Analyse d'images et apprentissage machine pour la détection des altérations des pierres des monuments historiques

Koubouratou Idjaton

UNIVERSITÉ D'ORLÉANS

ÉCOLE DOCTORALE MATHÉMATIQUES, INFORMATIQUE, PHYSIQUE THÉORIQUE ET INGÉNIERIE DES SYSTÈMES

LABORATOIRE : PRISME

THÈSE présentée par :

Koubouratou Olowountogni IDJATON

soutenance prévue le : **26 Septembre 2022**

pour obtenir le grade de : **Docteur de l'Université d'Orléans**

Discipline/Spécialité : **Sciences et Techniques de l'Ingénieur - Mathématiques et Informatique (STI-MI)**

Analyse d'images et apprentissage machine pour la détection des altérations des pierres des monuments historiques.

Thèse dirigée par :

Sylvie TREUILLET	MCF HDR, Université d'Orléans, Laboratoire PRISME
Xavier DESQUESNES	MCF, Université d'Orléans, Laboratoire PRISME
Xavier BRUNETAUD	MCF HDR, Université d'Orléans, LaMé

RAPPORTEURS :

Catherine ACHARD	PRU, Université Paris-Sorbonne, ISIR CNRS, Présidente du jury
Abderrahim ELMOATAZ	PRU, Université de Caen, GREYC UMR 6072 CNRS

JURY :

Laure TOUGNE	PRU, Université de Lyon 2, LIRIS UMR 5205 CNRS
Frédéric BOSHÉ	Senior Lecturer, University of Edinburgh, CyberBuild Lab
Rachid HARBA	PRU, Université d'Orléans, Laboratoire PRISME
Sylvie TREUILLET	MCF HDR, Université d'Orléans, Laboratoire PRISME
Xavier DESQUESNES	MCF, Université d'Orléans, Laboratoire PRISME
Xavier BRUNETAUD	MCF HDR, Université d'Orléans, LaMé

Remerciements

C'est avec un profond sentiment de gratitude que j'exprime ici mes remerciements à toutes les personnes qui ont contribué de près ou de loin à la réalisation de ce modeste travail.

J'adresse toute ma gratitude à ma directrice de thèse Sylvie TREUILLET pour m'avoir accordé l'opportunité d'apprendre à ses côtés, sa disponibilité, sa patience et sa confiance. Merci aussi à mes encadrants Xavier DESQUESNES et Xavier BRUNETAUD pour leurs précieux conseils qui ont contribué à enrichir mes réflexions tout au long de ce travail.

Je remercie aussi les membres du jury pour avoir accepté d'évaluer mon travail. Merci pour les échanges, commentaires et conseils riches et éclairés.

Mes remerciements à tous mes collègues du laboratoire PRISME. Tout particulièrement Fabrice ATREVI, Dian BAH et Asma BOUGRINE. Vous avez rendu cette expérience exceptionnelle. Merci aux autres collègues que j'ai pu côtoyer : Myriam, Asma, Narech, Evelyn, Rania, Khaouter, Doha, Amine, Marouane, Antonio, Gisèle. Les pauses Thés et Café, les discussions enrichissantes et conviviales me manqueront certainement.

Je souhaiterais aussi remercier mes proches, amis, frères et soeurs de la vie qui sont à Orléans, Paris, Cotonou, Porto-Novo, Chicago, Montréal et plus. Merci Fabrice ATREVI, Solange OLOUDE, Gisèle AÏSSOUN, Mariam TIDIGA et Samuel BATAILLE. Votre amitié et votre soutien permanent m'ont procuré réconfort et courage dans les moments les plus éprouvants de ces dernières années. Je vous suis très reconnaissante. Merci aux Power Girls : Gisèle, Léa, Ornella, Sedami, Geneviève et Bénédicta. Nos conversations profondes et nos escapades m'ont permis de rester connecter à la réalité.

Une pensée particulière à ma famille, mes parents. Merci à mon frère Tohid, ma maman et mon papa. "Vous êtes au creux de moi. Dans ma vie, dans tout ce que je fais." CD. Vous remercier ne sera jamais assez.

Je ne saurais finir sans remercier aussi tous les enseignants qui ont contribué à me former pendant tout mon cursus académique. Vos enseignements m'ont conduit d'une façon ou d'une autre à cet accomplissement.

Merci infiniment.

Résumé

La préservation des monuments historiques exige une surveillance de leur état pour assurer la sécurité des visiteurs et planifier au mieux les opérations de restauration. Cette surveillance repose essentiellement sur une observation visuelle réalisée par des experts sur site et un lourd et fastidieux travail d'inventaire et d'annotations manuelles sur des photos. Cette thèse propose de nouvelles méthodes d'analyse automatiques basées sur l'apprentissage machine pour faire de l'aide au diagnostic assisté par ordinateur à partir les images acquises pour la modélisation 3D complète de deux châteaux emblématiques du Val de Loire construits en tuffeau. Les travaux abordent deux problématiques : la segmentation pierre-à-pierre et la détection des altérations des pierres.

Pour la segmentation pierre-à-pierre, une base d'images labellisées a été créée grâce à une méthode ad-hoc utilisant des méthodes traditionnelles de seuillage et détection de contours dans les images couleurs. Cette base a ensuite servi à l'entraînement des meilleurs réseaux neuronaux de segmentation de l'état de l'art (FCN, SegNet, U-Net et DeepLab). Une étude comparative permet de valider le réseau DeepLab comme le plus performant et robuste face aux images complexes (avec une similarité de 98% avec la vérité terrain mesurée par l'indice de Jaccard).

Pour la détection des altérations des pierres, une seconde base d'apprentissage a été créée en s'appuyant sur une approche originale d'augmentation des données par la reprojection automatique des annotations des experts réalisées sur des orthophotos vers les images couleurs de haute résolution d'origine. La nouvelle architecture proposée, qui insère des modules de transformers dans le réseau YOLO, permet d'obtenir des performances qui surpassent l'état de l'art pour la détection d'altérations des pierres (avec un taux de recouvrement avec la vérité de 80%).

Enfin, une plateforme Web opérationnelle permet aux experts d'appliquer les algorithmes proposés et visualiser les résultats de segmentation automatique via

une interface conviviale, pour faciliter leur diagnostic sur les façades des châteaux. Ces contributions ouvrent la voie au développement d'un outil convivial multi acteurs pour un diagnostic plus précis des grands monuments.

Abstract

The preservation of historic monuments requires monitoring their condition to ensure the safety of visitors and to plan restoration operations as well as possible. This monitoring relies mainly on visual observation by experts on site and a heavy and tedious work of inventory and manual annotations on photos. This thesis proposes new automatic analysis methods based on machine learning for computer-assisted diagnosis from images acquired for the complete 3D modeling of two emblematic castles of the Loire Valley built in tufa. The work addresses two issues : stone-by-stone segmentation and stone alterations detection.

For stone-by-stone segmentation, a database of labelled images was created thanks to an ad-hoc method using traditional thresholding and edge detection methods in color images. This database was then used to train the best state-of-the-art segmentation neural networks (FCN, SegNet, U-Net and DeepLab). A comparative study validates DeepLab as the most efficient and robust network for complex images (with a similarity of 98% with the ground truth measured by the Jaccard index).

For the detection of stone alterations, a second learning base was created using an original approach of data augmentation by automatic reprojection of expert annotations made on orthophotos to the original high resolution color images. The new architecture proposed, which inserts transformer modules in the YOLO network, allows to obtain performances that surpass the state of the art for the detection of stone alterations (with a truth recovery rate of 80%).

Finally, an operational Web platform allows experts to apply the proposed algorithms and visualize the segmentation results via a user-friendly interface, to facilitate their diagnosis on castle facades. These contributions pave the way for the development of a user-friendly multi-actor tool for a more accurate diagnosis of large monuments.

Sommaire

Liste des tableaux

Liste des figures

Liste des sigles et Abréviations

CNN : Convolutional Neural Network ([FR] Réseaux de neurones convolutifs)

CPU : Central Processing Unit

CSP : Cross Stage Partial network.

CWT : Continuous Wavelet Transform ([FR] Transformée en ondelettes continue)

FPN : Feature Pyramid Network

GPU : Graphics Processing Unit ([FR] Processeur Graphique)

ICOMOS-ISCS : Glossaire illustré sur les formes d'altération de la pierre

mAP : mean average precision ([FR] précision moyenne globale)

MNT : Modèle Numérique de Terrain

NMS : non-maximum suppression

R-CNN : Region-Based Convolutional Neural Network ([FR] Réseau neuronal convolutif basé sur les régions)

RAM : Random Access Memory (ou mémoire vive)

RoI : region of interest ([FR] Région d'Intérêt)

SGD : Stochastic Gradient Descent ([FR] descente de gradient stochastique)

SfM : Structure-from-Motion

Sliding windows : Fenêtres glissantes

TLS : Terrestrial Laser Scanner ([FR] Scanner Laser Terrestre)

YOLO : You Only Look Once

LiDAR : Light detection and ranging ([FR] détection et télémétrie par la lumière)

VMHB : VISION FOR MONITORING HISTORICAL BUILDINGS

Introduction

Les châteaux de la vallée de la Loire ont été construits, pour la plupart, entre le Moyen-Âge et la Renaissance, et sont des symboles architecturaux qui représentent une part importante du patrimoine culturel en France et en Europe. Les châteaux de style Renaissance marquent la transition entre l'architecture militaire et l'architecture d'apparat, comme en témoignent les grandes fenêtres, les nombreuses cheminées et l'amélioration de leur confort sanitaire. La pierre reste le matériau privilégié pour la construction de ces édifices prestigieux.

Les châteaux de la Loire sont bâtis en pierres de taille de tuffeau, une craie tendre qui permet la réalisation des fines ornementations qui ont fait la réputation de ces châteaux. Le tuffeau est aussi malheureusement une pierre tendre et poreuse, très sensible aux variations d'humidité, qui entraînent une dégradation dans le temps.

Ainsi, au fil des années, du fait des intempéries climatiques et parfois de l'action humaine, le suivi et l'entretien de ces bâtiments du patrimoine culturel nécessitent de plus en plus de ressources. Ils restent essentiels pour préserver la sécurité des visiteurs et sauvegarder ces héritages du passé qui représentent un facteur identitaire de civilisation.

Pour planifier au mieux les opérations de restauration, les experts établissent un carnet de santé du monument. Parmi les différents points de contrôle, l'état des dommages des murs en pierres font l'objet de surveillances régulières. Elle se fait généralement à partir d'observations visuelles sur site, d'analyses qualitatives et d'un relevé manuel des zones d'altérations sur des orthophotos. Il s'agit d'une tâche assez minutieuse, très longue à l'échelle d'un grand château et pas toujours exhaustive en raison de la taille des façades et des parties du bâtiment qui restent

inobservables par un expert depuis le sol.

Grâce aux nouvelles techniques d'acquisition d'images et de numérisation (drones, scanner laser et la photogrammétrie), il est aujourd'hui possible de réaliser un scan complet des châteaux sous forme d'images ou de nuages de points qui permettent d'obtenir la reconstruction photogrammétrique complète et précise du château. Les orthophotos des façades sont d'ailleurs réalisées à partir de ces images rectifiées. Mais, la quantité de données est telle qu'il est nécessaire de développer de nouveaux outils pour aider à mieux les exploiter afin de faciliter le travail des experts.

L'objectif de cette thèse est de développer de nouvelles méthodes automatiques d'analyse pour apporter une aide au diagnostic assisté par ordinateur à partir des images acquises pour la modélisation 3D complète de deux châteaux emblématiques du Val de Loire, le château de Chambord et le château de Chaumont-sur-Loire.

Nos travaux permettent de tirer profit des images couleur et orthomosaïques existantes issues de la modélisation 3D complète de ces châteaux pour faire une analyse automatique du pierre-à-pierre et des altérations des pierres.

La suite de ce manuscrit de thèse s'organise autour de quatre (04) chapitres. Le premier décrit les données exploitées dans cette thèse. Elles sont issues d'acquisitions 3D réalisées dans le cadre des projets VALMOD et DIANE.

Le deuxième s'attaque au problème de la segmentation pierre-à-pierre, en dressant un état de l'art des méthodes existantes. Des méthodes classiques de seuillage et détection de contours dans les images couleur sont explorées et appliquées pour la proposition d'une méthode ad-hoc. Cette dernière a permis de créer une base d'apprentissage avec vérité-terrain pour l'étude des réseaux d'apprentissage. Une étude comparative permet d'analyser les performances et la robustesse des méthodes classiques et des réseaux d'apprentissage étudiés.

Le troisième chapitre est consacré à la détection des altérations des pierres. Il présente une seconde base d'apprentissage créée par une approche nouvelle d'augmentation des données par la reprojection automatique des annotations des experts réalisées sur des orthophotos vers les images couleur de haute résolution d'origine.

Une nouvelle architecture de réseau qui insère des modules de transformateurs dans le réseau YOLO, est proposée et permet d'obtenir des performances qui améliorent l'état de l'art.

Le dernier chapitre présente une plateforme Web développée pour permettre aux experts d'appliquer les méthodes étudiées à l'échelle d'un château et visualiser leurs résultats via une interface conviviale, pour faciliter le diagnostic sur les façades. Les détails conceptuels, choix techniques de conception de cette plateforme sont décrits.

Enfin, la conclusion fait la synthèse des travaux et présente les perspectives envisagées à ces travaux.

Chapitre 1

Acquisition des données et Bases de données

> La science est une bénédiction pour qui la saisit et une malédiction pour qui la fuit, mais on ne doit jamais être orgueilleux de son propre savoir puisqu'il n'y a pas de limite dans la science et que personne ne peut arriver à la perfection.
>
> Ptahhotep

Sommaire

1.1 Introduction

Le développement récent des technologies numériques a encouragé les experts en patrimoine à produire une large variété de données (images à haute résolution, scanner laser terrestre, nuage de points 3D) pour documenter et étudier le patrimoine culturel. Les quantités de données devenant de plus en plus importantes, il naît alors la nécessité de développer de nouveaux outils en vue d'exploiter ces données numériques pour faciliter les travaux de surveillance, de maintenance et de restauration du patrimoine culturel pour les experts.

L'acquisition des données exploitées dans cette thèse, s'est déroulée au cours de plusieurs campagnes sur deux châteaux emblématiques du Centre Val de Loire : le château de Chambord et le château de Chaumont-sur-Loire. Le développement des outils a consisté à tirer profit des images existantes issues de la modélisation 3D complète des châteaux pour faire une analyse du pierre-à-pierre et des altérations.

Dans ce chapitre, nous décrivons l'ensemble des données exploitées ainsi que les campagnes et techniques d'acquisition de ces données. La section 1.2, présente les châteaux et les projets qui ont permis les acquisitions de données. Puis la section 1.3 vient expliquer les principales techniques d'acquisition, les procédés de conception et les spécificités des données produites. Enfin la section 1.4 constitue une analyse détaillée et une exploration des ensembles de données et vérités-terrain associées, pour l'étude et l'élaboration de méthodes de segmentation pierre-à-pierre et de détection des altérations.

1.2 Campagnes d'acquisition

1.2.1 Les châteaux

Les châteaux de Chambord et Chaumont-sur-Loire sont deux monuments patrimoniaux emblématiques de la Région Centre-Val de Loire et de l'histoire de France.

Le **château de Chambord** est situé à l'est de la ville de Blois, le long du Cosson. Il est un symbole du style architectural de la Renaissance. Le château est érigé sur le Domaine de Chambord [1] d'une superficie de 5440 hectares, le plus grand

1. https ://www.chambord.org/

parc clos d'Europe [1]. Le mur d'enceinte du domaine est long de 32 kilomètres et celui du château mesure environ 600 mètres de long et $8,5$ mètres de hauteur pour la partie basse et 20 mètres de hauteur pour la partie haute du mur.

La figure 1.1 montre une vue globale du château. Un bâtiment carré avec une tour à chaque angle et en son sein un Donjon centré sur la façade nord-ouest. Au coeur du Donjon se trouve un ingénieux escalier à doubles révolutions.

Les campagnes d'acquisition au château de Chambord, représentent approximativement 6000 photographies prises au sol en extérieur pour les façades, sans drone. Les drones ont été utilisés uniquement pour les toitures. Quelques intérieurs ont aussi été réalisés au scanner laser.

FIGURE 1.1 – Photographie du château de Chambord [1].

Le **château de Chaumont-sur-Loire** est situé au sud-est de la ville de Blois, le long de la Loire. Il illustre l'architecture défensive de l'époque gothique et l'architecture d'agrément de la Renaissance. Le château est érigé sur le domaine de

Chaumont d'une superficie de 32 hectares[2], d'où il domine la Vallée de la Loire.

La figure 1.2 présente une vue générale du château. Il comporte 15 pièces ouvertes au public, une petite chapelle de style gothique, une cour interne faite de trois façades donnant sur la Loire et un grand escalier en colimaçon dans le centre du château.

Durant les campagnes d'acquisition à Chaumont, les photographies des façades ont été essentiellement prises au sol, en plus de quelques photographies prises par drones pour certaines façades, la cour intérieur et les toitures. Au total, environ 4000 photographies du château de Chaumont-sur-Loire. Le scanner laser a été réalisé sur près de 45% des intérieurs avec environ 600 stations laser.

FIGURE 1.2 – Photographie du château de Chaumont-sur-Loire [2].

2. https ://domaine-chaumont.fr/

1.2.2 Les projets

Les projets VALMOD et DIANE ont permis la mise en oeuvre des campagnes d'acquisition, et s'inscrivent dans le programme Intelligence des Patrimoines, un programme Ambition Recherche Développement (ARD) soutenu par la Région Centre-Val de Loire et porté par le Centre d'études supérieures de la Renaissance de Tours [3].

Le **projet VALMOD** a permis de produire des contenus innovants sur l'architecture du château de Chambord, pour être utilisés à la fois dans des actions scientifiques liées à la conservation du patrimoine, mais aussi comme support de médiation pour du contenu touristique. Il n'avait pas vocation à décrire les altérations des pierres. Les équipes impliquées dans la réalisation de ces projets proviennent des unités de recherche du LaMé (Université d'Orléans), du CESR, du LI (Université de Tours) et du Domaine national de Chambord.

Les principales tâches du projet ont été : la modélisation 3D de l'état actuel du monument historique, l'inventaire historique et architectural ; l'Elaboration des scénarii de valorisation ; enfin les transfert et dissémination auprès des publics. Il a permis de réaliser plusieurs campagnes d'acquisition de relevé 3D et de modélisation du Château de Chambord.

Le **projet DIANE** se concentre sur la conservation et la valorisation scientifique et touristique du Domaine régional de Chaumont-sur-Loire. Il a permis de réaliser le relevé des altérations pour produire le carnet de santé du château de Chaumont-sur-Loire. Les acteurs investis dans ce projet sont les unités de recherche du LaMé, l'équipe Image-Vision du laboratoire PRISME, le service Patrimoine et Inventaire de la Région Centre-Val de Loire et le Domaine de Chaumont-sur-Loire.

Les grands travaux du projet ont été : la caractérisation de l'état du château pour enrichir son diagnostic sanitaire, la synthèse documentaire de son histoire, et la modélisation 3D du Chaumont-sur-Loire.

Ces projets, VALMOD et DIANE, ont permis d'élaborer des méthodologies pour l'utilisation complémentaire des techniques de scanner laser terrestre et de photogrammétrie.

3. https ://intelligencedespatrimoines.fr/

1.3 Techniques d'acquisition et Représentation des données

Les campagnes d'acquitions réalisées dans les projets de recherches VALMOD et DIANE ont permis de créer des données pour la génération des nuages de points 3D en utilisant des techniques de scanner laser et de photogrammétrie.

1.3.1 Scanner Laser Terrestre (TLS)

Le scanner laser est une technique d'acquisition qui acquiert les coordonnées (x, y, z) de nombreux points sur une cible en émettant des impulsions laser vers ces points et en mesurant la distance entre le dispositif et la cible [17]. Elle est robuste aux variations de conditions d'éclairage et aux surfaces à faible texture. C'est une méthode active qui génère sa propre source de lumière, ce qui lui permet de travailler même dans l'obscurité.

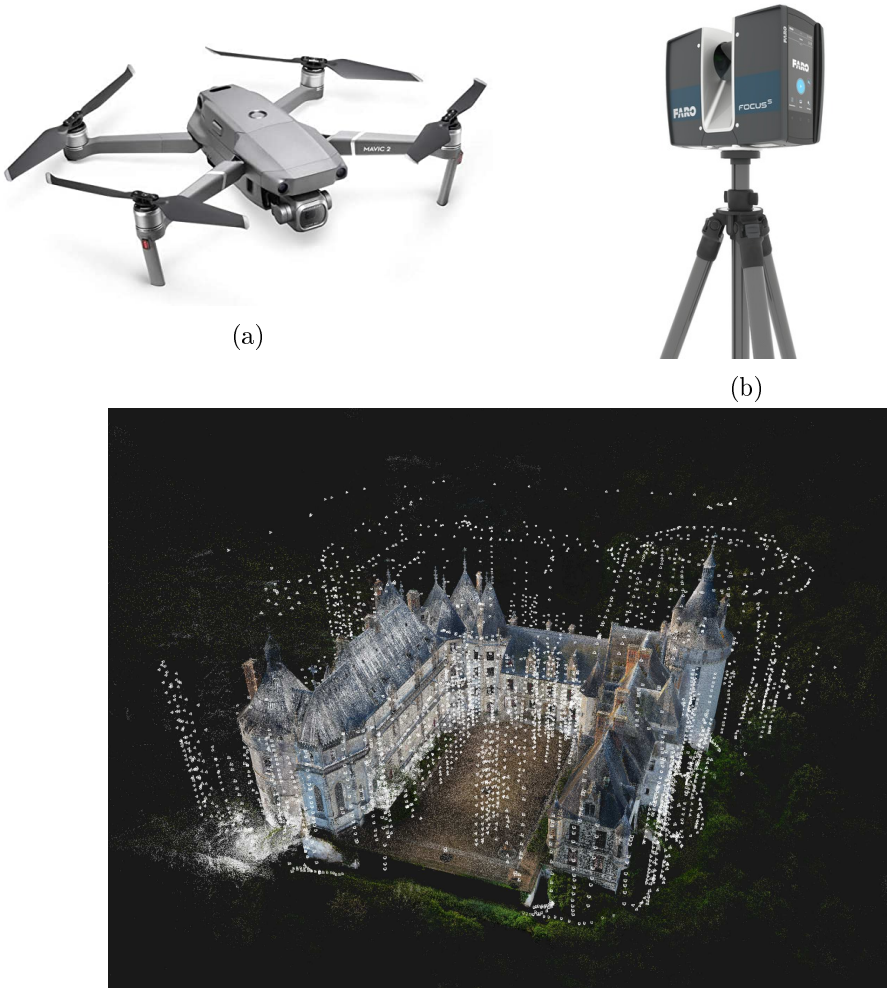
Il existe les scanners laser mobile ou MLS, les scanner laser terrestre ou TLS et la télédétection par laser (en anglais, LIght Detection And Ranging) ou LIDAR embarqués sur drones. Le TLS FARO Focus 120 1.3b a été employé lors des acquitions pour les châteaux de Chambord et de Chaumont-sur-Loire. Il a été installé sur un trépied et positionné à des emplacements dénommés stations de scans pour scanner les façades des monuments historiques. La prise en compte de coordonnées de points localisés dans le modèle 3D permet à terme de localiser les stations de scans.

La figure 1.3 présente une vue du scanner laser terrestre FARO Focus 120, le drone Dji Mavic Pro 2 et les emplacements des équipements lors d'une campagne d'acquisition pour l'enregistrement numérique de l'entièreté du château de Chaumont-sur-Loire avec caméra sur drone.

Une fois le scanner laser et son trépied positionné, il réalise les acquisions en opérant un scan ou balayage du faisceau laser, avec un pas angulaire constant, dans toutes les directions à l'exception de celle du trépied. Le scanner laser contient un miroir qui sert à orienter le faisceau laser vers un point dans l'espace, avec une orientation connue. Le retour du faisceau laser dans la même direction permet de calculer le temps de vol ou décalage temporel. Ainsi, pour chaque orientation, le scanner laser calcule une distance à l'objet. En ajoutant tous les points, on obtient

un jeu de données correspondant à un objet sphérique où chaque point possède un rayon différent. Le scanner laser permet de réaliser des nuages de points 3D.

Contrairement au processus photogrammétrique, qui peut échouer pour différentes raisons, le scanner laser sera toujours en capacité de produire un résultat. Il tire cet avantage de sa capacité à générer sa propre lumière et produire nativement la donnée 3D. La certitude d'obtenir des acquisitions contribue à rendre cette technique de plus en plus adoptée pour la documentation et l'analyse des monuments historiques par les experts en patrimoine [18].



(a) (b)

(c)

FIGURE 1.3 – (a) Le drone Dji Mavic Pro 2. (b) Le scanner laser terrestre FARO Focus 120. (c) Acquisition 3D de l'entièreté du Château de Chaumont-sur-Loire (France) avec toutes les prises de vue par caméra sur drone. Les positions de la caméra sont localisées par des polyèdres blancs. Cette image à été réalisée par les experts en patrimoine du Laboratoire LaMé.

1.3.2 Photogrammétrie

La Photogrammétrie est une technique utilisée dans le traitement des données du partimoine historique, qui constiste à recréer des modèles 3D à partir de photographies. Pour ce faire, on capture une grande quantité de photographies du monument historique suivant une large variété de points de vue. Dans ce processus d'acquition, il est fondamental de maintenir un important taux de recouvrement spatial des photographies et aussi de garder la focale constante durant l'acquisition des photographies.

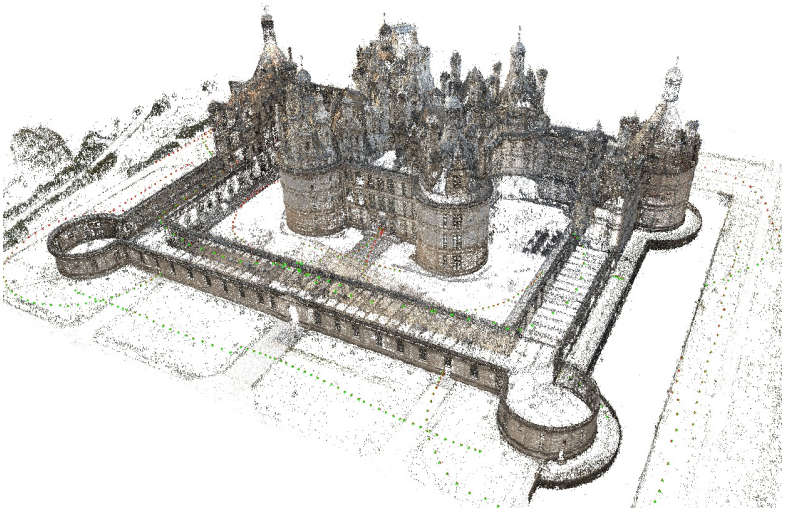
L'ouverture de l'apareil de prise de vue est aussi un paramètre important pour la réalisation de la photogrammétrie. Il permet de déterminer la profondeur de champs de prise de vue. Ce qui définit la zone où la meilleure netteté sera obtenue. Il permet aussi une bonne exposition de la photographie.

Les photos acquises sont traitées à l'aide de logiciels de calculs photogramétriques tels que Reality Capture [19], MicMac [20], qui se basent sur l'identification de pixels commun entre les photos de différents points de vue. Pour deux photographies par exemple, le processus de photogrammétrie consiste à détecter les pixels en commun entre elles. Les positions relatives de ces pixels sont analysées sur l'ensemble des photographies pour déduire leurs positions dans l'espace sous forme de coordonnées (x, y, z). Un modèle 3D est conçu à partir des positions dans l'espace de tous les pixels. Ce processus peut être généralisé à un grand nombre de photographies avec des points de vue différents permettant de fiabiliser les calculs de position de chaque pixel, en réduisant l'erreur tout en augmentant l'étendue du modèle 3D.

La photogrammétrie permet de réaliser des modèles 3D d'où on peut extraire une carte de profondeur à l'aide d'un choix de primitive. La carte de profondeur peut être utilisée pour projeter un modèle 3D sous la forme d'une orthomosaïque [21].

Pour réaliser les orthomosaïques des façades extérieures des châteaux, la méthode de prise de vue linéaire a été adoptée. Dans ce cas, les photographies sont prises en déplaçant le drone ou l'appareil photo sur son trépied, parallèlement à la façade photographiée. Une première série de photos éloignées et couvrant toute

(a)



(b)

FIGURE 1.4 – (a) Acquisition de photographie pour la photogrammétrie de la cour intérieure du château de Chaumont-sur-Loire (France). Les polyèdres blancs représentent la localisation des positions des caméras. Cette image à été réalisée par les experts en patrimoine du Laboratoire LaMé. (b) Numérisation du château de Chambord réalisée avec le logiciel MicMac par les étudiants du master PPMD de l'ENSG en tant que travail pédagogique de terrain durant les automnes 2014 et 2015.

la hauteur de la façade est prise. Elles servent de photos d'orientation. Enfin, une deuxième série de photos rapprochées est prise. Celles-ci permettent de définir plus précisément la texture. Un recouvrement de 60% a été respecté pour les photos d'orientation et 75% pour les photos de texture ou de détails. Ainsi, après chaque prise de vues, il est observé un déplacement de 40% ou de 25% environ par rapport

à la vue précédente.

La figure 1.4 présente les emplacements de l'appareil photo pour les prises de vues réalisées lors d'une campagne d'acquisition sur les façades de la cour intérieure du château de Chaumont-sur-Loire.

Les techniques TLS et photogrammétrie sont souvent combinées pour produire des modèles 3D complets et réaliser des orthomosaïques. Les géométries de base, murs et structures principales, sont déterminées par photogrammétrie, tandis que les détails fins et les formes plus complexes parfois moins accessibles sont produites à l'aide d'un scanner laser et parfois par drone [22].

Le scanner est souvent plus pertinent pour les intérieurs qui disposent d'un éclairage souvent inadapté et limitent le recouvrement, en raison de la subdivision des espaces, ce qui peut faire échouer un processus photogrammétrique. La photogrammétrie, pouvant être aéroportée (drone), sera par contre préférée pour les extérieurs difficiles d'accès tels que des parties en élévation, ou des toitures.

1.3.3 Nuage de points 3D

Le nuage de points 3D est une représentation prisées pour la numérisation des monuments historiques. Il permet de représenter dans un espace tridimensionnelle avec des millions de points de coordonnées (x, y, z). Dans cette représentation, les points ne sont pas liés entre eux, ils sont représentés individuellement pas leurs coordonnées.

Généralement, les nuages de points sont obtenus par des opérations d'acquitions effectuées avec un scanner laser. Il est aussi possible de réaliser des nuages de points par photogrammétrie.

Les nuages de points 3D servent aussi à réaliser la segmentation sémantique des éléments constitutifs d'un bâtiment historique [18]. Ils servent aussi à l'extraction de caractéristiques plus détaillées et plus précises pour la classification de tout ou partie d'un monument historique [23].

1.3.4 Orthomosaïque

Les orthomosaïques sont utilisées par les experts du patrimoine pour surveiller et documenter la santé des grands bâtiments historiques. Une orthomosaïque ou or-

thoimage peut servir de support cartographique plan : la géométrie a été redressée en exploitant un modèle numérique de terrain (représentation de la topographie). L'inclinaison de la prise de vue, les variations de relief et les déformations optiques des objectifs sont corrigés de sorte que l'orthoimage semble être prise à la verticale de tous les points qu'elle figure, ces points étant situés sur un terrain parfaitement plat. Pour couvrir de grandes surfaces comme les façades des monuments, les orthomosaïques sont créées à partir d'un grand nombre de photos couleur rectifiées, assemblées, et égalisées radiométriquement.

Les orthomosaïques sont obtenues à l'aide du logiciel photogrammétrique libre et gratuit MicMac ([24]), à partir de la modélisation 3D, et offrent une représentation cartographique plane de l'ensemble d'une façade ou d'une tour à une échelle uniforme. Elles sont en général moins résolues que les images couleur d'origine utilisées pour les créer. Cette représentation cartographique est très pratique pour surveiller les grands bâtiments historiques, pour annoter et localiser les altérations [4]. La Figure 1.5 présente l'orthomosaïque de la façade sud du château de Chambord à l'échelle uniforme de $5mm^2$ par pixel.

FIGURE 1.5 – Orthomosaïque de l'enceinte basse du château de Chambord

Les sections suivantes présentent une exploration des diverses données utilisées pour les travaux de segmentation pierre-à-pierre et les travaux de détection des altérations. Les images couleur prisent sur site et les orthomosaïques sont les deux types de données qui nous ont servi à l'étude et au développement des méthodes proposées dans la suite de cette thèse.

1.4 Vérités-terrain

Les vérités-terrain exploitées dans cette thèse, ont été réalisées sur les données disponibles au moment des travaux. Elles ont été produites dans le cadre de la conception du carnet de santé des châteaux, et ne sont pas spécifiquement acquises pour la segmentation pierre-à-pierre ou la détection des altérations.

1.4.1 Segmentation pierre-à-pierre

Les données de segmentation pierre-à-pierre sont composées de 6 orthomosaïques représentant des façades différentes du château de Chambord. Elles sont sous la forme d'images couleur représentées en trois canaux (RGB). La résolution de ces orthomosaïques est de $5mm^2$ par pixel.

La figure 1.6 présente le plan simplifié du château de Chambord. On peut y voir, le Donjon en forme carré avec ses quatres angles marqués par des tours : Tour Dieudonné, Tour François 1^{er}, Tour Caroline de Berry et Tour Henri V.

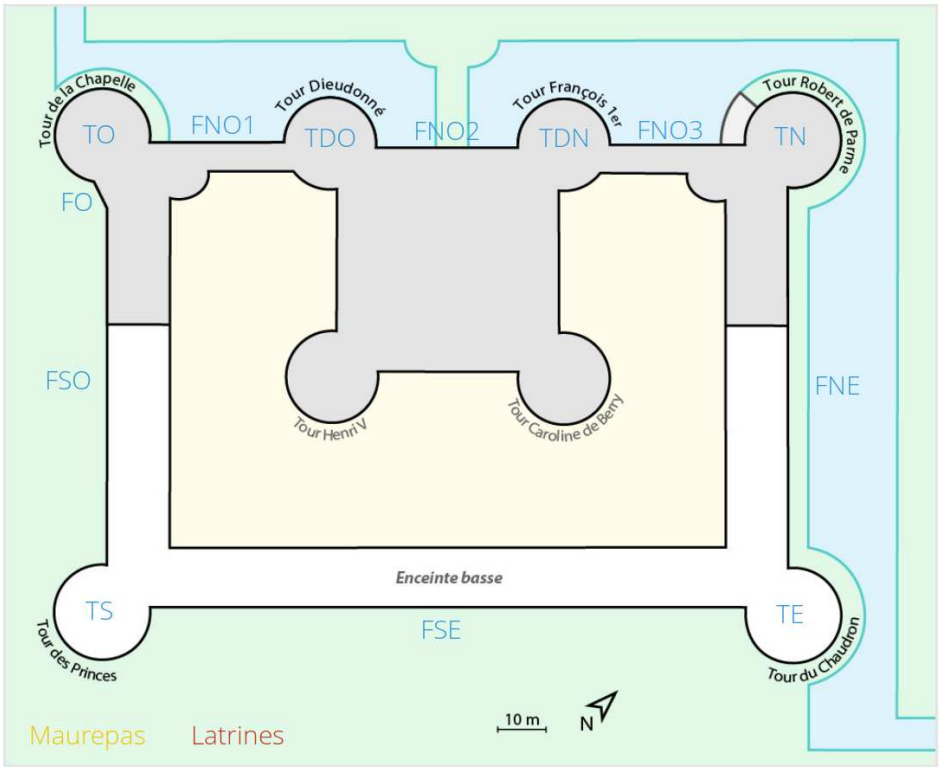


FIGURE 1.6 – Plan simplifié du château de Chambord [3].

Le tableau 1.1 présente pour chaque orthomosaïque les hauteurs et largeurs en pixel. Les codes des orthomosaïques sont référencés à leur position sur le plan. Les façades Nord Est, Sud Est et Sud Ouest sont les plus larges car elle s'étentendent sur toute la longueur du château. La façade nord ouest quant à elle est ponctuée par deux des quatres tours du Donjon. Elle a donc été réalisée en trois parties : FNO1, FNO2 et FNO3. Le nombre de photos prises sur site pour la réalisation de chaque orthomosaïque est aussi souligné dans le tableau.

Tableau 1.1 – Dimensions de l'ensemble des données de segmentation pierre-à-pierre.

Code	Orthomosaïque	Largeur (pixel)	Hauteur (pixel)	Photos
FSE	Façade Sud Est (enceinte basse)	41864	3828	109
FSO	Façade Sud Ouest	27605	11275	82
FNE	Façade Nord Est	24989	10513	104
FNO1	Façade Nord Ouest 1	15465	9217	44
FNO2	Façade Nord Ouest 2	10436	8818	48
FNO3	Façade Nord Ouest 3	7247	9075	42

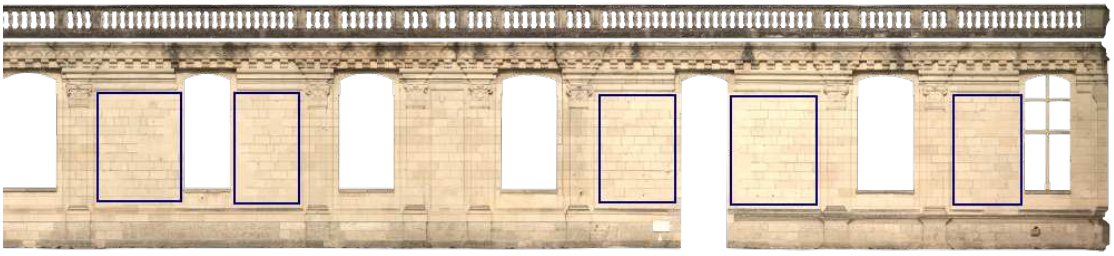


FIGURE 1.7 – Partie de l'orthomosaïque de l'enceinte basse du château de Chambord. Les cadres bleues représentent les parties de murs découpées.

Tableau 1.2 – Proportions joints/surface des pierres sur les 5 découpages de la figure 1.7.

	Aire (pixel)	Aire (%)
Joints	639885	2,40
Surface des pierres	26034015	97,60
Total	26673900	100

La figure 1.7 présente une partie de l'orthomosaïque de l'enceinte basse du château de Chambord. Les cadres bleus désignent les parties des murs considérées pour constituer la base de données pour entrainer les réseaux d'apprentissage. Les moulures et motifs n'ont pas été considérés car ils peuvent présenter un aspect localement similaire à celui d'un joint.

La figure 1.8 présente quelques exemples d'images de parties du mur : deuxième, troisième et cinquième cadre bleu sur la figure 1.7. Pour chaque image, la vérité terrain manuellement réalisée est associée. Le tableau 1.2 montre les proportions, représentées par l'aire en pixel et en pourcentage, des joints et surfaces des pierres

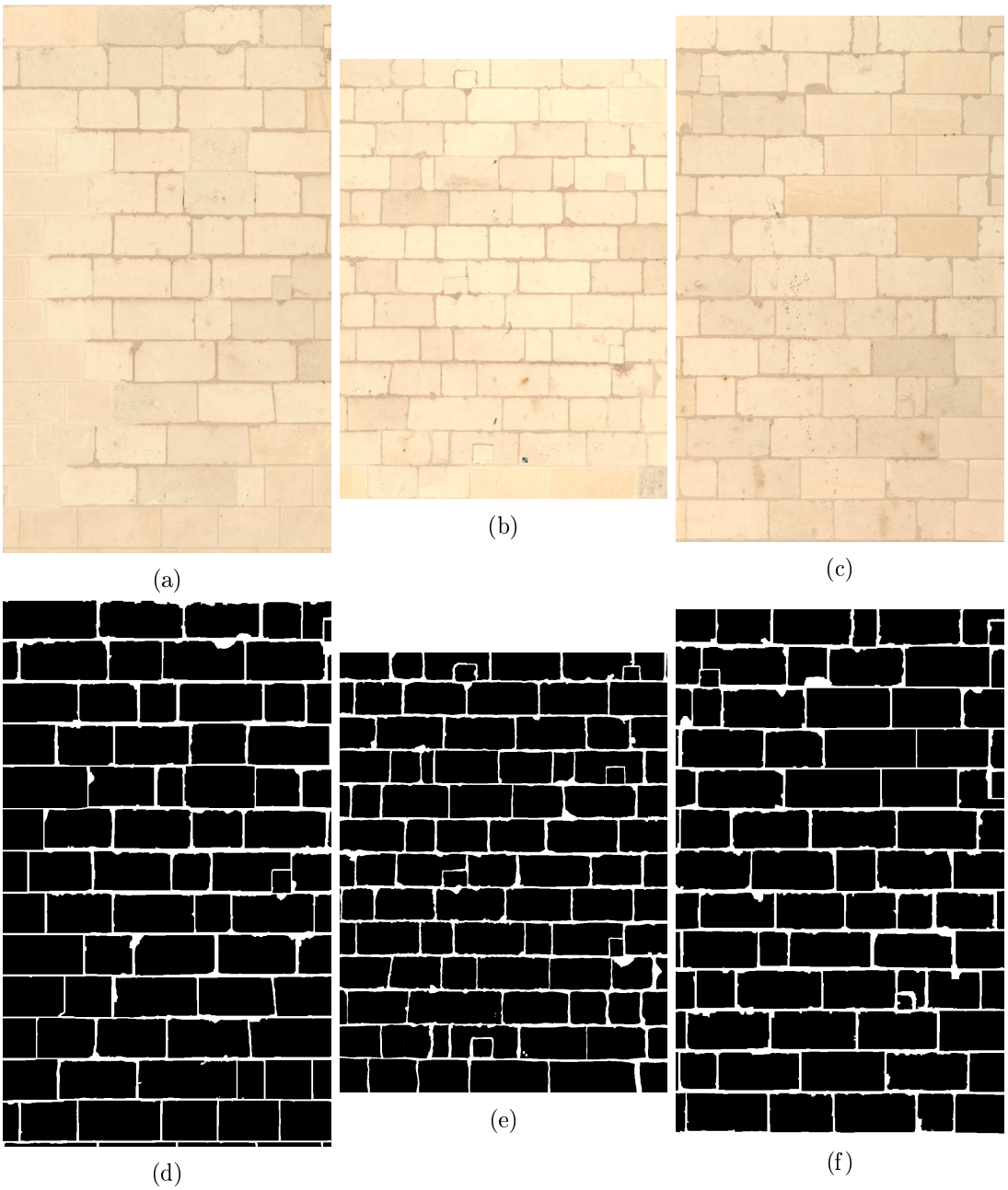


FIGURE 1.8 – Parties d'othophotos et leurs vérités-terrain : (a),(b) et (c) : deuxième, troisième et cinquième cadre bleue sur la figure 1.7. (d), (e) et (f) vérités-terrain réalisées manuellement.

sur les 5 découpages de la figure 1.7. Les joints représentent $2,40\%$ et la surface des pierres $97,60\%$ de la surface totale. Ce qui traduit un déséquilibre entre les classes joints et surface des pierres.

1.4.2 Détection des altérations

Les données de détection des altérations sont composées de 3 orthomosaïques représentant les différentes façades de la cour intérieure du château de Chaumont-sur-Loire. Il s'agit d'images en couleur représentées en trois canaux (RGB). La résolution de ces orthomosaïques est de $2,5mm^2$ par pixel. Le tableau 1.3 présente pour chaque orthomosaïque les largeurs et hauteurs en pixel.

Tableau 1.3 – Dimensions de l'ensemble des données de détection des altérations.

Orthomosaïque	Largeur (pixel)	Hauteur (pixel)
Façade cour Est	17025	11955
Façade cour Sud	11735	10624
Façade cour Ouest	17506	9587

Pour réaliser la vérité-terrain, les experts procèdent à l'annotation des altérations sur les orthomosaïques à partir d'observations et évaluations visuelles minutieuses opérées sur site. Les annotations sont reportées numériquement sur les orthomosaïques avec le logiciel QGIS [25].

Il convient de préciser que les orthomosaïques ne sont qu'un support graphique pour la cartographie des altérations, leur résolution est insuffisante pour que les experts puissent faire une détection de qualité directement sur l'orthomosaïque. Ainsi, pour documenter certaines altérations peu visibles à l'échelle d'une orthomosaïque, il arrive que les experts prennent des photos complémentaires réalisées à courtes portées, centrées sur les altérations. Cependant, même dans ce cas, la détection se fait à l'oeil sur site, ces photos complémentaires ne servent qu'à documenter la délimitation des contours. Ces photos zoomées sur les altérations n'ont pas été étudiées comme vérité terrain car elles ne sont pas représentatives des photos massivement acquises systématiquement durant une campagne de numérisation 3D par photogrammétrie.

La figure 1.10 montre le résultat des annotations manuelles d'altérations réalisées par les experts sur l'orthomosaïque de la façade de la cour Est du château de Chaumont.

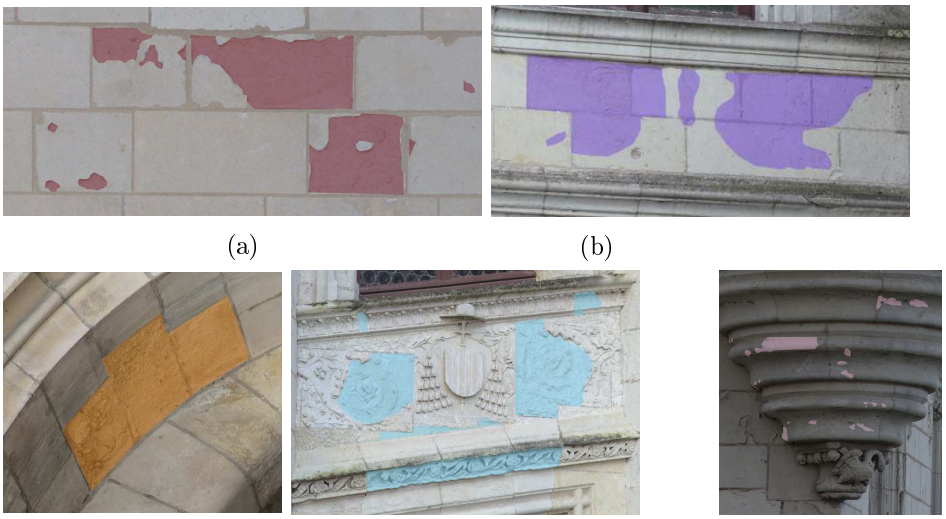


FIGURE 1.9 – Cinq types d'altérations : (a) desquamation en plaque. (b) desagrégation sableuse. (c) desquamation en feuillet. (d) erosion. (e) pelage. Cette image a été réalisée par les experts en patrimoine du Laboratoire LaMé.

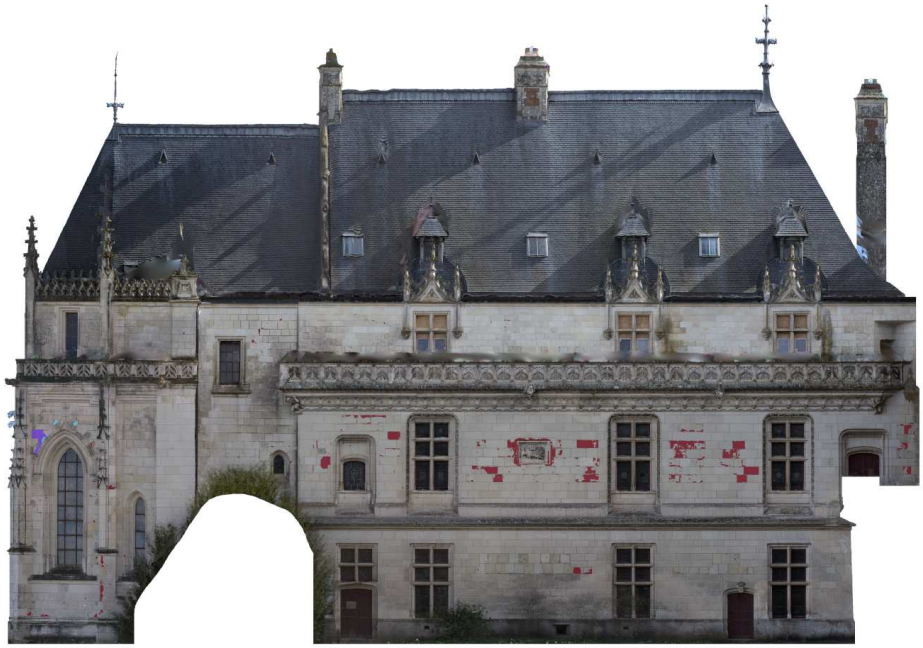


FIGURE 1.10 – Orthomosaïque de la façade cour Est avec les annotations manuelles d'altérations réalisées par les experts.

Cinq types d'altérations ont été identifiés sur l'ensemble des données. Il s'agit de la desquamation en plaque, la desquamation en feuillet, l'érosion, la désagrégation sableuse et le pelage (Voir figure 1.9).

L'apprécation des caractérisques et différences entre altérations est subjectif et dépendant de la maitrise de chaque expert. Le glossaire ICOMOS [26] sert

de référence terminologique pour une identification plus précise et commune des différents types d'altérations.

Tableau 1.4 – Proportions de surface saine et de surface d' altérations.

	Aire (pixel)	Aire (%)
Altérations	96868944	22,03
Surface saines	342828656	77,97
Total	439697600	100

Sur les trois façades, les annotations des experts révèlent que l'ensemble des altérations représentent environ 22% des surfaces des murs. Les détails des proportions sont consignés dans le tableau 1.4.

Tableau 1.5 – Proportions de surface de chaque type d'altérations.

Altérations	Aire (pixel)	Aire (%)
Desquamation en plaque	95321512	98,40
Desagrégation sableuse	154626	0,16
Desquamation en feuillet	185434	0,19
Erosion	1078428	1,11
Pelage	128944	0,14
Total des altérations	96868944	100

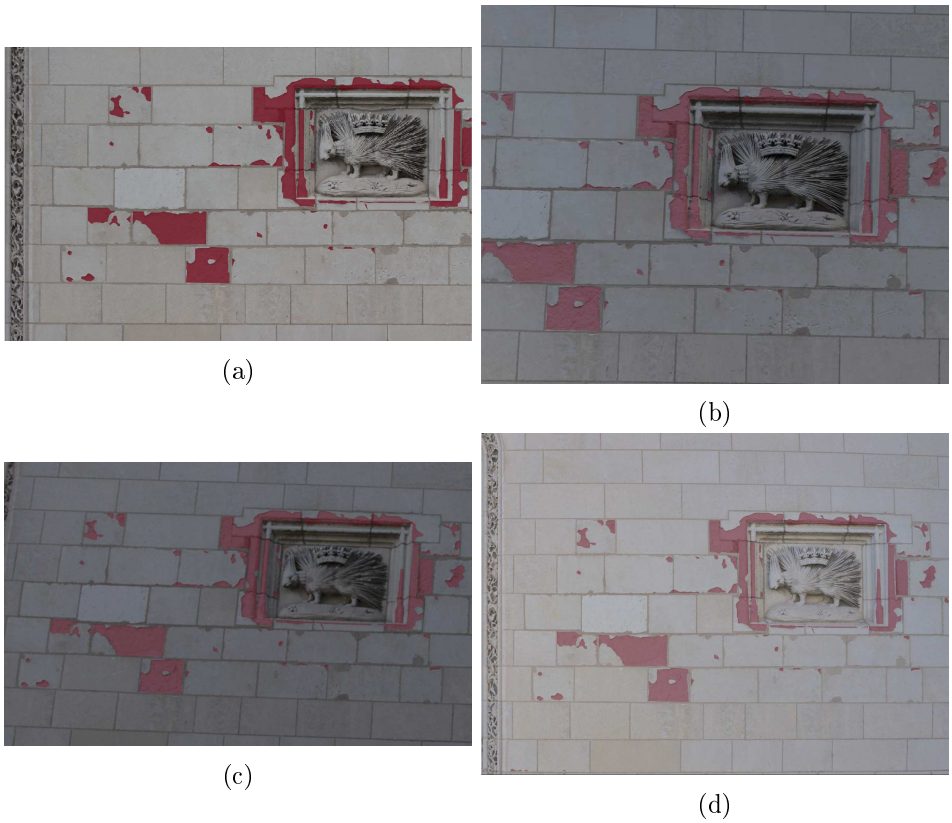
Le tableau 1.5 résume les proportions de chaque type d'altérations. Les desquamations en plaques couvrent $98,40\%$ de la surface d'altération. L'erosion représente 1% environ. Les autres types daltérations sont chacune sur une surface d'environ 0.2%, une présence très faible voir inexistante. Les desquamations en plaques restent un type d'altération très présent sur les châteaux du style Renaissance de la vallée de la Loire.

Dans le développement des réseaux d'apprentissage machine étudiés, il est important de disposer d'une base de vérité-terrain suffisamment large pour obtenir des réseaux de détection d'altérations plus perfomants. Pour ce faire, des techniques d'augmentation des images par générations d'images artificielles [27, 28] ou transformations des images existantes (rotation, contraste, retournement, etc.)

sont couramment utilisées [29, 30].

Une nouvelle approche plus adaptée consiste à la rétroprojection des annotations réalisées sur l'orthomosaïque sur les images couleur originales ayant servis à concevoir l'orthomosaïque. Cette approche tire avantage du grand recouvrement des images couleur originales, pour obtenir une augmentation d'images cohérente, avec plusieurs points de vue différents de la même zone d'altérations. De plus, cela permet de travailler directement sur les images couleur originales qui ont une meilleure résolution que l'orthomosaïque avec un rendu plus précis des textures.

La figure 1.11 présente un exemple de l'augmentation des données par reprojection sur les images couleur des altérations annotées par les experts sur les orthomosaïques. Chaque annotation, reprojetée reproduit fidèlement l'annotation effectuée par l'expert sur l'orthomosaïque. On obtient dans cet exemple, trois images avec une résolution plus élevée, des points de vue et des expositions variables.



(a) (b) (c) (d)

FIGURE 1.11 – Augmentation des données par reprojection, sur les images couleur, des altérations annotées par les experts sur les orthomosaïques. (a) : zoom sur l'othophoto présentée dans la figure 1.10. (b), (c) et (d) : résultats de la reprojection sur 3 images couleur originales correspondantes (focale 100mm, focale 24mm, drone). Ces images ont été réalisées par les experts en patrimoine du Laboratoire LaMé.

Il est important de préciser que les acquisitions réalisées pour concevoir cet ensemble de données n'étaient pas dans un but de détection d'altération. Les techniques de prises de vues n'ont donc pas spécialement tenu compte des altérations comme étant le centre d'intérêt de l'opération. Il n'y a donc pas eu de zoom ni de cadrages spécifiques sur les altérations. La forte présence d'altération sur les façades a encouragé l'exploitation de cet ensemble de données dans nos travaux à des fins de détection d'altération. Ce choix ouvre donc la voix à de nouvelles applications pour des données existantes démontrant les mêmes observations.

1.5 Bilan du chapitre

Dans ce chapitre, nous avons exploré les deux ensembles de données : l'un exploité pour la segmentation pierre-à-pierre qui est constitué d'orthomosaïques du château de Chambord ; et l'autre exploité pour la détection d'altérations qui comporte des orthomosaïques du château de Chaumont-sur-Loire.

Pour les processus d'acquisitions et de conception des données, les experts ont adopté les techniques de photogrammétrie et scanner laser terrestre. Les campagnes d'acquisitions ont été effectuées sur site à différents moments dans le temps commençant par le château de Chambord puis le château de Chaumont quelques années plus tard.

Cinq types d'altérations sont annotés sur les othophotos couvrant 22% des façades. Parmis ceux-ci, la desquamation en plaque constitue la plus dominante représentant 98% de la surface des altérations. Elle est aussi un type d'altération très préjudiciable à l'acceuil des touristes du fait de la chute de matière qui en découle.

Dans les chapitres suivants, nous décrivons l'utilisation de ces données pour l'évaluation des méthodes proposées.

Chapitre 2

Segmentation pierre-à-pierre

> La machine analytique n'a pas de prétention à donner naissance à quoi que ce soit. Elle peut faire ce que nous savons lui apprendre à faire.
>
> Ada Lovelace

Sommaire

2.1 Introduction

Pour documenter et étudier le patrimoine historique bâti, les experts produisent des cartographies où sont tracés à main levée les contours des pierres et des joints de la maçonnerie [4]. La figure 2.1 montre un exemple de cette segmentation pierre-à-pierre réalisée manuellement. La segmentation pierre-à-pierre représente une étape importante pour la mise en place d'outils de surveillance de l'état de santé des châteaux. Elle reste possible à réaliser manuellement pour de petite zones, mais serait trop laborieuse à réaliser à l'échelle d'un grand château comme Chambord. D'où la nécessité de développer un outil de segmentation automatique.

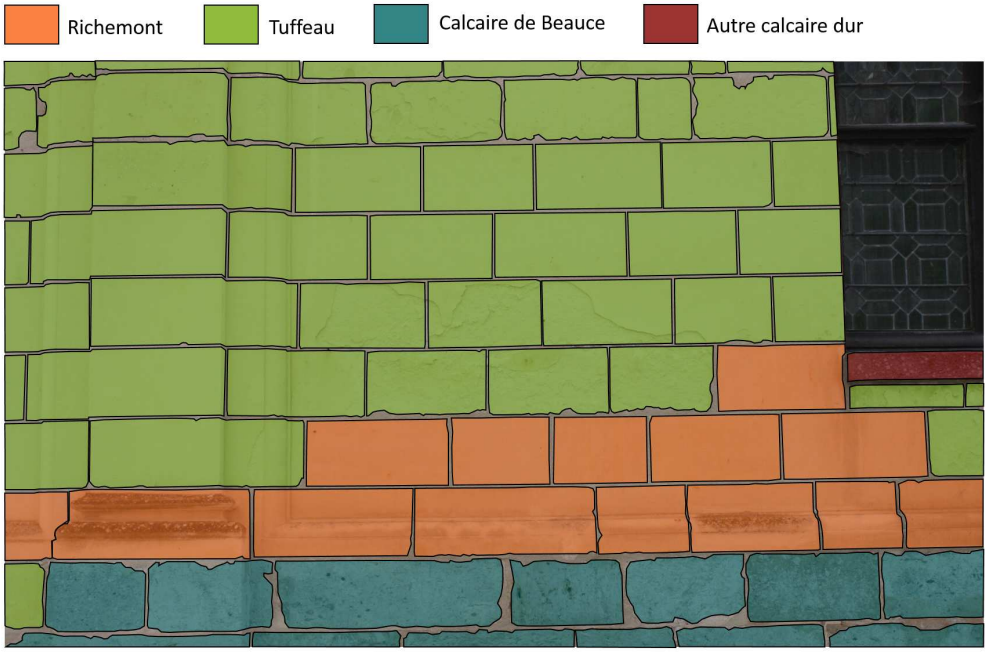


FIGURE 2.1 – Exemple de segmentation pierre-à-pierre réalisée manuellement par une experte du patrimoine [4].

Ce chapitre dresse un état de l'art des méthodes utilisées pour la segmentation pierre-à-pierre automatique des maçonneries à partir d'images ou de nuage de points (section 2.2). Les approches classiques appliquées dans le cadre de ces travaux sont détaillées dans la section 2.3. La section 2.4 présente la méthode ad-hoc adoptée pour produire notre base de données labélisées nécessaire à l'apprentissage machine. Puis les principales architectures de réseaux de neurones utilisées pour la segmentation sont décrites dans la section 2.5. Enfin, dans la section 2.6, sont

exposés et discutés les résultats des expérimentations sur les approches de l'état de l'art et les méthodes par apprentissage profond. La section 2.7 fait le bilan du chapitre.

2.2 État de l'art de la segmentation pierre-à-pierre

Les méthodes présentées dans la littérature pour la segmentation pierre-à-pierre sur des patrimoines historiques bâtis peuvent être regroupées en deux catégories : celles s'appliquant sur les nuages de points 3D, d'une part, et celles s'appliquant sur les images couleur, d'autre part. On peut aussi les regrouper en techniques de segmentation d'images classiques, d'une part, et les méthodes utilisant les algorithmes d'aprentissage machine, d'autre part.

L'agencement des pierres est détecté par les différences de relief ou de courbure lorsque leur résolution est suffisante. Les nuages de points 3D provenant de LiDAR ou de scanner laser terrestre (TLS) sont largement utilisés pour la modélisation virtuelle et la surveillance de monuments historiques. Les auteurs de [31], [32] utilisent des données 3D TLS pour la surveillance des façades de bâtiments historiques faits de pierres taillées. Elles permettent aussi d'identifier les anomalies de structure, les signes de dislocation et d'effondrements [33, 34] ainsi que la classification des matériaux de construction courants à l'aide de l'apprentissage automatique [35]. Les nuages de points 3D ont aussi servi à réaliser la reconnaissance de régions détériorées [36, 37] ; détecter l'état de corrosion des bâtiments historiques [38, 39] et effectuer la détection de texture pour la classification des monuments historiques [40], [41],(voir [42] pour une revue des techniques de modélisation 3D avec leurs limites et leurs potentialités).

Parmi les premiers travaux abordant la problématique spécifique de la segmentation pierre-à-pierre, Valero et al. [43, 5] utilisent des nuages de points 3D provenant de la façade sud de la chapelle royale de Stirling Castel (Écosse) qui possède des murs de maçonnerie en moellons (voir figure 2.2).

Ils adopent une approche basée sur la transformée en ondelette continue en passant par une carte de profondeur [44]. La carte de profondeur est une image à deux dimensions dont chaque pixel code la distance à la surface observée. Sur cette

carte, est appliquée la transformée en ondelette continue, puis un post-traitement utilisant la dilatation et l'algorithme de l'enveloppe convexe. Le résultat de la segmentation est ensuite reprojeté sur le nuage de points 3D pour la visualisation. La précision de cette approche, en terme de segmentation, n'a pas fait l'objet d'une évaluation quantitative par les auteurs, elle a plutôt servi à identifier des régions d'intérêts pour l'analyse des dégradations sur la surface des pierres.

FIGURE 2.2 – Une partie du mur maçonnerie en moellons de la chapelle royale de Stilrling Castel [5].

Hess et al. [45] utilisent aussi des nuages de points 3D obtenus par photogrammétrie. Les auteurs développent une approche améliorant l'outil brosse à dessin existant au sein du langage de programmation javascript. Un point de départ est choisi par l'utilisateur sur le nuage de points comme appartenant aux joints et, par croissance de régions, tous les points du nuage de points respectant les contraintes de couleur et d'angle vis-à-vis du point de départ sont agrégés comme appartenant aux joints. Pour la contrainte de couleur, le point courant doit être dans un intervalle de plus ou moins trois écarts types autour de la couleur du point initial. Pour la contrainte d'angle, un seuil minimal est fixé sur le produit scalaire des normales du point courant et du point de départ.

Kajatin et al. [6] proposent une approche avec un robot ratisseur de joint (voir figure 2.3) construit sur mesure pour leur étude. Avec ce robot, ils parcourent au plus près le mur et enregistrent des vidéos, desquelles ils extraient les images couleur de 848×480 pixels et les données de profondeur. Ces images sont en-

suite pré-traitées par un filtrage à moyenne mobile exponentielle. Puis segmentées avec huit classifieurs différents par apprentissage machine : k plus proches voisins [46], analyse discriminante quadratique [47], modèle bayésien [48], séparateur à vaste marge [49], arbres de décision, forêt aléatoire [50], AdaBoost [51] et UNet [9]. Les huit résultats de segmentation sont individuellement affinés par des opérations morphologiques. Enfin, la segmentation finale est la moyenne pondérée des segmentations de chaque classifieur.

FIGURE 2.3 – Robot ratisseur de joints [6].

Les méthodes précédentes exploitent le relief issu du nuage de points 3D pour distinguer les pierres et les joints ou des différences de teintes entre les deux. Elles sont donc adaptées aux maçonneries qui présentent des différences de relief suffisamment marquées vis-à-vis de la résolution des techniques d'acquisition (TLS, photogrammétrie). C'est le cas des maçonneries en moellons de la Chapelle médiévale de Stirling. Cependant, le style architectural des châteaux de la Loire en pierres taillées se caractérise, au contraire, par une recherche d'uniformité de surface et de couleur entre les joints des pierres. Ces derniers ne présentent donc qu'un très léger relief, quasi imperceptible et une teinte très proche de celle de la surface des pierres. Les opérations de restauration pour remplacer les parties abîmées de certaines pierres sont aussi réalisées de façon à masquer au maximum ces ajouts. Ainsi, les nouveaux joints sont plus fins, sans relief et de même teinte que la surface des pierres avoisinantes.

D'autres travaux abordent la segmentation sur des images couleur. Les auteurs de [7] ont réalisé une segmentation semi-automatique des pierres dans le but d'extraire des caractéristiques pour la classification des bâtiments historiques en trois catégories selon l'arrangement des pierres sur les façades. La figure 2.4 montre des exemples pour chacune des catégories. Ils utilisent la transformée de Hough probabiliste pour trouver des segments de droites sur les joints et contours des pierres. Les propriétés géométriques de ces segments fournissent, selon leur étude, les caractérisques nécessaires pour la tâche de classification. Les images couleur utilisées dans cette étude sont des photos de murs sur différents sites historiques du Pays basque : Durango et Urdaibai. Chaque image couleur est convertie en niveaux de gris et subdivisée en régions d'intérêt (ROI) de taille 300×300 pixels. Les ROI sont traitées, chacune indépendamment par une égalisation d'histogramme avant d'extraire les segments de droites. La combinaison des segments de droites de toutes les ROI permet d'obtenir la répartition des segments pour l'image complète.
Cette approche, en plus d'être semi-automatique, ne fournit pas assez de segments de droites pour détecter tous les joints, et les segments ne contiennent pas assez de détails pour réaliser la segmentation pierre-à-pierre avec précision.

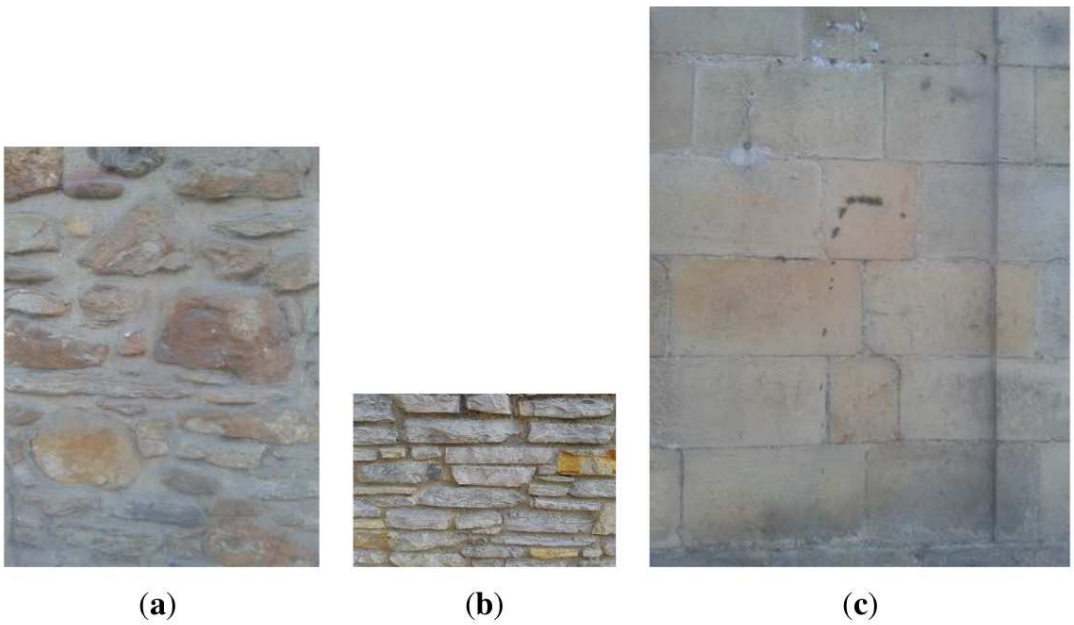


FIGURE 2.4 – Exemples des trois catégories considérées pour la classification des bâtiments historiques [7].

Dans [52] et [53], les auteurs utilisent l'intensité du LiDAR (réflectivité ponctuelle de l'objet touché par l'impulsion laser) pour analyser la voûte d'un pont en maçonnerie, et les murs d'une fortification Muraille médiévale d'Alcaçova au Portugal. Ils génèrent une image monochrome qui est ensuite pré-traitée par des filtres médian et gaussien avant d'appliquer l'opérateur sobel pour l'extration des contours. Le résultat de cette segmentation sert à produire un masque de départ pour la segmentation par ligne de partage des eaux. Cette approche peut être adaptée à des images couleurs après conversion en niveaux de gris. Cependant, la qualité de la segmentation finale, reste fortement dépendante du masque initial et très sensible, au bruit ainsi qu'aux altérations et variations de luminosité dans l'image.

Pour pallier cette insuffisance, Ibrahim et al. [54] proposent d'utiliser l'apprentissage profond et le réseau U-Net pour générer le masque de départ. Leur base de données contient 162 images de taille 512×512 pixels, qu'ils ont réparties comme suit : 117 pour l'entrainement, 13 pour la validation et 32 pour les tests. L'entrainement de U-Net est réalisé avec la fonction d'optimisation Adam. La robustesse du réseau de neurones a permis d'obtenir un masque plus fiable et de réaliser des scores de bonne segmentation de plus de 80% (F1-score de $81,75$ %; Precision $81,16$ %; Recall $82,14$ %).

Dans la section suivante, nous décrivons les méthodes classiques utilisées dans les différentes approches de segmentation pierre-à-pierre de l'état de l'art et testées sur nos images.

2.3 Méthodes classiques de segmentation pierre-à-pierre

2.3.1 Détection de contours

Lorsque les images des murs présentent un contraste ou une discontinuité de luminance entre les pierres et les joints, une première approche de segmentation consiste à appliquer une technique de seuillage ou de détection de contours.

L'algorithme d'**Otsu** est une méthode très populaire de seuillage automatique [55]. La valeur seuil est automatiquement définie à partir de l'histogramme de l'image pour donner la meilleure segmentation en deux classes (contours et sur-

faces) maximisant la séparabilité tout en maintenant la variabilité intra-classe. Malgré sa simplicité, cette méthode donne d'assez bons résultats sur un large éventail de cas.

Le filtre de **Canny** est un détecteur de contours très populaire pour les tâches de segmentation d'images [56]. Deux paramètres critiques sont les seuils lors de l'étape du seuillage par hystérésis. Le réglage de ces deux seuils (haut et bas) permet de garder uniquement les pixels des contours signicatifs. Typiquement, si l'intensité du gradient d'un pixel est supérieure au seuil haut, il sera conservé comme signicatif ; si elle est inférieure au seuil bas, il sera supprimé ; les pixels situés entre les deux seuils ne seront conservés que s'ils sont contigus à des pixels de contour déjà retenus (en cherchant dans la direction perpendiculaire à leur gradient).

2.3.2 Transformée de Hough

Lorsque la maçonnerie présente un agencement linéaire, le découpage des pierres peut être opéré en appliquant la transformée de Hough. Cet algorithme a été inventé par Paul Hough pour détecter des droites dans une image [57]. Il cherche toutes les droites possibles passant par les pixels de contour en utilisant un accumulateur dans le plan polaire discretisé. Pour l'ensemble des pixels de contour dans l'image binaire, on obtient une multitude de courbes sinusoïdales dans le plan polaire, qui s'entrecroisent. Les maxima d'accumulations correspondent aux droites recherchées.

Au delà des droites, l'algorithme de la transformée de Hough a été généralisé pour détecter de nombreux autres types des formes telles que les cercles, les ellipses, les rectangles [58, 59, 60]. Il nécéssite beaucoup de ressources de calcul pour parcourir et évaluer chaque pixel de l'image. Pour surmonter cette limitation plusieurs versions de l'algorithme appelé transformée de Hough probabiliste ont été proposées [61, 62, 63]. Dans celles-ci, l'algorithme évalue un échantillon aléatoire des pixels de contours, ce qui réduit le temps et les ressources de calcul. Ils permettent aussi d'obtenir les extrémités des segments de droites.

2.3.3 Transformée en ondelette continue

En traitement d'image, la transformée en ondelette est utilisée pour la compression d'images [64, 65] et aussi pour la détection et la caractérisation des contours dans une image [66, 5]. Une ondelette est une fonction de base correspondant à une petite oscillation (similaire à la transformée de Fourier à court terme) utilisée pour décomposer un signal [66]. La transformée en ondelettes continue est l'une des variantes de la transformée en ondelette qui consiste à décomposer un signal par des convolutions à plusieurs échelles avec des ondelettes [44, 67], en utilisant le plus souvent l'ondelette en chapeau mexicain.

2.3.4 Segmentation par ligne de partage des eaux

La segmentation par ligne de partage des eaux est une des méthodes classiques de segmentation en régions. L'algorithme a été défini dans sa première version par Meyer et al. [68, 69]. L'image en niveaux de gris est considérée comme une surface topographique avec des crêtes et des vallées. Les pixels à valeurs minimales contiguës sont progressivement agrégés, selon le principe de la montée des eaux (la méthode tire son nom de cette analogie). Il faut éviter de fusionner l'eau de différentes vallées. Ainsi, à chaque endroit où des regroupements de pixels tendent à fusionner, il faut construire une barrière. Les barrières construites représentent les contours des objets dans l'image et le résultat de la segmentation.

La segmentation par ligne de partage des eaux est très sensible au bruit ou toutes autres irrégularités dans l'image, ce qui conduit à une sursegmentation. Pour pallier cette sensibilité, la principale solution est la définition de marqueurs pour définir les vallées devant être fusionnées. Plusieurs auteurs ont proposé différentes approches pour la définition de marqueurs optimaux [70, 71, 72].

La section suivante propose une première approche semi-automatique ad-hoc combinant des méthodes classiques pour segmenter les images de murs en pierres de calcaire taillées des châteaux de la vallée de la Loire. L'objectif est de fournir des images pré-labellisées qui seront par la suite corrigées et validées par les experts en vue de l'apprentissage machine supervisé.

2.4 Approche de segmentation pierre-à-pierre par détection de contours et opérations morphologiques

La chaine de traitements proposée est illustrée sur la figure 2.5. Elle consiste en une série de pré-traitements de l'image suivie d'une première détection des contours des pierres puis d'une seconde détection qui vient affiner la première ; ensuite nous procédons à une application de traitements morphologiques visant la fermeture des contours, le remplissage de la surface des pierres ; puis nous procédons à la suppression des très petites régions en s'appuyant sur des connaissances a priori ; et enfin nous réappliquons un traitement morphologique pour réduire l'épaisseur des joints détectés.

L'étape de pré-traitements consiste à effectuer plusieurs filtrages : médian 5x5 puis Wiener 3x3 [73] sur l'image en niveaux de gris pour réduire considérablement les bruits sur la surface des pierres, puis à renforcer les contours, et enfin le contraste de l'image par une égalisation d'histogramme adaptatif [74].

Après pré-traitements, la détection grossière des contours est effectuée par la méthode de Canny [56]. Afin de paramétrer automatiquement le seuillage par hystérésis, on utilise le seuil de binarisation automatique SB fourni par le détecteur de Sobel [75] comme seuil d'hystérésis haut, et $0,4 \times SB$ comme seuil bas. Cette combinaison produit un résultat plus précis que les seuils laissés par défaut, comme le montre la figure 2.6.

La fermeture des contours consiste en une dilatation avec deux éléments structurants, l'un horizontal et l'autre vertical, qui tiennent compte de la dimension approximative des pierres. Le but visé par cette étape est de recouvrer les parties de joints manquants et consolider les joints très fins.

Puis, nous appliquons une opération morphologique de remplissage des trous, qui supprime tous les minimas qui ne sont pas connectés aux contours dans l'image ou, de manière équivalente, impose l'ensemble des minimas qui sont connectés aux contours dans l'image [76]. Cette opération permet de supprimer les artéfacts provenant des taches sur la surface des pierres. Enfin, toutes les régions dont l'aire est inférieure au tiers de la surface standard d'une pierre, sont supprimées. Ces régions sont généralement des zones situées sur des joints un peu larges ou des

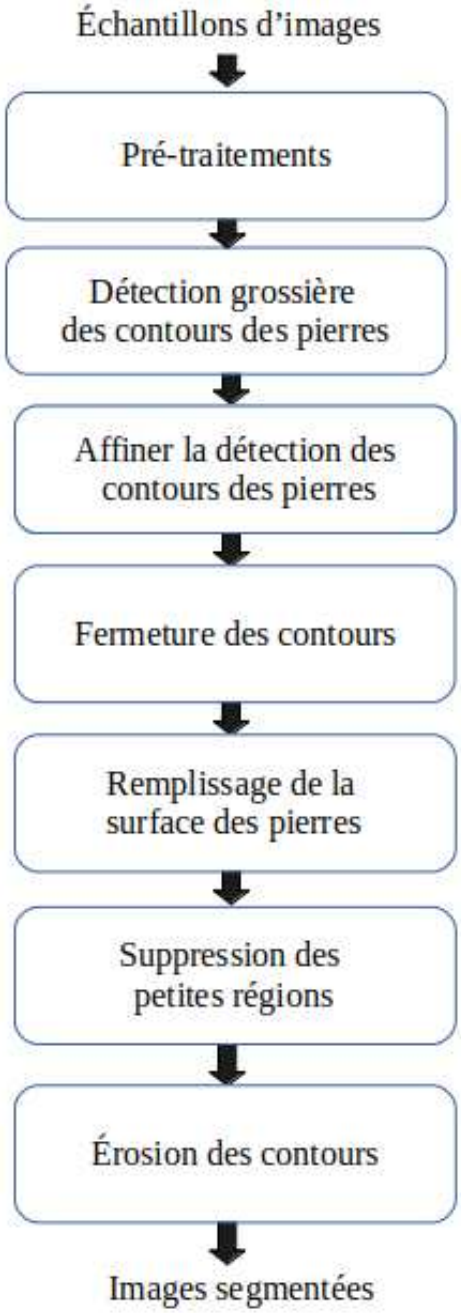


FIGURE 2.5 – Chaine de traitements proposée pour la segmentation des pierres.

pierres incomplètes en bordure de l'échantillon.

Cette première approche ad-hoc sera exploitée pour la production d'une base de données pour l'apprentissage et les tests des réseaux de neurones étudiés dans la section suivante. La figure 2.7 montre quelques exemples d'images couleur, le

résultat de l'approche et le résultat final après correction par les experts.

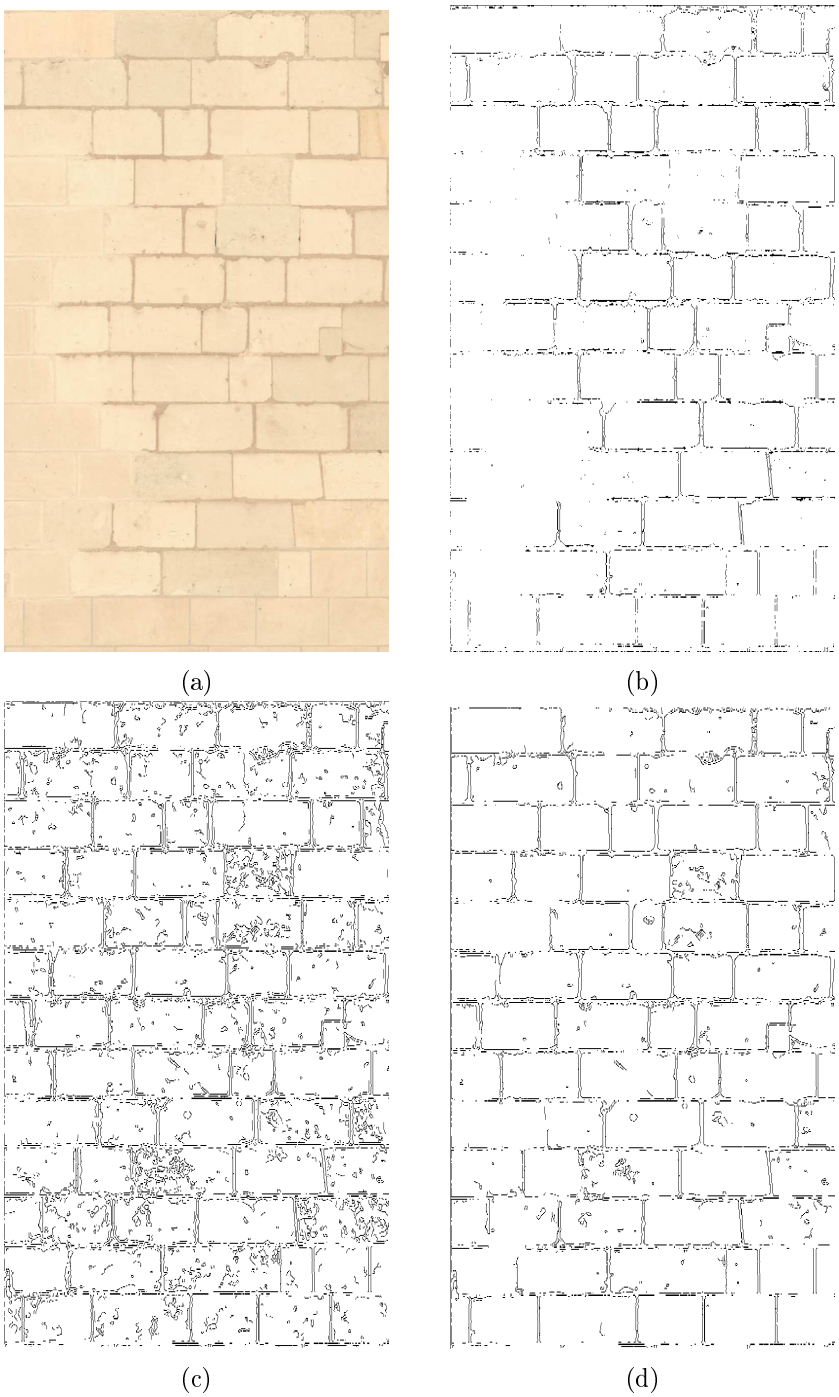


FIGURE 2.6 – Exemple de détection des contours : (a) Image couleur ; (b) Sobel ; (c) Canny ; (d) Approche proposée avec Sobel et Canny

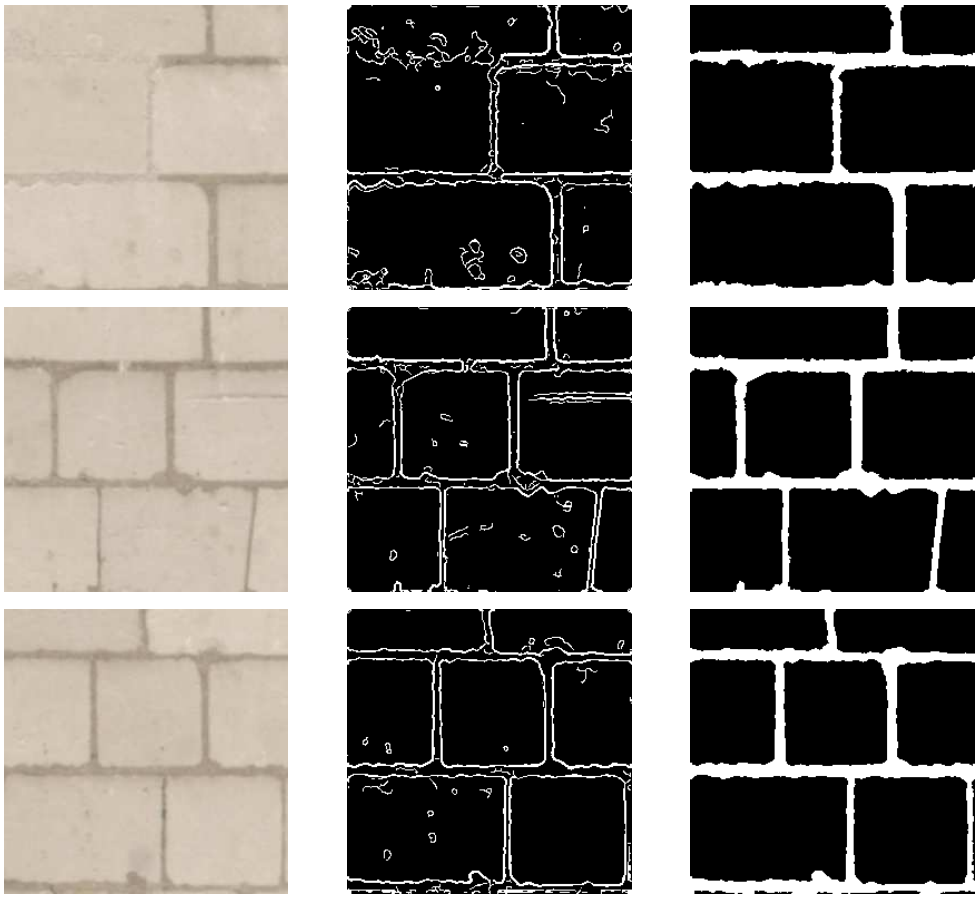


FIGURE 2.7 – Exemples d'utilisation de l'approche proposée pour la réalisation de la base de donnée pour l'apprentissage. **Colonne 1 :** Images couleur ; **Colonne 2 :** Résultats de la détection de contours avec l'approche proposée ; **Colonne 3 :** Résultats finaux après correction et validation par les experts.

2.5 Méthodes de segmentation par apprentissage profond

L'apprentissage profond est une forme d'apprentissage automatique exploitant des réseaux de neurones dans lequel le nombre de couches est considérablement élevé. Dans le domaine du traitement d'image, les réseaux de neurones convolutifs (CNN pour Convolutional Neural Network, en anglais) sont inspirés du fonctionnement du cortex visuel des animaux [77]. Récemment, ces réseaux ont démontrés des performances remarquables dans plusieurs applications de vision comme la détection d'objets [78], la classification [79], ou la segmentation d'images [10], [8], [80].

Les CNN pour la segmentation sémantique sont bien adaptés à notre problématique puisqu'ils fournissent un label pour chaque pixel selon sa classe d'ap-

partenance. Ils permettent d'aborder notre problème de segmentation comme une classification binaire : pixels appartenant à la classe *joint* et pixels appartenant la classe *pierre*. Les CNN sont également connus pour être plus robustes aux grandes variations de luminosité présentes dans nos images.

De manière générale, l'architecture d'un réseau de neurones convolutifs profond est composée de plusieurs couches empilées. La première couche reçoit en entrée l'image couleur, puis, chaque couche suivante reçoit en entrée la sortie de la couche précédente. Une couche constitue une étape où s'effectue des opérations. Ces opérations peuvent être des convolutions, l'application de fonctions d'activation (ReLU, sigmoid, tangente hyperbolique) ou des opérations de pooling (échantillonnage) [81]. La convolution est la multiplication d'une matrice par un noyau de convolution de taille 1×1, 3×3, 5×5 ou autres. On parle de convolution 1×1, 3×3, 5×5 ou $n \times n$, avec $n \in \bar{\mathbb{N}}$.

Une fonction d'activation effectue un calcul au niveau de chaque neurone du réseau pour déterminer sa sortie. La fonction de rectification unitaire (ReLU : $f(x) = max(0, x)$) est la plus largement utilisée car elle permet une convergence plus rapide lors de l'entrainement du réseau en optimisant les ressources.

La sortie d'une couche est une matrice de caractériques de taille $h \times l \times c$ avec h la hauteur, l la largeur et c le nombre de canaux de la matrice. Les dimensions h et l se réduient par le jeu des convolutions au fur et à mesure que le réseau devient profond. Le nombre de canaux est conditionné par le nombre de noyaux de convolutions appliqués. Des opérations de concaténation des sorties entre couches éloignées permettent également de fusionner des caractéristiques à plusieurs échelles .

La dernière couche est l'application d'une fonction de sortie, par exemple softmax, qui prend en entrée la matrice de caractériques provenant des couches précédentes et produit en sortie, soit un masque de segmentation, soit un label, soit les coordonnées de la position des objets dans l'image d'entree, suivant qu'il s'agit d'un réseau de segmentation d'images, de classification d'images ou de détection d'objets, respectivement.

Cette section présente les architectures des réseaux d'apprentissage profond

testées dans notre étude de segmentation pierre-à-pierre.

2.5.1 FCN

FCN (Fully Convolutionnal Network) est un réseau entièrement convolutif introduit en 2015 par Long et al. [8] pour la segmentation sémantique d'une image. Il transforme les architectures des réseaux AlexNet [82], GoogLeNet [83] et VGG16 [84], initialement conçues pour prédire la classe correspondante à une image en remplaçant dans ces réseaux, la couche entièrement connectée par une couche de convolution. La couche entièrement connectée reçoit en entrée un vecteur de caractéristiques de taille fixe k, avec k le nombre de classes. La couche de convolution quant à elle, reçoit en entrée des matrices de caractéristiques de toutes tailles, et produit en sortie des matrices de caractéristiques de taille identique à celles fournies en entrée.

L'architecture du réseau FCN est représentée dans la figure 2.8. Elle prend en entrée une image de dimension $h \times l \times c$, avec h la hauteur, l la largeur et c le nombre de canaux dans l'image, et prédit en sortie un masque de segmentation de dimension $h \times l \times 1$. L'image d'entrée subit des convolutions 3×3 et rectifications linéraires (ReLU), suivies d'une opération de sous-échantillonnage 2×2 avec un pas (stride) de 2. Pour compenser cette réduction de taille, le masque de segmentation est obtenu en appliquant une opération de sur-échantillonnage 2×2 avec un pas (stride) de 2, sur la matrice de caractéristiques issue des convolutions.

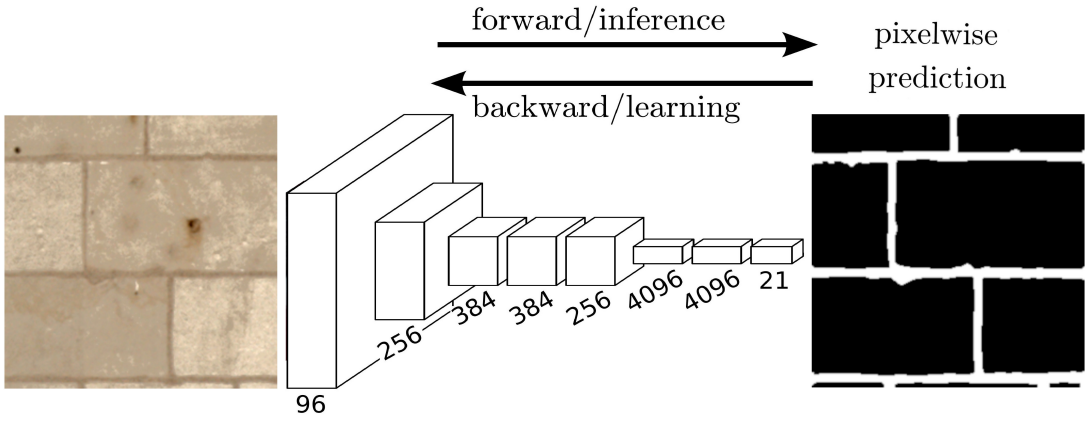


FIGURE 2.8 – Architecture du réseau FCN [8]

2.5.2 U-Net

U-Net est un réseau de neurones convolutif introduit en 2015 pour la segmentation d'images biomédicales [9]. Il a depuis lors été largement utilisé dans la littérature pour diverses tâches de segmentation d'images hors du domaine médical. U-Net est connu pour ses bonnes performances avec des bases d'apprentissage relativement petites. Son architecture (voir Figure 2.9) dont la forme en **U** a inspiré le nom, est composée de deux parties : l'encodeur (branche gauche du U) et le décodeur (branche droite du U).

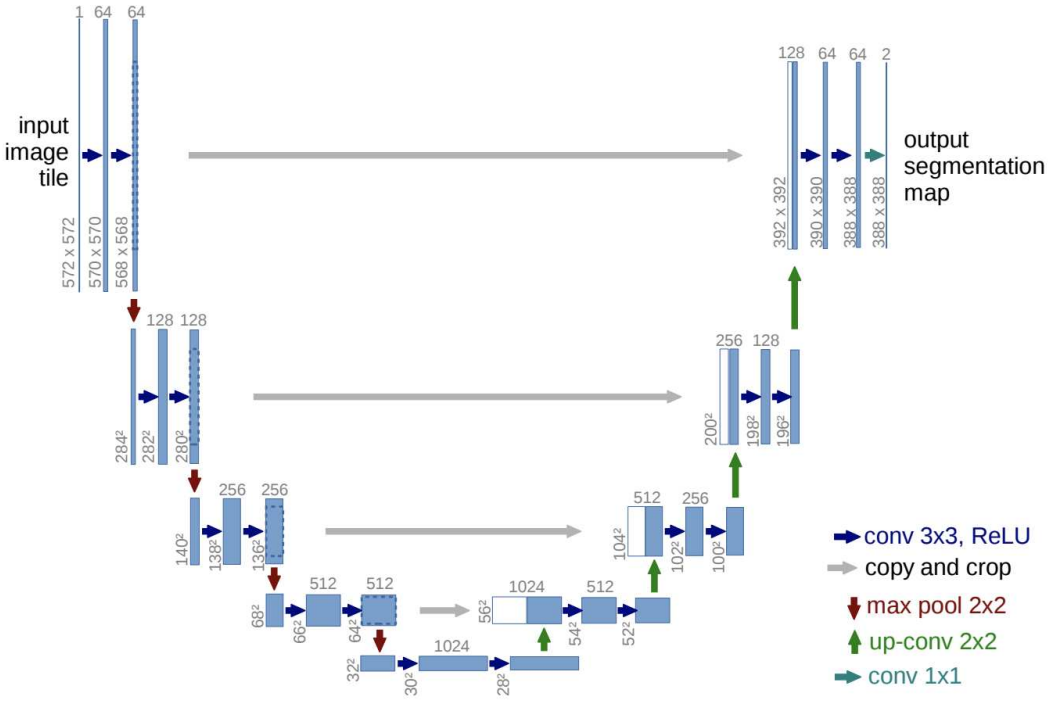


FIGURE 2.9 – Architecture du réseau U-Net [9]

L'encodeur prend en entrée l'image couleur qu'il passe à travers des couches successives pour produire en sortie une matrice de caractérisques qui sera exploitée par le décodeur. Chaque ensemble de couches est constitué d'opérations de convolution 3×3 suivies par une opération de ReLU, puis d'une opération de sous-échantillonnage (max-pooling) 2×2 avec un pas (stride) de 2. Le décodeur prend en entrée la matrice des caractéristiques provenant de l'encodeur et les concatène avec les caractéristiques de la couche précédente. Ensuite des opérations de convolution 3×3 puis de ReLU, et une opération de sur-échantillonnage 2×2. La

couche finale du décodeur réalise une convolution 1×1 pour prédire le masque de segmentation.

2.5.3 SegNet

SegNet est un réseau de neurones convolutif introduit en 2015 [10]. Le réseau est basé sur une architecture encodeur-décodeur (figure 2.10). Il contient une partie encodeur similaire à l'architecture du FCN décrite à la section 2.5.1 et une partie décodeur suivie d'une couche de classification des pixels. La partie encodeur est constituée des couches de convolution provenant du réseau VGG16 [84]. Il s'agit de 13 couches de convolution 3×3 disposées en cinq blocs. Chaque bloc de convolution est suivi d'une couche de sous-échantillonnage (max pooling) 2×2, avec un pas (stride) de 2, dont les indices sont stockés.

Les différents sous-échantillonnages effectués dans la partie encodeur entraînent progressivement une réduction de la résolution spatiale de la matrice des caractéristiques. Le décodeur y remédie en utilisant un sur-échantillonnage. Il prend en entrée la matrice de caractéristiques résultant de l'encodeur et est constitué de bloc de convolutions précédés d'une couche de sur-échantillonnage par un facteur de 2. Le sur-échantillonnage est effectué en utilisant les indices précédemment stockés à l'issue de chaque max pooling. Cette réutilisation des indices permet de limiter la perte de l'information spatiale causée par le max pooling. Enfin, une couche de classification des pixels utilisant le classifieur softmax, prédit le masque de segmentation.

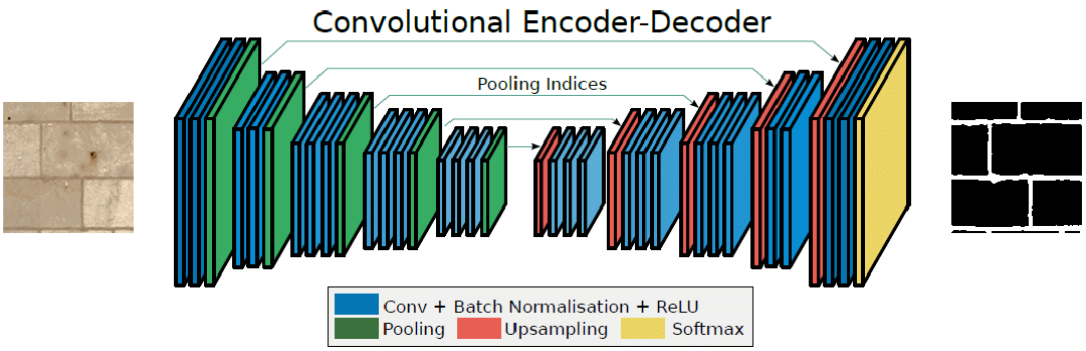


FIGURE 2.10 – Architecture du réseau SegNet [10]

2.5.4 DeepLab

Le réseau de segmentation sémantique Deeplab a été introduit en 2015 par Chen et al [85] puis fait l'objet de différentes versions successives [86, 87, 88]. Dans nos travaux, nous avons utilisé la version dénomée DeepLabv3+, la plus récente et plus performante [88]. La spécificité de l'architecture de DeepLabv3+ est l'utilisation de convolution à trous (figure 2.11). La convolution à trous permet d'élargir les champs perceptifs des filtres pour incorporer un contexte plus large sans augmenter le nombre de paramètres ou la quantité de calcul.

La partie encodeur est constituée de couches de convolution provenant du réseau ResNet-18 [13]. Il contient 18 couches de convolution et encode des informations contextuelles multi-échelles en appliquant la convolution à trous à plusieurs échelles. La matrice de caractéristiques provenant de la dernière couche de l'encodeur subit une convolution 1×1 pour réduire le nombre de canaux. La sortie de la convolution 1×1 est concaténée avec la taille de la matrice des caractéristiques correspondante dans la partie décodeur. Ensuite, une opération de convolution 3×3 et un sur-échantillonnage par facteur 4 sont appliqués pour prédire le masque de segmentation.

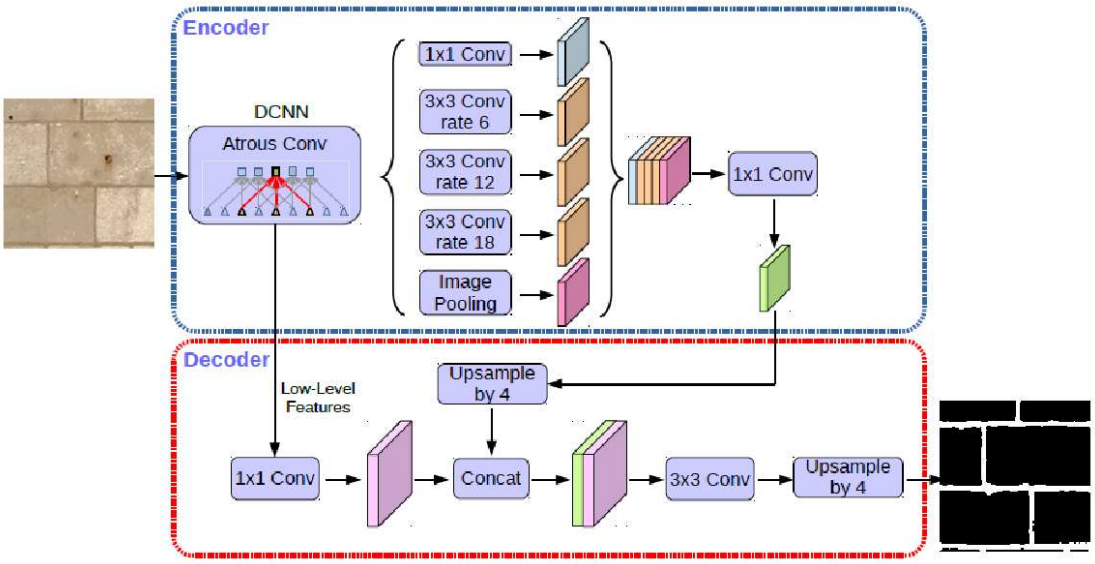


FIGURE 2.11 – Architecture du réseau DeepLabv3+ [11]

2.6 Expérimentations et résultats

2.6.1 Base de données PAP

La base de données PAP est constituée de 245 imagettes de taille 256×256 pixels extraites des orthomosaïques des murs d'enceinte du château de Chambord (figure 1.5). La vérité de terrain est fournie par une labellisation semi-automatique effectuée par des experts. Dans un premier temps, l'approche décrite à la section 2.4 est utilisée pour réaliser une première annotation ensuite elle est manuellement corrigée et complétée par les experts pour obtenir la vérité-terrain.

Nous avons appliqué différentes techniques d'augmentation des données : deux types de variations de luminosité, deux types de variations de contraste, retournement et flou. Dans le choix des différentes techniques appliquées, nous prenons en considération une intégrité de l'image dépendant du domaine, les joints étant orientés horizontalement ou verticalement, les rotations ont été utilisées en maintenant cette contrainte. Comme les campagnes d'acquisition des images sont souvent réalisées dans des conditions d'éclairage diverses, les techniques telles que la variation de la luminosité et du contraste ont été davantage privilégiées.

La première variation de contraste consiste à modifier l'intensité des pixels de chaque canal de couleur dans la plage $[0,3; 0,9]$. La seconde convertit l'image dans l'espace de couleur HSV puis met à l'échelle la valeur des pixels V par un facteur de distribution uniforme sur l'interval $[-2; 4]$. Deux changements de luminosité sont également appliqués : un assombrissement de la valeur des pixels par l'ajout d'une valeur dans la plage $[-0,6; -0,1]$, un éclaicissement par l'ajout d'une valeur dans la plage $[0,1; 0,6]$. Ensuite, l'image subit une réflexion affine sur l'axe X ou l'axe Y. Enfin, un processus d'ajout de flou est appliqué à l'image à l'aide d'un filtre gaussien.

La base de données ainsi augmentée est composée de 1715 imagettes de taille 256×256 pixels. La figure 2.12 présente quelques exemples de la base. Les imagettes présentent des caractéristiques assez variées en terme de luminosité, de contraste, et d'altérations de teintes semblables à celles des joints. La base de données a été séparée en 70% pour l'apprentissage des réseaux, 15% pour la validation et 15% pour les tests. Nous avons utilisé des versions pré-entrainées de chaque réseau sur

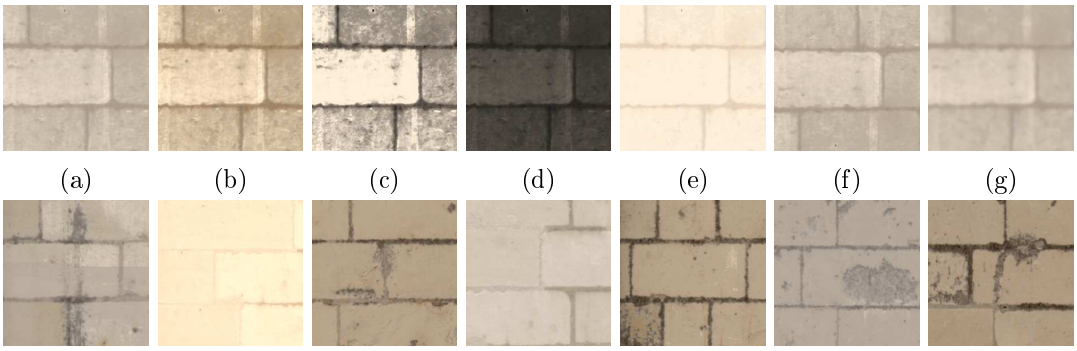


FIGURE 2.12 – Exemple d'imagettes de la base PAP et des techniques d'augmentation appliquées.
Première ligne : (a) Image originale. (b) après la première modification du contraste. (c) après la deuxième modification du contraste. (d) après la première modification de la luminosité. (e) après la deuxième modification de la luminosité. (f) après réflexion selon l'axe X. (g) après floutage par filtre gaussien.
Deuxième ligne : Quelques images complexes de la base de données.

la base de données publique Visual Object Classes Challenge 2012 (VOC2012) [89]. Les poids de ces modèles ont ensuite été affinés par un apprentissage sur notre base de données PAP.

2.6.2 Métriques d'évaluations

Afin de quantifier la performance d'une méthode de segmentation, il est d'usage de comparer le résultat obtenu avec la vérité-terrain produite par les experts à l'aide de quatre métriques : la précision, le rappel, le F1-score et l'IoU (Intersection over Union).

Ces métriques se basent sur les taux de vrai positif (TP ou True Positive en anglais), de faux positif (FP ou False Positive en anglais) et de faux négatif (FN ou False Negative en anglais). Le taux de vrai positif, représente le nombre de pixels de joint bien détectés par le modèle. Le taux de faux positif, représente le nombre de pixels détectés par le modèle comme appartenant aux joints mais qui en réalité appartiennent à la surface des pierres. Le taux de faux négatif, représente le nombre de pixels appartenant aux joints et qui ne sont pas détectés par le modèle.

La **precision** réflète la proportion des prédictions du modèle qui sont correctes : l'exactitude du modèle sur les joints détectés. Elle est déterminée par l'équation suivante :

$$P = \frac{TP}{TP + FP} \tag{2.1}$$

Le **rappel**, indique la sensibilité du modèle, c'est-à-dire la proportion des détections correctes sur l'ensemble des pixels à détecter. Il est défini par l'équation suivante :

$$R = \frac{TP}{TP + FN} \tag{2.2}$$

Le **F1-score** ou F-measure est la moyenne harmonique de la précision et du rappel. Le F1-score se définit comme suit :

$$F1 - score = 2 \times \frac{P \times R}{P + R} \tag{2.3}$$

L'**Intersection over Union** ou IoU est le rapport des surfaces communes entre prédictions et vérité-terrain sur l'ensemble des surfaces. L'IoU est calculé par :

$$IoU = \frac{TP}{TP + FP + FN} \tag{2.4}$$

2.6.3 Environnement de tests et implémentations

L'entrainement des réseaux s'est effectué sur une station de travail Dell Precision Tower fonctionnant sous le système d'exploitation Linux Ubuntu 20.04. Il est équipé d'un processeur CPU Intel core I7 de 3,4 GHz ; 4 coeurs et une mémoire vive de 32 Go. La station de travail comporte aussi un processeur GPU sur la carte graphique NVIDIA GeForce de 16 Go de mémoire. Le code source utilise Python 3.7, CUDA 11.2, cuDNN 8.1, Torch 1.9, Torchvision 0.10, entre autres.

Les tests ont été effectués sur un ordinateur portable Dell Precision 3551 équipé d'une carte graphique NVIDIA Quadro P5000 de 4 Go de mémoire graphique, d'une RAM de 16 Go et d'un processeur CPU Intel Core i5, 4 coeurs avec une fréquence du processeur de 2,6 GHz.

Les principaux paramètres d'apprentissage des réseaux ont été ajustés empiriquement. Les images d'entrée sont de taille 256×256 pixels. Le paramètre du taux d'apprentissage est fixé à $1e-02$. La fonction d'optimisation utilisée est l'algorithme de la descente de gradient stochastique (ou stochastic gradient descent, SGD) [90, 91] avec un momentum de $0,9$ et un batch size de 4 images. Le paramètre pour la régularisation L2 est $\lambda = 1e-04$. La régularisation est une contrainte appliquée à la fonction d'optimisation, qui renforce la minimisation de

l'erreur et contribue à une meilleure convergence de l'apprentissage. La régularisation L2 ($L2Reg$) utilise la norme L2 ou norme Euclidienne (voir équation 2.5).

$$L2Reg = \lambda \sum_{i=1}^{n} |w_i|^2 \tag{2.5}$$

Avec, w les poids du réseau et n le nombre total de poids à mettre à jour dans le réseau.

2.6.4 Comparaison des résultats

Les méthodes basées sur les algorithmes classiques de détection de contours et de seuillage présentent la particularité d'être très sensibles au bruit et aussi aux zones d'altérations existantes sur la surface des pierres. La figure 2.13 montre un exemple de zone d'altération sur la surface de la pierre de teinte très proche de celle des joints.

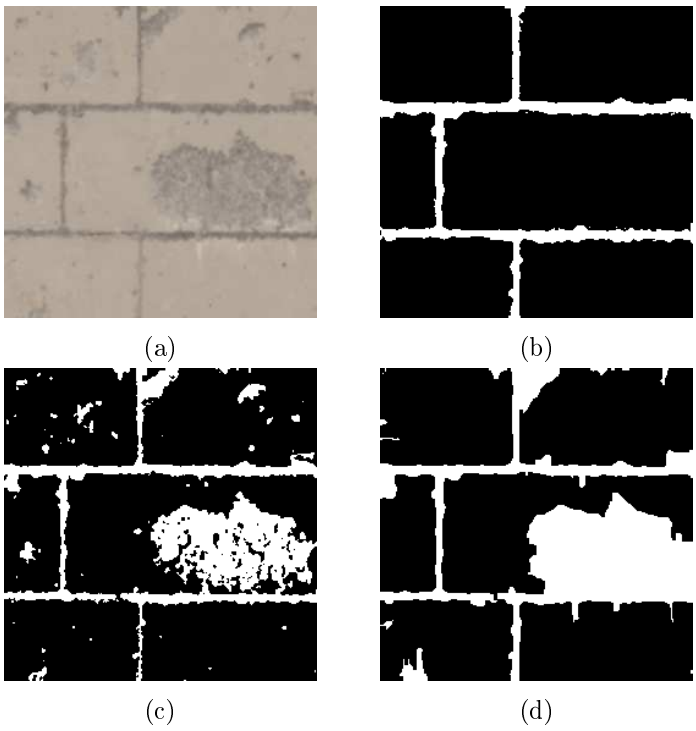


FIGURE 2.13 – Résultat de la segmentation sur image contenant une zone de forte altération. (a) Image ; (b) Vérité-terrain ; (c) Résultat de la segmentation par une méthode basée OTSU ; (d) Résultat de la segmentation par la méthode basée Canny et Sobel (section 2.4).

La méthode basée sur la transformée de Hough n'a pas été incluse dans ce comparatif car elle n'a pas produit des résultats probants : elle manque de précision

autour des joints et n'est pas adaptée aux ruptures de linéarités produites par les petites pierres incluses lors de restaurations précédentes ou les arcs que l'on trouve généralement autour des fenêtres et des portes. La méthode basée sur la transformée en ondelette continue n'est pas adaptée à l'uniformité entre pierres et joints caractéristique des châteaux de la renaissance étudiés.

Le tableau 2.1 présente une comparaison entre les différentes approches basées sur des réseaux de neurones : U-Net, SegNet et DeepLabv3+, et les algorithmes plus classiques de seuillage Canny et OTSU.

Tableau 2.1 – Comparaison des performances quantitatives des méthodes de segmentation pierre-à-pierre.

Methodes	Précision	Rappel	F1-score
Canny	0.535	0.308	0.390
OTSU	0.526	0.306	0.387
SegNet model	0.961	0.897	0.927
U-Net model	0.862	0.927	0.893
DeepLabv3+ model	**0.978**	**0.982**	**0.980**

De manière générale, il est observé que les méthodes basées sur des réseaux de neurones ont de bien meilleures perfomances que les approches classiques de seuillage, détection de contours et opérations morphologiques.

La segmentation par le réseau DeepLabv3+ permet d'obtenir les meilleures performances, avec un F1-score de 98 %, une précision de 98% et un rappel de 97%, sans doute grâce à la convolution à trous incluse dans son architecture et qui permet de réduire considérablement la sensibilité au bruit du réseau. Le réseau SegNet réalise une précision de 96%, un rappel de 89% er un F1-score de 92.7%. Le pourcentage de rappel plus bas de ce réseau est lié au fait qu'un nombre plus élévé de pixels appartenant aux joints n'ont pas pu être détectés. Les performances du réseau U-Net sont plus faibles que celles des précédents. Il obtient un F1-score de 89%, une précision de 86% et un rappel de 92%. La faible précision vient du fort taux de faux positifs. Elle montre la forte sensibilité du réseau aux fins détails dans l'image ce qui induit la détection de plusieurs pixels de surface de pierre comme appartenant aux joints.

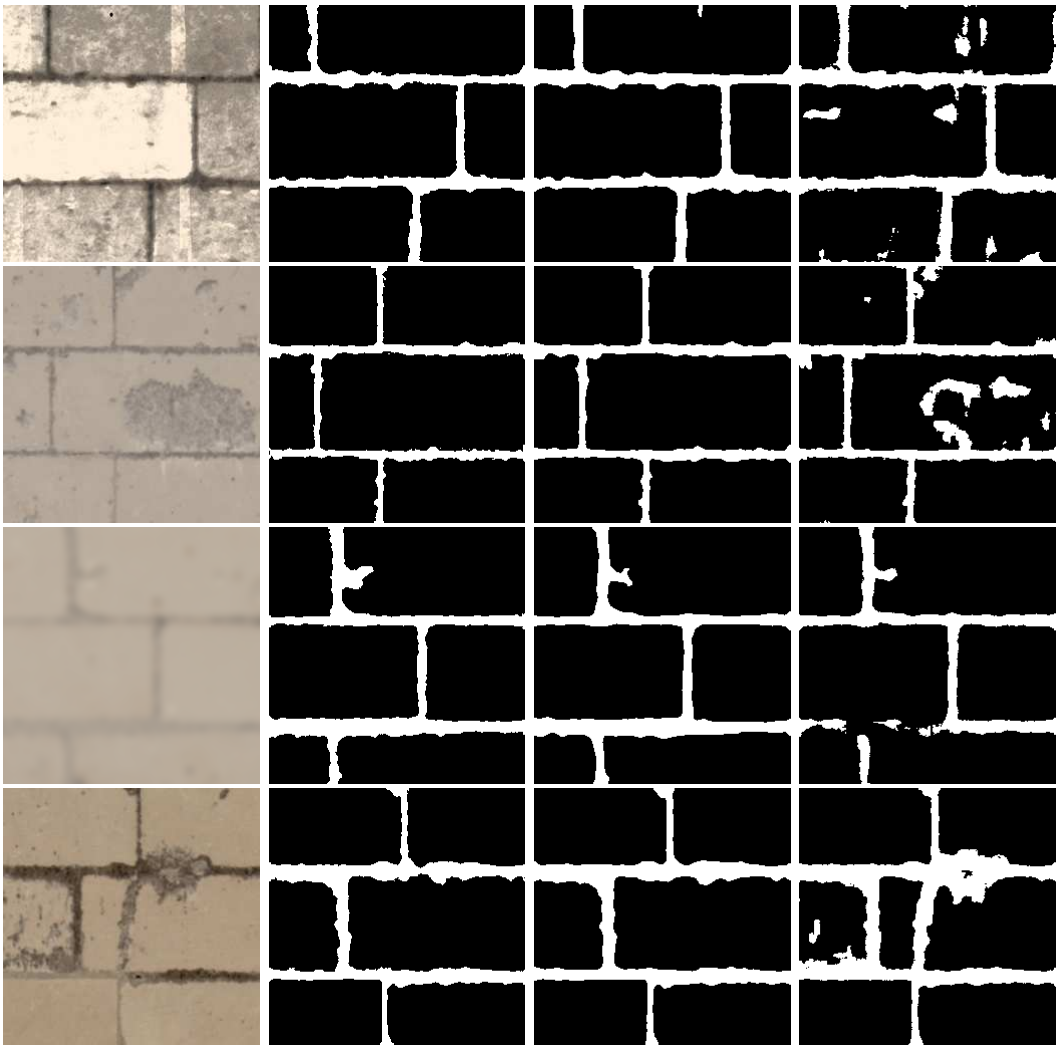


FIGURE 2.14 – Exemple de segmentation **Colonne 1 :** Images originales ; **Colonne 2 :** Vérités terrains ; **Colonne 3 :** résultats de segmentation avec le réseau DeepLabv3+ ; **Colonne 4 :** résultats de segmentation avec le réseau Segnet.

Pour la métrique d'IoU, le SegNet réalise un $IoU = 0.833$; le U-Net obtient un $IoU = 0.751$ et le DeepLabv3+ un $IoU = 0.916$.

Tableau 2.2 – Matrice de confusion de la segmentation avec le réseau SegNet

	joint	pierre
joint	0.896	0.104
pierre	0.036	0.964

Les Tables 2.2, 2.3 et 2.4, présentent les matrices de confusion obtenues sur l'ensemble des images de tests de la base de données. Pour le modèle SegNet, nous observons que seulement 3,6% des pixels faisant partie de la surface des

Tableau 2.3 – Matrice de confusion de la segmentation avec le réseau U-Net

	joint	pierre
joint	0.861	0.138
pierre	0.06	0.932

Tableau 2.4 – Matrice de confusion de la segmentation avec le réseau DeepLabv3+

	joint	pierre
joint	0.982	0.018
pierre	0.021	0.978

pierres ont été mal classés, et que 10,4% des pixels faisant partie des joints ont été mal classés. Les bruits dans l'image et le polissage des joints pour les rendre indiscernables peuvent être des facteurs qui expliquent ces erreurs. L'architecture du modèle peut également y contribuer car certains détails de l'image d'entrée sont perdus lors du traitement à travers les différentes couches de convolutions de la partie encodeur.

Le tableau 2.3, permet de constater que U-Net présente 13,8% de pixels mal détectés pour les joints et 6% de pixels mal détectés pour la surface des pierres. Ces forts taux confirment la sensibilité du réseaux aux détails sur la surfaces des pierres induisant des proximités entre les caractéristiques de certains pixels de la surface des pierres avec des pixels de joints.

Le tableau 2.4, confirme que DeepLabv3+ présente les meilleurs résultats avec seulement 2,1% de pixels mal détectés pour la surface des pierres et 1,8% de pixels mal détectés pour les joints. Les facteurs comme la robustesse au bruit dû à la convolution à trous, les couches de traitement multi-échelles de l'architecture du réseau DeepLab peuvent expliquer ses meilleures performances.

La capacité de généralisation de l'approche basée sur le réseau DeepLabv3+ a été testée sur des portions d'orthomosaïque d'un autre château de la vallée de la loire : celui de Chaumont-sur-Loire. La figure 2.15 montre les résultats. L'approche présente de bonnes capacités de généralisation pour réaliser une segmentation automatique pierre-à-pierre sur différents bâtiments du patrimoine culturel de style Renaissance.

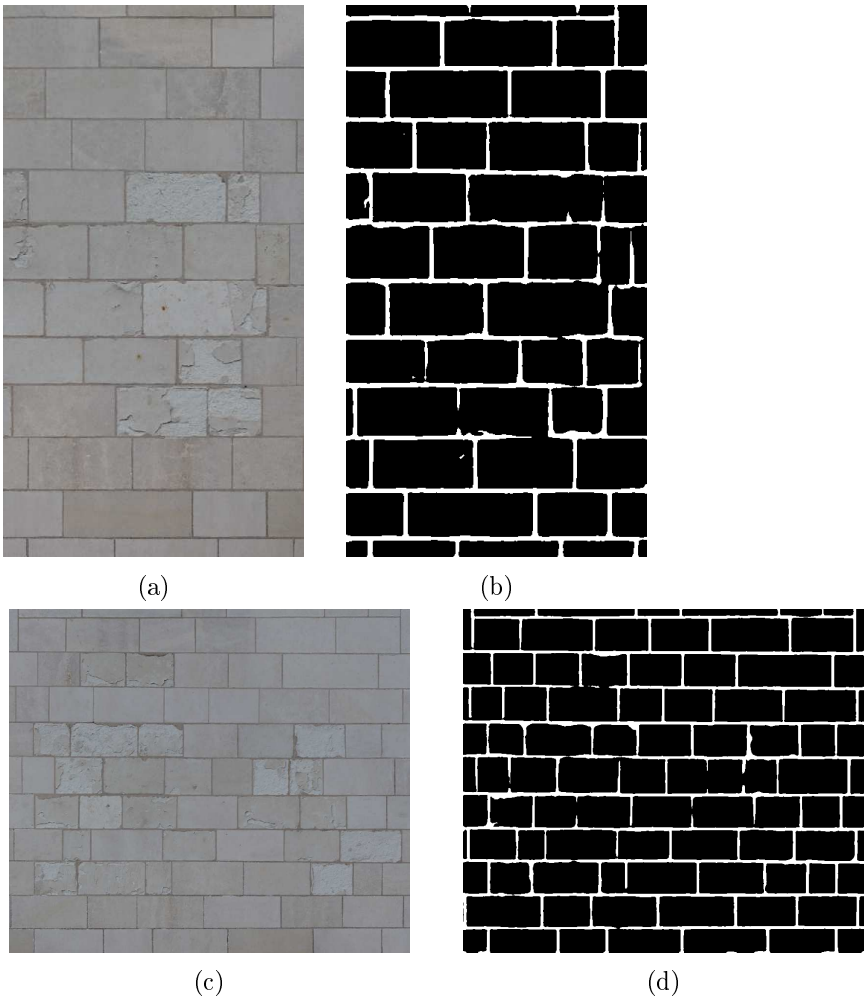


FIGURE 2.15 – (a) et (c) : Image du château de Chaumont-sur-Loire ; (b) et (d) : Résultats de segmentation avec le réseau DeepLabv3+.

2.7 Bilan du chapitre

Dans ce chapitre, nous avons étudié les principales approches de l'état de l'art. Elles sont de deux ordres : celles basées sur les algorithmes classiques de détection de contours, segmentation de région et seuillage : Canny, Sobel, transformée de Hough et transformée en ondelette continue. Puis celle basée sur le réseau de neurones U-Net pour la segmentation sémantique.

Pour améliorer l'existant, une nouvelle approche de segmentation pierre-à-pierre a été introduite à partir du réseau de neurones SegNet. Cette approche a démontré de meilleures performances. Cependant elle présente aussi quelques limites dues à la perte de détails en descendant dans les profondeurs des couches du réseau.

Ensuite, le réseau de neurones DeepLabv3+ a été exploité et permet d'obtenir les meilleures performances. Il tire sa robustesse des spécificités de son architecture comme les modules de convolution à trous, sa capacité résiduelle et l'usage d'un procédé d'extraction de caractéristiques multiéchelles dans l'image traitée.

L'étude comparative de ces approches démontre qu'elles améliorent notablement l'état de l'art pour la segmentation pierre-à-pierre sur les images des châteaux du style de la renaissance de la vallée de la Loire.

La mise en place de la base de données PAP utilisée pour l'entrainement et les tests des réseaux constitue aussi une contribution de cette étude. La vérité-terrain de cette base de données a été réalisée de façon semi-automatique à l'aide d'une première approche ad-hoc de segmentation pierre-à-pierre proposée. Les résultats de segmentation de cette approche ont ensuite été affinés par des experts en patrimoine pour obtenir la vérité-terrain. Ce procédé a permis de réduire la complexité de la phase d'annotation des orthomosaiques pour les experts.

Chapitre 3

Détection d'altérations des pierres

> Qu'on se le dise pourtant, les machines, si puissantes et si sophistiquées soient-elles, demeurent très spécialisées.
>
> Yann Le Cun

Sommaire

3.1 Introduction

Les orthomosaïques offrent une représentation cartographique plane suivant une échelle métrique (centimètre carré par pixel) uniforme de chaque façade des bâtiments historiques. C'est notamment pourquoi, elles représentent le support de prédilection utilisé par les experts pour documenter et surveiller l'état des bâtiments et monuments historiques. Ils procèdent à des annotations manuelles sur les images couleur, suite à des observations visuelles sur site. Le processus est détaillé dans le Chapitre 1. La figure 3.1 montre les altérations relevées manuellement par les experts sur l'orthomosaïque de la façade de la cour Ouest du château de Chaumont-sur-Loire. Quatre types d'altérations ont été relevés sur cette façade.

FIGURE 3.1 – Exemple d'annotations d'altérations réalisées manuellement par les experts en patrimoine : desquamation en plaque (annotée en rouge), desquamation en feuillet (annotée en orange), erosion (annotées en bleu clair) et pelage (annotées en rose clair).

Dans ce chapitre nous étudions la détection automatique des desquamations en plaque sur les images couleur. Elles sont aussi potentiellement la cause de décrochement de matière sources de risques pour les visiteurs. Ce type d'altération est le plus couramment observé sur les châteaux de la vallée de la loire. Le reste du chapitre est structuré comme suit. La section 3.2 présente les différentes approches de la littérature utilisant les réseaux de neurones profonds pour la détection d'altérations des pierres. Ces méthodes sont approfondies dans la section 3.3 qui présente leurs architectures. Ensuite, la section 3.4 développe la nouvelle architecture pro-

posée utilisant YOLO et les transformers. Dans la section 3.5, sont décrits les détails d'implémentations et les résultats de la méthode proposée et une comparaison à l'état de l'art. Enfin, la section 3.6 fait la synthèse de cette étude et fournit quelques perspectives.

3.2 État de l'art de la détection d'altération par traitement d'images

Les premiers travaux sur la détection des dégradations sur les images couleur de pierres sont basés sur des méthodes classiques de traitement d'images, comme le seuillage des histogrammes couleurs ou la segmentation 3D basée sur la texture de l'image, tenant compte des paramètres de luminance, des couleurs dominantes et de l'algorithme de Canny [36].

Par la suite, Valero et al. [43] ont proposé des techniques d'apprentissage machine pour une plus grande robustesse en joignant l'exploitation du modèle 3D à celle de l'image couleur du mur en moellons de la Façade sud de la chapelle royale de Stirling Castel (Écosse). Partant du nuage de points 3D issus d'acquisitions par un scanner laser terrestre (TLS) et de la photogrammétrie, ils effectuent une segmentation pierre-à-pierre comme présentée dans le chapitre précédent [31]. Puis, pour chaque pierre segmentée, des caractéristiques liées à la géométrie et la variation des couleurs sont extraites et utilisées pour entrainer un classifieur à l'identification des altérations.
La perte de matière est identifiée par le relief en supposant que la surface des pierres est relativement plane : toute région distante en profondeur de la surface est considérée comme une zone d'altération. De plus, des caractéristiques liées à l'homogénéité de couleur sont identifiées en supposant que les sous-régions distantes en couleur de la valeur moyenne de la couleur sur toute la pierre sont de probables zones d'altérations.
Enfin, les caractéristiques extraites sur l'ensemble des pierres sont utilisées en entrée d'un réseau de classification par régression qui détecte le type d'altérations : défauts géométriques (érosion, délitage, dommages mécaniques) et altérations chromatiques.

D'autres travaux ont par la suite utilisé des réseaux de neurones convolutifs (CNN) qui ont récemment démontré de grandes performances pour la détection d'objets dans les images. La détection d'objet consiste à prédire la classe et les coordonnées de la boîte englobante autour de chaque instance d'objet dans une image. Les réseaux de détections d'objets sont entraînés sur de grandes bases de données [92, 93, 94] d'images génériques de la vie quotidienne (personne, chat, chien, vélo, voiture, bateau, avion...). Ces réseaux peuvent être utilisés pour détecter des altérations des pierres sur une façade. Pour ce faire, il faut renforcer leur apprentissage par des entraînements supplémentaires sur une base de données spécifiques incluant des images d'altérations des pierres.

A ce titre, Wang et al. [95] ont utilisé le réseau de neurones Faster R-CNN exploitant ResNet101 comme extracteur de caractéristiques (features extractors ou backbone) pour détecter les desquamations en plaque et efflorescence sur les murs du Palace Museum, en Chine. Ce réseau a été entrainé sur une base de données de 500 images de 500×500 pixels et contenant 1484 instances d'altérations, sous la forme de boîtes englobantes, extraites de deux orthomosaïques.

Ces travaux représentent une amélioration d'une approche précédente où les auteurs ont utilisé AlexNet [82] et GoogleLeNet [83] pour classer des images de briques du mur de la Cité Interdite en quatre catégories : pierres non-altérées, desquamations en plaque, fissures et efflorescences [96]. Ces réseaux ont été entrainés sur une base de données de 5145 images de pierres découpées automatiquement par fenêtres glissantes (sliding windows), de tailles 480×105 pixels et 210×105 pixels, provenant d'une même orthomosaïque. Ainsi, ils supposent qu'une seule catégorie d'altération est présente par image.

Kwon et al. [97] ont également utilisé le réseau Faster R-CNN mais avec Inception comme extracteur de caractéristiques, un modèle moins performant que ResNet101 car il n'utilise pas les informations résiduelles. Ce réseau a été entrainé sur une base d'images issues du rapport périodique sur les biens culturels de 2017 réalisé par l'Administration du patrimoine culturel de Corée du Sud. Le but est de détecter quatre types d'altérations : fissure, perte, détachement, colonisation biologique. La base est composée de 400 images (100 images pour chaque type

d'altération).

Tableau 3.1 – Récapitulatif des principales références bibliographiques pour la détection d'altérations des pierres par traitements d'images

Auteurs	Données	Approches
Wang et al. 2018	Orthomosaïque provenant de la Cité Interdite en Chine, 5145 images de 480 × 105 pixels et 210 × 105 pixels. Quatre types d'altérations : pierres non-altérés, desquamations en plaque, fissures et efflorescences.	Détecte un seul type d'altération par classification d'images : AlexNet, GoogleLeNet.
Valero et al. 2019	Façade sud de la chapelle royale du Stirlling Castel (Ecosse). Nuage de points 3D issues d'acquisitions TLS et de photogrammétrie. Détecte les défauts géométriques (érosion, délitage, dommages mécaniques) et les altérations chromatiques	Segmentation pierre-à-pierre + caractéristiques géométriques et de couleur de la surface des pierres pour détecter les altérations par réseaux de classification par régression.
Wang et al. 2019	2 orthomosaïques provenant du Palace Museum en Chine, 500 images de 500 × 500 pixels contenant 1484 instances d'altérations. Deux types d'altérations : desquamations en plaque et efflorescence.	Faster R-CNN utilisant ResNet101 pour détecter les 2 types d'altérations.
Kwon et al. 2019	400 images issues du rapport périodique sur les biens culturels réalisé en 2017 par l'administration du patrimoine culturel en Coré du Sud ; 4 types d'altérations : fissure, perte, détachement, colonisation biologique.	Faster R-CNN utilisant InceptionV2

3.3 Méthodes d'apprentissage testées pour la détection d'altérations

Les réseaux de détection d'objets entrainés sur de grandes bases d'images génériques [93, 92] ont pour but de prédire une boîte englobante au plus près de chaque instance d'objet présent dans une image et de donner la classe correspondante. Tandis que les réseaux de segmentation d'instances prédisent, en plus, le masque correspondant au découpage plus précis de l'objet au sein de la boite englobante.

Outre Faster R-CNN testé par Wang [95] et Kwon [97] pour la détection d'altérations sur les maçonneries, d'autres réseaux peuvent être adaptés à la détection d'altérations de desquamation sur les pierres calcaires. Les sections suivantes présentent plus en détails les approches testées dans cette étude.

3.3.1 Faster R-CNN

Faster R-CNN est un réseau de détection d'objets introduit en 2015, qui a remporté la 1ère place au Microsoft COCO Challenge [93]. C'est une version améliorée de Fast R-CNN [98] où l'algorithme de recherche sélective [99] est remplacé par un plus performant proposant les régions d'intérêt (le Region Proposal Network, RPN) [12]. Faster R-CNN a une architecture de détection à deux étapes. D'une part, les cartes de caractéristiques issues des couches convolutionnelles sont introduites dans un RPN, qui prédit plusieurs régions d'intérêt (RoI) et un score de probabilité que chacune des régions contienne un objet. D'autre part, un classifieur détecte la classe correspondant aux objets présents dans chaque région d'intérêt ayant le score de probabilité le plus élevé (figure 3.2).

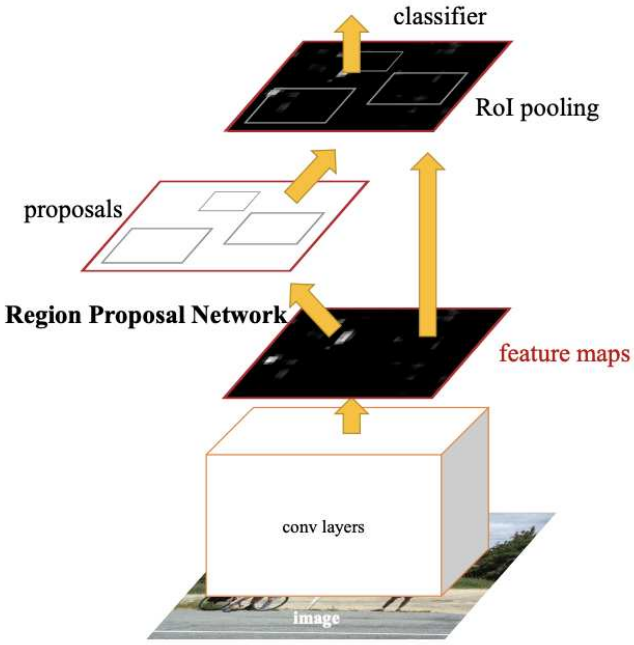


FIGURE 3.2 – Architecture du réseau Faster R-CNN [12].

Pour produire les cartes de caractéristiques à partir de l'images en entrée,

plusieurs réseaux de convolutions peuvent être utilisés, par exemple ResNet [13], Inception [100, 101] ou encore Mobilenet.

ResNet a été introduit pour la classification d'images [13] et a remporté le concours ImageNet Large Scale Visual Recognition Challenge (ILSVRC) en 2015. Le réseau a été proposé pour résoudre le problème de gradient évanescent. En effet, lorsque les réseaux de neurones convolutifs deviennent très profonds, le gradient rétro-propagé devient de plus en plus petit, ce qui empêche la mise à jour proportionnelle des poids du réseau [102]. Pour pallier ce problème, ResNet utilise un bloc résiduel (figure 3.3) qui ajoute une connexion raccourcie qui saute une ou plusieurs couches, aux sorties des couches de convolutions obtenant ainsi une carte de caractéristiques résiduelles $F(x) + x$.

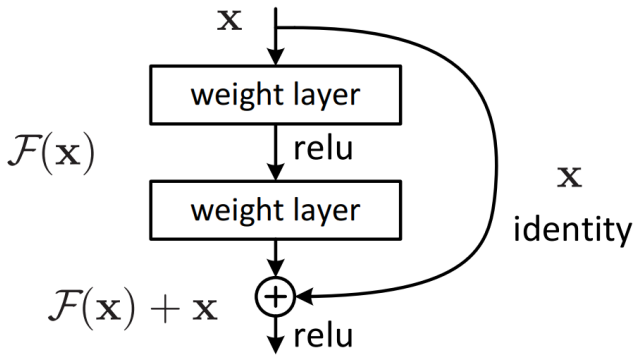


FIGURE 3.3 – Architecture du module résiduel de ResNet [13].

Mobilenet est un réseau de classification d'images introduit pour réduire la taille du modèle et le coût de calcul tout en maintenant les performances. Il est constitué de couches de blocs résiduels inversés et de couches de convolutions séparables en profondeur (voir Figure 3.4) [14]. MobileNet est conçu pour être utilisé dans un environnement aux ressources informatiques limitées et peut être utilisé comme extracteur de caractéristiques dans un réseau de détection d'objets embarqué, par exemple.

Il existe d'autres architectures de réseaux de neurones profonds qui pourraient aussi convenir à la détection d'altérations des pierres comme Mask R-CNN et YOLO.

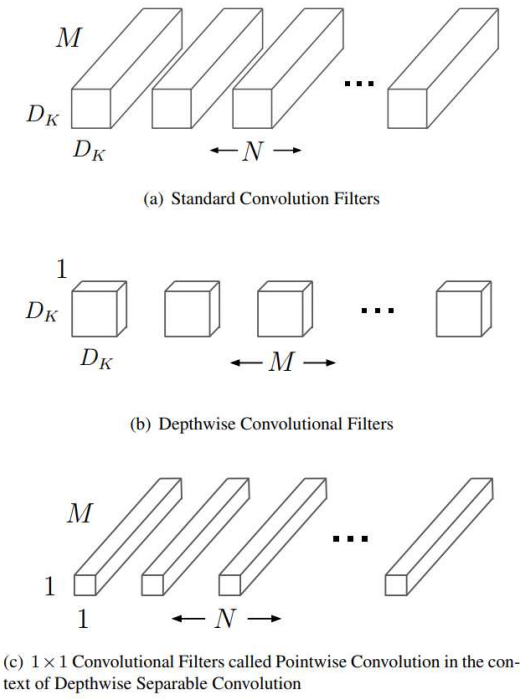


FIGURE 3.4 – Architecture module de convolution séparable en profondeur [14].

3.3.2 Mask R-CNN

Mask R-CNN a été introduit en 2017 et s'inscrit dans la suite des réseaux convolutifs basés sur la région (R-CNN) [103]. L'architecture du réseau (figure 3.5) est une extension de Faster R-CNN (section 3.3.1) pour détecter les boîtes englobantes auquel il rajoute une série de couches de convolutions [104] qui détectent un masque de segmentation sémantique à l'intérieur de chaque boîte englobante.

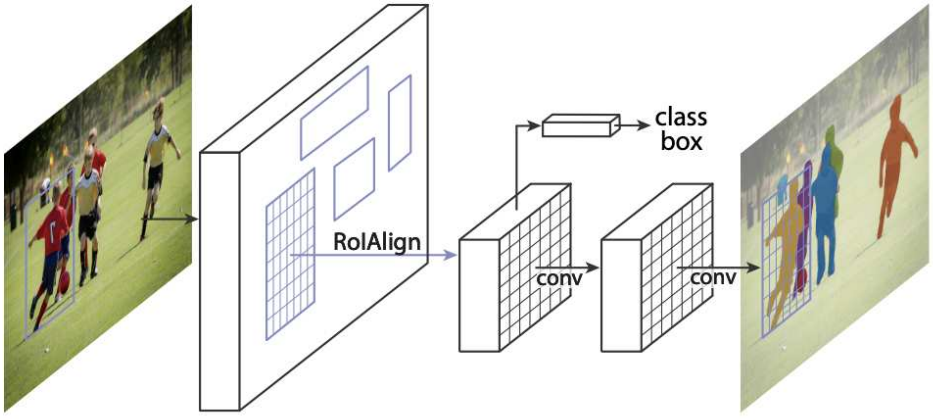


FIGURE 3.5 – Architecture du réseau Mask R-CNN [15].

3.3.3 YOLO

Le réseau de détection d'objets YOLO a été développé en 2016 [105]. C'est la première architecture qui détecte les boîtes englobantes et prédit les classes des objets dans une image en une seule étape. YOLO existe en plusieurs versions [106, 107, 108, 109]. La version YOLOv5 [106], la plus récente est constituée de quatre parties principales (figure 3.6) : le backbone, le neck, le head et le detect. Le backbone utilise le réseau CSP bottleneck (cross stage partial) à trois convolutions [110], pour extraire et agréger les caractéristiques pertinentes à différentes granularités ou échelles. Les CSP sont des réseaux très profonds basés sur DenseNet [111] et qui relient des couches pour atténuer le problème d'évanescence du gradient et renforcer la propagation des caractéristiques, encourager le réseau à réutiliser les caractéristiques et réduire le nombre de paramètres du réseau. Le neck utilise un réseau pyramidal de caractéristiques [112, 113]. Puis, le head où les matrices des caractéristiques du backbone sont traitées pour, enfin, dans le detect servir à la prédiction des coordonnées de la boîte englobante, la classe et le score de confiance de la détection de chaque objet dans l'image.

Comme Faster R-CNN, YOLO est un réseau de détection d'objets basé sur les ancres. Les ancres sont les boîtes englobantes initiales de tailles prédéfinies. Elles sont générées en très grand nombre puis réduites progressivement pour obtenir les boîtes englobantes au plus près des objets. Ainsi, dans son processus d'entrainement, le réseau générera pour chaque image en entrée, des milliers d'ancres de tailles multiples, parmi lesquelles seront retenues celles qui ont une forte probabilité de contenir un objet [114].

Le processus de génération des ancres (ou boîtes englobantes initiales) est généralement configuré en définissant les paramètres de taille et de ratio des ancres par défaut sous forme de liste en début du processus d'entrainement. Afin de s'adapter à une base de données spécifique, les dernières versions de YOLO permettent de faire un apprentissage automatique de ces paramètres par l'algorithme de k-means [115, 116]. Plus la taille de l'ancre s'adapte à la variabilité des boîtes englobantes présentes dans l'ensemble des images de la base de données, plus le réseau est précis dans la détection et mieux il apprend. Il s'agit de déterminer par regroupement

la taille (hauteur et largeur) et de ratio de l'ensemble des boîtes englobantes dans l'ensemble des données. Ceci est très important pour les tâches personnalisées, car la distribution des tailles et des emplacements des boîtes englobantes peut être très différente de celle des boîtes englobantes prédéfinies dans le jeu de données COCO. Des travaux récents démontrent que ce procédé permet d'améliorer les capacités d'entrainement des réseaux de détection d'objets [117, 118].

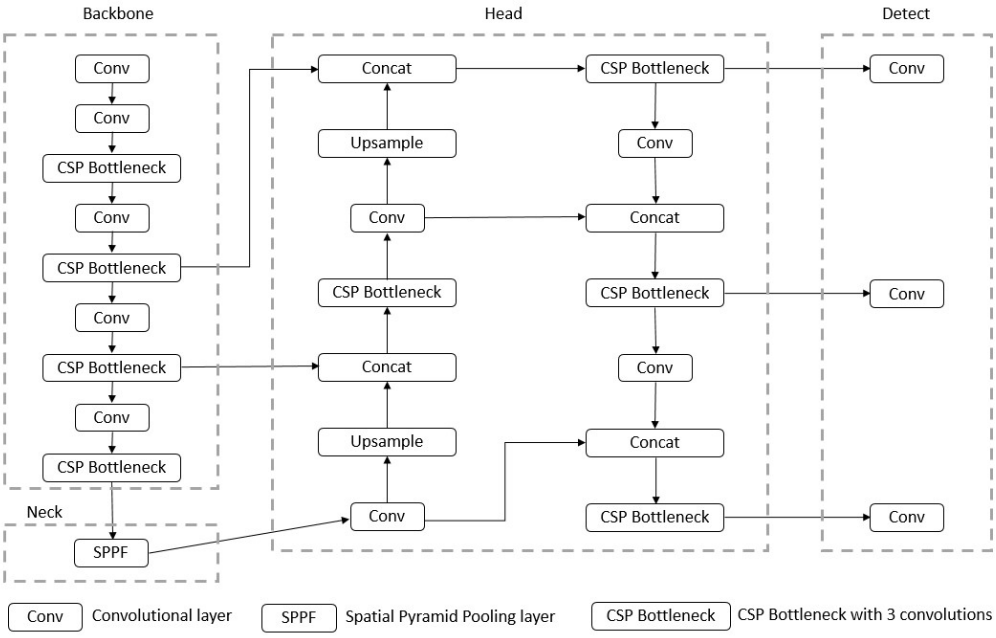


FIGURE 3.6 – Architecture du réseau YOLOv5.

Pour se comparer à l'état de l'art pour la détection des altérations des pierres, nous avons expérimenté Faster R-CNN sur les murs extérieurs des châteaux de style Renaissance dans la vallée de la Loire. Pour améliorer les résultats, nous proposons une nouvelle architecture de réseau de neurones profond combinant le réseau YOLO et les transformers.

Les transformers rendent les réseaux neuronaux plus puissants et plus performants par l'introduction de blocs d'auto-attention qui supplantent les convolutions et prennent en compte des informations contextuelles [119, 120, 121]. Les transformers ont été introduits pour la compréhension automatique du langage afin d'améliorer l'interprétation d'un mot dans une phrase [16] ; chaque mot pouvant avoir plus d'un sens selon le contexte et son emplacement dans la phrase. Ils ont

supplanté les réseaux neuronaux récurrents qui traitent l'ordre des mots dans une phrase. Le modèle de transformer, en revanche, n'utilise pas de réseaux récurrents mais un mécanisme d'attention, qui fonctionne en comparant chaque mot de la phrase à tous les autres mots de la phrase, y compris lui-même (auto-attention), et en pondérant sa pertinence contextuelle pour relativiser sa signification dans la phrase. Ce mécanisme permet au réseau neuronal de concentrer son attention sur une ou plusieurs entrées particulières et d'ignorer les autres. Le mécanisme d'attention et son application aux images seront plus détaillés dans la section suivante.

3.4 Architecture proposée pour la détection d'altérations avec YOLO et transformers

D'autres auteurs [122, 123] ont récemment proposés de nouvelles architectures de réseau combinant aussi YOLO et transformers.
Zhu et al. [123] ont introduit, dans le cadre du challenge VisDrone 2021, le réseau TPH-YOLO. Dans ce réseau les transformers sont intégrés au niveau du neck et du head. Dans le neck, ils servent à rallonger le nombre de couches par rapport au head dans l'architecture de base de YOLOv5. Ceci augmente le coût du calcul et la mémoire, mais permet de rendre le réseau plus robuste aux très petits objets fortement présents dans les images de drone. Dans le head, les transformers remplacent totalement les couches de convolutions. Zhang et al. [122] ont aussi proposé, en 2021, le reseau ViT-YOLO dans lequel des transformers sont combinés aux modules de convolution CSP au niveau du backbone.

Nous avons introduit les transformers [119] au niveau du head de l'architecture de base de YOLOv5 en remplacement des réseaux CSP. L'architecture, présentée dans la figure 3.7, utilise le backbone de YOLOv5. Le head, est composé de convolutions 3×3, suivies d'un sur-échantillonnage ou up-sampling et d'une concaténation à la carte de caractéristiques de résolution identique issue du CSP du backbone. Le feature map de convolution ainsi obtenu est mis en entrée d'un module transformer.

Le mécanisme d'attention d'un transformer est illustré sur la figure 3.8. La transposition des modèles de transformer au traitement des images se fait au tra-

vers d'une grille régulière où chaque patch joue le rôle des mots dans le traitement du langage. Une valeur d'attention correspondant à une requête Q (query) et une paire clé-valeur K-V (Key-Value) est calculée par des opérations d'algèbre linéaire et un jeu de poids W affinés lors de l'apprentissage du réseau.

Pour une entrée X, qui correspond à un patch de la carte des caractérisques, on calcule :

$$\begin{aligned} Attention(Q, K, V) &= softmax(QK^T) \times V \\ &avec \\ Q &= XW_q \\ K &= XW_k \\ V &= XW_v \end{aligned} \tag{3.1}$$

Où $W_{q,k,v}$ sont les matrices de poids. La requête Q est comparée à toutes les clés possibles K de l'ensemble des données au moyen d'un produit scalaire. Q et K doivent donc avoir la même dimension. Ces scores d'alignement sont normalisés par une fonction $softmax$, puis la sortie est calculée par une somme pondérée. La dimension de V est donc celle de la sortie souhaitée. En ajustant les pondérations W, le système peut apprendre la zone des entrées sur laquelle se concentrer.

La version d'attention à tête multiples (multi-head) consiste à utiliser plusieurs modules d'attention en parallèle qui partagent la même entrée (query) mais ont des poids propres ; de sorte que les n valeurs d'attention obtenues en parallèle peuvent être concaténées. Cela permet différentes projections linéaires de la même entrée.

Dans l'architecture proposée HBSpall-TransYOLO, l'intégration des transformers est faite au niveau du head, après les couches de convolution du backbone. Cela permet de traiter la carte des caractérisques issue des opérations de convolution du backbone et de réduire la dimensionnalité. En conséquence, pour atteindre de bonnes performances, notre architecture ne nécessite pas de larges base d'apprentissage pour entrainer les transformers, contrairement à celles où les transformers sont directement appliqués aux données.

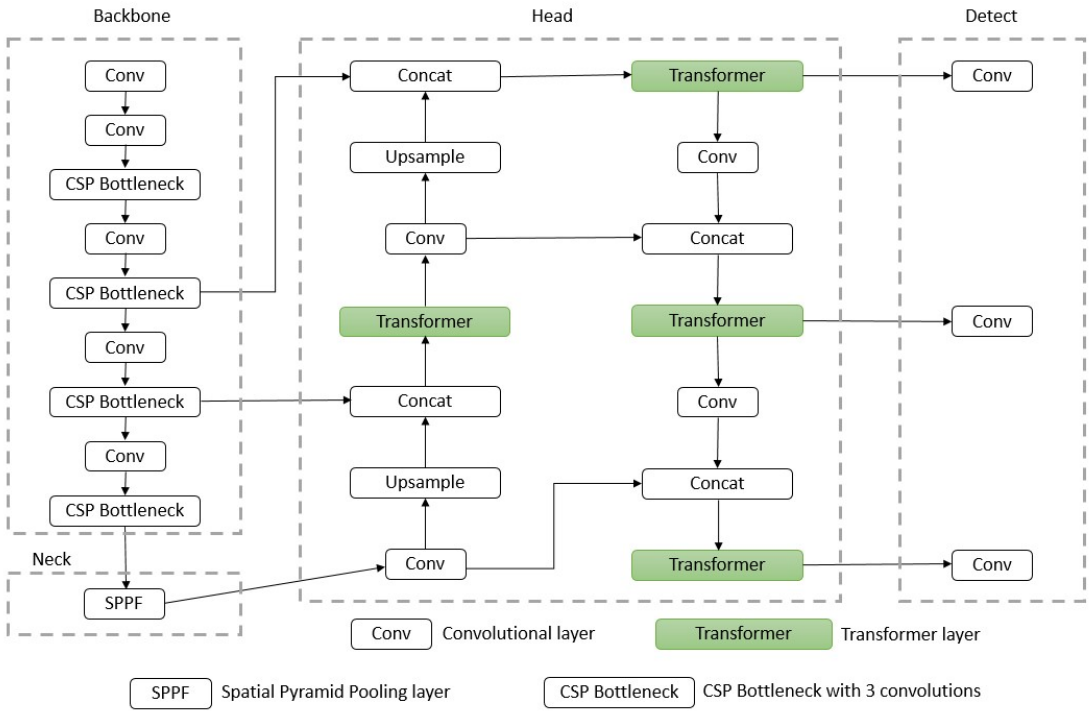


FIGURE 3.7 – Architecture du réseau de détection d'altérations avec YOLO et transformers.

3.4.1 Suppression des chevauchements (filtrage des boîtes englobantes)

Comme évoqué précédemment, une grande quantité de boîtes englobantes sont générées au cours du processus de détection, elles sont souvent redondantes. Lorsque l'on considère que les objets détectés ne peuvent pas se chevaucher, cela signifie qu'en cas de détections multiples, une seule doit être considérée comme correcte et que les autres sont fausses. On peut, dans ce cas, appliquer l'approche de la suppression non maximale (NMS en anglais pour Non-Maximum Suppression) [114] : seule la boîte englobante la plus pertinente est conservée, au sens d'un chevauchement optimal. Le critère de chevauchement utilisé est généralement l'index de Jaccard ou l'intersection sur l'union (intersection over union, IoU) entre les boîtes englobantes. La suppression se fait en fonction d'un seuil d'IoU prédéfini, qui peut affecter considérablement les performances d'apprentissage du réseau. Des travaux récents proposent, plusieurs variantes du NMS dont le GIoU-NMS [124], DIoU-NMS [125], ou le Soft-NMS [126]. Le réseau HBSpall-TransYOLO proposé utilise

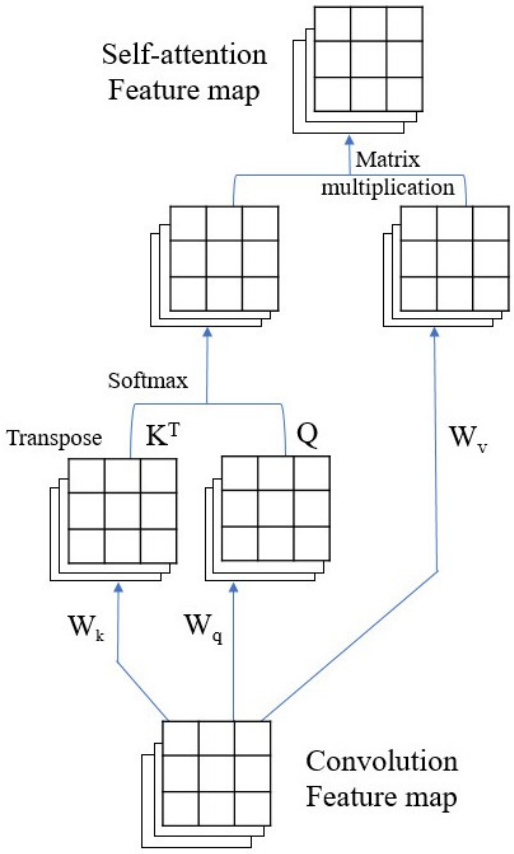


FIGURE 3.8 – Architecture of a self-attention building block [16].

le DIoU-NMS qui prend en compte le critère de chevauchement, mais également la distance entre les centres des boîtes englobantes. Cette approche permet au réseau de converger plus rapidement qu'avec un NMS classique ou d'autres variantes.

3.4.2 Fonction de perte

La fonction de perte évalue les erreurs entre la prédiction et la vérité terrain pendant le processus d'entrainement du réseau. Le réseau proposé utilise une activation sigmoïde suivie d'une perte d'entropie binaire croisée. La fonction d'activation sigmoïde ajuste la sortie de prédiction du réseau dans la plage $[0, 1]$. La fonction de perte d'entropie binaire croisée (ou binary cross-entropy) compare le score de probabilité de la prédiction du réseau p ($p \in [0, 1]$) avec la vérité terrain y suivant la formule 3.2.

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} -(y_i log(p_i) + (1 - y_i) \times log(1 - p_i)) \tag{3.2}$$

3.5 Implémentation et résultats

Les sections suivantes précisent les détails d'implémentation et les réglages du réseau proposé avant de présenter l'évolution de ses performances et une comparaison aux différents approches existantes.

3.5.1 Base de données BD-Altérations

La base de données utilisée pour la détection des altérations a été constituée à partir des images couleur provenant des murs extérieurs du château de Chaumont-sur-Loire (figure 3.9). Ces images couleur sont issues d'un ensemble d'images préalablement acquises pour produire des modèles 3D de l'ensemble du château et non spécifiquement acquises pour la détection d'altérations. Ce qui implique qu'elles ne bénéficient pas de cadrage spécifique sur les zones d'altérations, ni zoom ou d'attentions particulières visant à mieux reconnaitre les caractéristiques des differents types d'altérations.

La vérité terrain est obtenue par la labellisation manuelle effectuée par les experts sur les orthomosaïques ; cette labellisation est ensuite reprojetée sur les images couleurs d'origine de plus haute résolution, acquises pour la conception des orthomosaïques (plus de détails dans le chapitre 1). Nous avons fait le choix de travailler sur les images couleur car elles présentent une résolution bien meilleure que celle des orthomosaïques disponibles.

Il convient de noter qu'aucune technique classique d'augmentation de la base d'images, n'a été effectuée pour les phases d'entraînement (rotation, variation de contrastes, etc.) [127, 29] car une grande redondance existe de fait, de par la contrainte d'un chevauchement des images de plus de 60% pour créer le modèle 3D.

Une seule classe est considérée dans cette étude : la desquamation en plaques. La base est constituée de 1012 imagettes de taille 256×256 pixels. Le tableau 3.2 présente la distribution des zones de desquamation en plaques dans la base de

FIGURE 3.9 – Orthomosaïque de la façade intérieure Est du château de Chaumont-sur-Loire.

Tableau 3.2 – Distribution des zones de desquamation en plaques dans la base de données BD-Altérations

Base de données	Imagettes	Zones de desquamation en plaques
Train set	759	2859
Test set	253	1096
Total	1012	3955

données. Au total, 3955 zones de desquamation en plaques ont été labéllisées sur les 1012 images de la base de données. L'ensemble d'apprentissage représente 75% de la base de données ; soit 759 images contenant 2859 zones de desquamation en plaques. L'ensemble de test quand à lui, correspond à 25% de la base de données ; soit 253 images contenant 1096 zones de desquamations en plaques.

3.5.2 Détails d'implémentation

Le réseau YOLOv5 est préalablement entraîné sur la base de données COCO de Microsoft [93]. Nous affinons ensuite les poids obtenus en poursuivant l'apprentissage sur notre base de données [128]. Quand au réseau de détection d'altérations

avec YOLO et transformers proposé, une correspondance des couches a été appliquée. Ainsi, les poids obtenus de l'apprentissage préalable sur la base de données COCO sont maintenus dans les couches correspondantes du nouveau réseau et les couches de transformers sont entrainées à partir d'une initialisation automatique avec trois itérations de réchauffement (warm-up) et un momentum initial de 0,8. Il s'agit, de démarrer l'apprentissage du réseau avec une valeur de momentum plus petite qui sera incrémentée progressivement jusqu'à atteindre la valeur de momentum paramètrée. Cette incrémentation progressive se fait sur un nombre d'itérations défine dénomée les itérations de warm-up. Cette technique permet d'éviter que les gradients ne divergent trop fortement en début d'apprentissage, ce qui aide à réduire les instabilités et contribue à une meilleure convergence [129, 130].

La phase d'apprentissage des réseaux s'est effectuée sur le serveur CaSciModOT de la Région Centre-Val de Loire. C'est une station de travail fonctionnant avec le système d'exploitation OpenSUSE Leap 15.2. Le nœud GPU Intel dispose d'un processeur Intel Xeon Gold 6248 de 2.5 GHz avec quatre (04) cartes graphiques NVIDIA Tesla V100 de 32 GB de mémoire chacune. Le nœud contient aussi 40 cœurs de CPU avec 192 Go de RAM. L'environnement du code source est python 3.7, CUDA 11.2, cuDNN 8.1, Torch 1.9, Torchvision 0.10, entre autres.

Les tests quand à eux ont été effectués sur un ordinateur portable avec 16 Go de mémoire graphique, 64 Go de RAM, un CPU Xeon E5-2620 v4 et une seule NVIDIA Quadro P5000. Ainsi, nous avons pu observer et confirmer la capacité des réseaux proposés à être exploité par des experts ou tout autre acteur du patrimoine culturel sans avoir besoin d'une puissance de calcul particulièrement élevée.

Les hyperparamètres d'apprentissage ont été ajustés empiriquement. La taille des images en entrées est de 256×256 pixels. Le paramètre du taux d'apprentissage est fixé à $0,01$. Nous utilisons l'algorithme de la descente de gradient stochastique (ou stochastic gradient descent, SGD) [90, 91] avec un momentum de 0,937, un weight decay de 0,0005, un batch size de 12 images et nous utilisons le taux d'apprentissage à un cycle (ou one cycle learning rate) [131].

Le taux d'apprentissage cyclique, est une technique de paramétrage des réseaux de neurones qui sert à ajuster le taux d'apprentissage de façon optimale tout au

long des itérations, en l'augmentant progressivement à partir d'une valeur initiale jusqu'à un maximum, puis en le dimunuant jusqu'à sa valeur initiale. Ce cycle est répété tout au long de l'apprentissage. Ce procédé optimise considérablement la convergeance du réseau [132].

3.5.3 Métriques d'évaluations

Les métriques choisies pour évaluer les algorithmes de détections et de segmentation d'instances sont : la précision moyenne (ou mean average precision, mAP) [92] et la courbe de Precision-Rappel.

3.5.3.1 Précision moyenne (mAP)

La Précision moyenne (mAP) est calculée pour un seuil spécifique de chevauchement avec la vérité-terrain. Le même indice d'IoU décrit en section 3.4.1 est utilisé ici entre les boîtes englobantes prédites par le réseau et les boîtes englobantes de la vérité-terrain. Il varie de $[0, 1]$. Ainsi, un modèle dont toutes les prédictions sont fausses a une $mAP = 0$ et à l'inverse, un modèle dont toutes les prédictions sont correctes a une $mAP = 1$.

La précision est définie suivant la formule $P = TP/(TP + FP)$. Avec :

— TP, True Positive ou Vrai Positif, le nombre de boîtes englobantes détectées avec $IoU \geq seuil$.

— FP, False Positive ou Faux Positif, le nombre de boîtes englobantes détectés avec $IoU < seuil$.

— FN, False Negative ou Faux Négatif, le nombre de boîtes englobantes présentes dans la vérité terrain mais non détectées.

La valeur seuil de l'IoU généralement utilisée dans la littérature est fixée à $0,5$. La précision moyenne peut aussi être évaluée en faisant varier le seuil sur un intervalle $0,50 : 0,05 : 0,95$.

3.5.3.2 Courbe Précision-Rappel

Pendant la phase d'exécution des reseaux entrainés, il arrive fréquemment d'obtenir pour une même boîte englobante de la vérité terrain, plusieurs boîtes englobantes prédites. Dans ce cas, un seuil d'Intersection Over Union (IoU) est défini.

Toutes les boîtes englobantes possédant une valeur $IoU \geq seuil$ sont considérées vrai positif et toutes les autres ($IoU < seuil$) sont considérées faux positif. Le choix de ce seuil s'avère donc délicat et peut affecter les performances en précision et rappel du modèle.

La précision représente en détection d'objet, la proportion des boîtes englobantes prédites qui est correcte (avec $IoU > seuil$). Elle reflète l'exactitude du modèle à détecter des boîtes englobantes très proches de celles de la vérité-terrain.

Le rappel exprime la proportion de boîtes englobantes de la vérité-terrain qui ont été correctement prédites. Il indique la capacité du modèle à détecter les réelles altérations présente dans la vérité-terrain.

La courbe de précision-rappel, est obtenue en faisant varier la valeur du seuil de chevauchement IoU considéré sur l'intervalle $[0, 1]$ et calculer pour chaque valeur de seuil la précision et le rappel. La paire $(precision, rappel)$ obtenue pour chacune des valeurs de seuil dans l'intervalle, représente un point de la courbe avec en abscisse, le rappel et en ordonnée, la précision. La courbe idéale doit s'approcher des valeurs unitaires pour la précision et le rappel.

3.5.4 Réglages et évaluation du réseau proposé

Pour atteindre les meilleures performances, plusieurs combinaisons de paramètres d'apprentissage ont été testées. Le tableau 3.3 présente quelques résultats en fonction du nombre d'itérations. Après 300 itérations, le réseau n'atteint pas sa précision optimale. Cela peut s'expliquer par le fait qu'une partie des couches de l'architecture du réseau (les transformeurs) est entrainée à partir d'une initialisation aléatoire des poids. Au delà de 1200 itérations, aucune amélioration significative des performances n'a été observée. Le réseau entrainé à 1200 itérations réalisant un mAP de $0,81$ a donc été retenu dans la suite des tests.

La figure 3.11 présente quelques exemples de détections des desquamations en plaques par le réseau proposé. On observe que le réseau est robuste aux variations de luminosité présentes sur certaines images de la base de données, ainsi que dans certains cas plus rares où les images ont été prises avec un mur partiellement ombragé.

Tableau 3.3 – Performance du réseau HBSpall-TransYOLO pour différents nombres d'itérations

Itérations	Boîtes englobantes mAP at IoU=0,50	Temps d'Inference (minute :seconde)
300	0,77	00 :35
1200	**0,81**	00 :34
2100	0,79	00 :31
3000	0,77	00 :31
5000	0,77	00 :32

Cependant, le réseau est moins performant dans le cas d'images complexes, comprenant une multitude de petites zones de desquamation en plaques. Les images (g), (h) et (i) de la figure 3.11, présentent des exemples de prédictions complexes du réseau. Notons que certaines zones de desquamation en plaques de taille réduite ne sont pas détectées par le réseau. Afin d'évaluer ce problème, nous proposons dans la section 3.5.6, une étude plus détaillée de l'évolution de la précision moyenne de la détection, en fonction de la taille des zones de desquamation.

La figure 3.10 montre la courbe de précision-rappel du réseau.

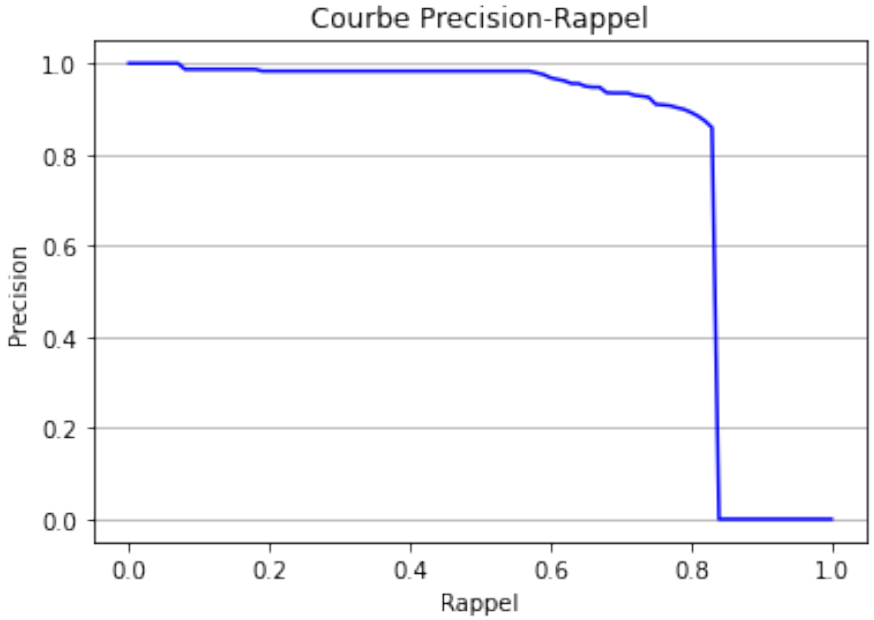


FIGURE 3.10 – Courbe Précision-Rappel du réseau HBSpall-TransYOLO pour 1200 epoch.

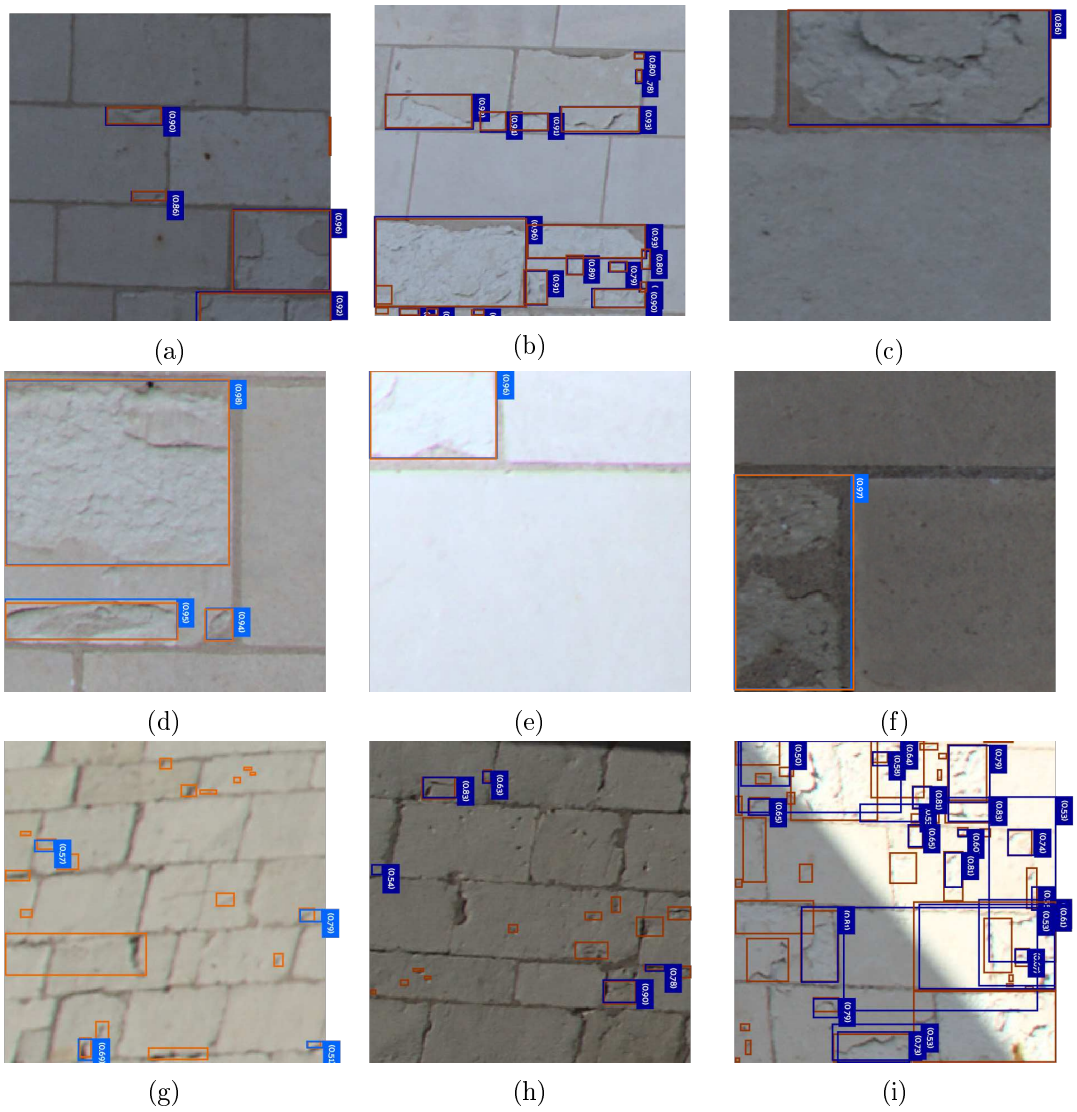


FIGURE 3.11 – Exemples de détections de desquamations en plaque réalisées avec le réseau HBSpall-TransYOLO. Les boîtes englobantes de la vérité-terrain sont en rouge et les boîtes englobantes prédites avec les scores de confiance respectifs sont en bleu. (a), (b), (c), (d), (e) et (f) : Exemples de bonnes détections ; (g),(h) et (i) : Exemple de détections plus complexes.

3.5.5 Comparaison à l'état de l'art

Pour mieux évaluer la méthode proposée, nous comparons ses performances avec les méthodes existantes dans la littérature utilisant l'apprentissage profond pour la détection des altérations dans les images couleur, à savoir : Faster R-CNN avec Resnet101 [97]. La méthode utilisant Faster R-CNN avec Inception [97] n'a pas abouti sur notre base de données, obtenant des précisions très faibles $mAP \leq 0.128$. Les auteurs n'ayant pas mis à disposition les codes source pour

cette approche, il est difficile de réaliser une comparaison objective.

Nous avons élargi la comparaison en expérimentant Faster R-CNN avec d'autres backbones largement adoptés en détection d'objets : Resnet50-FPN, Resnet50 et Mobilenet. Notre approche est également comparée à l'architecture classique de YOLOv5 pour évaluer l'apport des transformers introduits. Tous les réseaux sont testés sur la même base de test après une phase d'apprentissage sur un sous-ensemble d'images de notre base de données BD-Altérations. Les résultats sont résumés dans le tableau 3.4.

Tableau 3.4 – Comparaison des réseaux pour la détection des desquamations en plaques.

Réseau	F1-score	Boîtes englobantes mAP, IoU 0,50	Temps d'Inférence (minute :seconde)
HBSpall-TransYOLO	**0,85**	**0,81**	00 :34
YOLOv5	0,83	0,73	**00 :24**
Faster R-CNN & Resnet101[97]	0,77	0,73	01 :51
Faster R-CNN & Resnet50-FPN	0,77	0,76	01 :17
Faster R-CNN & Resnet50	0,68	0,61	01 :37
Faster R-CNN & Mobilenet	0,72	0,70	00 :51

Les F1-scores des réseaux basés sur Faster-R-CNN restent inférieurs à 80%. En observant les performances des réseaux avec Resnet50 et Resnet50-FPN, on peut noter que l'introduction d'un Feature Pyramid Network (FPN) améliore nettement les résultats. Notamment, un réseau moins profond Resnet50 avec FPN obtient des performances similaires à ceux de [97] avec Resnet101, plus profond. De manière générale, les réseaux basés sur YOLO ont permis d'obtenir les meilleurs performances. Notre architecture incluant les transformers améliore le F1-score de 2% et le mAP de 8%, par rapport au YOLOv5 classique.

Le temps d'inférence est d'environ trente secondes pour les réseaux basés sur YOLO. Cela confirme le caractère temps-réel, particularité de ces réseaux. Faster R-CNN avec Mobilenet fait cinquante secondes environ. Tandis que les réseaux Faster R-CNN avec des backbones plus complexes réalisent l'inférence en moins de deux (2) minutes sur la totalité des images de l'ensemble de test. Ces durées

restent extrêmement rapides comparées au temps d'annotation manuelle par un expert.

3.5.6 Variation de la précision en fonction de la surface des zones d'altérations

Une part importante des 3955 zones d'altérations de la base de données BD-Altérations sont de petites surfaces. L'histogramme de la figure 3.12 montre le pourcentage cumulatif des zones d'altérations en fonction de leur surface (en pixels). La surface est estimée par l'aire de la boîte englobante. Ainsi, lorsque l'on filtre les zones d'altération pour ne retenir que celles dont la surface est d'au moins 155 pixels, on observe que cela représente environ 60% des zones. De plus, seulement 10% des zones d'altérations ont une surface d'au moins 5000 pixels.

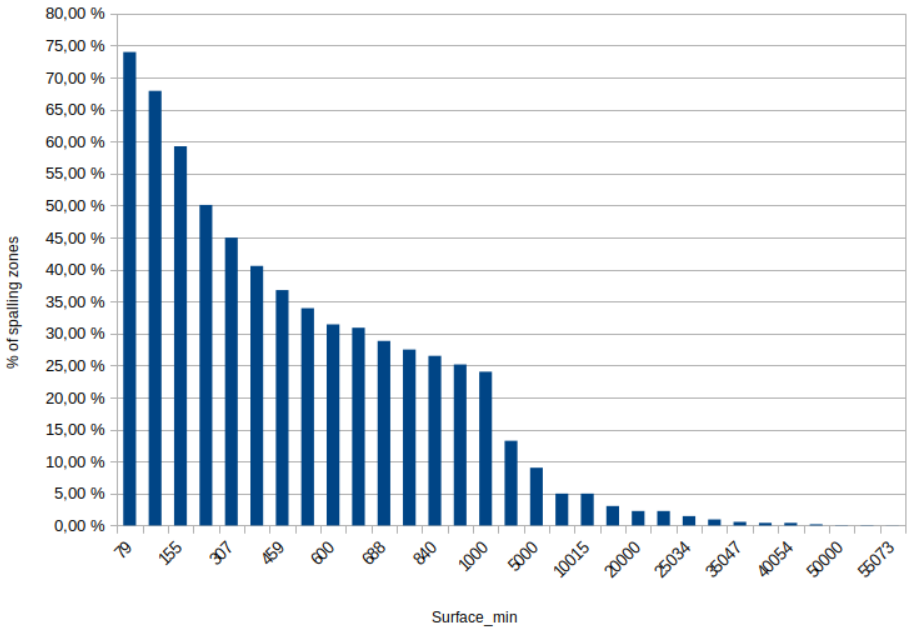


FIGURE 3.12 – Histogramme cumulatif des zones d'altérations en fonction de leur surface (en pixels) pour les 3955 zones de la base de données.

La figure 3.13 présente une étude de l'impact de la surface d'altération sur les performances du réseau en terme de précision moyenne (mAP). Pour tous les réseaux, la tendance générale est croissante ; plus la surface des zones d'altérations est grande, plus le mAP est élevé. Cependant, les réseaux Faster-RCNN avec ResNet50 et Faster-RCNN avec MobileNet n'arrivent pas à détecter les zones d'altération dont la surface est de plus de 45000 pixels, ce qui représente des zones

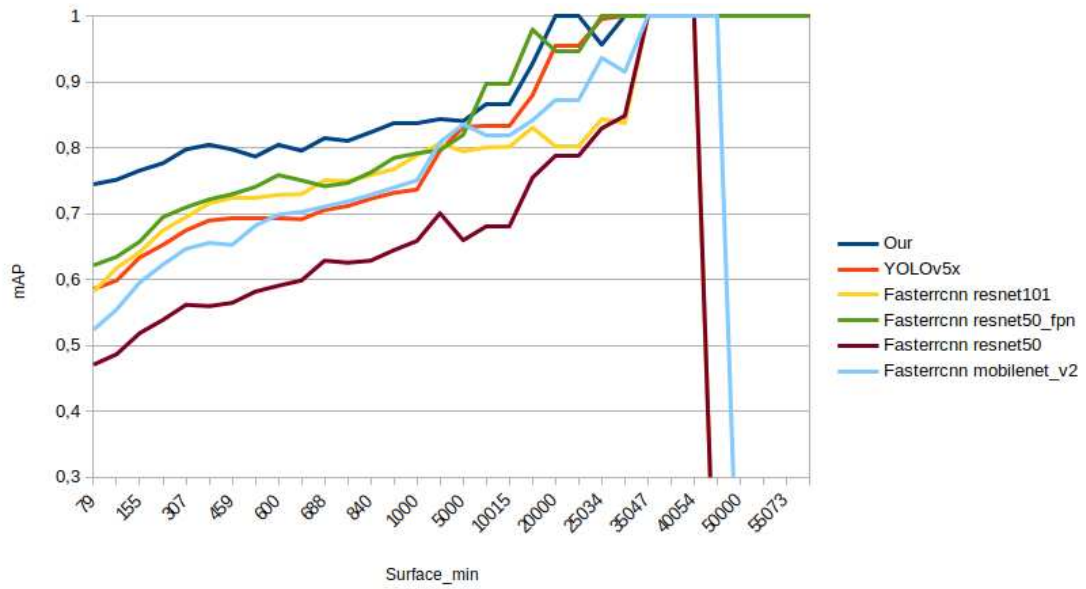


FIGURE 3.13 – Variation de la précision moyenne de la détection (mAP) en fonction de la surface minimale des zones d'altérations pour les 3955 zones de la base de données.

d'altérations qui couvrent environ 70% ou plus d'une image de 256×256 pixels.

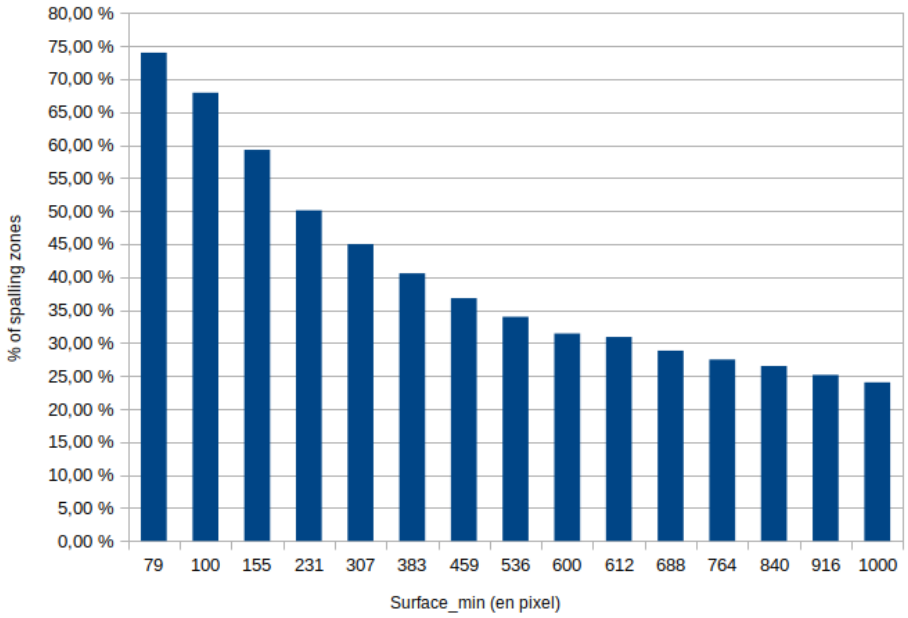


FIGURE 3.14 – Histogramme cumulatif des 3000 les plus petites.

La figure 3.15 analyse plus en détails le comportement des réseaux sur les 3000 plus petites zones en terme de surface. L'histogramme de la figure 3.14 montre le pourcentage cumulatif de ces zones d'altérations en fonction de leur surface en pixels.

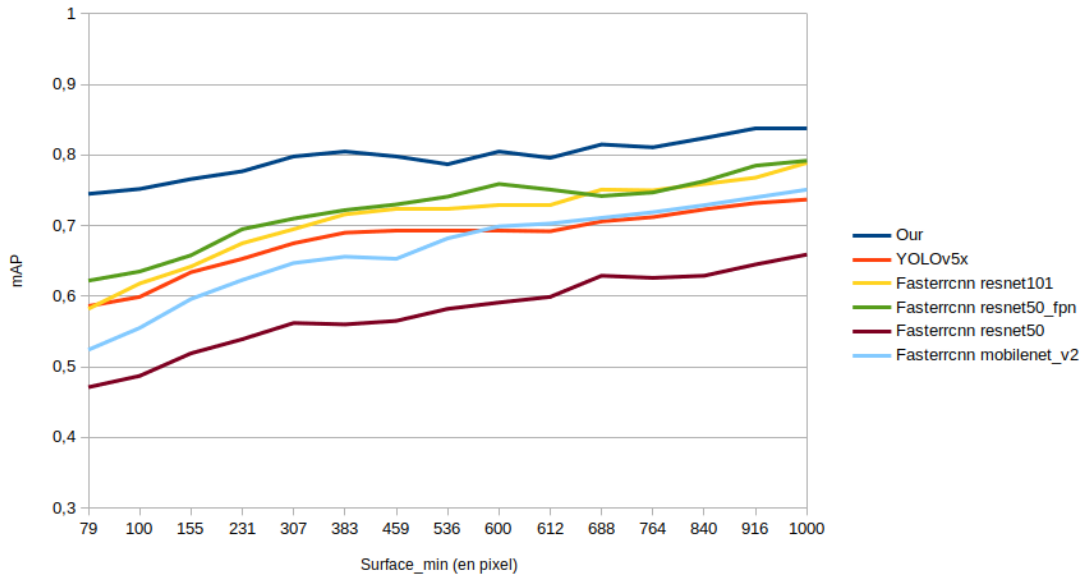


FIGURE 3.15 – Variation de la précision moyenne de la détection (mAP) en fonction de la surface minimale des zones d'altérations pour les 3000 plus petites zones de la base de données.

La figure 3.15 distingue plus précisement, les performances de chacun des réseaux sur les courbes. Le réseau proposé a un mAP qui évolue significativement au-dessus des autres tirant avantage de ses couches de transformers qui produisent des caractéristiques permettant une detection plus précise.

3.5.7 Segmentation des masques par zone d'altération

Le réseau Mask R-CNN permet, en sus de détecter les boîtes englobantes des zones de desquamations, de fournir le masque de segmentation au plus près de la zone. Il a été entrainé avec le backbone Resnet50-FPN qui a donné les meilleurs résultats. Le mAP des masques de segmentation est calculé de façon simultanée aux boîtes englobantes mais en considérant la taille des masques de segmentation dans le calcul de l'IoU.

La figure 3.16 présente quelques exemples de détection réalisés avec le réseau Mask R-CNN. La vérité-terrain est étiquetée avec des couleurs distinctes pour chaque zone de desquamations en plaque. Les tests nous démontrent que, d'une manière générale, le réseau produit de bons résultats sur les images présentant peu de zones de desquamations en plaques. Mais les performances diminuent sur les images présentant de très nombreuses zones de desquamations en plaques. Le

réseau reste robuste sur les images présentant des éclairages non optimaux (surexposition ou sous exposition) comme le montrent les troisième et quatrième colonnes de la figure 3.16.

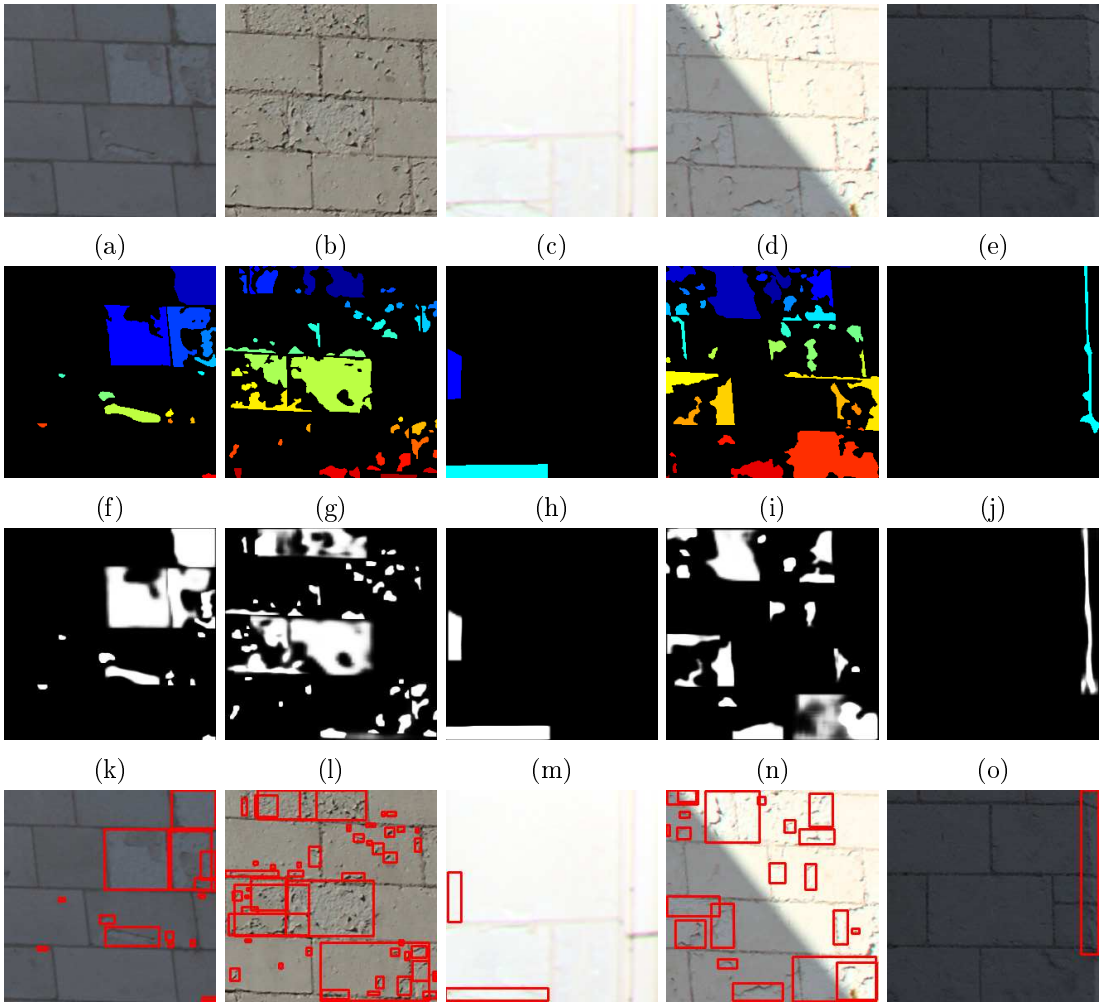


FIGURE 3.16 – Exemple de détection de desquamations en plaque avec le réseau Mask-RCNN. **Première ligne :** Images couleur ; **Deuxième ligne :** Vérité terrain présentant les zones de desquamation en plaque labellisées par les experts sur orthomosaïque et reprojettés sur ces images ; **Troisième ligne :** Masque de segmentation prédit par le réseau Mask-RCNN ; **Quatrième ligne :** Boîtes englobantes détectées par le réseau Mask-RCNN.

Le tableau 3.5 présente les performances du réseau sur les images de test. La précision moyenne de la prédiction des masques de segmentation et de la détection des boîtes englobantes sont toutes les deux autour de 84%, pour un seuil d'IoU de $0,50$. Le temps d'inférence est environ trois fois plus élevé que pour les réseaux précédents qui ne font que la détection des boîtes englobantes sans la segmentation.

Néanmoins, il reste inférieur à deux minutes pour l'ensemble de la base de test, ce qui reste extrêmement rapide, comparé au temps passé par un expert pour produire les annotations manuelles.

Tableau 3.5 – Performances du réseau Mask R-CNN pour la détection des boîtes englobantes et la segmentation des masques.

Réseau	Masque de Segmentation mAP, IoU 0,50	Boîtes englobantes mAP, IoU 0,50	Temps d'Inference (minute :seconde)
Mask R-CNN & Resnet50-FPN	0,843	0,844	01 :27

3.6 Bilan du chapitre

Dans ce chapitre, nous avons proposé une nouvelle architecture de réseau qui combine la plus récente version du réseau YOLO et des modules de transformers. Nous renforçons les capacités d'apprentissage du réseau en tirant profit des dernières innovations en apprentissage, telles que la définition automatique des ancres par l'algorithme k-means et le filtrage des boîtes englobantes par la méthode de DIoU-NMS. Une fois entrainé sur notre base d'images d'altérations, ce réseau permet de détecter les zones de desquamation en plaque avec une précision supérieure à 80%.

Ces performances surpassent celles des approches utilisées jusque-là dans ce domaine basées sur les architectures Faster R-CNN.

L'étude de l'influence de la surface des zones d'altérations sur la précision des réseaux a permis d'observer des limites dans la détection de zones de desquamations en plaques de petite surface. L'approche proposée reste néanmoins moins impactée que les autres. Une discussion avec les experts sur la pertinence du maintien de ces très petites zones d'altérations, en lien avec la stratégie de reprojection est en cours.

Le réseau Mask R-CNN a également été testé sur notre base d'altérations. Il démontre une très bonne aptitude à prédire un masque au plus près des contours de chaque zone de desquamation en plaques avec une précision de près de 85%.

Chapitre 4

Analyse globale d'images à l'échelle d'un château

> Sans la curiosité de l'esprit que serions-nous ?
> Telle est bien la beauté et la noblesse de la science : désir sans fin de repousser les frontières du savoir, de traquer les secrets de la matière et de la vie sans idée préconçue des conséquences éventuelles.
>
> Marie Curie

Sommaire

4.1 Introduction

Rappelons que notre objectif est de développer des méthodes de traitement d'images pour l'analyse des monuments emblématiques du patrimoine historique bâti. Ces méthodes serviront à fournir aux experts des outils pour renforcer la surveillance et le diagnostic sur l'état de santé des monuments. Ce chapitre expose un aperçu pour l'application à l'échelle d'un château des approches développées dans les chapitres précédents. Il présente aussi l'application web développée pour l'analyse et le traitement des images et orthomosaïques d'un patrimoine et la collaboration entre experts.

Le reste du chapitre est subdivisé en trois sections. Dans la première, nous discutons l'application à l'échelle d'une grande orthomosaïque des approches proposées (voir section 4.2). La deuxième section présente l'application web développée, qui intègre les approches proposées, avec une interface conviviale pour faciliter l'application sur des orthomosaïques de grandes dimensions à l'échelle d'un château (voir section 4.3). Enfin la troisième section fait le bilan du chapitre.

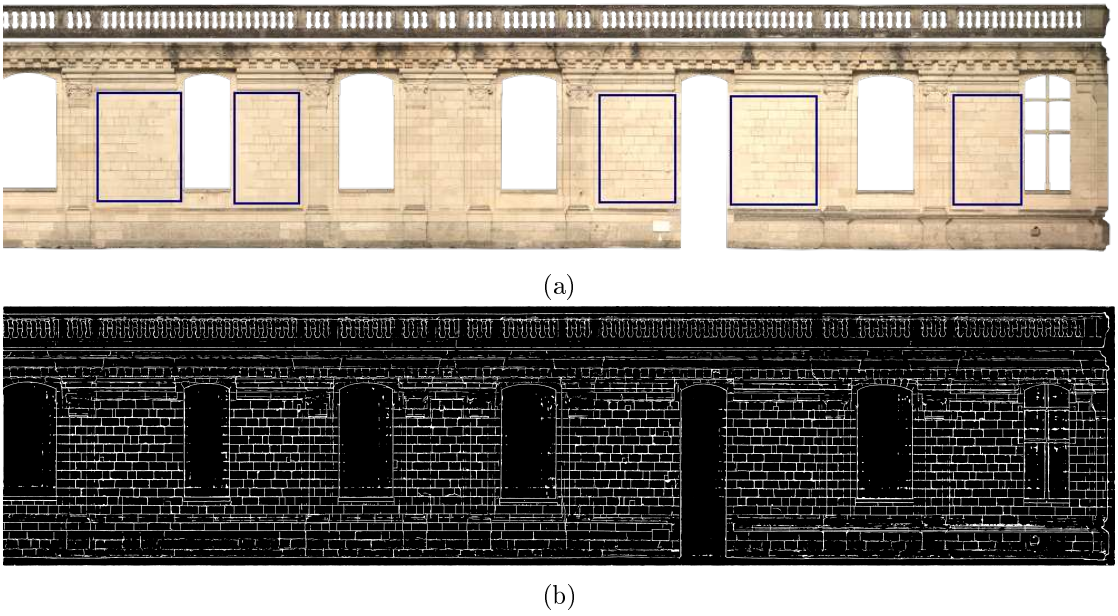
4.2 Analyse sur de grandes orthomosaïques

Pour le traitement des larges images, différents mécanismes sont communément utilisés en traitement d'images avec les réseaux de neurones convolutifs. L'un consiste à découper l'image en imagettes par fenêtre glissante. Chaque imagette est traitée séparément dans la phase d'entrainement des réseaux. Dans la phase de test, les imagettes sont ensuite assemblées pour obtenir le résultat. Cette approche présente l'avantage d'agrandir la base de données de par la division d'une grande image en imagettes plus petites et plus nombreuses.

L'autre revient à réduire l'échelle de l'image afin d'avoir à traiter une image plus petite. Dans la phase d'apprentissage elle pourrait impliquer une base de données plus petite pour l'entrainement des réseaux. Cette approche pourrait aussi engendrer des baisses de performance des réseaux qui pourraient s'expliquer par la perte de précision des contours lors de la réduction d'échelle de l'image [133]. Il est aussi possible de réduire l'echelle tout en découpant l'image en imagettes pour

traitement adaptatif à différentes résolutions [134].

Les méthodes adoptées varient suivant la problématique et les spécificités des images. Dans les chapitres précédents, pour les phases d'entrainement des réseaux, nous avons adopté la stratégie de la fenêtre glissante. Ainsi, les larges orthomosaïques ont été découpées en imagettes de 256×256 pixels. Pour les phases de test, la même stratégie est adoptée. La figure 4.1 montre le résultat de l'application du modèle basé sur le réseau DeepLabv3+ sur une partie de l'orthomosaïque de l'enceinte basse du château de Chambord.



(a)

(b)

FIGURE 4.1 – (a) Partie de l'orthomosaïque de l'enceinte basse du château de Chambord. Les cadres bleues représentent les parties de murs découpées. (b) Résultat de la Segmentation pierre-à-pierre.

4.3 VMHB : Application web

Dans le cadre de cette thèse, il a été développé une application web de partage et de traitements des données, essentiellement des orthomosaïques de grandes dimensions à l'échelle d'un Château. L'application est dénommée VISION FOR MONITORING HISTORICAL BUILDINGS (VMHB).

Le choix de la conception d'une application web a été fortement motivé par l'accessibilité. Elle permet de faciliter la collaboration entre experts quelque soit leur situation géographique. Il suffit de disposer d'un navigateur et d'une connexion

internet.

Cependant, la fluidité de la navigation dans une application web est dépendante de la taille (évaluée en Kilo ou Mega Octet) des pages de l'application. Les orthomosaïques des données existantes sont de grande taille (en Mega Octet) et de grande dimension (voir chapitre 1). La problématique du maintien de la fluidité lors de la manipulation de ces orthomosaïques à travers l'application web s'est posée. Nous avons résolu cette problématique en implémentant le tuilage des orthomosaïques.

4.3.1 Tuilage d'une orthomosaïque

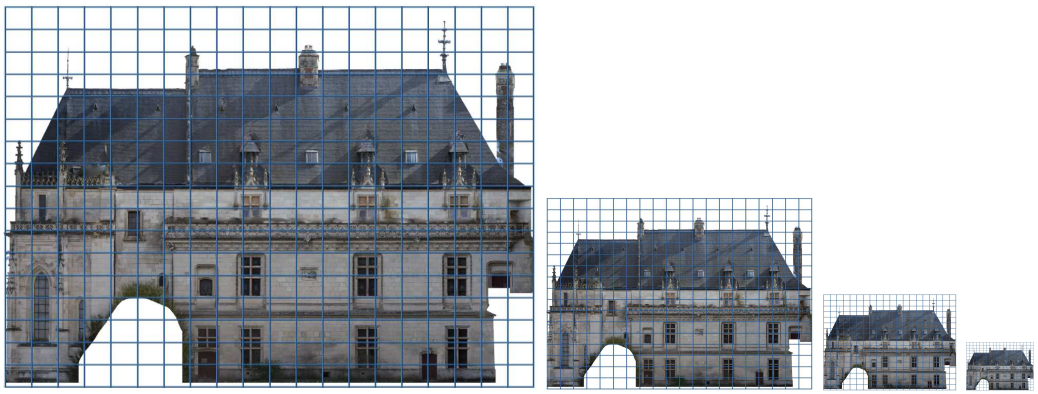
Le tuilage est une technique permettant la visualisation et la navigation dans les images de très grandes dimensions, généralement de larges cartographies numériques, à travers une interface web. Elle consite à prendre en entrée une image de très haute résolution et produire en sortie un grand nombre de tuiles. Chaque tuile représente une partie de l'image à un niveau de zoom donné [135].

Dans une premier temps, le serveur reçoit une requête d'affichage. Il détermine le niveau de zoom. Puis identifie et retourne les tuiles correspondantes au niveau de zoom et à la partie de l'image à afficher. Le chargement et l'affichage se font pour chaque tuile individuellement et plusieurs tuiles sont chargées en simultané sur l'interface web. Il revient ainsi beaucoup plus optimal de charger quelques dizaines de petits fichiers images de quelques kilo-octets et les afficher au fur et à mesure ; au lieu de charger un seul fichier image de plusieurs centaines de méga-octets et ne pouvoir l'afficher qu'à la fin de tout le chargement.

Il existe des librairies open source et libres telles que OpenSeadragon, Leaflet et Open Layers ; pour intégrer le processus de tuilage à une application web. Open Layers [136] et Leaflet [137] ont été conçues spécifiquement pour la cartographie numérique. OpenSeadragon [138] a été développée pour les grandes images du patrimoine culturel comme les bâtiments historiques, les tableaux d'art et autres. Cette liste n'est pas exhaustive, mais l'étude de ces principales librairies nous a permis de faire le choix de la librairie OpenSeadragon, plus adaptée à nos orthomosaïques.

Le tuilage est réalisé une seul fois côté serveur pour chaque nouvelle orthomosaïque lors de son chargement dans l'application web. Le nombre de niveau de zoom ($nbZoom$) considéré est défini par : $nbZoom = \log_2 2(max(w, h))$. Les tuiles obtenues sont sauvegardées sur le serveur. La taille des tuiles est fixée à 128×128 pixels. A chaque nouvelle requête d'affichage, elles sont chargées progressivement en fonction du niveau de zoom et de la navigation dans l'interface.

La figure 4.2 présente un exemple de découpage des tuiles sur l'orthomosaïque de la façade cour Est du château de Chaumon-sur-Loire. Elle a une largeur de 17025 pixels et une hauteur de 11955 pixels. Chaque cellule à bordures bleues, représente un exemple de découpage en tuile pour différents niveaux de zoom. Pour cette orthomosaïque, 14 niveaux de zoom ont été obtenus à l'issue de l'opération de tuilage.



(a) (b) (c) (d)

FIGURE 4.2 – Exemple de découpage des tuiles suivant différents niveaux de zoom sur l'orthomisaïque de la façade cour Est du château de Chaumon-sur-Loire : (a) niveau de zoom à 100%. (b) niveau de zoom à 50%. (c) niveau de zoom à 25%. (c) niveau de zoom à 12,5%

4.3.2 Architecture de l'application

L'application web VMHB a été développée suivant une architecture Client-Serveur. Ce type d'architecture permet d'isoler les traitements qui peuvent solliciter beaucoup de ressources sur un serveur dédié. La partie Client, se réalise essentiellement au regard de l'utilisateur avec l'exécution de l'application sur un ordinateur à travers un navigateur.

La figure 4.3 montre l'architecture de l'application web VMHB. Elle est divisée

en deux(02) parties : le front-end et le back-end. Le front-end, est l'ensemble de toutes les interfaces de l'application web qui s'exécutent dans la partie Client. Il est développé avec les languages HTML, CSS et Javascript. Le back-end, représente l'ensemble des traitements et processus de l'application web.

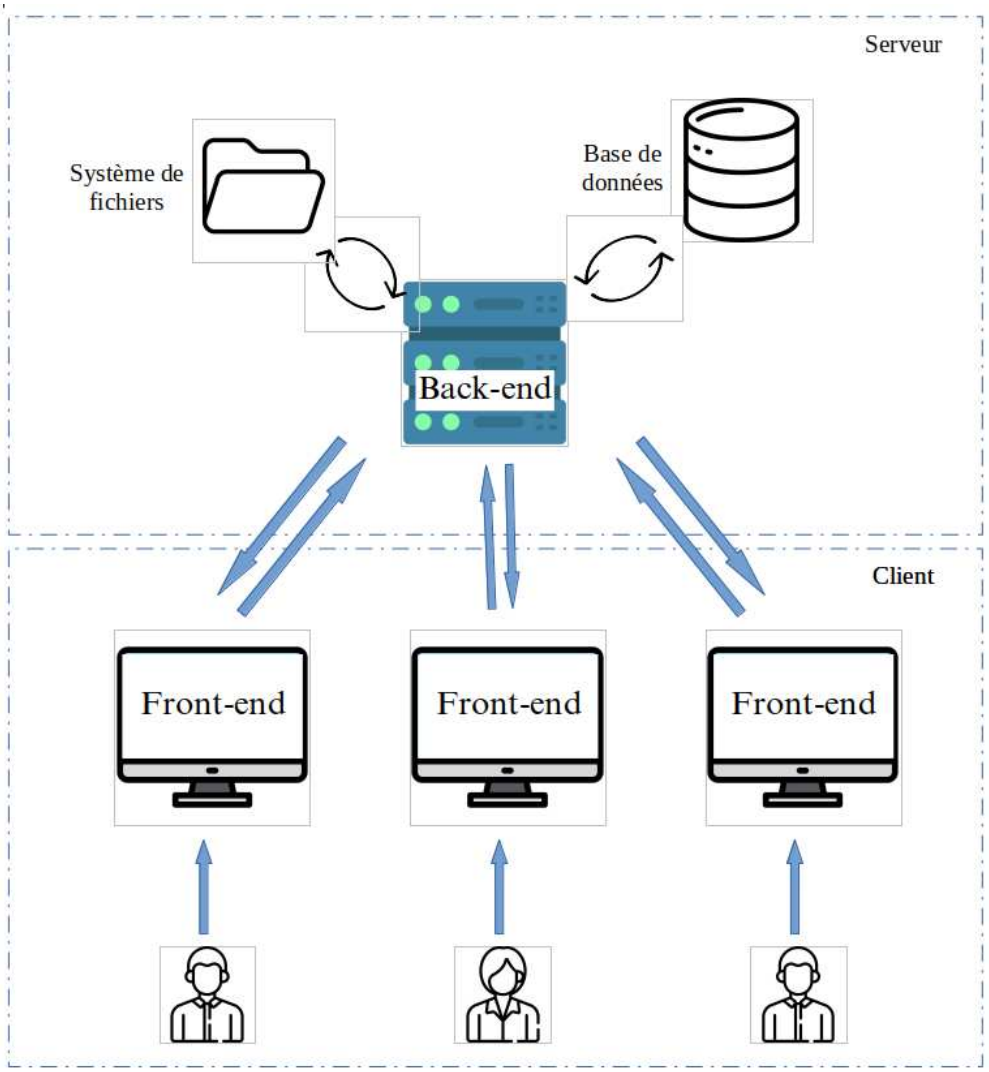


FIGURE 4.3 – Architecture de l'application web VMHB.

Le back-end de VMHB est une API Web constituée de collection de web services nécessaires pour répondre à chacune des interactions sur le front-end qui requiert un traitement. Il est développé en Python, Django et SQL pour la gestion de la base de données. Django est un framework en python, libre et open source, qui simplifie le développement d'applications web complexes et robustes [139, 140]. Le back-end interagit aussi avec le système de fichiers pour la lecture, l'écriture des

orthomosaïques et la sauvegarde des résultats de traitements des orthomosaïques sur le serveur.

4.3.3 Fonctionnalités et interface utilisateur

VMHB dispose de différentes fonctionnalités implémentées à travers quatre (04) principaux modules : la gestion des projets, la gestion des calques, la gestion des images et la gestion des utilisateurs.
La gestion des projets consiste à gérer l'environnement de travail d'un utilisateur et à l'associer aux données qu'il traitera sur l'application. Un projet représente un ensemble d'opérations sur des images et des calques effectués par un utilisateur sur un monument historique donné. Il peut être public ou privé. Un projet public est accessible par d'autres utilisateurs mais ne peut être supprimé que par son auteur. La figure 4.4 présente l'interface du tableau de bord de la gestion de projet de l'application. Sur l'interface, la section Mes projets récapitule les projets de l'utilisateur connecté et la section Projets publics regroupe les projets créés par d'autres utilisateurs et qui sont déclarés Publics.

La gestion des images porte sur l'ajout et la manipulation d'une image dans un projet. Une image peut être une image couleur originale prise sur site ou une orthomosaïque produite à partir de ces images couleur. Plusieurs images peuvent être ajoutées à un projet. Ainsi, des orthomosaïques de différentes façades peuvent être ajoutées au projet Chambord par exemple, y être traitées puis exportées.

La gestion des calques permet d'exécuter et sauvegarder les différents traitements effectués sur les images. Elle se compose de fonctionnalités comme la création, le déplacement, la modification, la suppression des calques ; et aussi le chargement des calques existants. Un calque est une zone de l'image, sélectionnée par l'utilisateur pour y appliquer une opération. L'enrégistrement de calques sauvegarde les résultats des opérations de segmentation pierre-à-pierre et de détection des altérations, réalisées par l'utilisateur sur un calque. La figure 4.5 montre l'interface de création des claques et d'applications des méthodes. Elle comporte une barre lattérale gauche qui regroupe toutes les opérations sur les images dans la section MÉTHODES, toutes les opérations sur les calques dans la section CALQUES

et tous les outils de visualisation des résultats à travers la superposition des calques et le réglage de l'opacité.

Enfin, la gestion des utilisateurs consiste à créer, authentifier et gérer les utilisateurs de VMHB. Elle permet la connexion multi-utilisateur, l'échange pour l'intervention de plusieurs acteurs sur un même projet.

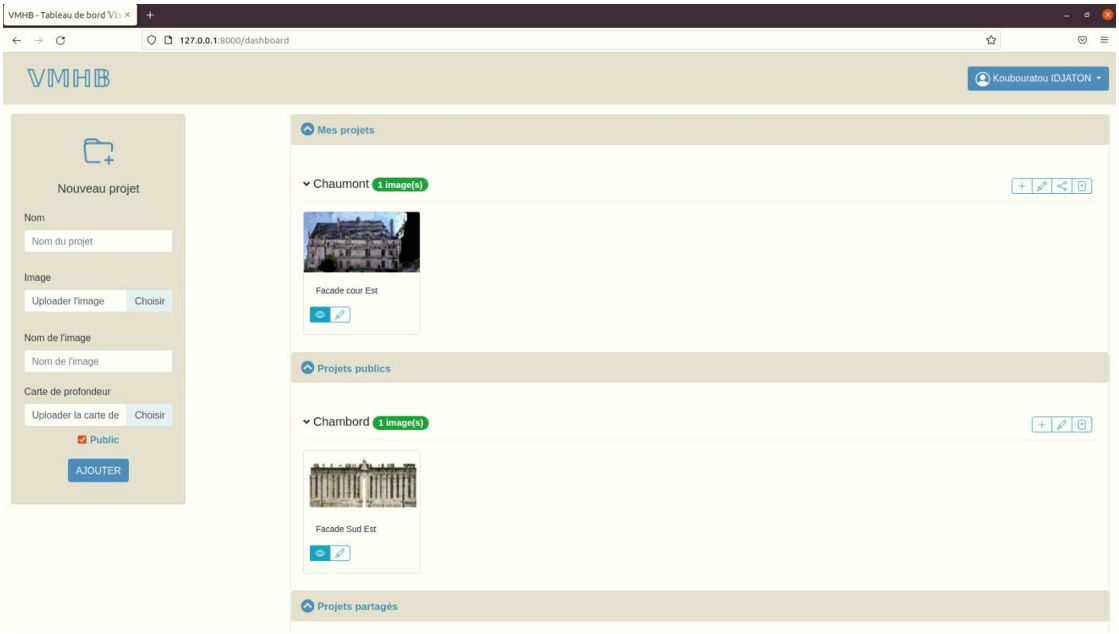


FIGURE 4.4 – Application web VMHB : Interface de création des claques et applications des méthodes.

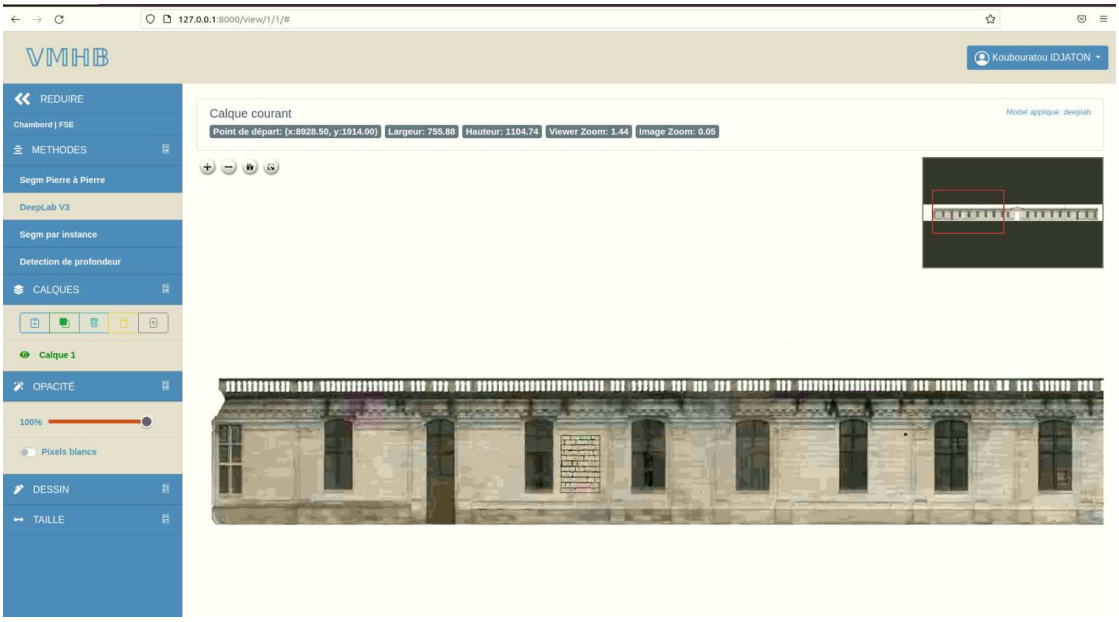


FIGURE 4.5 – Application web VMHB : Tableau de bord de la gestion de projets.

4.4 Bilan du chapitre

Dans ce chapitre, nous avons discuté l'application sur de grandes orthomosaïques à l'échelle d'un château des réseaux proposés. La méthode du découpage de grande orthomosaïque en imagettes est utilisée. Elle permet de conserver plus de détails des contours des joints ou des altérations.

L'application web VMHB développée sert à implémenter ces réseaux au travers une interface conviviale pour un usage multi acteurs. La technique du tuilage des images est employée pour faciliter la manipulation des grandes images via un navigateur web. L'application est structurée suivant une architecture client-server qui facilite le traitement des réseaux en arrière plan grâce à des services s'exécutant sur le serveur avec une base de données MySQL.

Conclusion et Perspectives

La préservation des bâtiments historiques exige d'expertiser et qualifier leur état. Une étape importante est le diagnostic de l'état de santé par une évaluation minutieuse et souvent fastidieuse. La segmentation des pierres et la détection des altérations des pierres sont des tâches cruciales de cette évaluation. Cette thèse a pour objectif de proposer des méthodes de traitement d'images pour l'aide au diagnostic des châteaux de la Loire, à partir des données existantes (images couleur et orthomosaïques), et en tirant profit d'un apprentissage machine.

Deux ensembles de données issues d'acquisitions 3D photogrammétriques des châteaux de Chambord et de Chaumont-sur-Loire ont été exploités. Ces ensembles de données ont été constitués lors de plusieurs campagnes d'acquisition entre 2014 et 2021.

Les travaux présentés dans ce manuscrit proposent plusieurs contributions :

— la création de deux bases d'apprentissage : l'une pour la segmentation pierre-à-pierre, l'autre pour la détection des altérations des pierres ;
— un modèle de segmentation automatique des joints sur les maçonneries en pierres de tuffeau exploitant l'un des meilleurs réseaux convolutionnels de l'état de l'art ;
— un modèle de prédiction des desquamations en plaques sur ces mêmes maçonneries introduisant une nouvelle architecture de réseau basée sur YOLO et des transformers ;
— enfin, une application WEB pour l'application des modèles à l'échelle d'un château à travers une interface conviviale.

La segmentation pierre-à-pierre est réalisée manuellement par les experts en patrimoine. Elle consistait à tracer méticuleusement les contours des joints qui dé-

tourent chaque surface de pierre. Ce procédé a inspiré les principales approches de l'état de l'art qui proposent des méthodes basées sur les algorithmes classiques de détection de contours, segmentation de région et seuillage comme Canny, Sobel, transformée de Hough et transformée en ondelette continue. Avec les prouesses des réseaux de neurones une nouvelle approche basée sur le réseau d'apprentissage profond pour la segmentation sémantique U-Net a aussi été avancée. Nous avons étudié ces approches sur notre base de données. Il ressort que les algorithmes classiques sont très sensibles au bruit dans l'image, aux zones d'altérations des pierres, aux différences de luminosités dues aux conditions d'acquisitions des images sur site. Une première approche ad-hoc de segmentation pierre-à-pierre a permis de réaliser de façon semi-automatique une base de données avec vérité-terrain en facilitant grandement la labélisation manuelle des experts. Pour améliorer les approches existantes, nous avons introduit deux nouvelles approches de segmentation pierre-à-pierre à partir des réseaux d'apprentissage profond SegNet et DeepLabv3+. Ces approches ont démontré de meilleures performances. Cependant celle basée sur SegNet présente quelques limites dues à la perte de détails en descendant dans les couches profondes de l'architecture du réseau. Celle basée sur DeepLabv3+ se démontre plus robuste et performante grâce aux modules de convolution à trous, sa fonction résiduelle et l'extraction de caractéristiques multi-échelles dans l'image traitée. L'étude comparative présentée dans le chapitre 2 sur les performances quantitatives et qualitatives démontrent que ces nouvelles approches améliorent l'état de l'art pour la segmentation pierre-à-pierre sur les images des châteaux du style de la renaissance.

La détection d'altération des pierres est réalisée par une annotation manuelle sur les orthomosaïques suite à plusieurs observations des façades sur site par les experts. Cinq types d'altérations couvrant 22% de la surface des façades sont relevés. Parmi eux, la desquamation en plaque, causant la chute de matière, est très préjudiciable à l'accueil des touristes, et est la plus prolifique couvrant 98% de la surface des altérations. Nous avons proposé une nouvelle architecture de réseau qui combine la plus récente version du réseau YOLO et des modules de transformers. Elle permet de détecter les zones de desquamation en plaque avec

une précision supérieure à 80%. Ces performances surpassent celles des approches de la littérature qui se basent sur les architectures Faster R-CNN sur la base de données BD-Altérations. Le réseau proposé et les approches existantes présentent des limites dans la détection de zones de desquamations en plaques de petite surface. Nous avons étudié la variation de précision de chacun des réseaux en fonction de la surface des zones d'altérations. On observe que le modèle proposé reste moins impacté que les autres. Pour être toujours plus précis dans la détection des différentes zones d'altérations, une autre approche a également été testée, en utilisant le réseau Mask R-CNN. Elle permet de prédire un masque au plus près des contours de chaque zone de desquamation en plaques avec une précision de près de 85%.

Les approches proposées durant nos travaux ont été intégrées dans une application web dénommée VMHB, vision for monitoring historical building. Elle sert à implémenter ces réseaux au travers d'une interface conviviale pour un usage multi-acteur. La complexité de manipulation des orthomosaïques de très grandes tailles dans un navigateur web a été réglée avec l'implémentation de la technique du tuilage des images. Elle facilite la manipulation des grandes images comme les cartes, les images spatiales ou dans notre cas, les images de patrimoines dans un navigateur web. Les opérations de segmentation et de détection avec les réseaux dans l'application se font en arrière-plan grâce à des services s'exécutant sur le serveur avec une base de données MySQL suivant la structuration client-server de VMHB. Enfin, la méthode du découpage de grande orthomosaïque en imagettes permettant de conserver plus de détails des contours des joints ou des altérations est utilisée pour l'application des réseaux proposés sur des grandes orthomosaïques à l'échelle d'un château.

Dans les travaux futurs, une évaluation plus élargie des approches proposées pour la segmentation pierre-à-pierre serait souhaitable. Même si elles donnent de bonnes performances sur notre base de données, leurs capacités de généralisation à d'autres monuments historiques de la renaissance devront être étudiées et renforcées au besoin.

Pour la détection d'altération, même si le HB-Spall-TransYOLO a démontré

de bonnes performances sur la base de données BD-Altérations, il serait souhaitable de simplifier son architecture en proposant une nouvelle architecture centrée sur l'attention et les modules de spatial pyramid pooling. De plus, une discussion avec les experts sur la pertinence du maintien des très petites zones d'altérations, en lien avec la stratégie de reprojection est en cours. Développer un modèle de détection multi-classes identifiant plusieurs types d'altérations est aussi une piste de travaux futurs. Une évaluation plus approfondie de ces réseaux sur une base de données avec plus de types d'altérations pourrait révéler des observations intéressantes. Cela nécessitera que les experts procèdent à des acquisitions complémentaires d'images sur site et des annotations manuelles. En attendant la possibilité de nouvelles campagnes d'acquisition, une base de données d'images artificielles mais réalistes pourrait être mise en place par clonage des altérations sur des imagettes de murs sans altérations. Cela permettrait de renforcer la capacité du réseau à détecter différents types d'altérations moins prédominantes que les desquamations en plaques.

Enfin, l'application VMHB, même si elle est opérationnelle, peut être améliorée pour permettre aux experts d'affiner les résultats de détections ou de segmentation lorsqu'elles sont moins précises.

Publications

Au cours de cette thèse, les travaux effectués ont débouché sur des publications scientifiques, des participations à des conférences nationales et internationales ainsi qu'une revue internationale en cours de révision. Des activités de vulgarisation scientifique ont aussi été menées. Nous présentons ci-dessous une liste complète des diverses publications et activités de vulgarisation scientifique.

Revue Internationale

— IDJATON, Koubouratou, JANVIER, Romain, DESQUESNES, Xavier, BRUNETAUD, Xavier et TREUILLET, Sylvie. Automatic detection of limestone spalling in 3D survey images of a Loire Valley castle using deep learning. Automation in construction, 2022 (en cours de révision)

Conférence Internationale

— IDJATON, Koubouratou, DESQUESNES, Xavier, TREUILLET, Sylvie et BRUNETAUD, Xavier. Transformers with YOLO network for damage detection in limestone wall images. 21st International Conference on Image Analysis and Processing. Springer, 2022.

— IDJATON, Koubouratou, DESQUESNES, Xavier, TREUILLET, Sylvie et BRUNETAUD, Xavier. Stone-by-Stone Segmentation for Monitoring Large Historical Monuments Using Deep Neural Networks. 25th International Conference on Pattern Recognition. Springer, 2021. p. 235-248.

Conférence Nationale

— IDJATON, Koubouratou, DESQUESNES, Xavier, TREUILLET, Sylvie et BRUNETAUD, Xavier. Segmentation semi-automatique diimages pour le

diagnostic de monuments historiques. XXVIIème Colloque GRETSI, Lille, Août 2019.

Vulgarisation Scientifique

— Journée du groupe RTR DIAMS Novembre, 2021 (Poster)
— Journée Patrimoine RTR DIAMS Juillet, 2021 (Oral)
— Concours MT180, 2021 (Finaliste Université d'Orléans et participation à la finale régionale)
— Journée des doctorants de l'École Doctorale MIPTIS, 2020 (Oral)
— Programme Edifice, 2019 (Encadrement de lycéens pour la découverte de la recherche scientifique)

Bibliographie

[1] Domaine national de Chambord. Chambord, dossier de présentation. https://www.chambord.org/fr/wp-content/uploads/sites/2/2016/11/DOMAINE_CHAMBORD_DOSSIER_PRESENTATION_2017-version-finale.pdf, 2017.

[2] Office du Tourisme Blois-Chambord. Le château de chaumont-sur-loire, belvédère sur la loire sauvage. https://www.bloischambord.com/explorer/les-chateaux/le-chateau-de-chaumont-sur-loire.

[3] A. Pinte, R. Héno, M. Pierrot-Deseilligny, X. Brunetaud, S Janvier-Badosa, and R. Janvier. Orthoimages of the outer walls and towers of the château de Chambord. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-5/W3 :243–250, aug 2015.

[4] Sarah Janvier-Badosa, Xavier Brunetaud, Kévin Beck, and Muzahim Al-Mukhtar. Kinetics of stone degradation of the castle of chambord in france. *International Journal of Architectural Heritage*, 10(1) :96–105, 2016.

[5] Enrique Valero, Frédéric Bosché, and Alan Forster. Automatic segmentation of 3d point clouds of rubble masonry walls, and its application to building surveying, repair and maintenance. *Automation in Construction*, 2018.

[6] Roland Kajatin and Lazaros Nalpantidis. Image segmentation of bricks in masonry wall using a fusion of machine learning algorithms. In *International Conference on Pattern Recognition*, pages 446–461. Springer, 2021.

[7] Noelia Oses and Fadi Dornaika. Image-based delineation of built heritage masonry for automatic classification. In *International Conference Image Analysis and Recognition*, pages 782–789. Springer, 2013.

[8] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, jun 2015.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net : Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[10] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet : A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12) :2481–2495, dec 2017.

[11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn : towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6) :1137–1149, 2016.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets : Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv :1704.04861*, 2017.

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[17] George Vosselman and Hans-Gerd Maas. *Airborne and terrestrial laser scanning*. CRC press, 2010.

[18] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds : A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(12) :4338–4364, 2020.

[19] Reality capture software. https ://www.capturingreality.com/, 2020.

[20] Institut géographique national et École nationale supérieure des sciences géographiques. Micmac open source software. https ://micmac.ensg.eu/, 2005.

[21] Valeria Cappellini, Chiara Stefani, Nicolas Nony, and Livio De Luca. Surveying masonry structures by semantically enriched 2.5 d textures : a new approach. In *Euro-Mediterranean Conference*, pages 729–737. Springer, 2012.

[22] A Guarnieri, F Remondino, and A Vettore. Digital photogrammetry and tls data fusion applied to cultural heritage 3d modeling. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci*, 36 :1–6, 2006.

[23] Xian-Feng Hana, Jesse S Jin, Juan Xie, Ming-Jie Wang, and Wei Jiang. A comprehensive review of 3d point cloud descriptors. *arXiv preprint arXiv :1802.02297*, 2, 2018.

[24] Deseilligny M. Pierrot. Producing orthomosaic with a free open source software (micmac), application to the archeological survey of meremptah's tomb., 2014.

[25] Fondation Open Source Geospatial (OSGeo). Qgis : Système d'information géographique. `https://www.qgis.org/`, 2020.

[26] ICOMOS International Scientific Committee for Stone (ISCS). *ICOMOS-ISCS : Glossaire illustré sur les formes d'altération de la pierre*, volume XV of *Monuments & Sites*. ICOMOS, Paris, 2008.

[27] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Synthetic data augmentation using gan for improved liver lesion classification. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 289–293. IEEE, 2018.

[28] Sheng-Wei Huang, Che-Tsung Lin, Shu-Ping Chen, Yen-Yi Wu, Po-Hao Hsu, and Shang-Hong Lai. Auggan : Cross domain adaptation with gan-based data augmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 718–731, 2018.

[29] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1) :1–48, 2019.

[30] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2918–2928, 2021.

[31] Enrique Valero, A Forster, F Bosché, Camille Renier, Ewan Hyslop, and Lyn Wilson. High level-of-detail bim and machine learning for automated masonry wall defect surveying. In *Proceedings of the International Symposium on Automation and Robotics in Construction, Berlin, Germany*, pages 20–25, 2018.

[32] Luis Javier Sánchez-Aparicio, Susana Del Pozo, Luís F. Ramos, Andrés Arce, and Francisco M. Fernandes. Heritage site preservation with combined radiometric and geometric analysis of TLS data. *Automation in Construction*, 85 :24–39, jan 2018.

[33] C. Marson, G. Sammartano, A. Spanò, and M. R. Valluzzi. Lidar data analyses for assessing the conservation status of the so-called baths-church in Hierapolis of Phrygia (TR). *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W11 :823–830, may 2019.

[34] Zhikun Hou, Mohammad Noori, and R St Amand. Wavelet-based approach for structural damage detection. *Journal of Engineering mechanics*, 126(7) :677–683, 2000.

[35] Liang Yuan, Jingjing Guo, and Qian Wang. Automatic classification of common building materials from 3D terrestrial laser scan data. *Automation in Construction*, 2020.

[36] Anna Maria Manferdini, Valentina Baroncini, and Cristiana Corsi. An integrated and automated segmentation approach to deteriorated regions recognition on 3d reality-based models of cultural heritage artifacts. *Journal of cultural heritage*, 13(4) :371–378, 2012.

[37] Juan Corso, Josep Roca, and Felipe Buill. Geometric analysis on stone façades with terrestrial laser scanner technology. *Geosciences*, 7(4) :103, 2017.

[38] P Kapsalas, M Zervakis, P Maravelaki-Kalaitzaki, ET Delegou, and A Moropoulou. Machine vision schemes towards detecting and estimating the state of corrosion. In *Pattern Recognition and Signal Processing in Archaeometry : Mathematical and Computational Solutions for Archaeology*, pages 146–165. IGI Global, 2012.

[39] Matthias Hemmleb, Friederike Weritz, A Schiemenz, A Grote, and Christiane Maierhofer. Multi-spectral data acquisition and processing techniques for damage detection on building surfaces. *Image engineering and vision metrology, Dresden, Germany*, 2006.

[40] E Grilli, D Dininno, G Petrucci, and F Remondino. From 2d to 3d supervised segmentation and classification for cultural heritage applications. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 42(2), 2018.

[41] A. Murtiyoso and P. Grussenmeyer. Point cloud segmentation and semantic annotation aided by GIS data for heritage complexes. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume 42, pages 523–528. Copernicus GmbH, jan 2019.

[42] Fabio Remondino. Heritage recording and 3D modeling with photogrammetry and 3D scanning. *Remote Sensing*, 3(6) :1104–1138, jun 2011.

[43] Enrique Valero, Alan Forster, Frédéric Bosché, Ewan Hyslop, Lyn Wilson, and Aurélie Turmel. Automated defect detection and classification in ashlar masonry walls using machine learning. *Automation in Construction*, 106 :102846, 2019.

[44] Ingrid Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

[45] M. R. Hess, V. Petrovic, and F. Kuester. Interactive classification of construction materials : feedback driven framework for annotation and analysis of 3D point clouds. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W5 :343–347, aug 2017.

[46] Naomi S Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3) :175–185, 1992.

[47] Santosh Srivastava, Maya R Gupta, and Béla A Frigyik. Bayesian quadratic discriminant analysis. *Journal of Machine Learning Research*, 8(6), 2007.

[48] Harry Zhang. Exploring conditions for the optimality of naive bayes. *International Journal of Pattern Recognition and Artificial Intelligence*, 19(02) :183–198, 2005.

[49] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3) :273–297, 1995.

[50] Leo Breiman. Random forests. *Machine learning*, 45(1) :5–32, 2001.

[51] Trevor Hastie, Saharon Rosset, Ji Zhu, and Hui Zou. Multi-class adaboost. *Statistics and its Interface*, 2(3) :349–360, 2009.

[52] B Riveiro, B Conde-Carnero, H González-Jorge, P Arias, and JC Caamaño. Automatic creation of structural models from point cloud data : the case of masonry structures. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 2, 2015.

[53] Belén Riveiro, Paulo B. Lourenço, Daniel V. Oliveira, Higinio González-Jorge, and Pedro Arias. Automatic Morphologic Analysis of Quasi-Periodic Masonry Walls from LiDAR. *Computer-Aided Civil and Infrastructure Engineering*, 31(4) :305–319, apr 2016.

[54] Yahya Ibrahim, Balázs Nagy, and Csaba Benedek. Cnn-based watershed marker extraction for brick segmentation in masonry walls. In *International Conference on Image Analysis and Recognition*, pages 332–344. Springer, 2019.

[55] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1) :62–66, 1979.

[56] John Canny. A computational approach to edge detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-8 :679 – 698, 12 1986.

[57] Paul VC Hough. Method and means for recognizing complex patterns. *US patent*, 3(6), 1962.

[58] Mohamed Rizon, Haniza Yazid, Puteh Saad, Ali Yeon Md Shakaff, Abdul Rahman Saad, Masanori Sugisaka, Sazali Yaacob, M Rozailan Mamat, and M Karthigayan. Object detection using circular hough transform, 2005.

[59] Virendra Kumar Yadav, Saumya Batham, Anuja Kumar Acharya, and Rahul Paul. Approach to accurate circle detection : Circular hough transform and local maxima concept. In *2014 International Conference on Electronics and Communication Systems (ICECS)*, pages 1–5. IEEE, 2014.

[60] HK Yuen, John Princen, John Illingworth, and Josef Kittler. Comparative study of hough transform methods for circle finding. *Image and vision computing*, 8(1) :71–77, 1990.

[61] J Matasyx and C Galambosyand J Kittlery. Progressive probabilistic hough transform. In *Proceedings of the British Machine Vision Conference, Southampton, UK*, pages 14–17, 1998.

[62] Nahum Kiryati, Yuval Eldar, and Alfred M Bruckstein. A probabilistic hough transform. *Pattern recognition*, 24(4) :303–316, 1991.

[63] Richard S Stephens. Probabilistic approach to the hough transform. *Image and vision computing*, 9(1) :66–71, 1991.

[64] P Mathieu, M Barlaud, and M Antonini. Compression d'images par transformée en ondelette. In *12° Colloque sur le traitement du signal et des images, FRA, 1989*. GRETSI, Groupe d'Etudes du Traitement du Signal et des Images, 1989.

[65] Macarena Boix and Begoña Cantó. Wavelet transform application to the compression of images. *Mathematical and Computer Modelling*, 52(7) :1265–1270, 2010. Mathematical Models in Medicine, Business & Engineering 2009.

[66] Christophe Damerval. *Ondelettes pour la détection de caractéristiques en traitement d'images. Application à la détection de région d'intérêt.* Theses, Université Joseph-Fourier - Grenoble I, May 2008.

[67] Yves Meyer. Ondelettes et opérateurs. *I : Ondelettes*, 1990.

[68] Fernand Meyer. Topographic distance and watershed lines. *Signal processing*, 38(1) :113–125, 1994.

[69] Fernand Meyer. The watershed concept and its use in segmentation : a brief history. *arXiv preprint arXiv :1202.0216*, 2012.

[70] Richard Beare. A locally constrained watershed transform. *IEEE transactions on pattern analysis and machine intelligence*, 28(7) :1063–1074, 2006.

[71] MA Gonzalez and VL Ballarin. Automatic marker determination algorithm for watershed segmentation using clustering. *Latin American applied research*, 39(3) :225–229, 2009.

[72] Dadong Wang and Pascal Vallotton. Improved marker-controlled watershed segmentation with local boundary priors. In *2010 25th International Conference of Image and Vision Computing New Zealand*, pages 1–6. IEEE, 2010.

[73] Jae S Lim. Two-dimensional signal and image processing. *Englewood Cliffs, NJ, Prentice Hall, 1990, 710 p.*, 1990.

[74] Karel Zuiderveld. Contrast limited adaptive histogram equalization. In *Graphics gems IV*, pages 474–485. Academic Press Professional, Inc., 1994.

[75] I Sobel. An isotropic 3× 3 gradient operator, machine vision for three–dimensional scenes. *Freeman, H., Academic Pres, NY*, page 376379, 1990.

[76] Pierre Soille. *Morphological image analysis : principles and applications.* Springer Science & Business Media, 2013.

[77] Y.L. Cun. *Quand la machine apprend : La révolution des neurones artificiels et de l'apprentissage profond.* Odile Jacob, 2019.

[78] Ross Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, dec 2015.

[79] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[80] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, and Jose Garcia-Rodriguez. A review on deep learning techniques applied to semantic segmentation. *arXiv preprint arXiv :1704.06857*, 2017.

[81] Shervin Minaee, Yuri Y Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning : A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[82] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25 :1097–1105, 2012.

[83] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[84] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv :1409.1556*, 2014.

[85] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv :1412.7062*, 2014.

[86] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab : Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4) :834–848, 2017.

[87] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv :1706.05587*, 2017.

[88] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[89] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2012 (voc2012) results. http ://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[90] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv :1609.04747*, 2016.

[91] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[92] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2) :303–338, 2010.

[93] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco : Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[94] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet : A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009.

[95] Niannian Wang, Xuefeng Zhao, Peng Zhao, Yang Zhang, Zheng Zou, and Jinping Ou. Automatic damage detection of historic masonry buildings based on mobile deep learning. *Automation in Construction*, 103 :53–66, 2019.

[96] Niannian Wang, Qingan Zhao, Shengyuan Li, Xuefeng Zhao, and Peng Zhao. Damage classification for masonry historic structures using convolutional neural networks based on still images. *Computer-Aided Civil and Infrastructure Engineering*, 33(12) :1073–1089, 2018.

[97] Dohyung Kwon and Jeongmin Yu. Automatic damage detection of stone cultural property based on deep learning algorithm. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42(2/W15), 2019.

[98] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[99] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2) :154–171, 2013.

[100] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[101] Sergey Ioffe and Christian Szegedy. Batch normalization : Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[102] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth inter-*

national conference on artificial intelligence and statistics, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[103] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 38(1) :142–158, 2015.

[104] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[105] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once : Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[106] Glenn Jocher, K Nishimura, T Mineeva, and R Vilariño. Yolov5. *Code repository https ://github. com/ultralytics/yolov5*, 2020.

[107] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4 : Optimal speed and accuracy of object detection. *arXiv preprint arXiv :2004.10934*, 2020.

[108] Joseph Redmon and Ali Farhadi. Yolov3 : An incremental improvement. *arXiv preprint arXiv :1804.02767*, 2018.

[109] Joseph Redmon and Ali Farhadi. Yolo9000 : better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[110] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. Cspnet : A new backbone that can enhance learning capability of cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 390–391, 2020.

[111] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. Densenet : Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv :1404.1869*, 2014.

[112] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9) :1904–1916, 2015.

[113] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[114] Peiyuan Jiang, Daji Ergu, Fangyao Liu, Ying Cai, and Bo Ma. A review of yolo algorithm developments. *Procedia Computer Science*, 199 :1066–1073, 2022. The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021) : Developing Global Digital Economy after COVID-19.

[115] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[116] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2) :129–137, 1982.

[117] Yuanyi Zhong, Jianfeng Wang, Jian Peng, and Lei Zhang. Anchor box optimization for object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1286–1294, 2020.

[118] Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin. Region proposal by guided anchoring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2965–2974, 2019.

[119] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words : Transformers for image recognition at scale. *arXiv preprint arXiv :2010.11929*, 2020.

[120] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[121] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[122] Zixiao Zhang, Xiaoqiang Lu, Guojin Cao, Yuting Yang, Licheng Jiao, and Fang Liu. Vit-yolo : Transformer-based yolo for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2799–2808, 2021.

[123] Xingkui Zhu, Shuchang Lyu, Xu Wang, and Qi Zhao. Tph-yolov5 : Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2778–2788, October 2021.

[124] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union : A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019.

[125] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss : Faster and better learning for bounding box regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12993–13000, 2020.

[126] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-nms – improving object detection with one line of code. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[127] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv :1712.04621*, 2017.

[128] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1) :43–76, 2020.

[129] Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[130] Fengxiang He, Tongliang Liu, and Dacheng Tao. Control batch size and learning rate to generalize well : Theoretical and empirical evidence. *Advances in Neural Information Processing Systems*, 32, 2019.

[131] Leslie N Smith. A disciplined approach to neural network hyper-parameters : Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv :1803.09820*, 2018.

[132] Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE, 2017.

[133] Laurent Guigues, Jean Pierre Cocquerez, and Hervé Le Men. Scale-sets image analysis. *International Journal of Computer Vision*, 68(3) :289–317, 2006.

[134] Vijay S Kumar, Tahsin Kurc, Jun Kong, Umit Catalyurek, Metin Gurcan, and Joel Saltz. Performance vs. accuracy trade-offs for large-scale image analysis applications. In *2007 IEEE International Conference on Cluster Computing*, pages 100–109. IEEE, 2007.

[135] Adam Smith. Introducing zoomify image, 2007.

[136] Open layers. `https://openlayers.org/`, 2006.

[137] Vladimir Agafonkin. Leaflet. `https://leafletjs.com/`, 2010.

[138] CodePlex Foundation and OpenSeadragon contributors. Openseadragon. `https://openseadragon.github.io/`, 2011.

[139] Jeff Forcier, Paul Bissex, and Wesley J Chun. *Python web development with Django*. Addison-Wesley Professional, 2008.

[140] Django Software Foundation. Django project. https://www.djangoproject.com/, 2005.

Koubouratou IDJATON

Analyse d'images et apprentissage machine pour la détection des altérations des pierres des monuments historiques.

La préservation des monuments historiques exige une surveillance de leur état pour assurer la sécurité des visiteurs et planifier au mieux les opérations de restauration. Cette surveillance repose essentiellement sur une observation visuelle réalisée par des experts sur site et un lourd et fastidieux travail d'inventaire et d'annotations manuelles sur des photos. Cette thèse propose de nouveaux outils d'aide au diagnostic assisté par ordinateur basés sur l'intelligence artificielle pour réaliser une détection automatique des altérations des pierres dans des images couleurs. Une plateforme Web opérationnelle permet aux experts d'appliquer les algorithmes proposés et de visualiser les résultats de segmentation pierre-à-pierre et de détection des altérations des pierres via une interface conviviale, pour faciliter leur diagnostic sur les façades des châteaux. Ces contributions ouvrent la voie au développement d'un outil convivial multi acteurs pour un diagnostic plus précis des grands monuments.

Mots clés : Analyse d'images couleur, vision 3D, intelligence du patrimoine, détection d'altération, segmentation de pierre, apprentissage machine

Image processing and machine learning for stone alterations detection in historical monuments

The preservation of historic monuments requires monitoring their condition to ensure the safety of visitors and to plan restoration operations as well as possible. This monitoring relies mainly on visual observation by experts on site and a heavy and tedious work of inventory and manual annotations on photos. This thesis proposes new computer-assisted diagnostic tools based on artificial intelligence to perform automatic detection of stone alterations in color images. An operational web platform allows experts to apply the proposed algorithms and to visualize the results of stone-to-stone segmentation and stone alteration detection via a user-friendly interface, to facilitate their diagnosis on castle facades. These contributions pave the way for the development of a user-friendly multi-actor tool for a more accurate diagnosis of large monuments.

Keywords : 3D models, Image processing, Cultural Heritage, alteration detection, stone segmentation, machine learning

Laboratoire PRISME, Site de Polytech Galilée, 12 Rue de Blois 45067 Orléans, France