

Rapport de fin de mission

Contrat de professionnalisation



Orange Labs

2 Avenue de Belle Fontaine
35510 Cesson-Sévigné

Elève-ingénieur

Benoît Quero

Spécialité Imagerie
Numérique

Tuteur entreprise

Christophe Daguet

Ingénieur de recherche

Tutrice académique

Ewa Kijak

Enseignante-chercheuse

Année universitaire 2019-2020

Remerciements

Je tiens à remercier toutes les personnes qui ont contribué à mon intégration et à la réussite de mon expérience chez Orange Labs à Cesson-Sévigné.

Je voudrais tout d'abord remercier mon tuteur Christophe Daguet, ingénieur de recherche au sein de l'équipe Content Video Audio, qui m'a choisi pour intégrer son équipe projet et participer à ses activités de recherche. Il a su et sait me faire confiance dans les différentes étapes du projet et me donne une certaine indépendance dans les différentes tâches.

Je tiens également à remercier Matthieu Gendrin, Sylvain Kervadec, Sandra Nabil et Thomas Houé, avec qui je travaille sur la thématique de l'acquisition et de la modélisation de scène en trois dimensions, ainsi que mon manager Didier Tattevin et le reste de l'équipe Content Video Audio, pour leur accueil et leur bonne humeur.

Enfin, je remercie Ewa Kijak, ma tutrice ESIR, pour son suivi tout au long de cette année et ses retours sur les différents rapports.

Table des matières

Introduction	1
I. Présentation de l'entreprise	2
A. Le groupe Orange	2
B. Orange Labs : Recherche et Développement	3
C. L'équipe Content Video Audio	3
D. Veille sociétale	4
II. Mes missions au sein de l'équipe	5
A. Contexte	5
B. Structure from Motion et Multi-View Stereo	7
C. Fonctionnalités apportées à ces librairies	11
D. Alternatives au Multi-View Stereo pour le calcul des cartes de profondeur	15
E. Conclusion sur le projet	21
Conclusion	22
Bibliographie	23

Lexique

Nuage de points : ensemble de points dans un espace en trois dimensions, représentant généralement un objet 3D.

Mesh : objet 3D constitué de sommets, d'arêtes et de faces. Les faces sont représentées la plupart du temps par des triangles.

Réseau de neurones : ensemble de neurones artificiels interconnectés permettant de résoudre des problèmes complexes comme la reconnaissance d'objet dans une image. Le réseau de neurones a la capacité d'apprendre lui-même en modifiant ses poids en fonction des données qu'il a à traiter.

Dataset : jeu de données permettant à un réseau de neurones de s'entraîner, c'est-à-dire à apprendre à résoudre le problème voulu.

Carte de profondeur : image représentant par différentes couleurs ou nuances de gris la distance par rapport à la caméra pour chaque pixel.

Carte de disparité : image représentant par différentes couleurs ou nuances de gris la distance entre les pixels de deux images d'un même objet de point de vue différent. Elle est similaire à la carte de profondeur mais ne prend pas en compte la distance avec la caméra. Elle ne donne donc pas la profondeur réelle.

Introduction

Le contrat de professionnalisation que je réalise actuellement pour ma troisième et dernière année à l'Ecole Supérieure d'Ingénieurs de Rennes se déroule à Orange Labs à Cesson-Sévigné, depuis le 30 septembre 2019 et jusqu'au 30 septembre 2020.

Orange est une entreprise avec une grande renommée et est très présente dans le bassin rennais. Cela m'a tout d'abord amené à postuler à Orange Labs pour mon stage de deuxième année sur un sujet qui me semblait intéressant. Bien que ma candidature n'avait pas été retenue, j'ai pu garder contact et ainsi avoir la possibilité d'intégrer les équipes de recherche d'Orange pour une année d'alternance sur un sujet tout aussi passionnant.

Dans un premier temps, nous présenterons le groupe Orange, ses activités, sa place sur le marché, sa division de Recherche et Développement ainsi qu'une veille sociétale. Puis, dans un second temps, nous verrons mon rôle au sein de l'équipe et les missions que j'ai réalisées.

I. Présentation de l'entreprise

A. Le groupe Orange

Le groupe Orange est l'un des principaux opérateurs de télécommunications en France et dans le monde. L'entreprise est présente dans 26 pays et employait en 2019 148 000 personnes, dont 88 000 en France. Son nombre de clients est de 266 millions dans le monde et elle a réalisé cette année là un chiffre d'affaire de 42 milliards d'euros.

En 2000, la société France Télécom, créée en 1988, rachète en majorité l'entreprise britannique Orange Plc et fusionne alors ses activités mobiles en une entité nommée "Orange SA". Afin d'homogénéiser le nom des services de l'entreprise, la ligne fixe de France Télécom est rebaptisée Orange en 2012, ainsi que son nom juridique en 2013.

Les activités du groupe sont présentes dans cinq principaux secteurs :

- les services de communication résidentiels : téléphonie fixe, internet (ADSL, fibre optique), télévision numérique, contenus multimédia (vidéo à la demande), téléphonie IP et visiophonie,
- les services de communication personnels, à savoir la téléphonie mobile,
- les services de communication d'entreprise, grâce à Orange Business Services,
- les services bancaires, depuis 2017, avec Orange Bank,
- les services domotiques, avec les offres Maison Protégée et Maison Connectée.

Orange opte sur une diversification de ses activités en investissant dans les contenus, le multimédia, les services ou encore les solutions télécoms à destination des entreprises. Le groupe investit beaucoup à l'étranger, notamment en Afrique et au Moyen-Orient, afin de limiter sa dépendance au marché français.

En 2019, Orange avait 266 millions de clients dans le monde (Europe, Afrique, Antilles, Moyen-Orient), dont 201 millions en téléphonie mobile et 18,2 millions en haut débit fixe. L'entreprise est leader sur la fibre en Europe avec 40 millions de foyers raccordables.

Orange dispose du réseau haut débit le plus étendu de France : 99,6% de la population vit dans une zone éligible à l'ADSL d'Orange. L'opérateur possède également le plus grand réseau de fibre optique sur le territoire français : 14,6 millions de foyers raccordables. En téléphonie mobile, Orange dispose de la meilleure couverture internet mobile : 99% de la population en bénéficia.

Ses principaux concurrents dans le secteur des télécommunications en France sont SFR (2ème opérateur en France avec environ 20% de part de marché), Free (6,5 millions d'abonnés internet) et Bouygues Telecom (12% de d'augmentation des ventes en 2019).

B. Orange Labs : Recherche et Développement

Bien qu'Orange soit avant tout une entreprise de télécommunication, l'entreprise investit fortement dans la Recherche et le Développement, se hissant même à la 19ème place des entreprises les plus innovantes au monde en 2018, selon un classement élaboré par le Boston Consulting Group.

La Technology and Global Innovation (TGI) est la division Recherche et Développement d'Orange. Le groupe y a consacré environ 700 millions d'euros en 2019. Cette division compte près de 8000 employés répartis dans 19 centres, dont Orange Labs Cesson-Sévigné.

Les activités de recherche du groupe viennent soutenir les activités actuelles d'Orange mais permettent aussi de les faire évoluer. On retrouve par exemple parmi elles des travaux sur la télévision Ultra Haute Définition, la fibre optique, la 5G et les objets connectés.

Les activités de recherche peuvent être divisées en 9 domaines :

vie personnelle numérique	société numérique	pays émergents numériques
entreprises numériques	connectivité ambiante	infrastructure logicielle
Internet des Objets	données et connaissances	confiance et sécurité

Figure 1 : les 9 domaines de recherche d'Orange Labs

Mes activités chez Orange Labs, comme celles de mon équipe, ont pour but de développer de nouveaux algorithmes et technologies se basant beaucoup sur l'Intelligence Artificielle. Elles se placent donc le domaine « données et connaissances ».

C. L'équipe Content Video Audio

Mes activités à Orange se déroulent au sein de la direction Home d'Orange Labs Services, et plus précisément dans l'équipe Content Video Audio (CVA) qui a pour but de mener des activités de recherche en technologie sur les contenus

audiovisuels, comme la compression numérique ou l'étude et la mise en oeuvre de formats vidéo immersifs. L'équipe est composée d'une trentaine de personnes, comptant parmi elle des managers, des chercheurs, des ingénieurs, des doctorants ou post-doctorants, des alternants et des stagiaires.

Je partage la thématique de l'acquisition et de la modélisation de scène en trois dimensions avec Christophe Daguet, Sylvain Kervadec, ingénieurs de recherche, Matthieu Gendrin, développeur, Thomas Houé, stagiaire, et Sandra Nabil, post-doctorante jusqu'à mars dernier.

Mes travaux s'inscrivent dans le projet ICCE (Immersive Communication, sound Capture & Entertainment). Ce projet vise à étudier la captation sonore et la communication immersive, pour les rendre plus propres, interactives et économies en bande passante. Ces technologies pourraient être intégrées sur des produits de communication ou de divertissement comme l'assistant vocal Djingo ou les technologies immersives comme la réalité virtuelle 360. Elles ont comme clients potentiels tout type d'entreprises ainsi que les particuliers.

D. Veille sociétale

Pour répondre aux enjeux sociaux, sociétaux et environnementaux, Orange s'engage à respecter cinq engagements majeurs dans tous les pays où le groupe est implanté :

- ▶ Œuvrer pour le respect des libertés fondamentales : droits humains, liberté d'expression, protection des données personnelles et de la vie privée.
- ▶ Favoriser l'inclusion numérique et le développement des territoires : fournir l'accès au plus grand nombre et lutter contre l'exclusion numérique
- ▶ Net zéro carbone d'ici 2040 : agir en faveur de la transition énergétique et écologique
- ▶ Proposer des produits et des services responsables
- ▶ Développer l'employabilité et le bien-être au travail

Orange a signé en 2014 l'accord mondial Santé-Sécurité afin d'intégrer la SST dans tous les pays où Orange est présent. Il a pour but de d'améliorer les conditions de santé, de sécurité, et de qualité de vie au travail des salariés :

- déploiement d'un plan de formation des Comités Santé-Sécurité,
- état des lieux mené par des représentants de la Direction et des organisations syndicales dans les filiales du groupe,
- création d'indicateurs (KPI) permettant le suivi de l'accord, tels que le nombre d'accidents de travail ou de trajet ayant entraîné un arrêt de travail, etc.

II. Mes missions au sein de l'équipe

A. Contexte

J'ai intégré l'équipe CVA afin de contribuer aux travaux de recherche autour des formats vidéo immersifs et plus particulièrement sur la mise au point d'un système de captation multi-caméras, permettant l'acquisition de contenus vidéo ainsi que d'informations complémentaires sur le relief des scènes filmées. Ces informations servent ensuite à la reconstruction d'une modélisation (nuage de points, mesh) permettant à l'utilisateur final de visualiser la scène sous différents points de vue.

Mon rôle était de tester, comprendre et améliorer des logiciels open-source de reconstruction 3D (des librairies de Structure from Motion et Multi-View Stereo principalement) pour les adapter à un système multi-caméras permettant d'isoler et de reconstruire en 3D une ou des personnes dans une scène. Cette technologie permettrait de réaliser des vidéos 3D de personnes, comme des sportifs, que l'on pourrait visualiser avec un casque de réalité virtuelle par exemple.

Un studio est actuellement en cours de construction dans le nouveau bâtiment Orange Atalante. Il sera équipé d'un système d'une vingtaine de caméras permettant de capter tous les points de vues d'une personne ou un objet.



Figure 2 : Futur studio d'Orange Atalante

Pour obtenir des informations sur le relief d'une scène, il existe plusieurs technologies possibles. Le LiDAR utilise le principe de télédétection laser par télémétrie. Il envoie une impulsion de lumière et mesure le temps que cette impulsion met pour revenir jusqu'au récepteur. Les scanners laser 3D embarquent cette technologie et permettent de reconstruire des scènes en 3D. L'inconvénient de cette technique est qu'elle n'offre pas encore assez de détails sur les scènes en mouvement.

Pour calculer la profondeur d'une scène, il est aussi possible d'utiliser des caméras de profondeur comme Intel RealSense ou Microsoft Azure Kinect. Ces caméras sont équipées de caméras RGB, de capteurs infrarouges et d'émetteurs infrarouges. L'émetteur va projeter sur la scène des points lumineux (en infrarouge) qui seront répartis géométriquement les uns par rapport aux autres. En fonction des positions de ces points, perçues différemment par les capteurs infrarouges, la disparité de chacun de ces points est calculée, et on en déduit la profondeur de la scène. Bien que les résultats de ces caméras sont plutôt bons, une trop forte luminosité dans la scène peut perturber les caméras et donner des résultats trop bruités.

Ces deux techniques demandent donc un matériel spécifique. Pour reconstruire une scène en 3D seulement à partir d'images de différents points de vue, nous allons utiliser deux techniques de photogrammétrie : Structure from Motion (SfM) et Multi-View Stereo (MVS). La première permet d'obtenir un premier nuage de points (dit « sparse ») et la seconde un nuage de points plus complet (dit « dense »). MVS permet également d'obtenir un mesh à partir du nuage de points dense. Aujourd'hui, elles sont principalement appliquées aux domaines des bâtiments, à la topographie ou à la reconstruction d'objets en 3D.

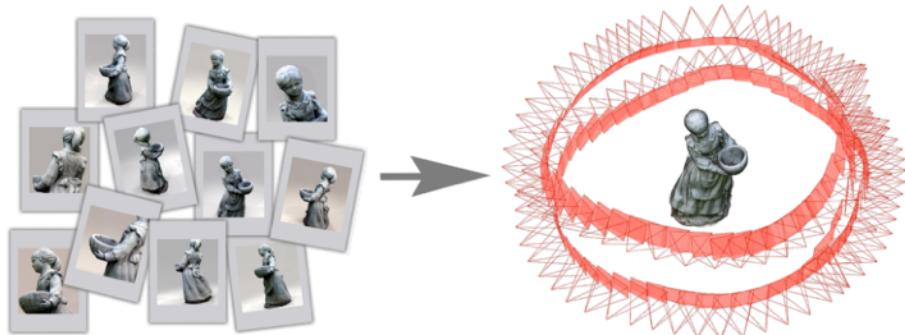


Figure 3: Principe du Structure from Motion et du Multi-View Stereo : partir d'images 2D pour créer un objet 3D¹

¹ source de l'image : Bianco, S.; Ciocca, G.; Marelli, D. Evaluating the Performance of Structure from Motion Pipelines. J. Imaging 2018, 4, 98.



Figure 4 : Exemples de nuages de points obtenus grâce au Structure from Motion et au Multi-View Stereo²

L'avantage de ces techniques, par rapport à l'utilisation d'un LiDAR ou de caméras de profondeur comme Realsense ou Kinect, est qu'elles n'ont besoin que de photographies de la scène, prises par n'importe quel appareil photo. Elles sont donc plus faciles d'utilisation et bien moins coûteuses. En revanche, leurs inconvénients sont le manque de points dans les régions homogènes (zones avec peu de variation d'intensité lumineuse) des objets et la difficulté de reconstruire une surface réfléchissante ou transparente. De plus, les points du nuage de points final ne sont pas indexés, c'est-à-dire que l'on ne peut pas isoler rapidement un objet ou une personne de la scène.

Les séquences d'images multi-vues qui ont servi de test dans ce rapport sont composées de 4 points de vue différents et ont été prises dans les locaux d'Orange Labs (cf annexes).

B. Structure from Motion et Multi-View Stereo

Il existe plusieurs solutions open-source de Structure from Motion et Multi-View Stereo : Colmap, PMVS2, VisualSfM, OpenMVG + OpenMVS...

Après de nombreux tests, notre choix s'est porté sur OpenMVG et OpenMVS. Ils donnent de meilleurs résultats, leurs codes sont facilement réutilisables et très bien documentés.

1. OpenMVG : Structure from Motion

OpenMVG s'occupe la partie Structure from Motion. Cette librairie prend en entrée une séquence de photos d'une scène, d'un bâtiment, d'un objet ou d'une personne,

² source de l'image : Li, Y.; Li, Z. A Multi-View Stereo Algorithm Based on Homogeneous Direct Spatial Expansion with Improved Reconstruction Accuracy and Completeness. *Appl. Sci.* 2017, 7, 446.

et produit en sortie un nuage de points « sparse » ainsi que les paramètres intrinsèques et extrinsèques des caméras, c'est-à-dire leur distance focale, les coordonnées de la projection du centre optique, leur positions dans l'espace, etc.

L'algorithme de Structure from Motion est le suivant (voir *Figure 5*):

1. détection de points d'intérêt (features) dans chaque image.

L'algorithme SIFT (Scale-invariant feature transform) permet de détecter des informations telles que les angles et les bords dans les images. Les descripteurs SIFT sont invariants au zoom, au cadrage, à l'angle d'observation et à l'exposition, ce qui permet de détecter les même points d'intérêt dans chacune des images.

2. mise en correspondance des points d'intérêt entre les images.

Les descripteurs SIFT détectés précédemment sont ici comparés et associés si similaires. Pour supprimer les points d'intérêt mal associés, l'algorithme RANSAC est utilisé (cf Annexes). Pour avoir un maximum de correspondances, il faut donc avoir des photos ayant un point de vue suffisamment proche.

3. estimation des paramètres caméras.

Les paramètres intrinsèques et extrinsèques sont estimés grâce au calcul de la matrice Fondamental et de la matrice Essentiel (cf annexe) à partir des correspondances trouvées à l'étape précédente.

4. génération d'un nuage de points sparse par triangulation

Pour chaque point d'intérêt ayant une correspondance dans d'autres images, on réalise une triangulation dans l'espace 3D, c'est-à-dire que l'on va projeter le point depuis chaque caméra où il apparaît et le point d'intersection des directions donnera ses coordonnées dans l'espace.

5. bundle adjustment : correction des paramètres caméras et des points 3D

Dans cette étape, les paramètres caméras et les points 3D sont ajustés en minimisant l'erreur de reprojection entre les différents points de vue.

Nous obtenons donc un nuage de points composé des points d'intérêt détectés dans les images des différents points de vue de la scène. (voir *Figure 6*)

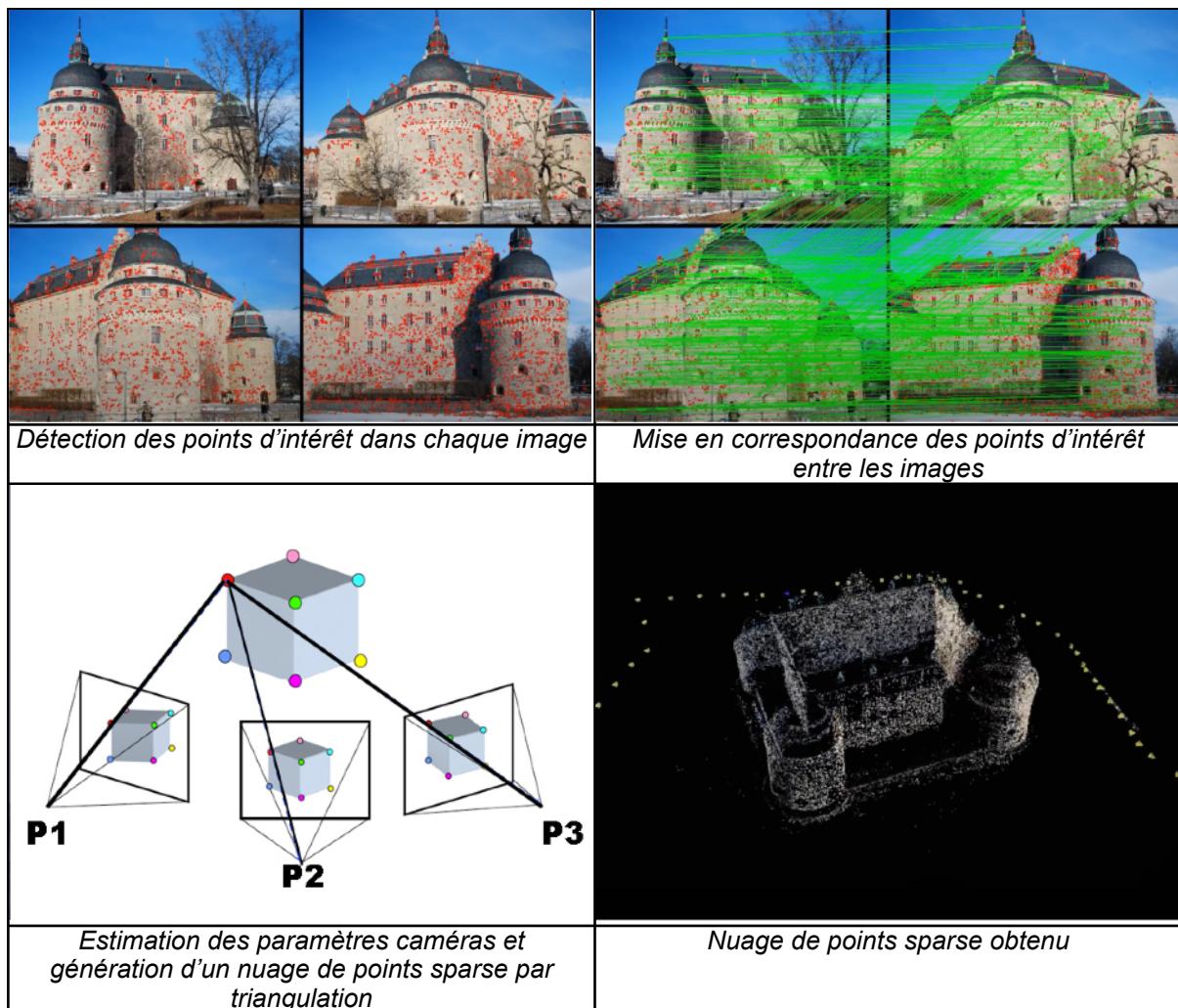


Figure 5 : Étapes du Structure From Motion³

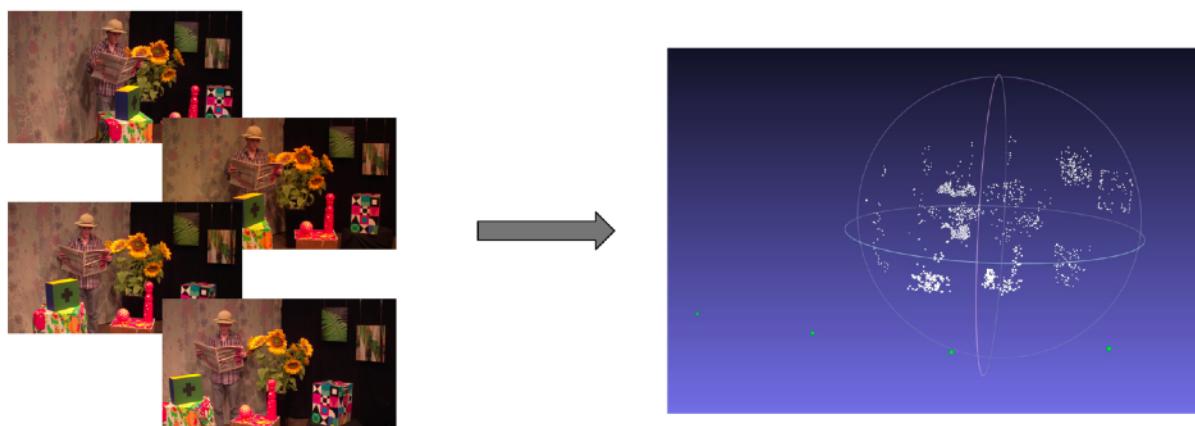


Figure 6 : Nuage de points « sparse » obtenu avec OpenMVG sur une de nos séquences (4 points de vue) incluant une personne (1 600 points)

³ Source des images : https://fr.wikipedia.org/wiki/Structure_from_motion, <https://www.youtube.com/watch?v=i7ierVkJYa8>

2. OpenMVS : Multi-View Stereo

L'algorithme de Structure from Motion permet de calculer les informations nécessaires au Multi-View Stereo [9], qui va recréer notre scène en 3D sous forme d'un nuage de point dense ou d'un mesh. La qualité de reconstruction dépend énormément du Multi-View Stereo, et de nouveaux algorithmes voient le jour chaque année. Il y a plusieurs catégories de MVS : les méthodes basées sur la propagation qui étendent le nuage de points autour des zones d'intérêts, les méthodes basées voxel (pixel en 3D) qui transforment l'espace 3D en grille régulière et estiment si chaque voxel appartient à la surface du modèle, ou encore les méthodes se basant sur la minimisation d'une fonction pour affiner petit à petit le modèle 3D final. Mais la méthode la plus efficace reste celle de la fusion de cartes de profondeur. Pour chaque point de vue, des cartes de profondeur sont estimées, puis filtrées, pour être enfin fusionnées afin de créer un nuage de points.

OpenMVS est une implémentation logicielle du Multi-View Stereo. Il prend en entrée le nuage de points sparse et les paramètres caméras calculés avec OpenMVG. OpenMVS permet de générer en sortie un nuage de points dense et un mesh. La méthode utilisée par OpenMVS est celle de la fusion de carte de profondeur.

Les étapes d'OpenMVS sont les suivantes :

1. calcul des cartes de profondeurs pour chaque image

- A. initialisation des cartes de profondeur avec le nuage de points sparse

Le nuage de points sparse nous donne déjà des informations de profondeur. Les cartes de profondeur de chaque vue peuvent donc être initialisées grâce à celui-ci.

- B. algorithme du PatchMatch Stereo

Le PatchMatch Stereo [10] va ensuite calculer la carte de profondeur en projetant des patchs (petites fenêtres de pixels définis par leur centre et leur normale) dans la scène depuis un des points de vue. L'algorithme ajuste ensuite la position et l'orientation du patch pour que celui-ci corresponde depuis chaque point de vue. On répète l'opération pour chaque point d'intérêt du SfM et on en créer de nouveaux en propageant ceux que l'on vient de calculer. (cf annexes)

2. Filtrage des cartes de profondeur

Les cartes de profondeur ainsi obtenues sont ensuite filtrées pour lisser les surfaces planes et supprimer les valeurs aberrantes. (voir *Figure 7*)

3. Fusion des cartes de profondeurs

Les cartes sont enfin fusionnées pour obtenir un nuage de points dense. (voir *Figure 8*)

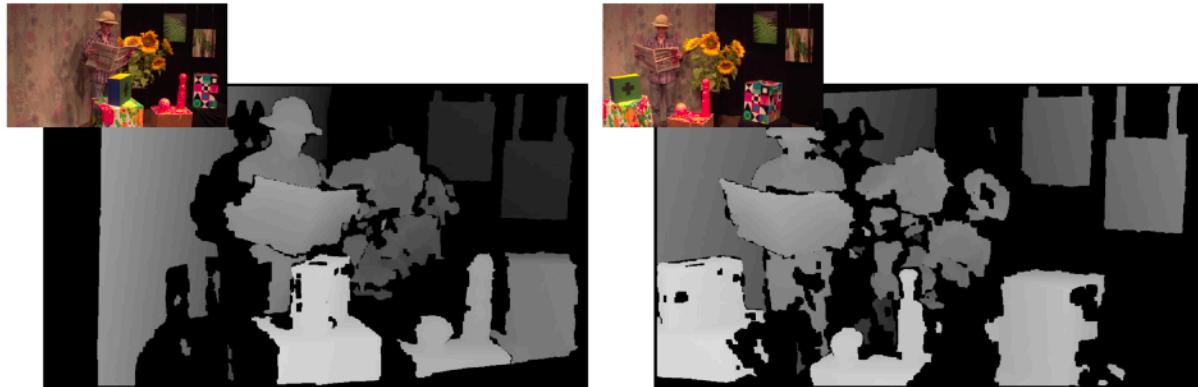


Figure 7 : Cartes de profondeur obtenues avec OpenMVS

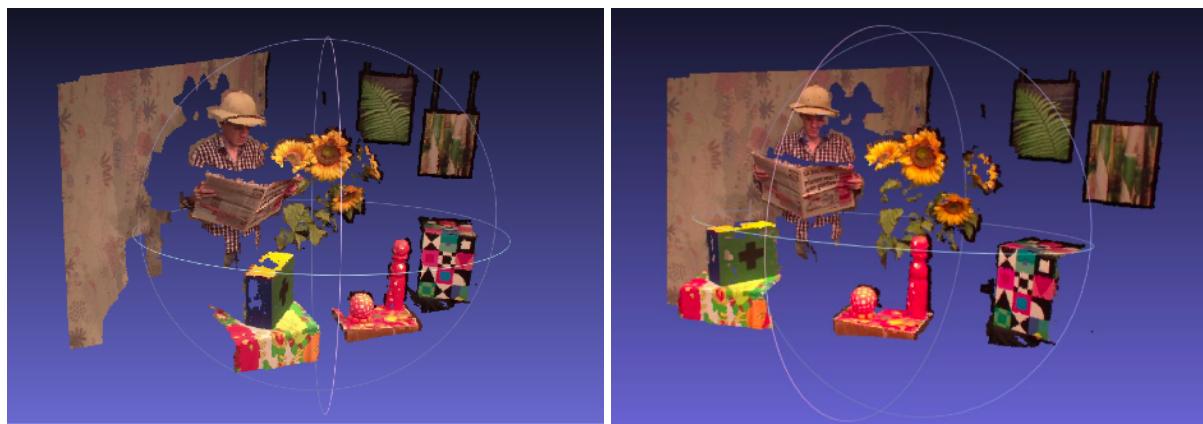


Figure 8 : Nuage de points « dense » obtenu avec OpenMVS (240 000 points)

C. Fonctionnalités apportées à ces librairies

Pour pouvoir utiliser ces librairies dans notre contexte, il faut leur apporter quelques modifications.

1. Détection et segmentation de personnes

Le nuage de points dense obtenu grâce à OpenMVG et OpenMVS n'est pas indexé, nous ne pouvons donc pas isoler facilement une personne dans une scène. Pour cela, nous avons décidé d'utiliser Detectron2. Detectron2 est une librairie de Facebook AI Research implémentant l'état de l'art des algorithmes de détection et de segmentation basés sur du Deep-Learning. On y retrouve des modèles comme Faster R-CNN, RetinaNet ou Mask R-CNN.

Mask R-CNN [11] est une extension de Faster R-CNN. Faster R-CNN est un réseau de neurones convolutifs permettant de détecter un objet, l'entourer par une bounding box (boîte englobante) et indiquer à quelle classe il appartient (voiture, personne, etc). Mask R-CNN permet en plus d'estimer un masque, c'est-à-dire de segmenter l'objet au pixel près (cf annexes). C'est celui-ci qui va nous intéresser.



Figure 9 : Résultat de Mask R-CNN

Voici la méthode utilisée pour créer un nuage de points avec une personne isolée :

1. Lancement du SfM avec comme images sources les images complètes (sans segmentation)
2. Segmentation et isolation des personnes dans chaque image
3. Lancement du MVS avec comme images sources les images segmentées, mais avec les paramètres caméras et le nuage de points sparse obtenus avec les images complètes

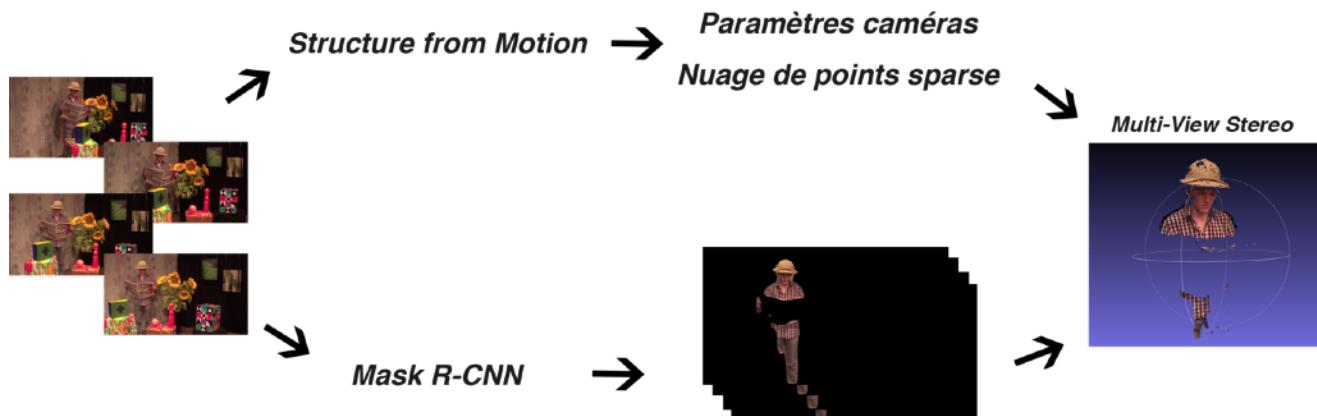


Figure 10 : Schéma résumant la méthode utilisée pour créer un nuage de points avec une personne isolée

Il est théoriquement possible de faire les étapes SfM et MVS directement avec les images segmentées, mais en pratique, moins de points d'intérêt sont trouvés et les paramètres caméras sont donc moins bien estimés. Cela aboutit à un moins bon nuage de points final.

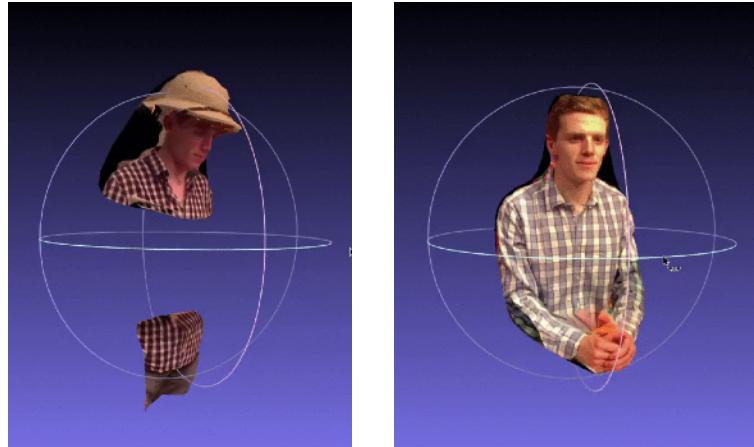


Figure 11 : Meshs obtenus avec OpenMVS + Mask RCNN

2. Estimation et réutilisation des paramètres caméras

La deuxième fonctionnalité ajoutée part du principe que les caméras utilisées dans le futur studio seront fixes. Il n'y a donc pas besoin de calculer les paramètres caméras à chaque reconstruction.

De plus, nous avons pu remarquer que si la scène capturée était pauvre en détails, les paramètres intrinsèques et extrinsèques des caméras étaient mal estimés. La solution est donc de calculer les paramètres caméras sur des scènes riches en détails pour pouvoir par la suite les réutiliser sur des scènes pauvres en détails.

OpenMVG offre déjà la possibilité de lancer le Structure from Motion à partir de paramètres caméras connus. Il a donc suffit de créer une fonction permettant de sauvegarder les paramètres calculés, et une autre permettant de les relire.

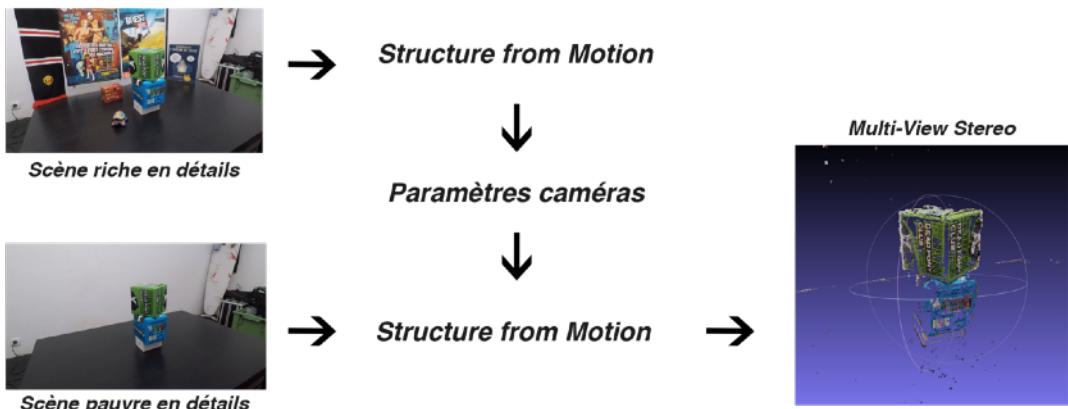


Figure 12 : Schéma résumant le principe de réutilisation des paramètres caméras

Cette modification permet de gagner du temps quand il y a beaucoup de caméras, car leurs paramètres ne sont pas recalculés, et d'améliorer légèrement le nuage de points final. Par exemple, on peut voir sur la *Figure 13* que la forme du cube est mieux reconstruite que sur la *Figure 14*.

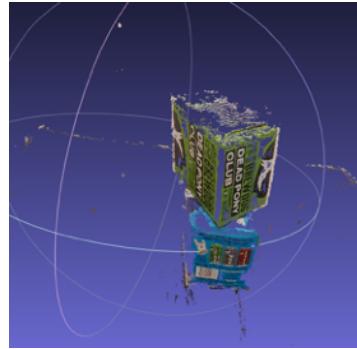


Figure 13 : Reconstruction avec paramètres caméras calculés sur une scène pauvre en détails

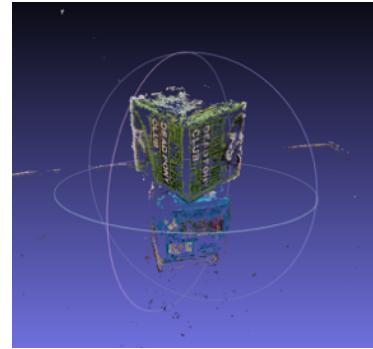


Figure 14 : Reconstruction avec paramètres caméras calculés sur une scène riche en détails : forme plus fidèle

3. Multi-View Stereo multi-pass multi-résolution

Dans le but d'améliorer la qualité de la reconstruction finale, nous avons imaginé un OpenMVS multi-pass multi-résolution.

Multi-pass signifie que nous lançons plusieurs fois OpenMVS :

- une première fois normalement, à partir du nuage de points sparse obtenu par OpenMVG,
- une seconde fois, à partir du nuage de points dense obtenu précédemment.

Multi-résolution signifie que :

- à la première passe, la résolution des images sources est divisée par 4,
- à la deuxième passe, on garde la résolution originale.

En théorie, comme OpenMVS part du nuage de points sparse d'OpenMVG pour calculer les cartes de profondeur et le nuage de points dense, la deuxième passe devrait augmenter le nombre de points et donc réduire les trous observés. En pratique, les trous sont toujours présents, mais nous avons quand même pu remarquer que la répartition des points étaient plus homogènes et qu'il y avait légèrement moins de bruit (moins de points aberrants). Sur la *Figure 16*, on peut aussi remarquer que certains trous se sont agrandis sur après la seconde passe.

Nous avons également testé en multi-résolution pour que le nombre de points du point cloud dense de la première passe soit moins important et ait une probabilité

d'outliers plus faible, pour ne pas influer négativement sur la deuxième passe. Ceci n'a pas non plus amélioré significativement le résultat.



Figure 15 : Nuage de points dense après la 1ère passe, 1/4 de résolution



Figure 16 : Nuage de points dense après la 2nde passe, pleine résolution

D. Alternatives au Multi-View Stereo pour le calcul des cartes de profondeur

Comme nous avons pu le voir, le Multi-View Stereo permet d'obtenir de bons résultats quand les éléments de la scène sont texturés et riches en détails. En revanche, quand il y a peu de détails, dans des zones dites homogènes, il est plus difficile pour l'algorithme de déterminer une valeur de profondeur. Il y a donc des trous dans les cartes de profondeur et dans le nuage de points.

Pour pallier à ces différents défauts, j'ai essayé quelques moyens alternatifs de calcul et de complétion de cartes de profondeur.

1. Calcul des cartes de profondeur par Deep Learning

Malgré l'efficacité des méthodes de Multi-View Stereo dites « conventionnelle », les méthodes basées deep-learning prennent de plus en plus d'importance et permettent d'obtenir des résultats visuellement très fidèles à la réalité. J'ai pu tester quelques réseaux de neurones : R-MVSNet et DeepMVS. Ces réseaux prennent en entrée les images des différentes vues et la position des caméras, calculées précédemment par Structure from Motion.

R-MVSNet [13] et DeepMVS [11] utilisent des réseaux de neurones convolutifs, qui sont spécialement conçus pour traiter des images en entrée. Ils ont la particularité d'être composés d'un extracteur de features. Plus de détails sur MVSNet [12], R-MVSNet et DeepMVS en annexe.

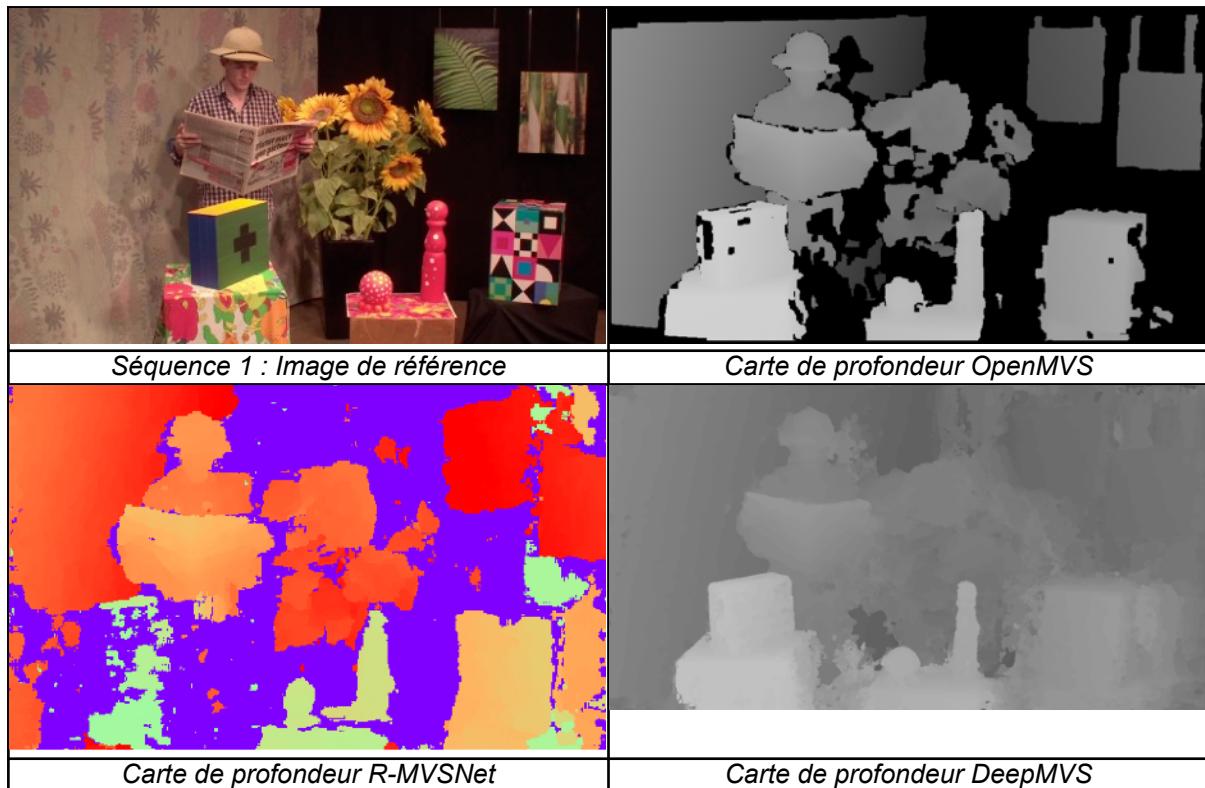


Figure 17 : Résultats obtenus par les différentes méthodes de calcul de carte de profondeur - séquence 1

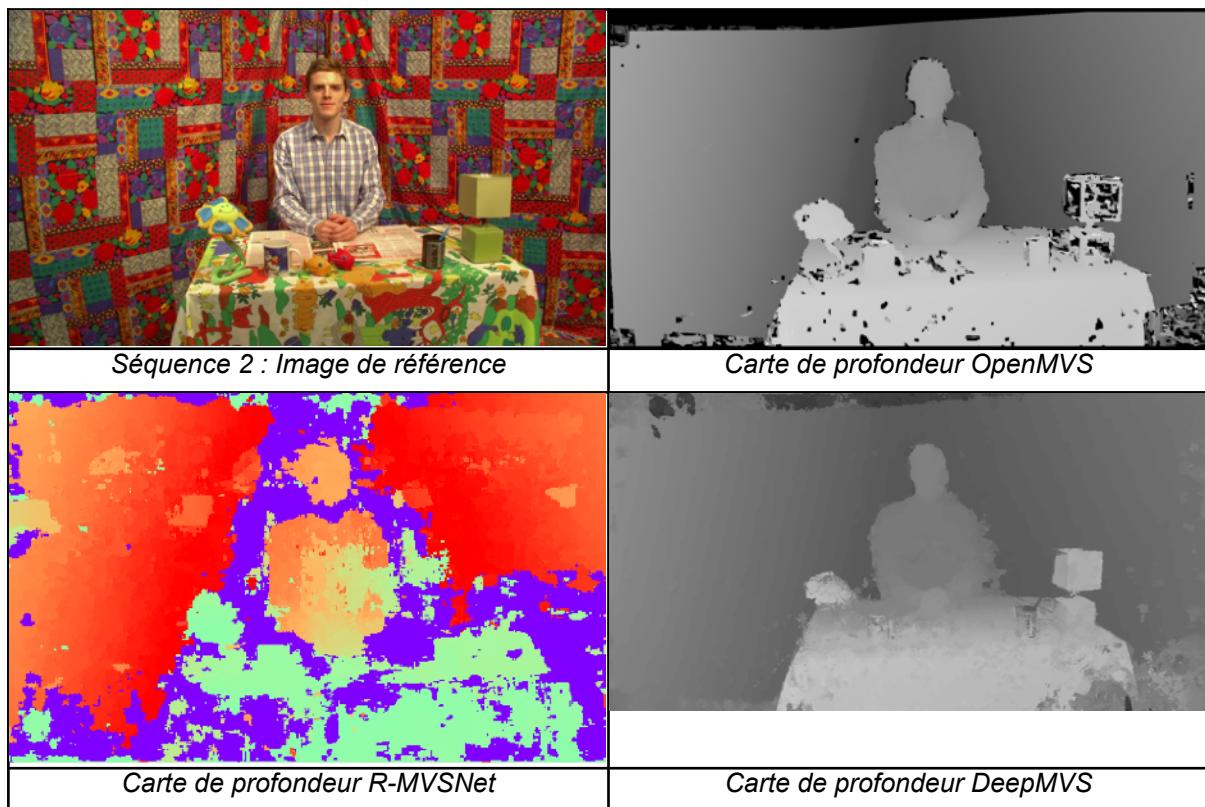


Figure 18 : Résultats obtenus par les différentes méthodes de calcul de carte de profondeur - séquence 2

On peut observer que R-MVSNet ne donne pas de bons résultats sur les séquences d'images testées. Il y a beaucoup de trous, de bruit et les contours des objets ne sont pas bien respectés. DeepMVS donne de meilleurs résultats : il n'y a pas de trous, par rapport à OpenMVS et R-MVSNet, mais les profondeurs et les contours ne semblent pas toujours bien respectés. On peut le voir sur la *Figure 17* autour du chapeau notamment. On remarque également que R-MVSNet et DeepMVS estiment moins bien les zones planes que OpenMVS.

Notre but est de reconstruire en 3D des personnes, or R-MVSNet et DeepMVS sont entraînés sur des bâtiments, extérieurs et intérieurs, et des objets du quotidien, ce qui peut expliquer pourquoi la qualité des résultats n'est pas aussi bonne que prévu.

2. Complétion des cartes de profondeur d'OpenMVS

Comme nous avons pu le voir, les méthodes de deep-learning ne permettent pas de tout le temps obtenir des résultats satisfaisants. Les techniques de complétion de carte de profondeur (depth completion) sont donc un compromis entre méthodes conventionnelles et méthodes basées deep-learning.

Les caméras de profondeur ou les scanners laser sont de plus en plus utilisés pour estimer les reliefs d'une scène, dans des domaines comme la conduite autonome ou la robotique. Mais ces appareils ne permettent pas d'obtenir des cartes de profondeur très précises et sont souvent incomplètes. De nombreux chercheurs se sont penchés sur le sujet et plusieurs techniques de complétion sont apparues pour améliorer la qualité des cartes de profondeur.



Figure 19 : Exemple de l'utilisation de depth completion⁴

La majorité des solutions se basent sur la carte de profondeur « sparse » obtenue avec LiDAR ou caméra de profondeur et son image RGB associée. Il existe des techniques d'inpainting, en observant les valeurs des pixels autour des trous pour les boucher, et de nombreuses techniques basées sur du deep-learning qui permettent d'obtenir de très bons résultats.

Dans notre cas, nous n'utilisons pas de caméra de profondeur ni de scanner 3D pour obtenir des cartes de profondeur, mais le Multi-View Stereo. Les cartes de

⁴ sources des images : Ma, F., Cavalheiro, G., & Karaman, S. (2019). Self-Supervised Sparse-to-Dense: Self-Supervised Depth Completion from LiDAR and Monocular Camera. 2019 International Conference on Robotics and Automation (ICRA), 3288-3295.

profondeurs obtenues avec MVS sont souvent plus bruitées, et les techniques citées ci-dessus ne sont pas prévues pour cela. De plus, elles ne se basent que sur une carte de profondeur et son image RGB associée, et non sur les images des différentes vues.

Bien que non adapté au Multi-View Stereo, j'ai pu tester sur nos cartes de profondeur obtenues avec OpenMVS le code basé sur le papier *Deep depth completion of a single RGB-D image* [14]. Celui-ci propose une méthode de complétion basée sur un réseau de deep-learning. À partir de l'image RGB, un réseau de neurones convolutifs basé sur VGG-16 va prédire la normale à la surface pour chaque pixel ainsi que les bords des objets. Ensuite, ces deux informations sont combinées avec la carte de profondeur obtenue avec OpenMVS dans une fonction de coût. Cette fonction est minimisée en donnant plus ou moins d'importance à ces trois paramètres : respect de la profondeur d'origine, respect de l'homogénéité entre pixels voisins et respect de la normale d'origine.

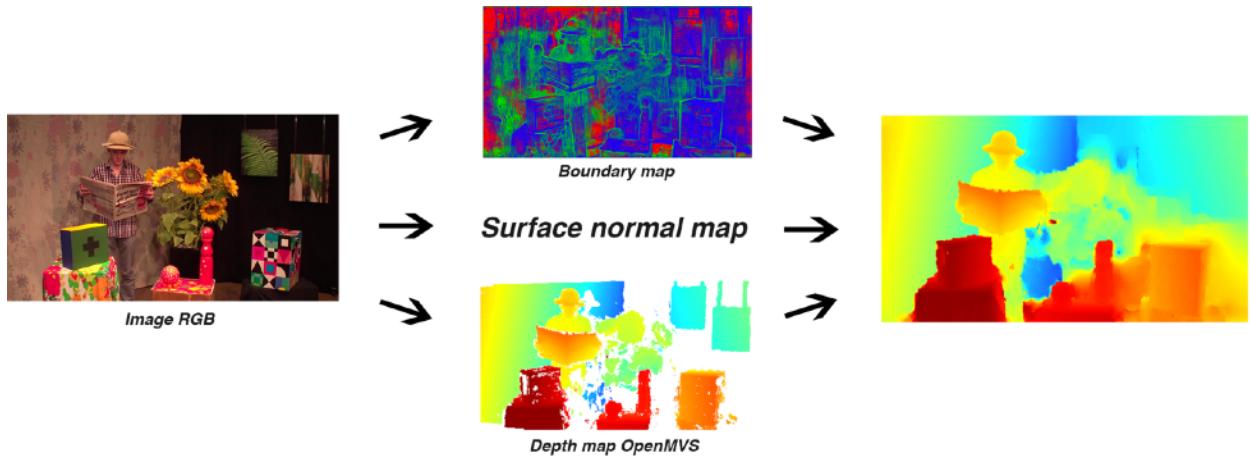


Figure 20 : Étape de la méthode de complétion de carte de profondeur

Le dataset utilisé pour entraîner ce réseau est obtenu à partir du dataset Matterport3D [17], contenant des photos d'intérieur de bâtiments associées à leur carte de profondeur, obtenue avec une caméra de profondeur. Ces nombreuses photos étant prises dans les mêmes bâtiments, les cartes de profondeur ont pu être utilisées pour reconstruire l'intérieur des bâtiments sous forme de meshs. En regénérant des cartes de profondeur à partir de ces meshs, on obtient une vérité terrain (ground truth). Ce dataset n'est donc pas non plus adapté à notre usage car il ne permet pas d'améliorer des cartes de profondeur contenant des personnes.

Après avoir donné à ce réseau une de nos images et sa carte de profondeur associée obtenue avec OpenMVS, on obtient une carte où les zones sans profondeur ont été complétées. Sur cet exemple *Figure 22*, le résultat est plutôt bon,

même si quelques zones ont mal été estimées, notamment autour des objets en bas à droite.

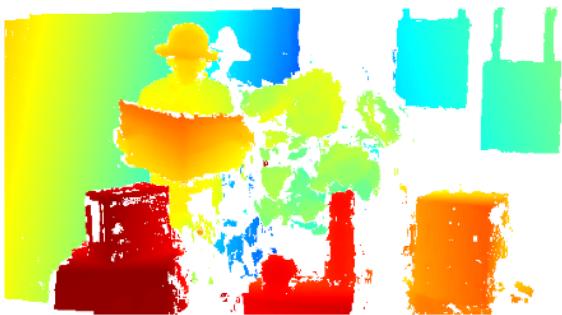


Figure 21 : En entrée : carte de profondeur
OpenMVS

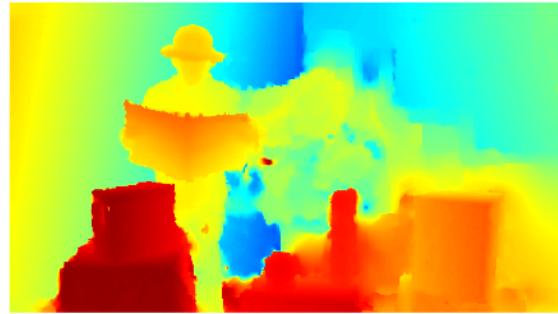


Figure 22 : En sortie : carte de profondeur
complétée

L'article *Depth-map completion for large indoor scene reconstruction* [15] se base sur la méthode ci-dessus pour l'adapter au Multi-View Stereo. Il réalise sensiblement les mêmes opérations que dans l'article précédent puis ajoute la possibilité d'améliorer les cartes de profondeur en projetant chaque pixel dans l'espace 3D pour vérifier que sa valeur de profondeur concorde avec celle des autres points de vue. Cette méthode n'a pas pu être testé car le code n'était pas disponible.

3. Apprentissage de réseaux de Deep Learning sur des datasets de personnes

Comme nous avons pu le voir, les datasets sur lesquels sont entraînés ces réseaux de neurones ne sont pas adaptés à notre usage qui est la reconstruction 3D de personnes.

Parmi les datasets de personnes (corps entiers), beaucoup ont été créé grâce à un système multi-caméras entourant une ou des personnes. Dans Human3.6M [18] ou HumanEva [21], les personnes prennent différentes postures et réalisent diverses activités du quotidien (boire, s'assoir, téléphoner, etc). The Panoptic Studio dataset [20] contient quant à lui des images multi-vues de personnes en pleine interaction sociale (discutant, jouant, se déplaçant) avec parfois des occlusions par des objets ou d'autres personnes. Dans le COCO dataset [19], la solution utilisée est la correspondance image-à-surface, c'est-à-dire un modèle 3D humain par dessus une image unique.

HUMBI est un dataset paru en juin 2020 [16]. Il contient des images multi-vues prises par 107 caméras, dont environ 26 millions d'images de personnes, corps entiers, en mouvement. Il contient également les meshs associés à ces images.

Pour entraîner un modèle à la complétion de carte de profondeur, nous pouvons utiliser ce dataset. Il nous faut donc :

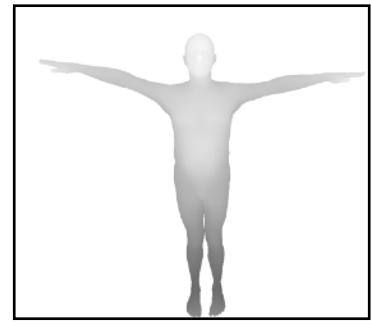
- les images RGB de chaque vue,
- des cartes de profondeur incomplètes, que l'on peut obtenir grâce à OpenMVG + OpenMVS en ne considérant que quelques vues parmi les 107 disponibles,
- des cartes de profondeur Ground Truth, que l'on peut obtenir en projetant la profondeur des meshs sur les différentes vues.



Image RGB issue de HUMBI



*Carte de profondeur incomplète
obtenue avec OpenMVS*



*Carte de profondeur Ground
Truth*

Figure 23 : Données à utiliser pour entraîner un réseau pour la depth completion

Cette technique pour créer un dataset pour la complétion de carte de profondeur est celle utilisée dans l'article *Deep depth completion of a single RGB-D image* [14] vu précédemment.

Pour exploiter ce dataset, deux solutions s'offrent à nous. Il est possible de faire un apprentissage entièrement avec des images d'humains, mais cela requiert une quantité importante de données et un temps d'apprentissage conséquent. Il existe également une technique qui permet d'utiliser un réseau ayant déjà appris une tâche similaire : l'apprentissage par transfert. En simplifiant, comme les réseaux de neurones sont des empilements de couches, chacune apprenant de la précédente, il suffirait de modifier les dernières couches pour l'adapter à notre problème. Nous pourrions par exemple réutiliser le réseau appris sur des datasets de bâtiments, pour l'adapter à la complétion de carte de profondeur de personnes. Nous aurions alors besoin de moins de données et de moins de temps.

E. Conclusion sur le projet

Ce projet de reconstruction 3D de personnes, qui s'inscrit dans les travaux d'Orange sur l'étude et la mise en œuvre de formats vidéo immersifs, n'est encore qu'à ses débuts et ne va cesser d'évoluer pour améliorer de plus en plus la qualité, la facilité et la vitesse de reconstruction.

Pour l'instant, il faut tout de même respecter quelques contraintes pour que le résultat soit correct : contenu texturé, sans surfaces transparentes ou réfléchissantes, caméras proches pour le Structure from Motion et Multi-View Stereo, modèle entraîné sur un dataset de personnes pour les réseaux de deep-learning, etc. Comme nous avons pu le voir, si ces conditions sont respectées, les reconstructions 3D de personnes peuvent être très réussies.

Les recherches scientifiques dans ce domaine avancent vite, surtout en deep-learning, et vont donc permettre à l'équipe de faire évoluer rapidement le projet. De plus, comme le futur studio d'Orange Atalante sera équipé d'une vingtaine de caméras, cela permettra d'améliorer la qualité de reconstruction ainsi que donner la possibilité de reconstruire en 3D une personne sous tous ses angles.

Conclusion

Orange est un grand acteur des télécommunications et du multimédia en France. J'ai pu intégrer Orange Labs dans une équipe de Recherche et Développement sur les contenus immersifs audio et vidéo, ce qui m'a permis de découvrir un peu plus le monde de l'entreprise, ainsi que celui de la recherche.

Ce projet m'a permis d'apprendre de nombreuses choses sur les plans technique et scientifique car travailler à plein temps sur une même thématique permet de monter en compétence rapidement. J'ai été amené à lire de nombreux articles scientifiques, ce qui m'a permis d'étendre ma culture dans le domaine de l'imagerie numérique. Le fait de travailler en autonomie et en télétravail pendant plusieurs mois m'a également appris à bien m'organiser, planifier mes tâches, et à gérer les problèmes par moi-même.

La thématique des contenus immersifs sur laquelle j'ai travaillé m'a permis de mettre à profit les connaissances techniques que j'ai acquises durant mes trois années à l'ESIR. Elle était en adéquation parfaite avec les matières enseignées en spécialité Imagerie Numérique et avec mon projet professionnel, qui est de continuer à travailler dans ce domaine une fois mon diplôme obtenu.

Bibliographie

- [1] France Telecom Orange : présentation et histoire. (2013, 2 avril). In finance. <https://www.infinance.fr/articles/entreprise/societe-cotee-en-bourse/article-france-telecom-orange-presentation-et-histoire-455.htm>
- [2] Orange, fournisseur historique d'accès internet et de téléphonie mobile. Selectra. <https://selectra.info/telecom/fournisseurs/orange>
- [3] Les activités de recherche d'Orange. (décembre 2016). Dossier de presse. https://www.orange.com/sirius/edossiers/pdfs/activites-de-recherche-orange-fr_dp_activite_recherche_fr_full.pdf
- [4] Derniers résultats consolidés. (2020, 30 avril). Orange. <https://www.orange.com/fr/Investisseurs/Resultats-et-presentations/Folder/Derniers-resultats-consolides>
- [5] Répondre aux enjeux du développement durable. (2019, 7 août). Orange. <https://www.orange.com/fr/Human-Inside/Notre-engagement-societal>
- [6] Santé et sécurité : des droits fondamentaux des salariés. (2015, 8 avril). Orange. <https://www.orange.com/fr/actus-courtes-tuiles/responsabilite/actions/Confiance/Accord-UNI-sur-la-SST>
- [7] Accord mondial Santé-Sécurité-Qualité de Vie au Travail: où en sommes-nous ? CFECGC Orange. <https://www.cfecgc-orange.org/201804176359/conditions-de-travail-et-sante/accord-mondial-sante-securite-qualite-de-vie-au-travail-ou-en-sommes-nous.html>
- [8] Furukawa, Y., & Hernández, C. (2015). Multi-View Stereo: A Tutorial. Found. *Trends Comput. Graph. Vis.*, 9, 1-148.
- [9] Bleyer, M., Rhemann, C., & Rother, C. (2011). PatchMatch Stereo - Stereo Matching with Slanted Support Windows. *BMVC*.
- [10] He, K., Gkioxari, G., Dollár, P., & Girshick, R.B. (2017). Mask R-CNN. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2980-2988.
- [11] Huang, P., Matzen, K., Kopf, J., Ahuja, N., & Huang, J. (2018). DeepMVS: Learning Multi-view Stereopsis. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2821-2830.

- [12] Yao, Y., Luo, Z., Li, S., Fang, T., & Quan, L. (2018). MVSNet: Depth Inference for Unstructured Multi-view Stereo. *ECCV*.
- [13] Yao, Yao and Luo, Zixin and Li, Shiwei and Shen, Tianwei and Fang, Tian and Quan, Long. Recurrent MVSNet for High-resolution Multi-view Stereo Depth Inference. *Computer Vision and Pattern Recognition (CVPR)*. 2019.
- [14] Zhang, Y., & Funkhouser, T. (2018). Deep Depth Completion of a Single RGB-D Image. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 175-185.
- [15] Liu, H., Tang, X., & Shen, S. (2020). Depth-map completion for large indoor scene reconstruction. *Pattern Recognit.*, 99.
- [16] Yu, Z., Yoon, J.S., Lee, I., Venkatesh, P., Park, J., Yu, J., & Park, H. (2020). HUMBI: A Large Multiview Dataset of Human Body Expressions. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2987-2997.
- [17] Chang, A.X., Dai, A., Funkhouser, T., Halber, M., Nießner, M., Savva, M., Song, S., Zeng, A., & Zhang, Y. (2017). Matterport3D: Learning from RGB-D Data in Indoor Environments. 2017 International Conference on 3D Vision (3DV), 667-676.
- [18] Ionescu, C., Papava, D., Olaru, V., & Sminchisescu, C. (2014). Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36, 1325-1339.
- [19] Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C.L. (2014). Microsoft COCO: Common Objects in Context. *ArXiv*, *abs/1405.0312*.
- [20] Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B.C., Matthews, I., Kanade, T., Nobuhara, S., & Sheikh, Y. (2015). Panoptic Studio: A Massively Multiview System for Social Motion Capture. 2015 IEEE International Conference on Computer Vision (ICCV), 3334-3342.
- [21] Sigal, L., Balan, A., & Black, M.J. (2009). HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion. *International Journal of Computer Vision*, 87, 4-27.

Annexes

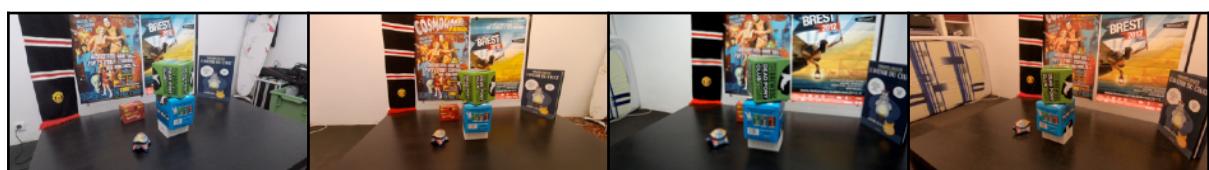
Annexe 1 : Séquences de test utilisées



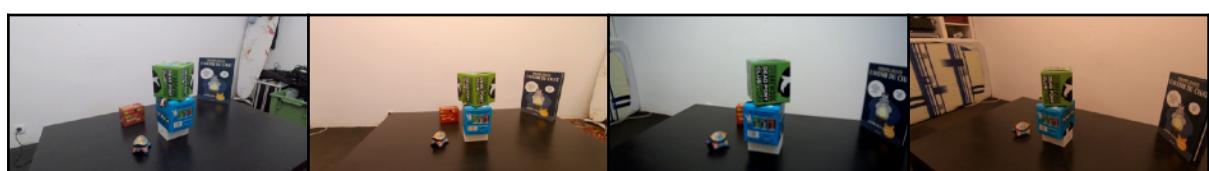
Séquence 1



Séquence 2



Séquence 3 : cubes 1/3



Séquence 4 : cubes 2/3



Séquence 5 : cubes 3/3

Annexe 2 : Matrice Fondamentale et matrice Essentielle

La matrice Fondamentale est la matrice 3x3 qui va relier l'ensemble des points d'intérêt d'une paire d'images. Pour trouver les valeurs composant F , il suffit de résoudre l'équation suivante :

$$\begin{bmatrix} x'_i & y'_i & 1 \end{bmatrix} \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = 0$$

$$x_i x'_i f_{11} + x_i y'_i f_{21} + x_i f_{31} + y_i x'_i f_{12} + y_i y'_i f_{22} + y_i f_{32} + x'_i f_{13} + y'_i f_{23} + f_{33} = 0$$

Comme la matrice Fondamentale, la matrice Essentielle va relier l'ensemble des points d'intérêt d'une paire d'images, mais en prenant en compte les propriétés intrinsèques des caméras : focales, facteurs d'agrandissement, coordonnées de la projection du centre optique, etc. Elle peut être trouvée grâce à l'équation suivante :

$$\mathbf{E} = \mathbf{K}^T \mathbf{F} \mathbf{K}$$

où \mathbf{k} est la matrice des paramètres intrinsèques.

Annexe 3 : Algorithme RANSAC

RANSAC (RANdom SAmple Consensus) va permettre de retirer les points aberrants (outliers) de l'ensemble des points d'intérêt.

Algorithm 1 RANSAC

- 1: Select randomly the minimum number of points required to determine the model parameters.
 - 2: Solve for the parameters of the model.
 - 3: Determine how many points from the set of all points fit with a predefined tolerance ϵ .
 - 4: If the fraction of the number of inliers over the total number points in the set exceeds a predefined threshold τ , re-estimate the model parameters using all the identified inliers and terminate.
 - 5: Otherwise, repeat steps 1 through 4 (maximum of N times).
-

Annexe 4 : Algorithme PatchMatch Stereo

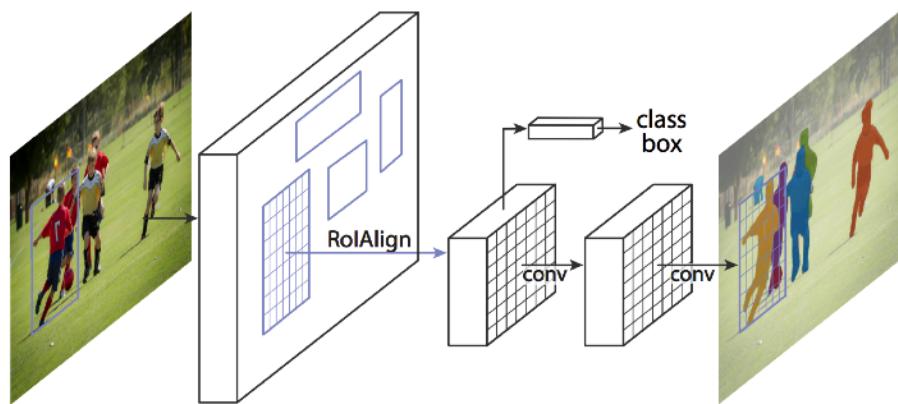
Définition du patch: petite fenêtre de pixels défini par son center et sa normale

1. feature matching initial grâce au sparse point cloud
2. projection des patches dans chaque images (features = centres des patchs)
3. ajustement de la position et de l'orientation des patchs pour que leur projection dans chaque vue corresponde
4. répéter N fois :
 - A. expansion : propager les patchs existants dans les espaces vides pour créer de nouveaux patchs en se basant sur la similarité d'intensité entre les pixels
 - B. filtrage : retirer les mauvais patchs en se basant sur la cohérence de visibilité

Annexe 5 : Mask R-CNN

Le fonctionnement de Mask R-CNN est le suivant :

- extraction des points d'intérêt de l'image grâce au réseau de neurones ResNet,
- la feature map obtenue est passée à un RPN (Region Proposal Network) qui retourne les bounding boxes,
- celles-ci sont ensuite passées dans une couche de pooling pour leur donner la même taille, puis dans une couche entièrement connectée (fully-connected layer) pour prédire la classe,
- enfin, les bounding boxes sont passées dans un FCN (fully convolutional network) pour créer un masque pour chaque objet.



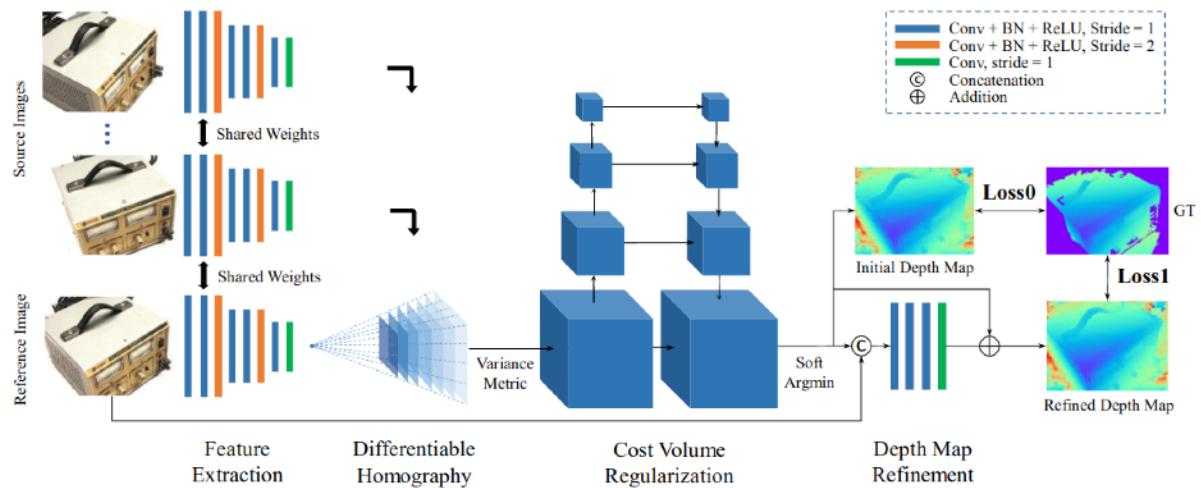
Mask R-CNN

Annexe 6 : Architecture MVSNet

Yao, Yao and Luo, Zixin and Li, Shiwei and Fang, Tian and Quan, Long. *MVSNet: Depth Inference for Unstructured Multi-view Stereo*. European Conference on Computer Vision (ECCV). 2018.

La première étape du réseau de MVSNet est d'extraire les features sous forme de features map grâce à des couches de CNN (réseaux de neurones convolutifs). La deuxième étape consiste à créer un volume de coût 3D (cost volume) grâce à l'homographie différentiable. Ensuite, une carte de profondeur initiale est créée grâce à des convolutions 3D, puis améliorée grâce à son image RGB associée (image de référence) pour obtenir la carte de profondeur finale.

L'entraînement de ce réseau s'est fait sur le dataset DTU, fournissant des images de points de vue différents d'objets et de bâtiments extérieurs, ainsi que leur carte de profondeur associée.



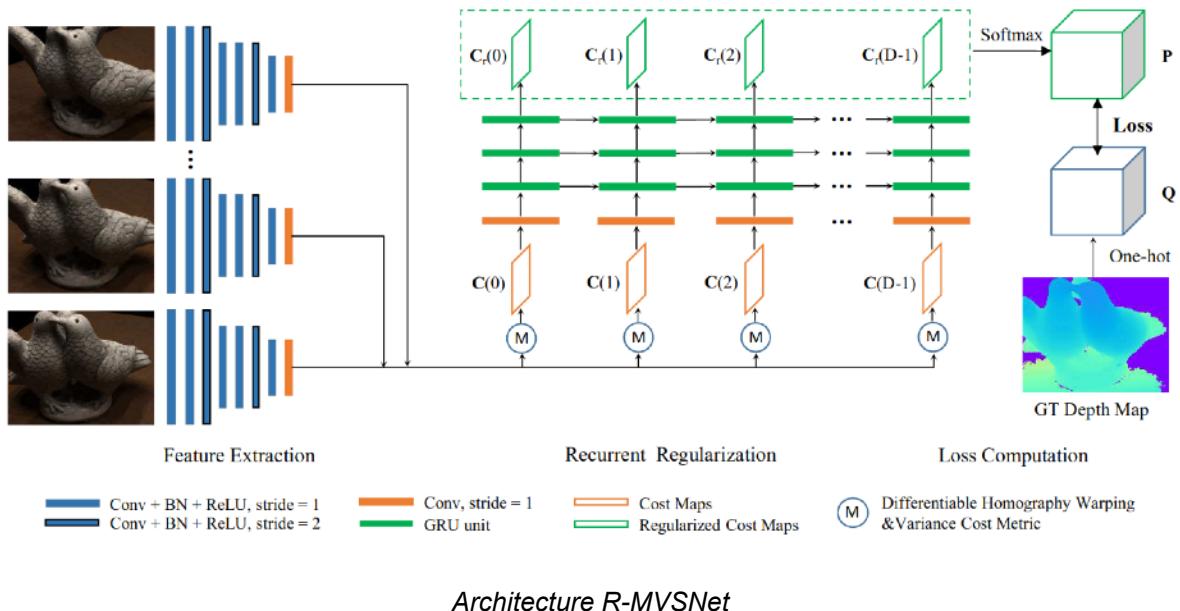
Architecture MVSNet

Annexe 7 : Architecture R-MVSNet

Yao, Yao and Luo, Zixin and Li, Shiwei and Shen, Tianwei and Fang, Tian and Quan, Long. *Recurrent MVSNet for High-resolution Multi-view Stereo Depth Inference*. Computer Vision and Pattern Recognition (CVPR). 2019.

R-MVSNet (Recurrent-MVSNet) est basé sur l'architecture de MVSNet.

Recurrent-MVSNet remplace la régularisation du cost volume 3D par une régularisation séquentielle du cost volume 2D le long de l'axe de profondeur. Cela permet de réduire la complexité et de permettre d'utiliser des images de plus hautes résolutions.



Architecture R-MVSNet

L'entraînement de ce réseau s'est fait sur le dataset DTU, fournissant des images de points de vue différents d'objets et de bâtiments extérieurs, ainsi que leur carte de profondeur associée.

Annexe 8 : Architecture DeepMVS

Huang, P. and Matzen, K. and Kopf, J. and Ahuja, N. and Huang, J. DeepMVS: Learning Multi-View Stereopsis. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

DeepMVS est un réseau de neurones convolutifs. Après avoir choisi une image de référence parmi les différents points de vue, on crée pour chaque image des « plane-sweep volumes », c'est-à-dire des volumes composés de N plans, parallèles à l'image de référence. N correspond au nombre de niveaux de profondeur. Ensuite, l'architecture est composée d'un réseau de Patch Matching, qui va détecter les features pour chaque image et chaque niveau de profondeur, puis des réseaux U-Net et VGG-19, qui vont permettre rassembler les features de toutes les images par niveau de profondeur, et enfin d'une couche de max pooling et de deux couches convolutives, qui vont générer la carte de profondeur. Le résultat est ensuite corrigé et lissé.

Les datasets utilisés pour l'entraînement de DeepMVS contiennent des images

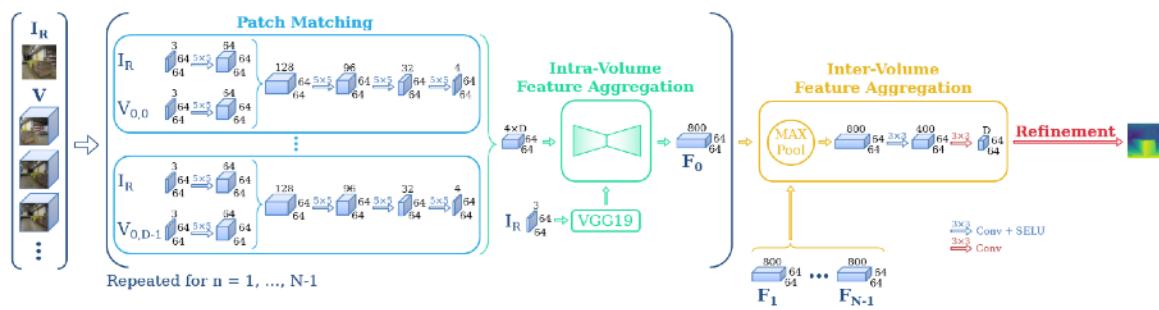
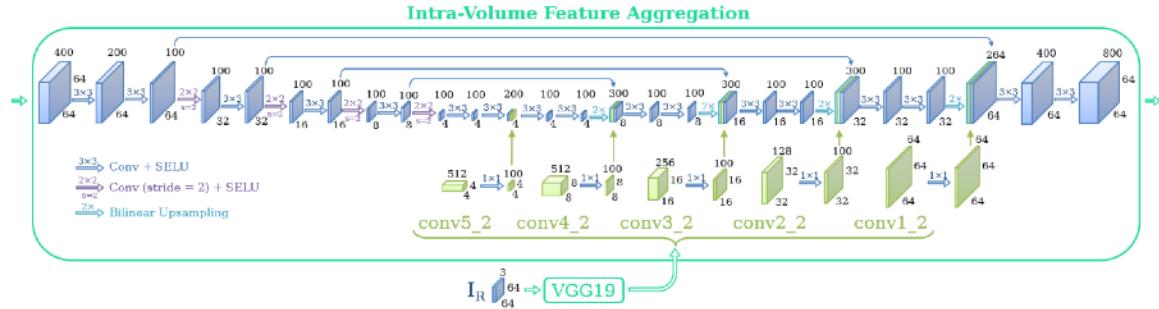


Figure 2. DeepMVS network architecture.



Architecture DeepMVS

réelles d'intérieur et d'extérieur de bâtiments avec leur carte de profondeur associée (SUND3D, RGB-D SLAM, CITYWALL et ACHTECK-TURM) et d'un dataset d'images synthétiques de ville issues d'un jeu vidéo.