niveau des petites parcelles on perd de précision.
Ceci rend les modèles d'U-net et LinkNet plus convaincant au niveau des prédictions et au niveau des performances. Entre les deux, au niveau des précisions, ils sont très proches mais U-net semble offrir des performances de prédiction légèrement meilleures.

5.1.6. Richesse des données en entrée (nombre de canaux)

Renvoyer ici aux figures à analyser.

Les figures montrent les résultats obtenus avec U-Net, en appliquant la focal loss avec les valeurs gamma=2 et alpha=0,75, avec resnet50 comme encodeur, pour les 3 tests d'entraînement respectifs : 3 bandes, 6 bandes, et 9 bandes.

L'ajout de bandes spectrales a eu un impact significatif sur les résultats de notre modèle de segmentation. Au départ, en utilisant seulement 3 bandes spectrales, nous avons constaté que la différenciation entre l'anacardier et le manguier était très difficile, avec des précisions de seulement 0 pour l'anacardier et 0,89 pour le manguier. De plus, le taux de confusion entre ces deux classes était élevé, atteignant 0,89, ce qui signifie que de nombreux pixels d'anacardier ont été incorrectement classés comme manguier. De même, le teck était fortement confondu avec la classe "forêts et savanes arborés", avec des précisions de seulement 0,06 pour le teck et 0,95 pour "forêts et savanes arborés", et un taux de confusion entre les deux atteignant 0,83.
Cependant, lorsque nous avons augmenté le nombre de bandes à 6, nous avons observé une amélioration significative des performances, en particulier pour les classes d'anacardier et de teck, qui ont atteint des précisions respectives de 0,15 et 0,31. Il est important de noter que cette augmentation du nombre de bandes a légèrement diminué la précision pour la classe des manguiers de 0,05 et pour la classe "forêts et savanes arborés" de 0,01.
Enfin, en utilisant 9 bandes spectrales, nous avons atteint des précisions considérablement améliorées pour toutes les classes qui étaient précédemment confondues. Les précisions sont passées à 0,79 pour l'anacardier, 0,83 pour le manguier, 0,75 pour le teck et 0,95 pour "forêts et savanes arborés". Cela a également eu un impact positif sur la précision globale du modèle, qui est passée à 0,9 par rapport à 0,74 avec 6 bandes et 0,69 avec seulement 3 bandes. Ces résultats soulignent l'importance de l'information spectrale supplémentaire et l'information texturale pour différencier efficacement ces classes de végétation dans les images satellitaires.

5.2. Application du modèle

Renvoyer ici aux figures à analyser.
Nous appliquons le modèle développé à une vaste zone sélectionnée à partir de l'image de référence. Cette expansion de notre démarche vers une échelle plus grande vise à évaluer la robustesse et la généralisation du modèle sur des terrains variés, sur lesquels on a aucune annotation. Cette extension de notre évaluation sur une vaste zone sélectionnée à partir de l'image de référence nous permettra de mieux comprendre comment le modèle se comporte sur des imagettes non masquées. Jusqu'à présent, notre démarche a impliqué un masquage des zones non labellisées pour améliorer la qualité des données d'entraînement et la validation des performances se fait sur des imagettes masquées.

5.3. Limites de l’approche

5.3.1. Annotations et image de référence

L'une des principales limitations réside dans la qualité de la base de données utilisée. Malgré nos efforts pour travailler avec les données disponibles, nous avons dû faire face à des contraintes en ce qui concerne l'image de référence et les annotations associées. Les annotations étaient peu nombreuses et leur qualité variait considérablement. Les contours des parcelles n'étaient pas toujours correctement tracés, ce qui entraînait des chevauchements entre les classes et des difficultés dans la segmentation précise. Parfois, des erreurs rares étaient présentes dans certaines annotations. Par exemple, dans une situation, une parcelle d'arbres

contenait une zone de sol nu. Cette erreur a entraîné une identification incorrecte de cette zone comme étant du sol nu, alors qu'il s'agissait clairement d'une parcelle d'arbres. Ainsi, nous avons fait usage d'un nombre limité d'imagettes pour notre travail. Bien que nous ayons eu seulement 443 imagettes de dimensions 512x512, ce qui est restreint en matière de deep-learning. De plus, chaque imagette était soumise à un masquage des zones inconnues et seules les imagettes contenant au minimum 25% de pixels porteurs d'informations pertinentes (c'est-à-dire dont la classe est connue) ont été conservées dans l'approche. Cela nous a empêché de tirer pleinement parti de l'ensemble des imagettes disponibles.

5.3.2. Adéquation de la tâche au jeu de données

La nature même de la tâche de segmentation des parcelles et à sa pertinence vis-à-vis du jeu de données peuvent aussi être discutées au regard de la variabilité des caractéristiques des parcelles à détecter. Certaines parcelles étaient facilement identifiables, tandis que d'autres suscitaient des hésitations quant à leur identification en tant que parcelles ou objets isolés. Par exemple, quelques arbres isolés étaient annotés et considérés comme des parcelles d'arbres, bien que cela ne corresponde pas à la définition de parcelles. En revanche, certaines parcelles étaient classées comme telles, mais contenaient également des objets isolés d'autres classes. Cette hétérogénéité dans la taille et la composition des parcelles a été un problème pour la segmentation précise. Aussi, les parcelles annotées ont des tailles très variées, certaines parcelles se son être y compris selon les classes, certaines pouvant être très grandes comme les surfaces en eau ou non végétalisées.

5.3.3. Hétérogénéité des classes d'intérêt

L'hétérogénéité des classes d'intérêt, en particulier les arbres, est aussi une grande source d'instabilité de la méthode. Rappelons que nous travaillons ici dans un contexte naturel soumis à diverses variables, telles que l'âge des arbres, leur taille de couronne, leur intensité de développement végétatif et de croissance, leur vitalité et leur phénologie. Ces variations naturelles intrinsèques à l'objet d'étude entraînent une grande diversité de signatures (spectrale, texturale, structurelle…) au sein d'une classe d'arbres donnée, qui s'étale donc largement dans l'espace des variables. Deux classes d'arbres distinctes peuvent aussi se chevaucher partiellement dans cet espace des variables, car deux arbres d'espèces différentes peuvent présenter des signatures similaires selon leur état physiologique. Cela entraîne donc une plus grande complexité dans la tâche de segmentation sémantique.
Par ailleurs, le choix des classes sélectionnées dans ce projet a été le résultat de compromis entre besoin thématique, capacités méthodologiques et résultats d'expérimentations, afin d'être à la fois pertinentes et adaptées au contexte de l'étude. Mais ce choix n'est pas figé et peut toujours bénéficier d'améliorations futures car il reste une marge d'amélioration en termes de définition et de délimitation des classes, qui pourrait notamment être revisitée à la lumière de nouvelles informations de référence acquises sur le terrain.

5.3.4. Découpage de l'image

La méthode de découpage des images en imagettes n'est pas totalement optimisée, et cela pour plusieurs raisons. Tout d'abord, les images ont été découpées sans chevauchement entre les imagettes. Bien que ce découpage sans chevauchement ait été motivé par la nécessité d'éviter que des parties des imagettes d'entraînement ne se retrouvent dans l'ensemble de test lors de la sélection aléatoire, il a abouti à la découpe des parcelles en morceaux répartis entre différentes imagettes. Cette approche pourrait potentiellement influencer les résultats, car certaines informations contextuelles pouvant améliorer la segmentation pourraient être perdues. De plus, lors de la sélection des imagettes, l'utilisation du seuil de 0.75 sur les pixels noirs a

supprimé un certain nombre d'imagettes ayant légèrement plus de pixels noirs mais pouvant contenir des zones d'intérêt pour l'apprentissage (appartenant par exemple à des classes peu représentées dans la base de référence). Cela a restreint d'autant plus la quantité de données disponibles pour l'apprentissage. En somme, ces limitations soulignent les défis rencontrés dans la conception du processus d'apprentissage et la nécessité de trouver un équilibre entre les choix méthodologiques et la qualité des données d'entraînement.

5.3.5. Validation croisée et fiabilité

Nous avons essayé d'entamer dans notre approche la validation croisée du modèle, mais ceci fut complexe. L'un des défis majeurs est de garantir que chaque partition de l'ensemble de données contienne une distribution représentative de toutes les classes, y compris les classes minoritaires. Dans le cas contraire, la performance du modèle peut être biaisée en faveur des classes majoritaires, car il sera moins exposé aux classes minoritaires lors de l'entraînement et de la validation. Cela peut conduire à une sous-estimation des performances réelles du modèle pour les classes moins fréquentes.
Une des limitations importantes de ne pas utiliser la validation croisée dans notre contexte, réside dans la fiabilité de l'évaluation des performances du modèle. La validation croisée est une technique essentielle qui permet de diviser le jeu de données en plusieurs ensembles d'entraînement et de validation, ce qui permet d'évaluer le modèle sur plusieurs partitions différentes des données. Cela signifie que les performances du modèle peuvent être biaisées par la manière dont les données ont été divisées en ensembles d'entraînement et de validation. Si, par exemple, le modèle est évalué sur un ensemble de validation qui est non représentatif de la diversité des données réelles, les résultats peuvent être sur-optimistes ou pessimistes.
Il faut noter que notre jeu de données de test utilisé tout au long de nos expérimentations respecte la représentativité des classes en termes de nombre, mais on ne sait rien au niveau de la représentativité de la diversité des données réelles.

5.3.5. L'instabilité des modèles

Nous avons constaté sur notre projet une instabilité des résultats au niveau des trois classes d'intérêt surtout, les performances sont généralement assez stables au niveau des autres classes, ce qui pourrait être inquiétant et moins fiable pour notre approche comparative. Cette instabilité pourrait être due au fait que les modèles ont été initialisés avec des poids différents avant l'entraînement ce qui peut avoir un impact significatif sur la convergence du modèle et ses performances finales. Ainsi, ceci pourrait être dû à la taille du jeu de données d'entraînement, en outre, une petite taille d'ensemble de données peut augmenter la variabilité des performances du modèle, car il peut ne pas être en mesure d'apprendre toutes les variations présentes dans les données.

6. Conclusion et perspectives

Ce projet a été une exploration en profondeur des réseaux de neurones exploitant l'apprentissage profond (deep learning) pour la classification des différentes catégories d'occupation du sol dans des images satellites. Notre objectif central était la reconnaissance des plantations arborées, notamment les anacardiers, les manguiers, le teck, ainsi que d'autres zones forestières. Notre démarche a impliqué une série de tests avec diverses architectures de réseau, en manipulant plusieurs paramètres clés.

Nous avons commencé par une phase préliminaire, où nous avons affiné le nombre de classes et leur regroupement, en nous basant sur des expérimentations pour trouver un équilibre entre le nombre d'imagettes et le pourcentage de pixels noirs. Une fois notre jeu de données établi, comprenant 443 imagettes finales

avec un masquage de 75% au maximum (ce qui signifie qu'au plus 75% des pixels dans une imagette sont noirs), et comprenant 7 classes, dont 3 principales ("anacardier", "manguier" et "teck"), nous avons entrepris une série de tests pour évaluer divers paramètres et leur impact sur les performances finales.

Notre exploration a commencé par la manipulation de la fonction de perte "focal loss", en variant ses paramètres alpha et gamma. Ensuite, nous avons ajusté la résolution spatiale à 1m20, suivi d'expérimentations avec le transfert de connaissances en utilisant un réseau pré-entraîné. Finalement, nous avons testé deux autres architectures spécifiquement conçues pour la segmentation, à savoir LinkNet et PSPNet. Cette démarche méthodique nous a permis de mieux comprendre les différents paramètres et de sélectionner les configurations optimales pour atteindre nos objectifs de classification d'images aériennes.

Il s'avère que le modèle U-Net avec resnet50, dans sa configuration avec utilisation de la focal loss (alpha=0.75, gamma=2) et non préentraîné, appliqué à 6 bandes spectrales et 3 indices de textures, permet de classifier les plantations d'anacardiers avec une précision de 0,79, les plantations de manguiers avec une précision de 0,83, les plantations de teck avec une précision de 0.75 et les forêts et savanes arborées avec une précision de 0.95. Les classes non arborées affichent quant à elles une précision respectivement de 1 pour l'eau et 0.99 pour les autres surfaces.

Il est important de noter que malgré l'instabilité des résultats observée dans notre approche, certaines tendances semblent se confirmer de manière cohérente. En particulier, il est généralement possible de distinguer avec succès les anacardiers et les manguiers des autres espèces forestières, y compris le teck. Cependant, la confusion entre les anacardiers et les manguiers peut varier en fonction des différentes configurations de modèle et de paramètres. De même, le teck se distingue nettement des autres classes, mais il présente toujours une certaine confusion avec les classes de forêts et de savanes arborées. Cette cohérence dans les tendances de distinction malgré l'instabilité globale peut être un point de départ pour des améliorations futures dans notre modèle

Bien que notre projet ait produit des résultats encourageants, il est important de noter quelques limitations. La qualité variable des annotations et le nombre limité d'imagettes (443) ont restreint la robustesse de notre modèle. L'adaptation aux caractéristiques diverses des parcelles et l'hétérogénéité des classes, en particulier les arbres, ont posé des défis de segmentation. De plus, le découpage sans chevauchement des images en imagettes a pu entraîner la perte d'informations contextuelles importantes. Enfin, l'absence de validation croisée soulève des questions sur la fiabilité de l'évaluation. Pour l'avenir, l'amélioration des annotations, une meilleure adaptation aux variations des parcelles, un découpage d'image plus optimal et une validation croisée robuste sont des axes à explorer pour surmonter ces limites.
Pour l'avenir, plusieurs axes d'amélioration s'offrent à nous. Tout d'abord, il est impératif d'améliorer la qualité des annotations, en mettant l'accent sur une délimitation plus précise des parcelles et une réduction des erreurs rares. Nous devons également adapter notre modèle pour mieux gérer la variabilité des caractéristiques des parcelles. Un découpage d'image plus optimal, avec un chevauchement contrôlé entre les imagettes, pourrait préserver les informations contextuelles importantes. De plus, il est crucial d'explorer des techniques de validation croisée robuste pour des évaluations plus fiables des performances.

ANNEXES

1 Fonctionnement de PSPNet

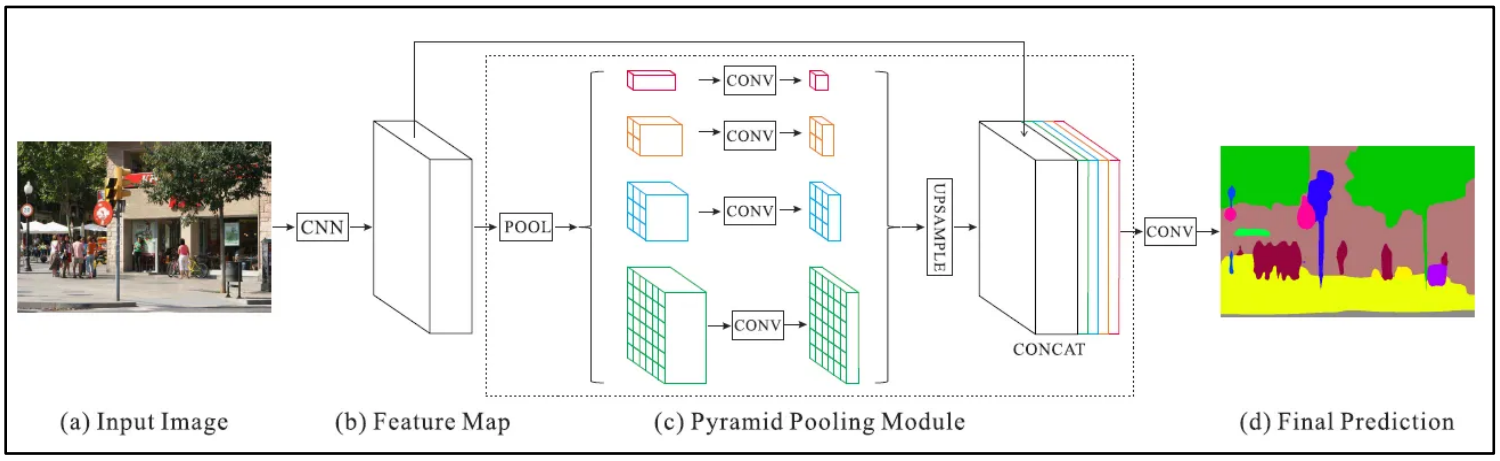


Fig. architecture de PSPNet

(a) et (b)

En (a), nous avons une image d'entrée.

En (b), ResNet est utilisé avec une stratégie de réseau dilaté utilisant la convolution dilatée (également connue sous le nom de convolution à trous ou convolution à pas dilaté, est une opération couramment utilisée dans les réseaux de neurones convolutionnels (CNNs) pour le traitement d'images. Elle permet d'augmenter la taille effective du champ réceptif d'une couche de convolution sans augmenter le nombre de paramètres. La convolution traditionnelle utilise un noyau de filtre (kernel) avec une taille fixe et se déplace de manière contiguë sur l'image d'entrée. En revanche, la convolution dilatée introduit un paramètre supplémentaire appelé "facteur de dilatation" ou "pas de dilatation" (dilated rate en anglais). Ce facteur de dilatation détermine l'espacement entre les échantillons échantillonnés dans le noyau de filtre lors de la convolution) pour extraire des fonctionnalités. La taille de la carte des caractéristiques est de 1/8 de l'image d'entrée ici.

(c).1. Mise en commun moyenne (average pooling)[4] des sous-régions

En (c), la mise en commun moyenne des sous-régions est effectuée pour chaque carte de caractéristiques.

Rouge : Il s'agit du niveau le plus grossier qui effectue une mise en commun moyenne globale (global average pooling[5]) sur chaque carte de caractéristiques, pour générer une sortie de bac unique.

Orange : Il s'agit du deuxième niveau qui divise la carte des caractéristiques en sous-régions 2 × 2, puis effectue une mise en commun moyenne pour chaque sous-région.

Bleu : Il s'agit du troisième niveau qui divise la carte des caractéristiques en 3 × 3 sous-régions, puis effectue une mise en commun moyenne pour chaque sous-région.

Vert : Il s'agit du niveau le plus fin qui divise la carte des caractéristiques en sous-régions 6 × 6, puis effectue une mise en commun pour chaque sous-région.

(c).2. Convolution 1 × 1 pour la réduction des dimensions

[4] Le pooling moyen est une opération de regroupement qui calcule la valeur moyenne des correctifs d'une carte de fonctionnalités et l'utilise pour créer une carte de fonctionnalités sous-échantillonnée (regroupée)

[5] Global Average Pooling est une opération de pooling conçue pour remplacer les couches entièrement connectées dans les CNN classiques. L'idée est de générer une carte de caractéristiques pour chaque catégorie correspondante de la tâche de classification dans la dernière couche

Ensuite, une convolution 1 × 1 est effectuée pour chaque carte de caractéristiques regroupée afin de réduire la représentation du contexte à 1/ N de celle d'origine (noire) si la taille du niveau de la pyramide est N.

Dans cet exemple, N =4 car il y a 4 niveaux au total (rouge, orange, bleu et vert).
Si le nombre de cartes de fonctionnalités en entrée est de 2048, alors la carte de fonctionnalités en sortie sera (1/4) × 2048 = 512, soit 512 nombres de cartes de fonctionnalités en sortie.
(c).3. Interpolation bilinéaire pour le suréchantillonnage
Une interpolation bilinéaire est effectuée pour suréchantillonner chaque carte de caractéristiques de faible dimension afin d'avoir la même taille que la carte de caractéristiques d'origine (noire).

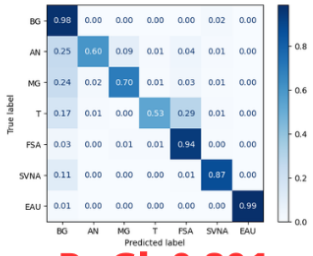
(c).4. Concaténation pour l'agrégation de contexte
Tous les différents niveaux de cartes de fonctionnalités suréchantillonnées sont concaténés avec la carte de fonctionnalités d'origine (noire). Ces cartes de fonctionnalités sont fusionnées en tant que priorité globale. C'est la fin du module de pooling pyramidal en (c).

(d)
Enfin, elle est suivie d'une couche de convolution pour générer la carte de prédiction finale en (d).

2 Figures des résultats

2.1 Résultats d'U-net avec les différentes configurations de la focal loss

0.5 gamma 0.25 alpha

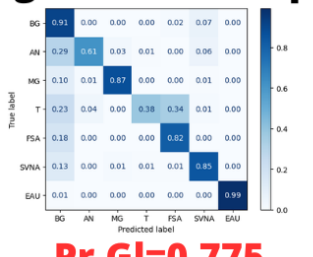


Pr.GI=0.801

0.5 gamma 0.5 alpha

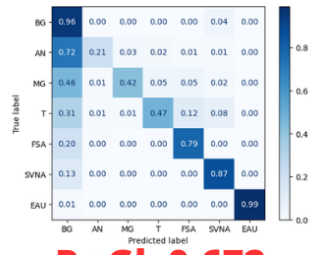
Pr.GI=0.724

0.5 gamma 0.75 alpha



Pr.GI=0.775

0.5 gamma 1 alpha

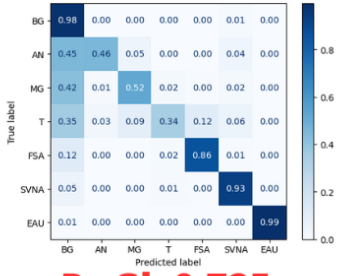


Pr.GI=0.672

1 gamma 0.25 alpha

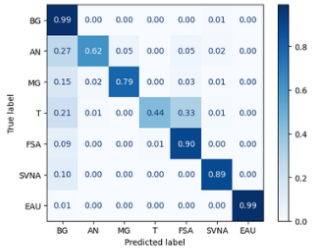
Pr.GI=0.750

1 gamma 0.5 alpha



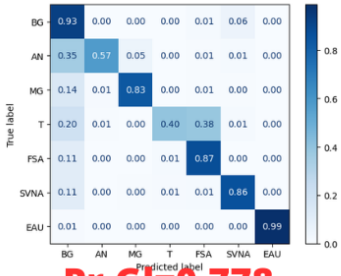
Pr.GI=0.725

1 gamma 0.75 alpha



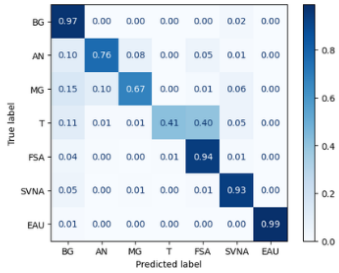
Pr.GI=0.802

1 gamma 1 alpha



Pr.GI=0.778

2 gamma 0.25 alpha

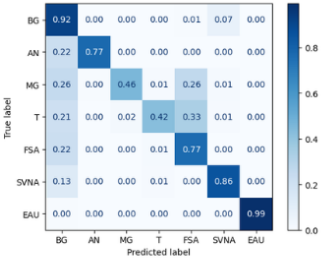


Pr.GI=0.810

2 gamma 0.5 alpha

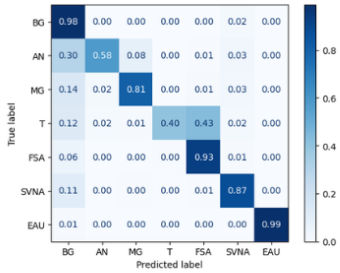
Pr.GI=0.784

2 gamma 0.75 alpha



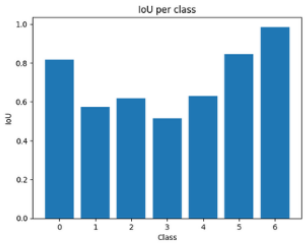
Pr.GI=0.741

2 gamma 1 alpha



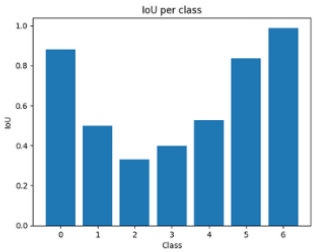
Pr.GI=0.794

0.5 gamma 0.25 alpha



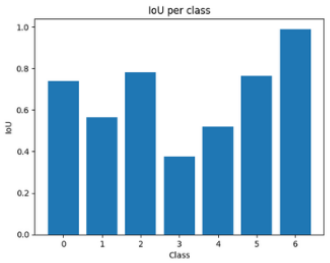
Pr.GI=0.801

0.5 gamma 0.5 alpha



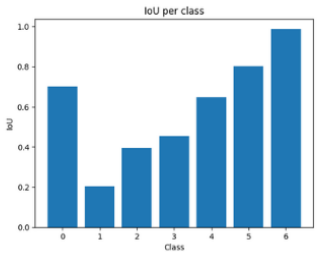
Pr.GI=0.724

0.5 gamma 0.75 alpha



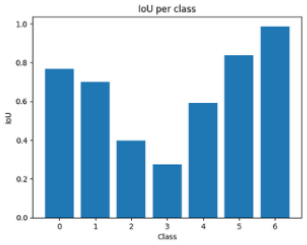
Pr.GI=0.775

0.5 gamma 1 alpha



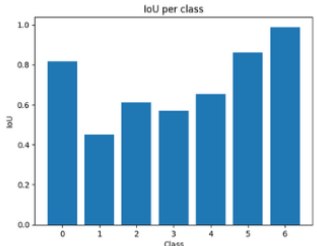
Pr.GI=0.672

1 gamma 0.25 alpha



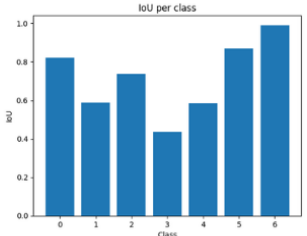
Pr.GI=0.750

1 gamma 0.5 alpha



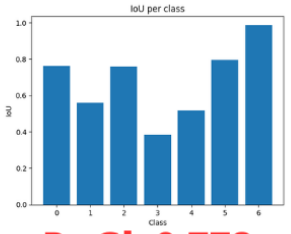
Pr.GI=0.725

1 gamma 0.75 alpha



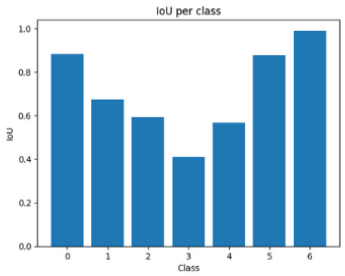
Pr.GI=0.802

1 gamma 1 alpha



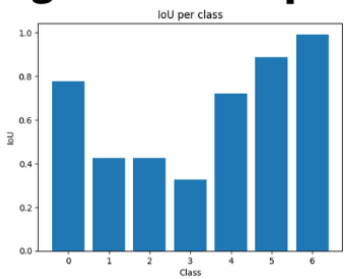
Pr.GI=0.778

2 gamma 0.25 alpha



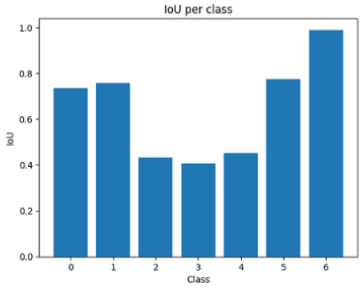
Pr.GI=0.810

2 gamma 0.5 alpha



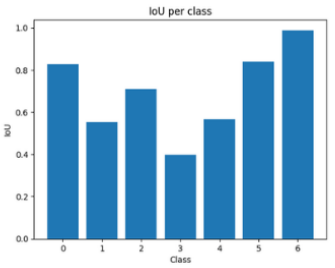
Pr.GI=0.784

2 gamma 0.75 alpha



Pr.GI=0.741

2 gamma 1 alpha

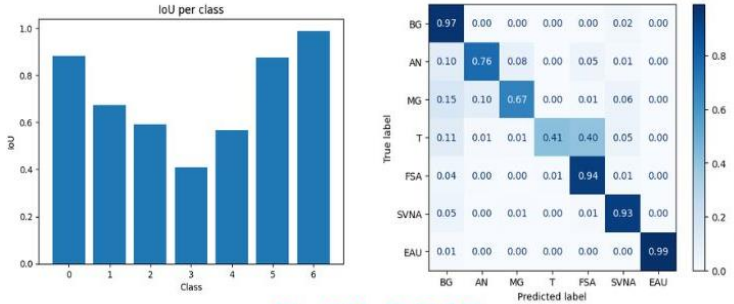


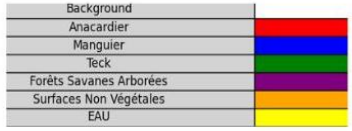
Pr.GI=0.794

2.2 Comparaison avec la fonction de perte d'entropie croisée catégorique

Comparaison de la meilleure configuration de la focal loss avec l'Entropie croisée catégorielle

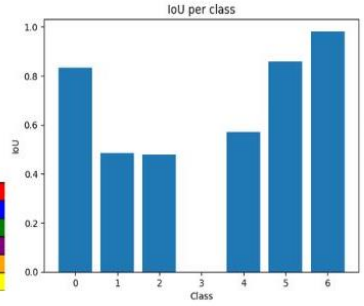
Modèle Unet avec la focal loss , avec les paramètres 2 gamma 0,25 alpha

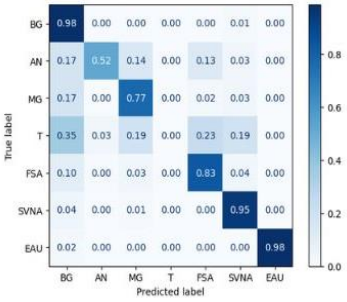




Pr.Gl=0.810

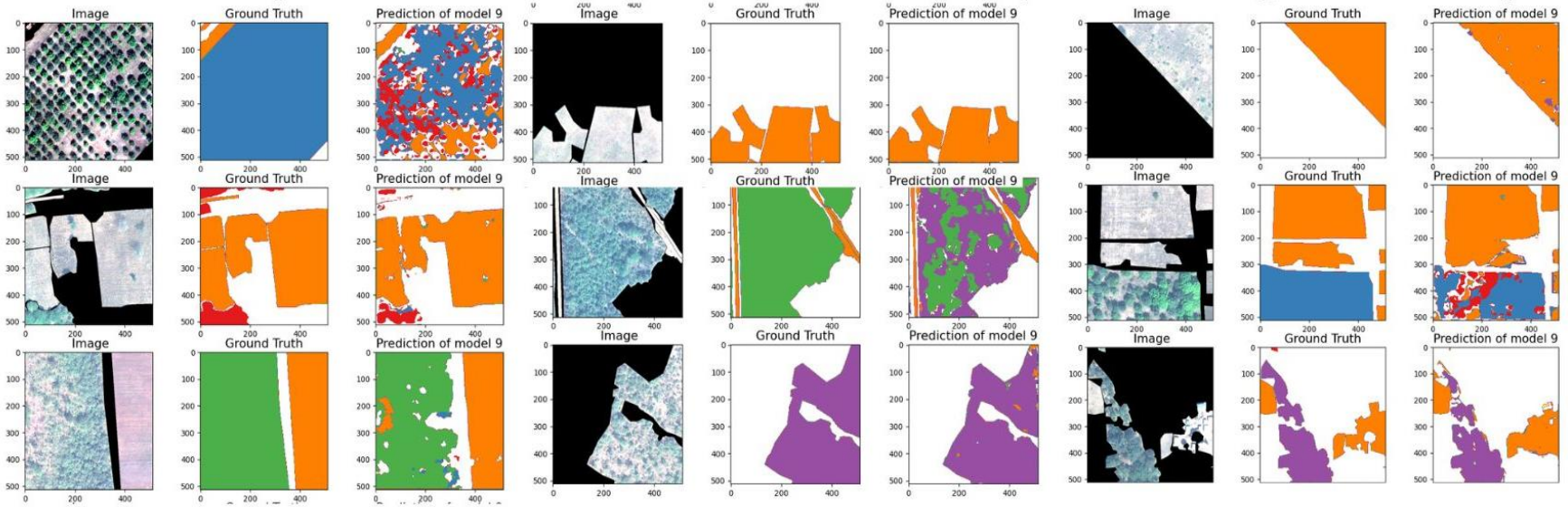
Modèle Unet sans focal loss





Pr.Gl=0.718

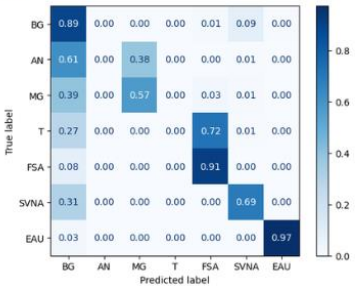
Modèle 9 correspond au modèle avec la focal loss , avec les paramètres 2 gamma 0,25 alpha



2.3 Résultats d'U-net avec deux résolutions différentes

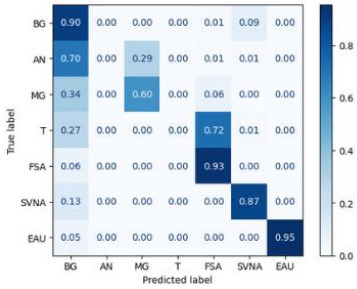
Modèles U-net entraînés sur les imagettes de 128*128 (résolution 1m20)

gamma 2 alpha 0,25



Pr.Gl=0.575

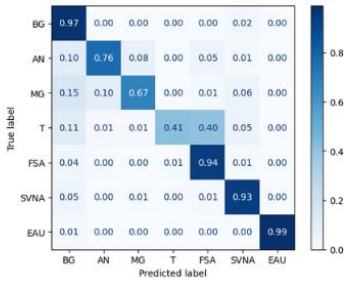
gamma 2 alpha 0,75



Pr.Gl=0.607

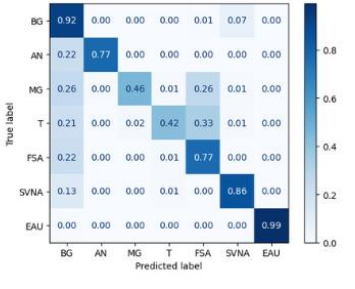
Modèles U-net , même configurations entraînés sur les imagettes de 512*512 (résolution 30cm)

gamma 2 alpha 0,25



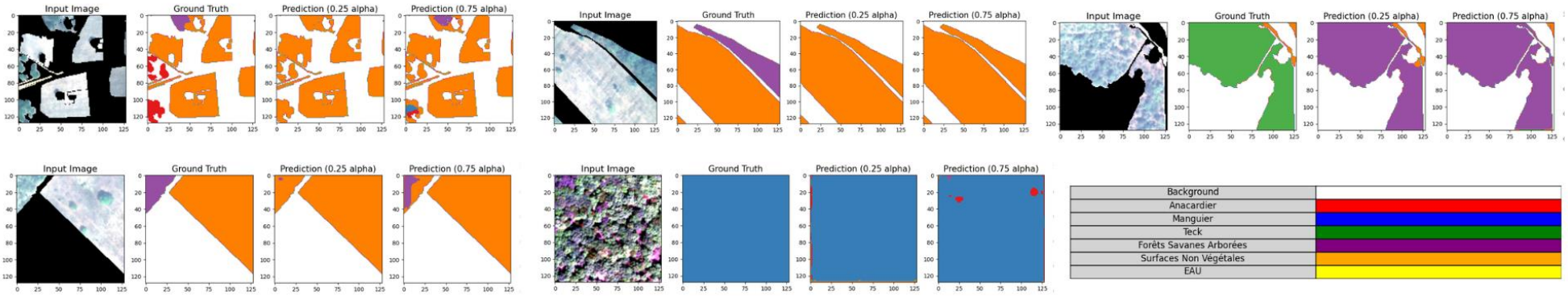
Pr.Gl=0.810

gamma 2 alpha 0,75



Pr.Gl=0.741

Quelques prédictions des modèles unet entraînés sur les imagettes de 128*128

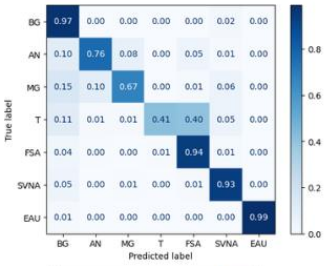


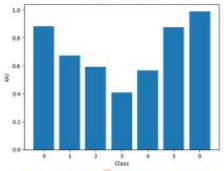
2.4 Résultats de préentraînement et de postprocess

Modèle Unet avec la focal loss , avec les paramètres 2 gamma 0,25 alpha,non préentraîné sur imagenet

Model 1

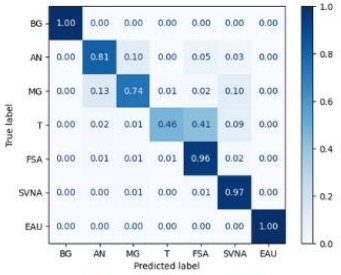
Avant post traîtement

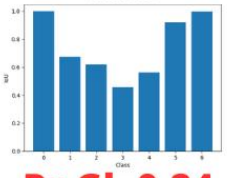




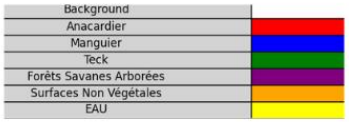
Pr.Gl=0.81

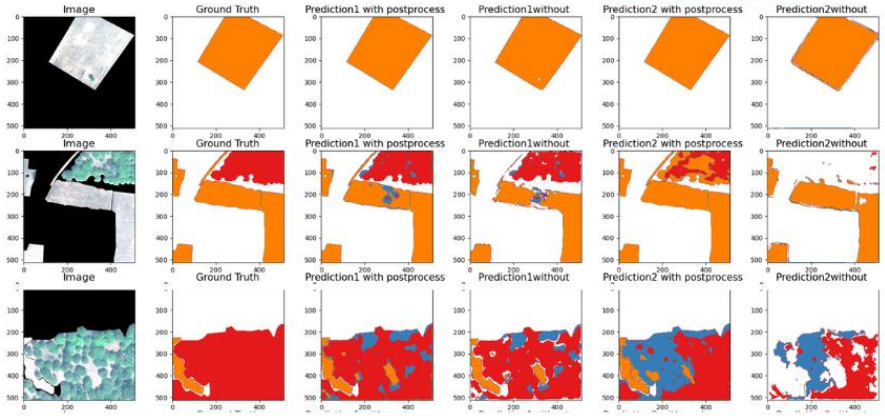
Après post traîtement





Pr.Gl=0.84

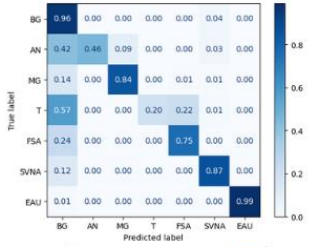


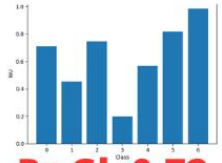


Modèle Unet avec la focal loss , avec les paramètres 2 gamma 0,25 alpha, préentraîné sur imagenet

Model 2

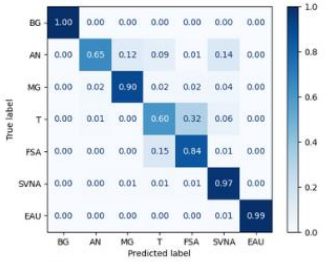
Avant post traîtement

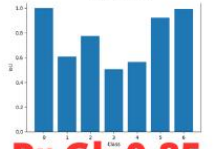




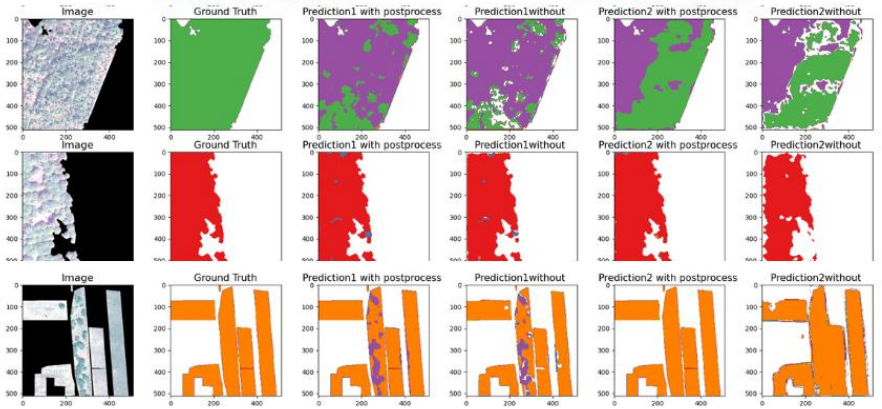
Pr.Gl=0.72

Après post traîtement

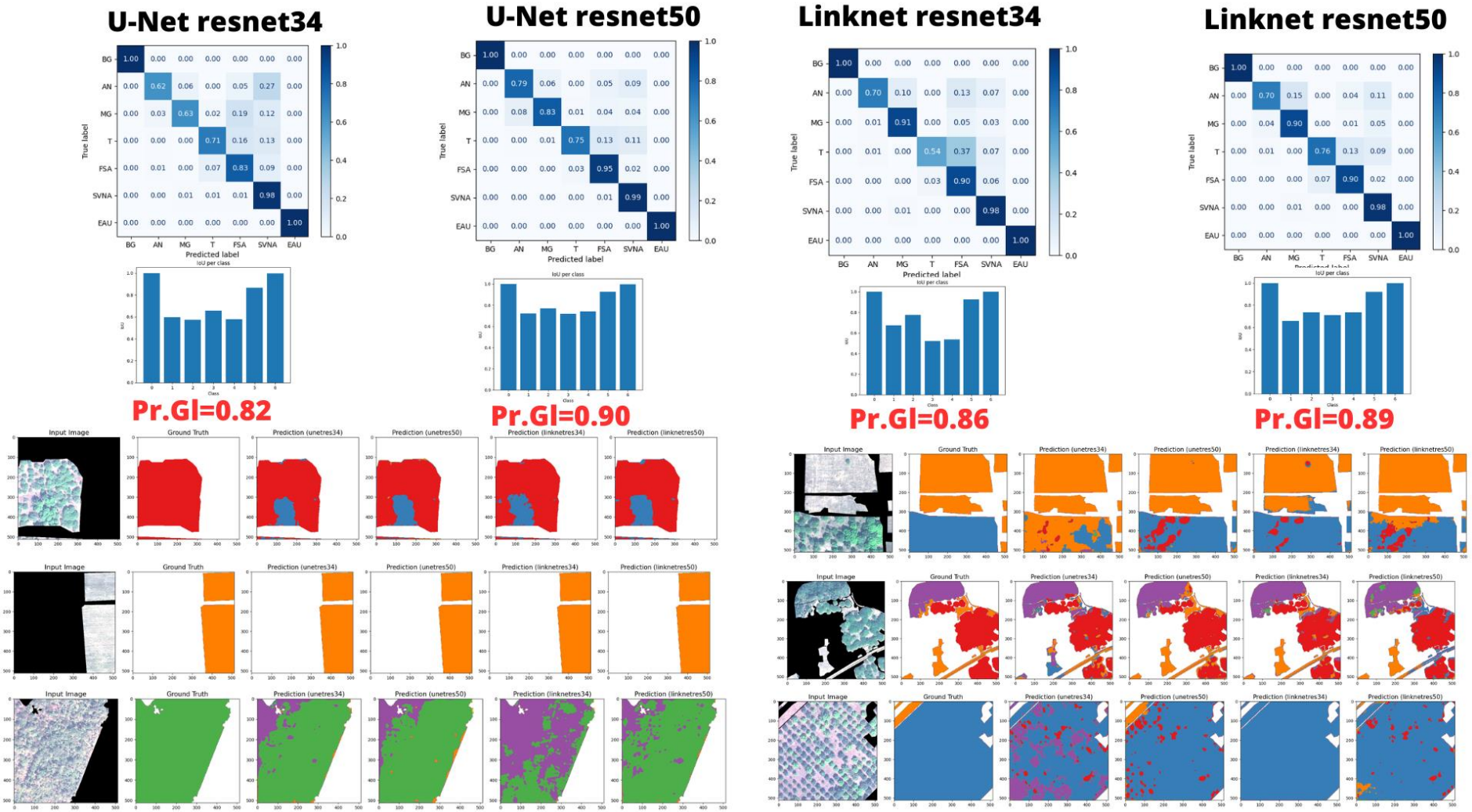




Pr.Gl=0.85

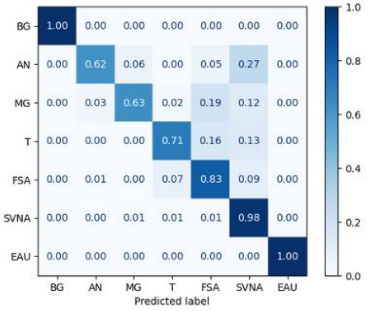


2.3 Résultats d'U-net et LinkNet



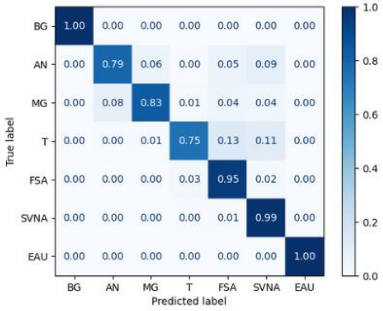
2.3 Résultats d'U-net et PSPNet

U-Net resnet34



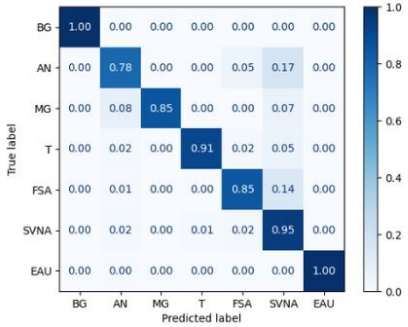
Pr.Gl=0.82

U-Net resnet50



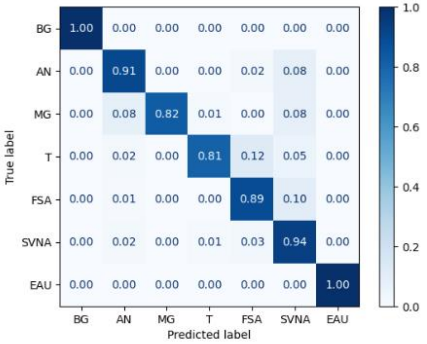
Pr.Gl=0.90

Pspnet resnet34



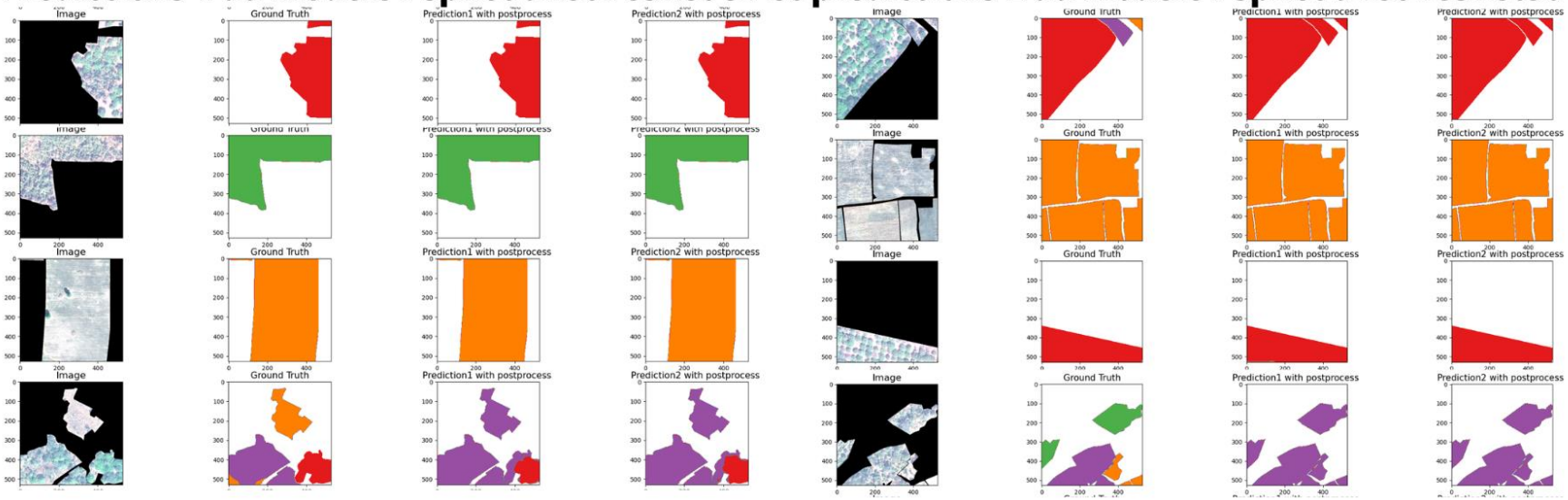
Pr.Gl=0.90

Pspnet resnet50



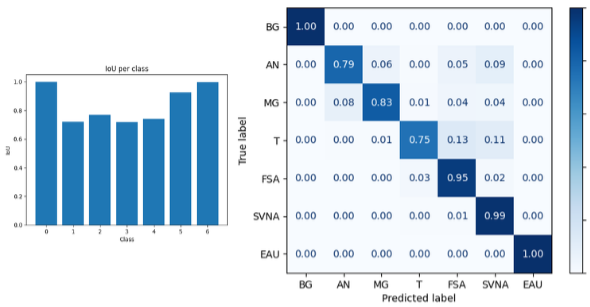
Pr.Gl=0.91

Prédictions 1 du modèle Pspnet avec resnet 34 et prédictions 2 du modèle Pspnet avec resnet50

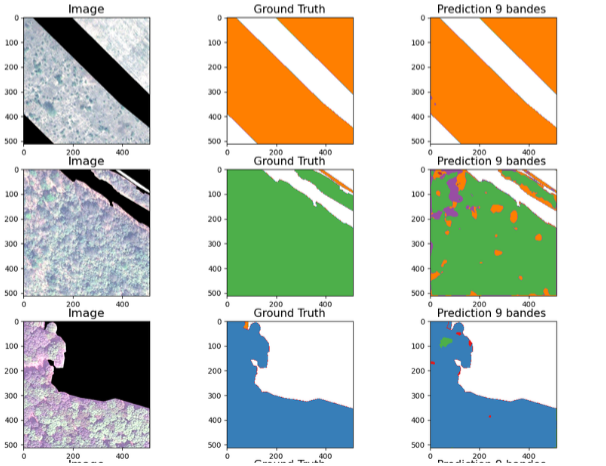


2.5 Résultats d'U-net avec différentes entrées de bandes

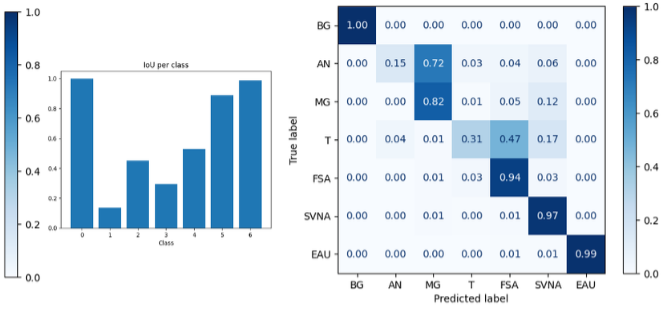
U-Net resnet50 avec 9 bandes



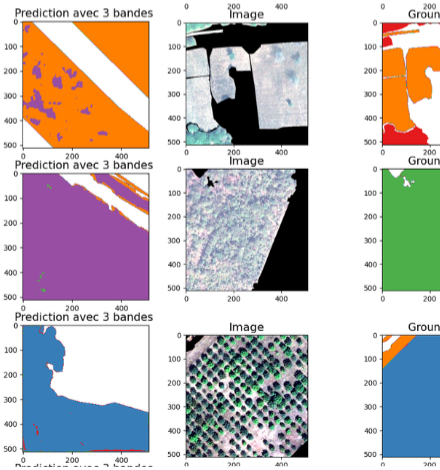
Pr.GI=0.90



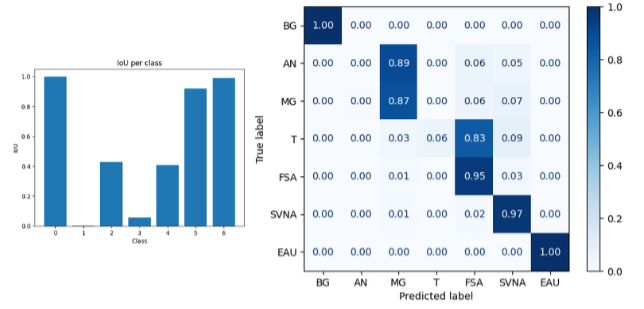
U-Net resnet50 avec 6 bandes



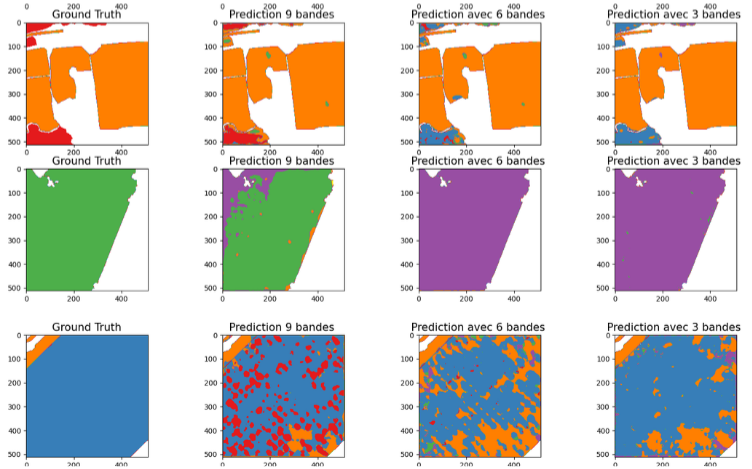
Pr.GI=0.74



U-Net resnet50 avec 3 bandes



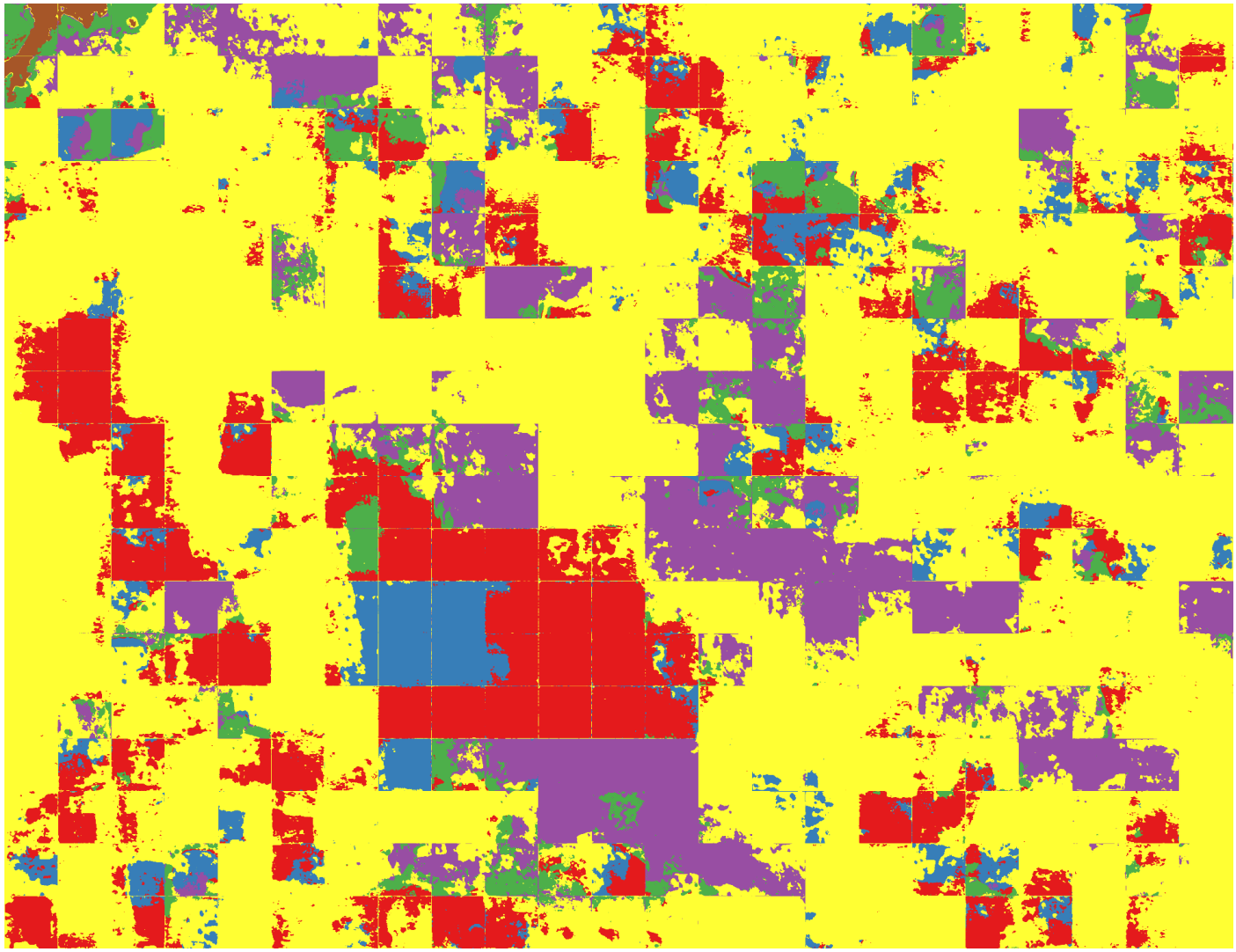
Pr.GI=0.69



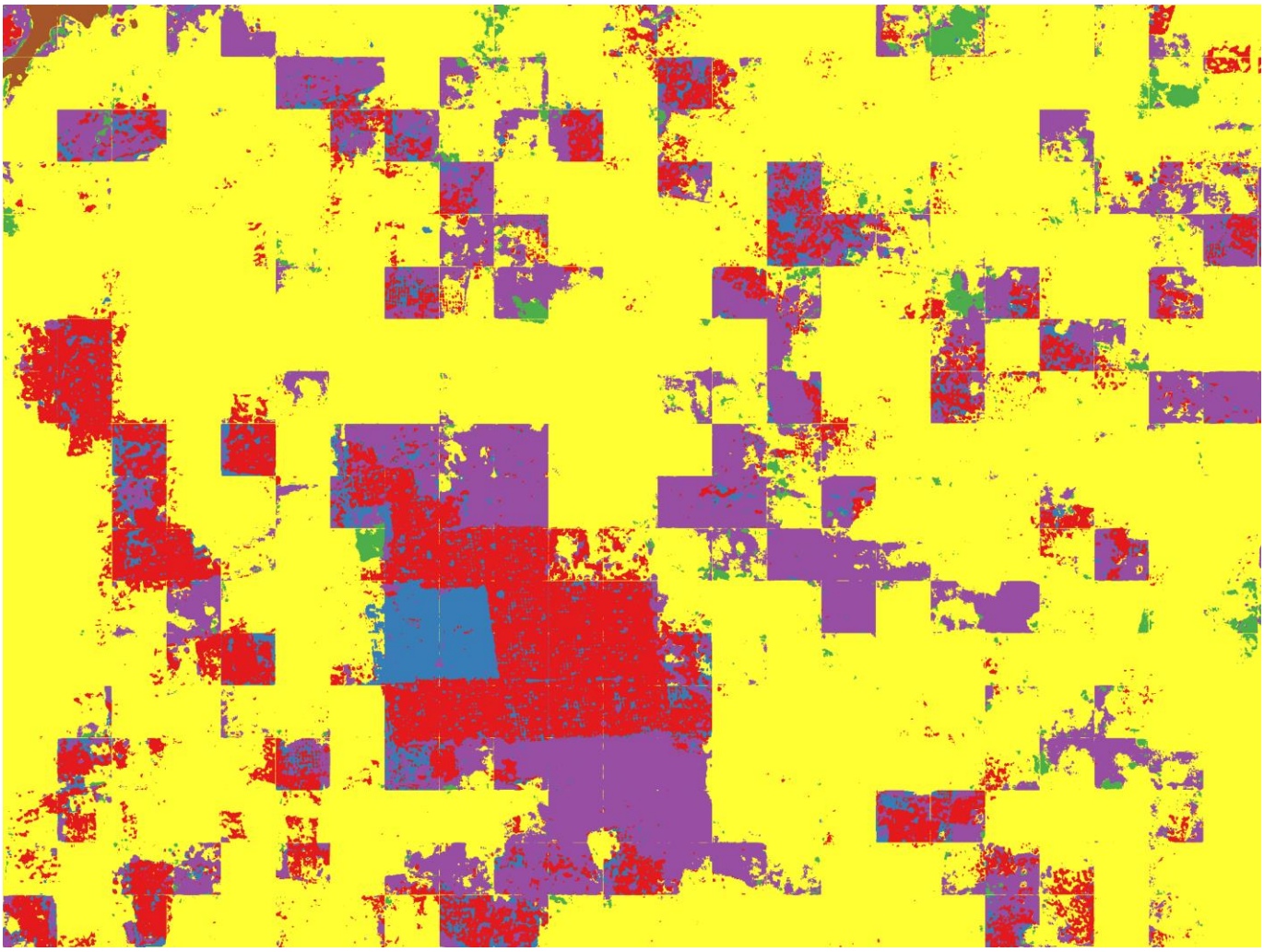
2.5 Application du modèle sur une large zone

Couleur	Classe
Rouge	Anacardiers
Bleu	Manguiers
Vert	Tecks
Mauve	Forêts et savanes arborées
Jaune	Surfaces végétales non arborées
Marron	Eau

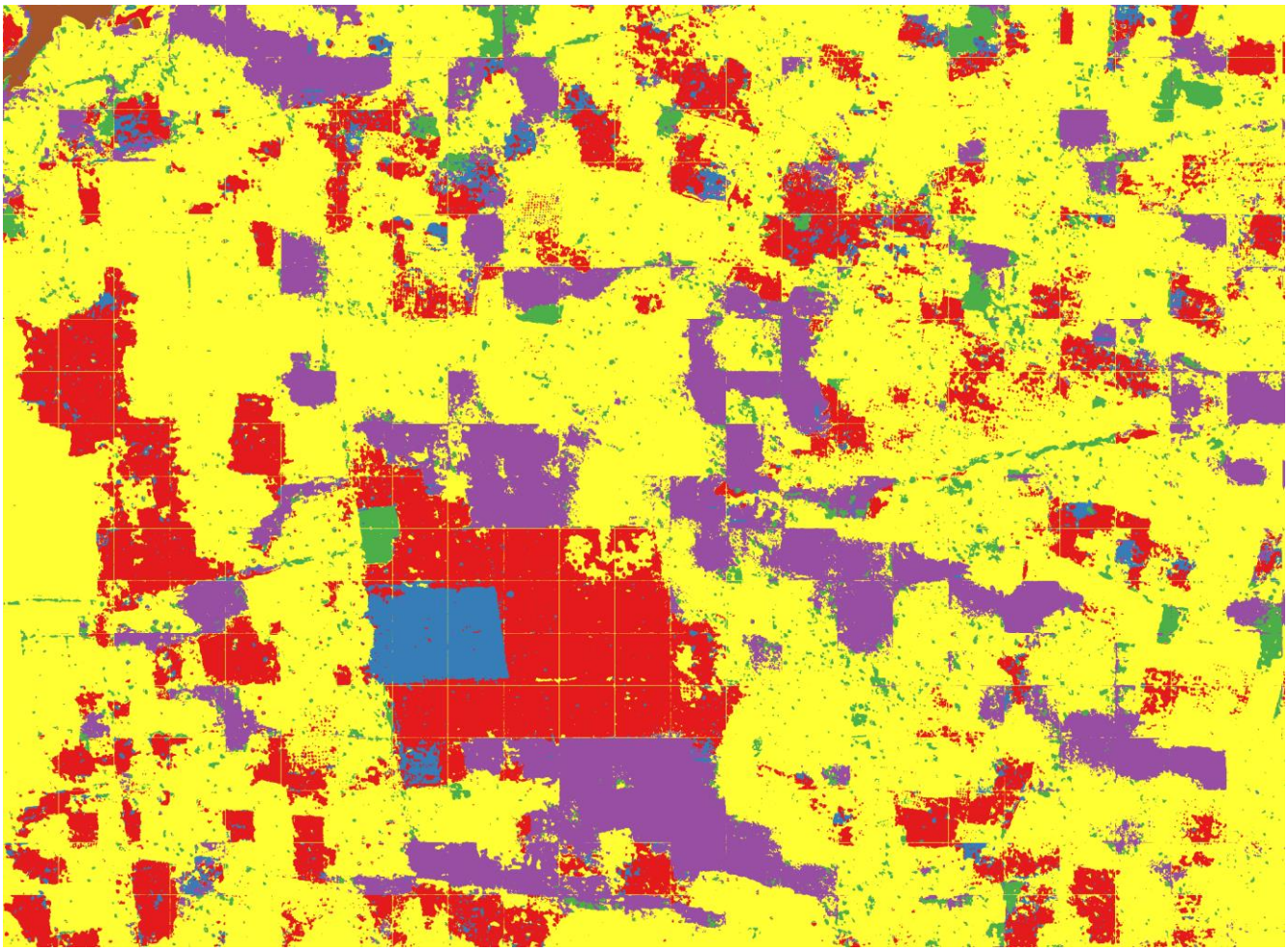
Prédiction avec U-Net et ResNet 34 avec préentraînement sur ImageNet :



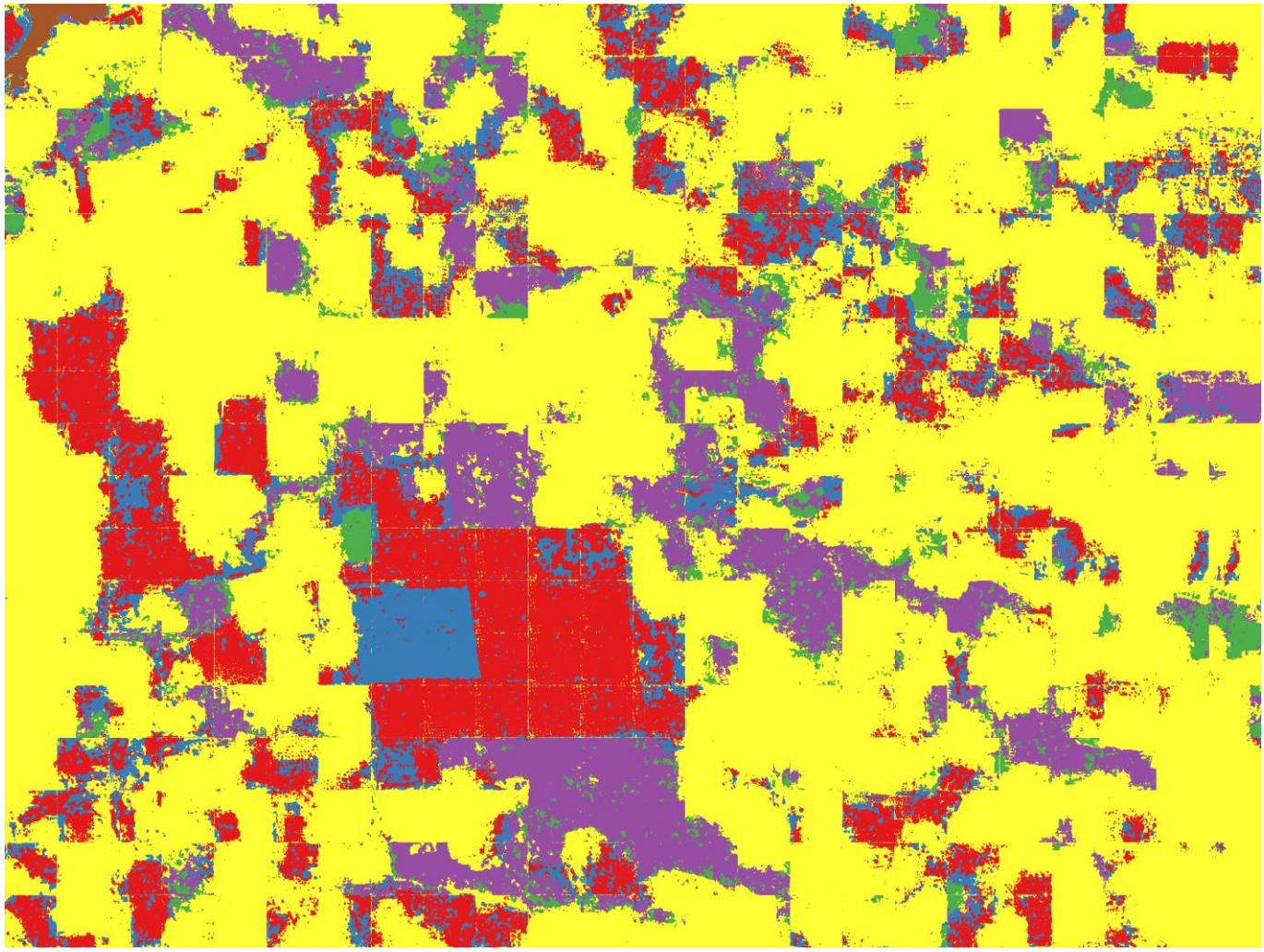
Prédiction avec U-Net et ResNet 34 sans préentraînement :



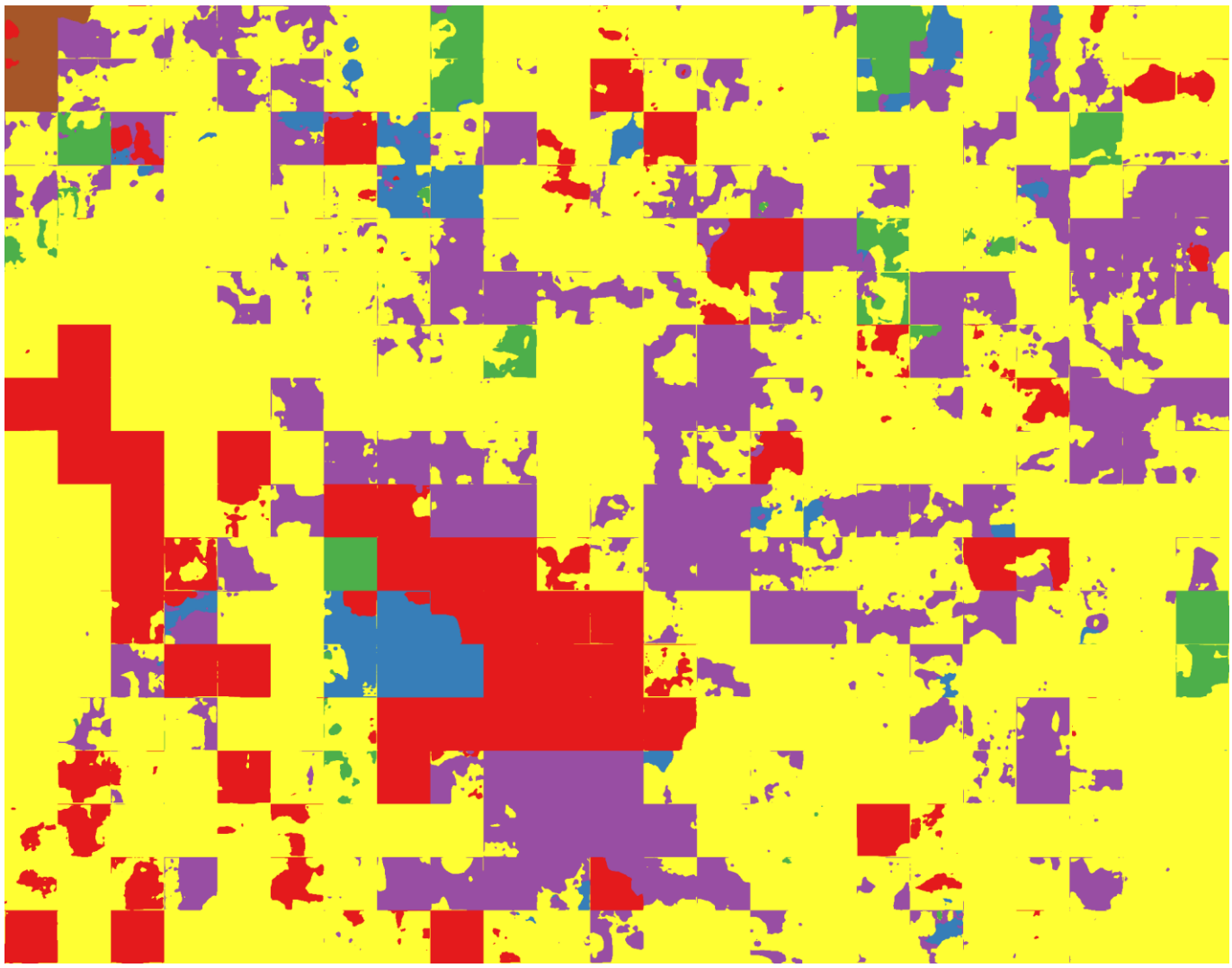
Prédiction avec U-Net et ResNet 50 :



Prédiction avec LinkNet et ResNet 50 :



Prédiction avec PSPNet et ResNet 50 :



Références

1https://ts2.space/fr/lia-et-levolution-de-la-technologie-dimagerie-par-satellite/
2https://github.com/satellite-image-deep-learning/techniques#2-segmentation
3.https://arxiv.org/abs/1505.04597
4.https://arxiv.org/pdf/1512.03385.pdf
5.https://neurohive.io/en/popular-networks/resnet/
6.https://arxiv.org/pdf/1505.04597.pdf // https://fr.wikipedia.org/wiki/ImageNet
7.https://github.com/ayushdabra/dubai-satellite-imagery-segmentation // https://github.com/YudeWang/U-Net-Satellite-Image-Segmentation
8.https://earth.esa.int/eogateway/missions/pleiades-neo
9.http://haralick.org/journals/TexturalFeatures.pdf
10.https://segmentation-models.readthedocs.io/en/latest/
11.https://www.academia.edu/71983889/Assessing_the_Accuracy_of_Remotely_Sensed_Data_Principles_and_Practices
12.**https://datascientest.com/matrice-de-confusion**
13.https://paperswithcode.com/method/focal-loss