Applications of multi-image remote sensing

Roger Marí Molas

▶ **To cite this version:**

Roger Marí Molas. Applications of multi-image remote sensing. Computer Vision and Pattern Recognition [cs.CV]. Université Paris-Saclay, 2022. English. ⟨NNT : 2022UPASM045⟩. ⟨tel-04143750⟩

HAL Id: tel-04143750

https://theses.hal.science/tel-04143750v1

Submitted on 28 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

université
PARIS-SACLAY

Applications of multi-image remote sensing

Applications de la télédétection multi-image

Thèse de doctorat de l'Université Paris-Saclay

École doctorale n° 574, mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées
Graduate School : Mathématiques
Référent : ENS Paris-Saclay

Thèse préparée dans l'unité de recherche **Centre Borelli (Université Paris-Saclay, CNRS, ENS Paris-Saclay)**, sous la direction de **Gabriele FACCIOLO**, Professeur des universités, et la co-direction d'**Enric MEINHARDT-LLOPIS**, Maître de conférences.

Thèse soutenue à Paris-Saclay, le 14 décembre 2022, par

Roger Marí Molas

Composition du jury

Membres du jury avec voix délibérative

Pascal MONASSE Professeur des universités, École des Ponts ParisTech	Président
Mathieu AUBRY Directeur de recherche, École des Ponts ParisTech	Rapporteur & Examinateur
Marc PIERROT-DESEILLIGNY Directeur de recherche, Institut national de l'information géographique et forestière (IGN)	Rapporteur & Examinateur
Pablo D'ANGELO Chargé de recherche, German Aerospace Center (DLR)	Examinateur
Coloma BALLESTER Professeure, Universitat Pompeu Fabra (UPF)	Examinatrice
Rafael GROMPONE VON GIOI Directeur de recherche, ENS Paris-Saclay	Examinateur

Thèse de doctorat

NNT: 2022UPASM045

université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Titre : Applications de la télédétection multi-image

Mots clés : reconstruction 3D multi-vues, images satellites, ajustement de faisceaux, rendu neuronal

Résumé : Cette thèse étudie le problème de la reconstruction 3D à partir d'une collection d'images satellites à haute résolution. La reconstruction 3D multi-vues par satellite nécessite un contrôle très fin de la géométrie d'acquisition, afin de garantir la cohérence des estimations d'altitude obtenues à partir de différentes vues. La première partie de la thèse est donc consacrée à l'optimisation de la représentation mathématique de la géométrie d'acquisition, qui se présente généralement sous la forme de modèles de caméras RPC. Nous proposons une méthodologie d'ajustement de faisceaux qui maximise la cohérence géométrique entre un ensemble de vues satellites et les caméras RPC associées. Cette méthodologie intègre un algorithme d'estimation de modèles RPC qui permet la composition directe des modèles originaux non raffinés avec des transformations correctives, sans utiliser de représentations intermédiaires approximatives. La deuxième partie de la thèse présente différentes applications pratiques de la télédétection multi-image, dont la plupart profitent du contrôle de la cohérence de la géométrie d'acquisition. Les différentes méthodes concernent les sujets suivants : la détection des changements de volume au niveau de la surface de la Terre à différentes dates ; la génération géométriquement cohérente de mosaïques à grande échelle construites à partir d'images satellites plus petites ; un réseau de rendu neuronal (NeRF) capable d'apprendre la géométrie d'une scène satellite et de synthétiser de nouvelles vues réalistes, avec la capacité de distinguer les ombres et les objets transitoires des structures permanentes ; et une comparaison entre les algorithmes classiques et les réseaux d'apprentissage profond supervisés pour la mise en correspondance stéréo. Comme résultat, cette thèse décrit une variété d'idées de pointe sur l'exploitation des images satellites optiques qui ont le potentiel d'améliorer les activités liées à la connaissance de la surface terrestre à grande échelle, comme la surveillance, la planification urbaine ou la gestion des ressources naturelles. Les méthodes présentées sont évaluées avec des images satellites WorldView-3 et SkySat à haute résolution. L'implémentation de la plupart des méthodes est également publiée en logiciel libre Python.

Title : Applications of multi-image remote sensing

Keywords : multi-view 3D reconstruction, satellite images, bundle adjustment, neural rendering

Abstract : This thesis studies the problem of 3D reconstruction from a collection of high-resolution satellite images. Satellite multi-view 3D reconstruction requires a very fine control of the acquisition geometry, in order to guarantee the consistency of altitude estimates obtained from different views. The first part of the thesis is therefore devoted to the optimization of the mathematical representation of the acquisition geometry, which usually takes the form of RPC camera models. We propose a bundle adjustment methodology that maximizes the geometric consistency between a set of satellite views and the associated RPC cameras. This methodology incorporates an RPC estimation algorithm that allows the direct composition of the original unrefined models with corrective transformations, without using approximate intermediate representations. The second part of the thesis presents different practical applications of multi-image remote sensing, most of which benefit from the consistency control of the acquisition geometry. The different methods concern the following topics : the detection of volume changes on the Earth's surface across different dates ; the geometrically consistent generation of large-scale mosaics built from smaller satellite images ; a neural rendering network (NeRF) capable of learning the geometry of a satellite scene in a self-supervised manner and also of synthesizing new realistic views, with the ability to distinguish shadows and transient objects from permanent structures ; and a comparison between classic algorithms and supervised deep learning networks for dense stereo matching. As a result, this thesis describes a variety of cutting-edge ideas on the exploitation of optical satellite images that have the potential to improve activities related to large-scale land surface knowledge, such as surveillance, urban planning or natural resource management. The presented methods are evaluated with high-resolution images from the WorldView-3 and SkySat constellations. The implementation of most methods is also released as open-source Python code.

Abstract

This thesis addresses the problem of computing accurate digital surface models from multiple optical images captured by Earth observation satellites. This problem has traditionally been addressed with stereo vision methods. The reason for this is the nature of the first Earth observation satellite constellations, which were composed of a few large units capable of imaging large areas at once, but limited to long revisit times. Remote sensing is evolving today towards new constellations of a large number of small satellites, which offer shorter revisit times and cover large areas with multi-image acquisitions. This context opens the door to revisit the process of 3D reconstruction from satellite imagery, to provide new tools that allow a natural treatment of multiple input views and minimize the overload of dealing with large numbers of stereo pairs. We present a series of methods for this purpose, which are evaluated with WorldView-3 and SkySat images at sub-meter resolution.

The opening chapter is an introduction to the origins and characteristics of optical satellite imagery, which is followed by a review of stereo-based digital surface modeling approaches based on these data. Multi-view stereo pipelines are usually structured around a selection of image pairs, which undergo subsequent stereo rectification, dense stereo matching and triangulation steps, resulting in a set of pairwise altitude estimates. The pairwise models are then merged into a single multi-view model, whose accuracy is less affected by occlusions or radiometric inconsistencies. The introduction concludes by presenting a common problem of these methods: the existence of inconsistencies between the altitude estimates of different pairs, due to inaccuracies in the image sensor models.

The first part of the thesis is devoted to the study of methods aimed at correcting the multi-view geometry inconsistencies caused by the satellite image sensor models. The RPC (*Rational Polynomial Coefficients*) camera model is adopted to represent the sensor models in a generic manner. A bundle adjustment methodology is then proposed to optimize an input collection of RPC models by addressing the main source of inaccuracies, which are errors related to the sensor orientation. The proposed strategy is contrasted with other conventional bundle adjustment methods, which rely on a local approximation of the sensor models, and with concurrent strategies that circumvent the sensor model inaccuracies by direct geometry-based co-registration of pairwise altitude estimates.

The second part of the thesis presents different applications related to multi-view digital surface modeling. The majority of the methods in this second part take advantage of the outcome of the first part and show the practical advantages of working with multiple geometrically coherent satellite images. The different methods concern the following topics: the detection of volume changes on the Earth's surface across different dates; the geometrically consistent generation of large-scale mosaics built from smaller satellite images; a neural rendering network (*NeRF*) capable of learning the geometry of a satellite scene in a self-supervised manner and also of synthesizing new realistic views, with the ability to distinguish shadows and transient objects from permanent structures; and a comparison between classic algorithms and supervised deep learning networks for dense stereo matching.

Résumé

Cette thèse aborde le problème de l'estimation de modèles numériques de surface précis à partir de multiples images optiques capturées par des satellites d'observation de la Terre. Ce problème a traditionnellement été abordé avec des méthodes de vision stéréo. Ceci s'explique par la composition des premières constellations de satellites d'observation de la Terre, constituées de quelques grandes unités capables de photographier de grandes zones en une seule image, mais limitées à de longues périodes de revisite. La télédétection évolue aujourd'hui vers de nouvelles constellations composées d'un grand nombre de petits satellites, qui offrent des périodes de revisite plus courtes et couvrent de grandes zones avec des acquisitions multi-images. Ce contexte ouvre la porte à une révision du processus de reconstruction 3D à partir d'images satellites, pour fournir des outils qui permettent une gestion naturelle de plusieurs vues d'entrée et minimisent la surcharge liée au traitement d'un grand nombre de paires stéréo. Nous présentons une série de méthodes à cette fin, qui sont évaluées avec des images WorldView-3 et SkySat à haute résolution sub-métrique.

Le premier chapitre est une introduction aux origines et aux caractéristiques de l'imagerie satellitaire optique, suivie par un aperçu des approches de modélisation numérique de surface basées sur ces données. Les chaînes stéréo multi-vues sont généralement structurées autour d'une sélection de paires d'images, qui sont ensuite soumises à des étapes de rectification stéréo, de mise en correspondance stéréo et de triangulation, ce qui permet d'obtenir un ensemble d'estimations d'altitude par paire. Les modèles obtenus de différentes paires stéréo sont ensuite fusionnés en un seul modèle multi-vues, dont la précision est moins affectée par les occlusions ou les incohérences radiométriques. L'introduction se termine par la présentation d'un problème commun à ces méthodes : l'existence d'incohérences entre les estimations d'altitude de différentes paires stéréo, dues à des imprécisions dans les modèles de caméra des images.

La première partie de la thèse est consacrée à l'étude des méthodes visant à corriger les incohérences de la géométrie multi-vues causées par les modèles de caméra des images satellites. Le modèle de caméra RPC (*Rational Polynomial Coefficients*) est adopté pour représenter les modèles de capteurs de manière générique. Une méthodologie d'ajustement de faisceaux est ensuite proposée pour optimiser une collection de modèles RPC en entrée en abordant la principale source d'inexactitudes, qui sont les erreurs liées à l'orientation du capteur. La stratégie proposée est contrastée par rapport à d'autres méthodes conventionnelles d'ajustement de faisceaux, qui reposent sur une approximation locale des modèles de caméra, et par rapport à des stratégies concurrentes qui contournent les inexactitudes des modèles de caméra par un recalage géométrique direct des estimations d'altitude stéréo.

La deuxième partie de la thèse présente différentes applications liées à la modélisation numérique de surface multi-vues. La plupart des méthodes de cette deuxième partie profitent des résultats de la première partie et montrent les avantages pratiques de travailler avec plusieurs images satellites géométriquement cohérentes. Les différentes méthodes concernent les sujets suivants : la détection des changements de volume au niveau de la surface de la Terre à différentes dates ; la génération géométriquement cohérente de mosaïques à grande échelle construites à partir d'images satellites plus petites ; un réseau de rendu neuronal (*NeRF*) capable d'apprendre la géométrie d'une scène satellite et de synthétiser de nouvelles vues réalistes, avec la capacité de distinguer les ombres et les objets transitoires des structures permanentes ; et une comparaison entre les algorithmes classiques et les réseaux d'apprentissage profond supervisés pour la mise en correspondance stéréo.

Remerciements

Mes remerciements vont en premier lieu à mon directeur de thèse, Gabriele Facciolo, pour son encadrement et son aide inestimable au fil de ces années. Travailler avec quelqu'un comme lui, brillant, toujours de bonne humeur et accessible, a été fondamental pour mon apprentissage mais aussi pour mon état d'esprit et je me sens très privilégié pour cela. Je remercie également Enric Meinhardt-Llopis, Carlo de Franchis et Thibaud Ehret, qui ont co-supervisé une partie des travaux présentés dans cette thèse.

Je tiens ensuite à remercier Mathieu Aubry et Marc Pierrot-Deseilligny qui m'ont fait l'honneur d'accepter d'être les rapporteurs de cette thèse, ainsi que le reste du jury : Pablo d'Angelo, Coloma Ballester, Rafael Grompone von Gioi et Pascal Monasse. Je ne peux que leur exprimer ma sincère gratitude.

Je remercie naturellement tous les membres du Centre Borelli. À Jean-Michel Morel pour sa confiance et pour avoir facilité mon arrivée au laboratoire, à Mariano et Philippe pour son accueil lors de mes premiers jours, à Jérémy pour son soutien et pour avoir constamment partagé son talent, à Marina avec qui j'ai eu le plaisir d'organiser plusieurs séminaires sur le traitement d'images, à Pablo J. pour les agréables discussions au bureau, et à tous les autres doctorants et chercheurs que j'ai côtoyés au laboratoire : Axel, Charles, Long, Marie, Max, Miguel, Pablo A., Quentin, Roland, Sébastien, Tina, Valéry (*et al.*, évidemment !), pour leur gentillesse et les bons moments passés ensemble. Mes meilleurs vœux aussi aux nouveaux arrivants qui commencent maintenant leur doctorat.

Je voudrais aussi remercier la Région Île-de-France pour le financement de cette thèse, qui m'a donné l'occasion de vivre à Paris et de mieux connaître la culture française. Cette thèse m'a également conduit à des conférences à Séoul, Bruxelles (virtuellement), Nice et la Nouvelle-Orléans. Je veux remercier le Centre Borelli de m'avoir permis de participer à des événements aussi stimulants, qui n'ont fait que renforcer mon admiration pour la communauté scientifique. Merci à Alina et à l'ensemble du secrétariat pour leur soutien logistique lors de ces missions.

Vull expressar també el meu agraïment a la professora Gloria Haro, per haver despertat el meu interès per la visió 3D durant els meus estudis a l'Universitat Pompeu Fabra i per haver estat un referent per a la meva formació. Gràcies també a la resta de boníssims professors i professionals que em van preparar a Barcelona per poder afrontar aquests darrers anys.

Com no pot ser d'una altra manera, gràcies a la meva família i als meus amics catalans, que sempre m'han rebut amb els braços oberts en les meves visites a casa. Especialment als meus pares Jaume i Carme, en qui he pensat molt. No els podré agrair mai prou la seva bondat i el gust per la vida i per la cultura que m'ensenyen contínuament.

Non posso che concludere ringraziando Alice, che è la scoperta più speciale che ho fatto in questo percorso e che mi ha sostenuto in modo straordinario. Grazie di cuore.

Contents

Abstract **5**

Résumé **7**

Remerciements **9**

1 Introduction **13**

1.1 Motivation . . . 13
1.2 Origins and characteristics of optical satellite imagery . . . 14
1.3 New small satellite constellations . . . 15
1.4 3D reconstruction from multi-view satellite images . . . 15
1.5 Short outline of the thesis . . . 22
1.6 Summary of contributions . . . 27
1.7 List of publications . . . 28

2 Introduction (en français) **31**

2.1 Motivation . . . 31
2.2 Origines et caractéristiques de l'imagerie satellitaire optique . . . 32
2.3 De nouvelles constellations de petits satellites . . . 33
2.4 Reconstruction 3D multi-vues à partir d'images satellites . . . 34
2.5 Présentation de la thèse . . . 42
2.6 Synthèse des contributions . . . 47
2.7 Liste des publications . . . 48

I Geometric modeling of multi-view satellite imagery **51**

3 Geolocation correction methods for satellite multi-view stereo **53**

3.1 Introduction . . . 53
3.2 Related work . . . 55
3.3 Correlation-based DSM alignment . . . 56
3.4 Bias-compensation bundle adjustment . . . 58
3.5 Orientation-compensation bundle adjustment . . . 60
3.6 Feature track refinement . . . 61
3.7 Evaluation with WorldView-3 image pairs . . . 62
3.8 Conclusion . . . 65

4 Bundle adjustment of RPC camera models **67**

4.1 Introduction . . . 67
4.2 Related work . . . 68
4.3 Method description: Tie point generation and optimization framework . . 71
4.4 Evaluation and comparison with geometry-based registration . . . 82
4.5 Conclusion . . . 88

5 Robust RPC camera modeling **91**

5.1 Introduction . . . 91
5.2 Related work . . . 92
5.3 RPC estimation via terrain-independent regularized least squares . . . 94
5.4 Evaluation with SAR and optical satellite imagery . . . 95
5.5 Conclusion . . . 97

II Applications of multi-view satellite imagery **99**

6 Automatic stockpile volume monitoring **101**

6.1 Introduction . . . 101
6.2 Related work . . . 102
6.3 Method description: Surface monitoring from recurrent satellite images . . 103
6.4 Experiments: Coal volume estimation using SkySat multi-image time series 106
6.5 Conclusion . . . 107

7 Perfect sensor localization for push-frame image stitching **109**

7.1 Introduction . . . 109
7.2 Related work . . . 110
7.3 L1B$^+$: A geometrically consistent mosaic methodology for RPC cameras . 111
7.4 Evaluation with SkySat imagery and application to stereo pipelines . . . 116
7.5 Conclusion . . . 118

8 Satellite NeRF **121**

8.1 Introduction . . . 121
8.2 Related work . . . 123
8.3 Sat-NeRF: Self-supervised neural rendering for multi-date satellite imagery 125
8.4 Evaluation and comparison with classic multi-view stereo . . . 131
8.5 Conclusion . . . 135

9 Disparity estimation networks **137**

9.1 Introduction . . . 137
9.2 Related work . . . 138
9.3 Pyramid Stereo Matching network (PSM) . . . 142
9.4 Hierarchical Stereo Matching network (HSM) . . . 144
9.5 Experiments: Altitude-based assessment using satellite image pairs . . . 146
9.6 Conclusion . . . 155

10 Conclusion **157**

III Appendix **161**

A Fundamentals of bundle adjustment optimization **163**

B Disparity estimation networks: Extras **167**

B.1 List of DFC2019 stereo pairs 167
B.2 Completeness values 168

Bibliography **171**

1 Introduction

1.1 Motivation

The term *remote sensing* refers to the science of obtaining information about a scene through the analysis of data acquired by a device that is not in contact with the target scene [LKC15]. We routinely use a simple form of remote sensing: our eyes act as sensors that capture the radiation emitted by the objects around us, and our brain analyzes and extracts knowledge from that data. Those who can see well from a distance are aware of the advantages of remote sensing: they can learn details that others cannot while enjoying a safe position and more time to react to events.

In remote sensing, the data acquired is usually electromagnetic radiation. While the human eye can only see a limited range of frequencies, artificial sensors can cover the entire electromagnetic spectrum. Depending on the origin of the radiation, remote sensing sensors can be divided into two categories:

- Active remote sensing: The measured radiation corresponds to an energy bounce originally generated by an artificial source directed at the scene. Classic examples are airborne lidar (LIght Detection Imagery And Ranging) scanners or radar systems like SAR (Synthetic-Aperture Radar) satellites.
- Passive remote sensing: The measured radiation is either reflected solar radiation (optical imaging) and/or emitted by the targets (thermal imaging). Classic examples are camera equipped satellites or UAVs (Unmanned Aerial Vehicles).

Optical satellite imagery is the topic of study in this thesis. This branch of data is of special interest for Earth observation for various of reasons: coverage of large areas, non-intrusive access to any location, independence from ground support or human guidance, stable and predictable trajectory or automatic revisit periods. In many respects, optical Earth observation satellites replicate the fundamental scheme of human visual perception. Instead of a camera system (the eyes) mounted on a ground device (the body), they use a camera system mounted on a floating, orbiting platform, observing from a distance of hundreds of kilometers. However, one key part is missing: the analysis methodology that converts the acquired data into useful information. Such methodology is a constantly evolving field, as is satellite technology and the Earth's surface. The need to review known methods and update them represents the starting point of this work.

In particular, this thesis focuses on the applications of optical satellite imagery related to Earth surface modeling. While optical images are two-dimensional, the third dimension, the altitude of the scene, usually represents essential information for a large number of applications. This is the case, for example, for activities such as natural disaster prevention

and response, land management, mining, waste disposal or urban planning. Change detection is usually at the core of such activities. Since changes occur in a three-dimensional space, having measurements in the same domain is an effective way to fully characterize them. At the same time, changes in image appearance may not correspond to actual changes in scene geometry, but to external factors such as sun exposure or weather [dF15]. As a result, we could say that for Earth observation tasks, a picture is probably worth a thousand words, but a good surface model may be priceless.

1.2 Origins and characteristics of optical satellite imagery

In 1972 the United States launched Landsat 1, the first Earth observation satellite that provided imagery in digital format and in multispectral form [TGH08]. Multispectral whisk broom scanners, such as those on the first Landsat units, were the prime Earth observation sensors during the 1970s into the 1980s. These instruments used an *across-track* scanning logic, with a rotating mirror that reflected light onto a single sensor and recorded one pixel at a time [LKC15]. However, the moving parts make this type of sensors expensive and subject to wear out. In 1982 the French Space Agency launched SPOT (in French, *Satellite Pour l'Observation de la Terre*), the first satellite in orbit to use a pushbroom scanner, following an *along-track* scanning logic. Pushbroom scanners use a linear array of sensors (usually charge-coupled devices or CCDs), that are arranged perpendicular to the direction of movement of the satellite and capture one line of pixels at a time. The sensors of the array can have varying sensibility, so they have to be perfectly calibrated to prevent stripe discontinuities in the output image. After SPOT, pushbroom scanners have become regularly used in optical Earth observation satellites (e.g., IKONOS, QuickBird, GeoEye or WorldView [PT12]).

Optical satellite images are usually characterized by the following magnitudes:

- Spectral resolution: Corresponds to the number and range of spectral bands. We distinguish between *panchromatic* images, comprising a single channel sensitive to a broad range of wavelengths; *multispectral* images, comprising a few broad bands such as blue, green, red and near infra-red; and *hyperspectral* images, comprising about a hundred or more contiguous and narrow spectral bands.
- Radiometric resolution: Corresponds to the number bits used to express numerical values in the image channels. E.g. 8-bit images have a range of 0 to 255 (2^8 values).
- Temporal resolution: In the case of orbiting satellites, temporal resolution is commonly understood as the amount of time that the satellite takes to recapture a particular geographical area. It is also referred to as revisit time or repeat cycle.
- Spatial resolution: Denotes the smallest physical object size that the sensor can distinguish. It is usually expressed in ground sampling distance (GSD), which refers to the dimensions of an image pixel measured on the ground, in meters per pixel.

The spectral, radiometric, temporal and spatial resolution of satellite images define their potential applications. For example, single channel panchromatic images usually offer better spatial resolution, but images with higher spectral resolution depicting colors or data not visible to the human eye may be of greater interest for tasks such as segmentation, soil moisture analysis, bathymetry or pollution monitoring.

In this thesis, temporal and spatial resolution play a major role, since they are critical

for modeling the Earth's surface in a recurrent and detailed manner. Fortunately for us, satellite technology has come a long way in the last fifty years. Landsat 1 offered a coarse resolution of about 60 to 80 m per pixel and a revisit time of 18 days [Tod77, WGA06]. SPOT 1 acquired images with a resolution of 10 m per pixel and paved the way for subsequent satellites with increasing spatial resolution [dF15]. IKONOS (from the Greek word *eikōn*, meaning *image*), launched in 1999, was the first to collect publicly available high-resolution imagery at 1 m per pixel [Gro01]. WorldView-3, launched in 2014, provides the highest resolution commercial imagery to date, reaching about 30 cm per pixel [TJ16]. Like spatial resolution, temporal resolution is also increasing from the order of days to hours. This trend is mainly due to new constellations of small satellites, which combine the passes of multiple units to reduce revisit times.

1.3 New small satellite constellations

Conventional satellite constellations typically consist of one or a few large units, over 1000 kg, orbiting at high altitudes between 600 and 800 km. E.g. Landsat, Sentinel, SPOT or WorldView-2 and WorldView-3 [ZSL+17]. Thanks to their size, these satellites can accommodate large sensors, covering wide swaths of tens to hundreds of kilometers, i.e. the area imaged on the surface [TJ16]. However, these sparse and heavy constellations are often expensive to develop and maintain over a lifetime of several years. In addition, the small number of units usually limits revisit times to a few days.

In this context, small satellites or *smallsats*, of less than 500 kg, are increasingly attracting the attention of the industry [SBD10]. The first experiments with tiny satellites date back to 1999 at Stanford University [HPSM+00], but the real breakthrough came in 2014 when the American company Planet (formerly Planet Labs) launched a fleet of 93 satellites about 5 kg each, which has evolved into the PlanetScope constellation [WBPVLdCGT17]. In 2022, PlanetScope counts with 130 units and is able to capture the entire surface of the Earth daily with a resolution of 3 to 5 m per pixel. Planet also controls other small satellite constellations, such as SkySat. SkySats have a mass of 110 kg, with 14 units in orbit in 2022, which provide a resolution of 0.6 to 0.8 m per pixel and an impressive sub-daily combined revisit time worldwide (multiple visits per day) [Pla22]. Other examples of small satellite constellations providing optical images are those of companies like Satellogic or BlackSky [CMT20].

Small satellites are economical and can orbit at low altitudes of 400 to 500 km without consuming much power, which is advantageous to observe things with a good level of detail. The main drawback is that these design choices restrict the sensors that can be mounted on a platform. As a result, high-resolution smallsat imaging systems like SkySat are constrained to small swaths. A single image is limited then to a small geographical footprint. To compensate for this, innovative acquisition modes are used to cover large areas. E.g. the SkySat *push-frame* acquisition mode captures strips of partially overlapping images as the satellite moves.

1.4 3D reconstruction from multi-view satellite images

3D reconstruction from multiple views is a classic computer vision problem. In general, multi-view 3D reconstruction methods can be divided into multi-view stereo (MVS) and true multi-view approaches [FH15]. MVS methods split the problem into multiple pairs of views, which are reconstructed independently, and then the pairwise geometric estimates

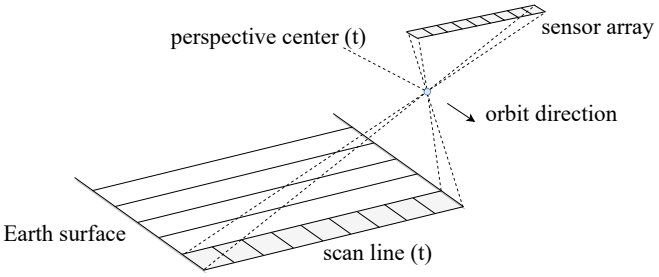


Figure 1.1: Schematic diagram of data acquisition using a pushbroom sensor.

are merged into a final multi-view reconstruction. True multi-view approaches, on the other hand, use all the available views at once to infer the underlying geometry.

This section first reviews the geometric camera models used for multi-view 3D reconstruction from satellite imagery. It then discusses the most common steps of satellite multi-view stereo, which is the predominant approach in production lines [dF15, Gon21].

1.4.1 The Rational Polynomial Coefficients (RPC) camera model

The algorithms presented in this thesis work with satellite images acquired with pushbroom sensors, whose geometry is illustrated in Figure 1.1. Satellite pushbroom sensors acquire one line of pixels at a time, implying that the orientation parameters and the center of perspective change as the satellite moves [Gro01]. Therefore, conventional pinhole camera models with a single center of projection, which are commonly used for close-range imagery, cannot be used to model the pushbroom acquisition process with exactitude.

When all the physical phenomena and components involved are known, the acquisition process can be represented as a chain of operations that model these factors, in what is known as a *physical* or *rigorous* sensor model [dFMLGF15]. However, physical sensor models are often complex and specific to each satellite platform, making it difficult to define generic algorithms with the potential to handle different image sources.

As a solution, after the launch of IKONOS in 1999, the Rational Polynomial Coefficients (RPC) camera model[1] has been widely adopted by the industry to work with pushbroom imagery [Gro01]. The RPC model is an abstract representation of the acquisition process, independent from the specific physical properties of the sensor. As the name indicates, the acquisition process is represented using rational polynomial functions. RPC models comprise a *projection* function $\mathcal{P} : \mathbb{R}^3 \to \mathbb{R}^2$, that maps 3D points to the image plane, and its inverse, the *localization* function $\mathcal{L} : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^3$. The projection function takes as input latitude, longitude and altitude coordinates, and maps them to rows and columns of the image; while the localization function takes as input row and column coordinates and a given altitude to produce the related latitude and longitude.

The RPC functions allow to triangulate the location of a 3D point that is observed in two or more images. If the point (x, y) of image i corresponds to the point (x', y') of image j, then the resulting 3D point is at some altitude h where

$$(x', y') = \mathcal{P}_j(\mathcal{L}_i(x, y, h)). \tag{1.1}$$

[1] Also referred to as Rational Polynomial Camera model or Rational Function Model in the literature.

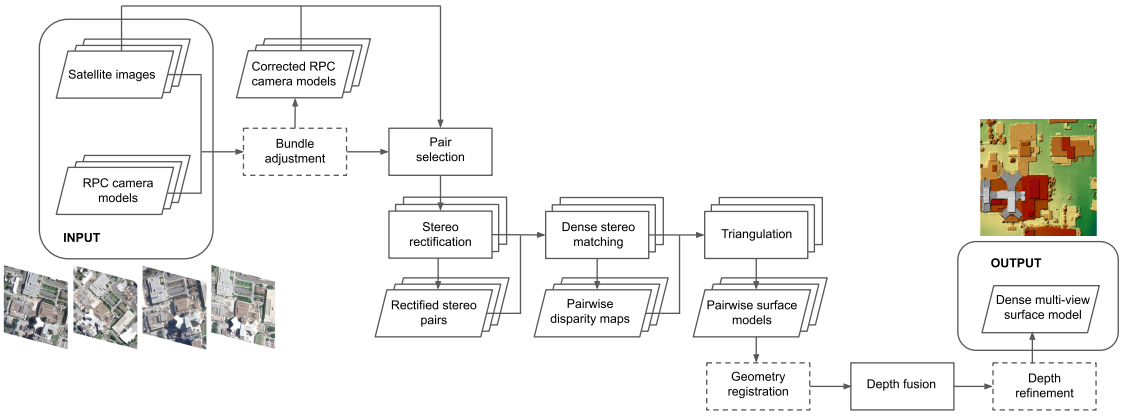


Figure 1.2: Classic steps of MVS digital surface modeling from multiple satellite images.

1.4.2 Satellite multi-view stereo pipelines

In the context of optical satellite imaging, 3D reconstruction is often equivalently referred to as digital surface modeling. The problem is usually addressed by means of multi-view stereo pipelines. This is the case, for instance, of the satellite stereo pipeline S2P by the Centre Borelli [dFMLM+14a], the NASA Ames stereo pipeline [BAM18], MicMac [RDPD17] by the the French National Geographic Institute (IGN), CARS [MSY+20] by the French Space Agency (CNES) or the CATENA [KdSG13] multi-stereo chain by the German Aerospace Center (DLR). Stereo-based digital surface modeling from satellite images has been extensively studied [Kus13, SAM+16, Qin17, GF18, LLJ+19, dR21] and previous works have shown that these methods can outperform complex true multi-view approaches [ODM+15, GRFvGG22]. Stereo vision methods acquire the relative depth information between two images by establishing corresponding image points. Dense correspondences are typically represented in the form of disparity maps, that contain the apparent motion from one image to the other [Sze10]. Built upon these axes, most satellite MVS pipelines share some fundamental steps, which are illustrated in the flowchart in Figure 1.2. The most common ones are reviewed in the following paragraphs.

Pair selection. This initial step aims to identify which are the most convenient stereo pairs to obtain an accurate geometry. A demonstration that using all possible pairs leads to worse results than using an adequate subset can be found in [FdFML17]. Assuming that all pairs cover the same exact geographical area, the most common choice to identify an optimal subset of pairs is to use the image metadata. The incidence angle of each view, the angle between the two views (also known as intersection angle), and the distance between the acquisition dates have been identified as key factors [dRM+14,FdFML17,Qin17,GF18]. Angles of incidence greater than 35-40 degrees are discarded to ensure a good view of the top of structures, and angles between views around 15-25 degrees are favored to work with a reasonable number of occlusions. Out of all pairs that satisfy the viewing angles criteria, short intervals between acquisition dates are prioritized, to minimize the possibility of major changes in the geometry and appearance of the scene. As pointed out in [GF18], there are exceptions to the latter: similar dates from different years may yield better results than distant dates of the same year; and dates belonging to the same season or not can also be a tie-breaking factor. Image similarity metrics or coarse feature matching can also be useful to complement criteria that relies on image metadata [MdFMLF19]. Another much more time-consuming criterion is to compute the geometry of all possible

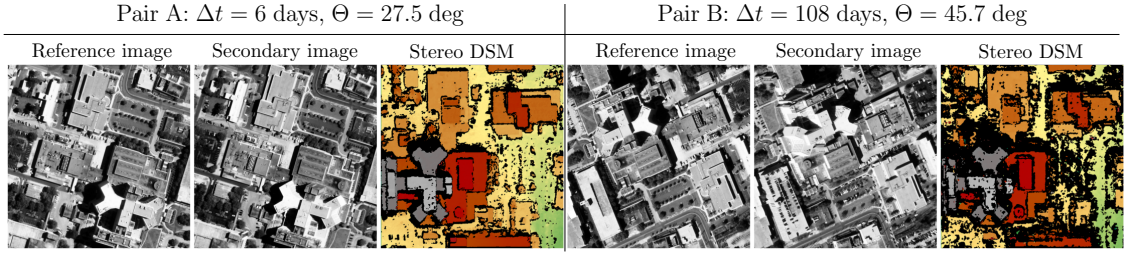


Figure 1.3: Effect of pair selection on stereo 3D reconstruction. The digital surface model (DSM) obtained with pair A is more accurate and complete than the equivalent for pair B, because the interval between acquisition dates (Δt) is smaller and the angle between views (Θ) is closer to the sweet spot (15-25 degrees). The DSM colors represent different altitude values, and incomplete areas are shown in black.

pairs and sort them according to some quality metric, such as less undefined values or error with respect to known data [FdFML17, Qin17]. The benefits of stereo pair selection are illustrated in Figure 1.3.

Stereo rectification. In close-range photography, stereo rectification uses epipolar geometry to turn lines containing corresponding points (epipolar lines) into horizontal lines with the same vertical or y-coordinate [HZ04]. This step reduces the search range for finding point correspondences between images, as subsequent matching algorithms only need to look in the x-axis direction. However, the geometry of satellite pushbroom sensors is different from the pinhole model commonly used for close-range imagery, causing epipolar lines to become curves. In addition, such curves are not conjugate: one curve in a satellite image does not have an equivalent curve in the other. Thus, traditional rectification algorithms cannot be applied with RPC models, which has led to a lack of consensus in the literature on how to approach this step. When geometry priors are available, images can be orthorectified in a common coordinate system [SAM+16, BAM18]. In the absence of geometry priors, a variety of tricks has been proposed. The majority of strategies are based on the assumption that pushbroom sensors can be locally modeled as affine cameras, i.e. parallel projection models with the camera center at infinity [OAH93, FY04]. This allows rectification by means of the epipolar affine fundamental matrix, which converts the epipolar curves into nearly conjugate horizontal lines, only non-coincident by a small vertical displacement [dFMLM+14b, BAM18]. The image pairs in Figure 1.3 have been rectified based on the latter idea. Another solution is to use generic pushbroom epipolar resampling methods. Such methods do not necessarily rely on the affine model assumption: instead, they exploit the idea that pairs of approximately conjugate curves can be constructed at local level with small error [OLT+10, WHL10, GF17]. The pairs of curves can be found by localizing points from one image at reasonable altitudes (e.g. a coarse ground elevation) and reprojecting them to the other image. After that, resampling transformations are applied to force corresponding curves into nearly horizontal lines.

Dense stereo matching. After rectification, the disparity map of each pair is estimated by establishing dense correspondences from one image to the other. Of all classic algorithms, the Semi-Global Matching (SGM) originally introduced in [Hir07], is the standard choice for satellite imagery because of its computational efficiency and compelling strategy for exploiting spatial regularity [dFMLM+14a, RDPD17, WHP+12, BAM18, MSY+20]. SGM produces disparity maps by minimizing an energy function comprising a unary matching cost and a regularization term that aggregates the costs of neigh-

bor points along different cardinal directions. A good number of pipelines use SGM with the census transform [ZW94] as matching cost for efficiency and robustness to illumination changes [MSY+20, WHP+12, dR21]. Others employ variants of the SGM algorithm. E.g. MGM or More Global Matching [FdFM15, FdFML17] improves SGM by injecting information from the perpendicular direction to the aggregated costs along each direction; tSGM or tube-based SGM [RWFH12, GF16, LLJ+19] implements a hierarchical coarse-to-fine method to limit the disparity search range; SGBM or Semi-Global Block Matching finds correspondences between fixed-size windows instead of individual pixels [BK00, LRF+19]; 3SGM uses automatically generated building footprints to discard information from previous pixels along a direction when a building edge is encountered [DDSS22].

An emerging line of research is also exploring the use deep neural networks (DNNs) to perform stereo matching with satellite imagery. Stereo matching DNNs usually comprise three modules: the feature extraction, whose output is used to construct a cost volume; the cost volume regularization, that aggregates costs over different disparity levels; and the disparity regression, that outputs the disparity map with minimum cost. Different architectures are tested in [WVPDR21] on aerial imagery with encouraging results. GA-Net [ZPYT19], which uses layers that are a differentiable approximation of the SGM algorithm, was tested on satellite images in [GRFvGG22]. GA-Net-Pyramid [XdF+22] takes inspiration from tSGM to restructure GA-Net with a more efficient pyramid architecture. Unlike SGM and other classic algorithms, neural networks have the ability to exploit semantic cues to infer disparities in ill-posed regions, such as low-textured areas. However, the networks tested so far require long training times and large geometry ground truth datasets, as they follow supervised learning. Interestingly, [AQ22] shows that SGM can be used to fine tune DNN methods for better generalizability across unseen satellite scenes. True multi-view matching DNN approaches, mainly inspired by MVSNet [YLL+18], are also being investigated for satellite imagery (e.g. CasMVSNet [GFZ+20] in [GRFvGG22] or SatMVS in [GLJ21]), but they have less presence than stereo-based methods in the literature.

Triangulation. This term designates the process of determining the location of a point in 3D space from its 2D observations on two or more images. The correspondences given by the pairwise disparity map are triangulated using the camera models of each pair, resulting in a collection of pairwise point clouds. This step is harder to solve with satellite RPC models in comparison to conventional pinhole cameras, where direct linear triangulation methods minimizing the algebraic error can be applied [HZ04]. The Ames stereo pipeline [BAM18] estimates a coarse center of projection for each satellite view and backprojects rays emanating from the center of projection and intersecting with the 2D observations. In practice, such approximation leads to pairs of rays that rarely intersect in the object space. The 3D point closest to each pair of rays is taken as the solution, and the shortest distance between the rays is taken as a confidence measure. Similarly, weak perspective pinhole camera models are approximated locally in [ZSS19], where it is claimed that triangulation of satellite imagery with conventional 3D geometry tools like COLMAP [SF16, SZFP16] is possible, at the cost of small errors in altitude accuracy. The weak perspective pinhole camera local approximation is less reliable than that of affine cameras, and is limited to a small variation in scene depth. The S2P [dFMLM+14a] pipeline achieves higher triangulation accuracy than the previous methods by directly employing the RPC camera models. In particular, it uses a reprojection strategy that takes a reference 2D observation and iteratively approximates the related epipolar curve. The 3D point that draws the epipolar curve closest to the associated 2D observation on the other image of the pair is taken as the solution. Other methods employing the RPC camera

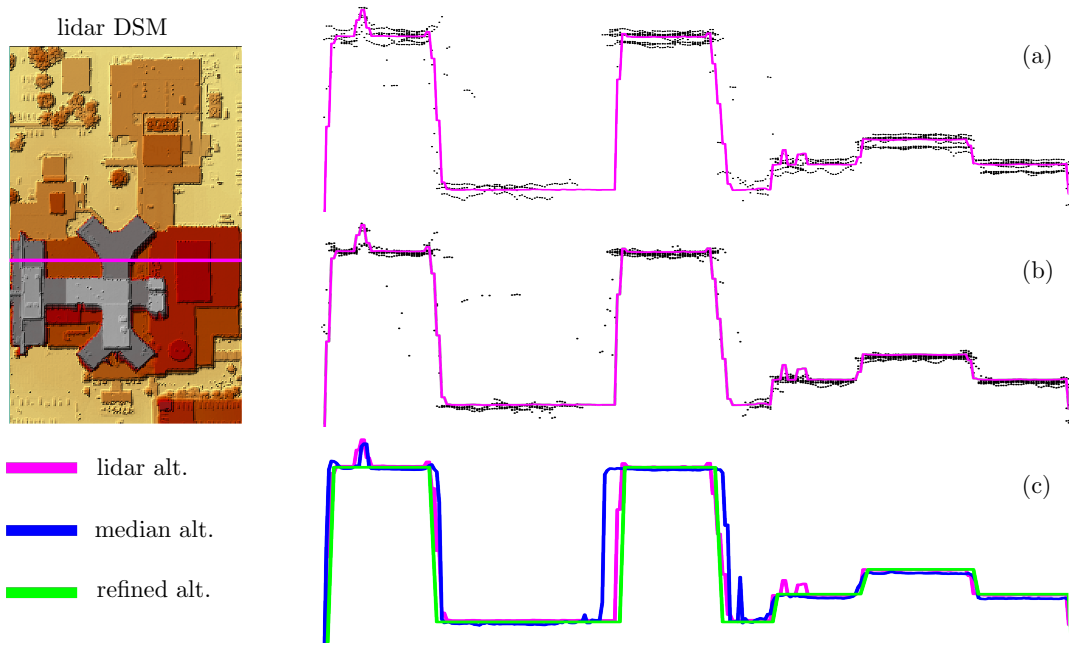


Figure 1.4: Depth fusion process. Scattered dots represent the pairwise altitude values derived from different stereo pairs and located on the magenta line shown on the lidar DSM. Pairwise altitude values are usually inconsistent due to small inaccuracies in the camera models of the satellite images, as in (a). Bundle adjustment and geometry registration methods can be used to align the pairwise altitudes, as in (b). Once the geometry of all pairs is consistent, they can be merged into a single multi-view model with a local filter (e.g., the median) and subsequent refinement steps (e.g., polygonalization), as in (c).

models for triangulation can be found in [TH02, ODM+14, ZWDF15].

Depth fusion. The point clouds obtained from individual stereo pairs are often sparse or subject to outliers. It is also possible that the individual pairs do not cover the entire area of interest. To obtain a more accurate, dense and complete model of the target area, the pairwise models are merged into a single digital surface model, i.e. the final multi-view model, as shown in Figure 1.4. Assuming that pairwise geometries are registered in the object space, it is usual to merge them by projecting the 3D point clouds into a digital surface model (DSM) and applying a point-wise filter. The DSM is a 2.5D image-like format in which the plane containing a surface is discretized into a 2D grid, where each 2D cell is assigned an altitude value. Multiple point clouds can be fused into a single DSM by merging the altitude values that fall inside the same DSM cells. Taking the median altitude is widely used because of its simplicity and robustness to large outliers [Kus13, GF18, dR21]. However, this strategy is not optimal in certain areas, such as in the DSM cells corresponding to building edges or vegetation. As observed in [FdFML17], such cells usually exhibit a distribution with two modes: one at low altitudes (e.g., the ground next to the building is observed or there are no leaves on the tree) and another at high altitudes (e.g., the roof of the building is observed or there are leaves on the tree). As a solution, altitude clustering-based strategies are proposed in [FdFML17], that prioritizes lower altitude centroids; or [RPDD18], that selects the output altitude by minimizing an energy function similar to SGM, in which altitude centroids are treated as candidate disparities. An adaptive depth fusion filter is introduced in [Qin17] to incorporate spatial information and color from orthophotos. [QLFR22] incorporates instead confidence

measures from dense matching algorithms (e.g., the point-wise costs of SGM): altitude candidates are ranked in decreasing confidence and divided in two groups, in such a way that the median of the highest-confidence group is used instead of the usual median if both values are significantly different. [KdG+17] investigates different variants of median altitude fusion, including TV-L_1 inspired spatial regularization terms, that lead to visual improvements in the output models. Another study in [dK12] demonstrates the benefits of pairwise matching with regularization and subsequent fusion versus early fusion of all matching costs and subsequent regularization.

Satellite MVS pipelines are usually completed with a series of pre-processing and post-processing steps that have a significant impact on the accuracy of the output 3D models. These important complementary steps are depicted in dashed boxes in Figure 1.2.

Bundle adjustment. Bundle adjustment approaches are a usual pre-processing step to correct geolocation inconsistencies between the RPC models of different satellite views of the same scene [Kus13, LLJ+19, dR21, AA20]. These geolocation inconsistencies are mainly caused by the lack of precision in the knowledge of the sensor orientation. In practice, they prevent a correct triangulation with RPC cameras, as they associate multi-image observations of the same point to different locations in the 3D space (see Figure 1.4(a)). In other words, the condition (1.1) cannot be satisfied. Bundle adjustment methods refine the polynomial coefficients of the RPC functions by minimizing the reprojection error of a set of reference points (or tie points) observed across multiple images [TMHF00]. Absolute correction strategies use Ground Control Points (GCP), i.e. sets of points whose latitude, longitude and altitude coordinates are known, while relative correction strategies usually replace GCP with correspondences of image features. The latter ensures that the geometry observed by the input views is at least self-consistent (see Figure 1.4(b)). The RPC correction methodologies can be direct, if the RPC coefficients are explicitly modified [XZ09]; or indirect, if the RPC functions are composed with complementary correcting functions [GD03, FH05]. It should be noted that bundle adjustment approaches are not the only existing solution to overcome the effects of RPC inaccuracies in digital surface modeling. However, the growth of optical satellite imagery and the emergence of new multi-image acquisitions, such as those delivered by small satellites, are increasing its weight in MVS pipelines, because of its suitability to handle multiple views at once. E.g. [BSAH21] reworked the Ames stereo pipeline to handle SkySat data. MicMac also comprises a bundle adjustment module [RDPD17].

Geometry registration. A small group of works has shown that it is possible to perform satellite multi-view 3D reconstruction even when the RPC camera models are subject to small inconsistencies. In these cases, the net effect of inaccuracies in the RPC functions is corrected locally for each pair, e.g., by carefully checking that only horizontal displacements persist after stereo rectification. The multi-view bundle adjustment is then discarded in favor of a step dedicated to the registration of pairwise models [FdFML17, Qin17]. This alignment or registration takes place before the fusion step and is based entirely on the geometry of the pairwise models, so that the camera models of the images can be ignored. Both [FdFML17] and [Qin17] align pairwise DSM by means of spatial shifts that are computed by optimizing some cost function (e.g., the normalized-cross correlation or the squared error) with respect to a reference DSM. The choice of a shift in the 3D space for geometry registration is justified by the local approximation of the sensor models as parallel projection models. The Ames stereo pipeline [BAM18] also includes 3D terrain alignment tools based on an Iterative Closest Point algorithm [PCSM13], which supports rotations between different models as well. The most common sources of error in DSM

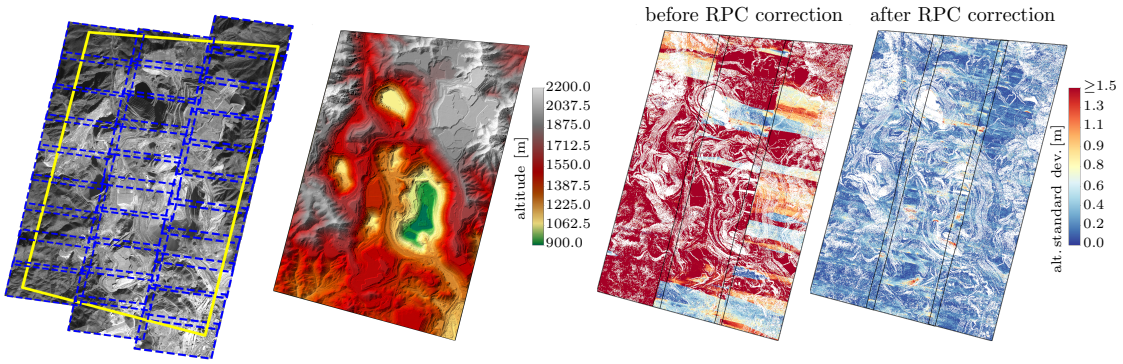
registration are reviewed in [NK11] and summarized as shifts, elevation-dependent biases, and higher-order sensor-specific biases.

Depth refinement. Occlusions, temporary objects (e.g., cars, vegetation, demolished structures) or erroneous correspondences from stereo matching can lead to 3D models with incomplete regions, local irregularities and outlier altitude values. Regularization in stereo matching can also lead to non-sharp object contours. These are undesirable features in digital surface modeling, since the world we live in is generally made of regular structures, especially in inhabited areas. Therefore, it is common for multi-view surface models to undergo some post-processing steps. Morphology, interpolation or inpainting routines can be used to fill incomplete DSMs [FdFM15, KGAL20]. Classic local filtering techniques can be used to remove noise and outliers by aggregating spatial information, but usually affect negatively detail sharpness and building edges, with more or less impact depending on the size of the spatial window [WW06]. One way to preserve and improve object contours is to incorporate segmentation labels. Early approaches for DSM segmentation used segment/contour detection and clustering algorithms. Building boundaries were typically identified by fitting 2D polygons [SdKR10] or heuristic classification methods based on altitude and surface normals [PY09]. Today, deep learning networks can be used to address this problem. A differential advantage of these methods is that they are not constrained to a limited list of building categories or geometric forms. The works of [BdKR18, BRK19, BKFR19, LBK20] propose different variants of a conditional adversarial generative network (cGAN) that improves building shapes in photogrammetric DSMs by encouraging the generation of equivalent DSMs difficult to discern from lidar-generated DSMs. Refined building shapes can be vectorized into 3D polygonal bodies [WZB21]. The concurrent ResDepth architecture [SS20] approaches the problem in a different way: instead of using adversarial learning and predicting absolute altitudes, the model is trained to learn the altitude differences (the residue) that bring the input DSM closer to a lidar-generated equivalent. Depth refinement with neural networks is remarkably improved by using orthophotos aligned with the input DSM, which encourages the identification of sharp contours. The main challenges remain the lack of large training databases and the ability to generalize across different urban landscapes.

The MVS flowchart in Figure 1.2 could be extended with further steps, such as changes in 3D representation (e.g., point could to mesh conversion [Kus13, LLJ+19]), semantic segmentation or texturing [Kus13, MLd19]. However, such additional tasks are beyond the scope of this thesis, as they are not strictly necessary in the 3D geometry estimation process.

1.5 Short outline of the thesis

This thesis is divided in two parts. The first part, comprising Chapters 3 to 5, studies the geometric modeling of multi-view satellite imagery and the crucial role of adequate sensor modeling to achieve accurate multi-view consistency. The characterization and correction of RPC camera models are the focus of these chapters, which discuss in detail the bundle adjustment and geometry registration steps of the satellite multi-view stereo pipeline (Figure 1.2). The second part, comprising Chapters 6 to 9, investigates different applications and problems related to digital surface modeling from multiple satellite images. A good number of the tasks covered in the second part take advantage of the geometric modeling concepts of the first part to address other issues of the MVS pipeline, such as the number of pairs needed to reconstruct an area of interest or the matching algorithm.



(a) SkySat acquisition

(b) MVS digital surface model, obtained from 144 stereo pairs

(c) Standard deviation between altitude values from different individual stereo pairs

Figure 1.5: The Morenci mine (United States) is reconstructed using multiple SkySat stereo pairs in Chapter 4. The bundle adjustment correction of the RPC camera models prior to the stereo reconstruction leads to multi-view geometry with sub-meter consistency.

Chapter 3: Geolocation correction methods for satellite multi-view stereo
This chapter investigates which strategy, out of the bundle adjustment and the geometry registration steps in Figure 1.2, leads to a higher accuracy in satellite MVS digital surface modeling. We compare three different methods to overcome the geolocation inconsistencies introduced by the input RPC camera models. Apart from this correction strategy, the other steps of the MVS pipeline are the same. The first method avoids taking action on the camera models by explicitly co-registering the altitude values derived from different pairs. The other two methods use bundle adjustment variants to correct the camera models. Both bundle adjustment algorithms rely on the local approximation of satellite cameras as affine projection models, but each one modifies the RPC models with a different corrective transformation. One bundle adjustment uses an offset or bias correction, in which a translation defined in the image domain is applied after the RPC projection function; and the other uses a rotation in the object space, which is applied before the projection function. The three methods are tested on different areas of interest, using 50 input pairs of WorldView-3 panchromatic images and a reference lidar DSM. The results reveal that the geometry registration method is most robust when the bundle adjustment tie points do not form a single connected component in the connectivity graph, involving subsets of disconnected images. Otherwise, bundle adjustment strategies achieve the best results, which is explained by the fact that geometry registration methods are sensitive to incomplete parts or outliers in the altitude models derived from single pairs.

Chapter 4: Bundle adjustment of RPC camera models
This chapter delves into the problem of correcting the inaccuracies of RPC camera models in satellite multi-view stereo pipelines. An improved bundle adjustment methodology that builds upon one of the bundle adjustment strategies previously introduced in Chapter 3 is proposed. In this case, multi-view consistency is achieved by correcting sensor orientation-related errors by direct composition of the original RPC functions with a corrective rotation matrix in the object space. By directly employing the original RPC functions, the local approximation of affine projection models is avoided. Other significant improvements include an optimal tie point selection criterion to handle larger multi-view collections efficiently, and an RPC fitting algorithm that is used to ensure that the output camera models follow the same format as the input ones. The stereo pair selection criterion is also adapted to work with SkySat images, which can form pairs with limited

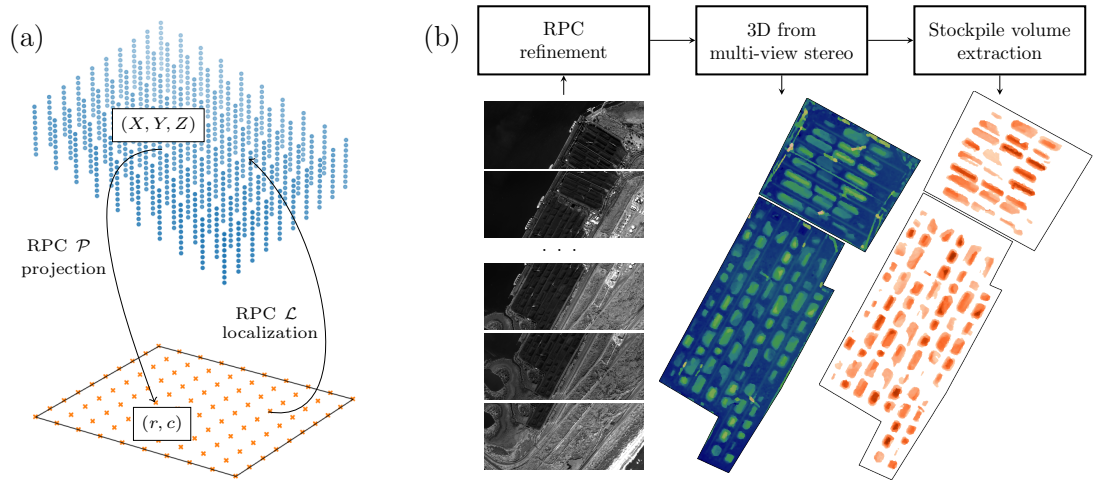


Figure 1.6: (a) The RPC fitting algorithm described in Chapter 5 estimates the geolocation functions based on a grid of virtual correspondences between 3D points (X, Y, Z) and 2D image coordinates (r, c). (b) The surface monitoring approach in Chapter 6 uses the strategy in (a) to correct the RPC functions of SkySat acquisitions captured on different dates over a coal terminal. The corrected camera models are used to build an accurate 3D model for each acquisition date and monitor changes in coal stockpile volume over time.

overlap and small baseline, as different images from the same acquisition are captured almost simultaneously. The resulting method is tested using SkySat acquisitions over two different areas of interest, and the output RPC models are subsequently used in a satellite stereo pipeline to extract digital surface models. In the absence of ground truth models of the observed areas, the deviation between altitude values from different pairwise models is used as evaluation metric, as shown in Figure 1.5. The results indicate that the proposed bundle adjustment leads to better pairwise geometry registration against explicit geometry registration algorithms. Tie point reprojection errors reach subpixel values and the altitude standard deviation is reduced from the order of meters to tens of centimeters.

Chapter 5: Robust RPC camera modeling
This chapter describes in detail the RPC fitting algorithm used in the bundle adjustment pipeline presented in Chapter 4. The algorithm follows a terrain-independent regularized least squares approach to find the coefficients of the RPC functions that relate a set of virtual correspondences between image space and object space. The input virtual correspondences have to follow a regular grid structure, as shown in Figure 1.6(a). The grid has to cover the entire altitude range of the scene, and has to project onto image point observations that also cover the entire image. These virtual correspondences are generated using a different sensor representation, such as a physical sensor model or a previous RPC model corrected by some transformation. We test the algorithm using Sentinel-1 SAR and WorldView-3 optical satellite images. The performance of the method is assessed by varying the density of point correspondences and the size of the area that they cover. The output RPC models achieve reprojection errors on the order of 10^{-4} pixels in most configurations, which shows the ability of the algorithm to capture complex geolocation functions in the form of RPC models.

Chapter 6: Automatic stockpile volume monitoring
This chapter addresses a real use case of a satellite MVS pipeline. Recurrent SkySat push-frame imagery is used to monitor stockpile volume in a coal terminal, as shown in

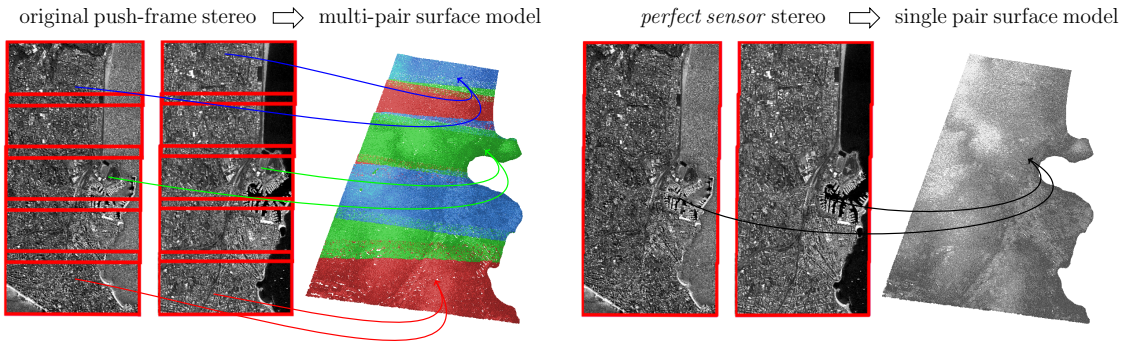


Figure 1.7: The image stitching method described in Chapter 7 combines strips of partially overlapping images (e.g., acquired by a push-frame system) into a single large-scale image. The geometry of the mosaic is described using a *perfect sensor* model, as if it had been acquired by an ideal pushbroom sensor. By constructing large-scale mosaics from smaller image collections, along with their perfect sensor model, satellite digital surface modeling can transition from a challenging multi-pair problem to a single pair problem.

Figure 1.6(b). The SkySat images are first grouped according to the acquisition date, and the RPC camera models of each group are corrected using the bundle adjustment method presented in Chapter 4. The MVS pipeline is then applied using the corrected RPC models, to obtain the digital surface model of each acquisition date. Thanks to the prior bundle adjustment, the geometry of the image pairs used at each date is registered with sub-meter accuracy. The resulting sequence of multi-view 3D models is used to measure the evolution of stockpile volume. In particular, the stockpile volume in each 3D model is obtained by subtracting a terrain model and integrating altitude values above ground level. The volume estimates are validated using measurements collected on site over the same time period. Both distributions are strongly correlated, thus highlighting the enormous potential of recurrent multi-image acquisitions with strong multi-view consistency.

Chapter 7: Perfect sensor localization for push-frame image stitching
This chapter addresses the problem of reducing the number of stereo pairs needed to reconstruct an area observed in push-frame strips of multiple images. Satellite images acquired by push-frame systems are partially overlapped along the direction of satellite motion. We propose a geometrically consistent image stitching method (also known as image mosaicing) that merges small images from the same push-frame strip into a single large-scale image, denoted L1B$^+$. An example of the input and output images of the methodology is shown in Figure 1.7. The output mosaic is obtained by warping each input image into a common image space with a projective transformation. The method also produces a *perfect sensor* model, in RPC format, that describes the geometry of the L1B$^+$ image, as if the mosaic had been acquired by an ideal pushbroom sensor. The RPC fitting algorithm described in Chapter 5 is used to derive the perfect sensor model, based on virtual correspondences between image space and object space that are found using the RPC models of the original push-frame strip after a bundle adjustment. Different pairs of SkySat push-frame strips of 3 and 5 images are used to test the method, and the output L1B$^+$ images are subsequently exploited to reconstruct the observed areas with a single stereo pair. The results indicate that the L1B$^+$ models achieve higher altitude accuracy with respect to MVS models derived from multiple pairs of the original push-frame images. This is mainly explained by the fact that L1B$^+$ models are not subject to registration errors between altitude values derived from different image pairs.

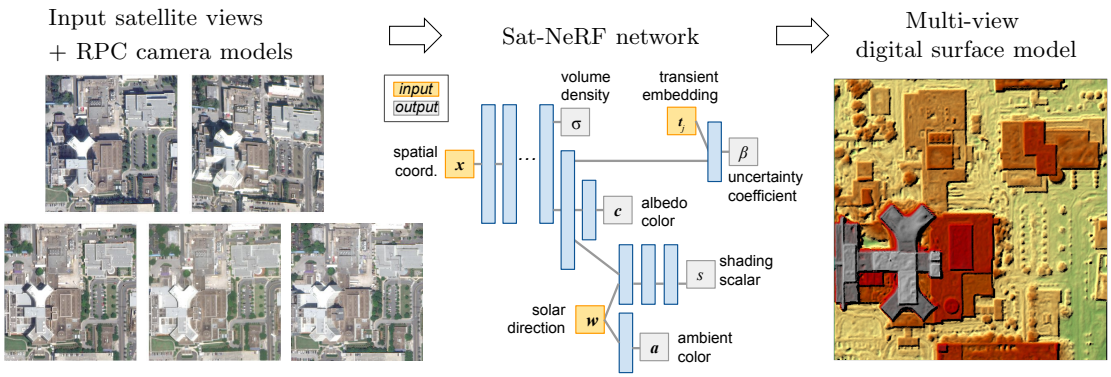


Figure 1.8: The Sat-NeRF network presented in Chapter 8 adapts state-of-the-art neural rendering networks to multi-date satellite imagery. Differently from conventional stereo-based pipelines, Sat-NeRF can be used for real multi-view digital surface modeling. Using N input views, the network learns the geometry and appearance at each 3D point of the scene, with robustness to transient objects and variable illumination.

Chapter 8: Satellite NeRF

This chapter aims to replace all the steps of the MVS flowchart shown in Figure 1.2, with the sole exception of an initial bundle adjustment of the RPC camera models. For this purpose, we take inspiration from state-of-the-art neural rendering networks and introduce the Satellite NeRF or Sat-NeRF model. The Sat-NeRF network is based on the original NeRF model for learning radiance fields, i.e. a continuous representation of geometry and appearance of a 3D scene, which is learned without explicit geometry supervision. Sat-NeRF is tailored to handle multi-date satellite imagery, it employs satellite RPC cameras instead of pinhole cameras and circumvents some major constraints of NeRF, that is limited to static scenes with invariant illumination. As illustrated in Figure 1.8, Sat-NeRF learns the geometry and appearance of permanent structures simultaneously using a main backbone of layers, that only takes spatial coordinates as input. Shadows and transient objects like cars are learned separately through secondary heads that have access to the direction of solar rays and image-specific embedding vectors describing transient attributes. The cost function is also adapted to favor dissociation between the permanent and the temporary. As a result, the method is able to predict, for each 3D point of the scene: the volume density, which defines the geometry; the color, which is a linear combination of an albedo with a shading based on the direct illumination from solar rays and the ambient illumination; and an uncertainty coefficient, that represents the probability that the point belongs to a transient object or not. Querying Sat-NeRF at multiple points in a given viewing direction allows realistic novel view synthesis and digital surface modeling. We train Sat-NeRF on different areas using 10 to 20 WorldView-3 RGB images and obtain similar altitude accuracy with respect to conventional satellite MVS pipelines with manual selection of input pairs. The main advantages of Sat-NeRF are the ability to achieve greater detail in surface modeling and the true multi-view nature of the method. The main drawbacks are training times, the need for multiple views and the presence of local irregularities in the learned geometry.

Chapter 9: Disparity estimation networks

This chapter focuses on the dense stereo matching block of MVS pipelines. Unlike SGM and other classic algorithms for disparity map estimation, neural networks have the ability to exploit semantic cues to resolve ill-posed regions, such as low-textured areas. Based

on this observation, we investigate whether neural networks for disparity estimation can be plugged into satellite stereo pipelines to improve the output geometry. Two supervised learning architectures of convolutional neural networks, Pyramid Stereo Matching (PSM) and Hierarchical Stereo Matching (HSM), are reviewed. Both networks follow a common structure of feature extraction, cost volume regularization and disparity regression modules. However, the way the cost volume is constructed in each case is significantly different. PSM builds a single cost volume by concatenating feature maps at different disparity levels, while HSM builds a multi-scale pyramid of 4 cost volumes using the absolute differences between feature maps. The explicitly multi-scale and less voluminous representation that HSM uses for cost volumes is intended to improve and accelerate the processing of large high-resolution images. For a fair comparison against concurrent classic algorithms, in the experimental part, we use pre-trained weights previously fine-tuned on an aerial image benchmark. The latter guarantees that the networks are familiar with Earth observation scenes, even if image resolution is not the same. Four sets of 30 WorldView-3 RGB and panchromatic image pairs are used as input to evaluate the PSM and HSM architectures, as well as the MGM variant of the classic SGM algorithm. A satellite stereo reconstruction pipeline is run with the different matching methods, and the output DSMs are evaluated with a reference lidar DSM. The results show that, under ideal conditions (e.g., suitable baseline, similar acquisition dates), deep learning methods provide higher numerical accuracy compared to classic SGM variants. PSM provides a higher level of detail, but becomes impractical as image size increases. The main weaknesses of the two networks are their limitation to a predefined disparity range, their sensitivity to the stereo rectification of the input pairs, and the need to adjust the image size to obtain optimal disparity maps.

1.6 Summary of contributions

This thesis revisits multi-view stereo pipelines for digital surface modeling from satellite images. The objective is to facilitate the use of multiple input views instead of a single stereo pair. For this purpose, the first contribution of this thesis is a bundle adjustment methodology to refine the RPC camera models of a set of satellite images (Chapter 4), which is the result of a review of the methods in the literature (Chapter 3). The bundle adjustment methodology includes the description of an RPC fitting algorithm (Chapter 5), that allows to obtain the refined camera models in the same format as the input ones. The refined camera models lead to consistent multi-view geometry with sub-meter accuracy. This opens the door to simplify multi-view stereo pipelines (no need to register altitude values derived from different pairs) and use real multi-view methods. As a result of the latter, the next contribution of this thesis is an image stitching algorithm named L1B$^+$ that merges a band of small, partially overlapping satellite images into a single large-scale image (Chapter 7). We show that it is possible to derive a single camera model for the output image, as if it had been acquired with an ideal pushbroom scanner, under the condition that the RPC models of the input images are geometrically consistent. The latter allows satellite digital surface modeling to transition from a challenging multi-pair problem to a single pair problem. Another contribution of this thesis is the Sat-NeRF network (Chapter 8), which is inspired by state-of-the-art self-supervised neural rendering approches. Sat-NeRF is a real multi-view method for digital surface modeling from multi-date satellite images, it does not rely on a selection of stereo pairs. Without any geometry supervision, the network is able to infer the surface observed in a collection of satellite images by learning the geometry and appearance of permanent structures simultaneously using a main backbone of layers, while shadows and transient objects are

learned separately using secondary heads. Lastly, as secondary contributions of this thesis, we find an application of satellite multi-view stereo to a real use case for stockpile volume monitoring (Chapter 6) and an experimental evaluation of different deep learning networks for dense stereo matching against classic algorithms similar to SGM (Chapter 9).

1.7 List of publications

Presented in this thesis

- Roger Marí, Carlo de Franchis, Enric Meinhardt-Llopis, and Gabriele Facciolo. To bundle adjust or not: A comparison of relative geolocation correction strategies for satellite multi-view stereo. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 2188–2196, 2019. https://doi.org/10.1109/ICCVW.2019.00274
- Roger Marí, Carlo de Franchis, Enric Meinhardt-Llopis, Jérémy Anger, and Gabriele Facciolo. A generic bundle adjustment methodology for indirect RPC model refinement of satellite imagery. *Image Processing On Line*, 11:344–373, 2021. https://doi.org/10.5201/ipol.2021.352
- Roland Akiki, Roger Marí, Carlo de Franchis, Jean-Michel Morel, and Gabriele Facciolo. Robust rational polynomial camera modelling for SAR and pushbroom imaging. In *2021 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 7908–7911, 2021. https://doi.org/10.1109/IGARSS47720.2021.9554583
- Roger Marí, Carlo de Franchis, Enric Meinhardt-Llopis, and Gabriele Facciolo. Automatic stockpile volume monitoring using multi-view stereo from SkySat imagery. In *2021 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 4384–4387, 2021. https://doi.org/10.1109/IGARSS47720.2021.9554482
- Roger Marí, Thibaud Ehret, Jérémy Anger, Carlo de Franchis, and Gabriele Facciolo. L1B+: A perfect sensor localization model for simple satellite stereo reconstruction from push-frame image strips. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 5-1-2022:137–143, 2022. https://doi.org/10.5194/isprs-annals-V-1-2022-137-2022
- Roger Marí, Gabriele Facciolo, and Thibaud Ehret. Sat-NeRF: Learning multi-view satellite photogrammetry with transient objects and shadow modeling using RPC cameras. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1310–1320, 2022. https://doi.org/10.1109/CVPRW56347.2022.00137
- Roger Marí, Thibaud Ehret, and Gabriele Facciolo. Disparity estimation networks for aerial and high-resolution satellite images: A review. *Image Processing On Line*, 12:501–526, 2022. https://doi.org/10.5201/ipol.2022.435

Others

- Manuel Lopez, Roger Marí, Pau Gargallo, Yubin Kuang, Javier Gonzalez-Jimenez, and Gloria Haro. Deep single image camera calibration with radial distortion. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11809–11817, 2019. https://doi.org/10.1109/CVPR.2019.01209

- Thibaud Ehret, Roger Marí, and Gabriele Facciolo. NeRF, meet differential geometry! *arXiv preprint arXiv:2206.14938*, 2022. https://arxiv.org/abs/2206.14938

2 Introduction (en français)

2.1 Motivation

Le terme *télédétection* désigne la science qui consiste à obtenir des informations sur une scène à partir de l'analyse de données acquises par un dispositif qui n'est pas en contact avec la scène observée [LKC15]. Nous utilisons régulièrement une forme simple de télédétection : nos yeux agissent comme des capteurs qui reçoivent le rayonnement émis par les objets qui nous entourent, que notre cerveau analyse pour en extraire des connaissances. Ceux qui ont une bonne vision de loin sont conscients des avantages de la télédétection : ils peuvent apprendre des détails que d'autres ne peuvent pas connaître, tout en bénéficiant d'une position sans risque et de plus de temps pour réagir aux événements.

En télédétection, les données acquises sont généralement des rayonnements électromagnétiques. Alors que l'œil humain ne peut voir qu'une gamme limitée de fréquences, les capteurs artificiels peuvent couvrir l'ensemble du spectre électromagnétique. Selon l'origine du rayonnement, les capteurs de télédétection peuvent être divisés en deux catégories :

- Télédétection active : Le rayonnement mesuré correspond à un rebond d'énergie initialement généré par une source artificielle dirigée vers la scène. Les exemples classiques sont les scanners lidar (LIght Detection Imagery And Ranging) aéroportés ou les systèmes radar comme les satellites SAR (Synthetic-Aperture Radar).
- Télédétection passive : Le rayonnement mesuré estle rayonnement solaire réfléchi (imagerie optique) et/ou le rayonnement émis par les cibles (imagerie thermique). Les exemples classiques sont les satellites ou les drones équipés de caméras.

L'imagerie optique par satellite est le sujet d'étude de cette thèse. Cette branche de données est d'un intérêt particulier pour l'observation de la Terre pour diverses raisons : par exemple, la couverture de grandes zones, l'accès non intrusif à toute localisation, l'indépendance par rapport au support terrestre ou au guidage humain, la trajectoire stable et prévisible ou les périodes de revisite automatique. D'une certaine manière, les satellites optiques d'observation de la Terre reproduisent le schéma fondamental de la perception visuelle humaine. Au lieu d'un système de caméra (les yeux) monté sur un dispositif terrestre (le corps), ils utilisent un système de caméra monté sur une structure flottante, en orbite, qui observe à des centaines de kilomètres de distance. Cependant, il manque un élément clé : la méthodologie d'analyse qui convertit les données acquises en informations utiles. Cette méthodologie est un domaine en constante évolution, comme le sont la technologie des satellites et la surface de la Terre. La nécessité de revoir les méthodes connues et de les mettre à jour constitue le point de départ de ce travail.

En particulier, cette thèse se focalise sur les applications de l'imagerie satellitaire optique liées à la modélisation de la surface terrestre. Alors que les images optiques sont bidimensionnelles, la troisième dimension, l'altitude de la scène, représente généralement une information essentielle pour un grand nombre d'applications. C'est le cas, par exemple, pour des activités telles que la prévention et la réponse aux catastrophes naturelles, la gestion du territoire, l'exploitation minière, la gestion des déchets ou la planification urbaine. La détection des changements est généralement au cœur de ces activités. Étant donné que les changements se produisent dans un espace tridimensionnel, disposer de mesures dans le même domaine est un moyen efficace de les caractériser pleinement. En même temps, les changements dans l'apparence des images peuvent ne pas correspondre à des changements réels dans la géométrie de la scène, mais à des facteurs externes tels que l'exposition au soleil ou la météo [dF15]. Par conséquent, nous pourrions dire que pour les activités liées à l'observation de la Terre, une image vaut probablement mille mots, mais un bon modèle de surface peut être inestimable.

2.2 Origines et caractéristiques de l'imagerie satellitaire optique

En 1972, les États-Unis ont lancé Landsat 1, le premier satellite d'observation de la Terre qui fournissait des images en format numérique et multispectral [TGH08]. Les caméras multispectrales *whisk broom*, telles que celles des premières unités Landsat, ont été les principaux capteurs d'observation de la Terre au cours des années 1970 et 1980. Ces instruments utilisaient une logique de balayage *across-track*, avec un miroir rotatif réfléchissant la lumière sur un seul capteur et enregistrait un pixel à la fois [LKC15]. Cependant, les pièces mobiles rendent ce type de capteurs coûteux et susceptibles d'être endommagés. En 1982, l'Agence spatiale française a lancé SPOT (*Satellite Pour l'Observation de la Terre*), le premier satellite en orbite à utiliser une caméra *pushbroom*, suivant une logique de balayage *along-track*. Les caméras pushbroom utilisent un capteur linéaire (généralement des dispositifs à transfert de charges ou CCD), qui est disposé perpendiculairement à la direction du mouvement du satellite et capture une ligne de pixels à la fois. Les différents dispositifs du capteur peuvent avoir une sensibilité variable, de sorte qu'ils doivent être parfaitement calibrés pour éviter des discontinuités dans l'image de sortie. Après SPOT, les capteurs pushbroom sont devenues régulièrement utilisées dans les satellites d'observation de la Terre (par exemple, IKONOS, QuickBird, GeoEye ou WorldView [PT12]).

L'imagerie satellitaire optique est généralement caractérisée par les grandeurs suivantes :

- Résolution spectrale : Elle correspond au nombre et à la gamme de bandes spectrales capturées. On distingue les images *panchromatiques*, comprenant un seul canal sensible à une large plage de fréquences ; les images *multispectrales*, comprenant quelques bandes de grande largeur telles que le bleu, le vert, le rouge et le proche infrarouge ; et les images *hyperspectrales*, comprenant une centaine ou plus de bandes spectrales contiguës et étroites.

- Résolution radiométrique : Elle correspond au nombre de bits utilisés pour exprimer les valeurs numériques dans les canaux de l'image. Par exemple, les images de 8 bits sont représentées par des valeurs de 0 à 255 (2^8 valeurs).

- Résolution temporelle : Dans le cas des satellites en orbite, la résolution temporelle est communément considérée comme le temps que prend le satellite pour capturer à nouveau une zone géographique particulière. Elle est également appelée période de revisite.

- Résolution spatiale : Désigne la plus petite taille d'objet physique que le capteur peut distinguer. Elle est généralement exprimée en distance d'échantillonnage au sol (GSD), qui fait référence aux dimensions d'un pixel mesurées au sol, en mètres par pixel.

Les résolutions spectrale, radiométrique, temporelle et spatiale des images satellites définissent leurs applications potentielles. Par exemple, les images panchromatiques à canal unique offrent généralement une meilleure résolution spatiale, mais les images à plus haute résolution spectrale décrivant des couleurs ou des données non visibles à l'œil humain peuvent être plus intéressantes pour des tâches telles que la segmentation, l'analyse de l'humidité du sol, la bathymétrie ou la surveillance de la pollution.

Dans cette thèse, les résolutions temporelle et spatiale jouent un rôle majeur, car elles sont essentielles pour modéliser la surface de la Terre de manière récurrente et détaillée. Heureusement, la technologie des satellites a fait de grands progrès au cours des cinquante dernières années. Landsat 1 offrait une résolution grossière d'environ 60 à 80 m par pixel et une période de revisite de 18 jours [Tod77, WGA06]. SPOT 1 a permis d'acquérir des images avec une résolution de 10 m par pixel et a ouvert la voie aux satellites suivants avec une résolution spatiale plus élevée encore [dF15]. IKONOS (du mot grec *eikōn*, qui signifie *image*), lancé en 1999, a été le premier à capturer des images à haute résolution à 1 m par pixel [Gro01] accessibles au public. WorldView-3, lancé en 2014, fournit l'imagerie commerciale à la plus haute résolution jusqu'à présent, qui atteint environ 30 cm par pixel [TJ16]. Comme la résolution spatiale, la résolution temporelle est également en train d'augmenter, passant de l'ordre de quelques jours à seulement quelques heures. Cette tendance est principalement due aux nouvelles constellations de petits satellites, qui combinent les passages de nombreuses unités afin de réduire les périodes de revisite.

2.3 De nouvelles constellations de petits satellites

Les constellations de satellites conventionnels se composent généralement d'une ou de quelques grandes unités, de plus de 1000 kg, en orbite à des altitudes élevées, entre 600 et 800 km. Par exemple, Landsat, Sentinel, SPOT ou WorldView-2 et WorldView-3 [ZSL+17]. Grâce à leur taille, ces satellites peuvent accueillir de grands capteurs, dont les images peuvent couvrir de larges bandes au sol de plusieurs dizaines ou centaines de kilomètres [TJ16]. Cependant, ces constellations de satellites de grande taille sont souvent coûteuses à développer et à maintenir sur une durée de plusieurs années. En outre, le nombre réduit d'unités limite généralement les périodes de revisite à quelques jours.

Dans ce contexte, les petits satellites ou *smallsats*, de moins de 500 kg, attirent de plus en plus l'attention de l'industrie [SBD10]. Les premières expériences avec de petits satellites remontent à 1999 à l'université de Stanford [HPSM+00], mais la véritable irruption a eu lieu en 2014 lorsque la société américaine Planet (anciennement Planet Labs) a lancé une flotte de 93 satellites d'environ 5 kg chacun, qui a évolué jusqu'à devenir la constellation PlanetScope [WBPVLdCGT17]. En 2022, PlanetScope est composée de 130 unités et est capable de capturer quotidiennement la totalité de la surface de la Terre avec une résolution de 3 à 5 m par pixel. Planet contrôle également d'autres constellations de petits satellites,

comme SkySat. Les satellites SkySat ont une masse de 110 kg, avec 14 unités en orbite en 2022. Ils fournissent une résolution de 0.6 à 0.8 m par pixel et une impressionnante période de revisite combinée inférieure à un jour dans le monde entier [Pla22]. Parmi les autres exemples de constellations de petits satellites fournissant des images optiques, on trouve celles des sociétés Satellogic ou BlackSky [CMT20].

Les satellites de petit format sont économiques et peuvent orbiter à de basses altitudes de 400 à 500 km sans consommer beaucoup d'énergie, ce qui est avantageux pour observer les choses avec un bon niveau de détail. Le principal inconvénient est que leur nature limite les capteurs qui peuvent être montés sur une telle structure. Par conséquent, les systèmes d'imagerie à haute résolution des petits satellites comme SkySat sont limités à l'observation d'une petite zone à la fois. Une seule image est alors limitée à une petite zone géographique. Pour compenser cela, des modes d'acquisition innovants sont utilisés pour couvrir de grandes zones. Par exemple, le mode d'acquisition *push-frame* des satellites SkySat capture des séquences d'images qui se chevauchent partiellement lorsque le satellite se déplace.

2.4 Reconstruction 3D multi-vues à partir d'images satellites

La reconstruction 3D multi-vues est un problème classique de la vision par ordinateur. En général, les méthodes de reconstruction 3D multi-vues peuvent être divisées en approches stéréo multi-vues (ou MVS, de l'anglais *multi-view stereo*) et approches réellement multi-vues [FH15]. Les méthodes MVS divisent le problème en plusieurs paires d'images, qui sont reconstruites indépendamment, puis fusionnées en un seul modèle final multi-vues. Par contre, les approches multi-vues réelles utilisent toutes les vues disponibles en même temps pour déduire la géométrie sous-jacente.

Cette section présente d'abord les modèles géométriques de caméra utilisés pour la reconstruction 3D multi-vues à partir d'images satellites. Elle aborde ensuite les étapes les plus courantes de la stéréo multi-vues à partir d'images satellite, qui est l'approche prédominante adoptée dans les chaînes de production [dF15, Gon21].

2.4.1 Le modèle de caméra RPC (Coefficients Polynomiaux Rationnels)

Les algorithmes présentés dans cette thèse fonctionnent avec des images satellites acquises par des capteurs pushbroom, dont la géométrie est illustrée dans la Figure 2.1. Les capteurs pushbroom acquièrent une ligne de pixels à la fois, ce qui implique que les paramètres d'orientation et le centre de perspective changent lorsque le satellite se déplace [Gro01]. Par conséquent, le modèle de caméra sténopé ou *pinhole* (en anglais), avec un seul centre de projection, qui est couramment utilisé pour l'imagerie conventionnelle, ne peut pas être utilisé pour modéliser le processus d'acquisition pushbroom avec précision.

Lorsque tous les phénomènes et éléments physiques impliqués dans le processus d'acquisition sont connus, celui-ci peut être représenté comme une chaîne d'opérations qui modélise ces facteurs, dans ce que l'on appelle un modèle *physique* ou *rigoureux* du capteur [dFMLGF15]. Cependant, les modèles physiques des capteurs sont souvent complexes et spécifiques à chaque plateforme satellitaire, ce qui rend difficile la définition d'algorithmes génériques capables de traiter différentes sources d'images.

Comme solution, depuis la mise en orbite de IKONOS en 1999, le modèle de caméra RPC (Coefficients Polynomiaux Rationnels) a été largement adopté par l'industrie pour

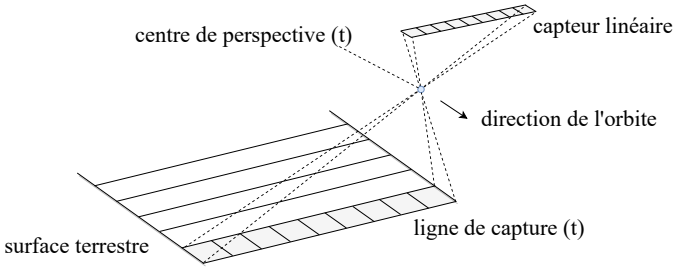


Figure 2.1: Schéma d'acquisition de données à partir d'un capteur pushbroom.

travailler avec les images satellites [Gro01]. Le modèle RPC est une représentation abstraite du processus d'acquisition, indépendante des propriétés physiques spécifiques du capteur. Comme son nom l'indique, le processus d'acquisition est représenté à l'aide de fonctions polynomiales rationnelles. Les modèles RPC comprennent une fonction de *projection* $\mathcal{P} : \mathbb{R}^3 \to \mathbb{R}^2$, qui projette les points 3D sur le plan de l'image, et son inverse, la fonction de *localisation* $\mathcal{L} : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^3$. La fonction de projection prend en entrée des coordonnées de latitude, longitude et altitude, et les fait correspondre aux lignes et colonnes de l'image ; tandis que la fonction de localisation prend en entrée des coordonnées de ligne et de colonne ainsi qu'une altitude donnée, et produit la latitude et la longitude correspondantes.

Les fonctions RPC permettent de trianguler l'emplacement d'un point 3D observé dans deux ou plusieurs images. Si le point (x, y) de l'image i correspond au point (x', y') de l'image j, alors le point 3D résultant se trouve à une certaine altitude h telle que

$$(x', y') = \mathcal{P}_j(\mathcal{L}_i(x, y, h)). \tag{2.1}$$

2.4.2 Stéréo multi-vues à partir d'images satellites

Dans le contexte de l'imagerie satellitaire optique, la reconstruction 3D est souvent appelée de manière équivalente modélisation numérique de surface. Ce problème est généralement abordé en utilisant des systèmes stéréo multi-vues (MVS). C'est le cas, par exemple, de la chaîne S2P du Centre Borelli [dFMLM+14a] ou des logiciels Ames *stereo pipeline* de la NASA [BAM18], MicMac [RDPD17] de l'Institut national de l'information géographique et forestière (IGN), CARS [MSY+20] de l'Agence spatiale française (CNES) ou CATENA [KdSG13] du Centre aérospatial allemand (DLR).

La modélisation numérique de surface basée sur la stéréo à partir d'images satellites a été largement étudiée [Kus13, SAM+16, Qin17, GF18, LLJ+19, dR21] et des travaux précédents ont montré que ces approches peuvent surpasser les approches multi-vues réelles plus complexes [ODM+15, GRFvGG22]. Les méthodes de vision stéréo acquièrent l'information relative de profondeur entre deux images en établissant des correspondances de points d'image. Ces correspondances sont généralement représentées sous la forme de cartes de disparité, qui contiennent le mouvement apparent d'une image à l'autre [Sze10]. Fondées sur ces axes, la plupart des chaînes MVS pour les images satellites partagent certaines étapes fondamentales, qui sont illustrées dans le diagramme de la Figure 2.2. Les plus courantes sont analysées dans les paragraphes suivants.

Sélection de paires. Cette étape initiale vise à identifier les paires stéréo les plus adéquates pour obtenir une géométrie précise. Une démonstration que l'utilisation de

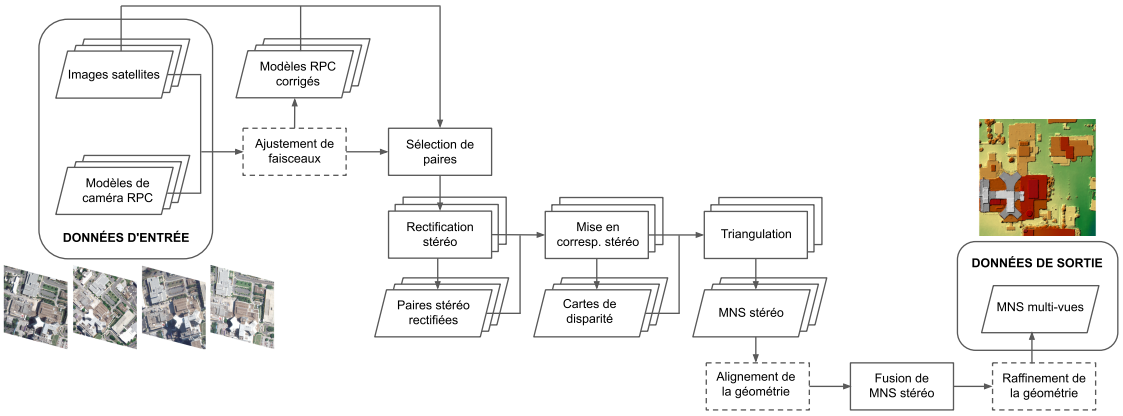


Figure 2.2: Étapes de la stéréo multi-vues (MVS) à partir d'images satellites. Les modèles numériques de surface (MNS) de plusieurs paires sont fusionnés pour obtenir un MNS final.

toutes les paires possibles conduit à des résultats moins précis que l'utilisation d'un sous-ensemble adéquat peut être trouvée dans [FdFML17]. En supposant que toutes les paires couvrent exactement la même zone géographique, le choix le plus commun pour identifier un sous-ensemble optimal de paires est d'utiliser les métadonnées des images. L'angle d'incidence de chaque vue, l'angle entre les deux vues (également connu comme l'angle d'intersection) et l'écart entre les dates d'acquisition ont été identifiés comme des facteurs clés [dRM+14, FdFML17, Qin17, GF18]. Les angles d'incidence supérieurs à 35-40 degrés sont écartés pour garantir une bonne vue du dessus des structures, et les angles d'incidence autour de 15-25 degrés sont préférés pour travailler avec une quantité raisonnable d'occlusions. Parmi toutes les paires qui répondent aux critères d'angle de vue, les intervalles les plus courts entre les dates d'acquisition sont privilégiés, afin de minimiser la possibilité de changements majeurs dans la géométrie et l'apparence de la scène. Comme indiqué dans [GF18], il y a toutefois des exceptions à cette règle : des dates similaires d'années différentes peuvent donner de meilleurs résultats que des dates éloignées de la même année ; et les dates appartenant ou non à la même saison peuvent être un élément important pour décider entre des intervalles similaires. Les mesures de similarité entre images ou la correspondance entre points caractéristiques peuvent également être utiles pour compléter les critères qui se basent sur les métadonnées des images [MdFMLF19]. Un autre critère qui prend plus de temps consiste à calculer la géométrie de toutes les paires possibles et à les classer en fonction d'une certaine métrique de qualité, telle que le moins de valeurs indéfinies ou d'erreurs par rapport aux données connues [FdFML17, Qin17]. Les avantages de la sélection de paires stéréo sont illustrés dans la Figure 2.3.

Rectification stéréo. Dans la photographie conventionnelle, la rectification ou alignement stéréo utilise la géométrie épipolaire pour transformer les lignes contenant des points correspondants (les lignes épipolaires) en lignes horizontales ayant la même coordonnée verticale y [HZ04]. Cette étape facilite la recherche de correspondances de points entre les images, car les algorithmes de mise en correspondance ultérieurs n'ont besoin de considérer que la direction de l'axe x. Cependant, la géométrie des capteurs pushbroom des satellites est différente du modèle de caméra sténopé utilisé pour l'imagerie conventionnelle, ce qui fait que les lignes épipolaires deviennent des courbes. En plus, ces courbes ne sont pas conjuguées : une courbe dans une image satellite n'a pas forcément une courbe équivalente dans l'autre. Les algorithmes de rectification traditionnels ne peuvent donc

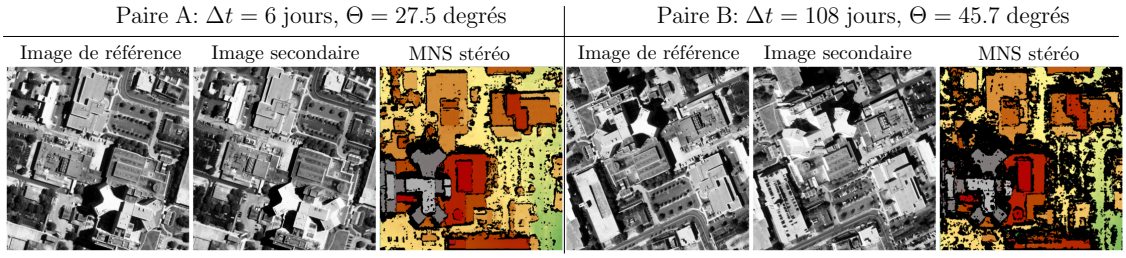


Figure 2.3: Effet de la sélection de paires sur la reconstruction 3D stéréo. Le modèle numérique de surface (MNS) obtenu avec la paire A est plus précis et plus complet que celui obtenu avec la paire B, car l'intervalle entre les dates d'acquisition (Δt) est plus petit et l'angle entre les vues (Θ) est plus proche de l'idéal (15-25 degrés). Les couleurs du MNS représentent différentes valeurs d'altitude, et les zones incomplètes sont indiquées en noir.

pas être appliqués aux modèles RPC, ce qui a conduit à un manque de consensus dans la littérature sur la façon d'aborder cette étape. Quand des valeurs de référence géométriques sont disponibles, les images peuvent être orthorectifiées dans un système de coordonnées commun [SAM+16, BAM18]. Autrement, une variété de techniques a été proposée. La majorité des méthodes reposent sur l'hypothèse que les capteurs pushbroom peuvent être modélisés localement comme des caméras affines, qui sont des modèles de projection parallèle avec le centre de la caméra à l'infini [OAH93, FY04]. Ceci permet une rectification à partir de la matrice fondamentale affine, qui convertit les courbes épipolaires en lignes horizontales presque conjuguées, non alignées seulement par un petit déplacement vertical [dFMLM+14b, BAM18]. Les images de la Figure 2.3 ont été rectifiées sur la base de cette dernière idée. Une autre solution consiste à utiliser des méthodes génériques de rééchantillonnage épipolaire pour les caméras pushbroom. Ces méthodes ne reposent pas nécessairement sur l'hypothèse du modèle affine : elles exploitent plutôt l'idée que des paires de courbes approximativement conjuguées peuvent être construites localement avec une faible marge d'erreur [OLT+10, WHL10, GF17]. Les paires de courbes peuvent être trouvées en localisant les points d'une image à des altitudes raisonnables (par exemple, une élévation approximative du niveau du sol) et en les reprojetant sur l'autre image. Des transformations de rééchantillonnage sont ensuite appliquées pour forcer les courbes correspondantes à être des lignes presque horizontales.

Mise en correspondance stéréo. Après rectification, la carte de disparité de chaque paire est estimée en établissant les correspondances de points entre les images. Parmi les algorithmes classiques, *Semi-Global Matching* (SGM), introduit dans [Hir07], est le choix standard pour l'imagerie satellitaire en raison de son efficacité de calcul et de sa stratégie avantageuse pour exploiter la régularité spatiale [dFMLM+14a, RDPD17, WHP+12, BAM18, MSY+20]. L'algorithme SGM produit des cartes de disparité en minimisant une fonction d'énergie comprenant un coût de correspondance et un terme de régularisation qui agrège les coûts des points voisins sur différentes directions cardinales. Un bon nombre de chaînes utilisent SGM avec la transformation census [ZW94] comme coût de correspondance pour des raisons d'efficacité et de robustesse aux changements d'illumination [MSY+20, WHP+12, dR21]. D'autres utilisent des variantes de l'algorithme SGM. Par exemple, MGM ou *More Global Matching* [FdFM15, FdFML17] améliore SGM en injectant dans les coûts agrégés sur chaque direction de l'information provenant de la direction perpendiculaire ; tSGM ou *tube-based SGM* [RWFH12, GF16, LLJ+19] met en œuvre une méthode hiérarchique pour délimiter la recherche de disparités ; SGBM ou

Semi-Global Block Matching met en correspondance des fenêtres de taille fixe au lieu de pixels individuels [BK00, LRF+19] ; 3SGM utilise des empreintes de bâtiments générées automatiquement pour écarter les informations des pixels précédents dans une direction lorsqu'un bord de bâtiment est rencontré [DDSS22].

Une nouvelle direction de recherche explore également l'utilisation de réseaux de neurones profonds pour effectuer la mise en correspondance stéréo avec des images satellite. Les réseaux de mise en correspondance stéréo comprennent généralement trois modules : l'extraction de caractéristiques, dont le résultat est utilisé pour construire une matrice de coûts 3D ou *cost volume* (en anglais) ; la régularisation du *cost volume*, qui agrège les coûts sur différents niveaux de disparité ; et la régression de la disparité, qui produit la carte de disparité basée sur le coût minimum. Différentes architectures ont été testées dans [WVPDR21] sur des images aériennes avec des résultats encourageants. GA-Net [ZPYT19], qui utilise des couches qui sont une approximation différentiable de l'algorithme SGM, a été testée sur des images satellites dans [GRFvGG22]. GA-Net-Pyramid [XdF+22] reprend les idées de tSGM pour restructurer GA-Net avec une architecture pyramidale plus efficace. Contrairement à SGM et les algorithmes classiques, les réseaux de neurones ont la capacité d'exploiter les indices sémantiques pour déduire les disparités dans les régions difficiles, comme les zones faiblement texturées. Cependant, les réseaux testés jusqu'à présent nécessitent de longs temps d'apprentissage et de grands ensembles de données géométriques de vérité terrain, car ils suivent un apprentissage supervisé. Fait intéressant, [AQ22] montre que SGM peut être utilisé pour raffiner l'apprentissage de ces réseaux afin d'obtenir une meilleure généralisation sur des images satellites non vues lors de l'entraînement. Les approches multi-vues par réseaux de neurones, principalement fondées sur MVSNet [YLL+18], sont également étudiées pour l'imagerie satellitaire (par exemple CasMVSNet [GFZ+20] dans [GRFvGG22] ou SatMVS dans [GLJ21]), mais elles sont moins fréquentes que les méthodes stéréo dans la littérature.

Triangulation. Ce terme désigne le processus de détermination de la localisation d'un point dans l'espace 3D à partir de ses observations 2D sur deux ou plusieurs images. Les correspondances données par les cartes de disparité sont triangulées en utilisant les modèles de caméra de chaque paire, ce qui donne une collection de nuages de points 3D. Cette étape est plus difficile à résoudre avec les modèles RPC par rapport au modèle sténopé conventionnel, qui permet d'utiliser des méthodes de triangulation linéaire directe qui minimisent l'erreur algébrique [HZ04]. Le chaîne stéréo Ames [BAM18] estime un centre de projection grossier pour chaque image satellite et trace les rayons émanant du centre de projection et intersectant les observations 2D. En pratique, une telle approximation conduit à des paires de rayons qui ne se croisent pas dans l'espace 3D. Le point 3D le plus proche de chaque paire de rayons est donc considéré comme la solution, et la distance la plus courte entre les rayons est prise comme mesure de confiance. De manière similaire, le modèle sténopé est approximé localement dans [ZSS19], où il est suggéré que la triangulation de l'imagerie satellitaire avec des outils de géométrie 3D conventionnels comme COLMAP [SF16,SZFP16] est possible, au détriment d'une petite perte de précision en altitude. L'approximation locale du modèle sténopé est encore moins fiable que celle des caméras affines, et se limite à une faible variation de la profondeur de la scène. La chaîne stéréo S2P [dFMLM+14a] atteint une plus grande précision de triangulation que les méthodes précédentes en utilisant directement les modèles RPC. Plus précisément, elle utilise une stratégie de reprojection qui prend une observation 2D de référence et approxime de manière itérative la courbe épipolaire associée. Le point 3D qui dessine la courbe épipolaire la plus proche de l'observation 2D associée sur l'autre image de la paire

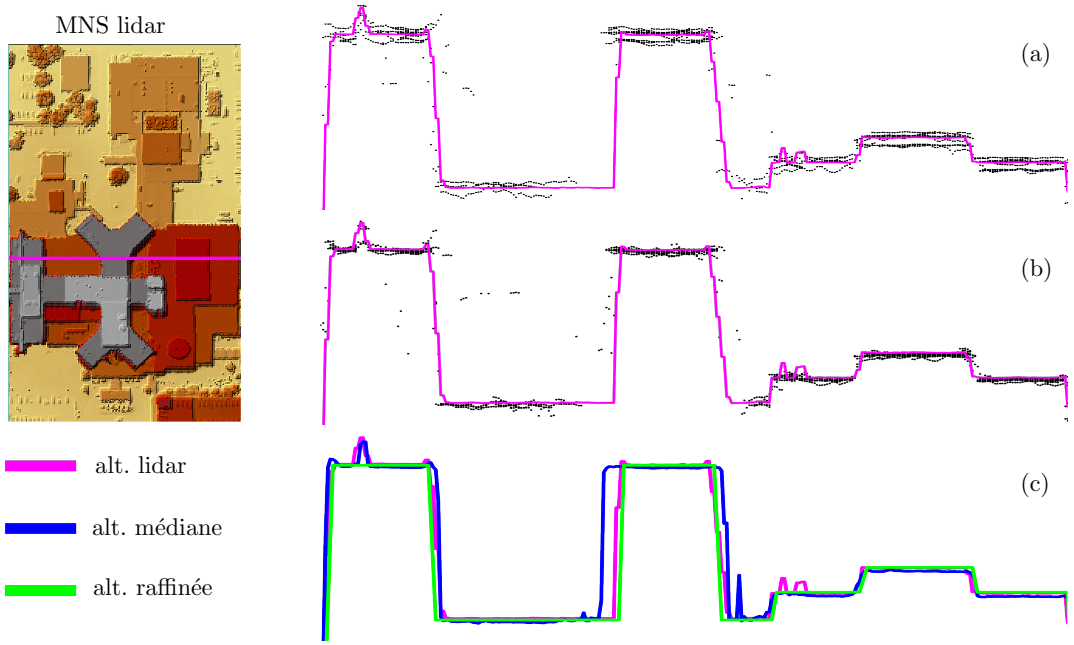


Figure 2.4: Processus de fusion de MNS stéréo. Les points épars représentent les valeurs d'altitude dérivées de différentes paires stéréo correspondants à la ligne magenta indiquée sur le MNS lidar. Les valeurs d'altitude de différentes paires sont généralement incohérentes en raison de petites imprécisions dans les modèles de caméra des images satellite, comme dans (a). Les méthodes d'ajustement de faisceaux et de recalage géométrique peuvent être utilisées pour les aligner, comme dans (b). Une fois que la géométrie de toutes les paires est cohérente, les valeurs d'altitude peuvent être fusionnées en un seul modèle multi-vues avec un filtre local (par exemple, la médiane) et des étapes de raffinement ultérieures (par exemple, de polygonalisation), comme dans (c).

est considéré comme la solution. D'autres méthodes utilisant des modèles RPC pour la triangulation peuvent être trouvées dans [TH02, ODM+14, ZWDF15].

Fusion de MNS stéréo. Les nuages de points 3D obtenus à partir de différentes paires stéréo sont souvent incomplets et contiennent des points aberrants. Il est également possible que les paires individuelles ne couvrent pas la totalité de la zone d'intérêt. Pour obtenir un modèle plus précis, dense et complet de la zone observée, les modèles stéréo sont fusionnés en un seul modèle multi-vues final, comme le montre la Figure 2.4. En supposant que les différents modèles stéréo sont alignés dans l'espace 3D, il est habituel de les fusionner en projetant les nuages de points 3D dans un modèle numérique de surface (MNS) et en appliquant un filtre local. Le MNS est un format 2.5D dans lequel le plan contenant la surface est discrétisé en une grille 2D, où chaque point 2D est assigné une valeur d'altitude. Les nuages de points 3D peuvent être fusionnés en un seul MNS en fusionnant les valeurs d'altitude qui se trouvent dans les mêmes points du MNS. Pour cela, l'altitude médiane est largement utilisée en raison de sa simplicité et de sa robustesse aux valeurs aberrantes [Kus13, GF18, dR21]. Cependant, une telle procédure n'est pas optimale dans certaines zones, telles que celles correspondant aux bords des bâtiments ou de la végétation. Comme observé dans [FdFML17], ces zones présentent généralement une distribution avec deux modes : un à basse altitude (si on observe le sol à côté du bâtiment ou s'il n'y a pas de feuilles sur l'arbre) et un autre à haute altitude (si on observe le toit du bâtiment ou s'il y a des feuilles sur l'arbre). Comme solution, des méthodes de partition-

nement de données ou *clustering* (en anglais) sont proposées dans [FdFML17], qui donne la priorité aux centroïdes de plus basse altitude ; ou [RPDD18], qui sélectionne l'altitude de sortie en minimisant une fonction d'énergie similaire à SGM, dans laquelle les centroïdes d'altitude sont traités comme des disparités candidates. Un filtre de fusion de profondeur adaptatif, qui incorpore l'information spatiale et la couleur des orthoimages est introduit dans [Qin17]. [QLFR22] incorpore des mesures de confiance issues des algorithmes de mise en correspondance stéréo (par exemple, les coûts associés à SGM) : les valeurs d'altitude sont classées par ordre de confiance décroissante et divisées en deux groupes, de telle sorte que la médiane du groupe ayant la confiance la plus élevée est utilisée au lieu de la médiane habituelle si les deux valeurs sont significativement différentes. [KdG+17] étudie différentes variantes de la fusion par altitude médiane, y compris des termes de régularisation spatiale inspirés de la méthode TV-L_1, qui conduisent à des améliorations visuelles dans les modèles de sortie. Un autre travail dans [dK12] montre les avantages de la mise en correspondance stéréo avec régularisation et une étape de fusion ultérieure par rapport à une fusion anticipée de tous les coûts de mise en correspondance suivie d'une étape de régularisation.

Les chaînes MVS pour les images satellites sont généralement complétées par une série d'étapes de prétraitement et de post-traitement qui ont un impact considérable sur la précision des modèles 3D de sortie. Ces étapes complémentaires mais également importantes sont représentées dans les cases à lignes pointillées de la Figure 2.2.

Ajustement de faisceaux. Les approches d'ajustement de faisceaux ou *bundle adjustment* (en anglais) sont une étape de prétraitement habituelle pour corriger les incohérences de géolocalisation entre les modèles RPC de différentes images satellites de la même scène [Kus13, LLJ+19, dR21, AA20]. Ces incohérences de géolocalisation sont principalement dues au manque de précision dans la connaissance de l'orientation du capteur. En pratique, elles empêchent une triangulation correcte avec les caméras RPC, car les observations multi-images du même point sont associées à des localisations différentes dans l'espace 3D (comme dans la Figure 2.4(a)). Par conséquent, la condition (2.1) ne peut être satisfaite. Les méthodes d'ajustement de faisceaux optimisent les coefficients polynomiaux des fonctions RPC en minimisant l'erreur de reprojection d'un ensemble de points de référence observés sur les images [TMHF00]. Les stratégies de correction absolue utilisent des points de contrôle au sol (GCP), dont les coordonnées de latitude, longitude et altitude sont connues, tandis que les stratégies de correction relative remplacent habituellement les GCP par des correspondances de points caractéristiques des images. Les corrections relatives permettent que la géométrie observée par les vues d'entrée soit au moins cohérente en elle-même (comme dans la Figure 2.4(b)). Les méthodes de correction des modèles RPC peuvent être directes, si les coefficients RPC sont explicitement modifiés [XZ09] ; ou indirectes, si les fonctions RPC sont composées avec d'autres fonctions de correction complémentaires [GD03, FH05]. Il convient de noter que les approches d'ajustement de faisceaux ne sont pas les seules solutions existantes pour surmonter les problèmes d'imprécision des modèles RPC dans la génération de modèles numériques de surface. Mais en même temps, la croissance de l'imagerie satellitaire optique et les nouvelles acquisitions multi-images fragmentées, telles que celles des petits satellites, sont en train d'augmenter l'importance de ces méthodes dans les chaînes MVS, en raison de sa capacité à traiter plusieurs vues à la fois. Par exemple, [BSAH21] a modifié la chaîne stéréo Ames pour traiter les données SkySat. MicMac comprend également un module d'ajustement de faisceaux [RDPD17].

Alignement de la géométrie. Certains travaux dans la littérature ont montré qu'il est possible d'effectuer une reconstruction 3D multi-vues même si les modèles de caméra RPC sont affectés par de petites incohérences. Dans ces cas, l'effet des imprécisions dans les fonctions RPC est corrigé localement pour chaque paire stéréo, par exemple en vérifiant soigneusement que seuls les déplacements horizontaux persistent après la rectification stéréo. L'ajustement de faisceaux est alors remplacé par une autre étape dédiée à l'alignement des modèles de surface obtenus à partir des différentes paires stéréo [FdFML17, Qin17]. Cet alignement ou recalage a lieu avant l'étape de fusion et est basé uniquement sur la géométrie, de sorte que les modèles de caméra des images peuvent être ignorés. Dans [FdFML17] et [Qin17] les MNS stéréo sont alignés en appliquant des déplacements dans l'espace 3D qui sont calculés en optimisant une fonction de coût (par exemple, la corrélation croisée normalisée ou l'erreur quadratique) par rapport à un MNS de référence. Le choix d'un déplacement dans l'espace 3D pour le recalage géométrique est justifié par l'approximation locale des modèles RPC comme des caméras affines. Le chaîne stéréo Ames [BAM18] comprend également des outils d'alignement de modèles de surface basés sur l'algorithme *Iterative Closest Point* [PCSM13], qui supporte aussi les rotations entre différents modèles. Les sources d'erreur les plus communes dans le recalage des MNS sont analysées dans [NK11] et identifiées comme des décalages, des biais liés à l'altitude et des biais d'ordre supérieur spécifiques au capteur.

Raffinement de la géométrie. Les occlusions, les objets temporaires (par exemple, les voitures, la végétation, les structures démolies) ou les correspondances stéréo erronées peuvent conduire à des modèles 3D avec des régions incomplètes, des irrégularités locales et des valeurs d'altitude aberrantes. La régularisation dans la mise en correspondance stéréo peut également conduire à des contours d'objets flous. Ce sont des caractéristiques indésirables dans les modèles numériques de surface, car le monde dans lequel nous vivons est généralement constitué de structures régulières, surtout dans les zones habitées. Il est donc fréquent que les modèles de surface multi-vues soient soumis à des étapes de post-traitement. Des routines de morphologie, interpolation ou *inpainting* peuvent être utilisées pour remplir des MNS incomplets [FdFM15, KGAL20]. Les techniques classiques de filtrage local peuvent être utilisées pour supprimer le bruit et les valeurs aberrantes en agrégeant les informations spatiales, mais elles affectent généralement de manière négative la netteté des détails et les bords des bâtiments, avec plus ou moins d'impact selon la taille de la fenêtre spatiale [WW06]. Une façon de préserver et d'améliorer les contours des objets est d'incorporer des labels de segmentation. Les premières approches de segmentation de MNS utilisaient des algorithmes de détection de segments ou de contours et des algorithmes de *clustering*. Les contours des bâtiments étaient généralement identifiés en ajustant des polygones 2D [SdKR10] ou en utilisant des méthodes de classification heuristiques basées sur l'altitude et les normales de surface [PY09]. Aujourd'hui, les réseaux de neurones profonds peuvent être utilisés pour résoudre ce problème. Un avantage différentiel de ces méthodes est qu'elles ne sont pas contraintes à une liste limitée de catégories de bâtiments ou de formes géométriques. Les travaux de [BdKR18, BRK19, BKFR19, LBK20] proposent différentes variantes d'un réseau cGAN (en anglais, *conditional adversarial generative network*) qui améliore les formes des bâtiments dans les MNS photogrammétriques en encourageant la génération de MNS équivalents difficiles à discerner de MNS obtenus à partir de données lidar. Les formes raffinées des bâtiments peuvent être vectorisées en polygones 3D [WZB21]. L'architecture concurrente ResDepth [SS20] aborde le problème de manière différente : au lieu d'utiliser l'apprentissage antagoniste ou *adversarial* et de prédire les altitudes absolues, le modèle est entraîné pour apprendre les différences d'altitude (le résidu) qui rapprochent le MNS d'entrée d'un équivalent provenant de don-

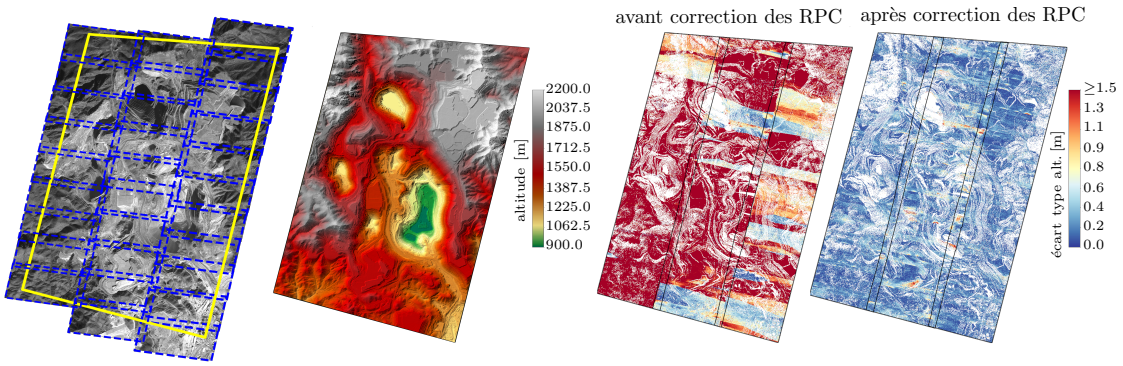
nées lidar. Le raffinement de la profondeur avec les réseaux de neurones est remarquablement amélioré par l'utilisation d'orthoimages alignées avec le MNS d'entrée, ce qui favorise l'identification des contours. Les principaux défis restent le manque de grandes bases de données d'entraînement et la capacité de généraliser aux différents paysages urbains.

Le diagramme MVS de la Figure 2.2 pourrait comprendre d'autres étapes, telles que des changements dans la représentation 3D (par exemple, la conversion de points 3D en maillage ou *mesh* [Kus13, LLJ+19]), la segmentation sémantique ou la texturation [Kus13, MLd19]. Cependant, ces tâches supplémentaires ne sont pas strictement nécessaires dans le processus d'estimation de la géométrie 3D.

2.5 Présentation de la thèse

Cette thèse est divisée en deux parties. La première partie, du Chapitre 3 au Chapitre 5, étudie la modélisation géométrique de l'imagerie satellite multi-vues et le rôle crucial d'une modélisation adéquate des capteurs pour obtenir une géométrie multi-vues précise. La caractérisation et la correction des modèles de caméra RPC sont le sujet principal de ces chapitres, qui traitent en détail les étapes d'ajustement de faisceaux et de recalage géométrique de la stéréo multi-vues à partir d'images satellite (Figure 2.2). La deuxième partie, du Chapitre 6 au Chapitre 9, aborde différentes applications et problèmes liés à la génération de modèles numériques de surface à partir de plusieurs images satellites. Un bon nombre des problèmes traités dans la deuxième partie utilisent les concepts de modélisation géométrique de la première partie pour aborder d'autres questions de la chaîne stéréo multi-vues (MVS), telles que le nombre de paires stéréo nécessaires pour reconstruire une zone d'intérêt ou les algorithmes de mise en correspondance stéréo.

Chapitre 3 : Méthodes de correction des erreurs de géolocalisation pour la stéréo multi-vues à partir d'images satellites
Ce chapitre vise à déterminer quelle stratégie, parmi l'ajustement de faisceaux ou l'alignement direct de la géométrie (Figure 2.2), permet d'obtenir une plus grande précision dans les chaînes MVS pour l'imagerie satellitaire. Nous comparons trois méthodes différentes pour surmonter les incohérences de géolocalisation introduites par les modèles de caméra RPC en entrée. En dehors de cette stratégie de correction, les autres étapes de la chaîne MVS sont les mêmes. La première méthode évite toute intervention sur les modèles de caméra en alignant explicitement les valeurs d'altitude dérivées de plusieurs paires stéréo. Les deux autres méthodes utilisent des ajustements de faisceaux différents pour corriger les modèles RPC. Les deux algorithmes d'ajustement de faisceaux reposent sur l'approximation locale des caméras satellites comme des modèles de projection affine, mais chacun modifie les modèles RPC avec une fonction de correction différente. La première méthode d'ajustement de faisceaux utilise une correction de décalage ou de biais, dans laquelle une translation définie dans le plan de l'image est appliquée après la fonction de projection RPC. L'autre utilise une rotation dans l'espace 3D appliquée avant la fonction de projection. Les trois méthodes sont testées sur différentes zones d'intérêt, en utilisant 50 paires d'images panchromatiques WorldView-3 et un MNS lidar de référence. Les résultats montrent que la méthode d'alignement de la géométrie est la plus robuste lorsque les points utilisés par l'ajustement de faisceaux ne forment pas une composante unique dans le graphe de connectivité, impliquant des sous-ensembles d'images déconnectées. Dans les autres cas, les stratégies d'ajustement de faisceaux obtiennent les meilleurs résultats, ce qui s'explique par le fait que les méthodes de recalage géométrique direct sont sensibles



(a) Acquisition SkySat (b) Modèle numérique de surface MVS, obtenu à partir de 144 paires stéréo (c) Écart type entre les valeurs d'altitude de différentes paires stéréo

Figure 2.5: La mine Morenci (États-Unis) est reconstruite à partir de plusieurs paires stéréo SkySat dans le Chapitre 4. L'ajustement de faisceaux des modèles de caméra RPC avant la reconstruction stéréo permet d'obtenir une géométrie multi-vues cohérente.

aux parties incomplètes ou aux valeurs aberrantes dans les modèles d'altitude obtenus à partir de paires stéréo individuelles.

Chapitre 4 : Ajustement de faisceaux de modèles RPC
Ce chapitre approfondit le problème de la correction des inexactitudes des modèles de caméra RPC dans les chaînes MVS pour l'imagerie satellitaire. Nous proposons une méthodologie améliorée d'ajustement de faisceaux qui est basée sur l'une des stratégies d'ajustement de faisceaux précédemment présentées dans le Chapitre 3. Dans ce cas, la cohérence géométrique est obtenue en corrigeant les erreurs liées à l'orientation des capteurs par la composition directe des fonctions RPC originales avec une matrice de rotation dans l'espace 3D. En utilisant directement les fonctions RPC originales, l'approximation locale des modèles de projection affine est évitée. Les autres améliorations significatives comprennent un critère de sélection permettant d'utiliser un sous-ensemble optimal de points dans l'ajutement de faiseaux, ce qui permet de traiter efficacement de grandes données multi-vues; ainsi qu'un algorithme d'estimation de modèles RPC qui permet d'obtenir les modèles de caméra de sortie dans le même format que ceux d'entrée. Le critère de sélection de paires stéréo est également adapté pour travailler avec des images SkySat, qui peuvent former des paires avec un chevauchement et une distance entre les prises de vue très limités, car les différentes images de la même acquisition sont capturées presque simultanément. La méthode résultante est testée avec des acquisitions SkySat sur deux zones d'intérêt différentes, et les modèles RPC de sortie sont ensuite utilisés dans une chaîne stéréo pour extraire des modèles numériques de surface. L'écart type entre les valeurs d'altitude obtenues à partir de différentes paires est utilisé comme métrique d'évaluation, comme illustré dans la Figure 2.5, car nous ne disposons pas de modèles de vérité terrain des zones observées. Les résultats indiquent que l'ajustement de faisceaux proposé conduit à un meilleur alignement géométrique par rapport aux algorithmes de recalage géométrique direct. L'erreur de reprojection des points utilisés par l'ajustement de faisceaux atteint des valeurs sub-pixéliques et l'écart type entre les valeurs d'altitude est réduit de l'ordre de quelques mètres à quelques dizaines de centimètres.

Chapitre 5 : Estimation robuste de modèles RPC
Ce chapitre décrit en détail l'algorithme d'estimation de modèles RPC utilisé dans la méthodologie d'ajustement de faisceaux présentée dans le Chapitre 4. L'algorithme utilise

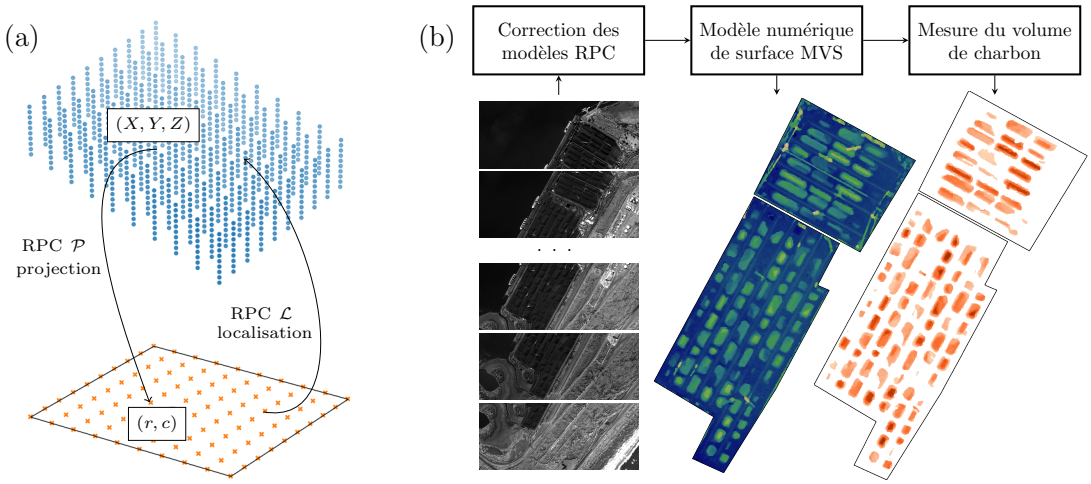


Figure 2.6: (a) L'algorithme d'estimation de modèles RPC décrit dans le Chapitre 5 estime les fonctions de géolocalisation à partir d'une grille de correspondances virtuelles entre des points 3D (X, Y, Z) et des coordonnées image 2D (r, c). (b) Le système de suivi des surfaces du Chapitre 6 utilise le concept illustré en (a) pour corriger les fonctions RPC des acquisitions SkySat capturées à différentes dates sur un terminal charbonnier. Les modèles de caméra RPC corrigés sont utilisés pour construire un modèle 3D précis pour chaque date d'acquisition et suivre l'évolution du volume de charbon.

une méthode indépendante du terrain de type moindres carrés régularisés pour trouver les coefficients des fonctions RPC qui relient un ensemble de correspondances virtuelles entre l'espace image et l'espace 3D. Les correspondances virtuelles d'entrée doivent suivre une structure de grille régulière, comme illustré dans la Figure 2.6(a). La grille 3D doit couvrir la totalité de l'altitude de la scène, et doit se projeter sur des observations 2D qui couvrent également la totalité de l'image. Ces correspondances virtuelles sont générées à l'aide d'une représentation différente du capteur, telle qu'un modèle physique ou un modèle RPC précédent corrigé par une autre fonction. Nous testons l'algorithme en utilisant des images satellites SAR et optiques, provenant respectivement de Sentinel-1 et WorldView-3. La robustesse de la méthode est évaluée en faisant varier la densité des correspondances des points en entrée et la taille de la zone géographique couverte. Les modèles RPC obtenus permettent d'atteindre des erreurs de reprojection de l'ordre de 10^{-4} pixels dans la plupart des configurations, ce qui montre la capacité de l'algorithme à capturer des fonctions de géolocalisation complexes sous la forme de modèles RPC.

Chapitre 6 : Système de suivi automatique des volumes de stockage
Ce chapitre présente un cas réel d'utilisation d'images satellite dans une chaîne MVS. Des acquisitions push-frame récurrentes d'images SkySat sont utilisées pour suivre le volume des stocks dans un terminal charbonnier, comme illustré dans la Figure 2.6(b). Les images SkySat sont d'abord regroupées en fonction de leurs dates d'acquisition, et les modèles de caméra RPC de chaque groupe sont corrigés par la méthode d'ajustement de faisceaux du Chapitre 4. La chaîne MVS est ensuite appliquée en utilisant les modèles RPC corrigés, pour obtenir le modèle numérique de surface de chaque date d'acquisition. La géométrie des paires stéréo utilisées à chaque date est alignée avec une précision inférieure au mètre grâce à l'ajustement de faisceaux précédent. La séquence de modèles 3D multi-vues résultante est utilisée pour mesurer l'évolution du volume des stocks. Plus précisément, le volume des stocks dans chaque modèle 3D est obtenu par soustraction d'un modèle de ter-

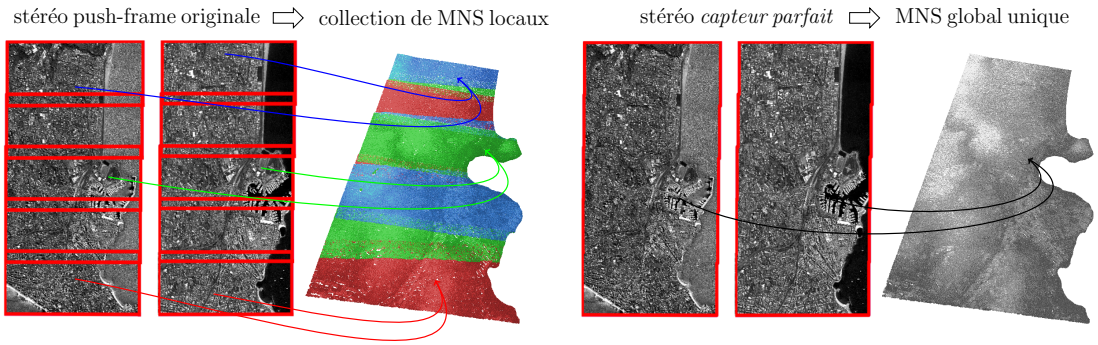


Figure 2.7: La méthode d'assemblage d'images du Chapitre 7 combine des séquences d'images qui se chevauchent partiellement (comme celles acquises par un système push-frame) en une seule mosaïque à grande échelle. La géométrie de la mosaïque est décrite avec un modèle de *capteur parfait*, comme si elle avait été acquise par un capteur pushb-room idéal. En construisant des mosaïques à grande échelle à partir de collections d'images plus petites, ainsi que leur modèle de capteur parfait, la reconstruction stéréo multi-vues de modèles numériques de surface (MNS) peut passer d'un problème complexe avec de nombreuses paires d'entrée à un problème avec une seule paire stéréo.

rain et par intégration des valeurs d'altitude au-dessus du sol. Les estimations de volume sont validées en utilisant des mesures effectuées sur le terrain pendant la même période. Les deux distributions présentent une forte corrélation, soulignant ainsi l'énorme potentiel des acquisitions multi-images récurrentes avec une géométrie multi-vues cohérente.

Chapitre 7 : Modélisation de capteur parfait pour l'assemblage d'images push-frame

Ce chapitre vise à réduire le nombre de paires stéréo nécessaires pour reconstruire une zone observée dans des séquences d'images push-frame. Les images satellite acquises par les systèmes push-frame se chevauchent partiellement dans la direction de déplacement du satellite. Nous proposons une méthode d'assemblage d'images géométriquement cohérente qui fusionne les images de petite taille provenant de la même séquence push-frame en une seule mosaïque à grande échelle, que nous appelons L1B$^+$. Un exemple d'images d'entrée et de sortie de la méthodologie est montré dans la Figure 2.7. La mosaïque de sortie est obtenue en projetant chaque image d'entrée dans un espace image commun à l'aide d'une transformation projective. La méthode produit également un modèle de *capteur parfait*, sous la forme d'un modèle RPC, qui décrit la géométrie de l'image L1B$^+$, comme si la mosaïque avait été acquise par un capteur pushbroom idéal. L'algorithme d'estimation de modèles RPC décrit dans le Chapitre 5 est utilisé pour déterminer le modèle de capteur parfait, sur la base d'un ensemble de correspondances entre l'espace image et l'espace 3D établi à partir des modèles RPC de la séquence d'images d'entrée après un ajustement de faisceaux. Nous avons testé la méthode sur différentes paires de séquences push-frame SkySat de 3 et 5 images, et les images L1B$^+$ obtenues sont ensuite utilisées pour reconstruire les zones observées avec une seule paire stéréo. Les résultats indiquent que les modèles L1B$^+$ atteignent une précision d'altitude supérieure à celle des modèles 3D obtenus à partir de plusieurs paires stéréo des images originales. Cela s'explique principalement par le fait que les modèles L1B$^+$ ne sont pas affectés par les erreurs de recalage entre les valeurs d'altitude obtenues à partir de différentes paires stéréo.

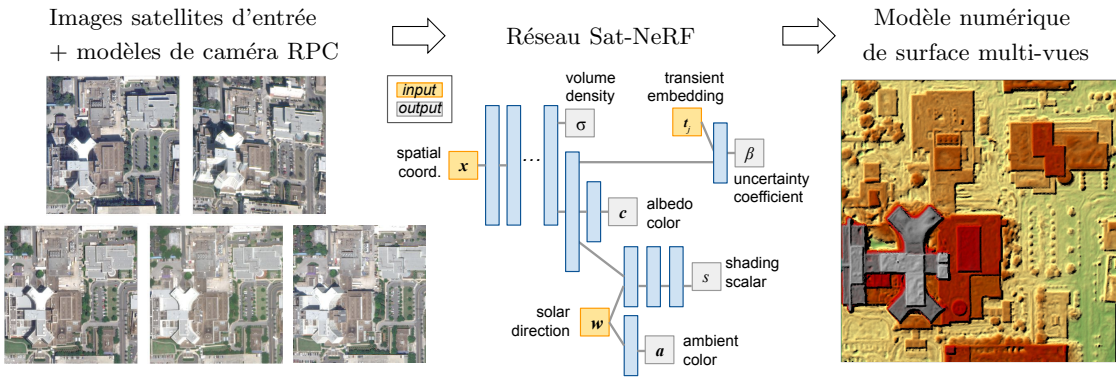


Figure 2.8: Le réseau de neurones Sat-NeRF décrit dans le Chapitre 8 est une adaptation des approches de rendu neuronal à l'imagerie satellitaire multi-dates. Contrairement aux chaînes de production classiques basées sur la stéréo, le réseau Sat-NeRF permet une reconstruction 3D multi-vues réelle. En utilisant N vues d'entrée, le réseau apprend la géométrie et l'apparence à chaque point 3D de la scène, de manière robuste aux objets transitoires et aux variations d'illumination.

Chapter 8 : Satellite NeRF

Ce chapitre vise à remplacer toutes les étapes de la stéréo multi-vues présentées dans la Figure 2.2, avec la seule exception de l'ajustement de faisceaux initial pour corriger les modèles de caméra RPC. À cette fin, nous présentons le modèle *Satellite NeRF*, ou Sat-NeRF, pour lequel nous nous sommes inspirés des méthodes de pointe de rendu neuronal. Le réseau Sat-NeRF est basé sur le modèle NeRF original pour l'apprentissage des champs de rayonnement neuronal : une représentation continue de la géométrie et de l'apparence d'une scène 3D, qui est apprise sans supervision directe de la géométrie. Sat-NeRF est conçu pour travailler avec des images satellites multi-dates, il utilise des modèles de caméra RPC au lieu du modèle sténopé classique et contourne certaines contraintes de NeRF, comme la limitation aux scènes statiques avec illumination fixe. Comme indiqué dans la Figure 2.8, Sat-NeRF apprend simultanément la géométrie et l'apparence des structures permanentes en utilisant une structure principale multi-couches, qui ne prend que les coordonnées spatiales en entrée. Les ombres et les objets transitoires comme les voitures sont appris séparément par des couches secondaires qui ont accès à la direction des rayons solaires et à des vecteurs de caractéristiques spécifiques à chaque image. La fonction de coût est également adaptée pour favoriser la dissociation entre les éléments permanents et temporaires. La méthode est ainsi capable de prédire, pour chaque point 3D de la scène : la densité, qui définit la géométrie ; la couleur, qui est une combinaison linéaire d'un albédo avec l'illumination directe des rayons solaires et l'illumination ambiante ; et un coefficient d'incertitude, qui représente la probabilité que le point appartienne ou non à un objet transitoire. L'échantillonnage de Sat-NeRF le long des rayons selon une direction donnée permet de générer la vue de la scène et les altitudes correspondantes. Dans la partie expérimentale, Sat-NeRF est entraîné sur différentes zones en utilisant 10 à 20 images RGB WorldView-3, et atteint une précision d'altitude similaire à celle des chaînes MVS classiques avec sélection manuelle des paires stéréo d'entrée. Les principaux avantages de Sat-NeRF sont la capacité d'obtenir des modèles de surface plus détaillés et la nature véritablement multi-vues de la méthode. Les points faibles de la méthode sont les temps d'apprentissage, la nécessité de disposer de plusieurs images d'entrée et la présence d'irrégularités locales dans la géométrie apprise.

Chapter 9 : Réseaux de neurones pour l'estimation de cartes de disparité
Ce chapitre aborde l'étape de mise en correspondance stéréo des chaînes MVS pour les images satellites. Contrairement à SGM et les algorithmes classiques, les réseaux de neurones ont la capacité d'exploiter des indices sémantiques pour déduire les disparités dans les régions difficiles, comme les zones faiblement texturées. Sur cette base, nous étudions la possibilité d'intégrer des réseaux de neurones pour l'estimation de cartes de disparité dans les chaînes MVS, afin d'améliorer la géométrie de sortie. Deux architectures d'apprentissage supervisé de réseaux de neurones convolutifs, *Pyramid Stereo Matching* (PSM) et *Hierarchical Stereo Matching* (HSM), sont évaluées. Les deux réseaux suivent une structure commune de modules d'extraction de caractéristiques, de régularisation des coûts et de régression de la disparité. Cependant, la manière dont la matrice de coûts ou *cost volume* est construite est différente dans chaque cas. PSM construit une structure unique en concaténant les vecteurs de caractéristiques à différents niveaux de disparité, tandis que HSM construit une pyramide multi-échelle de coûts en utilisant les différences absolues entre les vecteurs de caractéristiques. La représentation explicitement multi-échelle et moins volumineuse utilisée par HSM est destinée à améliorer et à accélérer le traitement des grandes images à haute résolution. Pour une comparaison plus juste avec les algorithmes classiques nous utilisons dans la partie expérimentale des poids pré-entraînés et ajustés en utilisant des images aériennes. De cette manière, les réseaux sont familiarisés avec des scènes d'observation de la Terre, même si la résolution des images n'est pas la même. Quatre séries de 30 paires stéréo RGB et panchromatiques WorldView-3 sont utilisées pour évaluer les architectures PSM et HSM, ainsi que la variante MGM de l'algorithme SGM classique. Une chaîne de reconstruction stéréo est utilisée avec les différentes méthodes de mise en correspondance stéréo, et les modèles 3D obtenus sont évalués avec un modèle 3D lidar de référence. Les résultats montrent que, dans des conditions idéales (comme par exemple, une distance appropriée entre les caméras et des dates d'acquisition similaires), les méthodes d'apprentissage profond offrent une précision numérique supérieure à celle des algorithmes classiques. PSM atteint un niveau de détail plus élevé, mais devient peu pratique lorsque la taille des images d'entrée augmente. Les points faibles des deux réseaux sont leur limitation à un intervalle de disparité prédéfini, leur sensibilité à la rectification stéréo des paires d'entrée et la nécessité d'ajuster la taille des images pour obtenir des cartes de disparité optimales.

2.6 Synthèse des contributions

Cette thèse passe en revue les chaînes stéréo multi-vues pour la génération de modèles numériques de surface à partir d'images satellites. L'objectif est de faciliter l'utilisation de plusieurs vues d'entrée au lieu d'une seule paire stéréo. À cette fin, la première contribution de cette thèse est une méthodologie d'ajustement de faisceaux pour corriger les modèles de caméra RPC d'un ensemble d'images satellites (Chapitre 4), qui est le résultat d'une étude des méthodes de la littérature (Chapitre 3). La méthodologie d'ajustement de faisceaux comprend la mise en place d'un algorithme d'estimation de modèles RPC (Chapitre 5), qui permet d'obtenir les modèles de caméra corrigés dans le même format que ceux d'entrée. L'ajustement de faisceaux des modèles de caméra RPC permet d'obtenir une géométrie multi-vues cohérente avec une précision inférieure au mètre. Cela offre la possibilité de simplifier les chaînes stéréo multi-vues (car il n'est plus nécessaire d'aligner les valeurs d'altitude de différentes paires stéréo) et d'utiliser des méthodes multi-vues réelles. En conséquence, la contribution suivante de cette thèse est une méthode d'assemblage d'images, nommée L1B$^+$, qui fusionne une séquence de petites images satel-

lites qui se chevauchent partiellement en une seule image à grande échelle (Chapitre 7). Nous montrons qu'il est possible de calculer un modèle de caméra unique pour l'image de sortie, comme si elle avait été acquise avec un capteur idéal, à condition que les modèles RPC des images d'entrée soient géométriquement cohérents. Ainsi, la reconstruction stéréo multi-vues peut passer d'un problème complexe nécessitant de nombreuses paires stéréo à un problème simplifié avec une seule paire en entrée. Une autre contribution de cette thèse est le réseau Sat-NeRF (Chapitre 8), une architecture d'apprentissage auto-supervisé qui est une adaptation des méthodes de rendu neuronal les plus avancées. Sat-NeRF est une méthode multi-vues réelle pour la génération de modèles numériques de surface à partir d'images satellites multi-dates, qui ne s'appuie pas sur une sélection de paires stéréo. Sans aucune supervision de la géométrie, le réseau est capable d'inférer la surface observée dans un ensemble d'images satellites en apprenant simultanément la géométrie et l'apparence des structures permanentes avec une structure principale multi-couches, tandis que les ombres et les objets transitoires sont appris séparément avec des couches secondaires. Pour terminer, parmi les contributions secondaires de cette thèse, on peut trouver une application de la stéréo multi-vues à partir d'images satellites récurrentes pour le suivi automatique des volumes de stockage (Chapitre 6), et une évaluation expérimentale de différents réseaux de neurones profonds pour la mise en correspondance stéréo, avec une comparaison par rapport aux des algorithmes classiques similaires à SGM (Chapitre 9).

2.7 Liste des publications

Présentées dans cette thèse

- Roger Marí, Carlo de Franchis, Enric Meinhardt-Llopis, and Gabriele Facciolo. To bundle adjust or not: A comparison of relative geolocation correction strategies for satellite multi-view stereo. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 2188–2196, 2019. https://doi.org/10.1109/ICCVW.2019.00274

- Roger Marí, Carlo de Franchis, Enric Meinhardt-Llopis, Jérémy Anger, and Gabriele Facciolo. A generic bundle adjustment methodology for indirect RPC model refinement of satellite imagery. *Image Processing On Line*, 11:344–373, 2021. https://doi.org/10.5201/ipol.2021.352

- Roland Akiki, Roger Marí, Carlo de Franchis, Jean-Michel Morel, and Gabriele Facciolo. Robust rational polynomial camera modelling for SAR and pushbroom imaging. In *2021 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 7908–7911, 2021. https://doi.org/10.1109/IGARSS47720.2021.9554583

- Roger Marí, Carlo de Franchis, Enric Meinhardt-Llopis, and Gabriele Facciolo. Automatic stockpile volume monitoring using multi-view stereo from SkySat imagery. In *2021 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 4384–4387, 2021. https://doi.org/10.1109/IGARSS47720.2021.9554482

- Roger Marí, Thibaud Ehret, Jérémy Anger, Carlo de Franchis, and Gabriele Facciolo. L1B+: A perfect sensor localization model for simple satellite stereo reconstruction from push-frame image strips. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 5-1-2022:137–143, 2022. https://doi.org/10.5194/isprs-annals-V-1-2022-137-2022

- Roger Marí, Gabriele Facciolo, and Thibaud Ehret. Sat-NeRF: Learning multi-view satellite photogrammetry with transient objects and shadow modeling using RPC cameras. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1310–1320, 2022. https://doi.org/10.1109/CVPRW56347.2022.00137
- Roger Marí, Thibaud Ehret, and Gabriele Facciolo. Disparity estimation networks for aerial and high-resolution satellite images: A review. *Image Processing On Line*, 12:501–526, 2022. https://doi.org/10.5201/ipol.2022.435

Autres

- Manuel Lopez, Roger Marí, Pau Gargallo, Yubin Kuang, Javier Gonzalez-Jimenez, and Gloria Haro. Deep single image camera calibration with radial distortion. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11809–11817, 2019. https://doi.org/10.1109/CVPR.2019.01209
- Thibaud Ehret, Roger Marí, and Gabriele Facciolo. NeRF, meet differential geometry! *arXiv preprint arXiv:2206.14938*, 2022. https://arxiv.org/abs/2206.14938

Part I

Geometric modeling of multi-view satellite imagery

3 Geolocation correction methods for satellite multi-view stereo

In satellite multi-view stereo (MVS) pipelines, the altitude values derived from different pairs are fused to obtain the final mutli-view 3D model. The fusion step usually requires that the altitude values (and therefore the camera models) are registered in the same reference system. This is not what happens in practice, since depending on the satellite geopositioning equipment the camera models may contain errors of up to tens of meters on the ground. This chapter compares three different methods employed in MVS pipelines to overcome the geolocation inconsistencies introduced by the camera models of multiple views. The first method avoids taking action on the camera models by explicitly co-registering the geometry of pairwise models before the fusion step. The other two methods use bundle adjustment approaches to correct the camera models at the beginning of the MVS pipeline. Both bundle adjustment strategies identify a set of tie points, automatically derived from image features, and correct the camera models so that each tie point projects to the corresponding feature observations. The main difference lies on the transformation that is applied to correct each camera model. One bundle adjustment uses an offset or bias correction, in which a translation defined in the image domain is applied after the projection function; and the other uses a rotation defined in the object space, which is applied before the projection function. The three methods are tested on different areas of interest, using input pairs of WorldView-3 images. The results indicate that, with a sufficient number of tie points, bundle adjustment approaches provide superior performance. The latter is explained by the fact that geometry registration methods are subject to incomplete parts or outliers in the altitude models derived from single pairs.

3.1 Introduction

In the past few years, the advances in satellite technology have resulted in a remarkable increase of high-resolution imagery of the Earth surface, with many areas being captured on a daily basis or multiple times per year. Current satellite imagery allows the use of photogrammetry to build accurate 3D digital surface models (DSMs) in a periodic manner, providing extremely valuable up-to-date information about the evolution of the terrain and the human activity on it. Alternative technologies such as airbone lidar and aerial image acquisitions are limited by their narrow swath and their cost, making it difficult to update surface models on a regular basis.

Satellite images are typically provided along with a Rational Polynomial Coefficients (RPC) camera model [FDG06], and other metadata such as the acquisition timestamp

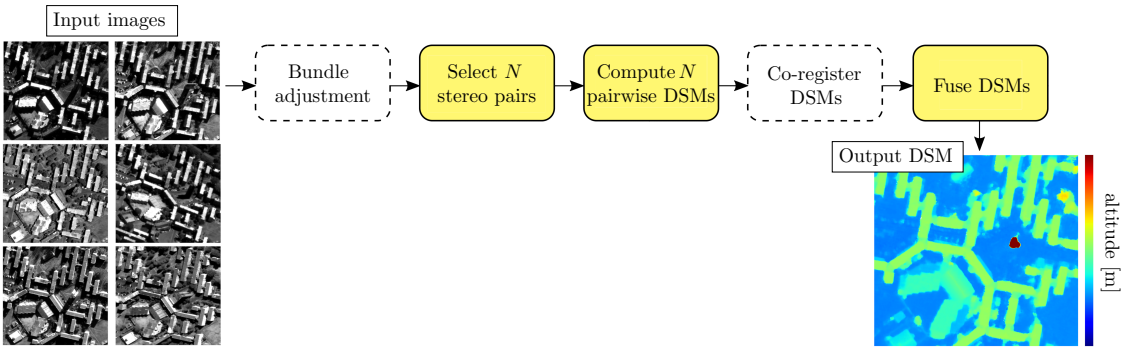


Figure 3.1: Geolocation correction methods are a key step to remove the undesired effects of RPC inaccuracies and thus obtain accurate multi-view digital surface models (DSMs) from satellite imagery. Bundle adjustment or explicit co-registration of pairwise models represent two different approaches to this problem. As illustrated by the dashed boxes, bundle adjustment methods are applied before the stereo reconstruction of multiple pairs, while geometry registration methods are applied after.

or the pixel size. The RPC camera model, previously introduced in Section 1.4.1, is composed by two rational polynomial functions that approximate the mapping from 3D space points to 2D image pixels (i.e. the *projection* function) and its inverse (i.e. the *localization* function). The RPC functions allow to model complex camera systems, such as a push-broom scanner, independently of the physical properties of the system. Given a stereo correspondence found across two or more satellite images, the associated RPCs can be used to triangulate and retrieve the 3D point that projects on the corresponding points.

Although RPCs are expected to be precise enough, the complex system they encode is subject to measurement errors in the satellite geopositioning equipment, mainly related to the attitude angles. Such inaccuracies, also referred to as *pointing errors* [ODM+14], can be of the order of tens of pixels in the image domain. This implies that different satellite views are typically not consistent in a common frame of reference (i.e. each 3D point projects to a slightly different location in the images). Hence it is imperative to use some correction strategy to prevent pointing errors from affecting the triangulation of 3D points and the accuracy of the output DSMs.

Multi-view stereo (MVS) pipelines, previously introduced in Section 1.4.2 and illustrated in Figure 3.1, are the most popular choice for digital surface modeling from multiple satellite images [dFMLM+14a, FdFML17, RPDD18, DL16, dK12]. These pipelines use a variety of strategies to prevent the propagation of RPC inaccuracies into the output 3D models. Such strategies can be classified into bundle adjustment or image-based methods, applied before the stereo reconstruction of multiple image pairs; and geometry-based methods, applied after the pairwise reconstruction process.

The objective of this chapter is to investigate the performance of bundle adjustment versus geometry-based methods to overcome geolocation errors from multiple satellite images. For this purpose, we use three satellite MVS pipelines, each using a different method to correct geolocation inaccuracies. Two of the pipelines reviewed in this chapter use a bundle adjustment procedure, each optimizing a different set of parameters to correct the RPCs: 2D translations in the image domain [ODM+14] or the 3D rotation in object space [TMHF00]. Oppositely, the third approach omits any kind of prior bundle adjustment and aligns independent models based on geometry correlation [FdFML17]. To

perform a fair comparison, the different blocks of the MVS pipelines (yellow boxes in Figure 3.1) are all common apart from the geolocation correction method applied in each case.

It should be noted that in all the methods studied in this chapter, the geolocation correction is relative, i.e., the geometry derived from the multiple views is recorded in a single reference frame, but not in absolute coordinates of longitude, latitude and altitude.

Our evaluation focuses on the reconstruction of small areas of interest, which are of relevance to many applications. For example, monitoring tasks, where it is necessary to reconstruct a particular area in detail from the available images. In addition, modern small satellite constellations, previously introduced in Section 1.3, favor smaller footprints and revisit capacity to large swaths. We employ the *IARPA Multi-View Stereo 3D Mapping Challenge* dataset [BKHB16], which includes 47 DigitalGlobe WorldView-3 images, with a resolution of 30 cm per m at nadir, collected between 2014 and 2016 over Buenos Aires; and the *2019 IEEE GRSS Data Fusion Contest* (DFC2019) dataset [BFC+19, LSYH+19] with 26 DigitalGlobe WorldView-3 images collected between 2014 and 2016 over Jacksonville (among others). Panchromatic images are used in both cases. The completeness (percentage of points where the absolute difference is less than 1 meter) of the output models with respect to a lidar ground truth DSM is used as the main evaluation metric to assess the performance of the different methods.

3.2 Related work

The computer vision community has proposed different methods to correct the pointing error of RPC camera models. Bundle adjustment based solutions are a generally accepted practice that consist in detecting inter-image tie points and applying a compensating function to the original RPCs so that the back-projections of the tie points are coincident in the 3D world [GD03, dR12, ODM+14, GF18, LLJ+19]. We refer to the set of image points from different views that correspond to the same 3D point as a *feature track*.

3.2.1 Fundamentals of bundle adjustment approaches for RPC camera correction

Bundle adjustment is the problem of jointly optimizing the viewing parameters (external and/or internal) of multiple cameras and the 3D locations of the objects they observe [TMHF00]. Given an initial set of K 3D points $\{\boldsymbol{X}_k\}_{k=1,\dots,K}$, and their 2D observations $\boldsymbol{x}_{mk}$ across M cameras with projection functions $\{\mathcal{P}_m\}_{m=1,\dots,M} : \mathbb{R}^3 \to \mathbb{R}^2$, bundle adjustment finds the optimal solution by minimizing the reprojection error of the setting

$$E(\mathcal{P}_m, \boldsymbol{X}_k) = \sum_{k=1}^{K} \sum_{m=1}^{M} \|\boldsymbol{x}_{mk} - \mathcal{P}_m(\boldsymbol{X}_k)\|^2, \tag{3.1}$$

where $\boldsymbol{X}_k$ and $\mathcal{P}_m$ contain the variables to be adjusted. Following (3.1), the reprojection error is typically defined as the sum of squared Euclidean distances between the estimated reprojected 3D points, $\mathcal{P}_m(\boldsymbol{X}_k)$, and their actual 2D observations $\boldsymbol{x}_{mk}$ in the images.

In the case of satellite images, since cameras are far from the Earth's surface (typically above 500 km), the main component of the reprojection error comes from the inaccurate knowledge of the satellite orientation [GD03]. Thus, the energy in (3.1) could ideally be minimized by composing each projection map $\mathcal{P}_m$ with a 3D transformation $\mathcal{R}_m$, whose

objective is to reorient the camera with respect to the object coordinates. This way the problem can be reformulated as

$$E(\mathcal{R}_m, \boldsymbol{X}_k) = \sum_{k=1}^{K} \sum_{m=1}^{M} \|\boldsymbol{x}_{mk} - \mathcal{P}_m(\mathcal{R}_m(\boldsymbol{X}_k))\|^2. \tag{3.2}$$

In [GD03] it was shown that the net effect of RPC inaccuracies related to the attitude angles reduces to a 2D translation (also termed as *correction offset*) in images covering lengths up to 50 km. Based on this observation, *bias compensation* approaches for such scenarios have in common the optimization of a 2D correction offset for each RPC model [dR12, FH05, ODM+14]. Following the notation from (3.1), this amounts to finding the 2D translations $\mathcal{T}_m$ in the image domain that minimize

$$E(\mathcal{T}_m, \boldsymbol{X}_k) = \sum_{k=1}^{K} \sum_{m=1}^{M} \|\boldsymbol{x}_{mk} - \mathcal{T}_m(\mathcal{P}_m(\boldsymbol{X}_k))\|^2. \tag{3.3}$$

Note that (3.2) inserts the correction transformation before the original RPC projection, while (3.3) does it after. More generally, RPCs can also be corrected by composition with polynomial functions depending on the image or object tie point coordinates. This allows modeling other error sources beyond attitude inaccuracies, such as time-dependent drift or lens distortion. Although they are normally orders of magnitude smaller, these errors may not be negligible for certain scenarios (e.g., large images covering hundreds of kilometers) [GD03, TLW10].

3.2.2 Geometry-based registration of dense surface models

For smaller areas (e.g., 2×2 km), RPCs can be modeled as affine cameras using the first order Taylor approximation [SAM+16, FdFML17, dFMLM+14b]. This allows to correct the effect of bias on the triangulated points with a shift in the object space. Based on this idea, [FdFML17] proposed a new approach capable of producing high-quality reconstructions without needing a bundle adjustment. Independent DSMs are registered based on geometry-based correlation [FdFML17] instead of image information, to overcome the problem of geolocation inaccuracies. The main motivation for the latter is that is that finding a sufficient amount of tie points can be an issue with multi-date images, especially when restricted to small areas of interest. Significant differences due to noise or radiometric changes may cause image matches not to be accurate enough, set aside the impact of human activity, weather phenomena or seasonal changes.

3.3 Correlation-based DSM alignment

The first MVS pipeline studied in this chapter uses the geometry-based registration method originally proposed in [FdFML17]. Therefore, it aggregates 3D point clouds independently computed from N different stereo pairs, without any prior RPC correction. The method is summarized in Algorithm 1. Our implementation is based on the open-source satellite stereo pipeline S2P [dFMLM+14a].

The input images are cut into tiles covering small areas, where RPCs can be locally approximated as an affine camera model (see Section 3.2). The reconstruction of each tile starts by rectifying the image crops (I, I') of each stereo pair, to make epipolar lines horizontal. A variant of the SGM algorithm [FdFM15] with a cost based on the census

Algorithm 1: *MVS with NCC based DSM alignment*

Input : M views of a small area of interest (AOI) cropped from multi-date satellite images + associated RPCs

Output: High-quality DSM of the input AOI

- Select N stereo pairs: $\{ (I_n, I'_n) \}_{n=1,\dots N}$

for *each stereo pair* (I_n, I'_n) **do**

- Affine approx. of raw RPCs $\rightarrow P_n, P'_n$
- Epipolar rectification
- Dense stereo matching
- Triangulate using P_n, P'_n to get 3D point cloud
- Project point cloud to DSM_n
- Post-process DSM_n

- DSM alignment via 3D translations maximizing the NCC
- Fuse all DSMs via point-wise median

transform is then used to compute a disparity map robust to lighting changes. Only consistent disparities passing the left-right check are kept. The correspondences given by the disparity map are re-expressed in the original images domain and triangulated using the affine projection matrices P, P' of the pair to compute a dense 3D point cloud.

After running the previous process for the N input stereo pairs, the objective is to merge the output point clouds to obtain a high-quality reconstruction. To this end, each point cloud is projected on a geographic grid, thus producing different DSMs as in [BKHB16, FdFML17]. Morphological filters are then used to refine the DSMs, which may contain small holes due to the sampling step and larger ones due to mismatches and occlusions. Closing with a 3×3 structuring element is applied to fill small holes, followed by interpolation using a low value (the 5th percentile) on the boundaries to reduce larger holes. This interpolation strategy assumes that occluded parts are at ground level.

At this point, the post-processed DSMs are not aligned due to the pointing error in the satellite RPCs, which prevents any kind of fusion. This is where the differential part of the pipeline takes place: the DSMs are aligned via a 3D translation that maximizes the normalized cross correlation (NCC) between them, defined as

$$\text{NCC}(\mathbf{u}, \mathbf{v}) := \frac{1}{|\widehat{\Omega}|} \sum_{t \in \widehat{\Omega}} \frac{(\mathbf{u}_t - \mu_{\mathbf{u}}(\widehat{\Omega}))(\mathbf{v}_t - \mu_{\mathbf{v}}(\widehat{\Omega}))}{\sigma_{\mathbf{u}}(\widehat{\Omega})\sigma_{\mathbf{v}}(\widehat{\Omega})}, \tag{3.4}$$

where $\widehat{\Omega} := \Omega_{\mathbf{u}} \cap \Omega_{\mathbf{v}}$ is the intersection of the sets of known points in two DSMs $\mathbf{u}$ and $\mathbf{v}$. The mean and standard deviation of $\mathbf{u}$ on $\hat{\Omega}$ are denoted respectively $\mu_{\mathbf{u}}(\hat{\Omega})$ and $\sigma_{\mathbf{u}}(\hat{\Omega})$.

According to [FdFML17], this is motivated by two observations:

- The misalignment induced by the satellite pointing error is mainly a translation [GD03, SAM+16, dFMLM+14b].
- As long as the 3D geometry of the area does not change too much, matching surface models is more stable over time than using tie points across multi-date images.

The maximum correlation translations are employed to register all DSMs to the frame of reference of the first input stereo pair, which is expected to be the best according to the selection criterion used (see Section 3.7.1). After the alignment, the point-wise median is used to perform the DSMs fusion, as in [GF18]. Remark that the fusion is done using the

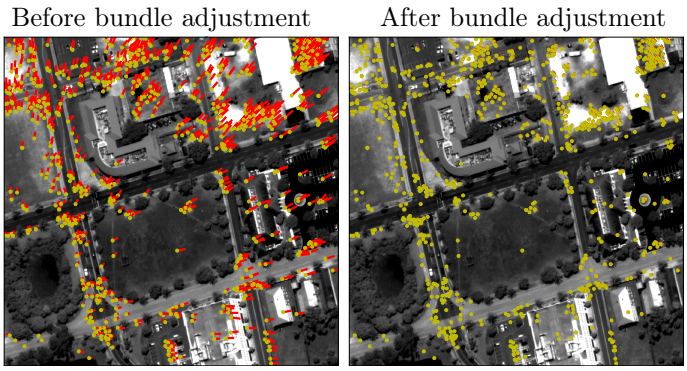


Figure 3.2: Effect of pointing error before and after bundle adjustment. Green dots represent the detected features, and red vectors the distance to the reprojected locations (i.e. reprojection error). After bundle adjustment, the reprojection error reaches subpixel magnitude, implying that the RPCs have been corrected.

DSMs previous to interpolation. Therefore, in practice, the interpolated DSMs are only used to compute the alignment transformations. This is done to minimize possible biases due to large areas of unknown values, as detailed in [FdFML17].

3.4 Bias-compensation bundle adjustment

This MVS pipeline (Algorithm 2), based on [ODM+14], was selected to test a bias-compensation bundle adjustment. The RPCs are corrected previous to the triangulation of stereo correspondences, implying that all DSMs are aligned before the fusion step. The approach aims to find the 2D translations (or correction offsets) in the image domain that compensate the pointing error of each view (see Figure 3.2). Bias correction methods are one of the simplest forms of bundle block adjustment approaches, which can also include polynomial terms in the corrective function [GD03, FH05].

The first step of the pipeline consists in the detection of feature tracks across the set of input images. The feature tracks employed in our experiments result from pairwise matches of SIFT keypoints [Low04]. We apply a distance ratio test as in [Low04] with a threshold of 0.6, and also perform geometric filtering using the Fundamental matrix to minimize the presence of outliers. The union-find algorithm from [MM12] is used to extend pairwise matches to unordered tracks of arbitrary length in an efficient way.

The feature tracks are used to initialize a sparse 3D point cloud and a correction offset for each image, providing this way the necessary inputs for the bundle adjustment. Algorithm 3, introduced in [ODM+14], details how to initialize the correction offsets and the 3D points from the feature tracks.

Initializing the set of 3D points is not a straight-forward task: the direct triangulation of feature track observations would result in different 3D coordinates due to the pointing errors. Similarly to [ODM+14] we use the unrefined RPC localization functions, to back-project a ray from each track observation to different altitudes $\{Z_{\min}, ..., Z_{\max}\}$, given by a series of ΔZ increments and covering the altitude range of the scene. The altitude Z where the multiple back-projections are less scattered (i.e. minimum σ_Z (3.5)) defines the initial altitude of the 3D point associated to the track. The (X, Y) coordinates are given by the mean (μ_X, μ_Y) of all back-projections at altitude Z.

Algorithm 2: *MVS with bias-compensation bundle adjustment*

Input : M views of a small area of interest (AOI) cropped from multi-date satellite images + associated RPCs
Output: High-quality DSM of the input AOI

- Feature track detection across the M input images
- Run Algorithm 3 to compute:
  (1) All possible correction offsets for all images
  (2) An initial value for the 3D point of each track, i.e. point cloud X
for *each image* I_m **in** $\{I_1, ..., I_M\}$ **do**
  - RANSAC to select an offset $\boldsymbol{\rho}_m$ with a large support
- Bias-compensation bundle adjustment to refine the M correction offsets and X, $\rightarrow \{\boldsymbol{\rho}_{\mathrm{BA}_m}\}_{m=1,...M},\ X_{\mathrm{BA}}$
- Select N stereo pairs: $\{(I_n, I'_n)\}_{n=1,...N}$
for *each stereo pair* (I_n, I'_n) **do**
  - Affine approx. of corrected RPCs $\rightarrow P_{\mathrm{BA}_n}, P'_{\mathrm{BA}_n}$
  - Epipolar rectification
  - Dense stereo matching
  - Triangulate using $P_{\mathrm{BA}_n}, P'_{\mathrm{BA}_n}$ to get 3D point cloud
  - Project point cloud to DSM_n
  - Post-process DSM_n
- Fuse all DSMs via point-wise median

Algorithm 3: *Correction offsets from feature tracks*

Input : M views of a small area of interest (AOI) cropped from multi-date satellite images + associated RPCs
Feature track k detected across the input images
Output: M correction offsets, one per image
3D point $\boldsymbol{X}_k$ corresponding to feature track k

for *each altitude* Z **in** $\{Z_{\min}, ..., Z_{\max}\}$ **do**
  for *each image* I_m **in** $\{I_1, ..., I_M\}$ **do**
    - Pick the 2D observation of track k in $I_m \rightarrow \boldsymbol{x}_{mk}$
    - Localize $\boldsymbol{x}_{mk}$ at altitude Z via $\mathrm{RPC}_m \rightarrow (X, Y, Z)$
    - Add (X, Y, Z) to the list of 3D point candidates for track k at altitude $Z \rightarrow \mathrm{LIST}_k$
  - Compute σ_Z of LIST_k using (3.5)
- Define the 3D point of track k as $\boldsymbol{X}_k = \min_{\sigma_Z} \mathrm{LIST}_k$
for *each image* I_m **in** $\{I_1, ..., I_M\}$ **do**
  - Project $\boldsymbol{X}_k$ via $\mathrm{RPC}_m \rightarrow \hat{\boldsymbol{x}}_{mk}$
  - Compute the correction offset $\rightarrow \boldsymbol{\rho}_{mk} = \hat{\boldsymbol{x}}_{mk} - \boldsymbol{x}_{mk}$

The scatter value σ_Z at altitude Z for a given feature track is defined as

$$\sigma_Z = \sqrt{\sum_i (X_i - \mu_X)^2 + \sum_i (Y_i - \mu_Y)^2}, \tag{3.5}$$

where X_i and Y_i are the coordinates of the back-projection of the i-th observation of the feature track at altitude Z.

To initialize the correction offsets of all input images, all possible offsets per image are computed using each feature track. Note that error-free tracks should ideally generate the same offsets for all images if the model holds. Adaptive RANSAC can be applied then to pick a single correction offset per image with the largest consensus [HZ04]. The RANSAC

Algorithm 4: *MVS with orientation-compensation bundle adjustment*

Input : M views of a small area of interest (AOI) cropped from multi-date satellite images + associated RPCs
Output: High-quality DSM of the input AOI

- Feature track detection across the M input images
- Affine approx. of all raw RPCs $\rightarrow \{ P_m \}_{m=1,\dots M}$
- Initialize sparse point cloud from feature tracks $\rightarrow X$
- Bundle adjustment to refine the M rotation matrices and $X \rightarrow \{ P_{\mathrm{BA}_m} \}_{m=1,\dots M},\ X_{\mathrm{BA}}$
- Select N stereo pairs: $\{ (I_n, I'_n) \}_{n=1,\dots N}$

for *each stereo pair* (I_n, I'_n) **do**
- Epipolar rectification
- Dense stereo matching
- Triangulate using $P_{\mathrm{BA}_n}, P'_{\mathrm{BA}_n}$ to get 3D point cloud
- Project point cloud to DSM_n
- Post-process DSM_n

- Fuse all DSMs via point-wise median

threshold used to declare inliers was set to 3 pixels. Further refinement of the tracks and the initial location of the 3D points is done in [ODM+14] by preserving, for each image, only those observations that contributed to an inlier offset. For simplicity, we kept all raw feature tracks and 3D points as output by Algorithm 3 to feed the bundle adjustment.

As mentioned in Section 3.1, apart from the relative geolocation correction method, the rest of the blocks of the approach do not change in comparison to the other MVS pipelines reviewed in this study (i.e. steps of lines 10–19 in Algorithm 2 are the same in Algorithm 1).

3.5 Orientation-compensation bundle adjustment

This pipeline is presented as an alternative to traditional bias-compensation approaches, without renouncing the correction of RPCs. It is a prototype of the RPC refinement methodology introduced in Chapter 4. Instead of optimizing a set of correction offsets, we propose to use bundle adjustment to directly address the orientation of the satellite cameras, by modeling the 3D transformation $\mathcal{R}_m$ that compensates the pointing error (Section 3.2.1) as a rotation matrix. The approach is outlined in Algorithm 4.

As in the previous MVS pipelines, since the area to reconstruct is assumed to be small, the RPCs can be locally modeled as affine camera projection matrices. The affine camera model can be decomposed as

$$P_{3\times 4} = \begin{pmatrix} M_{2\times 3} & \boldsymbol{t}_{2\times 1} \\ 0 & 1 \end{pmatrix}, \tag{3.6}$$

where $M_{2\times 3} = K_{2\times 2} R_{2\times 3}$, being K the calibration matrix and R and $\boldsymbol{t}$ the camera rotation and position respectively. The method aims to refine the rotation matrix R of each camera.

Note that a small rotation of a camera far away from a scene, as it is the case for satellite imagery, amounts in practice to a translation on the image domain [GD03]. This observation suggests that optimizing the rotation matrices R with bundle adjustment should be, at least, equivalent to the traditional offset correction. The rest of parameters of the affine projection matrices are fixed.

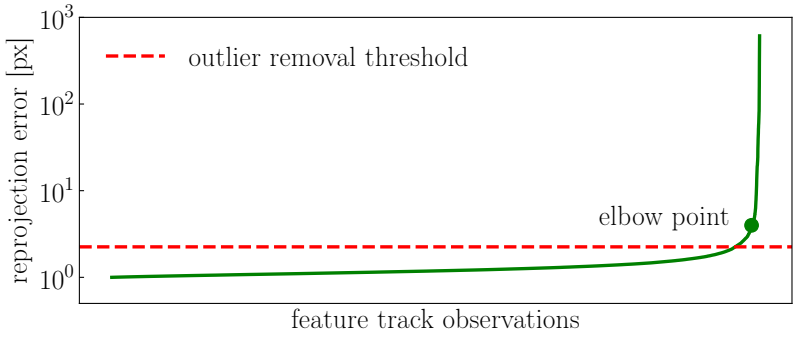


Figure 3.3: Feature track refinement. After an initial bundle adjustment with soft L^1 loss, sorting the reprojection errors typically results into a well-defined elbow-shaped function (due to the large errors caused by a small subset of outliers). We set an outlier removal threshold to the 95th percentile of all values below the elbow point (i.e. the point with largest distance with respect to the line defined by the minimum and maximum errors).

Since it is known that reducing the number of parameters to be optimized aids the bundle adjustment process, we encode all rotation matrices using the Euler angles as a 3-parameter representation. The rotation matrix R_m of the m-th camera is therefore decomposed into three elemental rotations, where the Euler angles ϕ, θ, α are the angles of rotation around the canonical axes:

$$R_m = R_x(\phi_m)R_y(\theta_m)R_z(\alpha_m) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\phi_m & -\sin\phi_m \\ 0 & \sin\phi_m & \cos\phi_m \end{pmatrix} \begin{pmatrix} \cos\theta_m & 0 & \sin\theta_m \\ 0 & 1 & 0 \\ -\sin\theta_m & 0 & \cos\theta_m \end{pmatrix} \begin{pmatrix} \cos\alpha_m & -\sin\alpha_m & 0 \\ \sin\alpha_m & \cos\alpha_m & 0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (3.7)$$

In Algorithm 4, feature tracks are detected following the same methodology from Section 3.4. Differently from Algorithm 3, the 3D points associated to the tracks are initialized by triangulating all pairwise matches per track and taking the mean of the resulting 3D locations, which is faster than the technique used in Algorithm 3.

3.6 Feature track refinement

It should be noted that the result of the bundle adjustment based approaches is highly dependent on the quality of the input feature tracks found across the input images.

The classic bundle adjustment loss considers the squared distances between projected and observed locations (i.e. $L^2(d) = d^2$, where d denotes the Euclidean distance), since the L^2 norm is well-posed for differential calculus and the optimization converges rapidly. However, this loss is very sensitive to the presence of outliers, which can cause the result to be biased according to erroneous tracks at the expense of good observations.

As a solution, a combination of the L^1 and L^2 losses such as the soft L^1 loss can be used. The soft L^1 loss is defined as

$$L^1_{\text{soft}}(d) = 2\left(\sqrt{1+d^2} - 1\right), \quad (3.8)$$

where d is the Euclidean distance between two points.

The soft L^1 loss (3.8) offers higher robustness to outliers. It behaves as a linear loss for large distances, which are likely to be caused by outliers; and as a quadratic loss for smaller distances around 1 pixel or below, likely to be caused by inliers.

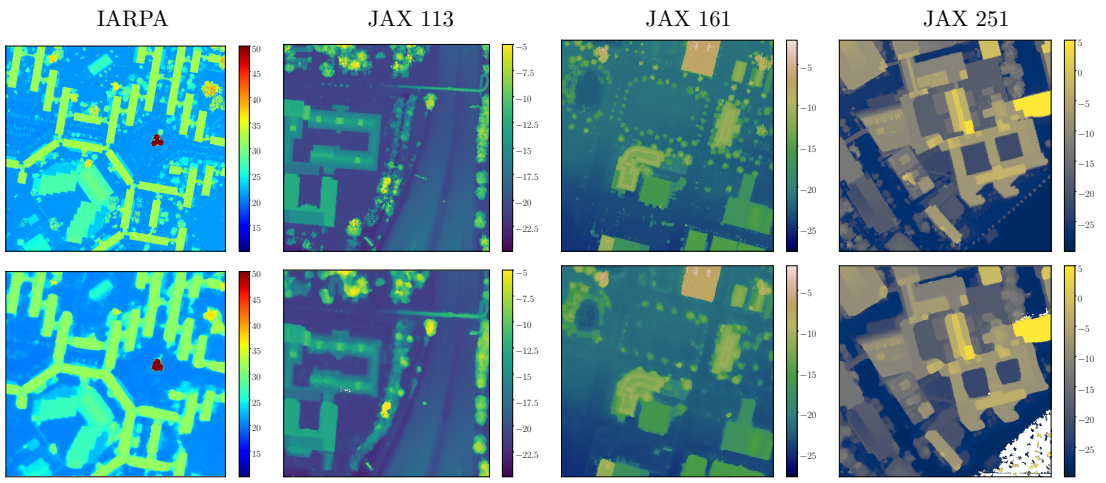


Figure 3.4: Lidar (top) versus photogrammetric DSMs from multi-date satellite images (bottom). For each AOI, the reconstructed DSM that obtained the highest completeness score is displayed in the bottom row. Colorbars on the right of the DSMs assign a color to each altitude, expressed in meters. White areas represent unknown values.

Based on to the previous, we employ the following procedure to improve the quality of the feature tracks. We run two successive bundle adjustment steps: the first one uses the soft L^1 loss and the second one the L^2 loss. Thanks to the soft L^1 loss, after the first run we can expect the gap in terms of reprojection error between inlier and outlier observations to increase. As shown in Figure 3.3, a threshold can be set to discard erroneous observations according to this error. The remaining tracks (presumably reliable inliers) are kept for the second run, using the L^2 loss, which yields the optimal estimator for Gaussian perturbations and quickly converges to a refined solution.

3.7 Evaluation with WorldView-3 image pairs

This section presents the conducted experiments and the results of our study. Table 3.1 summarizes the performance metrics for an area of interest (AOI) from the IARPA dataset and three AOIs from the DFC2019 dataset respectively. Examples of output DSMs are shown in Figure 3.4. Note that the reconstruction may contain unknown values, represented as white points in Figure 3.4, if no stereo pair finds a reliable correspondence for certain areas. The DSM resolution is set to 30 cm per pixel.

The reconstructed DSMs and the ground truth DSMs of each site may not be in the same frame of reference. We employ a translation that maximizes the correlation between both models to register them, following the procedure from Section 3.3. After this, the performance metrics are computed from the error between the two surfaces, which is defined as the point-wise absolute difference (in meters).

In Table 3.1, completeness represents the percentage of points whose error is less than 1 meter, with unknown values being counted as larger errors. The accuracy value, in meters, is the root mean square error (RMSE) of all known points. Both metrics are defined in [BKHB16]. Points within water bodies were not taken into account.

	IARPA	JAX 113	JAX 161	JAX 251
Oracle order				
Correlation-based DSM alignment	70.62 / 2.67	–	–	–
Bias-compensation BA - *naive*	64.39 / 2.72	–	–	–
Bias-compensation BA	70.63 / 2.74	–	–	–
Orientation-compensation BA - *naive*	64.50 / 2.71	–	–	–
Orientation-compensation BA	70.71 / 2.74	–	–	–
Heuristic order				
Correlation-based DSM alignment	68.08 / 2.69	77.72 / 2.00	82.75 / 1.70	74.87 / 2.90
Bias-compensation BA	69.73 / 2.74	77.74 / 2.04	82.53 / 1.73	76.86 / 2.91
Orientation-compensation BA	69.89 / 2.75	77.72 / 2.04	82.60 / 1.72	75.91 / 2.91
SIFT order				
Correlation-based DSM alignment	48.84 / 2.62	76.73 / 2.01	82.64 / 1.64	72.46 / 2.74
Bias-compensation BA	42.15 / 2.71	76.83 / 2.06	82.48 / 1.66	72.69 / 2.76
Orientation-compensation BA	42.15 / 2.71	76.79 / 2.04	82.44 / 1.66	71.14 / 2.78

Table 3.1: Completeness (%) / Accuracy (m) of the reconstructed DSMs for the IARPA (Buenos Aires, one AOI) and DFC2019 (Jacksonville, AOIs 161, 251 and 113) datasets. The *naive* label indicates that a single bundle adjustment (BA) run with classic L^2 loss for reprojection errors was used. Otherwise, the feature track refinement strategy from Section 3.6 was used in the bundle adjustment procedures.

3.7.1 Selection of input pairs

Previous work already highlighted the importance of the criterion used to select input stereo pairs for satellite MVS pipelines [FdFML17, GF18]. Poor choices lead to pairs of views sharing less visual content and output DSMs with larger errors and incomplete areas, making the correction of the pointing error harder both for image and geometry based approaches. We run the MVS pipelines using different criteria to assess the robustness of the geolocation correction methods:

- **Oracle order:** Obtained by computing the DSM of each possible stereo pair and sorting the pairs by decreasing completeness. It guarantees that the best pairs are selected, but it is expensive to compute and unrealistic since ground truth may not be available and is needed to compute the completeness of each pair [FdFML17, Qin17].
- **Heuristic order:** Detailed in [FdFML17], this criterion is an attempt to emulate the oracle order based on the metadata of the satellite images. Stereo pairs are sorted according to the intersection angle, incidence angle and proximity of acquisition date.
- **SIFT order:** The number of pairwise matches can be interpreted as a measure of the shared visual content between two images. This order sorts all possible pairs in decreasing number of SIFT matches, therefore prioritizing pairs with a higher overlap of visual content.

In all experiments, the best (i.e. the first) 50 pairs according to each selection criterion were employed to reconstruct the AOIs. Due to the high computational cost, the oracle order was only computed for the IARPA dataset.

3.7.2 Results and discussion

Table 3.1 reflects the concepts anticipated in Section 3.6. First of all, we can verify that naive bundle adjustment (single run, classic L^2 norm for reprojection errors) produces worse DSMs compared to the rest, even when optimal stereo pairs are used, underlining the need of strategies to handle outliers in the feature tracks.

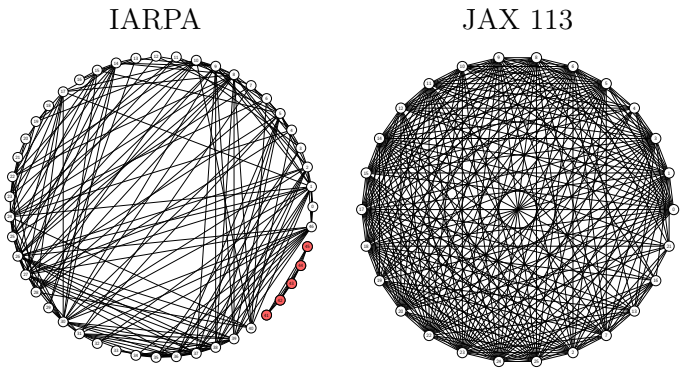


Figure 3.5: Connectivity graphs for the images of the IARPA and JAX 113 AOIs (edges link images with more than 40 matches).

Table 3.1 also supports the assumption that adjusting a 2D translation at image level or the 3D rotation of a satellite camera is almost equivalent, with the results being extremely similar in all experiments involving these strategies. It follows that complex methods to initialize the sparse point cloud input to the bundle adjustment, as Algorithm 3, can be replaced by simpler and more efficient methods such as the one employed in Section 3.5.

In all experiments, the overall RMSE of correlation based DSM alignment is slightly smaller. This bias seems natural considering that the maximum NCC minimizes the difference in location between all points of the DSMs; whereas bundle adjustment strategies only use a reduced amount of keypoints (i.e. feature tracks) to register them.

Given the similarity of the numerical values, the idea that seems to stand out is that both geometry-based and image-based solutions are valid for overcoming geolocation errors. Most of the experiments produced DSMs with completeness score above 65%. However, in certain scenarios, some of the pipelines experimented a loss of accuracy.

Failure prone cases for image-based corrections. In the IARPA AOI, the correlation based DSM alignment clearly outperformed bundle adjustment methods when using the number of SIFT matches to select input stereo pairs instead of the oracle or the heuristic orders.

The loss of accuracy of bundle adjustment methods, illustrated in Figure 3.6, can be explained by looking at the connectivity graph of the images according to the number of SIFT matches: it turns out that there is a group of 5 nodes weakly connected to the rest. The 5 nodes correspond to images taken from a similar viewpoint, with a large incidence angle. Consequently, they have very strong intra-similarity but less resemblance to the rest. If we only display edges accounting for more than 40 matches, the 5 nodes (in red) are disconnected from the others (see Figure 3.5, IARPA). The heuristic order does not use these views because they are too tilted. Still, red nodes exhibit a large amount of matches between them, so several pairs from the set are selected by the SIFT order. What happens then is that the bundle adjustment may end up putting all white nodes into a common frame of reference, while red nodes are adjusted to a different frame that fits better their particular set of observations. This idea can be verified by exploration of the DSMs obtained from pairs of white and red nodes.

Oppositely to bundle adjustment, the correlation based DSM alignment offers higher ro-

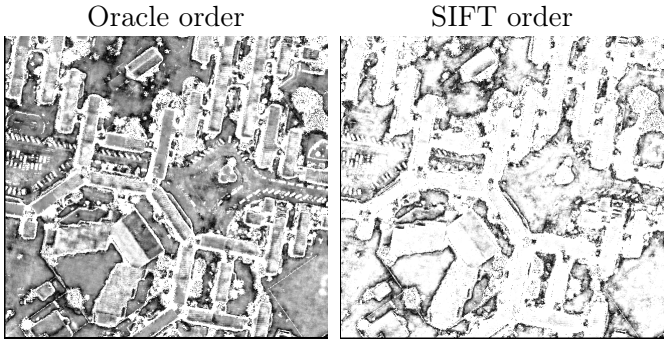


Figure 3.6: Reconstruction error of the DSM obtained with bundle adjustment of camera rotations (IARPA AOI), using the oracle and the SIFT orders. Brighter values account for larger errors. Errors above 1 m are clipped and correspond to white pixels.

bustness to this scenario, since it was conceived to deal with non registered DSMs. This case demonstrates that bundle adjustment algorithms require not only quality feature tracks to work properly, but in addition such tracks should connect the graph of input images in a consistent manner. Otherwise, it is better to avoid incorporating stereo pairs from disconnected sets.

In contrast to the IARPA case, the heuristic and SIFT orders yield similar results for the JAX AOIs. The connectivity graphs of the Jacksonville images are much more consistent (see Figure 3.5, JAX 113). Accordingly, the use of SIFT matches as a selection criterion seems appropriate. For JAX AOIs, more than 20 of the pairs selected by the SIFT order are also chosen by the heuristic order; for IARPA, only 2 pairs coincide in both orders.

Failure prone cases for geometry-based correction. In Table 3.1 we can see that several results obtained by DSM alignment achieve worse completeness with respect to the bundle adjustment counterparts (e.g., JAX 251 and IARPA with heuristic order). These errors can be attributed to the fact that correlation based alignment is sensitive to incomplete geometry, especially if the holes in the DSMs are relevant with respect to the size of the AOI, and outlier altitude values generated by the dense stereo matching process. Even in the absence of major radiometric changes, incomplete areas in the DSM may be caused by occlusions, which are common in areas with many structures, as in the concerned sites; or by water bodies, as in the lower right corner of JAX 251.

Altitude errors unrelated to RPC inaccuracies. There is an amount of error in the output DSMs that is unrelated to RPC inaccuracies. DSM errors at the edges of buildings or in vegetation areas are typically larger than 1 m (see Figure 3.6, oracle order). This is not surprising since in these areas it is necessary to choose between two extremely different modes: a 3D point belongs either to the floor or to a rooftop/tree. Therefore, the use of a median filter to merge the altitude values in these parts of the DSMs is suboptimal [FdFML17].

3.8 Conclusion

All approaches proved to be valid solutions to the problem. In general, they achieved similar evaluation metrics, but certain differences revealed weak points of each method. The geometry-based strategy emerged as a more robust solution when stereo pairs from

weakly connected subsets of images (in terms of feature matching) are used as input. Under a sufficient number of feature tracks, bundle adjustment approaches provided superior performance instead, because geometry-based solutions struggle with incomplete geometry models and outlier altitude values generated by the dense stereo matching process.

The conducted experiments highlighted that suboptimal selections of stereo pairs hinder the success of geolocation correction strategies. It was also shown that considering the distribution of pairwise matches can be a valuable additional source of information to discard inconvenient views or pairs. This insight could complement the heuristics used in the literature to select input pairs, mainly focused on the image metadata. The fact that the metadata affects the entire satellite images and is not specific to particular areas of interest argues in favor of this idea when working with small areas of interest.

We empirically verified that optimizing a 3D rotation matrix to compensate RPC inaccuracies leads to similar performance compared to bias-compensation bundle adjustment commonly used for RPC correction. Chapter 4 extends this idea and shows that a direct composition of the original RPC functions with a corrective rotation matrix is possible, thus removing the limitation to small areas of interest where the input cameras are approximated as local affine models.

Last but not least, it should be noted that the investigated methods are not incompatible. Pre-stereo bundle adjustment strategies could be followed by an ideally redundant post-stereo geometry registration, to further ensure robustness against geolocation errors.

4 Bundle adjustment of RPC camera models

This chapter delves into the problem of correcting the inaccuracies of Rational Polynomial Coefficients (RPC) camera models in satellite multi-view stereo pipelines. We propose a complete bundle adjustment methodology that builds upon one of the strategies introduced in Chapter 3, which compensates sensor orientation-related errors with a rotation in the object space. A set of tie points, automatically derived from feature tracks detected across the input images, is used to find the optimal corrective rotations by minimization of the reprojection error. Significant improvements are introduced: the local approximation of affine projection models is avoided in favor of directly using the original RPC functions, optimal tie point selection is incorporated to handle larger multi-view collections efficiently, and an RPC fitting algorithm is used to ensure that the output camera models follow the same format as the input ones. The stereo pair selection criterion is also adapted to work with SkySat images, which can form pairs with limited overlap and small baseline, as different images from the same acquisition are captured almost instantaneously. The resulting method is tested using SkySat acquisitions over two different areas of interest, and the output RPC models are subsequently used in a satellite stereo pipeline to extract digital surface models. In the absence of ground truth geometry models, the deviation between altitude values from different pairwise models is used as evaluation metric. The results indicate that the proposed bundle adjustment leads to better pairwise geometry registration against explicit geometry registration algorithms. Tie point reprojection errors reach subpixel values and the altitude standard deviation is reduced from the order of meters to tens of centimeters.

4.1 Introduction

As previously introduced in Section 1.4.1, the Rational Polynomial Coefficients (RPC) camera model is a generic sensor model, which is widely used to describe the acquisition process of optical satellite images independently from the specific physical properties of the sensor. The RPC model associated with each satellite image relates 3D object space coordinates to 2D image coordinates.

The RPC functions allow to triangulate the location of a 3D point observed in two or more images. Although RPCs provided by image vendors are expected to be precise enough, the complex system they encode is subject to measurement errors, which are mainly related to the attitude angles that define the sensor orientation. Such inaccuracies can easily cause a 3D point to project tens of pixels away from its real location on the image plane. This behavior implies that different views of a scene are typically not consistent in a common frame of reference. The RPC models may therefore introduce systematic errors in the triangulation of corresponding image points, substantially degrading the accuracy

of 3D models reconstructed from multiple views.

Nowadays, small push-frame satellites such as SkySat from Planet are capable of covering large areas of the Earth by acquiring long strips of partially overlapping frames with a small geographic footprint [MSS+14]. In stereo acquisition mode, a second (or even a third) strip of images is captured during the same passage with a different inclination after reorienting the satellite. RPC inconsistencies are especially problematic when it comes to using such fragmented acquisitions for 3D reconstruction. The reason is that the relief of the entire area of interest can only be obtained by fusing large collections of local and partially overlapping 3D models derived from the small footprint images. RPC inaccuracies produce different coordinates for corresponding local surfaces, which enormously increases the difficulty of the fusion process. In this context, designing efficient RPC correction strategies has become essential to effectively harness push-frame satellite imagery for large-scale multi-view 3D reconstruction, as illustrated in Figure 4.1.

This chapter describes a bundle adjustment methodology to refine the RPC models of a group of satellite images observing a certain area of interest (AOI) from different viewpoints. The main contributions are:

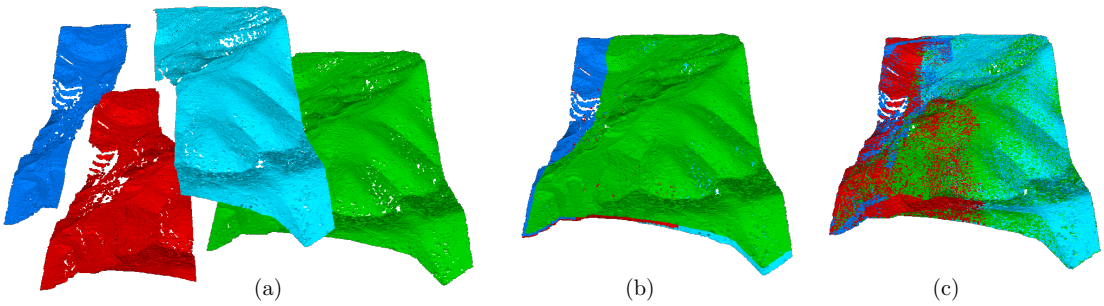
- A minimalist and open-source solution, which seeks to estimate a light and non-redundant amount of parameters to address the root of geolocation inaccuracies, i.e. the sensor orientation. The only inputs that are strictly necessary are the initial RPC models and the corresponding georeferenced images. The code is available at `https://github.com/centreborelli/sat-bundleadjust`.

- A standalone design. The refined cameras are output using the same standard as the input cameras, the RPC model, which makes them directly pluggable to any satellite stereo pipeline. Open-source RPC correction tools sometimes employ different formats to represent the corrected camera models (e.g., NASA Ames stereo pipeline [BAM18]) which is useful for internal use but not for reproducibility.

- A performance evaluation based on the practical impact in a multi-view stereo framework. We compute the deviation between altitude values derived from different pairs of SkySat images, before and after the RPC correction, and compare with the equivalent result obtained with a concurrent geometry-based registration method.

4.2 Related work

Over the past two decades, researchers have thoroughly investigated the problem of RPC refinement. The literature can be broadly divided into *indirect* or *direct* approaches. Indirect refining methods propose complementary functions, defined in image or object space, which are composed with the original RPC functions to correct the camera models. Oppositely, direct methods seek to explicitly modify the RPC polynomial coefficients. The variables involved in the RPC correction process are usually tuned based on a set of Ground Control Points (GCPs), i.e. tie points seen in the images whose absolute location in the 3D object space is known [GD03, TLW10, XZ09]. In absence of GCPs, keypoint feature detectors can be used to infer a substitute set of tie points between two or more images and perform a relative correction (not absolute) between the input cameras [GF17, GF18, WHP+12]. Due to the complexity of the task, direct methods require a larger number of tie points and are prone to poorer stability and accuracy with respect to indirect methods, which are the most common practice [XZ09].



(a) (b) (c)

Figure 4.1: Toy example illustrating the benefits of RPC refinement on multi-view 3D reconstruction. (a) Independent point clouds computed from four different pairs of satellite images. (b) Relative location of the models in the object space, using unrefined RPC models. (c) Relative location of the models in the object space, using RPC models corrected with the proposed methodology. RPC inaccuracies result in non-consistent reconstructions, which accumulate in different layers, as shown in (b). The refined RPCs solve this problem and provide accurately aligned models, as shown in (c).

Bundle adjustment methods based on the minimization of the reprojection error of the available tie points are a common solution for RPC correction in multi-view stereo pipelines, which are the preferred choice for 3D reconstruction from satellite images (Section 1.4.2). The projection of each tie point across the different cameras after bundle adjustment should ideally result in exactly corresponding 2D locations in the images, implying that the corrected RPCs are consistent in a common frame of reference.

An interesting alternative formulation of the bundle adjustment problem was proposed in [SLB09]. Bundle adjustment creates a significant number of variables to be estimated when the number of observed points is large. To avoid this growth in search variables, [SLB09] introduced an alternative cost function based on the cross product of the projection rays. In particular, the classic reprojection error is replaced by a sum of squares of the minimum distance between the projection rays for each point. In this formulation the 3D point locations are implicit and the search space is reduced to the camera parameters only; however, the authors conclude that the standard bundle adjustment formulation provides similar convergence properties for most applications.

4.2.1 Fundamentals of bundle adjustment approaches for RPC camera correction.

This section is a restatement of Section 3.2 from Chapter 3 and is only included here so that the rest of this chapter can be read in a self-contained manner.

Bundle adjustment is the problem of jointly optimizing the viewing parameters (external and/or internal) of multiple cameras and the 3D locations of the objects they observe [TMHF00]. Given an initial set of K 3D points $\{\boldsymbol{X}_k\}_{k=1,\dots,K}$, and their 2D observations $\boldsymbol{x}_{mk}$ across M cameras with projection functions $\{\mathcal{P}_m\}_{m=1,\dots,M} : \mathbb{R}^3 \to \mathbb{R}^2$, bundle adjustment finds the optimal solution by minimizing the reprojection error of the setting

$$E(\mathcal{P}_m, \boldsymbol{X}_k) = \sum_{k=1}^{K} \sum_{m=1}^{M} \|\boldsymbol{x}_{mk} - \mathcal{P}_m(\boldsymbol{X}_k)\|^2, \tag{4.1}$$

where $\boldsymbol{X}_k$ and $\mathcal{P}_m$ contain the variables to be adjusted. Following (4.1), the reprojection

error is typically defined as the sum of squared Euclidean distances between the estimated reprojected 3D points, $\mathcal{P}_m(\boldsymbol{X}_k)$, and their actual 2D observations $\boldsymbol{x}_{mk}$ in the images.

In the case of satellite images, since cameras are far from the Earth's surface (typically above 500 km), the main component of the reprojection error comes from the inaccurate knowledge of the satellite orientation [GD03]. Thus, the energy in (4.1) could ideally be minimized by composing each projection map $\mathcal{P}_m$ with a 3D transformation $\mathcal{R}_m$, whose objective is to reorient the camera with respect to the object coordinates. This way the problem can be reformulated as

$$E(\mathcal{R}_m, \boldsymbol{X}_k) = \sum_{k=1}^{K} \sum_{m=1}^{M} \|\boldsymbol{x}_{mk} - \mathcal{P}_m(\mathcal{R}_m(\boldsymbol{X}_k))\|^2. \tag{4.2}$$

In [GD03] it was shown that the net effect of RPC inaccuracies related to the attitude angles reduces to a 2D translation (also termed as *correction offset*) in images covering lengths up to 50 km. Based on this observation, *bias compensation* approaches for such scenarios have in common the optimization of a 2D correction offset for each RPC model [dR12, FH05, ODM+14]. Following the notation from (4.1), this amounts to finding the 2D translations $\mathcal{T}_m$ in the image domain that minimize

$$E(\mathcal{T}_m, \boldsymbol{X}_k) = \sum_{k=1}^{K} \sum_{m=1}^{M} \|\boldsymbol{x}_{mk} - \mathcal{T}_m(\mathcal{P}_m(\boldsymbol{X}_k))\|^2. \tag{4.3}$$

Note that (4.2) inserts the correction transformation before the original RPC projection, while (4.3) does it after. More generally, RPCs can also be corrected by composition with polynomial functions depending on the image or object tie point coordinates. This allows modeling other error sources beyond attitude inaccuracies, such as time-dependent drift or lens distortion. Although they are normally orders of magnitude smaller, these errors may not be negligible for certain scenarios (e.g., large images covering hundreds of kilometers) [GD03, TLW10].

4.2.2 To bundle adjust or not?

Chapter 3 investigated the net effect of RPC inaccuracies in small areas of interest, covering lengths up to some hundreds of meters or a few kilometers. RPCs can be locally modeled then as affine cameras, using the first order Taylor approximation [dFMLM+14b, FdFML17, GD03, SAM+16]. Under this approximation, a translation in the image domain is equivalent to a translation in the object world. This implies that triangulation errors due to RPC inaccuracies can ideally be corrected by a 3D translation, which can be estimated after the 3D reconstruction stage, e.g. with a point cloud registration algorithm. Based on this idea, certain satellite stereo pipelines use geometry-based alignments at a local scale to elude any prior RPC correction [FdFML17, HWG+20, MdFMLF19]. Such strategies may offer poor scalability as the amount of models to register increases, as geometry-based alignment is subject to incomplete areas and outlier altitude values derived from the dense stereo matching (see Chapter 3, Section 3.8). Running a bundle adjustment to refine the RPC models seems a more natural and less-constrained solution to the problem, as long as enough tie points are available. To further explore this issue, in the experimental part we also compare the bundle adjustment methodology proposed in this chapter with a geometry-based registration method that co-registers the altitude values derived from independent stereo pairs (Section 4.4.4).

4.2.3 Mathematical formulation of the RPC camera model

As the name indicates, the RPC camera model represents the acquisition process using rational polynomial functions. Both the 3D to 2D mapping (the *projection* function) and its inverse (the *localization* function) are expressed as a ratio of cubic polynomials.

The projection function $\mathcal{P}$ of an RPC model can be expressed as

$$(r, c) = \mathcal{P}(X, Y, Z) = \left(\frac{a(X, Y, Z)}{b(X, Y, Z)}, \frac{e(X, Y, Z)}{f(X, Y, Z)} \right), \tag{4.4}$$

where a, b, e, f are cubic polynomials; X, Y, Z represent the longitude, latitude and altitude of a 3D point; and r, c are the row and column of its projection on the image plane. In practice (X, Y, Z) and (r, c) are expressed in normalized coordinates, within a range of $[-1, 1]$, for numerical stability [AMdF+21]. Normalized coordinates are obtained using two scalars, an offset and a scale factor: $X_n = (X - X_{\text{offset}})/X_{\text{scale}}$, where X_n denotes the normalized value of X, and X could be r, c, X, Y or Z in (4.4).

The RPC projection function $\mathcal{P}$ (4.4) is characterized by the cubic polynomials a, b, e, f. Each of these polynomials has 20 coefficients and is defined as

$$\begin{aligned} p\,(X, Y, Z) = {} & p_0 + p_1 Z + p_2 Y + p_3 X + p_4 ZY + p_5 ZX + p_6 YX + p_7 X^2 + p_8 Y^2 \\ & + p_9 Z^2 + p_{10} ZYX + p_{11} Z^2 Y + p_{12} Z^2 X + p_{13} Y^2 Z + p_{14} Y^2 X \\ & + p_{15} ZX^2 + p_{16} YX^2 + p_{17} Z^3 + p_{18} Y^3 + p_{19} X^3, \end{aligned} \tag{4.5}$$

where p_i is the i-th coefficient of polynomial p, and p can be a, b, e, or f.

4.3 Methodology description: tie point generation and optimization framework

We present a bundle adjustment methodology to refine the RPC models of a group of satellite images observing a certain AOI from different viewpoints. Our objective is to produce a set of corrected RPC models, consistent with each other, which will result in the automatic alignment of altitude estimates derived from different stereo pairs using a satellite stereo reconstruction pipeline. Differently from the algorithms reviewed in Chapter 3, this method does not approximate the input RPC models as local affine cameras.

A block diagram of our method is shown in Figure 4.2. The first part of the pipeline is dedicated to feature tracking, which serves to initialize the image and object coordinates of a set of tie points observed across the input images. The tie points and their image observations, together with the input camera models, are input to the second part of the pipeline, where the bundle adjustment optimization problem is solved. Based on the solution, a corrected RPC model is finally fitted for each input image. The different blocks of the chain are explained in detail in the following subsections.

4.3.1 Feature tracking

As stated in Section 4.2.1, bundle adjustment problems are solved based on the reprojection of a set of points across the input cameras. The list of 2D coordinates containing the observation of a 3D point across multiple images is known as a *feature track* (Figure 4.3). In particular, the feature tracks employed in our methodology result from pairwise correspondences of distinctive keypoints. This section describes the feature tracking blocks of our pipeline, which are listed in Figure 4.2.

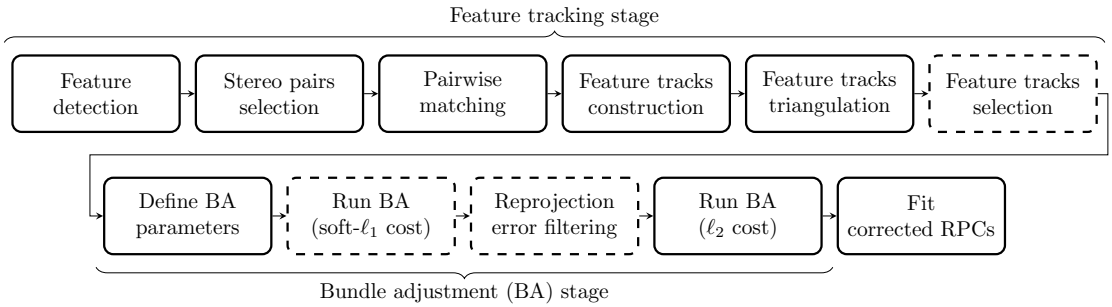


Figure 4.2: Block diagram of the proposed bundle adjustment methodology for RPC correction. Dashed blocks indicate optional but recommended processing, intended to improve accuracy and/or efficiency.

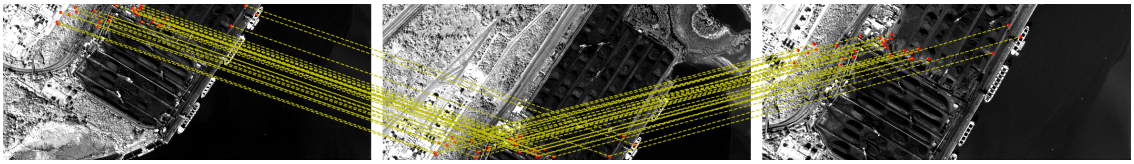


Figure 4.3: Example of 30 feature tracks detected across 3 overlapping views.

Feature detection

We employ the SIFT method to extract distinctive keypoints from the input images [Low04, ROD14]. Each SIFT keypoint is identified by a descriptor of 128 values, invariant to image translation, rotation, and scaling. SIFT descriptors are also robust to s slight changes of viewpoint, noise, blur, contrast changes or scene deformation.

For large images, limiting the maximum number of keypoints per view can be useful to regulate the cost of the matching step, which usually represents the bottleneck of feature tracking algorithms. To this end, we order the SIFT features of each image from the coarsest to finest scale and select the first N_{kp} keypoints to proceed to the matching step. The rationale is that coarse keypoints are less affected by image noise, thus more likely to be seen in the other images. By default, we consider a large value for N_{kp}, 60000 keypoints are allowed per image. Lower values of N_{kp}, e.g. a few thousand points, can be used to speed up the feature tracking stage, at the risk of a possible loss of accuracy. This is because keypoints from the coarser scales are located with less precision. In the same way, fewer keypoint correspondences can be expected to be found as N_{kp} decreases.

Stereo pairs selection

A key aspect of feature tracking is to define which pairs of images are suitable for matching purposes and which pairs can be omitted. Avoiding unnecessary pairs saves computational time and prevents undesired mismatches, which would result in erroneous tie points for bundle adjustment. Algorithm 5 is employed to select suitable image pairs to match from all possible stereo pairs based on the projected area overlap, in UTM coordinates, between the two images. We denote the list of pairs to match as *pairs_to_match*. Out of all the pairs in this list, the subset of pairs offering a well-posed baseline-to-altitude ratio is stored in *pairs_to_triangulate*. The pairs in *pairs_to_triangulate* are used to define the 3D tie point associated with each feature track (Section 4.3.1). Matches from pairs in *pairs_to_match* that are not part of *pairs_to_triangulate* are kept to contribute to

Algorithm 5: Stereo pairs selection

input : list of initial stereo pairs, *input_pairs* [default value = all possible pairs]
minimum image footprint overlap ratio, t_{overlap} [default value = 0.1]
minimum baseline-to-altitude ratio, $t_{\text{baseline-alt}}$ [default value = 1/4]
orbit altitude, in meters [default value = 500000] *SkySat orbit altitude ≈ 500 km*

output : list of stereo pairs suitable to match, *pairs_to_match*
list of stereo pairs suitable to triangulate, *pairs_to_triangulate*

Initialize *pairs_to_match* and *pairs_to_triangulate* as empty lists

for each pair **in** *input_pairs* **do**
  Get the 2 images of the current pair, i.e. I_{ref}, I_{aux}
  Intersect footprints of I_{ref} and I_{aux} in UTM coordinates
  if $\frac{\text{UTM intersection area}}{\text{UTM area } I_{ref}} \geq t_{\text{overlap}}$ **then**
    Add pair to *pairs_to_match*
    Compute baseline between I_{ref} and I_{aux}
    if $\frac{\text{baseline}}{\text{orbit altitude}} \geq t_{\text{baseline-alt}}$ **then**
      Add pair to *pairs_to_triangulate*

the reprojection error of the bundle adjustment stage if they are part of a feature track containing observations in at least one triangulation pair.

Pairwise matching

For each stereo pair in the *pairs_to_match* list, SIFT keypoints are matched using an automatic Fast Approximate Nearest Neighbors algorithm [ML14]. The search for matches is restricted to the regions of the images where the geographic footprints intersect.

Mismatches in the feature track observations can strongly undermine the bundle adjustment performance, leading to failure or strong biases [ASM+12, MdFMLF19]. We remove erroneous matches by means of a distance ratio test with a relative threshold of 0.6 [Low04]. Optionally, if the images cover small geographic footprints and the local affine camera approximation holds, as explained in Section 4.2, this can be followed by a RANSAC geometric filtering using the fundamental matrix[1] [HZ04].

Lastly, the distance d_{geo} between the approximate geographic coordinates of matched keypoints is also used to further remove mismatches. d_{geo} is computed, in meters, as

$$d_{\text{geo}} = \|\text{UTM}(\mathcal{L}_{\text{ref}}(x_{\text{ref}}, y_{\text{ref}}, z_{\text{ref}})) - \text{UTM}(\mathcal{L}_{\text{aux}}(x_{\text{aux}}, y_{\text{aux}}, z_{\text{aux}}))\| \tag{4.6}$$

where $x_{\text{ref}}, y_{\text{ref}}$ and $x_{\text{aux}}, y_{\text{aux}}$ are the image coordinates of a keypoint and its match, respectively. The approximate geographic coordinates of each point are obtained using the localization functions of the two cameras, denoted as $\mathcal{L}$, evaluated at a certain reference altitude, i.e. z_{ref}. The reference altitude is arbitrarily set as the SRTM altitude at the center of the image footprints. The geographic coordinates are converted to the UTM system to compute d_{geo}.

[1] A generic fundamental matrix is used in practice, computed with the 7-point algorithm [HZ04]. In many respects the epipolar geometry of two affine cameras is identical to that of two perspective cameras, with some simplifications due to the fact that the epipoles are at infinity. The RANSAC rejection threshold to fit the fundamental matrix is set to 0.3 pixels.

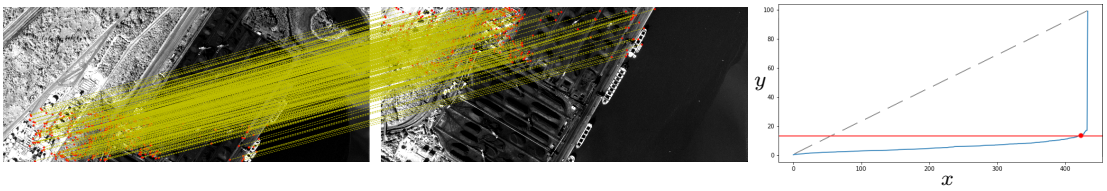


Figure 4.4: Example of pairwise matching. The plot on the right shows, in blue, the sorted approximated geographic distances d_{geo} (y-axis) for each match of a stereo pair (x-axis), computed using (4.6). An outlier rejection threshold τ, in red, is automatically set using the elbow point. The elbow point is defined as the point with largest distance with respect to the dashed line, i.e. the line defined by the minimum and maximum y values.

The values of d_{geo} should concentrate around an offset induced by the RPC errors. In presence of outliers, such distances, if sorted, typically describe a well-defined elbow shaped function. Based on this observation, an outlier rejection threshold τ is automatically set using the elbow point (i.e. the point with largest distance with respect to the line defined by the minimum and maximum d_{geo} values), as shown in Figure 4.4. If τ rejects more than 20% of the matches it is assumed that the distribution does not follow an elbow shape and no thresholding takes place, i.e.

$$\tau = \begin{cases} \tau, & \text{if } \tau \text{ is larger than the 80th percentile of } d_{\text{geo}} \\ \infty, & \text{otherwise.} \end{cases} \tag{4.7}$$

All pairwise matches where $d_{\text{geo}} > \tau$ are discarded.

Feature tracks construction

The recursive union-find algorithm from [MM12] is used to efficiently extend pairwise matches to feature tracks of arbitrary length (Algorithm 6). Keypoints that belong to the same track are assigned a common *parent* value. This strategy is independent of the order in which the stereo pairs are matched.

Algorithm 6: Feature tracks construction

input : list of pairwise matches connecting N keypoints, *pairwise_matches*

output : list of feature tracks, $\{T_j\}_{j=1,\dots,N_{\text{tracks}}}$

Create a unique id for each matched keypoint from 1 to N

Initialize an empty vector *parents* with length N, where each position corresponds to an id

Define the union-find functions:

Function `Find`(*parents, id*)*:*
    parent_id = *parents*[*id*], i.e. the *id*-th value of *parents*
    return *id* **if** *parent_id* is empty; **else** `Find`(*parents, parent_id*)

Function `Union`(*parents,* id_{left}*,* id_{right})*:*
    $parent_id_{\text{left}}$ = `Find`(*parents,* id_{left}) $parent_id_{\text{right}}$ = `Find`(*parents,* id_{right})
    if $parent_id_{\text{left}} \neq parent_id_{\text{right}}$ **then** $parents[parent_id_{\text{left}}] = parent_id_{\text{right}}$

for each pairwise match **in** *pairwise_matches* **do**
    `Union` (*parents*, left keypoint id, right keypoint id)

N_{tracks} = number of unique values in *parents*

Each track T_j corresponds to the set of keypoints whose ids have the same value in *parents*

Algorithm 7: Feature tracks triangulation

input : list of stereo pairs suitable to triangulate, *pairs_to_triangulate*
list of feature tracks, $\{T_j\}_{j=1,\dots,N_{\text{tracks}}}$
output : list of 3D points corresponding to the input feature tracks, $\{X_j\}_{j=1,\dots,N_{\text{tracks}}}$
for each track T_j **do**
  Initialize $\boldsymbol{X}_j = (0, 0, 0)$ — *$\boldsymbol{X}_j$ is the 3D point projecting on the j-th feature track*
  Set $w_j = 0$ — *w_j counts how many triangulation pairs have been used to compute $\boldsymbol{X}_j$*
for each pair **in** *pairs_to_triangulate* **do**
  Get the 2 images of the current pair, i.e. I_{ref}, I_{aux}
  for each track T_j **do**
    if T_j is seen in I_{ref} and I_{aux} **then**
      Get the 2D point observations of T_j in I_{ref} and I_{aux}, i.e. $\boldsymbol{x}_{\text{ref}j}$, $\boldsymbol{x}_{\text{aux}j}$
      $\boldsymbol{X} = \texttt{Triangulate}\ (\boldsymbol{x}_{\text{ref}j}, \boldsymbol{x}_{\text{aux}j})$ — *Triangulate using the RPCs of I_{ref} and I_{aux}*
      $\boldsymbol{X}_j = (w_j \boldsymbol{X}_j + \boldsymbol{X}) \,/\, (w_j + 1)$ — *Update the coordinates of the j-th 3D point*
      $w_j = w_j + 1$ — *Update the counter of pairs employed to compute the j-th 3D point*

Feature tracks triangulation

Similarly to the strategy in Section 3.5, the object coordinates of the tie point that projects onto each feature track are initialized by triangulating all the pairwise matches of the track and taking the mean of the resulting 3D locations (Algorithm 7). The RPC triangulation algorithm from [dFMLM+14a], denoted as `Triangulate` in Algorithm 7, is employed to triangulate keypoint correspondences with the RPC camera models.

Feature tracks selection

In large-scale scenes, using an optimal subset of tracks may benefit bundle adjustment in two ways:

1. Speed up the process and reduce memory usage by removing redundant constraints.
2. Increase the calibration accuracy due to the rejection of tracks with higher localization error.

Our pipeline employs the feature tracks selection method from [CSH17], summarized in Algorithm 8, to select an optimal subset of feature tracks (or equivalently tie points) for the bundle adjustment. Algorithm 8 represents the input cameras and the feature tracks connecting them with an epipolar graph (EG). In the EG each camera is a node and edges represent the epipolar relationship between pairs of nodes, encoded by the fundamental matrices estimated during the pairwise matching step. The weight of each edge can be seen as the number of pairwise matches between two cameras, while the weight associated with the i-th camera, denoted as C_i, is defined as

$$w(C_i) = n(C_i) + e^{-cost(C_i)}, \tag{4.8}$$

where $n(C_i)$ is the number of neighbor cameras of C_i and $cost(C_i)$ is computed as

$$cost(C_i) = mean(C_i) + \mu \cdot std(C_i), \tag{4.9}$$

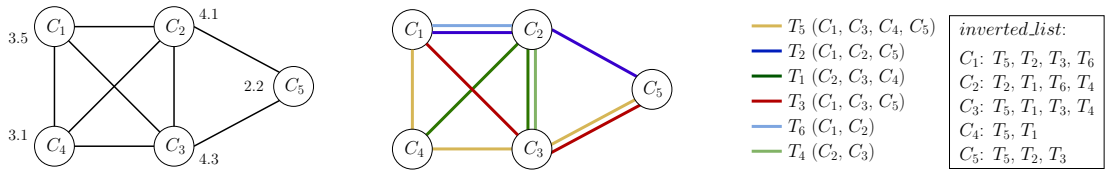
where $mean(C_i)$ and $std(C_i)$ correspond to the mean and the standard deviation of the average reprojection errors of visible tracks in C_i. The scalar μ is a balancing factor set to 3 [CSH17].

Algorithm 8: Feature tracks selection

input : epipolar graph, $EG = \{V, E\}$ *V and E denote the nodes/cameras and edges of EG*
list of tracks, $\{T_j\}_{j=1,\dots,N_{\text{tracks}}}$
number of spanning trees, K_{EG} [default value = 60]

output : subset of tracks, S

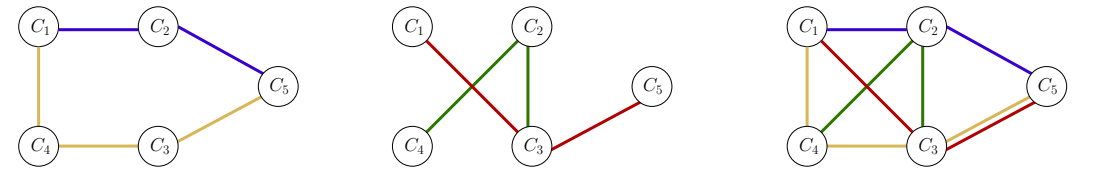
T_{ranked} = set of ranked tracks in T in decreasing priority (*length-scale-cost* criterion)
Initialize $k = 0$ *k is the counter of spanning trees*
Initialize $S = \{\}$ *S is the output subset of tracks*
while $k < K_{\text{EG}}$ **do**
  Compute camera weights *Equation (4.8)*
  C_{root} = camera with largest weight,
  Create *inverted_list* using T_{ranked}
    inverted_list is a list that sorts the tracks seen in each camera using T_{ranked}
  Set $l = 1$ *l counts the layers in the current tree*
  Initialize $H_k(l) = \{C_{\text{root}}\}$ *$H_k(l)$ is the set of cameras in the l-th layer of the current tree*
  Set $h = 1$ *h counts the number of cameras in $H_k(l)$*
  Initialize $S_k = \{\}$ *S_k is the set of tracks selected by the current tree*
  Initialize $I_k = \{C_{\text{root}}\}$ *I_k is the set of cameras connected by the current tree*
  while $V - I_k \neq \emptyset$ **and** $h > 0$ **do**
    Initialize $H_k(l+1) = \{\}$
      $H_k(l+1)$ is the set of cameras to visit in the next layer of the tree
    for each node C_q in $H_k(l)$ **do**
      for each track T_q in *inverted_list* visible in node N_q **do**
        if $T_q \notin S_k$ **then**
          W_q = cameras where T_q is seen
          R_q = neighbor cameras of C_q according to the EG
          $Z_q = W_q \cap R_q$ *Z_q = neighbor cameras of camera C_q where track T_q is seen*
          if $Z_q \not\subset I_k$ **and** $Z_q \neq \emptyset$ **then**
            Add $\{Z_q - Z_q \cap I_k\}$ to $H_k(l+1)$ *neighbor cameras part of the track and not yet visited will be visited in the next layer*
            $S_k = S_k \cup T_q$ *add track to the set of tracks selected by the current tree*
            $I_k = I_k \cup Z_q$ *update the set of cameras visited by the current tree*
    $l = l + 1$
    h = number of cameras in the l-th layer of tree H_k
    Sort cameras in the l-th layer in descending order according to their camera weight
  $S = S \cup S_k$ *add all tracks selected by the current tree to the output subset*
  $T_{\text{ranked}} = T_{\text{ranked}} - S_k$ *remove tracks selected by the current tree from the ranked set*
  $k = k + 1$ *update the counter of spanning trees, a new tree will be explored next*



(a) Epipolar graph (EG), with the weight $w(C_i)$ next to each camera C_i

(b) Input feature tracks, where each track T_j is assigned a different color

(c) Ranked feature tracks from highest to lowest priority and inverted list



(d) First spanning tree selection, i.e. S_k with $k = 0$

(e) Second spanning tree selection, i.e. S_k with $k = 1$

(f) Output set of feature tracks S for $K = 2$, T_6 and T_4 are discarded

Figure 4.5: Toy example of Algorithm 8 applied to 5 cameras connected by 6 feature tracks.

The selection process is a sequential procedure that constructs a number of *spanning trees*, denoted as K_{EG}. Each spanning tree selects a group of feature tracks, from those not chosen before, that can be used to connect all nodes in the EG. An example of EG, with the corresponding camera weights, is shown in Figure 4.5(a). Given the set of feature tracks in Figure 4.5(b), two different spanning tree selections are shown in Figure 4.5(d) and 4.5(e). The final result of the algorithm is obtained from the union of the feature tracks selected by all spanning trees, as illustrated in Figure 4.5(f).

Each spanning tree is constructed following a logic that establishes which cameras of the EG have to be visited first and which of the tracks visible in each camera have to be considered first. In general terms, cameras connected to a larger number of nodes are visited first, and longer tracks are selected first. In case of a tie (e.g. if two cameras are connected to the same number of nodes or if two tracks have the same length), other secondary characteristics are used to decide the order.

More specifically, the aforementioned logic is obtained by sorting the cameras in descending order according to their weight (4.8) and by ranking all input feature tracks from highest to lowest priority. Tracks are ranked using the *length-scale-cost* criterion: first, comparing the number of images where they are visible (*track length*), i.e. longer tracks are given higher priority; then, for those tracks with the same length, the average scale of the keypoints (*track scale*) is considered, i.e. lower scale tracks, susceptible of being located with higher precision, have higher priority; finally, for those tracks equal in length and scale, the ones with lower cost have a higher priority (*track cost*). Similarly to (4.9), the cost associated with the j-th track, denoted as T_j, is defined as

$$cost(T_j) = mean(T_j) + \mu \cdot std(T_j), \tag{4.10}$$

where $mean(T_j)$ and $std(T_j)$ correspond to the mean and the standard deviation of the reprojection errors of all feature observations in track T_j. Based on the tracks ranking, an inverted list, which sorts the visible tracks per camera in decreasing priority, can be generated. An example of an inverted list derived from a set of ranked tracks is shown in Figure 4.5(c).

The only hyper-parameter of Algorithm 8 that needs to be set is K_{EG}. Ultimately, K_{EG} indirectly controls the number of selected tracks, since higher values will result in the union of subsets found by a wider variety of spanning trees. Further details and other toy examples similar to Figure 4.5 describing Algorithm 8 can be found in [CSH17].

4.3.2 Bundle adjustment for RPC model refinement

Bundle adjustment is essentially a nonlinear weighted least squares problem, which can be written in generic form as

$$\min_{\mathbf{x}} \frac{1}{2} \sum_{i=1}^{N} w_i r_i^2(\mathbf{x}), \tag{4.11}$$

where w_i are the weights and r_i is the nonlinear function employed to compute the reprojection residual of the i-th feature track observation, i.e. $r_i = \|\boldsymbol{x}_{mk} - \mathcal{P}_m(\boldsymbol{X}_k)\|$ in (4.1).

The vector $\mathbf{x}$ in (4.11) contains all the parameters to estimate, which consist of

1. The parameters that are necessary to correct each camera model.
2. The coordinates of the tie points in object space.

The initial values of 1 define a set of rotations compensating for RPC inaccuracies (Section 4.3.2), and the initial values of 2 are obtained after the feature track triangulation step (Section 4.3.1).

Given the solution of (4.11), the corrected tie points, projected with the corrected camera models, should ideally match the corresponding feature track observations. Since SIFT keypoints are located with a certain degree of noise, a minimum subpixel error should be expected.

The next section describes the camera correction parameters and the details of the optimization process related to the bundle adjustment stage in Figure 4.2.

Camera correction parameters

The presented pipeline aims to refine the input RPC models by means of a corrective rotation around the camera center, which compensates errors in the attitude angles defining the satellite orientation. Given a 3D point $\boldsymbol{X}$, its bundle adjusted projection is obtained as

$$\boldsymbol{x}_{\text{BA}} = \mathcal{P}(R(\boldsymbol{X} - \boldsymbol{C}) + \boldsymbol{C}), \tag{4.12}$$

where $\boldsymbol{x}_{\text{BA}} = (r_{\text{BA}}, c_{\text{BA}})$, i.e. the bundle adjusted image coordinates, $\mathcal{P}$ is the unrefined projection function of the input RPC model and R is the corrective rotation estimated by bundle adjustment. The camera center $\boldsymbol{C}$ is derived by regressing a projective model from each RPC model, and it remains fixed during the optimization[2]. The corrective rotation R is initialized as the identity matrix and it is represented using Euler angles, which entails three values to optimize per camera:

$$R_m = R_x(\phi_m)R_y(\theta_m)R_z(\alpha_m) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\phi_m & -\sin\phi_m \\ 0 & \sin\phi_m & \cos\phi_m \end{pmatrix} \begin{pmatrix} \cos\theta_m & 0 & \sin\theta_m \\ 0 & 1 & 0 \\ -\sin\theta_m & 0 & \cos\theta_m \end{pmatrix} \begin{pmatrix} \cos\alpha_m & -\sin\alpha_m & 0 \\ \sin\alpha_m & \cos\alpha_m & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{4.13}$$

[2]Each projective model is obtained by creating a 3D grid in the scene space and finding its projection with the RPC. A DLT resectioning algorithm is used to estimate a projective matrix using the 3D to 2D correspondences [HZ04]. The center $\boldsymbol{C}$ of the projective model is used as center of rotation in (4.12). This is a coarse approximation, similar to [BAM18], but from our experience the exact value of $\boldsymbol{C}$ has little impact on the bundle adjustment solution.

where ϕ_m, θ_m and α_m are the three Euler angles to optimize for the m-th camera. This formulation is similar to the one introduced in [MdFMLF19]: the key distinction is that in this chapter the estimated rotation corrects the original RPC model itself instead of a local affine camera that approximates it.

Numerical optimization

The bundle adjustment problem is iteratively solved using a Trust Region Reflective method for large-scale bound-constrained nonlinear minimization problems [BCL99]. We use an implementation from the open-source Python library SciPy [Sci20], which employs the LSMR solver [FS11a, FS11b]. Instead of explicitly computing the derivatives of the problem, the algorithm relies on a finite difference estimation using the sparsity structure of the Jacobian to speed up the optimization process [CPR74, AMB+94].

Trust region methods can be seen as an evolution of Levenberg-Marquardt algorithms (both being *damped* Newton step methods) which are able to follow the negative curvature of the objective function for faster convergence [Ber04, LA05, TMHF00]. Further details on damped Newton methods and the construction of large sparse Jacobian matrices for bundle adjustment problems are given in Appendix A.

The numerical optimization stops iterating when the change of the cost function falls below a certain threshold controlled by a f_{tol} tolerance value. Similarly, other tolerances for termination checking the change of the independent variables (x_{tol}) and the norm of the gradient (g_{tol}) are used as well. The default values of f_{tol}, x_{tol} and g_{tol} are set to 10^{-4}, 10^{-10} and 10^{-8} respectively.

Cost functions and reprojection error based filtering

In addition to the outlier rejection tests conducted during the pairwise matching step (Section 4.3.1), a wide variety of compatible strategies have been studied in the literature to further minimize the impact of erroneous observations in bundle adjustment problems. The rejection of points whose reprojection error exceeds a certain threshold after a series of initial iterations is a popular strategy [SF16, VI18, MdFMLF19]. Other approaches use cost functions that enforce robustness to occasional large residuals, or employ iterative reweighting according to the reprojection error of each observation [TMHF00, ASM+12, Zac14].

In line with the previous ideas, the proposed bundle adjustment optimization starts with some initial iterations, employing a soft-ℓ_1 cost function (a variant of the Huber loss). Equation (4.11) then takes the form

$$\min_{\mathbf{x}} \frac{1}{2} \sum_{i=1}^{N} 2 \left(\sqrt{1 + r_i^2(\mathbf{x})} - 1 \right). \tag{4.14}$$

After this, a second series of iterations is applied using a classic sum of squared ℓ_2 differences, which yields the optimal estimator for Gaussian distributed perturbations. If we assume uniform weights:

$$\min_{\mathbf{x}} \frac{1}{2} \sum_{i=1}^{N} r_i^2(\mathbf{x}). \tag{4.15}$$

At the end of each series of iterations, the average reprojection distance ρ per feature track observation is computed as the average value of r_i, i.e. $\rho = \frac{1}{N} \sum_{i=1}^{N} r_i$

Parameter	Meaning [default value]
N_{kp}	max number of keypoints considered per image (larger scale points have priority) [60000]
t_{overlap}	min geographic area overlap ratio between two images to match, from 0 to 1 [0.1]
$t_{\text{baseline-alt}}$	min baseline-to-altitude factor between two images to triangulate [1/4]
$t_{\text{sift-ratio}}$	SIFT matching relative threshold between nearest and second nearest descriptors [0.6]
K_{EG}	number of spanning trees of the epipolar graph covered after feature track selection [60]
K_{ranking}	ranking criteria for feature track selection, in decreasing priority [*length-scale-cost*]
f_{tol}	tolerance for BA termination by the change of the cost function [10^{-4}]
x_{tol}	tolerance for BA termination by the change of the independent variables [10^{-10}]
g_{tol}	tolerance for BA termination by the norm of the gradient [10^{-8}]
$N_{\text{iter soft-}\ell_1}$	max BA iterations with a soft-ℓ_1 cost function [50]
$N_{\text{iter }\ell_2}$	max BA iterations with a squared ℓ_2 cost function [300]

Table 4.1: Main hyper-parameters of the feature tracking and bundle adjustment blocks.

Since (4.14) offers higher robustness to outliers than (4.15), after the first iterations we expect the gap in terms of reprojection distance between inlier and outlier observations to increase. In presence of outliers, at the end of the soft-ℓ_1 iterations, the sorted observation-wise reprojection distances for each image generally describe an elbow function. The method illustrated in Figure 4.4 can then be used to automatically set a rejection threshold τ_ρ, following the same logic as (4.7). If $r_i > \tau_\rho$, the i-th feature track observation is discarded. The remaining tracks, containing at least two observations, are fed to the second series of iterations, that quickly converges to the final solution.

We set the maximum number of iterations $N_{\text{iter soft-}\ell_1}$ and $N_{\text{iter }\ell_2}$ to 50 and 300, for (4.14) and (4.15) respectively. A summary of all parameters listed so far, their meaning and their default values, can be found in Table 4.1.

4.3.3 RPC model fitting

After the bundle adjustment optimization, a new RPC model is constructed for each input image. The new model encodes the corrected projection function $\mathcal{P}_{\text{BA}}$, which results from the composition of the estimated rotation matrix R and the input RPC projection function $\mathcal{P}$, as defined in (4.12).

The RPC fitting error derived from the i-th tie point with known 3D coordinates (X, Y, Z) and image coordinates (r, c), can be expressed as a vector of two components $(\delta r_i, \delta c_i)$,

$$\delta r_i = \frac{a\,(X,Y,Z)}{b\,(X,Y,Z)} - r = \frac{\boldsymbol{a}^T \boldsymbol{u}_i}{\boldsymbol{b}^T \boldsymbol{u}_i} - r \tag{4.16}$$

$$\delta c_i = \frac{e\,(X,Y,Z)}{f\,(X,Y,Z)} - c = \frac{\boldsymbol{e}^T \boldsymbol{u}_i}{\boldsymbol{f}^T \boldsymbol{u}_i} - c, \tag{4.17}$$

where δr_i and δc_i are the differences in pixel units between the projected and the real image coordinates. The vectors $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{e}$, $\boldsymbol{f}$ contain 20 coefficients each, defining the cubic polynomials a, b, e, f that characterize the RPC projection function (4.4), i.e.

$$\begin{aligned} \boldsymbol{a} &= [a_0 \;\; a_1 \;\; a_2 \;\; \cdots \;\; a_{19}]^T, \quad \boldsymbol{b} = [b_0 \;\; b_1 \;\; b_2 \;\; \cdots \;\; b_{19}]^T \\ \boldsymbol{e} &= [e_0 \;\; e_1 \;\; e_2 \;\; \cdots \;\; e_{19}]^T, \quad \boldsymbol{f} = [f_0 \;\; f_1 \;\; f_2 \;\; \cdots \;\; f_{19}]^T \end{aligned}$$

The vector $\boldsymbol{u}_i$, in turn, contains the variables associated with each polynomial coefficient in (4.5),

$$\begin{aligned} \boldsymbol{u}_i = [&1 \;\; Z \;\; Y \;\; X \;\; ZY \;\; ZX \;\; YX \;\; Z^2 \;\; Y^2 \;\; X^2 \;\; ZYX \\ &Z^2Y \;\; Z^2X \;\; Y^2Z \;\; Y^2X \;\; ZX^2 \;\; YX^2 \;\; X^3 \;\; Y^3 \;\; X^3]. \end{aligned}$$

Algorithm 9: RPC model fitting using bundle adjustment solution (for one camera)

input : input RPC model, with unrefined projection function $\mathcal{P}$ and localization function $\mathcal{L}$
bundle adjusted corrective rotation, R
camera center, $\boldsymbol{C}$
altitude range, $[Z_{\min}, Z_{\max}]$ *this range can be set based on the tie point 3D coords.*
number of samples per dimension, n [default value = 10] *defines a grid of* $n \times n \times n$

output : corrected RPC model, with a new projection function $\mathcal{P}_{\text{BA}}$

Set a certain margin (in pixel units), $m = 10$

Define a regular grid of $n \times n$ 2D points, G_{2D}, covering the input image $+ m$ in each dimension

Define n elevation layers $\{l_i\}_{i=1,\dots,10}$, uniformly distributed within $[Z_{\min}, Z_{\max}]$

Define a grid of $n \times n \times n$ 3D points, G_{3D}, by localizing G_{2D} at each elevation layer l using $\mathcal{L}$

Obtain the corrected image coords. $(r_{\text{BA}}, c_{\text{BA}})$ of each point (X, Y, Z) in G_{3D} *Equation (4.18)*

if the convex hull of the set $(r_{\text{BA}}, c_{\text{BA}})$ covers the entire input image **then**
  Fit $\mathcal{P}_{\text{BA}}$ with [AMdF+21] using the n^3 correspondences between (X, Y, Z) and $(r_{\text{BA}}, c_{\text{BA}})$
else
  Dilate the grid of fitting points by repeating all previous steps using $m = 2m$

Following (4.16) and (4.17), the new RPC coefficients are estimated by solving the optimization problem $\min_{\boldsymbol{a},\boldsymbol{b}} \sum_i |\delta r_i|^2$ and $\min_{\boldsymbol{e},\boldsymbol{f}} \sum_i |\delta c_i|^2$ with a sufficient amount of 3D to 2D point correspondences, which are known as Control Points (CNPs). This is done in the presented pipeline by means of the regularized least squares algorithm described in Chapter 5 [AMdF+21]. The projection function $\mathcal{P}_{\text{BA}}$ obtained from the resulting RPC coefficients encodes the corrected mapping given by the bundle adjustment solution with high precision. The average δr and δc errors are of the order of 10^{-4} pixels or less.

The set of CNPs employed to fit each output RPC model results from a virtual regular grid of $n \times n \times n$ samples in the object space. This grid is projected to the image space using the corrected mapping, to establish the 3D to 2D correspondences that are necessary to fit $\mathcal{P}_{\text{BA}}$ (Algorithm 9).

Drift correction in object space

The proposed methodology does not set any constraint to fix the bounding box that contains the 3D coordinates of the tie points employed by the bundle adjustment. The optimization process is therefore free to move the scene through the object space to find a better fit between the input cameras. As a result, a certain drift may exist between the initial and output location of the scene. To minimize any possible drift, we compose the corrected projection mapping of each camera with a translation in object space, $\boldsymbol{T}_{\text{drift}}$, which is common for all cameras. The corrected projection function that maps a 3D point $\boldsymbol{X}$ to its location $\boldsymbol{x}_{\text{BA}}$ in the m-th image evolves from (4.12) to

$$\boldsymbol{x}_{\text{BA}} = \mathcal{P}_m(R_m(\boldsymbol{X} + \boldsymbol{T}_{\text{drift}} - \boldsymbol{C}_m) + \boldsymbol{C}_m), \tag{4.18}$$

where $\boldsymbol{T}_{\text{drift}}$ is given by the solution of

$$\min_{\boldsymbol{T}_{\text{drift}}} \sum_{k=1}^{K} \|\boldsymbol{X}_k - (\hat{\boldsymbol{X}}_k + \boldsymbol{T}_{\text{drift}})\|^2. \tag{4.19}$$

Following (4.18), $\boldsymbol{T}_{\text{drift}}$ is applied previously to the corrected projection mapping found by bundle adjustment. In (4.19), $\hat{\boldsymbol{X}}_k$ represents the k-th tie point initial 3D location, while

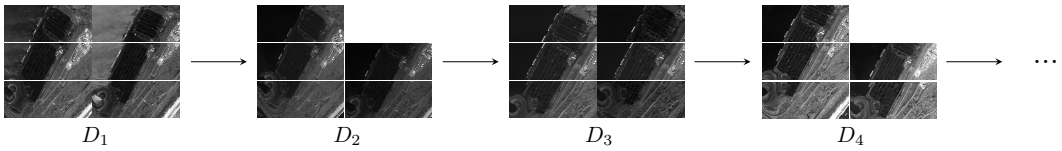


Figure 4.6: Illustration of a time series of SkySat imagery. A collection of partially overlapping views covering a certain area of interest is available at each acquisition date D_i of the time series.

$\boldsymbol{X}_k$ is its location after the bundle adjustment. Therefore $\boldsymbol{T}_{\text{drift}}$ seeks to bring the tie points from the initial unrefined scene location to that location in the object space where the bundle adjustment found the optimal consistence between the input cameras. Note that the absolute location of the corrected camera models remains subject to the absolute localization error of the input RPC functions, but our objective is to ensure consistency between cameras so that they can be exploited for multi-view 3D reconstruction, not their exact absolute location.

4.4 Evaluation and comparison with geometry-based registration

4.4.1 Data

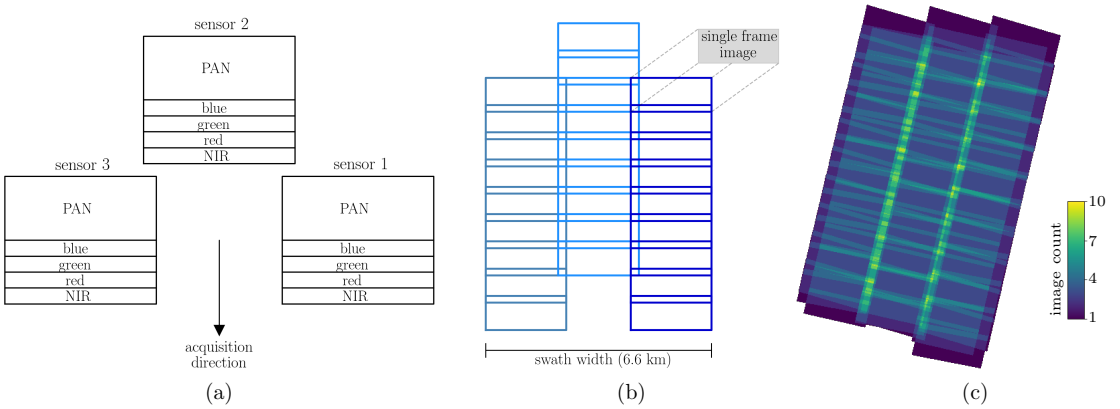
We test our method on two time series of SkySat panchromatic L1B frames, each covering a specific AOI. Each time series consists of groups of partially overlapping images acquired at different dates (Figure 4.6). The SkySat L1B frames have a nadir resolution of ∼0.72 m per pixel and a total size of 1349 × 3199 pixels. The main characteristics of SkySat acquisitions and each time series are detailed next.

SkySat acquisition geometry

SkySat satellites have 3 staggered sensors in their acquisition platform (Figure 4.7(a)), and each sensor acquires a continuous strip of single frame images at a time (Figure 4.7(b)). The minimum swath width across the strips amounts to ∼6.6 km on the ground [Pla19]. There is a small overlap between images of different strips and between images of the same strip [dKR14, AA20]. SkySats can provide stereo and tri-stereo acquisitions, which can be used for 3D reconstruction [Rag06]. The tri-stereo acquisition mode produces image triplets, consisting of a forward, nadir and backward view of each area, taken from the same orbit, with a time difference of a few seconds. Thus, the content of each strip is seen at least three times in the tri-stereo data (Figure 4.7(c)).

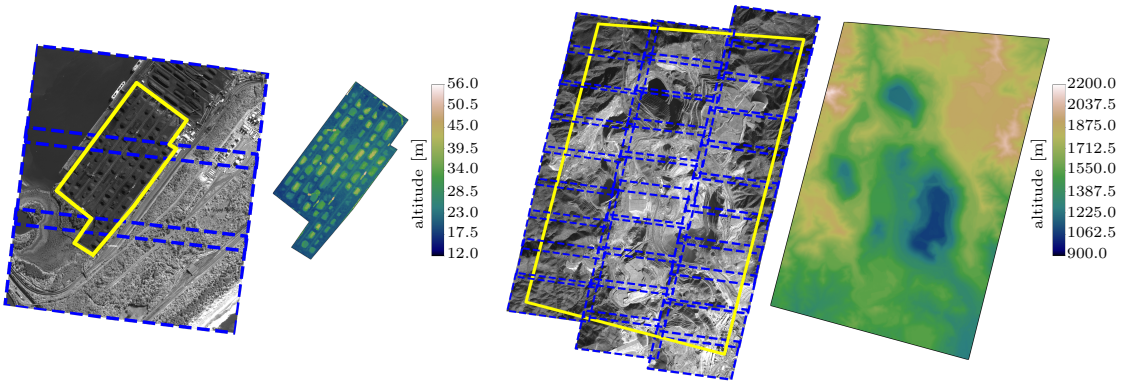
Time series description

- **Richards Bay:** 136 images, distributed in groups of 5-6, over 24 acquisition dates from January to May 2020. The AOI measures 1 km^2 and covers part of the Richards Bay coal terminal (Figure 4.8(a)). The maximum time interval between consecutive dates is 20 days (Figure 4.9(a)).
- **Morenci Mine:** 3 entire tri-stereo acquisitions, from 3 different dates in January 2019. Each tri-stereo acquisition amounts to ∼100 images, resulting in a total of 303



(a) (b) (c)

Figure 4.7: (a) Scheme of a SkySat focal plane. The upper half of the sensors is used for panchromatic capture (PAN), the lower half is divided into 4 bands with the blue, green, red and near infra-red (NIR) filters. (b) Scheme of a SkySat acquisition, where the strips of frames acquired by each sensor are outlined in different colors. (c) Number of overlapping views of a SkySat tri-stereo acquisition.



(a) Richards Bay site (South Africa) (b) Morenci Mine site (United States)

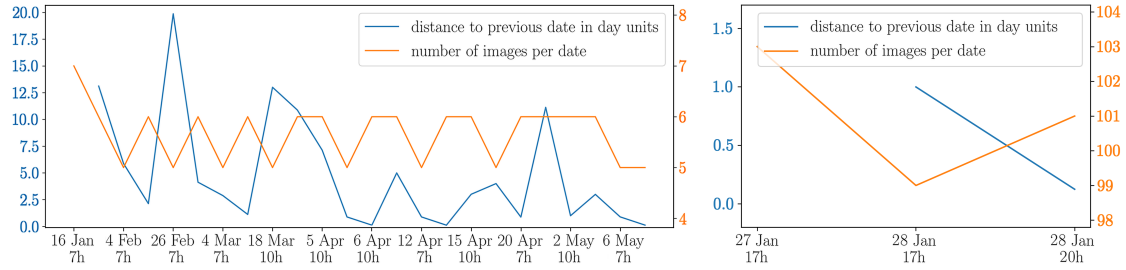
Figure 4.8: On the left, the selected area of interest (AOI) of each time series, marked in yellow, covered by a collection of SkySat L1B frames outlined in dashed blue lines. On the right, a digital surface model (DSM) of the AOI, obtained by multi-view stereo reconstruction after RPC correction.

frames. The AOI (Figure 4.8(b)) measures 62 km^2 and covers the Morenci copper mine. The time interval between consecutive dates is of a few hours (Figure 4.9(b)).

Note that all images in the Richards Bay series were acquired by the same sensor (sensor 2 in Figure 4.7(a)), while the Morenci Mine acquisitions contain the 3 strips per date, i.e. images from the 3 SkySat sensors. Also, note that all the images are free from clouds in the concerned AOIs.

4.4.2 Evaluation procedure

The RPC correction is done in a date-wise manner, as in [MdFMLF21], solving an independent bundle adjustment problem for the group of cameras of each date in the time series. This ensures that the scene geometry is coherent for each group, which is not guaranteed if multiple dates are treated at once.



(a) Richards Bay time series. (b) Morenci Mine time series.

Figure 4.9: Number of images per acquisition date and time distance between consecutive dates.

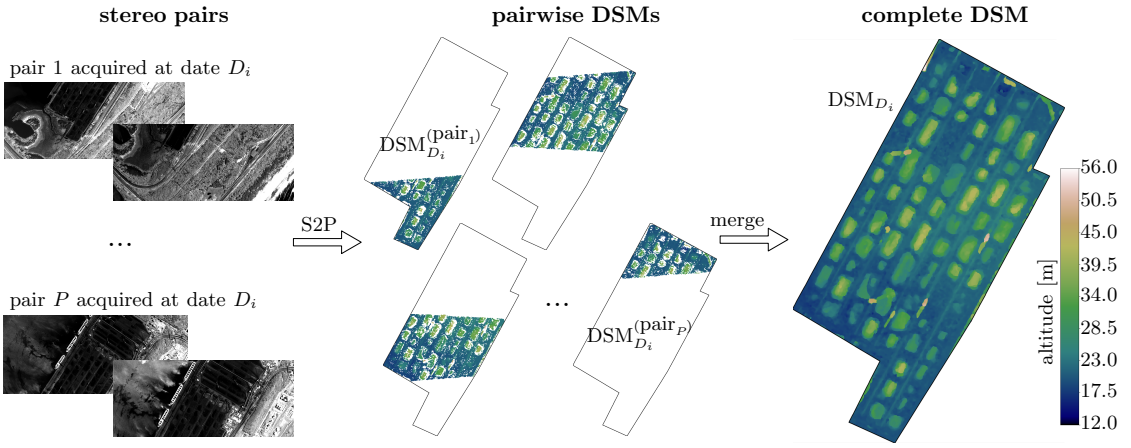


Figure 4.10: Multi-view stereo reconstruction process run at each acquisition date of the time series. The DSMs from different stereo pairs obtained with S2P [dFMLM+14a] using the corrected RPC models are merged into a complete DSM of the entire AOI by taking the average altitude value at each cell.

We evaluate the performance of our method by measuring its practical impact on the registration of 3D reconstructions from different stereo pairs of the same acquisition date. To this end, for each stereo pair, a digital surface model (DSM) is generated with the satellite stereo pipeline SP2 [dFMLM+14a, CC20] using the corrected RPC models resulting from the bundle adjustment. The resulting pairwise models are restricted to the part of the AOI that is visible in each stereo pair, as shown in Figure 4.10. If the RPC correction was successful, the pairwise DSMs of each date should appear automatically registered in the object space. This allows us to easily merge them into a denser and highly accurate model of the entire AOI by taking the average altitude value at each DSM cell. Figure 4.10 illustrates the complete multi-view stereo reconstruction procedure that is carried for each acquisition date of the time series. The DSM resolution is set to 1 m per pixel.

We compute the average evaluation metrics across all dates of each time series. Note that the proposed date-wise evaluation procedure does not seek to register DSMs from different acquisition dates. By splitting the time series into independent bundle adjustment problems, we aim to measure the degree of consistency between cameras that can be attained in a single date scenario, without the risk of reaching a solution that is affected by changes in the geometry of the AOI over time.

4.4.3 Evaluation metrics

The following metrics are employed to assess the functioning of bundle adjustment:

- **Number of feature tracks:** denoted as N_{tracks}, it corresponds to the number of feature tracks employed to perform the bundle adjustment.
- **Number of iterations:** denoted as N_{iters}, it corresponds to the number of iterations completed by the numerical optimization. This amount results from the addition of the number of iterations required to minimize the cost functions (4.14) and (4.15).
- **Reprojection error:** denoted as ρ, it amounts to the average reprojection error, in terms of Euclidean distance and measured in pixels, that is obtained from all feature track observations.
- **DSM altitude deviation:** denoted as σ_{alt}, this measure quantifies the consistency of corresponding altitude values retrieved from different stereo pairs of the same acquisition date. It results from computing the average point-wise standard deviation in altitude, measured in meters, of the 2D points that are seen in more than one of the pairwise DSMs (i.e. middle image in Figure 4.10). Ideally, these DSMs should coincide exactly, with a standard deviation equal to zero. In practice, a small amount of noise derived from the computation of depth from stereo can be expected, but this indicator should reach values as close to zero as possible. Visual examples of σ_{alt} are shown in Figure 4.11. Small holes or unfilled parts in Figure 4.11 correspond to points of the AOI that are not seen by at least two pairwise DSMs, where σ_{alt} cannot be computed.

4.4.4 Comparison with geometry-based DSM registration

We compare the proposed bundle adjustment methodology to an alternative geometry-based strategy that explicitly aligns dense surface models, described in Algorithm 10. As mentioned in Section 4.2, working with images with small geographic footprint allows the local approximation of RPCs as affine cameras. Geolocation errors related to RPC inaccuracies can then be corrected by applying 3D translations in the object space, to register the 3D models derived from different stereo pairs [FdFML17, SAM+16].

Algorithm 10 is different from the maximum normalized cross correlation method reviewed in Chapter 3 (Section 3.3). In this case, the independent DSMs are also aligned based on their geometry, but information from the panchromatic frames originally employed to generate the DSMs is additionally used in the process. Using image based features to compute the transformations for point cloud registration is a well established methodology in 3D vision [SCM10, BJIC17, PA17]. Under reasonable viewing angles, strategies of this kind can offer certain advantages (e.g. fast solution, robustness to different cloud formats or a coarse initialization) compared to classic exhaustive point cloud registration methods that exclusively rely on the 3D coordinates, such as the Iterative Closest Point (ICP) algorithm and its variants [BM92].

As noted in [SAM+16], DSM registration strategies offer little scalability for very large-scale scenarios, with possibly hundreds of surface models to align. In line with the evaluation procedure introduced in Section 4.4.2, we run Algorithm 10 in a date-wise manner, to register the DSMs obtained from stereo pairs of the same acquisition date. The resulting altitude deviation, σ_{alt}, is compared to the value obtained by the analogous experiments relying on the proposed RPC correction pipeline.

Algorithm 10: Geometry based DSM registration (N-inputs)

input : N pairwise DSMs, i.e. $\{DSM^{(n)}\}_{n=1,\dots,N}$
the pairs of images used to compute each DSM, i.e. $I_{\text{ref}}^{(n)}$, $I_{\text{aux}}^{(n)}$ for $DSM^{(n)}$

output : N co-registered DSMs, i.e. $\{DSM_{\text{r}}^{(n)}\}_{n=1,\dots,N}$

Compute all pairwise DSM intersections using the georeferenced bounds of the models

for each DSM intersection **do**

- Retrieve the indices of the intersecting DSMs, denoted as i and j
- Retrieve the reference image of each DSM, denoted as $I_{\text{ref}}^{(i)}$ and $I_{\text{ref}}^{(j)}$
- Find a set of matches, i.e. $\{\boldsymbol{x}_m^{(i)}, \boldsymbol{x}_m^{(j)}\}_{m=1,\dots,M}$, between SIFT keypoints in $I_{\text{ref}}^{(i)}$ and $I_{\text{ref}}^{(j)}$
- Retrieve $\{\boldsymbol{X}_m^{(i)}, \boldsymbol{X}_m^{(j)}\}_{m=1,\dots,M}$, where $\boldsymbol{X}_m^{(i)}$ are the 3D coordinates of $\boldsymbol{x}_m^{(i)}$ as given by $DSM^{(i)}$
- Define the equations derived from the M 3D matches resulting from the previous step:
  - The m-th correspondence amounts to $\boldsymbol{A}_m y = \boldsymbol{b}_m$, where $\boldsymbol{b}_m = \boldsymbol{X}_m^{(j)} - \boldsymbol{X}_m^{(i)}$,
  - $\boldsymbol{A}_m$ is a sparse vector of length N with 1 and -1 at i-th and j-th positions,
  - and y is the solution matrix with shape $N \times 3$

Build the linear system $Ay = b$, where A and b result from stacking all rows $\boldsymbol{A}_m$ and $\boldsymbol{b}_m$

Solve $\min_y ||Ay - b||$ to find the optimal solution y

for each input DSM **do**

- Obtain $DSM_{\text{r}}^{(n)}$ by applying the 3D translation $\boldsymbol{t}^{(n)}$ to all the 3D points in $DSM^{(n)}$, where $\boldsymbol{t}^{(n)}$ is the n-th row of y

4.4.5 Results

We conducted experiments using three different configurations for the proposed bundle adjustment methodology. BA-v1 corresponds to the presented bundle adjustment pipeline using only its non-optional blocks as shown in Figure 4.2. BA-v2 incorporates the two optional blocks from the bundle adjustment stage, dedicated to filtering outlier tie point observations. BA-v3 additionally activates the feature tracks selection, thus employing all blocks in Figure 4.2. The resulting evaluation metrics are shown in Table 4.2. The metrics obtained with the original unrefined RPC models and Algorithm 10, which eludes any RPC correction, are also reported.

Similarly to [AA20], for the Morenci Mine data, the images of each SkySat sensor are treated independently, i.e. the AOI is divided into three smaller AOIs corresponding to each sensor strip and we provide the average evaluation metrics across the three strips. This excludes from the evaluation all possible inconsistencies between the intrinsic parameters of the three sensors (e.g. lens distortion).

The most important ideas reflected in Table 4.2 are discussed below.

BA RPC correction results in accurate DSM registration. The bundle adjusted RPC models succeed to accurately align in the object space the DSMs reconstructed from different stereo pairs of the same acquisition date. This is reflected by the low deviation of altitude values, σ_{alt}, which reaches average values of a few tens of centimeters for both series, normally below half a meter. We obtain similar σ_{alt} for Richards Bay, regardless of the bundle adjustment configuration (BA-v1, BA-v2 or BA-v3). This seems reasonable as Richards Bay represents a not particularly challenging scenario, where the AOI is small

		N_{tracks}	N_{iter}	ρ_{init} [px]	ρ_{end} [px]	σ_{alt} [m]
Richards Bay	Unrefined RPC models	-	-	-	-	1.472
	Geometry-based DSM registration	-	-	-	-	0.304
	Basic BA RPC correction (BA-v1)	4310	17	1.327	0.133	0.291
	BA-v1 + filter after soft-ℓ_1 iter. (BA-v2)	4309	50	1.326	0.132	0.290
	BA-v2 + feature tracks select. (BA-v3)	197	35	1.697	0.167	0.295
Morenci Mine	Unrefined RPC models	-	-	-	-	2.675
	Geometry-based DSM registration	-	-	-	-	0.588
	Basic BA RPC correction (BA-v1)	76943	8	1.352	0.178	0.662
	BA-v1 + filter after soft-ℓ_1 iter. (BA-v2)	76594	94	1.291	0.129	0.434
	BA-v2 + feature tracks select. (BA-v3)	1094	56	1.829	0.206	0.370

Table 4.2: Quantitative results obtained for the two time series of SkySat imagery. Geometry-based DSM registration refers to Algorithm 10. BA-v1 corresponds to the presented bundle adjustment pipeline using all its non-optional blocks (Figure 4.2). BA-v2 incorporates the two optional blocks from the bundle adjustment stage, dedicated to filtering outlier tie point observations. Finally, BA-v3 additionally activates the feature tracks selection, thus employing all blocks in Figure 4.2.

and mostly flat. Things are different for the Morenci Mine, which covers a vast and mountainous AOI, with a more irregular relief. Thus, a wider range of altitudes and a larger amount of feature tracks and cameras needs to be handled. The σ_{alt} associated with the Morenci Mine strongly improves from BA-v1 to BA-v3, showing the benefits of the reprojection error filtering after some initial iterations (BA-v2), as well as the feature track selection (BA-v3). Using the complete pipeline (BA-v3) reduces σ_{alt} almost to the half of what is obtained with the basic blocks (BA-v1).

BA RPC correction offers higher robustness than geometry-based DSM registration. Using RPC models corrected with the complete presented pipeline (BA-v3) provided lower σ_{alt} than the equivalent experiments using geometry-based DSM registration (Algorithm 10). Again, the results were close for the Richards Bay data, but stronger differences are spotted for the Morenci Mine, whose size requires the alignment of a larger set of independent pairwise DSMs (Figure 4.11). Visual inspection reveals that this is because the registration methods exclusively based on the geometry of dense models are subject to poor results derived from the stereo reconstruction process. Algorithm 10 compensates geolocation errors derived from the camera models by assuming that the input DSM geometries are coherent, which may not always be true, e.g. due to outlier points or incomplete parts. Oppositely, our bundle adjustment methodology for RPC correction is applied before the stereo reconstruction process, in a totally independent manner, and optimizes the camera models jointly with the geometry of the scene. The fact that bundle adjustment can natively handle initial inaccurate geometry estimates is a key advantage.

The proposed pipeline may struggle in challenging scenarios for stereo matching, which is necessary to guarantee that all cameras are connected in terms of tie point observations. However, dense stereo reconstruction is also prone to failure in front of difficulties to find stereo matches. By extension, this weakness is indirectly shared with any geometry-based registration method, as is the case of Algorithm 10.

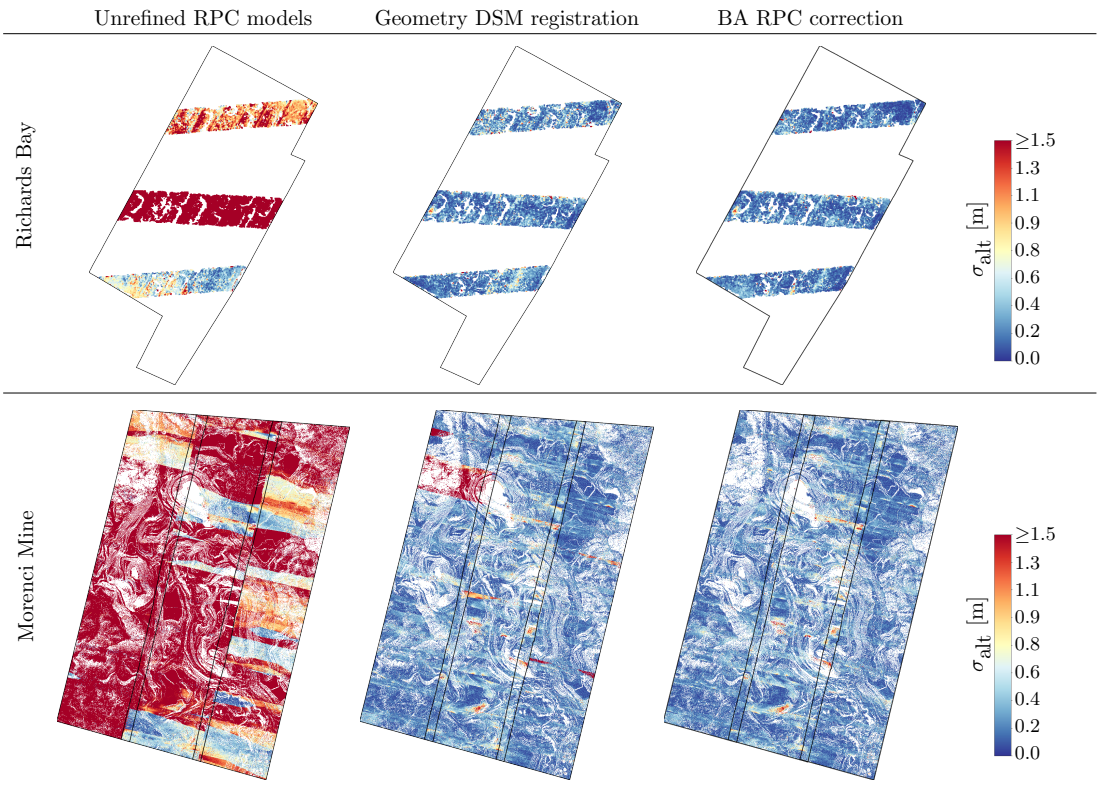


Figure 4.11: Deviation of corresponding DSM altitude values, σ_{alt}, for an acquisition date of each time series. 144 DSMs from different stereo pairs are used to cover the Morenci Mine AOI (48 DSMs for each sensor strip), while 4 pairwise DSMs are enough for the smaller AOI in Richards Bay. σ_{alt} is computed where at least 2 pairwise DSMs overlap, which only occurs in some bands for Richards Bay. Geometry registration after depth from stereo fails to align some DSMs near the top-left corner of the Morenci Mine, which is corrected when bundle adjusted RPCs are used to compute the models.

Feature track selection improves bundle adjustment convergence and accuracy. The average number of bundle adjustment iterations is reduced significantly when a suitable subset of feature tracks is selected using Algorithm 8. Observe how N_{iter} decreases from BA-v2 to BA-v3 in Table 4.2. For Richards Bay, BA-v3 attains very similar DSM registration (σ_{alt}) to the equivalent run where all tracks are used (BA-v2). For the Morenci Mine, σ_{alt} improves significantly using BA-v3. The more cameras to be registered, the more important it seems to prioritize long tracks and the use of a balanced set across all images. In Algorithm 8 we set $K_{\text{EG}} = 60$, to ensure at least 60 track observations are used per camera, if available. The reprojection error filtering step of our pipeline (Section 4.3.2) may further reduce the tie point observations per camera after feature track selection, so we chose a more generous value for K_{EG} than the one suggested in [CSH17], where $K_{\text{EG}} = 30$.

4.5 Conclusion

This chapter presented a generic pipeline to refine the RPC camera models of multi-view satellite image collections. The proposed methodology represents an improved version of the approach previously introduced in Section 3.5 of Chapter 3. However, in this

case, the approximation of RPC models as local affine cameras is avoided in favor of directly using the original RPC functions. For each input image, inaccuracies in the associated RPC functions due to inexact knowledge of the sensor orientation are corrected with a 3D rotation transformation around the camera center. The method relies on the identification of distinctive point correspondences between images, which are employed to automatically infer a set of tie points observed across the input cameras. The corrective rotation associated with each RPC model is found by running a bundle adjustment that minimizes the reprojection error of the tie points. A strategy for selecting an optimal subset of tie points was also detailed to remove redundant constraints from the bundle adjustment problem and improve its performance.

Based on the bundle adjustment solution, the presented pipeline outputs a corrected RPC camera model for each input image. The output cameras are highly consistent, in a common frame of reference, and can therefore be used for 3D multi-view reconstruction in a straightforward manner. This is validated by testing the method with multi-image SkySat L1B acquisitions over two different areas of interest and using the resulting RPC camera models in a satellite stereo pipeline to reconstruct both areas. The standard deviation between corresponding altitude values of surface models computed from different image pairs of the same acquisition date decreases from the order of a few meters to tens of centimeters, using the original and the refined RPC functions, respectively.

Finally, the presented method was compared with a different state-of-the-art solution, which eludes any RPC correction step by directly registering the geometry of independent local models. When a sufficient amount of tie points connecting all cameras is available, we obtain that bundle adjustment for RPC correction exhibits better performance. This is mainly because bundle adjustment offers robustness to initial inaccurate geometry estimates, as a consequence of simultaneously optimizing not only the camera models but also the scene geometry. Oppositely, 3D model registration methods based on geometry alone are subject to errors derived from dense stereo reconstructions.

5 Robust RPC camera modeling

The Rational Polynomial Coefficients (RPC) camera model can be used to describe a variety of image acquisition systems in remote sensing, notably optical and Synthetic-Aperture Radar (SAR) sensors. RPC functions relate 3D to 2D coordinates and vice versa, regardless of physical properties of the sensor, which has made them an essential tool to harness satellite images in a generic way. This chapter describes a terrain-independent algorithm to accurately estimate an RPC camera model from a set of 3D to 2D point correspondences based on a regularized least squares fit. The performance of the method is assessed by varying the density of point correspondences and the size of the area that they cover. We test the algorithm on Sentinel-1 SAR and WorldView-3 optical data, to derive RPCs from physical sensor models or a composition of inaccurate RPC models with complementary corrective functions. The output RPC models achieve reprojection errors on the order of 10^{-4} pixels in most configurations, which shows the ability of the algorithm to capture complex geolocation mappings in the form of RPC functions.

5.1 Introduction

The development of remote sensing applications requires geolocation tools, which relate the 3D world coordinates to the 2D image. Geolocation is usually represented by means of a *projection* function $\mathcal{P} : \mathbb{R}^3 \to \mathbb{R}^2$, that maps 3D points to the image plane, and its inverse, the *localization* function $\mathcal{L} : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^3$. When all the physical phenomena and components involved in the remote sensing acquisition process are known, the geolocation functions can be represented as a chain of operations that model these factors, in what is known as a *physical* or *rigorous* sensor model. Otherwise, the Rational Polynomial Coefficients (RPC) camera model can be adopted to describe the acquisition process of satellite images independently from the specific physical properties of the sensor. The reader is referred to Section 1.4.1 for more details.

This chapter describes a terrain-independent algorithm to fit the functions of an RPC camera model, from a set of 2D to 3D correspondences given by a different physical sensor model or another geolocation model (Figure 5.1). The main contributions are:

- An open-source implementation of the method as an easy-to-use Python package, which is available at `https://github.com/centreborelli/rpcfit`.
- An evaluation of the accuracy and robustness of the algorithm based on real scenarios. We test the method using Sentinel-1 and WorldView-3 images, to fit a Synthetic-Aperture Radar (SAR) physical sensor model or the composition of an existing RPC model with a complementary transformation.

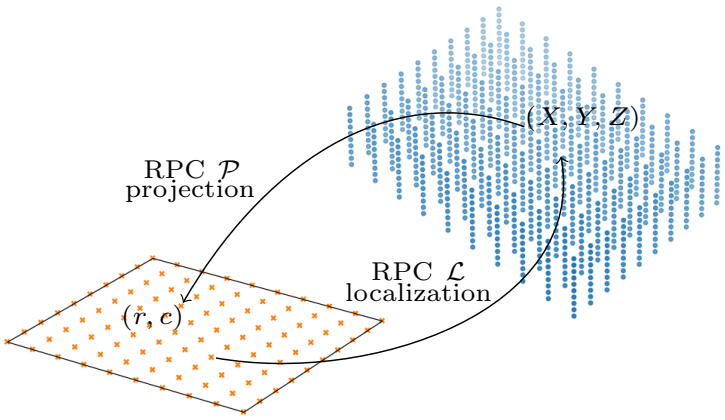


Figure 5.1: The RPC camera model coefficients are estimated using a grid of 3D CNPs (Control Points) and its projection onto the satellite image.

5.2 Related work

5.2.1 Pushbroom and SAR physical sensor models

Pushbroom scanners, previously introduced in Section 1.2 and 1.4.1, are the most common optical satellite image acquisition system [Gro01]. Physical sensor models for pushbroom scanners usually use internal parameters, which characterize the camera device (e.g., pixel size, focal length, principal point location, lens distortion coefficients); external parameters, which characterize the position and orientation of the camera with respect to the Earth (e.g., travel speed, attitude angle measurements, the position of the satellite on the orbit at the time of acquisition, orbit altitude and parameterization) and environmental parameters, such as atmospheric refraction. A detailed description of a simplified physical sensor model for pushbroom scanners can be found in [dFMLGF15].

In the case of Synthetic-Aperture Radar (SAR) images, the most used physical sensor model is the Range-Doppler model detailed in [Cur82]. SAR satellites send an electromagnetic wave that is reflected on the ground. The image is acquired line by line (similar to a pushbroom system), and the position of a ground patch in the image is related to its distance to the sensor, known as the *range*. The Range-Doppler model is constructed based on ephemeris data (time, position and velocity samples along the orbit) and the acquisition timing information. The ephemeris data needs to be interpolated to obtain continuous geolocation functions along the orbit.

5.2.2 Mathematical formulation of the RPC camera model

This section is a restatement of Section 4.2.3 from Chapter 4 and is only included here so that the rest of this chapter can be read in a self-contained manner.

As the name indicates, the RPC camera model represents the acquisition process using rational polynomial functions. Both the 3D to 2D mapping (the *projection* function) and its inverse (the *localization* function) are expressed as a ratio of cubic polynomials.

The projection function $\mathcal{P}$ of an RPC model can be expressed as

$$(r, c) = \mathcal{P}(X, Y, Z) = \left(\frac{a(X, Y, Z)}{b(X, Y, Z)}, \frac{e(X, Y, Z)}{f(X, Y, Z)} \right), \tag{5.1}$$

where a, b, e, f are cubic polynomials; X, Y, Z represent the longitude, latitude and altitude of a 3D point; and r, c are the row and column of its projection on the image plane. In practice (X, Y, Z) and (r, c) are expressed in normalized coordinates, within a range of $[-1, 1]$, for numerical stability [AMdF+21]. Normalized coordinates are obtained using two scalars, an offset and a scale factor: $X_n = (X - X_{\text{offset}})/X_{\text{scale}}$, where X_n denotes the normalized value of X, and X could be r, c, X, Y or Z in (5.1).

The RPC projection function $\mathcal{P}$ (5.1) is characterized by the cubic polynomials a, b, e, f. Each of these polynomials has 20 coefficients and is defined as

$$\begin{aligned} p(X, Y, Z) = {} & p_0 + p_1 Z + p_2 Y + p_3 X + p_4 ZY + p_5 ZX + p_6 YX + p_7 X^2 + p_8 Y^2 \\ & + p_9 Z^2 + p_{10} ZYX + p_{11} Z^2 Y + p_{12} Z^2 X + p_{13} Y^2 Z + p_{14} Y^2 X \\ & + p_{15} ZX^2 + p_{16} YX^2 + p_{17} Z^3 + p_{18} Y^3 + p_{19} X^3, \end{aligned} \tag{5.2}$$

where p_i is the i-th coefficient of polynomial p, and p can be a, b, e, or f.

5.2.3 RPC camera model fitting algorithms

RPC camera models have been widely used for high-resolution optical satellite imaging since the launch of Ikonos in 1999 [Gro01, TH01, FDG06, LJH15]. In the last decade, they have been proven to be extremely accurate for SAR acquisition systems as well [ZFL+10, ZHB+11]. Today, they are an essential tool to process satellite images in a generic way, from different sources and for multiple tasks, e.g. photogrammetry and radargrammetry based 3D reconstruction or image ortho-rectification and coregistration.

RPC fitting or estimation algorithms address the problem of finding the optimal coefficients for the polynomial functions of an RPC camera model. The RPC model of a satellite image can be constructed using a set of correspondences between image and object space coordinates. Depending on the nature of these correspondences, the literature can be classified into terrain-dependent or independent methods (or a combination of both). Terrain-dependent strategies use Ground Control Points (GCPs), whose object and image coordinates are known from manual labeling or on-site measurements. Oppositely, terrain-independent methods derive virtual sets of 2D to 3D point correspondences from other geolocation functions, usually a physical sensor model. Given the set of point correspondences, least squares algorithms are typically used to estimate the RPC coefficients that minimize the error between the projected 3D points and their image locations.

Several works have underlined the importance of using uniformly distributed points in sufficient amount, covering the different parts of the image and the whole altitude range of the scene [TH01, LJH15]. As a result, regularized least squares methods are commonly used to gain robustness to different configurations and enforce well-conditioned normal equations [TH01, LJH15, ZHB+11, WZS16]. Additionally, terrain-dependent strategies have explored the selection of optimal and balanced subsets of GCPs, e.g. [TH01] proposes a bucketing strategy and [WZS16] studies the benefits of encouraging correspondences located at building edges in urban scenarios. In contrast, terrain-independent methods can arbitrarily generate regular sets of points, but require special care to the boundaries and density of the structure, e.g. [ZHB+11] investigates the impact of different number of elevation layers for flat and mountainous areas.

5.3 RPC estimation via terrain-independent regularized least squares

This section describes the RPC fitting methodology proposed in this chapter. The objective is to fit the polynomial coefficients of the 4 cubic polynomials a, b, e, f of the RPC projection function $\mathcal{P}$, defined in (5.1). Each polynomial has 20 coefficients, where p_i is the i-th coefficient (5.2). However, since we set $p_0 = 1$ for the two denominator polynomials, a total of 78 coefficients need to be determined.

We follow a terrain-independent approach similar to that of [TH01], to fit the RPC function that emulates a different input geolocation model. The data used to fit the model consists of a 3D grid of uniformly distributed Control Points (CNPs) within some longitude, latitude and altitude boundaries (Figure 5.1). The 2D image point (r_i, c_i) corresponding to the CNP with 3D coordinates (X_i, Y_i, Z_i) can be obtained by projecting it with the input geolocation model, so that each sample results in five normalized coordinate values, i.e. $(X_i, Y_i, Z_i, r_i, c_i)$, where the subscript i denotes to the sample index.

Using the 3D and 2D coordinates of N CNPs and the RPC projection function (5.1), we follow the approach of [TH01] to build a system of equations, in matrix form, as

$$WT\boldsymbol{I} - W\boldsymbol{G} = 0, \tag{5.3}$$

where

$$\begin{aligned}
W &= \text{diag}\left[\frac{1}{b(X_1,Y_1,Z_1)}, ..., \frac{1}{b(X_N,Y_N,Z_N)}, \frac{1}{f(X_1,Y_1,Z_1)}, ..., \frac{1}{f(X_N,Y_N,Z_N)}\right] \\
T &= \text{block diag}\left[M_r, M_c\right] \quad M_r, M_c \in \mathbf{R}^{N\times 39} \\
&\quad i\text{-th row of } M_r = \left[1, Z_i, Y_i, ..., X_i^3, -r_iZ_i, -r_iY_i, ..., -r_iX_i^3\right] \\
&\quad i\text{-th row of } M_c = \left[1, Z_i, Y_i, ..., X_i^3, -c_iZ_i, -c_iY_i, ..., -c_iX_i^3\right], \\
\boldsymbol{I} &= \left[a_0, ..., a_{19}, b_1, ..., b_{19}, e_0, ..., e_{19}, f_1, ..., f_{19}\right]^T, \\
\boldsymbol{G} &= \left[r_0, ..., r_N, c_0, ..., c_N\right]^T.
\end{aligned}$$

In (5.3), W is a weight matrix with shape $2N \times 2N$, where $b(X_i, Y_i, Z_i)$ denotes the RPC denominator polynomial b evaluated with the 3D coordinates of the i-th CNP; T is the design matrix with shape $2N \times 78$; $\boldsymbol{I}$ is the solution vector with the 78 RPC polynomial coefficients necessary to determine the polynomial coefficients a, b, e and f of the projection function (5.1); and $\boldsymbol{G}$ is a vector of length $2N$ containing the CNPs image coordinates.

Equation (5.3) can be solved by least squares minimization to estimate $\boldsymbol{I}$, using the normal equation

$$T^TW^2T\boldsymbol{I} - T^TW^2\boldsymbol{G} = 0. \tag{5.4}$$

To increase numerical stability, ridge estimation regularization [WZS16, ZHB+11] is often added so that the normal equation becomes

$$(T^TW^2T + h^2E)\boldsymbol{I} - T^TW^2\boldsymbol{G} = 0, \tag{5.5}$$

where E is the identity matrix and h is a scalar controlling the regularization that is applied. To choose the best regularization factor h, the L-curve criterion was introduced in [ZHB+11]. This heuristic computes the log norm of the solution $(\log \|\boldsymbol{I}\|_h)$ versus the log norm of the residual $(\log \|WT\boldsymbol{I} - W\boldsymbol{G}\|_h)$ across different values of h that extend from the minimal to the maximal singular value of T. This curve usually has a L-shape, in

	SAR dataset	**Optical dataset**
platform	S1 A/B	WorldView-3
number of images	32	47
geographic area (lon, lat) center	Albania (18.82,41.02)	Argentina (-58.61,-34.47)
first acquisition date	2019-08-03	2014-11-15
last acquisition date	2020-02-05	2016-01-13
altitude boundaries [m]	$[-533, 2969]$	$[-513, 548]$

Table 5.1: Description of the SAR and optical datasets used in the experimental part. The altitude range of the area covered by each collection of images is defined using the minimum and maximum values from the corresponding SRTM [FRC+07] digital elevation model ±500 m to consider tall buildings or fine irregularities beyond ground level.

which the optimum corresponds to the maximum regularization parameter that achieves a small residual. The value corresponding to the corner of the curve, at the position of maximal curvature, is taken to set h automatically. For non-weighted regularized least squares (i.e. weights are set to the identity, $W = E$), the L-curve criterion is fast since the curvature can be computed with closed form expressions [HO93].

Therefore, we first set $W^{(0)} = E$ (where the superscript denotes the iteration number) and use the L-curve criterion to determine the optimal h and an initial solution $\boldsymbol{I}^{(0)}$. Then, for $i \geq 1$, $W^{(i)}$ is determined from $\boldsymbol{I}^{(i-1)}$ and is plugged in equation (5.5) to solve for $\boldsymbol{I}^{(i)}$ iteratively (the SVD least squares solver is used for stability). The iterations stop when the change in terms of RMSE between the RPC projected CNPs and their image coordinates becomes lower than a tolerance value. After convergence, some final ICCV (Iteration by Correcting Characteristic Value [ZHB+11]) iterations are computed to remove possible biases introduced by the regularization. The same stopping criterion based on the RMSE improvement is used. Each ICCV iteration k can be expressed as

$$(T^T(W^{(k)})^2T + E)\boldsymbol{I}^{(k)} = T^T(W^{(k)})^2\boldsymbol{G} + \boldsymbol{I}^{(k-1)}. \tag{5.6}$$

5.4 Evaluation with SAR and optical satellite imagery

5.4.1 Data description and use cases

This section characterizes the SAR and optical image datasets used to test the RPC fitting algorithm described in Section 5.3. Further details can be found in Table 5.1.

SAR. 32 Sentinel-1 SAR images. The data is in Interferometric Wide (IW) swath mode, each product contains 3 subswaths, and each subswath contains multiple bursts that need to be stitched together to get a continuous image. We construct the Range-Doppler physical sensor model (Section 5.2.1) and use the proposed methodology to fit an equivalent RPC camera model for each image of the dataset.

Optical. 47 WorldView-3 panchromatic images, from the 2016 *IARPA Multi-View Stereo 3D Mapping Challenge* [BKHB16]. The original RPCs of the images exhibit small inaccuracies, mainly due to inexact knowledge of the sensor orientation, which causes 3D points to project to non corresponding pixels across different images. As discussed in Chapters 3 and 4, bundle adjustment algorithms are a well-known approach to correct these RPC inaccuracies [FDG06, BAM18, MdFMLF19]. We apply a bundle adjustment similar to [BAM18] to correct the projection function $\mathcal{P}$ of each RPC into a new $\mathcal{P}_{\text{BA}}$,

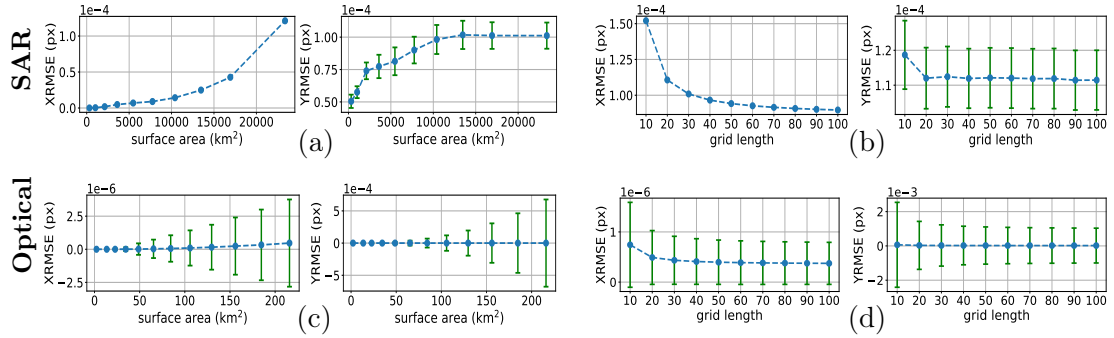


Figure 5.2: RPC fitting error varying grid length and surface area on the SAR (a-b) and optical data (c-d) described in Section 5.4.1. Each vertical bar corresponds to the $[-\sigma/2, \mu, \sigma/2]$ values of the root-mean-square error (RMSE) evaluated across the different images of the datasets, in each dimension of the image plane, where μ corresponds to the mean and σ to the standard deviation.

expressed as

$$\mathcal{P}_{\mathrm{BA}}(\boldsymbol{X}) = \mathcal{P}(R(\boldsymbol{X} - \boldsymbol{T} - \boldsymbol{C}) + \boldsymbol{C}), \tag{5.7}$$

where $\boldsymbol{X}$ is a 3D point. That is, each RPC is corrected by applying a translation $\boldsymbol{T}$ followed by a rotation R around an approximate camera center $\boldsymbol{C}$, before applying the original projection function $\mathcal{P}$. The center $\boldsymbol{C}$ is derived by regressing a projective model from each RPC model, as in Section 4.3.2. We use the RPC fitting algorithm to fit $\mathcal{P}_{\mathrm{BA}}$ from the composition of $\mathcal{P}$ with $R, \boldsymbol{T}$ and $\boldsymbol{C}$.

5.4.2 Performance assessment

For the experimental part, the algorithm is configured as follows. We set the altitude limits of the grid of CNPs using the altitude boundaries detailed in Table 5.1, a tolerance of 10^{-10} for the stopping criterion based on the RMSE improvement, and a maximum of 20 iterations for the weighted least squares and the ICCV iterations.

To evaluate the accuracy of the RPC fitting algorithm, we use a grid of Check Points (CKPs), which are located in the middle of each pair of consecutive CNPs. The RMSE between the image coordinates obtained by projecting the CKPs with the output RPC and the image coordinates obtained using the input geolocation model, measured in pixels, is used as evaluation metric.

In particular, two types of experiments were conducted to evaluate the performance and robustness of the method:

- *Varying surface area.* For each image, we fit different RPC models by gradually increasing the longitude and latitude limits of the grid of CNPs from a small square centered at the image center to a larger square covering the entire image. The number of CNPs is fixed, with 50×50 samples in the longitude and latitude dimensions and 10 elevation layers (25000 points in total).

- *Varying grid length.* For each image, we fit different RPC models, by increasing the number of CNPs in each elevation layer, i.e. $n \times n$ where n is the grid length. The longitude and latitude boundaries of the grid are fixed using the georeferenced limits of the image plane.

Overall, the results presented in Figure 5.2 show that the RPCs constructed with our method approximate the geolocation models with very high accuracy for the two datasets and the use cases outlined in Section 5.4.1. The RMSE is in the order of 10^{-4} pixels or less in both dimensions of the image plane for the majority of configurations that were tested, which emphasizes the robustness of the method.

The experiments with different grid lengths, reflected in Figure 5.2(b) and 5.2(d), show that 10 samples in the longitude and latitude directions is already a good choice and increasing this number beyond 20 does not result in significant improvements. The experiments with varying surface area, reflected in Figure 5.2(a) and 5.2(c) show that both the overall RMSE values and its variation across the different images increase with the size of the area being fitted. This is probably due to the fact that the SAR physical sensor model is less smooth for large neighborhoods. A similar behavior is obtained with the optical dataset, where the original RPCs are known to behave locally as an affine camera [FDG06, MdFMLF19], as discussed in previous chapters (Section 4.2.2).

5.5 Conclusion

This chapter described an automatic algorithm to fit the functions of an RPC camera model of a satellite image in a terrain-independent manner. The inputs of the method are a regular grid of 3D points (CNPs), distributed in multiple elevation layers, and the 2D points corresponding to their observation on the image plane. We evaluated the method on real scenarios using collections of SAR and optical satellite images, and assessed its performance by varying the configuration of input set of point correspondences. The results show that the output RPC models achieve reprojection errors on the order of 10^{-4} pixels in most configurations, which shows the ability of the algorithm to capture complex geolocation mappings in the form of RPC functions.

Part II

Applications of multi-view satellite imagery

6 Automatic stockpile volume monitoring

This chapter addresses a real use case of a satellite multi-view stereo (MVS) pipeline. We use recurrent SkySat multi-image acquisitions to monitor stockpile volume in a coal terminal. The SkySat images are first grouped according to the acquisition date, and the Rational Polynomial Coefficients (RPC) camera models of each group are corrected using the bundle adjustment method previously presented in Chapter 4. Each camera model is therefore refined by a rotation that compensates for errors due to inaccurate knowledge of the sensor orientation. The MVS pipeline is applied then using the corrected RPC models, to obtain the digital surface model (DSM) of each acquisition date. Thanks to the prior bundle adjustment, the geometry of the image pairs used at each date is registered with sub-meter accuracy. The resulting sequence of multi-view DSMs is used to measure the evolution of stockpile volume. In particular, the stockpile volume in each 3D model is obtained by subtracting a terrain model and integrating altitude values above ground level. The volume estimates are validated using measurements collected on site over the same time period. Both distributions are strongly correlated, thus highlighting the enormous potential of recurrent multi-image acquisitions with strong multi-view consistency.

6.1 Introduction

Stockpile measurement is a task of major importance in a wide variety of industrial activities involving the storage, treatment and transport of bulk materials, such as mining [RFF16] or landfill management [TGC+19]. Large-scale areas represent a particularly adverse scenario to measure vast piles of possibly irregular shapes. Primary methods, such as human-led topographic inspections, can pose a safety risk and are ineffective in terms of time and accuracy. This has raised the interest on technology-aided surveys employing ranging sensors and photogrammetry. Lidar scans and imagery from Unmanned Aerial Vehicles (UAVs), planes or robots, have proven to be highly accurate and have gained great popularity in recent years [RFF16, TGC+19, HCZH19, ZY19].

However, aerial or ground based solutions are often costly or not feasible in restricted areas, which may prevent frequent updates. Recurrent satellite imagery is emerging as an alternative, as it allows to classify, segment and reconstruct large areas without on-site actions [SMTR15, SJS+17, dFMLM+14a, dPM+20]. This trend is largely due to the growing number of small satellite constellations. Unlike conventional constellations, small satellite constellations can afford a large number of units, which are combined to obtain revisit times of a few hours. The PlanetScope and SkySat constellations are examples of the latter. The reader is referred to Section 1.3 for more details on small satellite constellations.

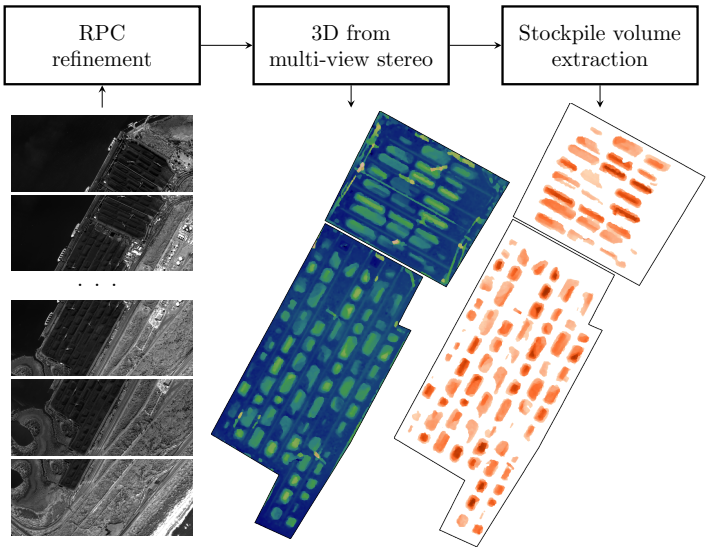


Figure 6.1: Diagram of our surface volume monitoring approach. The RPC camera models of a time series of SkySat multi-image acquisitions are refined and used to compute a highly accurate multi-view 3D model for each date, where stockpile volume can be measured.

Unlike traditional large-scale satellite imagery, SkySat image acquisitions are delivered as collections of multiple images, each covering a small geographical footprint. Individual images or frames are referred to as *scenes*. The SkySat push-frame acquisition mode is able to cover large areas of interest by ensuring a certain overlap between consecutive scenes as the satellite moves [Pla19]. Each scene is provided with a Rational Polynomial Coefficients (RPC) camera model, but the different RPCs of a push-frame acquisition typically exhibit non-negligible inaccuracies. Inconsistencies between camera models lead to systematic errors in altitude estimation from image correspondences, hindering multi-view 3D reconstruction. Bundle adjustment methods for multi-view RPC correction are indispensable to exploit such fragmented data [MdFMLF19, MdFML$^+$21], as discussed in Chapter 4. In addition, highly dynamic areas like open-air storage facilities pose an extra challenge for RPC correction: images acquired at different dates should be handled without assuming a static scene over time.

This chapter presents an automatic surface monitoring approach, outlined in Figure 6.1, to measure stockpile volume from a time series of SkySat multi-image acquisitions. The method works with stereo and tri-stereo acquisitions [Pla19]. The main contributions are:

- A generic date-wise refinement of RPC camera models, independent of satellite sensor specificities and able to handle areas of interest that change over time.
- A volume tracking strategy based on a time series of high-quality surface models obtained by multi-view stereo 3D reconstruction using the refined RPC models.
- A performance evaluation based on stock measurements acquired on site.

6.2 Related work

6.2.1 Stockpile volume monitoring

Most image based approaches to compute stockpile volume perform a 3D reconstruction of the area based on the dense matching of multiple views [RFF16, TGC$^+$19, HCZH19,

SMTR15]. Single image methods using site-specific heuristics or shape from shading have also been explored [dPM+20]. Once the 3D geometry of the scene is computed, stockpile volume can be measured in different ways. Cross-section methods model piles as big regular solids, while horizontal section methods divide them into layers following contour lines [TGC+19]. For finer estimations on irregular shapes, it is common to discretize the scene into small elementary 3D volumes (e.g., voxels, tetrahedrons, trigonal prisms [ZYF+18, ZY19]) or into Digital Surface Models (DSMs), i.e. a 2D grid where each cell is assigned an altitude value [RFF16, AM15]. The boundaries of the piles are usually obtained by subtraction of a bare terrain model, possibly combined with a segmentation step [HCZH19, TGC+19].

6.2.2 RPC camera model correction and satellite multi-view stereo

As previously introduced in 1.4.1, the RPC model is a generic camera model, independent of specific physical properties, widely used to describe the acquisition process of pushbroom scanners that capture optical satellite imagery. The RPC of a satellite image relates 3D space coordinates (latitude, longitude, altitude) to 2D image coordinates (pixel row and column). The 3D to 2D mapping and its inverse are known, respectively, as the *projection* and *localization* functions. Bundle adjustment methods [TMHF00, GD03] can be used to correct geolocation inaccuracies in the RPC functions and encourage multi-view geometry consistency in collections of satellite images that observe the same scene from different viewpoints. A review of the fundamental concepts of bundle adjustment approaches for RPC camera correction can be found in Section 3.2.1. The corrected RPC camera models provided by bundle adjustment methods are widely used in satellite multi-view stereo (MVS) reconstruction pipelines [BAM18, RDPD17, GF18], which represent the predominant approach for digital surface modeling from multiple satellite images. A review of satellite MVS pipelines can be found in Section 1.4.2.

6.3 Method description: Surface monitoring from recurrent satellite images

As illustrated in Figure 6.1, we propose a date-wise three-step methodology for stockpile volume monitoring from time series of multi-image acquisitions, such as the ones provided by small satellites with push-frame acquisition modes. First, the inaccuracies of RPC camera models are corrected using a bundle adjustment method; next, a satellite stereo pipeline is run for multiple input pairs to extract a multi-view stereo digital surface model (DSM) that covers the entire area of interest; finally, the multi-view DSM is used for stockpile volume measurement.

6.3.1 Bundle adjustment to correct camera orientation

Since we do not have access to Ground Control Points (GCPs) in the observed area, i.e. points whose object and image coordinates are known in advance, we perform a relative correction of the RPC camera models based on tie points derived from feature correspondences. The correction is done in a date-wise manner, solving an independent multi-view RPC correction problem for each multi-image acquisition. This ensures that the geometry seen by the cameras is coherent, which is not guaranteed if multiple dates are treated at once. As a result, feature mismatches are minimized and the accuracy of tie points increases.

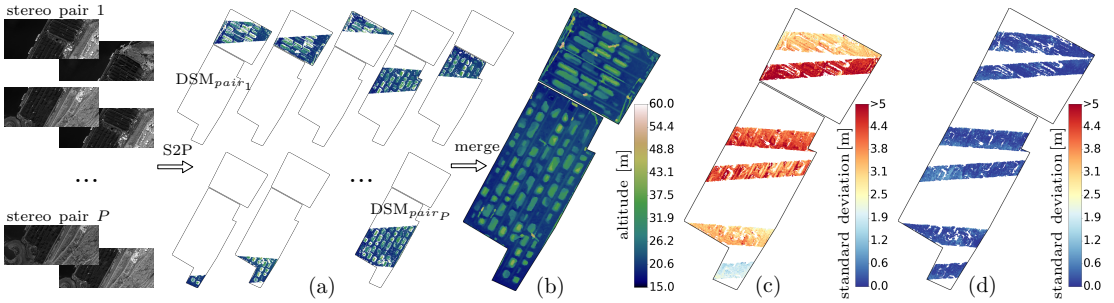


Figure 6.2: Illustration of the multi-view stereo process run to reconstruct the area of interest at each date. (a) Several DSMs of the different parts of the area are computed independently from different stereo pairs. (b) Thanks to RPC correction, the DSMs from different pairs are accurately registered and can be merged directly by taking the average height at each 2D cell. (c) and (d) show, respectively, the standard deviation, in meters, between altitude values of DSMs from different image pairs before and after RPC correction (the pairwise DSMs overlap only on the shown strips).

According to the above, for each date, the bundle adjustment methodology presented in Chapter 4 is applied to obtain the corrected RPC camera models of each date of the time series. The first part of the method is a feature tracking stage, where tie points are automatically found using SIFT keypoints [Low04], which are matched to keypoints of overlapping views with a sufficient baseline. The 3D coordinates of the tie point related to corresponding SIFT keypoints are initialized by triangulating the pairwise matches with the input RPC models, as in [dFMLM+14a], and taking the mean of the candidate 3D locations. The second part of the method solves the bundle adjustment problem, where tie point 3D coordinates and some camera correction parameters are simultaneously optimized to minimize the reprojection error of the set.

As detailed in Section 4.3.2, the proposed bundle adjustment corrects each RPC model by direct composition of the unrefined projection function $\mathcal{P}$ with a preceding rotation R in the object space, around an approximation of the camera center $\boldsymbol{C}$. Both $\mathcal{P}$ and $\boldsymbol{C}$ remain fixed during the optimization process. Each corrective rotation R is initialized as the identity matrix and characterized using the Euler angles representation [MdFMLF19].

The reprojection error is therefore expressed as

$$\min_{R_m, \boldsymbol{X}_k} \sum_{k=1}^{K} \sum_{m=1}^{M} \|\mathcal{P}_m(R_m(\boldsymbol{X}_k - \boldsymbol{C}_m) + \boldsymbol{C}_m) - \boldsymbol{x}_{mk}\|^2, \tag{6.1}$$

where $\boldsymbol{X}_k$ denotes the k-th tie point, $\boldsymbol{x}_{mk}$ its observation on the m-th image.

6.3.2 Multi-view stereo reconstruction

The corrected RPC camera models are used to run the open-source satellite stereo pipeline S2P [dFMLM+14a]. For each date, we select pairs of overlapping images using a criterion similar to [FdFML17]: the intersection angle must be between 5 and 35 degrees; and the bounding box of the triple intersection between the two image footprints and the area of interest must have both dimensions larger than 200 meters. The selected pairs are input to S2P to generate pairwise DSMs covering different parts of the area of interest, as shown in Figure 6.2(a).

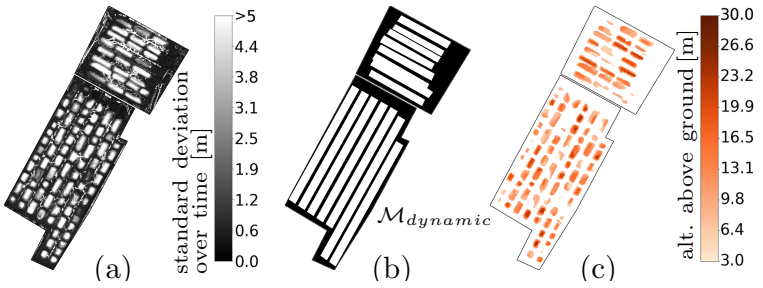


Figure 6.3: (a) Standard deviation of altitude values across the multi-view DSMs of the time series. (b) Mask of dynamic parts. (c) Example of normalized DSM or nDSM.

The altitude values of the pairwise DSMs are natively registered in the object space as a result of the previous RPC correction. This allows to easily merge them into a denser and highly accurate model, covering the entire area of interest, by taking the average altitude value at each DSM cell, as illustrated in Figure 6.2(b). Small holes in the DSMs, due to occlusions between views or lack of texture, are filled by a 5×5 median filter followed by cubic interpolation. The spatial resolution is set to 1 m per pixel for all DSMs.

Similarly to [FdFML17], the DSMs of different dates are aligned by a 3D translation that maximizes the Normalized Cross Correlation between their geometry. Even if the geometry may change between different dates, this alignment serves to minimize the standard deviation of DSM altitudes over time, which is subsequently exploited to determine a coarse mask of the dynamic parts of the area of interest (Figure 6.3).

Figure 6.2(c) and 6.2(d) highlight the impact of RPC correction: the standard deviation between altitude values of DSMs reconstructed from different image pairs is of the order of a few meters for the unrefined RPC models, while it decreases to tens of centimeters using the corrected ones.

6.3.3 Volume estimation

Time series of high-resolution DSMs from remote sensing can be effectively employed to monitor volumes changes over an area of interest. We propose an automatic system to measure stockpile volume in a real case scenario (Section 6.4.1).

For each DSM, a Digital Terrain Model (DTM) is subtracted to consider only the altitude values above ground. Since the area of interest in our case is located on a flat terrain, we model the DTM as a plane with altitude equal to the 25th percentile of DSM altitudes. Without loss of generality, cloth simulation methods can be used to model non-flat DTMs [ZQW+16]. Additionally, we determine a site-specific mask $\mathcal{M}_{\text{dynamic}}$ that delimits the dynamic parts of the area, shown in Figure 6.3(b). The labeling of $\mathcal{M}_{\text{dynamic}}$ is based on the standard deviation of altitude values over time, across the DSMs computed for each acquisition date, shown in Figure 6.3(a).

The normalized DSM (nDSM) containing the values of altitude above ground in the regions labeled as dynamic can be expressed as

$$\text{nDSM}(t) = \mathcal{M}_{\text{dynamic}}(\text{DSM}(t) \text{ - } \text{DTM}(t)), \tag{6.2}$$

where t represents the acquisition date of the time series.

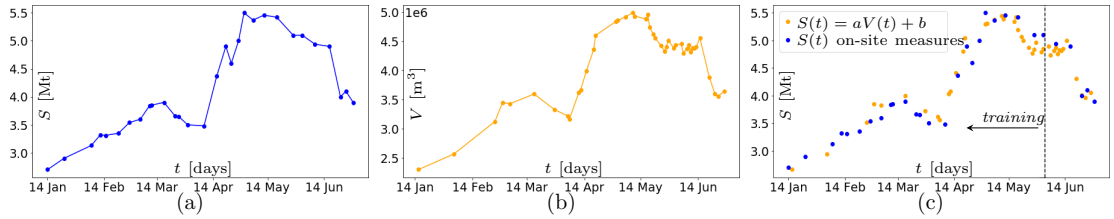


Figure 6.4: (a) On-site coal stock measurements, S, in mega tonnes (Mt). Both in this plot and (b) the line linking the set of discrete measures, represented as dots, is linearly interpolated. (b) Stockpile volume, V, as obtained from the time series of photogrammetric DSMs derived from satellite imagery. (c) Linear least squares regression to predict coal mega tonnes as a function of stockpile volume estimates.

Furthermore, only values in nDSM(t) between 3 and 30 meters are kept (Figure 6.3(c)). Values outside this range are likely to be due to noise, surface roughness or machinery and cranes working in the area. Note that both height thresholding and the use of $\mathcal{M}_{\text{dynamic}}$ are site and task specific post-processing steps aimed at ensuring that only stockpile altitude values are left in nDSM(t). In the absence of prior knowledge of the facilities or target volumes, these post-processing steps can be omitted at the cost of a possible loss of accuracy.

Finally, the stockpile volume in nDSM(t) is computed in cubic meters as the addition of all individual cell volumes:

$$V(\text{nDSM}(t)) = \sum_i l_i w_i h_i, \tag{6.3}$$

where l_i, w_i, h_i represent the length, width and height of the i-th nDSM cell. We use square cells with $l_i = w_i = 1$ m.

6.4 Experiments: Coal volume estimation using SkySat multi-image time series

6.4.1 Data

We test our method on a time series of SkySat panchromatic L1B scenes covering the Richards Bay Coal Terminal (RBCT) in South Africa, which has an open-air storage area of ~1.6 km^2. The RBCT is one of the world's leading coal export terminals. Tonnes of coal stockpiles are managed 24 hours a day to be shipped overseas.

The input time series comprises 43 acquisition dates, distributed non-uniformly between January and July 2020. The distance between consecutive dates oscillates between 1 and 20 days, falling below 1 week in most cases. Each date comprises 6 to 10 push-frame scenes, captured by the same sensor as the 3 installed in SkySats, with a difference of a few seconds. All images are cloud-free. The SkySat L1B scenes have a resolution of ~0.72 m per pixel at nadir and a total size of 1349 × 3199 pixels. Each scene is delivered with an unrefined RPC camera model, with an absolute geolocation accuracy of 30-50 m [Pla19].

6.4.2 Validation using on-site stock weight measurements

The volume V of coal stockpiles for each date t of the time series was computed using the presented method, resulting in the volume measurements shown in Figure 6.4(b). Coal

stock weight measurements, S, collected by on-site agents during the same time period, are shown in Figure 6.4(a) in mega tonnes (Mt). The correlation between both sets is recognizable to the naked eye.

To assess the performance of our system, we propose a simple approach to predict coal weight from stockpiles volume, so that the available measurements can be compared in equivalent units. The conversion is not straight-forward, as multiple date-dependent factors may be involved (e.g. humidity, non-uniform coal types or different pile densities). For this purpose, we linearly interpolate the two sets of measurements and apply a least squares regression to fit two coefficients a and b satisfying $S(t) = aV(t) + b$. The result is shown in Figure 6.4(c). We obtain $a = 1.02$, $b = 0.3$, where a can be interpreted as the best-fitting bulk density and b as a ground offset. Observe that 85% of the interpolated samples were used to fit a and b (denoted as *training* in Figure 6.4(c)), but the strong correlation extends to the rest of dates, stressing the robustness of the method.

It remains difficult to quantify the real accuracy of the system without exact time-matched measurements, especially since shipments or stock arrivals can occur within hours. In general, Figure 6.4(c) seems to indicate that the differences between our weight estimates derived from remote sensing and on-site measurements from neighboring dates are smaller than 0.3 Mt. Occasional larger differences can be due to noisy or interpolated data in the photogrammetric DSMs used for volume estimation, specific stockpile properties or a sudden strong activity in the area.

6.5 Conclusion

This chapter presented an automatic system for surface volume monitoring using a time series of multi-image acquisitions from a constellation of small satellites. The system is based on the generic RPC model correction methodology introduced in Chapter 4, which is applied to the image collections of each acquisition date in an independent manner. The corrected RPC camera models are input to an open-source satellite stereo pipeline to obtain a high-resolution multi-view DSM of the area of interest at each date of the time series, where the volume of altitudes above ground can be accurately measured by subtracting a terrain model. We used a time series of SkySat images distributed over ∼6 months to test the system on a real case scenario concerning the stockpile storage area of a coal export terminal. The volume estimates were validated using on-site weight measurements collected over the same time period. Both distributions are strongly correlated, thus highlighting the enormous potential of methods that encourage multi-view consistency in multi-image acquisitions from small satellites.

7 Perfect sensor localization for push-frame image stitching

This chapter presents a novel method to generate a single image product from a multi-image strip acquired by a push-frame satellite imaging system. The images of the push-frame strips are combined into a large-scale mosaic simulating a perfect sensor geometry, as if it had been acquired with an ideal pushbroom instrument. The Rational Polynomial Coefficients (RPC) camera models of the input push-frame images are corrected with the bundle adjustment method described in Chapter 4 and the RPC fitting algorithm detailed in Chapter 5 is used to produce the new perfect sensor localization model associated with the output mosaic. Among other applications, this simplifies the task of stereo reconstruction enormously: instead of dealing with multiple stereo pairs of small images, it is possible to reconstruct the entire area of interest covered by the push-frame acquisition using a single pair of mosaics incorporating all images. We test our method using strips of 3 and 5 SkySat L1B scenes to produce single output images, which we denote L1B$^+$. To evaluate the quality of the L1B$^+$ images and their localization models, the stereo reconstructions obtained with L1B$^+$ are compared with those obtained with L1B and with a lidar reference model. The results indicate that L1B$^+$ achieves higher altitude accuracy with respect to multi-view stereo models derived from multiple pairs of the original L1B scenes. This is mainly explained by the fact that L1B$^+$ models are not subject to registration errors between altitude values derived from different image pairs.

7.1 Introduction

In Section 1.3 we discussed the main characteristics of small satellite (or smallsat) constellations, such as SkySat from Planet [SBD10], which are emerging as an important economic alternative to conventional constellations. SkySats use a push-frame imaging mode to cover large areas of interest, beyond the limited footprint of their telescope. Push-frame systems acquire a continuous strip of small and partially overlapping images as the satellite moves [AA20, Pla22]. Despite their economic and technical appeal, the highly fragmented nature of push-frame acquisitions may discourage their use for some applications targeting areas of interest of several km^2 [XLG$^+$08].

This chapter describes an image stitching method to convert fragmented push-frame acquisitions into single large-scale images. The multiple images of a push-frame strip are combined into a geometrically consistent mosaic image, as if it had been acquired by an instrument with perfect sensor geometry (Figure 7.1). The local camera models of the input images are used to compute the localization model of the output mosaic. The resulting image and localization model can be treated as a new product, which we denote L1B$^+$.

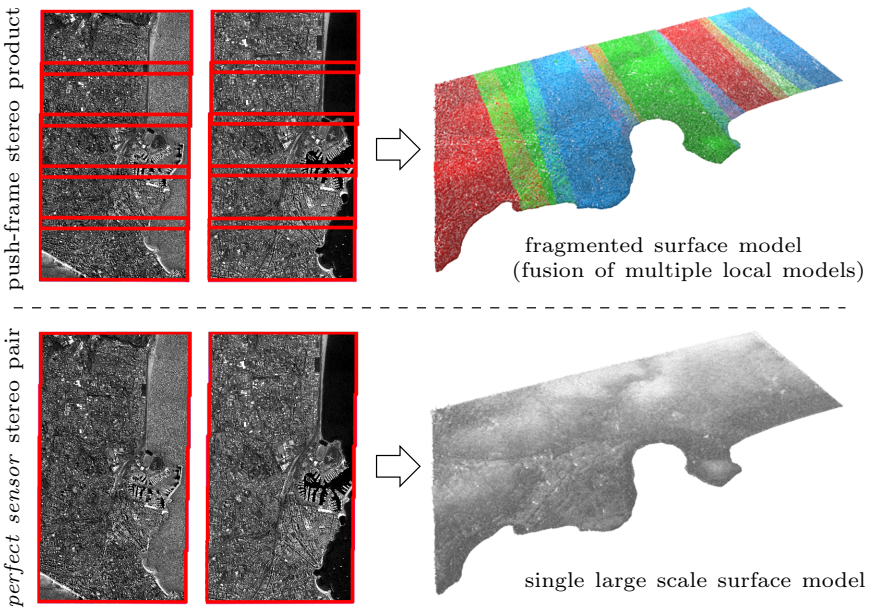


Figure 7.1: We combine the multiple images of each strip acquired by the push-frame system to produce large-scale mosaics. The resulting *perfect sensor* images and their localization models can be used for various applications. E.g., to simplify the task of large-scale stereo reconstruction from push-frame satellite imagery.

Our contributions are:

- A method capable of assembling the frames and camera models of strips of partially overlapping satellite images.
- An evaluation of the method based on an application of major interest: stereo reconstruction. We validate the method using strips of SkySat L1B scenes. The 3D models obtained with the L1B$^+$ products are compared with those obtained with the original L1B data and with a lidar reference model.

7.2 Related work

Smallsat push-frame imagery is used in a wide range of remote sensing applications, including topography extraction [AA20, BSAH21, dR21], super-resolution products [NAD$^+$21, AEdFF20] and various tasks demanding short revisit times, such as monitoring of natural phenomena [CSC21] or commercial assets [MdFMLF21, dPM$^+$20]. This chapter focuses on the task of 3D reconstruction, but the proposed method can be beneficial for any of these applications.

Topography extraction from multiple high-resolution satellite images is typically performed using stereo-based approaches [BAM18, dFMLM$^+$14a]. In Section 1.4.2 we introduced satellite multi-view stereo (MVS) pipelines and their main steps. When multiple input pairs are used, satellite MVS requires significant pre- and post-processing work. The pre-processing usually consists of a selection of suitable stereo pairs, to minimize the workload and avoid ill-posed pairs [FdFML17, GF18, MdFMLF19, GRFvGG22], and a correction of camera models associated with the satellite images, to make the multiple views geometrically consistent. The Rational Polynomial Coefficients (RPC) format introduced

in Section 1.4.1 is widely used to represent the camera models, which are usually corrected by means of bundle adjustment methodologies [TMHF00, GD03, MdFML$^+$21]. The post-processing tasks of satellite MVS revolve around the fusion step that merges the altitude estimates obtained from different pairs. The latter is usually done using local mean or median filters for simplicity, but several works have pointed out the suboptimal nature of this choice [FdFML17]. Heuristic filtering techniques [FdFML17, Qin17, QLFR22] or deep learning refinement strategies [SS20, BRK19] can be used to improve the result in sensitive areas, such as object boundaries or vegetation, and minimize errors due to artifacts or inaccurate geometric registration.

The use of push-frame imagery in satellite MVS pipelines represents a special case where the pre- and post-processing work requires special care, since the fragmented format of the acquisitions results in a large number of input pairs. Otherwise, the final 3D model may be subject to incomplete parts or altitude discontinuities. The proposed image stitching method aims to eliminate much of this pre- and post-processing work. By building large-scale mosaics from fragmented acquisitions, it allows the 3D reconstruction problem to transition from a challenging multi-pair scenario to a single pair scenario, which is the preferred input of stereo-based pipelines.

Our method characterizes the acquisition geometry of the output image mosaics using a perfect sensor geometry model, that takes inspiration from large pushbroom satellites such as in the Pléiades-HR constellation. Due to the complexity of the focal plane, the Pléiades-HR raw products should be considered as 25 sub-products with their local geometrical models [BLK09]. However, for the sake of usability, the final images emulate the geometry of an ideal pushbroom linear array. The Pléiades-HR perfect sensor geometry models are derived from the raw image, the rigorous sensor model of the satellite and a coarse elevation model of the area [BLK09]. Another existing tool that follows a similar philosophy is the `dg_mosaic` from the NASA Ames stereo pipeline [BAM18], which can be used to create a mosaic from multiple subscenes derived from the same parent Maxar pushbroom product, along with a new RPC model for it [SAM$^+$16]. In this chapter we address a more generic problem, as the input scenes do not originally belong to a common image.

7.3 L1B$^+$: A geometrically consistent mosaic methodology for RPC cameras

Given a push-frame strip S_1 of N small footprint images (also known as *scenes*), we propose a method to generate a single and equivalent large footprint image, denoted S_1^+ (i.e. the L1B$^+$ image), along with its perfect sensor localization model. We assume that the input images are radiometrically calibrated and cloud-free. As shown in Figure 7.2, the method consists of three main steps, which are detailed in the following subsections.

7.3.1 Correction of camera models

First of all, it is necessary to ensure that the local RPC camera models of all the scenes in S_1 are geometrically consistent. That is, 3D points in object space project to corresponding points in each image. Enforcing the geometric consistency will ease the subsequent mosaicing step and it will be fundamental to produce the perfect sensor localization model of the mosaic S_1^+. This is traditionally achieved by bundle adjustment in a multi-view setting. However, the baseline between consecutive cameras in a push-frame strip may

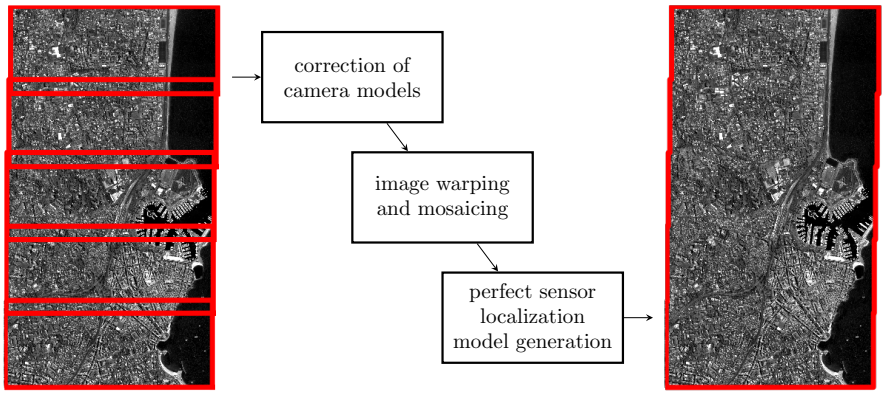


Figure 7.2: Diagram of the presented methodology.

be too small for 3D vision purposes, thus requiring some additional information. To this end, we employ a secondary strip S_2 covering the same area of interest observed by S_1 but from another point of view. SkySat stereo or tri-stereo products can provide the secondary strip S_2. Given the collection of $2{\cdot}N$ scenes in S_1 and S_2, the bundle adjustment methodology proposed in Chapter 4 is applied to perform a relative correction of their local RPC models [MdFML+21]. Each corrected RPC model results from the composition of the original projection function with a corrective rotation transformation around the estimated camera center. The corrective rotation compensates for the main source of inaccuracies, which is the inexact knowledge of the attitude angles. The reference points used by the bundle adjustment are automatically generated from correspondences of SIFT keypoints [Low04].

Note that the camera correction step does not restrict the scope of the presented method to stereo push-frame acquisitions. It is also possible to correct the camera models of S_1 without a secondary strip if a basemap with reference DSM or a set of ground control points (GCPs) are available. GCPs are points with known 3D coordinates whose position in the images is also available [GD03].

7.3.2 Image mosaicing

Once the camera models are corrected, each scene I_i in S_1 is warped into a common image space using a 2D projective transform H_i. A rough estimate of H_i is first obtained by establishing 2D point correspondences between the i-th scene and the central scene of the strip, i.e. the $\frac{N}{2}$-th scene. Instead of using classic feature matching techniques, we take advantage of the corrected RPC models computed in the previous step to initialize H_i (Section 7.3.2). Then, each H_i is refined into $\hat{H}_i$ using an image registration method [BM01, BFS18] (Section 7.3.2). Lastly, the output mosaic S_1^+ is obtained by averaging all the warped scenes as determined by the corresponding $\hat{H}_i$. A high-order Spline interpolation is used to perform the warping (we use order 5).

Image warping initialization

For each scene I_i, a regular grid of 10×10 2D points is localized in the 3D space using the i-th RPC localization function $\mathcal{L}_i$ and the average altitude of the area, h_{avg}. The average altitude h_{avg} of the area may be only an approximation, for instance a rough estimation can be obtained using the SRTM data [FRC+07]. Each grid point is then reprojected into

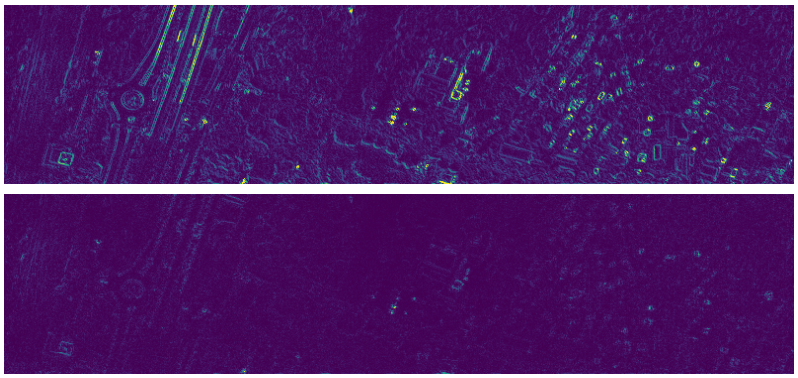


Figure 7.3: Residual difference in the overlap region after the alignment of two consecutive SkySat L1B scenes of the same push-frame strip. Top to bottom: using the initial transform H_i, using the refined transform $\hat{H}_i$. The same scaling and colormap have been used for both residuals.

the image space of the central scene of the strip, using the corresponding RPC projection function $\mathcal{P}_{\frac{N}{2}}$. The reprojection results in a set of 2D correspondences between a point $\boldsymbol{x}$ from each scene I_i and its homologous $\hat{\boldsymbol{x}}$, located in the mosaic image space, where the central scene of the strip remains in the center. Equation 7.1 summarizes the previous procedure:

$$\hat{\boldsymbol{x}} = \mathcal{P}_{\frac{N}{2}}(\mathcal{L}_i(\boldsymbol{x}, h_{\text{avg}})) \quad \text{where} \quad i \in [1, N]. \tag{7.1}$$

The correspondences $\boldsymbol{x} \leftrightarrow \hat{\boldsymbol{x}}$ are then used to fit (using a classic DLT algorithm [HZ04]) homographic transformations $H_i \in \mathbb{R}^{3\times3}$ such that $\hat{\boldsymbol{x}} = H_i\boldsymbol{x}$.

Note that in the case of the strip scenes that do not overlap with the central scene, we simply localize and reproject recursively along the neighboring frames, in the direction of the central scene, until we reach the latter.

Image warping refinement

The correspondences used to compute the transform H_i are inaccurate because the h_{avg} value used for the reprojection is not the exact altitude of the points seen in the image. However, the estimated transformations are useful to initialize an image-based registration method. The inverse compositional algorithm (ICA) is used to refine the coefficients of H_i, so that the warps of consecutive frames are precisely aligned. This assures the pixel consistency of the aligned scenes before merging them into the S_1^+ mosaic. The benefits of ICA to register push-frame satellite acquisitions has been previously studied by [AEdFF20, BFS18].

Consider two scenes I and I' of S_1 such that I' has to be aligned onto I to construct the mosaic S_1^+. Let H and H' be the initial homographies associated to I and I' respectively, which are computed as described in Section 7.3.2. Using ICA, we first refine $H' \circ H^{-1}$ such that the warped version of I' using this transformation is perfectly aligned onto I. This defines the refined homography R. Using the refined transform, we define the relative correction factor C such that $R = H' \circ C$. Figure 7.3 shows the residual difference after alignment with and without the refinement step. Observe that the refined transformation achieves a much better alignment. In this example, the RMSE without refinement is 19.37 and with refinement 5.71, thus confirming the visual result.

Algorithm 11: *Perfect sensor localization model generation*

Input : 1 strip S_1 of N partially overlapping scenes I_i,
with their local RPC_i camera models ($i \in [1, N]$).
S_1^+, the mosaic image of S_1,
and the matrices $\hat{H}_i$ that warp scene I_i onto S_1^+
Output: 1 $\text{RPC}_{S_1}^+$ camera model covering the entire S_1^+

1. Build G_{2D}, in the image space:
   Draw a regular grid of $(N{\cdot}M){\times}(3{\cdot}M)$ 2D points on S_1^+
2. Build G_{3D}, in the object space:
   Mark all the points of G_{2D} as non-visited
   for h **in** $3{\cdot}M$ uniformly spaced altitudes $\in [h_{\min}, h_{\max}]$
      for each scene I_i **in** S_1
         Localize at h the non-visited points of G_{2D} seen in
         scene I_i, using (7.2), and mark them as visited
3. Use the $(N{\cdot}M){\times}(3{\cdot}M){\times}(3{\cdot}M)$ correspondences $G_{2D} \leftrightarrow G_{3D}$ as input
   to run the RPC fitting algorithm described in Chapter 5 and obtain $\text{RPC}_{S_1}^+$

We then define by recurrence the set of refined transforms $\hat{H}_i = H_i \circ C_i$, where C_i corresponds to the composition of all necessary correcting factors from the reference image (the central scene) to the i-th image in S_1.

7.3.3 Perfect sensor geometry localization model

After completion of the mosaic S_1^+, the corrected camera models of the scenes that form the mosaic can be used to produce a perfect sensor localization model that follows the RPC standard and is valid throughout the entire S_1^+. The output camera model, denoted $\text{RPC}_{S_1}^+$, is generated by Algorithm 11.

The main idea of Algorithm 11 is to draw a regular grid of 2D points covering S_1^+, which is localized at different heights in the 3D space. By default, we use $(N \cdot M) \times (3 \cdot M)$ points, where N is the number of scenes and $M = 10$. Given a 2D point $\hat{\boldsymbol{x}} \in S_1^+$, the corresponding 3D point $\hat{\boldsymbol{X}}$ at height h is obtained as

$$\hat{\boldsymbol{X}} = \mathcal{L}_i(\hat{H}_i^{-1}\hat{\boldsymbol{x}}, h), \tag{7.2}$$

where $\hat{H}_i^{-1}$ is the inverse warping transform that transforms $\hat{\boldsymbol{x}}$ to its original small scene space, and $\mathcal{L}_i$ is the localization function of that scene. We set the range of altitudes $[h_{\min}, h_{\max}]$ by taking the maximum and minimum altitudes of the reference points used by the bundle adjustment [MdFML+21] and adding an extra margin of ± 100 meters.

Thanks to the corrected camera models, the different local RPC functions are highly consistent in the object space. This implies that the localization of the grid will result in a reasonably regular point cloud in the object space, without major discontinuities due to RPC inaccuracies.

Furthermore, when a point $\hat{\boldsymbol{x}}$ is seen in two overlapping scenes I_i and I_j, it can be localized in object space, with (7.2), using either $\mathcal{L}_i$ or $\mathcal{L}_j$. Since the scenes are registered and the RPC models adjusted, both choices will yield essentially the same 3D points for a range of altitudes centered around the surface. However, for points far from the surface we should start to observe a parallax due to the fact that the scenes are acquired from

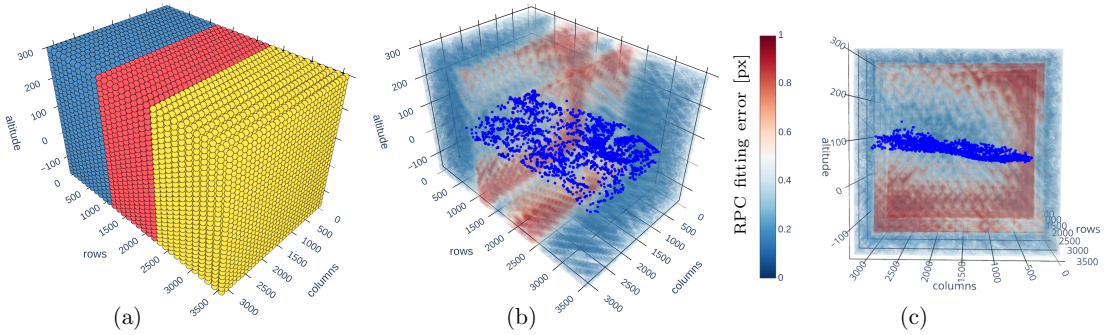


Figure 7.4: (a) 3D grid used for RPC fitting, for a strip of 3 scenes. The z coordinate is the altitude of the points in meters, while x and y correspond to their projection in the image plane. Point colors depend on the scene of the strip that was used to localize each 3D point. (b) Error of the perfect sensor localization model across the 3D grid, in pixel units. The reference points used by the bundle adjustment [MdFML$^+$21] in the prior correction of the local RPCs are shown in blue. (c) Front view of the error distribution.

different positions along the orbit. This parallax can be quantified. Assuming the SkySat parameters (baseline between views ∼1.5 km and altitude ∼500 km), in order to observe parallax of 1 pixel (assuming a resolution of 0.6 m/pixel) an elevation change of about $200\,\text{m} \approx \frac{500}{1.5} \cdot 0.6\,\text{m}$ should be present in the scene. This points to a limitation of the present method for images containing very large elevation changes. However, in our experiments (including the mountainous site of the Morenci mine, see Section 7.4) we did not observe any artifacts due to parallax.

The previous procedure results into a set of 2D-to-3D point correspondences between S_1^+ and the object space, thus the RPC fitting algorithm described in Chapter 5, can be applied to produce the final $\text{RPC}_{S_1}^+$ model [AMdF$^+$21]. Using the 2D-to-3D correspondences generated earlier with (7.2) we can define the fitting errors e associated to the $\text{RPC}_{S_1}^+$ model as the reprojection distances:

$$e = \|\hat{\boldsymbol{x}} - \mathcal{P}_{S_1^+}(\hat{\boldsymbol{X}})\|_2, \tag{7.3}$$

where e is in pixel units and $\mathcal{P}_{S_1^+}$ is the projection function of the perfect sensor localization model.

Figure 7.4(b) illustrates the usual distribution of errors e across the 3D points used to fit $\text{RPC}_{S_1}^+$. We observe that the error is small in the proximity of the surface, which can be inferred from the reference keypoints (seen as blue dots) used in the bundle adjustment (Section 7.3.1). This is reasonable as the surface points are registered in the merged product. Larger errors are observed approaching the altitude extrema of the volume, but only in bands that correspond to the overlap of two consecutive scenes. We attribute this to the fact that there is no guarantee that the RPCs of neighboring scenes are geometrically consistent away from the registered surface points. Inside the convex hull that contains all the reference points, e reaches average values ∼0.2 pixels, of the same order as the average reprojection error of the bundle adjustment (Chapter 4) [MdFML$^+$21]. Note that the 3D points used by the bundle adjustment highlight the part of the volume where the surface observed by S_1^+ is located.

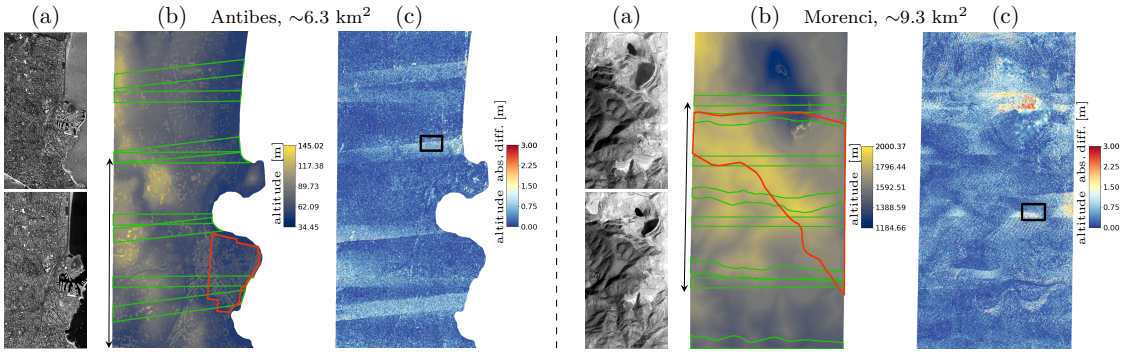


Figure 7.5: Qualitative results for SkySat stereo reconstruction, using 5 scenes from the input strips, i.e. $N = 5$. Left to right: (a) L1B$^+$ images. (b) L1B$^+$ derived DSMs. The size of the reconstructed area is indicated in square kilometers. The red boundary delimits the region for which a lidar model is available. The green boundaries delimit the regions of overlap between two local models used to derive the equivalent L1B DSM. The double-headed arrow indicates the length covered in the experiments with $N = 3$. (c) Absolute difference between the L1B$^+$ and L1B DSMs. The black rectangular outline delimits a subregion of interest inspected in Figure 7.6.

7.4 Evaluation with SkySat imagery and application to stereo pipelines

7.4.1 Data

We applied our method to two SkySat L1B stereo acquisitions, each providing two multi-image strips of the same area, with a time difference of a few seconds. One acquisition covers part of the city of Antibes (France) and the other covers the Morenci mine (United States). The two landscapes are very different: urbanized and flat terrain in the case of Antibes, as opposed to the mountainous and bare terrain of the mine.

The SkySat L1B scenes used in the experimental part have a nadir resolution of 0.58-0.86 m per pixel and a total size of 1349×3199 pixels. Each scene is delivered with an RPC camera model. The geometric accuracy of the L1B RPC models is of 30-50 m, with SkySats orbiting at altitudes of 400-500 km [Pla22]. Our proposed L1B$^+$ mosaics extend the footprint of the original L1B images and incorporate a consistent RPC model. In this chapter, we present experiments using $N = 3$ and $N = 5$ scenes per strip but the method can generalize to strips with more scenes. Using SkySat L1B scenes, we observed no deformation or misalignment in the output mosaics with strips with a number of scenes up to $N = 13$.

The SkySat acquisition platform has three staggered sensors, resulting in the push-frame system simultaneously acquiring three multi-image strips. Note that in the proposed method we assemble images from only one of the sensors at a time.

7.4.2 Stereo reconstruction based evaluation

To validate the quality of the L1B$^+$ images and their camera models, we evaluate them in the context of stereo reconstruction from two push-frame strips S_1 and S_2. For this pur-

		MAE [m]		
SkySat strip IDs	N	L1B$^+$-L1B	L1B$^+$-GT	L1B-GT
s107_20210705T131230Z s107_20210705T131300Z (Antibes)	3	0.325	1.214	1.270
	5	0.331	1.204	1.248
s4_20190127T175119Z s4_20190127T175154Z (Morenci)	3	0.391	0.781	0.854
	5	0.395	0.745	0.780

Table 7.1: Quantitative results of the stereo reconstruction based evaluation, using 3 or 5 scenes from the input strips. Left to right: MAE, in meters, between the L1B$^+$ and L1B derived DSMs, and MAE of each DSM with respect to a GT lidar model.

pose, we used the open-source satellite stereo pipeline S2P[1] [dFMLM$^+$14a], to reconstruct the areas of Antibes and the Morenci mine covered by the SkySat acquisitions, both using the original L1B scenes and the L1B$^+$ mosaics, i.e. S_1^+ and S_2^+.

As explained in Section 7.2, for the case of the L1B scenes the 3D reconstruction is a multi-view problem. We use the MVS methodology described in [MdFMLF21] to solve it. Following the selection of P suitable pairs of scenes, S2P is employed to reconstruct P independent local DSMs, using the corrected RPCs of the L1B scenes (Section 7.3.1). The P local models are lastly fused by taking the mean altitude at each cell of the DSM. In the conducted experiments, $P = 5$ for $N = 3$, while $P = 9$ for $N = 5$.

In contrast with the above, the L1B$^+$ products allow to reconstruct each area with a single execution of S2P, using as input the two perfect sensor images and their localization models.

7.4.3 Discussion

Table 7.1 lists the mean absolute error (MAE) between the DSMs obtained with the L1B$^+$ and L1B products. In addition, we computed the MAE of each model with respect to a ground-truth (GT) lidar model covering a subregion of the observed areas. Since the acquisition dates of the lidar and the SkySat images are not coincident, we manually annotated the parts of the surface models that are expected to be coherent.

Figure 7.5 shows the L1B$^+$ images for $N = 5$, the resulting L1B$^+$ DSMs and the absolute difference with respect to the equivalent L1B DSMs. In Figure 7.5(c), we can see that the absolute difference between L1B and L1B$^+$ DSMs is below 0.3 m in the majority of the surface points (the average corresponds to the L1B$^+$-L1B column of Table 7.1). However, there are parts of the area where this difference increases and approaches values close to 1 m. These traces are a consequence of the fusion of local models that is needed to generate the L1B DSM. In fact, the traces coincide in great measure with the green boundaries in Figure 7.5(b), which indicate the areas of overlap between the local models used to produce the L1B DSM. The local models do not match perfectly, because the correction of the camera models registers their geometries with an accuracy < 1 m, but a residual remains [MdFML$^+$21]. Consequently, the average altitude retained by the fusion process is subject to a certain degree of bias, especially in these overlapping zones. In the case of Antibes, the areas showing the largest differences follow a pattern of horizontal stripes, because the local geometries consist of overlapping planes (flat terrain), which

[1] https://github.com/centreborelli/s2p

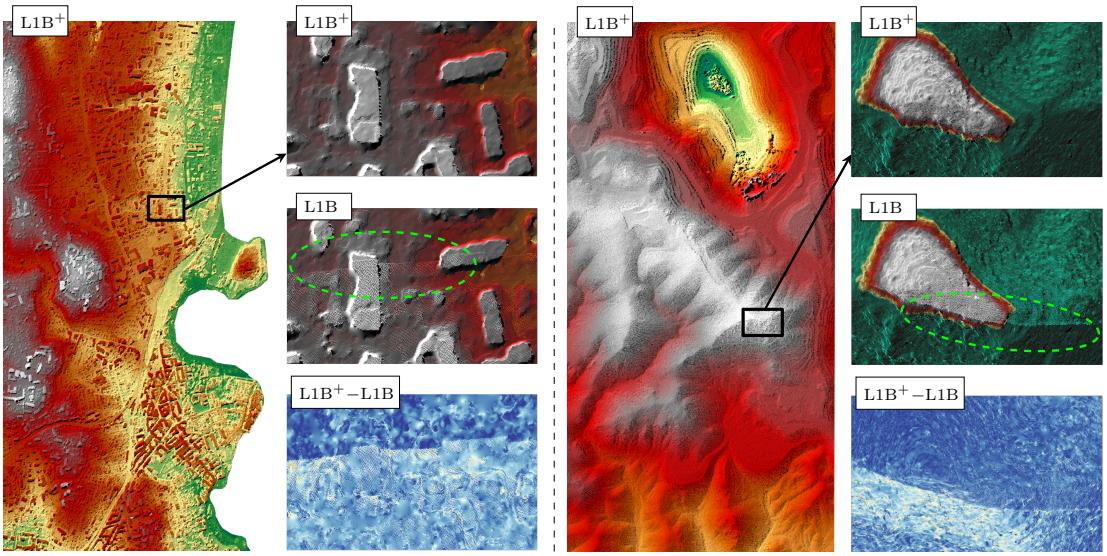


Figure 7.6: Detailed view of the L1B$^+$ DSMs and a subregion of interest. The inspection of the subregions shows that, in absence of outliers, the largest differences between L1B$^+$ and L1B DSMs coincide with small altitude discontinuities (circled in green) in the L1B model. The colormap is different in the subregion images with respect to the complete DSM to improve the contrast between local altitude values. The L1B$^+$−L1B images represent the absolute difference between altitudes, with the same colormap of Figure 7.5(c).

are stacked along the vertical axis. In the case of the Morenci mine the pattern is more irregular, with curves caused by non-exactly coincident mountain shapes and peaks due to the presence of outliers near the open pit (close to the upper left corner).

In Figure 7.6, we selected two subregions where the altitude differences between the L1B$^+$ and the L1B DSMs exhibit a strong increase. In accordance with the above observations, we can see that such differences are indeed caused by biases in the L1B DSM. By using L1B$^+$ products we eliminate the cause of such biases, i.e. the need to register and merge any local models, so the discontinuities disappear in a natural way.

The last two columns of Table 7.1 indicate that the L1B$^+$ DSMs improve the accuracy of the L1B ones, as they exhibit smaller differences with respect to the lidar. The MAE values obtained with L1B$^+$ are quite stable too, regardless of whether 3 or 5 images per strip are used. We attribute to the aforementioned misalignment between local geometries the fact that the MAE of L1B DSMs with respect to lidar tends to be slightly larger and more irregular.

Lastly, Table 7.1 shows that the MAE with respect to the lidar is higher for Antibes. This last observation is mainly explained by the presence of vegetation and the edges of the buildings, which are not as sharp in the photogrammetric DSM. Both vegetation and buildings are absent in the Morenci landscape.

7.5 Conclusion

We have presented an image stitching method to generate large-scale images from fragmented push-frame satellite acquisitions. A perfect sensor localization model is generated for the output mosaic images, using the RPC fitting algorithm of Chapter 5. This perfect

sensor geometry achieves geometric consistency between the image space and the 3D world thanks to a previous geometric correction of the input camera models using the bundle adjustment methodology proposed in Chapter 4. We denote the resulting product L1B$^+$. The method is validated using SkySat stereo acquisitions of 3 to 5 scenes.

The use of L1B$^+$ offers several advantages over SkySat L1B scenes. In this chapter we focused on the benefits for 3D reconstruction, which becomes significantly faster and simpler. Our method can be used to circumvent the need to handle multiple stereo pairs and the subsequent fusion of pairwise altitude models, a common drawback in 3D reconstruction from push-frame imagery. The L1B$^+$ products make it possible to reconstruct areas of interest of several km^2 with a single execution of a satellite stereo reconstruction pipeline. We also notice accuracy improvements in the L1B$^+$ derived DSMs, mainly due to the disappearance of any biases caused by the fusion of local models that is necessary with the original L1B scenes.

Future work could extend this methodology to handle parallax effects in areas of large elevation changes; or to assemble images from all three SkySat sensors at once, combining scenes from three multi-image strips instead of one. The generalizability of the method should also be investigated using push-frame acquisitions from other satellites.

8 Satellite NeRF

This chapter presents an end-to-end self-supervised learning multi-view approach as an alternative to satellite multi-view stereo pipelines. We introduce the Satellite NeRF network, or Sat-NeRF, for learning multi-view photogrammetry from multi-date satellite images. Sat-NeRF combines some of the latest trends in neural rendering with native satellite camera models, represented by Rational Polynomial Coefficients (RPC) functions. The proposed model is based on the original NeRF method for learning neural radiance fields, i.e. a continuous representation of geometry and appearance of a 3D scene. The neural radiance field learned by Sat-NeRF can be used to render new views and infer digital surface models of similar quality to those obtained with stereo-based pipelines. The input multi-date images may exhibit significant changes in appearance (e.g., due to varying shadows and transient objects such as cars and vegetation). Robustness to these challenges is achieved by a shadow-aware irradiance model and an uncertainty weighting of the loss function that minimizes the undesired effects of transient phenomena unrelated to the illumination. We train and evaluate Sat-NeRF on different locations using multi-date collections of 10-20 WorldView-3 images and stress the advantages of applying a bundle adjustment to the satellite camera models prior to training. This boosts the network performance and can optionally be used to extract additional cues for depth supervision. Our results indicate that Sat-NeRF achieves greater detail in surface modeling than conventional multi-view stereo pipelines, but is subject to the availability of multiple views and the presence of local irregularities in the learned geometry.

8.1 Introduction

As introduced in Section 1.1, high-resolution satellite imagery is a valuable resource for multiple economic activities, many of them based on knowledge of the geometry of the Earth's surface and its changes. This has triggered the development of a number of multi-view stereo pipelines capable of estimating altitude models with high accuracy from multiple satellite views [dK12,dFMLM+14a,FdFML17,RPDD18,GF19,SAM+16,BAM18]. The output large-scale 3D models are usually represented using discrete point clouds or digital surface models (DSMs) of a certain resolution.

The latest works in 3D modeling from multiple views show that it is possible to achieve a superior representation of a 3D object or scene by *learning* it as a continuous function or field $\mathcal{F}$ [TFT+20]. Neural rendering methods learn $\mathcal{F}$ by integrating differentiable rendering techniques into a neural network of fully-connected layers (often referred to as a multi-layer perceptron or MLP). The tasks of novel view synthesis and 3D reconstruction are then solved implicitly, as the network is trained to figure out which geometry and color radiance fit the camera projection mappings of a collection of input views.

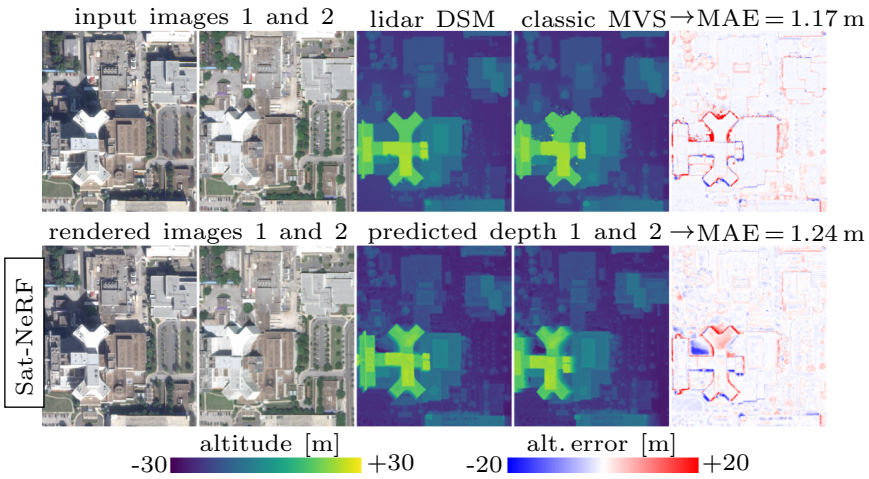


Figure 8.1: Images 1 and 2 exhibit color inconsistencies (e.g. shadows, cars, vegetation), hindering the direct use of NeRF. Sat-NeRF overcomes these problems and learns to render realistic views and underlying geometry. The digital surface model (DSM) derived from the network depth predictions is compared with a lidar equivalent, obtaining a mean absolute altitude error (MAE) similar to that of classic satellite multi-view stereo (MVS).

Neural radiance fields (NeRFs) have gained great popularity in the field of neural rendering [MST+20, TFT+20]. In this chapter, we introduce a NeRF variant architecture that achieves state-of-the-art results in novel view synthesis and 3D reconstruction from high-resolution satellite imagery in the wild. We refer to our variant as Satellite NeRF or Sat-NeRF. The original NeRF approach is not adapted to satellite images, e.g. because of the specificities of the camera models, the large distance between the cameras and the scene or the appearance inconsistencies of multi-date collections [MBRS+21, DI21] Sat-NeRF addresses these challenges using some of the latest advances in neural rendering [MBRS+21, DI21, DLZR22] and adapting well-known tools for satellite image processing. As a result, the model learns highly accurate 3D geometry, similar to that obtained with classic satellite multi-view stereo pipelines, as shown in Figure 8.1.

The main contributions of this chapter are:

- A NeRF variant that combines existing state-of-the-art methods to adapt to the satellite context. It is robust to the radiometric inconsistencies of multi-date satellite imagery, comprising shadows caused by a single non-static light source (the sun) and small transient objects (mainly trees or cars in open-air parkings). The code and data used are available at `https://github.com/centreborelli/satnerf`.
- A point sampling strategy adapted to satellite camera models. The Rational Polynomial Coefficients (RPC) camera model [Gro01, FDG06] of each input image is directly used to cast rays in the object space and query the network at different 3D points. This RPC-based strategy provides independence to the satellite acquisition system and improves the results obtained with approximate pinhole cameras.
- A study of the advantages of correcting multi-view RPC inconsistencies before training, e.g. by means of the bundle adjustment presented in Chapter 4 [MdFML+21]. We show that eluding this step leads to a drop in the performance of the model. In addition, we detail how to reuse the sparse point cloud employed in the bundle adjustment to improve geometry learning.

We evaluate Sat-NeRF on different areas of interest covering 256×256 m each, using about 10-20 RGB crops from multi-date WorldView-3 images for training [BFC+19, LSYH+19]. A lidar digital surface model of resolution 0.5 m per pixel is used as ground truth to assess the learned geometry. Sat-NeRF is compared to other NeRF variants [MST+20, DI21] as well as a state-of-the-art classic satellite stereo pipeline [dFMLM+14a].

8.2 Related work

Current state-of-the-art 3D reconstruction pipelines for satellite images typically follow multi-view stereo approaches, which can outperform sophisticated true multi-view software [ODM+15, GRFvGG22]. A review of satellite multi-view stereo pipelines can be found in Section 1.4.2. Due to the complexity of the task, these pipelines can still be improved in a number of aspects. Some of the most important limitations are:

- The 3D reconstruction usually follows the estimation of a dense disparity map using matching strategies derived from the Semi-Global Matching (SGM) algorithm [Hir07, dK12, dFMLM+14a, GF19, BAM18]. Therefore, human-crafted features and cost functions are at the core of the methodology.
- The selection of suitable stereo pairs to estimate disparity is another major challenge. Criteria based on image metadata (e.g. the acquisition dates or incidence angles) have proven to be useful, but do not guarantee the best choice [FdFML17, GF19, HWG+20].
- The lack of consensus on how the geometry derived from multiple stereo pairs should be refined or aggregated. Local point-wise operations are common to merge altitude values derived from different pairs, e.g. median [dK12, GF19, MdFMLF19] or k-medians [FdFML17]. However, recent work has shown that deep learning approaches can greatly improve the result, e.g. by exploiting geometry priors related to urban areas [BdKR18, BRK19, LBK20, SS20].
- Very often, it is necessary to make adjustments or parameter tuning to handle different sources or types of satellite images [ODM+15, ZSS19].

Neural rendering represents an opportunity to find a natural solution to the previous issues, as it allows to automatically learn the optimal features and operations adapted to each individual 3D scene. The main advantage of traditional pipelines is preserved as no explicit geometry supervision is required: the learning is self-supervised and based solely on the color of the input images. This is a key difference with respect to other state-of-the-art deep learning methods for DSM generation from satellite imagery [BRK19, SS20, GLJ21, GRFvGG22], which depend on the availability of ground truth geometry models.

8.2.1 Neural radiance fields (NeRF)

NeRF [MST+20] represents a static scene as a continuous volumetric function $\mathcal{F}$, encoded by a fully-connected neural network. $\mathcal{F}$ predicts the emitted RGB color $\mathbf{c} = (r, g, b)$ and a non-negative scalar volume density σ at a 3D point $\mathbf{x} = (x, y, z)$ of the scene seen from a viewing direction $\mathbf{d} = (d_x, d_y, d_z)$, i.e.

$$\mathcal{F} : (\mathbf{x}, \mathbf{d}) \mapsto (\mathbf{c}, \sigma). \tag{8.1}$$

Multi-view consistency is encouraged by restricting the network to predict the volume density σ based only on the spatial coordinates $\mathbf{x}$, while allowing the color $\mathbf{c}$ to be predicted

as a function of both $\mathbf{x}$ and the viewing direction $\mathbf{d}$. The dependency of $\mathbf{c}$ on the viewing direction allows to recreate specular reflections caused by to static light sources.

Given a set of input views and their camera poses, the training strategy is based on rendering the color of individual rays traced across the scene and projected onto the known pixels. Individual rays are chosen randomly, encouraging gradient flow at those ray intersections where the surface of the scene is susceptible of being located. Each ray $\mathbf{r}$ is defined by a point of origin $\mathbf{o}$ and a direction vector $\mathbf{d}$. The color $\mathbf{c}(\mathbf{r})$ of a ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ is computed as

$$\mathbf{c}(\mathbf{r}) = \sum_{i=1}^{N} T_i \alpha_i \mathbf{c}_i. \tag{8.2}$$

The rendered color $\mathbf{c}(\mathbf{r})$ results from the weighted integration of the colors $\mathbf{c}_i$ predicted at different points of the ray $\mathbf{r}$, which is discretized into N 3D points $\mathbf{x}_i$ between the near and far boundaries of the scene, t_n and t_f. Each point $\mathbf{x}_i$ in $\mathbf{r}$ is obtained as $\mathbf{x}_i = \mathbf{o} + t_i\mathbf{d}$, where $t_i \in [t_n, t_f]$.

Following (8.2), the weight given to the color predicted for each point $\mathbf{x}_i$ of $\mathbf{r}$ is defined by a transmittance factor T_i representing the probability that light reaches the point without hitting any other particle, and an alpha compositing value α_i encoding the opacity. Both T_i and α_i are set according to the volume density σ_i predicted for $\mathbf{x}_i$:

$$\alpha_i = 1 - \exp(-\sigma_i \delta_i); \quad T_i = \prod_{j=1}^{i-1} (1 - \alpha_j), \tag{8.3}$$

where δ_i is the distance between two consecutive points along the ray, i.e. $\delta_i = t_{i+1} - t_i$. Higher values of σ_i will result in larger opacity α_i, indicating that $\mathbf{x}_i$ possibly belongs to a non-transparent surface. Occlusions are handled by the transmittance T_i, equal to the cumulative product of the inverse opacity. Even if a point $\mathbf{x}_i$ is given a large opacity σ_i, the transmittance T_i only allows it to contribute decisively to the rendered color if it is not preceded by previous opaque points in the ray.

Given (8.3), the depth $d(\mathbf{r})$ observed in the direction of a ray $\mathbf{r}$ can be rendered in a similar manner to the color (8.2) [DLZR22, RBM+22] as

$$d(\mathbf{r}) = \sum_{i=1}^{N} T_i \alpha_i t_i. \tag{8.4}$$

NeRF is optimized by minimizing the mean squared error (MSE) between the rendered color and the real color of the input images, at the positions where the rays project:

$$\sum_{\mathbf{r} \in \mathcal{R}} \|\mathbf{c}(\mathbf{r}) - \mathbf{c}_{\mathrm{GT}}(\mathbf{r})\|_2^2, \tag{8.5}$$

where $\mathbf{c}_{\mathrm{GT}}(\mathbf{r})$ is the observed color of the pixel intersected by the ray $\mathbf{r}$, and $\mathbf{c}(\mathbf{r})$ is the color predicted by the NeRF using (8.2). $\mathcal{R}$ is the set of rays in each input batch.

8.2.2 NeRF variants

NeRF assumes that the density, radiance and illumination of the target 3D scene is constant. This is a strong limitation, as these conditions are rarely encountered outside laboratory settings. Many variants have been proposed to address this problem. In this section we briefly review three models that inspired our Sat-NeRF model.

NeRF-W [MBRS+21] or *NeRF in the Wild* gains robustness to radiometric variation and transient objects by learning to separate transient phenomena from the static scene. An extra head of fully-connected layers is used to predict a transient color $\mathbf{c}^\tau$ and volume density σ^τ for each input point, in addition to the usual $\mathbf{c}$ and σ. The transient outputs are linearly combined with the static ones to render the color of each ray. NeRF-W also uses the transient head to emit an uncertainty coefficient β, which measures the confidence of the network that a point belongs to a transient object. The value of β is used in the loss function to reduce the impact of transient/unreliable pixels in the learning process.

S-NeRF [DI21] or *Shadow NeRF* is, to the best of our knowledge, the first attempt to apply NeRF for multi-view satellite photogrammetry. S-NeRF showed the benefits in geometry estimation of simultaneously exploiting the direction of solar rays to learn the amount of sunlight s that reaches each point $\mathbf{x}$ of the scene. The direction of solar rays is a common metadata of satellite images. Sat-NeRF can be seen as an extension of S-NeRF that incorporates a modeling of transient objects similar to [MBRS+21] and a representation of the camera models more adapted to satellite data.

DS-NeRF [DLZR22] or *Depth Supervised NeRF* incorporates a depth supervision term to the loss function to accelerate the learning and reduce the amount of input images. The depth supervision term exploits a sparse set of 3D points that belong to the surface of the scene, which can be easily retrieved using structure-from-motion (SfM) pipelines [SSS06, SF16, MMPM17]. SfM is a common pre-processing step in NeRF frameworks, as it can estimate the camera poses needed to cast the input rays. A similar strategy to DS-NeRF is used in [RBM+22], which converts the sparse point clouds into dense depth priors.

Other recent NeRF variants are yet to be investigated in the context of satellite imagery. Some works are focused on achieving smoother scene representations or reducing the number of input views: e.g. DietNeRF [JTA21] introduces an auxiliary semantic loss to maximize similarity between high-level features instead of RGB colors; Mip-NeRF [BMT+21] prevents blurring and aliasing in collections of images with different resolutions; PixelNeRF [YYTK21] describes a framework that is trained across multiple scenes and learns priors that can generalize to unseen scenes with few available images. Recent undergoing research is also progressing to extend NeRFs to dynamic scenes [PSB+21, PSH+21, PCPMMN21, XAS21], to gain efficiency and reduce the training time [YLT+21, HSM+21, YZP+22, MESK22] or to handle complex illumination settings, under arbitrary, multiple light sources [BXS+20, SDZ+21] or near-darkness conditions [MHMB+22].

8.3 Sat-NeRF: Self-supervised neural rendering for multi-date satellite imagery

Sat-NeRF represents the scene as a static surface with an albedo color, i.e. the intrinsic color of static objects. The model learns to predict the geometry and the albedo color simultaneously with a set of additional outputs, which seek to explain the transient phenomena observed in the input images without inducing changes in the scene geometry.

We train the model following the ray casting strategy of NeRF (Section 8.2.1). Unlike the original NeRF (8.1), we assume a Lambertian surface and omit the color dependence on viewing angles. The inputs of Sat-NeRF are

- $\mathbf{x}$: 3-valued vector with the spatial coordinates of points located in the scene volume. Each point $\mathbf{x}$ is part of a ray $\mathbf{r}$.
- $\boldsymbol{\omega}$: 3-valued direction vector encoding the direction of solar rays. For each input image, $\boldsymbol{\omega}$ is extracted from the azimuth and elevation angles (θ, ϕ) that indicate the position of the sun in the satellite image metadata.
- $\mathbf{t}_j$: $N^{(\mathrm{t})}$-valued embedding vector, learned as a function of the image index j. The objective of $\mathbf{t}_j$ is to featurize the transient elements in the j-th view that cannot be explained by the position of the sun given by $\boldsymbol{\omega}$. We manually set $N^{(\mathrm{t})} = 4$.

The volumetric function of Sat-NeRF then writes $\mathcal{F} : (\mathbf{x}, \boldsymbol{\omega}, \mathbf{t}_j) \mapsto (\sigma, \mathbf{c}_\mathrm{a}, s, \mathbf{a}, \beta)$, where the outputs are

- σ: scalar encoding the volume density at location $\mathbf{x}$.
- $\mathbf{c}_\mathrm{a}$: albedo RGB color, which depends exclusively on the geometry, i.e. the spatial coordinates $\mathbf{x}$.
- s: shadow-aware shading scalar, learned as a function of $\mathbf{x}$ and the solar rays direction vector $\boldsymbol{\omega}$.
- $\mathbf{a}$: ambient RGB color, independent of scene geometry, that defines a global hue bias according to the position of the sun given by $\boldsymbol{\omega}$.
- β: uncertainty coefficient related to the probability that the color of $\mathbf{x}$ is explained by a transient object.

8.3.1 Shadow-aware irradiance model

This section describes how Sat-NeRF predicts the color $\mathbf{c}(\mathbf{r})$ of a ray $\mathbf{r}$ projected onto a certain pixel. We keep the rendering as in (8.2) and (8.4), with the transmittance and opacity factors as defined in (8.3), but adopt the shadow-aware irradiance model proposed in S-NeRF [DI21] to compute the color $\mathbf{c}$ at each point $\mathbf{x}$ of a ray $\mathbf{r}$:

$$\mathbf{c}(\mathbf{x}, \boldsymbol{\omega}, \mathbf{t}_j) = \mathbf{c}_\mathrm{a}(\mathbf{x}) \cdot \big(s(\mathbf{x}, \boldsymbol{\omega}) + (1 - s(\mathbf{x}, \boldsymbol{\omega})) \cdot \mathbf{a}(\boldsymbol{\omega})\big), \tag{8.6}$$

where $\mathbf{c}(\mathbf{x}, \boldsymbol{\omega}, \mathbf{t}_j)$ substitutes $\mathbf{c}_i$ in (8.2). The shading scalar $s(\mathbf{x}, \boldsymbol{\omega})$ takes values between 0 and 1 and is used to add shadows by darkening the albedo (Figure 8.2). Ideally, $s \approx 1$ in those 3D points directly illuminated by the sun, whose color should be entirely explained by the albedo $\mathbf{c}_\mathrm{a}(\mathbf{x})$.

In addition, (8.6) attempts to capture the bluish hues of shadows [AGA08, MQS08] by means of the ambient color $\mathbf{a}(\boldsymbol{\omega})$, which contributes to the points where s takes values closer to 0. In practice, we find that the direction of the solar rays $\boldsymbol{\omega}$ is narrowly related to the acquisition date (especially if the satellite passes at the same hours of the day), as shown in Figure 8.2. Thus, $\mathbf{a}(\boldsymbol{\omega})$ ends up capturing ambient irradiance due to a mixture of phenomena, which is related to $\boldsymbol{\omega}$ but also date-specific conditions like weather or seasonal changes.

As observed in S-NeRF [DI21], the shading scalar $s(\mathbf{x}, \boldsymbol{\omega})$ in (8.6) can produce unrealistic results for solar rays directions that are not seen in the training data. This can be minimized by adding a *solar correction* term to the loss:

$$L_\mathrm{SC}(\mathcal{R}_\mathrm{SC}) = \sum_{\mathbf{r} \in \mathcal{R}_\mathrm{SC}} \left(\sum_{i=1}^{N_\mathrm{SC}} (T_i - s_i)^2 + 1 - \sum_{i=1}^{N_\mathrm{SC}} T_i \alpha_i s_i \right), \tag{8.7}$$

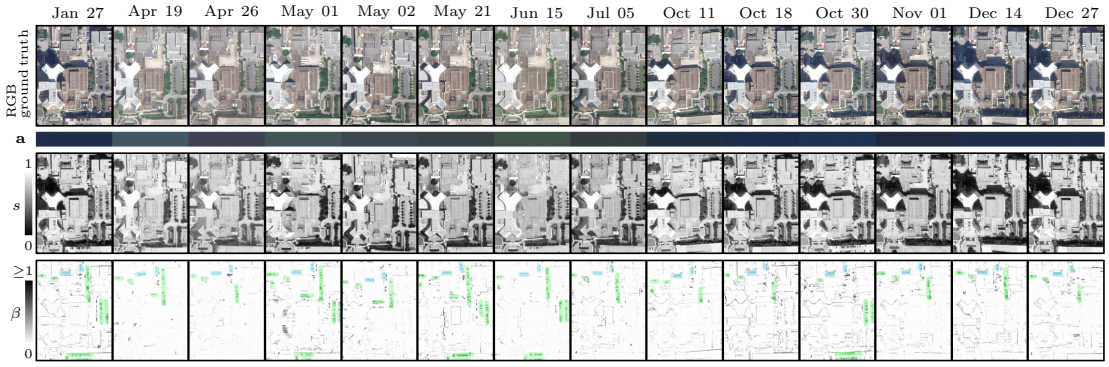


Figure 8.2: The shading scalar s related to the solar rays direction $\boldsymbol{\omega}$ learns shadows and material roughness. We observe that $\boldsymbol{\omega}$ is narrowly related to the acquisition date, causing the ambient color $\mathbf{a}$ associated with the low values of s (see (8.6)) to capture a mixture of phenomena, including seasonal changes reflected in the vegetation. The uncertainty prediction β does not affect shadows and concentrates on small color inconsistencies, mostly caused by cars in open-air parkings (green marks), large fans in rooftops (blue marks) or building edges.

where $\mathcal{R}_{\text{SC}}$ is a secondary batch of solar correction rays. Note that the rays in $\mathcal{R}_{\text{SC}}$ follow the direction of solar rays $\boldsymbol{\omega}$, while the rays in $\mathcal{R}$, in the main term of the loss (8.5), follow the viewing direction of the camera.

The solar correction term (8.7) uses the learned geometry, encoded by the transmittance T_i and opacity α_i (8.3), to further supervise the learning of the shadow-aware shading $s(\mathbf{x}, \boldsymbol{\omega})$. The first part of (8.7) enforces that, for each ray $\mathbf{r}$ in $\mathcal{R}_{\text{SC}}$, the s_i predicted at the i-th point should resemble T_i, i.e. high values before reaching the visible surface, low values afterwards (both s_i and T_i take values between 0 and 1). The second part of (8.7) encourages that the integration of s over $\mathbf{r}$ reaches 1, since non-occluded and non-shadow areas have to be mostly explained by the albedo in (8.6).

8.3.2 Uncertainty weighting for transient objects

Similarly to NeRF-W [MBRS+21], we use the task-uncertainty learning approach introduced in [KGC18] to gain robustness to transient objects by means of β. In our context, transient objects are punctual local changes across the input images that cannot be explained by the static surface or the available metadata, like the position of the sun. The irradiance model (8.6) does not handle transient objects explicitly. As a result, we observe that s and σ usually try to account for them, leading to wrong depth predictions, as shown in Figure 8.4. Thanks to β, Sat-NeRF is given some margin to ignore the color inconsistencies caused by these objects.

The uncertainty prediction β weights the contribution of each ray to the MSE between rendered and known colors:

$$L_{\text{RGB}}(\mathcal{R}) = \sum_{\mathbf{r}\in\mathcal{R}} \frac{\|\mathbf{c}(\mathbf{r}) - \mathbf{c}_{\text{GT}}(\mathbf{r})\|_2^2}{2\beta'(\mathbf{r})^2} + \left(\frac{\log \beta'(\mathbf{r}) + \eta}{2}\right), \tag{8.8}$$

where $\beta'(\mathbf{r}) = \beta(\mathbf{r}) + \beta_{\text{min}}$. In (8.8), we use $\beta_{\text{min}} = 0.05$ and $\eta = 3$ to avoid negative values in the logarithm. The role of the logarithm in L_{RGB} is to prevent β from converging to

infinity to solve the problem. In this way the model is forced to find a compromise between the uncertainty coefficients β and the differences of colors.

The $\beta(\mathbf{r})$ associated with a ray $\mathbf{r}$ is obtained by integrating the uncertainty predictions across the N points of $\mathbf{r}$:

$$\beta(\mathbf{r}) = \sum_{i=1}^{N} T_i \alpha_i \beta(\mathbf{x}_i, \mathbf{t}_j), \tag{8.9}$$

where $\beta(\mathbf{x}_i, \mathbf{t}_j)$, is the uncertainty coefficient predicted at the i-th point of $\mathbf{r}$. Sat-NeRF learns to predict the uncertainty β at each point of the scene based on its spatial coordinates $\mathbf{x}$ (some areas are more likely to exhibit transient objects, e.g. open-air parkings in Figure 8.2) and on the transient embedding vector $\mathbf{t}_j$ of each input training image. Depending on each view, the areas typically affected by transient objects will arbitrarily differ to a greater or lesser extent with respect to the albedo. Note that the embedding vector $\mathbf{t}_j$ is learned from the image index j during training.[1]

We find that it is better to start using β after the second epoch, when the shadow-aware shading s is already well initialized. Otherwise the model may use β to overlook shadow areas instead of trying to explain them with s. Thus, we replace (8.8) with (8.5) in the first two epochs.

8.3.3 Point sampling from satellite RPC models

Sat-NeRF casts rays directly using the RPC camera models of a set of satellite images. The RPC model is widely used for optical satellite imagery, as it allows to describe complex acquisition systems independently of satellite-specific physical properties [Gro01, AMdF+21]. Each RPC is defined by a projection function $\mathcal{P}$, to project 3D points onto image pixels, and its inverse, the localization function $\mathcal{L}$. More details on RPC models can be found in Section 1.4.1.

The use of RPCs in a NeRF framework represents an improvement with respect to previous work with satellite data. In S-NeRF [DI21] the RPC model of each input view is replaced with a custom simplified pinhole camera matrix, which is the common representation used in NeRF for close-range imagery [MST+20]. The RPC-based sampling described here corresponds to a more general approach, which also leads to better results (see Section 8.4).

We denote the minimum and maximum altitude of the scene as h_{min} and h_{max}, respectively.[2] The ray that crosses the scene and intersects the pixel $\mathbf{p}$ of the j-th image is modeled as a straight line between an initial and a final 3D point, i.e. $\mathbf{x}_{\text{start}}$ and $\mathbf{x}_{\text{end}}$. These boundary points are obtained by localizing the pixel $\mathbf{p}$ at h_{min} and h_{max}, using the RPC localization function $\mathcal{L}_j$ of the j-th image:

$$\mathbf{x}_{\text{start}} = \mathcal{L}_j(\mathbf{p}, h_{\text{max}})_{\text{ECEF}}; \quad \mathbf{x}_{\text{end}} = \mathcal{L}_j(\mathbf{p}, h_{\text{min}})_{\text{ECEF}}, \tag{8.10}$$

where the subindex ECEF indicates that the 3D points returned by the localization function $\mathcal{L}_j$ are converted to the Earth-centered, Earth-fixed coordinate system (or geocentric system), to work in a Cartesian system of reference.

[1] At test time, we use an arbitrary $\mathbf{t}_j$ selected from the training set.

[2] The altitude boundaries $[h_{\text{min}}, h_{\text{max}}]$ can be selected in various ways, e.g. from a large-scale elevation model extracted from low-resolution data.

Given $\mathbf{x}_{\text{start}}$ and $\mathbf{x}_{\text{end}}$, the origin $\mathbf{o}$ and direction vector $\mathbf{d}$ of the ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ that intersects the pixel $\mathbf{p}$ of the j-th image are expressed as

$$\mathbf{o} = \mathbf{x}_{\text{start}}; \quad \mathbf{d} = \frac{\mathbf{x}_{\text{end}} - \mathbf{x}_{\text{start}}}{\|\mathbf{x}_{\text{end}} - \mathbf{x}_{\text{start}}\|_2}. \tag{8.11}$$

The point of maximum altitude, $\mathbf{x}_{\text{start}}$, which is the closest to the camera, is taken as the origin $\mathbf{o}$ of the ray. The boundaries of the ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$, i.e. $[t_{\min}, t_{\max}]$, are set as $t_{\min} = 0$ and $t_{\max} = \|\mathbf{x}_{\text{end}} - \mathbf{x}_{\text{start}}\|_2$. Since working with ECEF coordinates is impractical, due to large values used in the representation, we normalize all ray points in the interval $[-1, 1]$ using an offset subtraction and scaling procedure similar to the one used in the RPC functions [Gro01]. The set of 3D points resulting from localizing all pixels in the input images at $h_{\min}$ and $h_{\max}$ is used to compute the offset and scale in each spatial dimension.

8.3.4 RPC refinement for improved performance

Bundle adjustment approaches, previously discussed in Chapters 3 and 4, are a common good practice in remote sensing to correct inconsistencies between a collection of RPC models observing the same scene [GD03, ODM+14, MdFMLF19, MdFML+21]. In particular, bundle adjustment methods correct the RPCs by minimizing the reprojection error of a set of corresponding points seen across the images [TMHF00].

In absence of a prior RPC refinement, a 3D point projected with different raw RPC functions often falls on non-coincident image points, by a distance of up to tens of pixels [GD03]. This would cause a systematic loss of accuracy in any NeRF methodology for satellite imagery, because rays traced from corresponding pixels of different views would not intersect at an exact point in the object space. To prevent this situation, before training Sat-NeRF, we apply the bundle adjustment method described in Chapter 4, which performs a relative correction of the RPC models of all input images [MdFML+21]. The set of points used by the bundle adjustment is derived from correspondences of SIFT keypoints [Low04].

While the refined RPC models directly increase the accuracy of the point sampling strategy described in Section 8.3.3, a prior bundle adjustment can also improve the Sat-NeRF performance in other ways. DS-NeRF [DLZR22] discussed how the training of a NeRF can benefit from a sparse set of previously known 3D points, under the idea that such points can be easily produced using SfM pipelines. In the case of satellite imagery, the bundle adjustment produces an equivalent set of sparse 3D points derived from image features [BNM+09, SAM+16, PHZC21, MdFML+21]. Based on this idea, we explore the benefits of adding the depth supervision term proposed in [DLZR22] to the loss of our Sat-NeRF model:

$$L_{\text{DS}}(\mathcal{R}_{\text{DS}}) = \sum_{\mathbf{r} \in \mathcal{R}_{\text{DS}}} w(\mathbf{r}) \left(d(\mathbf{r}) - \|\mathbf{X}(\mathbf{r}) - \mathbf{o}(\mathbf{r})\|_2\right)^2, \tag{8.12}$$

where $d(\mathbf{r})$ is the depth (8.4) predicted for a ray $\mathbf{r}$, whose origin point is $\mathbf{o}(\mathbf{r})$. If $\mathbf{r}$ intersects $\mathbf{X}(\mathbf{r})$, a known 3D point, then $\|\mathbf{X}(\mathbf{r}) - \mathbf{o}(\mathbf{r})\|_2$ is the target depth to be learned. $\mathcal{R}_{\text{DS}}$ denotes a batch of rays that intersects known 3D points. Since the pixel coordinates associated with these 3D points are already provided by the bundle adjustment, all rays in $\mathcal{R}_{\text{DS}}$ can be defined as explained in Section 8.3.3. Normalized coordinates between $[-1, 1]$ are used in (8.12) to represent points in the object space, for consistency with Section 8.3.3.

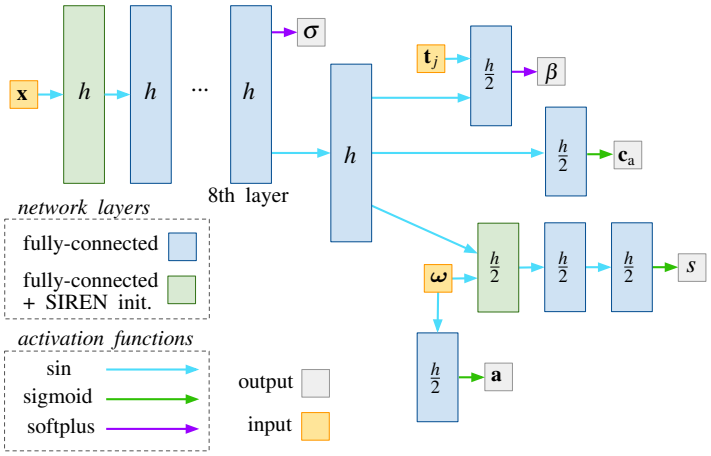


Figure 8.3: Sat-NeRF network architecture, where $\mathbf{x}$ are the input spatial coordinates, $\boldsymbol{\omega}$ is the direction of solar rays and $\mathbf{t}_j$ is the learned transient embedding of image j. The model predicts the volume density σ, the components of the irradiance model (8.6), i.e. albedo color $\mathbf{c}_\text{a}$, shading scalar s, ambient color $\mathbf{a}$, and an uncertainty coefficient β to weight the impact of transient objects.

Similarly to [DLZR22], we only use L_DS in the initial 25% of training iterations. In our experience, this proportion is usually enough to gain accuracy in the learned geometry. Observe that the contribution of each depth supervision ray $\mathbf{r}$ in $\mathcal{R}_\text{DS}$ is weighted by $w(\mathbf{r})$ in (8.12), where $w(\mathbf{r})$ is a scalar set according to the reprojection error of each point $\mathbf{X}(\mathbf{r})$ provided by the bundle adjustment.

8.3.5 Multi-task loss and network architecture

The main term of the Sat-NeRF loss function is the L_RGB defined in (8.8), which is complemented by the solar correction term L_SC (8.7) and the depth supervision term L_DS (8.12). The complete loss function can be expressed as

$$L = L_\text{RGB}(\mathcal{R}) + \lambda_\text{SC} L_\text{SC}(\mathcal{R}_\text{SC}) + \lambda_\text{DS} L_\text{DS}(\mathcal{R}_\text{DS}), \tag{8.13}$$

where λ_SC and λ_DS are an arbitrary weight given to each secondary term. We empirically find $\lambda_\text{SC} = 0.1/3$ and $\lambda_\text{DS} = 1000/3$ to provide good results, to keep the secondary terms sufficiently relevant but below the magnitude of L_RGB. For depth supervision, we used ∼2k-10k bundle adjustment points depending on the exact number used for each area of interest. $\mathcal{R}$, $\mathcal{R}_\text{SC}$ and $\mathcal{R}_\text{DS}$ have the same batch size.

The architecture of Sat-NeRF is shown in Figure 8.3. The main block of fully-connected layers, with h channels per layer, is dedicated to the prediction of the static properties of the scene: the volume density σ and the albedo color $\mathbf{c}_\text{a}$. A secondary head is added with fewer layers and half as many channels per layer to estimate the shading scalar s based on the direction of solar rays $\boldsymbol{\omega}$ and the vector of h geometry-related features learned by the main block. Lastly, two single-layer heads are used to predict the uncertainty coefficient β and the ambient color $\mathbf{a}$, from the transient embedding vector $\mathbf{t}_j$ and $\boldsymbol{\omega}$, respectively.

We employ SIREN layers with the initialization proposed in [SMB+20], as suggested in [DI21]. The use of a softplus function to predict σ was crucial to achieve satisfactory results. The uncertainty β is also produced by a softplus [MBRS+21], which yields a

Area index	004	068	214	260
number of train/test images	9/2	17/2	21/3	15/2
Altitude boundaries [m]	[-24, 1]	[-27, 30]	[-29, 73]	[-30, 13]

Table 8.1: Number of training and test images used for each area, and the altitude boundaries of the scene considered in each case.

smoother optimization problem compared to the usual ReLU [BMT+21]. The rest of the outputs result from sigmoid functions, since they are related to normalized RGB values and have to be in the interval $[0, 1]$. The value of h should be adjusted according to the resolution and the size of the observed area. In our experiments we set $h = 512$.

8.4 Evaluation and comparison with classic multi-view stereo

We evaluate Sat-NeRF on different areas of interest (AOI) of the *2019 IEEE GRSS Data Fusion Contest* [BFC+19, LSYH+19], which provides 26 Maxar WorldView-3 images collected between 2014 and 2016 over the city of Jacksonville, Florida, United States. We take as input a set of RGB crops of varying size, around 800×800 pixels, with a resolution of 0.3 m per pixel at nadir, covering 256×256 m for each AOI. The indices of the selected AOIs and the number of training and test images are listed in Table 8.1.

In all conducted experiments we use a single NeRF model, trained with an Adam optimizer starting with a learning rate of $5e^{-4}$, which is decreased at every epoch by a factor $\gamma = 0.9$ according to a step scheduler. The batch size is 1024 rays, and each ray $\mathbf{r}$ is discretized into 64 uniformly distributed 3D points. Training takes 300k iterations to converge, resulting in ∼10 h if a single batch of rays is used at each training iteration, or ∼20 h if a secondary term for solar correction or depth supervision is added to the loss (trained on a GPU with 16 GB RAM). We used bundle adjusted RPCs (Section 8.3.4) unless otherwise noted.

8.4.1 Ablation study

We evaluated the Sat-NeRF model starting from a simple NeRF and gradually adding new components. To this end, we propose three categories of experiments that are discussed below. Table 8.2 shows the quantitative results.

Category 1. Rows 0-3 are an ablation study dedicated to the irradiance model and the solar correction term described in Section 8.3.1. We verify that the S-NeRF irradiance model outperforms a basic NeRF and is strengthened by the solar correction term. Comparing our results with the ones reported in the original S-NeRF work [DI21] reveals the impact of the proposed RPC-based point sampling and bundle adjustment detailed in Section 8.3, to which we attribute the difference between the metrics of row 0 and row 3.

Category 2. Rows 4-5 assess our Sat-NeRF model, which incorporates the uncertainty prediction of β and employs (8.8) as main term of the loss function. These rows show that the uncertainty modeling improves both the learned geometry and the novel view synthesis, as illustrated in Figure 8.4. Compared to the best S-NeRF results (row 3), Sat-NeRF (row 4) provides higher PSNR/SSIM and similar or even smaller altitude MAE without requiring any solar correction. This insight could be exploited in settings that cannot afford additional training time to process a secondary batch of rays for solar correction. If we add the solar correction term (8.7) to the Sat-NeRF loss (8.8), the altitude MAE decreases even more: row 5 outperforms all previous configurations across all AOIs.

Area index	PSNR ↑ 004	068	214	260	SSIM ↑ 004	068	214	260	Altitude MAE [m] ↓ 004	068	214	260
0. S-NeRF + SC [DI21]	—	—	—	—	0.344	0.459	0.384	0.416	4.418	3.644	4.829	7.173
1. NeRF	17.93	10.26	15.26	14.95	0.559	0.536	0.736	0.443	3.327	2.591	2.691	3.257
2. S-NeRF	25.87	24.20	24.51	21.52	0.864	0.900	0.939	0.829	1.830	1.496	3.687	3.245
3. S-NeRF + SC	26.14	24.07	24.93	21.24	0.871	0.891	0.943	0.825	1.472	1.374	2.406	2.299
4. Sat-NeRF	26.16	24.80	25.54	21.88	0.876	0.903	0.951	0.840	1.416	1.275 ●	2.125	2.428
5. Sat-NeRF + SC	26.67	25.07	25.50	21.78	0.884	0.908	0.950	0.842	**1.288** ●	**1.249** ●	2.009	1.864
6. Sat-NeRF + SC (no BA)	21.55	22.87	24.53	20.96	0.571	0.874	0.942	0.816	1.577	1.392	2.176	1.875
7. Sat-NeRF + DS	26.43	25.27	25.69	21.94	0.879	0.913	0.952	0.842	1.420	1.298	1.714 ●	**1.624** ●
8. Sat-NeRF + DS + SC	26.62	25.00	25.66	21.66	0.881	0.909	0.952	0.839	1.366 ●	1.277	**1.676** ●	1.638 ●
S2P (10 pairs) [FdFML17]	—	—	—	—	—	—	—	—	1.370 ●	1.174 ●	1.811 ●	1.640 ●

Table 8.2: Numerical results using the test images (unseen in training). Rows 1-8 use the RPC-based sampling introduced in Section 8.3.3. $\lambda_{SC} = 0.05/3$ in row 3, otherwise $\lambda_{SC} = 0.1/3$ and $\lambda_{DS} = 1000/3$ where solar correction (SC) and depth supervision (DS) are used. The best altitude MAE values are given gold ● , silver ● and bronze ● medals. Sat-NeRF + SC and/or DS are the most awarded NeRF variants.

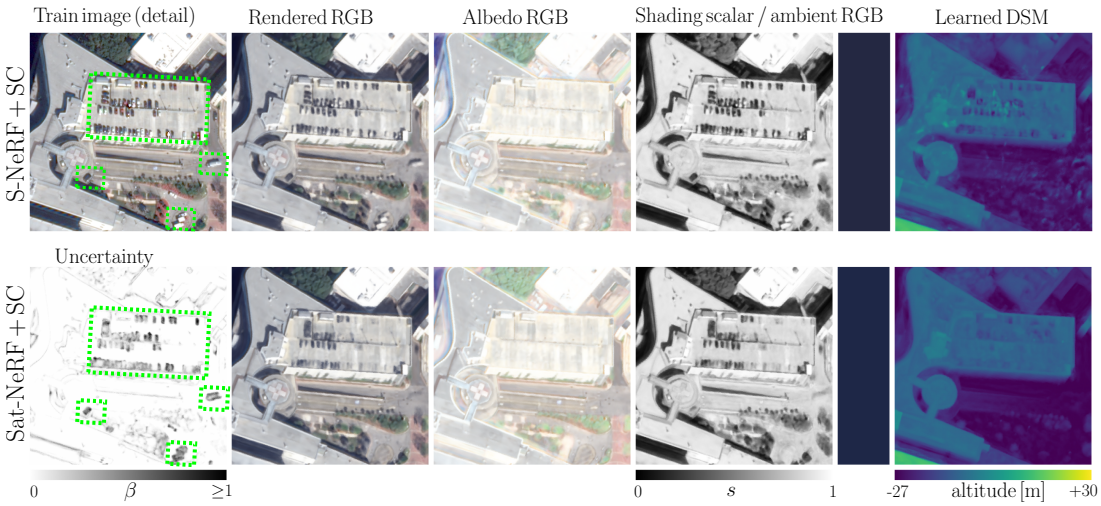


Figure 8.4: The uncertainty coefficient β learned by Sat-NeRF helps to improve the geometry learning with respect to S-NeRF [DI21]. In addition to shadows and textures, the shading scalar s of the irradiance model (8.6) usually attempts to account for transient objects (e.g. cars, marked in green). Sat-NeRF uses β to minimize the contribution of transient objects to the loss function (8.8), thus allowing the geometry and s to learn to ignore their color. In practice, we notice that s still retains some transient objects, but the learned geometry is much better, as shown in the corresponding DSM.

Category 3. Rows 6-8 demonstrate the benefits of using a prior bundle adjustment to refine the RPC models of the satellite images (Section 8.3.4). The comparison of rows 6 and 5 reveals that the use of unrefined RPCs induces a performance drop: both PSNR/SSIM and altitude MAE are worse. Lastly, rows 7-8 add the depth supervision term (8.12) to the Sat-NeRF loss (8.8) with and without solar correction, to leverage the sparse point cloud provided by the bundle adjustment. This strategy proves to be beneficial to further improve the altitude MAE in some areas (214 and 260). In other areas (004 and 068), the sparse point cloud may contain outliers, especially in multi-date collections, leading to slightly worse MAE compared to a plain Sat-NeRF loss.

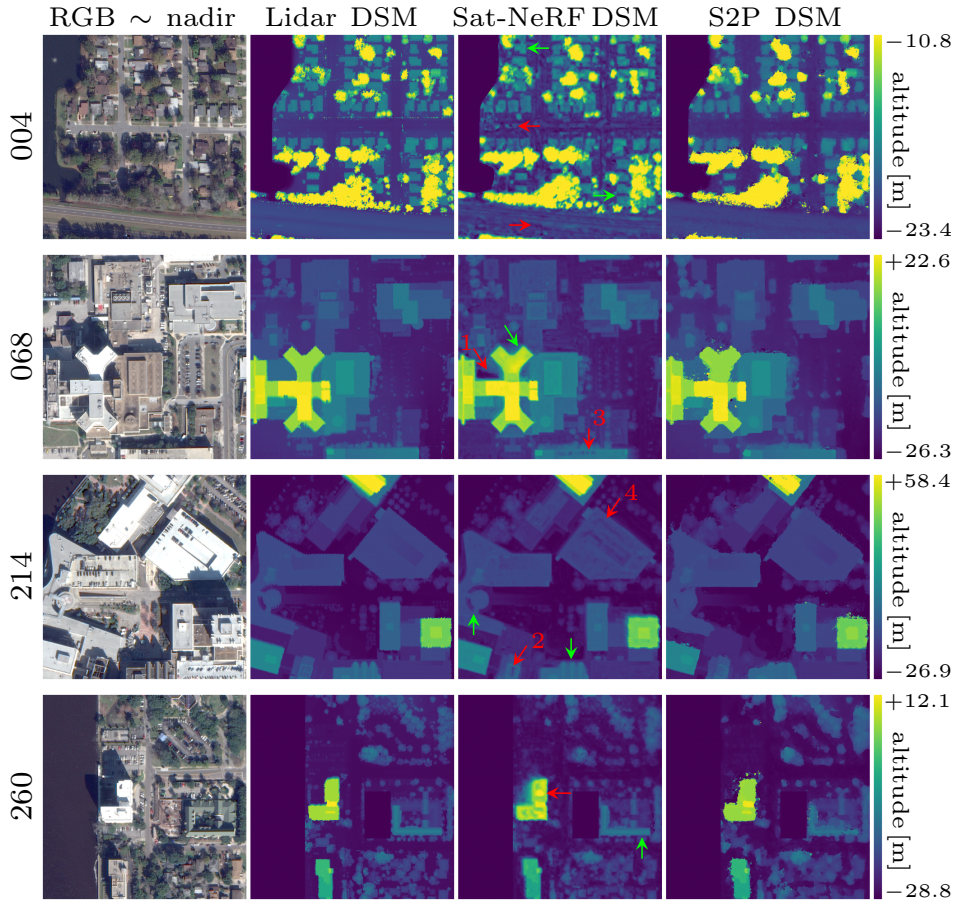


Figure 8.5: Left to right: ground truth lidar DSM, Sat-NeRF DSM and S2P DSM [dFMLM+14a]. The Sat-NeRF DSM corresponds to the model with lowest altitude MAE, in bold in Table 8.2. Compared to S2P, structures are sharper and more detailed in Sat-NeRF DSMs (green arrows), which are also free of single-point outliers. However, Sat-NeRF produces more local irregularities: roofs and roads are less flat (red arrows). Certain roofs exhibit holes, that can be explained by their constant changes across the training sequence. The uncertainty coefficient β can only absorb occasional inconsistencies, which does not include roofs under construction (arrows 1-2) or roofs that are unusually free of parked cars (arrows 3-4). For clarity, we provide an RGB view of the area from a near-nadir perspective. Water bodies are masked in the DSMs.

8.4.2 Comparison with traditional stereo-based pipelines

Sat-NeRF learns high quality 3D models, similar in accuracy to those obtained with satellite stereo pipelines relying on traditional algorithms for stereo matching [BAM18, dFMLM+14a]. We compare the DSMs produced by Sat-NeRF with a multi-view stereo DSM of the same area generated with S2P [dFMLM+14a, FdFML17], the satellite stereo pipeline that won the *2016 IARPA Multi-View Stereo 3D Mapping Challenge* [BKHB16].

We follow the methodology described in [FdFML17] to produce the S2P DSMs. For each AOI, we manually select 10 stereo pairs for disparity estimation. The selection criterion prioritizes pairs with an angle between views of 5 to 45 degrees, with a maximum incidence angle of 40 degrees for each view. Within this set, we take the 10 pairs with

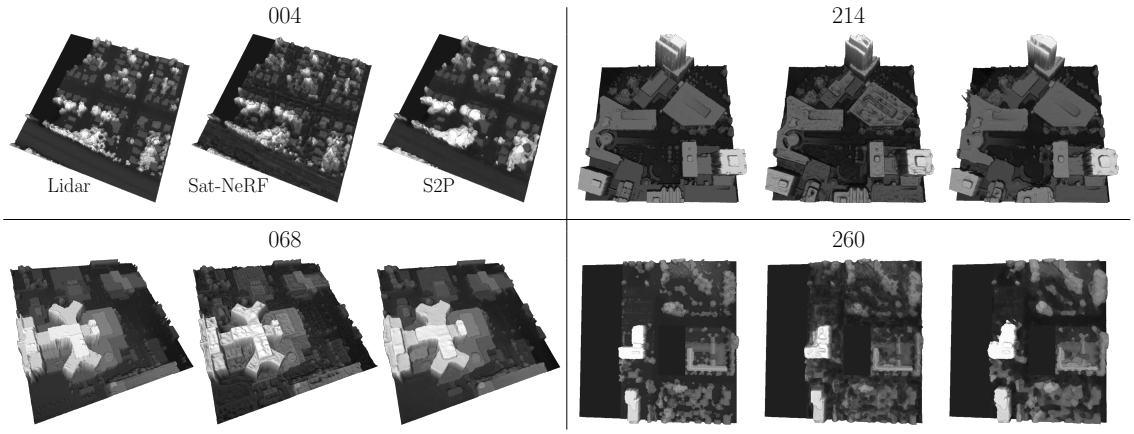


Figure 8.6: 3D visualization of the lidar, Sat-NeRF and S2P DSMs shown in Figure 8.5. Compared to classic multi-view stereo with manual pair selection, represented by S2P, Sat-NeRF provides finer details and sharper edges but exhibits local irregularities.

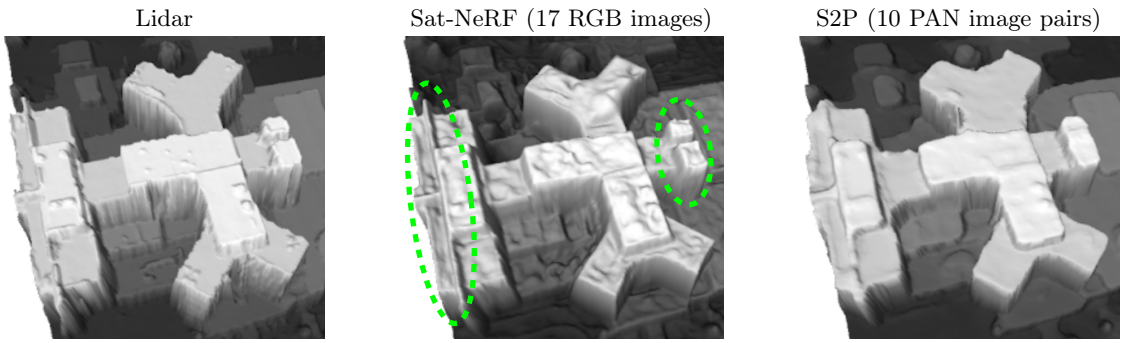


Figure 8.7: Detail of 3D models, area index 068. Sat-NeRF is able to learn sharp small altitude discontinuities, as in the highlighted areas, but is affected by local irregularities. We trained Sat-NeRF with RGB images and used panchromatic (PAN) images for S2P.

closer acquisition dates and run S2P. The RPCs used by S2P were the same used to train Sat-NeRF, i.e. all RPCs are corrected using the bundle adjustment presented in Chapter 4 [MdFML+21]. The 10 pairwise models are fused into a single DSM by taking the median altitude at each cell. To maximize the quality of the S2P DSMs we used the panchromatic product of the WorldView-3 images, instead of the RGB crops employed to train Sat-NeRF. Considering that the RGB images have a compressed dynamic (integer values in $[0, 255]$), i.e with less texture and more saturated areas, the Sat-NeRF DSMs are very encouraging compared to the state of the art with manual pair selection.

As shown in Figure 8.5 and 8.6, structures are more detailed in Sat-NeRF DSMs, but S2P provides more regular surfaces. For a better visualization, see the details in Figure 8.7. Numerically, in some cases, the global altitude MAE obtained with Sat-NeRF is slightly better with respect to the S2P DSMs (Table 8.2, last row), which are affected by isolated outliers. Future work points to hybrid methods or the aggregation of contour-preserving regularization techniques [NBM+22, EMF22].

8.5 Conclusion

This chapter introduced the Sat-NeRF neural network, a NeRF variant adapted to multi-view satellite images. Differently from other deep learning methods for DSM generation from satellite images, Sat-NeRF does not require explicit geometry supervision. Based on the color of the input images, a continuous representation of the 3D scene, between some altitude boundaries, is learned in a self-supervised manner. The geometry and appearance of permanent structures are simultaneously learned using a main backbone of fully-connected layers, while shadows and transient objects are learned by secondary heads to handle multi-date inconsistencies. The latter allows the use of multi-date images for training, such as the WorldView-3 images used in the experiments.

The proposed method achieves state-of-the-art results in novel view synthesis and 3D modeling from satellite imagery. It also highlights the benefits of incorporating well-known techniques for satellite image processing into a NeRF framework. In particular, we show how to represent the input cameras using the RPC models characteristic of satellite images, instead of the pinhole cameras commonly used in NeRF for close-range imagery. We also demonstrate the advantages of applying a bundle adjustment step before training time, such as the one proposed in Chapter 4, to improve reconstruction quality and, optionally, to provide additional cues for depth supervision. In comparison to classic satellite stereo pipelines, the main advantages of Sat-NeRF are the ability to achieve greater detail in surface modeling and the true multi-view nature of the method. The main drawbacks are training times, the need for multiple views and the presence of local irregularities in the learned geometry.

9 Disparity estimation networks

Unlike Semi-Global Matching (SGM) and other classic algorithms for disparity map estimation, neural networks have the ability to exploit semantic cues to resolve ill-posed regions, such as low-textured areas. Based on this observation, we take two neural networks for disparity estimation and use them for dense stereo matching in a satellite stereo pipeline, to verify if they lead to real improvements in geometry modeling. In particular, this chapter reviews the Pyramid Stereo Matching (PSM) and Hierarchical Stereo Matching (HSM) deep learning architectures. Both networks follow the usual structure of feature extraction, cost volume regularization and disparity regression modules, but differ in the way the cost volume is constructed. The performance of the two deep learning methods is compared to that of an SGM variant, using RGB and panchromatic image pairs of WorldView-3 acquisitions over urban landscapes unseen at training time. Pre-trained weights learned from a stereo matching benchmark for aerial imagery are used for both networks. The quality of the disparity maps output is assessed based on the subsequent surface models, which are evaluated using a lidar reference model. The results show that, under ideal conditions (e.g. a suitable baseline and similar acquisition dates), deep learning methods provide higher numerical accuracy compared to a classic SGM variants. PSM provides a higher level of detail, but becomes impractical as image size increases. The main weaknesses of the two networks are their limitation to a predefined disparity range, their sensitivity to the stereo rectification of the input pairs, and the need to adjust the image size to obtain optimal disparity maps.

9.1 Introduction

The concept of disparity refers to the horizontal displacement d between corresponding pixels of two images that observe the same scene from different viewpoints. Estimating disparity from stereo, i.e., knowing that a pixel (x, y) in the left image corresponds to a pixel $(x-d, y)$ in the right image, is a classic and well-known problem in computer vision. The disparity values that match a stereo pair of images are valuable information because they are inversely proportional to the depth of the scene. Assuming a simple pinhole camera model, the depth z of the point that corresponds to the 3D point denoted $\mathbf{x}$ and observed by the pixels (x, y) and $(x-d, y)$ is equal to

$$z = \frac{f \cdot B}{d}, \tag{9.1}$$

where d is the disparity, f is the focal length of the camera and B is the baseline length, corresponding to the segment between the two camera centers. The depth z in (9.1) represents the depth of $\mathbf{x}$ with respect to the baseline of the system.

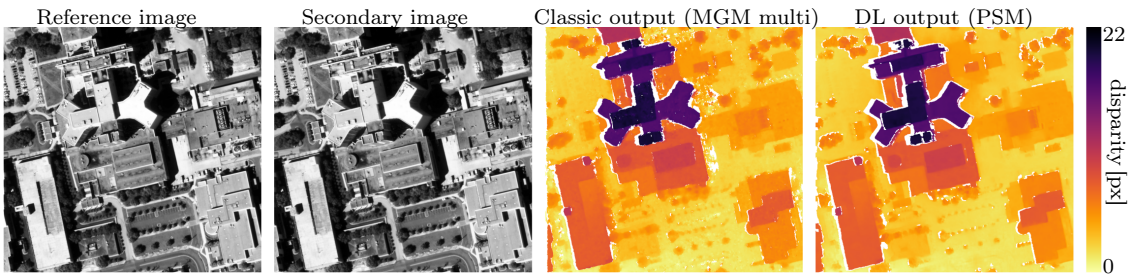


Figure 9.1: Example of input stereo pair [BFC+19] and the resulting disparity maps, in pixel units, obtained with a classic matching algorithm [FdFM15] and a deep learning (DL) network [CC18].

In this chapter we review two deep learning networks for disparity estimation from an input stereo pair: the Pyramid Stereo Matching (PSM) and Hierarchical Stereo Matching (HSM) architectures [CC18, YMHR19]. In particular, the objective is to evaluate the suitability of such methods for high-resolution remote sensing images. Figure 9.1 shows an example of the input and output data.

The majority of satellite stereo pipelines employ classic methodologies to construct disparity maps, which could be potentially replaced by a neural network. However, the lack of public benchmarks makes direct comparisons difficult [GRFvGG22, WVPDR21]. Most deep architectures are originally conceived and trained for synthetic or street-level scenes [GLU12, MG15, MIH+16], which raises questions about their performance in other fields, such as satellite imagery. It seems unfair to draw conclusions using networks that have never seen anything resembling a satellite image. For this purpose, we employ pre-trained weights that have been fine-tuned using a stereo matching benchmark for aerial imagery [WVPDR21]. Our choice is motivated by the fact that aerial images can be understood as fragments of very high-resolution satellite images, with a great similarity from the semantic point of view. Previous work has already demonstrated the advantages of using fine-tuned weights to work with remote sensing images [GRFvGG22, WVPDR21].

To assess the methods, we replace the matching algorithm of the satellite stereo pipeline S2P [dFMLM+14a] using each of them. Then, using the same rectified pairs and camera models, S2P is used to reconstruct multiple surface models, which are evaluated using a lidar-acquired ground truth. Since the only element that changes is the disparity map used to extract each surface model, the altitude errors are directly indicative of disparity accuracy.

9.2 Related work

The combination of global and local information is a widely studied subject in the field of disparity estimation. Both classic and deep learning methods have proposed various strategies to address this key issue. For each branch, this section reviews the most relevant algorithms related to our experiments.

9.2.1 Classic stereo matching

Local methods compute a matching cost (e.g. sum of absolute differences, normalized cross correlation, census) between a window centered on a pixel of the reference image

and an equivalent window centered on some pixel of the secondary image [SS02, FH15]. Epipolar geometry is used to reduce the amount of matching candidates [HZ04]. These methods are known to fail on ill-posed regions containing repetitive patterns, untextured regions or reflective surfaces.

Global methods overcome the limitations of local methods by approaching stereo matching as an energy minimization problem that adds a regularization term to the matching cost. The idea behind the regularization term is that neighbor pixels of the same object should have similar disparities. Most global matching energies take the generic form

$$E(d) = \sum_{\mathbf{p} \in I} C(d_{\mathbf{p}}) + \sum_{(\mathbf{p},\mathbf{q}) \in \xi} V(d_{\mathbf{p}}, d_{\mathbf{q}}), \tag{9.2}$$

where $C(d_{\mathbf{p}})$ is the local matching cost of assigning disparity $d_{\mathbf{p}}$ to pixel $\mathbf{p}$ and $V(d_{\mathbf{p}}, d_{\mathbf{q}})$ is the regularization term enforcing that $d_{\mathbf{p}}$ should be similar to $d_{\mathbf{q}}$, where $\mathbf{q}$ is a neighbor pixel of $\mathbf{p}$. The domain I comprises all nodes (or pixel coordinates) and ξ is the edge set pointing to the neighbor pixels taken into account. Usually, the graph $G = (I, \xi)$ is 4-connected or 8-connected.

The Semi-Global Matching (SGM) algorithm [Hir07] is a popular choice for satellite imagery [BAM18, dFMLM+14a, dR11]. It computes an approximate solution to the NP-hard problem (9.2) using a regularization term equal to

$$V(d, d') = \begin{cases} 0 & \text{if } d = d' \\ P1 & \text{if } |d - d'| = 1 \\ P2 & \text{otherwise.} \end{cases} \tag{9.3}$$

The regularization term (9.3) considers three different categories. A small penalty $P1$ is imposed for small disparity differences (up to 1 pixel), which are common on slanted surfaces. A larger penalty $P2$ (with $P2 > P1$) is given to stronger disparity discontinuities. Finally, there is no penalty if neighbor disparities d and d' are the same. Such categories are particularly suitable for terrestrial surface modeling, which consists mostly of flat or slanted terrain and roofs, with a minority of large discontinuities (e.g. cliffs or building boundaries).

The strategy adopted by SGM consists in dividing the original 2D problem into multiple 1D problems defined on scan lines, which are straight lines that run through the image in the 4 or 8 cardinal directions. Each scan line can be processed as an independent process, allowing for parallelization and high speed computation. For each direction $\mathbf{r}$, a cost volume $L_{\mathbf{r}}$ is computed recursively starting from the image borders. The cost $L_{\mathbf{r}}(\mathbf{p}, d)$ at pixel $\mathbf{p}$ along direction $\mathbf{r}$ at disparity level d is

$$L_{\mathbf{r}}(\mathbf{p}, d) = C(\mathbf{p}, d) + \min \begin{pmatrix} L_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, d), \\ L_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, d-1) + P1, \\ L_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, d+1) + P1, \\ \min_i L_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, i) + P2 \end{pmatrix}, \tag{9.4}$$

where

- $L_{\mathbf{r}}(\mathbf{p}, d)$ is the cost of assigning disparity d to pixel $\mathbf{p}$ following direction $\mathbf{r}$.
- $C(\mathbf{p}, d)$ is the matching cost of assigning disparity d to pixel $\mathbf{p}$.

- $L_{\mathbf{r}}(\mathbf{p}-\mathbf{r}, d)$ is the previous cost in $\mathbf{r}$ direction at disparity d.
- $L_{\mathbf{r}}(\mathbf{p}-\mathbf{r}, d-1)$ is the previous cost in $\mathbf{r}$ direction at disparity $d-1$.
- $L_{\mathbf{r}}(\mathbf{p}-\mathbf{r}, d+1)$ is the previous cost in $\mathbf{r}$ direction at disparity $d+1$.
- $\min_i L_{\mathbf{r}}(\mathbf{p}-\mathbf{r}, i)$ is the previous minimum cost in $\mathbf{r}$ direction, at any disparity level.

By combining (9.3) and (9.4), the cost (9.4) can be summarized as:

$$L_{\mathbf{r}}(\mathbf{p}, d) = C(\mathbf{p}, d) + \min_{d' \in D} \left(L_{\mathbf{r}}(\mathbf{p}-\mathbf{r}, d') + V(d, d') \right). \tag{9.5}$$

The different costs (9.5) computed in each direction $\mathbf{r}$ are added to obtain a single cost volume

$$C_{cost} = \sum_{\mathbf{r}} L_{\mathbf{r}}(\mathbf{p}, d), \tag{9.6}$$

and the final disparity for each pixel is selected using a Winner-Takes-All (WTA) evaluation [FH15] of $C_{cost}(\mathbf{p}, \cdot)$, i.e. the disparity d with minimum cost is taken for each pixel $\mathbf{p}$.

Multiple variants have been proposed to further improve the performance of SGM [PDJB+, RDPD17, FdFM15, AQ22]. We take the MGM variant as a reference for classic stereo matching. MGM or More Global Matching [FdFM15] improves SGM by injecting information from the perpendicular direction $\mathbf{r}^{\perp}$ to each cost along the direction $\mathbf{r}$. In particular, MGM modifies (9.5) as

$$L_{\mathbf{r}}(\mathbf{p}, d) = C(\mathbf{p}, d) + \sum_{\mathbf{x} \in (\mathbf{r}, \mathbf{r}^{\perp})} \frac{1}{2} \min_{d' \in D} \left(L_{\mathbf{r}}(\mathbf{p}-\mathbf{x}, d') + V(d, d') \right). \tag{9.7}$$

Expression (9.7) preserves its recursive nature and requires only minor adjustments in the parallelization process [FdFM15]. The difference is that the cost $L_{\mathbf{r}}(\mathbf{p}, d)$ not only considers the preceding points in a single scan line, but also uses the points of the preceding scan line (i.e. the pixel above). Such strategy improves the predicted disparities and prevents streaking artifacts, which are characteristic of SGM due to the 1D nature of the method.

9.2.2 Deep stereo matching

Like classic methods, deep learning architectures for disparity estimation also aim to obtain a good compromise between local and global matching costs. Feature extraction and the construction of cost volumes also constitute the usual steps in this branch of methods [LJBB20].

Convolutional neural networks (CNN) with encoder-decoder architectures, which had already proven successful in aggregating coarse-to-fine features for semantic segmentation (e.g., UNet), served as inspiration for the first end-to-end models for disparity regression. For example, DispNet [MIH+16] and CRL [PSR+17] reused hierarchical information by concatenating features from the encoder layers with those from the decoder layers. In these earlier models, the features extracted from the left and right image of an input pair were fused in the first layers of the contracting path, by building a correlation volume. Given some left and right feature maps, $\mathbf{f}_L$ and $\mathbf{f}_R$, the correlation volume C_{corr} is commonly computed as the normalized inner product at each disparity level d:

$$C_{corr}(x, y, d) = \frac{1}{F} \langle \mathbf{f}_L(x-d, y), \mathbf{f}_R(x, y) \rangle, \tag{9.8}$$

where F is the number of channels in the feature maps $\mathbf{f}_L$ and $\mathbf{f}_R$. The operation is done for each 2D position (x, y), resulting in an output volume of size $H \times W \times D$ where H and W are the height and width of the feature maps, and D is the disparity range. The resulting correlation volume can be forwarded as another feature map of D channels. The DispNet authors demonstrated that such strategy outperforms directly feeding the CNN with a stack containing both input images [MIH+16].

GC-Net [KMD+17] proposed a different strategy to merge the information from the input pair, by building a cost volume of features learned using a series of residual blocks [HZRS16]. Given the feature maps $\mathbf{f}_L$ and $\mathbf{f}_R$, the cost volume C_{cost} is built by concatenating the feature vectors at each disparity level d:

$$C_{cost}(x, y, d) = concat\left(\mathbf{f}_L(x - d, y), \mathbf{f}_R(x, y)\right), \tag{9.9}$$

where *concat* is the concatenation operation. Following the notation in (9.8) this strategy produces a 4D volume with size $H \times W \times D \times 2F$. The feature dimension F is preserved in this way, allowing the network to exploit contextual information in the later stages. The GC-Net then uses a 3D convolution encoder-decoder structure to regularize the cost volume at multiple scales, followed by a differentiable soft argmin operation to predict the disparity values.

GC-Net inspired the PSM [CC18] and HSM [YMHR19] networks that this chapter reviews in detail. Such models brought significant accuracy improvements by explicitly employing multi-scale features to construct the cost volume C_{cost}. Subsequent deep stereo matching models have also addressed other interesting issues. E.g. DeepPruner [DWM+] does not require a predefined disparity range to search for matches. The architecture is similar to that of PSM [CC18] but a differentiable PatchMatch algorithm is introduced to obtain a sparse cost volume, where the disparity search range is learned and adapted to each pixel. Pruning unlikely disparities provides a significant gain in efficiency. An image guided refinement module is also added to reduce noise and improve sharp boundaries: the disparity map predicted after cost regularization is coupled with features learned from the reference view, and used to feed a lightweight CNN that refines the disparity values.

GA-Net [ZPYT19, GRFvGG22] is another interesting method. It replaces some of the 3D convolutional layers used for cost regularization, which are computationally costly and memory-consuming, by using a Semi-global Guided Aggregation layer (SGA) which is a differentiable approximation of SGM, i.e. (9.5). The SGA layers are followed by a Local Guided Aggregation layer (LGA) to refine thin structures and object edges. The LGA filtering uses the cost values of adjacent disparity levels $(d-1, d, d+1)$ in a $K \times K$ spatial window to refine the cost of disparity d at the central point of the window.

It is common in end-to-end models for disparity and depth estimation [CC18, YMHR19, YLL+18] that the final predicted values are regressed by means of a differentiable soft argmin operation, as originally proposed in GC-Net [KMD+17]. In the last layer of such networks, the regularized costs C_{cost} are compressed from a 4D volume to a 3D volume by means of a single-channel 3D convolutional layer that reduces the feature dimension to a single value. The depth of the 3D volume C_{cost} is then equal to D, the disparity range. The predicted disparity $\hat{d}$ at a 2D point (x, y) is obtained as

$$\hat{d}(x, y) = \sum_{d=0}^{D} d \times softmax(-C_{cost}(x, y, d)). \tag{9.10}$$

All possible disparity levels $d \in D$ contribute to the prediction (9.10) according to a weight equal to the *softmax* function applied to the negative cost at position (x, y, d), to give higher weights to lower costs. Each weight represents the normalized probability of the corresponding disparity level.

This chapter reviews Pyramid Stereo Matching (PSM) [CC18] and Hierarchical Stereo Matching (HSM) [YMHR19], two deep learning architectures for disparity estimation from an input rectified stereo pair of images. Both models follow a structure that can be divided into a first part dedicated to feature extraction and a second part dedicated to the regularization of a cost volume constructed with the previously extracted features. Disparity values are then regressed based on the regularized costs as in (9.10).

9.3 Pyramid Stereo Matching network (PSM)

The PSM network [CC18] architecture is summarized in Figure 9.2(a). This model is inspired by GC-Net [KMD+17] but adds a major ingredient: a Spatial Pyramid Pooling (SPP) module at the end of the feature extraction path to further exploit the global information. This idea was motivated by the effectiveness of the SPP modules, originally introduced for semantic segmentation [ZSQ+17], to expand the receptive field and capture context information.

Feature extraction. Illustrated in Figure 9.2(b), this part consists of a CNN that is followed by a SPP module. The CNN is a contracting path consisting of 2D convolutional layers and a series of residual blocks. Downsampling takes place due to the use of a stride of 2 in certain layers. Some residual blocks use dilated convolutions to help increase the receptive field.

Given an input feature map $\mathbf{f}_i$, the SPP module of PSM applies an average pooling operation with four different kernel sizes, to explicitly generate features at different spatial scales. The multi-scale feature maps then undergo a 1×1 convolutional layer to compress the feature dimension, followed by an upsampling step employing bilinear interpolation, in such a way that their final height and width is the same as that of the input $\mathbf{f}_i$. The different levels of feature maps output by the SPP module are concatenated with $\mathbf{f}_i$ and fused using further convolutional layers (fusion block).

Cost regularization. Feature maps learned by the feature extraction part are used to build a 4D cost volume as explained in Section 9.2. This is done by concatenating the feature vectors of the two images at each disparity level, as in (9.9). The cost volume is regularized using a stacked hourglass architecture, detailed in Figure 9.2(c), that consists of a chain of three encoder-decoder modules of 3D convolutional and deconvolution layers. Each hourglass or encoder-decoder module generates a disparity map that contributes to the loss function, in what is referred to as intermediate supervision. The highest resolution is achieved in the last output, i.e. the third disparity map.

Stacked hourglass architectures with intermediate supervision exploit the possibility to reevaluate initial estimates [NYD16]. The idea is to give multiple opportunities to the network to produce coherent results at both local and global contexts. E.g. if the first hourglass module focuses on very local neighborhoods, subsequent modules will explore higher order spatial relationships.

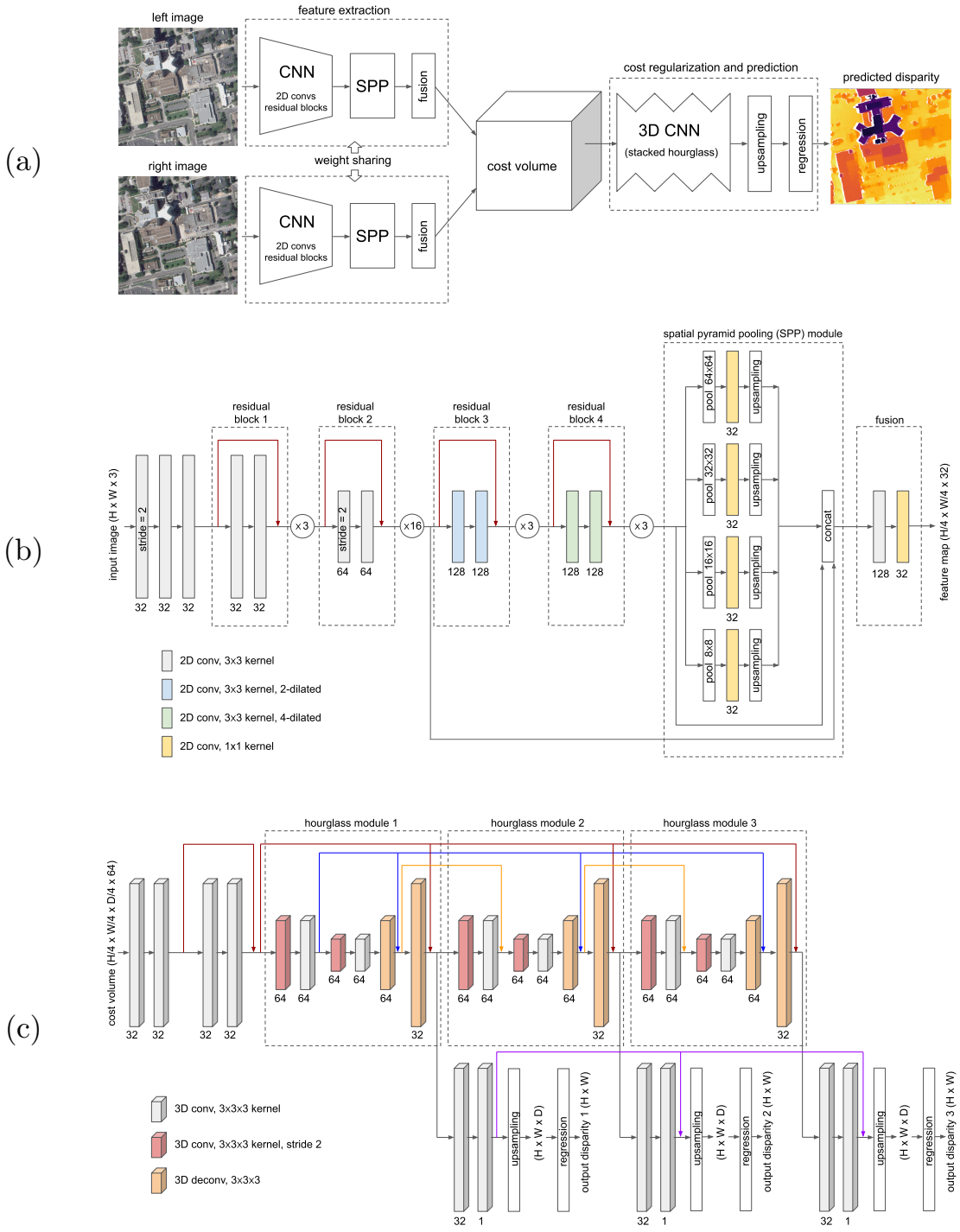


Figure 9.2: PSM network: (a) Overview. (b) Feature extraction path. The number of channels of each convolutional layer is shown below its rectangle. Circles indicate the number of residual blocks: e.g. residual block 1 is repeated 3 times. Bilinear interpolation is used for upsampling. The 2D conv layers are used with batch normalization, to gain stability, and ReLU activation to introduce non-linearities. Colored arrows represent additive skip connections. (c) Cost volume regularization path. This 3D CNN consists of three hourglass modules of 3D conv layers, which are used to aggregate the feature information along the disparity dimension. Trilinear interpolation is used for upsampling. Colored arrows represent additive skip connections. Each hourglass module predicts a disparity map, with increasing detail definition.

Loss function. The PSM network is supervised using a *smooth* L_1 loss function, chosen for its higher robustness to outliers with respect to the L_2 loss.

$$L(d_{\text{GT}}, \hat{d}) = \frac{1}{N}\sum_{i=1}^{N} smoothL_1(d_{\text{GT}} - \hat{d}), \quad \text{where} \quad smoothL_1(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise,} \end{cases} \tag{9.11}$$

where N is the number of pixels and d_{GT} and $\hat{d}$ are the ground truth and predicted disparities, respectively. Following the strategy of intermediate supervision of the stacked hourglass architecture, the loss terms obtained at each of the three hourglass modules are added to compute the final cost at each training iteration (9.12). The weight of each term is fixed, with increasing value according to the number of hourglass modules already covered

$$L_{\text{PSM}} = 0.5L_1 + 0.7L_2 + L_3, \tag{9.12}$$

where each L_i takes the form of (9.11) and the subscript $i = \{1, 2, 3\}$ refers to the index of the hourglass modules in Figure 9.2(c).

9.4 Hierarchical Stereo Matching network (HSM)

After the breakthrough of the GC-Net and PSM architectures, HSM [YMHR19] was developed with the purpose of gaining efficiency and accuracy when handling high-resolution image pairs, with a larger input image size. The HSM architecture is summarized in Figure 9.3(a).

Feature extraction. Like PSM, the feature extraction part of HSM starts with a CNN followed by a SPP module. The CNN also consists of an encoder structure of 2D convolutional layers and residual blocks. However, the number of residual blocks is significantly decreased with respect to PSM and max pooling and convolutions with a stride of 2 are used to further compress the spatial scale of the feature maps. Similarly to PSM, the SPP module of HSM applies average pooling with four different kernel sizes, but resulting features are merged by addition.

Instead of directly using the SPP output, denoted $\mathbf{f}_{SPP}$, to construct the cost volume, HSM reprocesses $\mathbf{f}_{SPP}$ by means of a decoder structure. This is shown in Figure 9.3(b). In particular, the decoder fuses features before and after SPP by concatenation and gradually upsamples the result using up-convolutions (convolution after upsampling), to produce coarse-to-fine feature maps that reach higher spatial resolutions. Four feature maps are obtained in the end, $\{\mathbf{f}_{SPP}^{(k)}\}$, where $k = \{0, 2, 4, 8\}$ is the upsampling factor with respect to the features provided by SPP. The four final feature maps $\{\mathbf{f}_{SPP}^{(k)}\}$ are compressed to 32 or 16 channels using 1×1 convolutional layers.

Cost regularization. Instead of building a single cost volume, HSM builds a multi-scale pyramid of four cost volumes using the four multi-scale $\{\mathbf{f}_{SPP}^{(k)}\}$ feature maps produced in the feature extraction part (see Figure 9.3(c)). Each cost volume of the pyramid has increasing spatial and disparity resolution. To control the size of the cost volumes, these are constructed in a different way with respect to PSM and GC-Net. Instead of concatenating the left $\mathbf{f}_L$ and right $\mathbf{f}_R$ features, as in (9.9), each C_{cost} volume is constructed using absolute differences:

$$C_{cost}(x, y, d) = |\mathbf{f}_L(x - d, y) - \mathbf{f}_R(x, y)|. \tag{9.13}$$

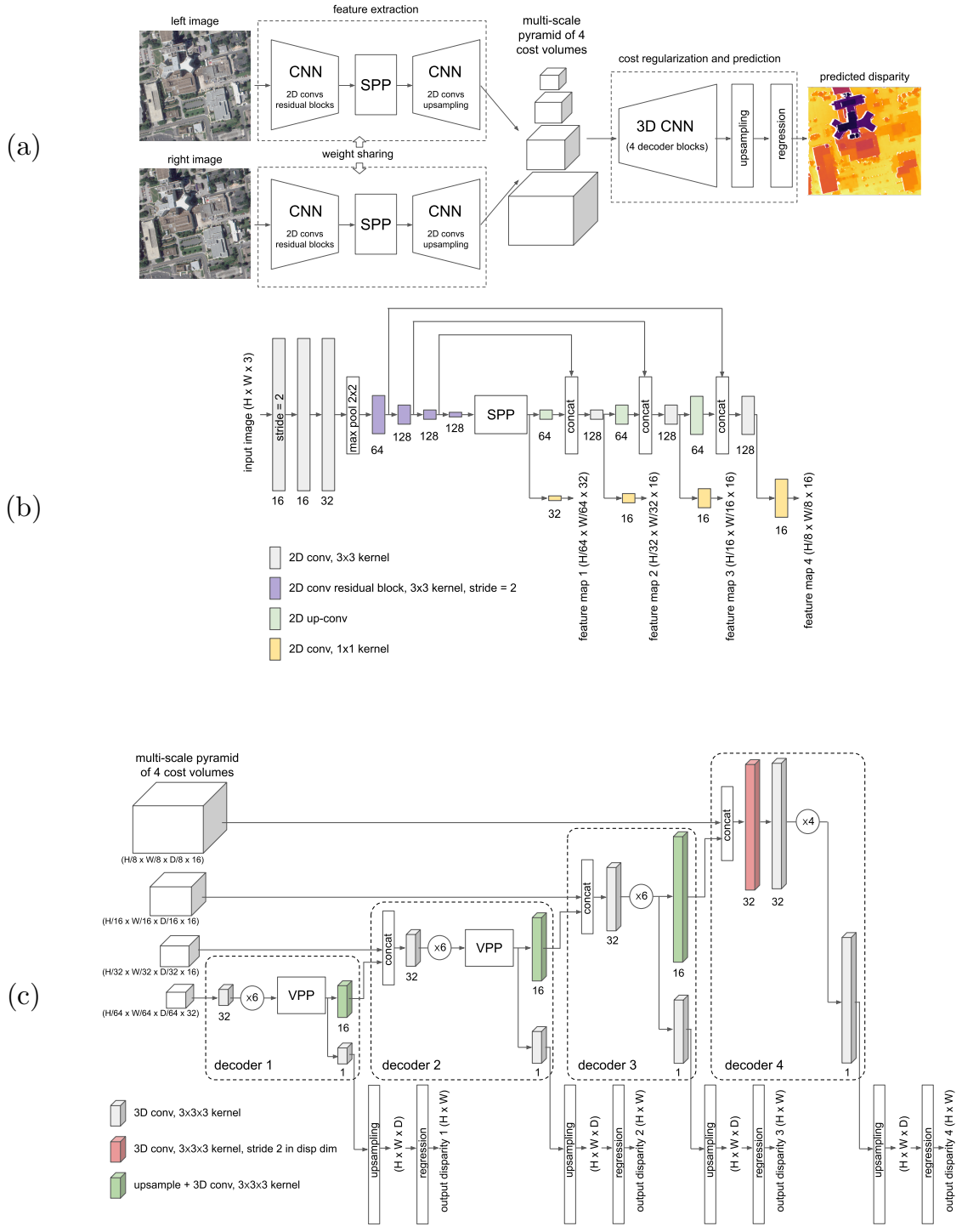


Figure 9.3: HSM network: (a) Overview. (b) Feature extraction path. The number of channels of each convolutional layer is shown below its rectangle. Bilinear interpolation is used for upsampling. The 2D conv layers are used with batch normalization, to gain stability, and ReLU activation to introduce non-linearities. (c) Cost volume regularization path. A chain of decoders, consisting of 3D convs and Volumetric Pyramid Pooling (VPP) blocks, are used to regularize a multi-scale pyramid of cost volumes. Trilinear interpolation is used for upsampling. Circles indicate how many times the preceding layer is repeated.

The set of multi-scale cost volumes is then regularized using a chain of 4 decoders, each devoted to one of the volumes, as shown in Figure 9.3(c). The decoders employ 3D convolutional layers and trilinear upsampling. Larger scale decoders take as input their corresponding cost volume concatenated with the filtered costs provided by the previous decoder. The first two decoders also contain a Volumetric Pyramid Pooling (VPP), which is the exact equivalent of the SPP module, but extended for 4D feature volumes, i.e. using 3D convolutional layers and 3D kernels for average pooling.

Four different disparity maps are regressed using the regularized costs output by the chain of decoders. The highest resolution is achieved in the latest output, i.e. the 4th disparity map.

Loss function. The HSM network is supervised in a very similar way to the PSM architecture, i.e. using a sum of smooth L_1 losses between the predicted and ground truth disparities. The weights assigned to the contribution of each disparity map increase exponentially according to the spatial resolution of the corresponding input cost volume

$$L_{\text{HSM}} = \frac{1}{2^6} L_1 + \frac{1}{2^4} L_2 + \frac{1}{2^2} L_3 + L_4, \tag{9.14}$$

where each L_i takes the form of (9.11) and the subscript $i = \{1, 2, 3, 4\}$ refers to the index of the decoder modules in Figure 9.3(c).

9.5 Experiments: Altitude-based assessment using satellite image pairs

The objective of this chapter is to evaluate the performance of PSM and HSM in the context of satellite imagery. We focus on urban areas, observed at viewing angles not far from nadir. Instead of training the deep learning models from scratch, we use pre-trained weights provided by the 2021 open-source stereo matching benchmark for aerial imagery [WVPDR21]. This ensures that the networks are already familiar with the usual elements of Earth observation images: building roofs, tilted facades, roads, trees, etc. Our experiments aim to assess the generalizability of the models and establish which architecture offers higher robustness to different input specificities.

9.5.1 Aerial stereo matching benchmark

The aerial stereo matching benchmark introduced in [WVPDR21] was built using the ISPRS Vaihingen dataset for urban classification and 3D building reconstruction [fPS12, RSJ+12].

The ISPRS Vaihingen dataset consists of 20 large-scale aerial images taken with an Intergraph/ZI digital mapping camera DMC [RSJ+12]. The aerial images are 16 bit pan-sharpened color-infrared[1] (CIR) images with a resolution of about 8 cm per pixel. A lidar point cloud of the Vaihingen area (Germany) is also available, with a point density varying between 4 and 7 points/m^2.

The 2021 aerial stereo matching benchmark [WVPDR21] provides 1092 pairs of 1024×1024 pixels, cropped from the CIR images and rectified using the MicMac library [PDJB+, RDPD17]. The corresponding ground truth disparity maps were generated from the lidar

[1] The blue channel of each image in the dataset has been replaced by the response of an infrared camera.

point cloud and are stored on 16 bits with the disparity value scaled by 256. Since the lidar data is very sparse, a density-based filter was used to account for occlusions. The authors of the benchmark used 585 training pairs to fine tune the PSM and HSM networks, originally trained on the KITTI dataset [GLU12, MG15].

9.5.2 Altitude-based evaluation and input data

For our experiments, we selected four areas of interest (AOIs) of 256×256 m from the *2019 IEEE GRSS Data Fusion Contest* (DFC2019) [BFC+19]. The DFC2019 dataset provides, among others, 26 WorldView-3 images, with a resolution of about 30 cm per pixel, acquired between 2014 and 2016 over the city of Jacksonville (Florida, US). We take image crops of varying size, around 800×800 pixels, covering each target AOI. The resulting images are used to form stereo pairs.

We select suitable stereo pairs according to the criterion of [FdFML17], with the objective to maximize the accuracy of output disparities. The selection criterion prioritizes pairs with an angle between views from 5 to 45 degrees and a maximum incidence angle of 40 degrees for each view. From this set, we take the 30 pairs with closest acquisition dates and use them as input. The lists of pairs selected for each AOI, as well as some example views, are included in the Appendix B.1.

As a substitute for ground truth disparities, our assessment is based on a digital surface model (DSM) of each AOI, derived from lidar and part of the DFC2019 data. The resolution of the lidar DSMs is 0.5 m per pixel. To convert disparity values to altitude, we take the satellite stereo reconstruction pipeline S2P [dFMLM+14a] and replace the matching algorithm with each deep learning model. The S2P tools are also used to rectify each input pair of images.

The altitude errors resulting from the disparity maps provided by the deep learning methods are compared with those achieved using the S2P baseline classic matching algorithms, i.e. MGM and an improved multi-scale version of MGM, both using the census transform [FdFM15, FdFML17, FH15].

9.5.3 Evaluation metrics

The evaluation metrics are the following:

- **MAE.** Mean Absolute Error, in meters of altitude, between a photogrammetric DSM and the ground truth lidar DSM.
- **Completeness.** Percentage of non-water points in a photogrammetric DSM where error is less than 1 meter with respect to the lidar DSM, with undefined values counted as larger errors.
- **NaNs.** Percentage of undefined values (Not a Number) in a photogrammetric DSM.
- **Successful pairs:** Number of stereo pairs, out of N originally given as input to a matching algorithm, that resulted in DSMs with less than 50% of undefined values.

9.5.4 Results

Using S2P, we compute (1) single-pair DSMs, i.e. the set of DSMs that result from each independent stereo pair of the 30 selected; and (2) multi-pair DSMs, which are denser and result from fusing all the successful single-pair DSMs using a median filter as in [GF18,

MdFML+21]. The camera models given to S2P have been previously corrected using the bundle adjustment presented in Chapter 4 [MdFML+21]. We distinguish between two types of experiments, according to the color space of the input images. Other subsections are devoted to complementary tests that study the performance of each method as a function of the distance between the acquisition dates and the baseline of the input pairs.

Panchromatic inputs

In these experiments, we use the panchromatic version of the DFC2019 images as input. Panchromatic images have a single channel with a wide range of intensity values, which allows them to be highly textured. This makes them well suited for classic matching algorithms. The numerical results are reported in Table 9.1. Qualitative results are shown in Figure 9.4.

Table 9.1 introduces some custom parameters, s and lr, which we observe to significantly affect the accuracy of the disparity maps generated with PSM and HSM (based on the subsequent altitude values). Otherwise, the default parameters of S2P are used, with the exception of the SIFT matching threshold, which is set to 0.5 to aim for very reliable matches for the rectification step.

Panchromatic inputs

		Single pair MAE [m] / NaNs [%] / Successful pairs			
Area index		004	068	214	260
MGM	$s=1, lr=1$	1.863 / 32.01 / 24	0.919 / 23.63 / 29	1.457 / 31.23 / 24	1.668 / 33.30 / 24
MGM multi	$s=1, lr=1$	1.531 / 34.63 / 18	0.886 / 25.74 / 28	1.288 / 31.95 / 23	1.663 / 34.48 / 21
PSM	$s=1, lr=1$	0.933 / 33.17 / 12	0.645 / 25.00 / 28	0.985 / 34.23 / 23	1.279 / 33.94 / 16
HSM	$s=1, lr=1$	1.545 / 34.64 / 20	0.838 / 28.90 / 28	1.229 / 34.89 / 23	1.580 / 31.72 / 24
PSM	$s=2, lr=2$	0.886 / 38.61 / 12	0.511 / 28.34 / 28	0.809 / 36.88 / 23	1.002 / 38.33 / 14
HSM	$s=2, lr=2$	1.110 / 37.63 / 16	0.605 / 28.09 / 28	0.911 / 35.40 / 23	1.189 / 32.98 / 19
PSM	$s=3, lr=3$	0.620 / 37.87 / 12	0.755 / 30.33 / 26	1.051 / 36.89 / 18	1.006 / 38.37 / 11
HSM	$s=3, lr=3$	0.885 / 37.15 / 12	0.635 / 28.88 / 28	0.972 / 36.14 / 23	1.163 / 34.49 / 18

		Multi-pair MAE [m] / NaNs [%]			
Area index		004	068	214	260
MGM	$s=1, lr=1$	1.806 / 0.65	1.030 / 0.62	1.627 / 0.39	1.474 / 0.69
MGM multi	$s=1, lr=1$	1.512 / 0.90	1.026 / 0.68	1.585 / 0.55	1.472 / 0.78
PSM	$s=1, lr=1$	1.327 / 2.59	0.813 / 0.06	1.453 / 0.28	1.297 / 1.08
HSM	$s=1, lr=1$	1.941 / 1.38	1.072 / 0.11	1.838 / 0.34	1.708 / 1.04
PSM	$s=2, lr=2$	1.401 / 6.13	0.780 / 0.25	1.347 / 1.10	1.154 / 3.39
HSM	$s=2, lr=2$	1.405 / 1.52	0.862 / 0.13	1.579 / 0.48	1.293 / 1.16
PSM	$s=3, lr=3$	1.115 / 7.08	0.838 / 0.22	1.677 / 1.26	1.155 / 4.72
HSM	$s=3, lr=3$	1.314 / 3.13	0.893 / 0.48	1.653 / 0.65	1.285 / 0.77

Table 9.1: Quantitative results using panchromatic images as input. The accuracy of the single pair (top) and multi-pair DSMs (bottom) is directly related to the accuracy of the disparity maps. For single-pair DSMs, all metrics are averaged across the successful pairs. Customized parameters: s, upsampling factor applied to the rectified images input to the matching algorithm; lr, left-right consistency check threshold in pixel units. The best altitude MAE values are highlighted in yellow.

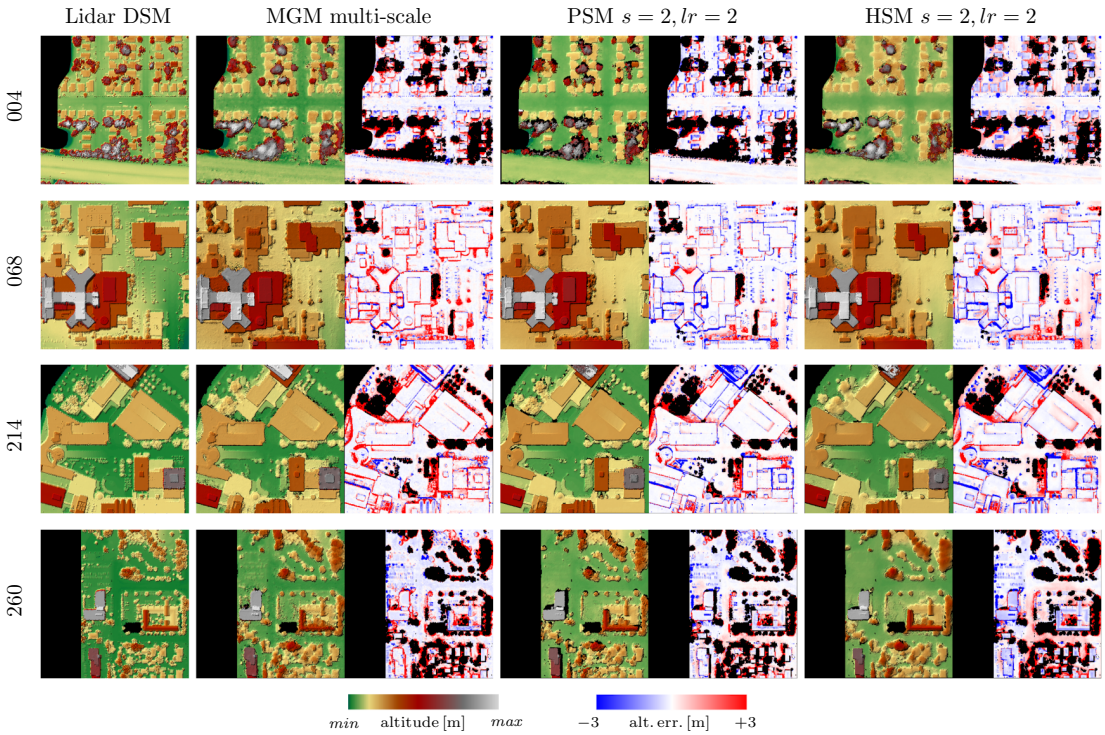


Figure 9.4: Visualization of the lidar DSMs, and the multi-pair photogrammetric DSMs obtained by merging the disparity maps of multiple stereo pairs obtained with MGM multi-scale, PSM and HSM. Panchromatic images used as input. Undefined values and water bodies are in black. Points corresponding to trees/vegetation in the lidar DSM are also masked in black in the the error images.

The following paragraphs discuss some of the main concepts related to Table 9.1 and the findings derived from it.

On the left-right consistency check. The left-right consistency check is a good practice to refine disparity maps: it filters disparity values that are not consistent when the left and right images of the stereo pair are exchanged [Hir07]. This step uses a consistency distance threshold lr, which strongly affects the percentage of NaN values in the disparity map and the subsequent DSM. Larger lr produces more complete DSMs in exchange of small inaccuracies.

On the scaling factor. The upsampling factor s is used to bring objects closer to the resolution of the training set, since aerial images have approximately four times the resolution of satellite images (8 cm vs. 30 cm per pixel). When $s > 1$, the lr threshold should be at least equal to s since the increase of disparity values is proportional to the size of the images. In general, we observe that $s = 2$ works best across the different AOIs, as shown in Figure 9.5. Using $s = 3$ only provides better results for both PSM and HSM in the AOI 004, a landscape of small houses. The rest of AOIs contain tall buildings, and $s > 2$ causes certain disparities related to skyscrapers to be larger than the upper limit of the disparity range set at training time (192 pixels). The latter downgrades the accuracy for some of the input stereo pairs. In addition to the resolution of the training set, the architecture of the networks is another factor that could explain the improvement in performance after upsampling, as shown in Figure 9.6.

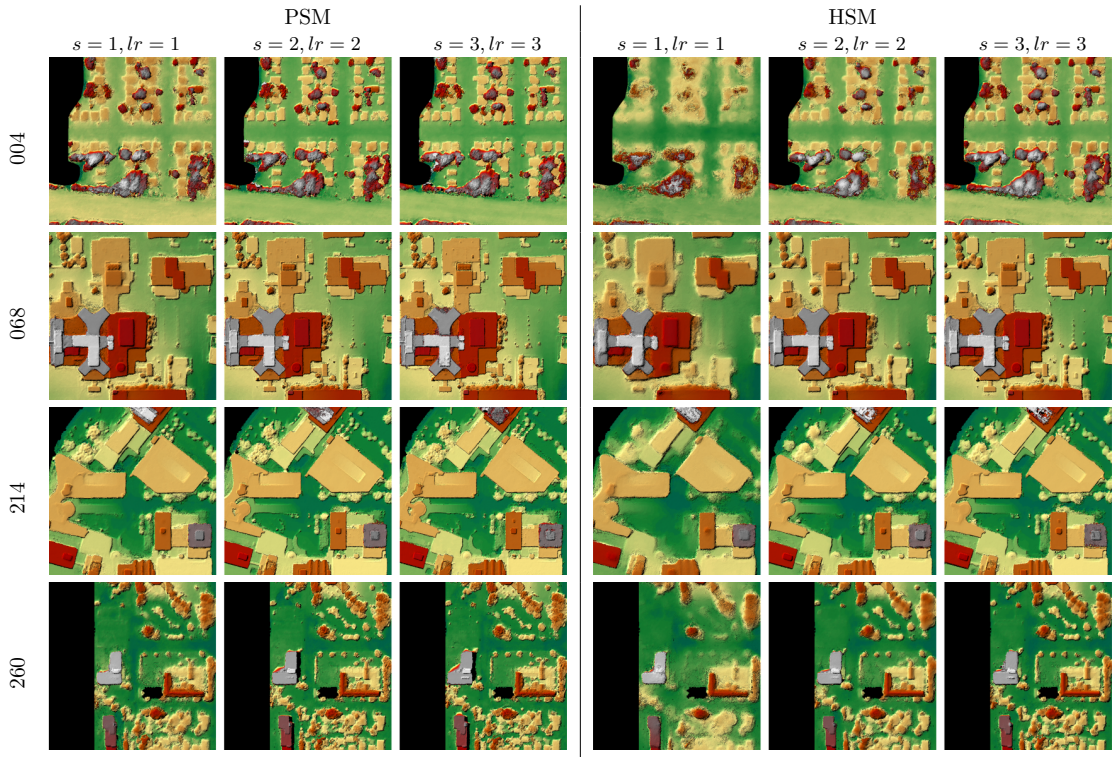


Figure 9.5: The resolution of the input images plays an important role in deep learning methods. By upsampling the input pair by a factor s, we simulate the resolution of the satellite images to be higher and closer to the resolution of the aerial image training set. We find that $s = 2$ works best, as $s = 1$ produces blurred edges and $s = 3$ introduces artifacts in tall buildings. HSM is more sensitive to resolution than PSM, as the difference in detail sharpness is stronger for different s values.

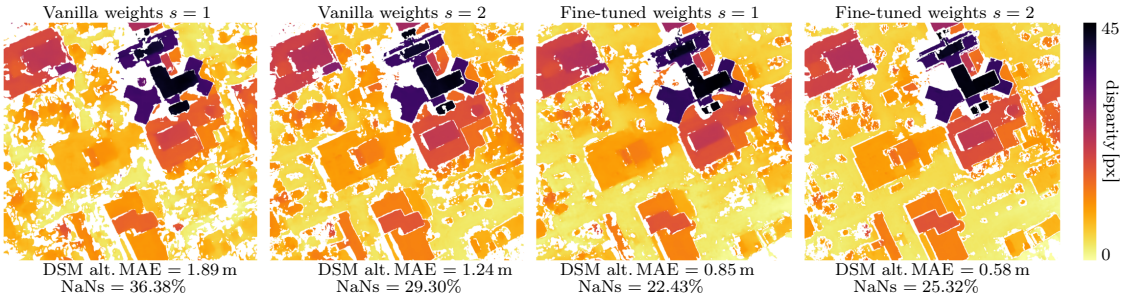
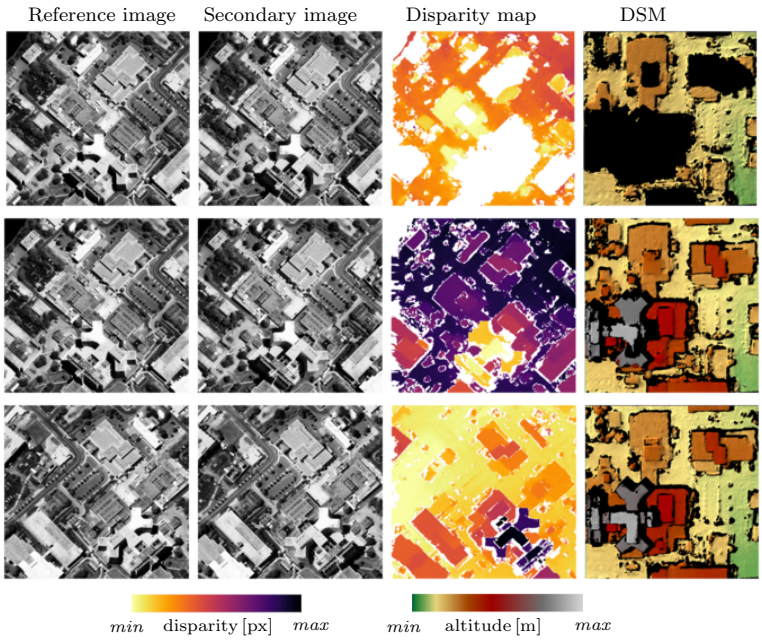


Figure 9.6: PSM vanilla vs. fine-tuned weights, same input pair. Vanilla weights were trained on KITTI2015 [MG15], while fine-tuned weights were refined using the 2021 aerial stereo matching benchmark [WVPDR21]. Fine-tuned weights systematically produce better disparities and less undefined (NaN) values after the left-right consistency check. However, upsampling the input pair by a factor $s = 2$ improves details and accuracy in both cases. This suggests that, in addition to the resolution of the training data, the architecture of the method may also be related to this behavior. Both PSM and HSM use non-learned upsampling operations before disparity regression, to reach the original input size. Thus, the actual costs used to predict the disparity map have lower spatial resolution, and upsampling the input pair could be seen as a bruteforce manner to increase it.



Failure example 1: The input pair is rectified allowing disparities to follow a negative or positive direction, indifferently. As a result, the network only works in areas with negative displacement, making the disparity map and the derived DSM largely incomplete.

Failure example 2: The input pair is rectified forcing all disparities to follow a negative direction, but allowing the background to have a higher disparity than the foreground. As a result, the accuracy of the disparity map and the derived DSM decreases, particularly in foreground objects (e.g. tall buildings).

Success example: The input pair is rectified as expected by the network. All disparities follow a negative direction, the background has lower disparity than the foreground, and all disparities fall within the expected range (e.g. 192 pixels). As a result, the accuracy of the disparity map and the derived DSM are maximized.

Figure 9.7: Deep learning methods are very sensitive to the way in which input pairs are rectified. Using the same pair and the PSM model, the top rows show example results obtained with suboptimal rectification. The bottom row shows the equivalent result when the rectification follows the learned format. Undefined values are filled in white in the disparity maps and black in the DSM.

On sensitivity to the rectification step. Deep learning methods are very sensitive to the way in which input pairs are rectified. The rectification step must satisfy several conditions to obtain optimal disparity maps: (1) All disparity values have to fall within the disparity range defined at training time. (2) All disparities from the reference image to the auxiliary image must point towards the left, following a negative displacement along each epipolar line. (3) Disparity values must be proportional to the altitude level. The networks expect the ground/background to exhibit small disparity, and the opposite for objects in the foreground or tall buildings.

All pairs considered in Table 9.1 were rectified by means of Algorithm 12, to ensure that conditions (2) and (3) were satisfied. If one of the conditions is not met in a given region, the networks fill it with NaN values or the accuracy is degraded (Figure 9.7). In contrast, classic methods such as SGM or MGM do not depend on any priors and can adapt the search range of disparities for each pair.

On the trade-off between accuracy and efficiency. Both PSM and HSM consistently achieve lower MAE than the standard MGM or its multi-scale version. Only HSM with $s = 1$ seems to perform clearly worse, probably because it is more specific for large and very high-resolution inputs. The behavior is the same if we take completeness as the main metric (see Appendix B.2, Table B.5). As shown in Figure 9.4, the superior performance of deep learning methods can be explained by their ability to produce better and sharper contours. Between the two networks, PSM achieves better MAE and completeness than HSM, but the difference narrows as the scaling factor s increases. Using $s \geq 2$, HSM might be the better choice for certain applications, as it requires less memory and is much faster for large inputs. E.g. with $s = 2$, on a CPU, most stereo pairs are processed within 5 seconds using HSM, while they take longer than 1 minute with PSM.

Algorithm 12: Rectify satellite stereo pair for disparity estimation network

input : two satellite images I_1 and I_2 and their respective camera models

output : rectified stereo pair I_1^R and I_2^R

1. Compute a set of *pairwise_matches* between I_1 and I_2
2. Use *pairwise_matches* to find the 3×3 rectifying homographies H_1 and H_2. *See comment 1*
3. Impose negative disparities by modifying H_2 into H_2'

$$H_2' = \begin{pmatrix} 1 & 0 & -t \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} H_2, \quad \text{where } t \geq \text{max disparity observed in } pairwise_matches$$

4. Check that disparity values are proportional to altitude. To do so, localize a point from I_1^R at two different altitudes: alt_1 and alt_2, such that $alt_1 < alt_2$. Then project the resulting locations onto I_2^R. The reprojection localized at alt_1 must have smaller disparity because $alt_1 < alt_2$. *See comment 2*

if step 4 is not satisfied **then**

5. Repeat step 3 imposing positive disparities, i.e. $t \leq$ min disparity in *pairwise_matches*
6. Apply a horizontal flip to both rectifying homographies, e.g

$$H_{flipped} = \begin{pmatrix} -1 & 0 & w \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} H, \quad \text{where } w \text{ is the image width}$$

7. Rectify I_1 using H_1 and I_2 using H_2'.

Comment 1: Steps 1 and 2 can be covered using the stereo-rectification method for pushbroom images described in [dFMLM+14b].

Comment 2: Let the rectifying homographies be denoted $\{H_1, H_2'\}$ and the rectified images be denoted $\{I_1^R, I_2^R\}$. Given the RPC camera models of the non-rectified satellite images $\{I_1, I_2\}$, which are characterized by a localization $\mathcal{L}$ and a projection function $\mathcal{P}$, a point (x, y) in I_1^R can be localized at altitude alt and reprojected to I_2^R as

$$(x', y) = H_2' \mathcal{P}_2(\mathcal{L}_1(H_1(x, y), alt))$$

resulting in a disparity $d = x' - x$

RGB inputs

In these experiments, we use the RGB version of the DFC2019 images as input. RGB images have three channels corresponding to red, green and blue values, which are compressed as integer values in $[0, 255]$. The compressed dynamics leads to a loss of texture and appearance of saturated areas. This makes RGB images more challenging than the panchromatic equivalent for matching purposes.

The numerical results obtained with RGB images are reported in Table 9.2. For each row in Table 9.2, we used the successful pairs from the equivalent row in Table 9.1 as input, with the same rectifying homographies. This ensures that any difference in the resulting disparity maps and DSMs is exclusively due to the change of color space. Example results are shown in Figure 9.8.

RGB inputs

		Single pair MAE (error increase) [m] / NaNs [%]			
Area index		004	068	214	260
MGM multi	$s = 1, lr = 1$	2.020 (+0.489) / 11.19	1.196 (+0.309) / 14.11	2.388 (+1.100) / 20.04	2.064 (+0.401) / 15.85
PSM	$s = 2, lr = 2$	1.358 (+0.472) / 27.69	0.601 (+0.089) / 26.25	1.128 (+0.318) / 36.12	1.297 (+0.294) / 30.81
HSM	$s = 2, lr = 2$	1.314 (+0.203) / 31.66	0.607 (+0.002) / 28.42	0.983 (+0.071) / 35.67	1.280 (+0.091) / 32.10

		Multi-pair MAE (error increase) [m] / NaNs [%]			
Area index		004	068	214	260
MGM multi	$s = 1, lr = 1$	2.132 (+0.620) / 0.68	1.297 (+0.270) / 0.58	2.445 (+0.860) / 0.21	1.728 (+0.256) / 0.59
PSM	$s = 2, lr = 2$	1.764 (+0.363) / 1.36	0.881 (+0.101) / 0.33	1.810 (+0.463) / 1.22	1.426 (+0.272) / 1.76
HSM	$s = 2, lr = 2$	1.606 (+0.201) / 0.71	0.857 (−0.005) / 0.11	1.631 (+0.052) / 0.62	1.440 (+0.147) / 0.88

Table 9.2: Quantitative results for single pair (top) and multi-pair DSMs (bottom), using RGB input images. The error increase refers to the increase in altitude MAE with respect to the MAE obtained with panchromatic images (Table 9.1). Customized parameters s and lr are the same as in Table 9.1.

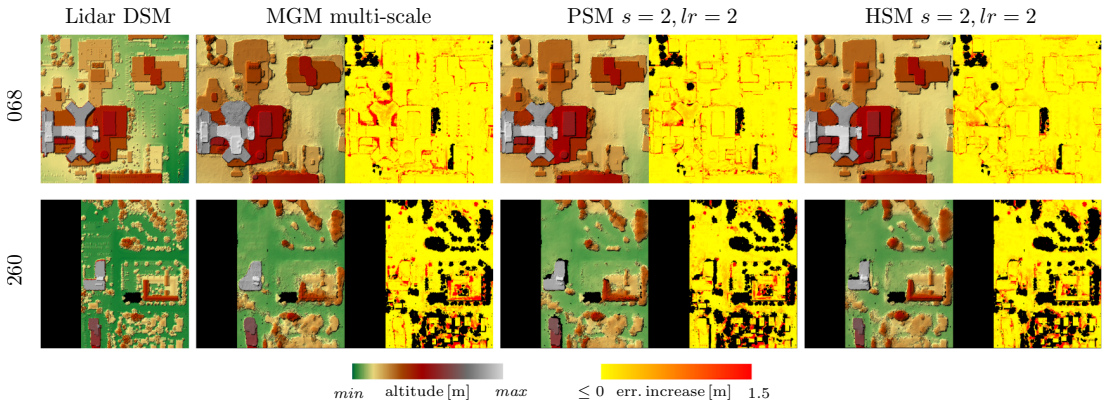


Figure 9.8: Examples of multi-pair photogrammetric DSMs obtained using RGB images as input instead of panchromatic as in Figure 9.4. The yellow-red images show the increase in altitude error with respect to the equivalent result obtained with panchromatic images. Masked points in black represent undefined points and water bodies.

Table 9.2 shows that RGB inputs lead to less accurate disparity maps across all methods, inducing larger altitude MAE in the subsequent DSMs. However, the increase in error is much larger for the classic algorithm (multi-scale MGM), as it cannot compensate for the loss of information with contextual semantic cues in the same way as deep learning networks do. Differently from the panchromatic scenario, MAE and completeness behave in a different way for RGB inputs. MGM multi-scale provided better completeness in the single pair disparity maps (Appendix B.2, Table B.6), but higher inaccuracy too (higher MAE). RGB compression also eliminates fine-scale noise, which reduces the amount of NaN values provided by MGM in some of the pairs.

For panchromatic inputs and $s = 2$, PSM was the most performing model both in terms of MAE (Table 9.1) and completeness (Table B.5). Using RGB inputs and $s = 2$, the performance of HSM and PSM seems to be more or less equal, with HSM being slightly more accurate in terms of MAE (Table 9.2) but also providing slightly lower completeness (Table B.6).

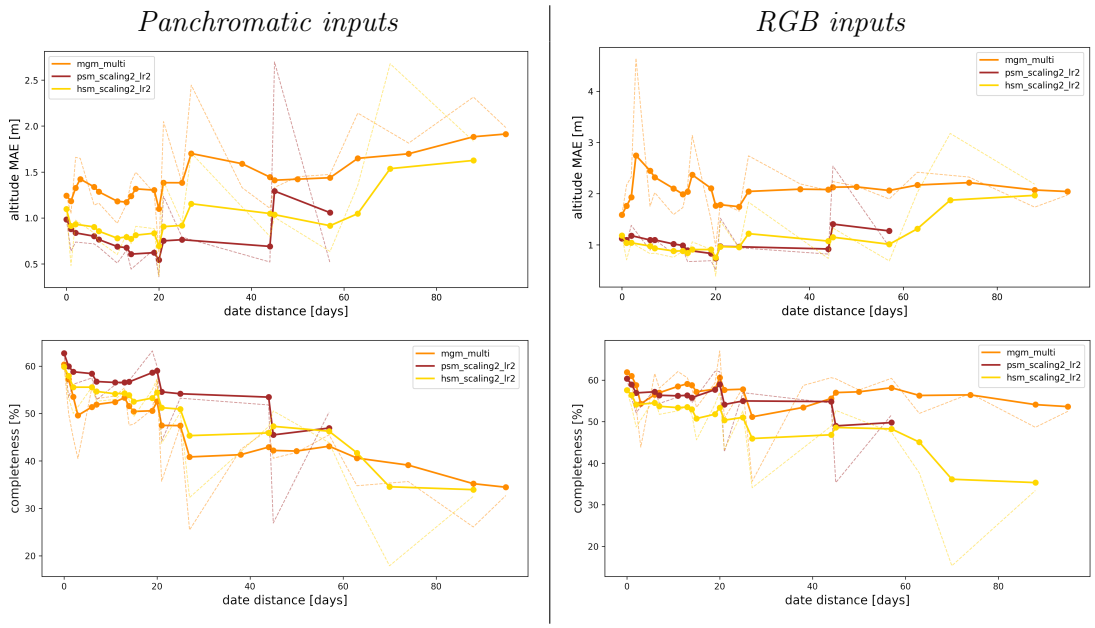


Figure 9.9: Average evaluation metrics as a function of the distance between the acquisition dates of each pair of images. Only successful pairs taken into account, i.e. those resulting in less than 50% of undefined altitude values. For better visualization, the opaque lines represent a smoothed version of the real dashed functions. Each smoothed value x'_i is obtained as $x'_i = 0.7x_{i-1} + 0.3x_i$.

Multi-date inputs

Figure 9.9 shows the altitude MAE values reported in Table 9.1 and Table 9.2 plotted as a function of the distance between the acquisition dates of each pair of images. The same is also shown using the completeness values (Tables B.5 and B.6). The more distant the dates are, the less radiometric correlation is expected (e.g., due to seasonal changes) and the more difficult it is to find correspondences.

For all methods, the disparity maps lose accuracy as the time distance increases, causing an increase in altitude error and a decrease in completeness. Using panchromatic inputs, the decline in performance seems to be slightly more pronounced for deep learning methods, as the gap between MGM multi-scale and the networks narrows towards the end of the plots. For RGB, the decline in performance of deep learning methods is more evident, especially in terms of completeness. In particular, PSM did not produce any successful pairs for time distances over 60 days.

To better understand Figure 9.9, note that the number of pairs with a given distance between acquisition dates is not uniformly distributed. In 90% of cases the distance is less than 30 days.

Small baseline inputs

Figure 9.10 shows multi-scale MGM, PSM and HSM tested on two additional panchromatic input pairs. The *Basilique Saint-Sernin* and *Prison Saint-Michel* (both in Toulouse, France) are not part of the DFC2019 data. We selected these input pairs because of the small baseline between the two cameras, which is a well-known challenging factor for

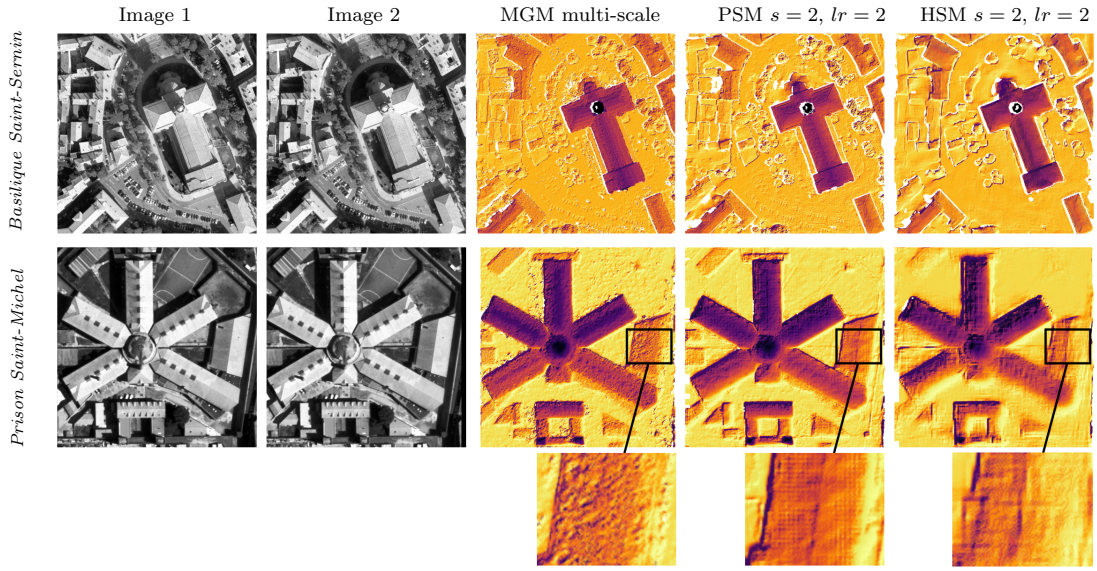


Figure 9.10: Small baseline experiments. Left to right: input pairs and output disparity. Moderate shading is applied to the disparity maps for better visualization. Undefined values are in white.

stereo vision [DR07, IPG+07].[2] Small baseline inputs are a common source of errors, because small changes in disparity can induce a large variation in the resulting depth (Equation (9.1)). At the same time, excessively large baselines reduce the amount of correspondences between the images, due to occlusions, and objects that appear and disappear from the scene. An adequate baseline must result from a compromise between these factors.

As shown in Figure 9.10, the disparity maps produced by multi-scale MGM are more complete. The PSM output appears less noisy, but some areas are affected by fine-scale checkerboard artifacts (see the detail in Figure 9.10). These fine-scale checkerboard artifacts are probably explained by the use of deconvolution layers in PSM [ODO16, SSK18] and have no connection with small baseline inputs. The HSM result is significantly worse with respect to the two previous methods: much less detailed, with large-scale artifacts and visibly smoothed contours, especially in *Prison Saint-Michel*.

9.6 Conclusion

This chapter reviewed the PSM and HSM convolutional neural networks for disparity estimation from an input stereo pair and investigated their applicability for satellite stereo reconstruction. The two methods were compared with a variant of the SGM algorithm, which is a classic matching strategy widely used for satellite images. We used pre-trained weights, fine-tuned using an aerial stereo matching benchmark. The quality of the disparity maps output by each method is assessed based on the subsequent surface models, which are evaluated using a lidar reference model.

The conducted experiments show that the deep learning methods provide higher accuracy than classic concurrent algorithms, and should therefore be preferred for satellite stereo

[2] The two pairs exhibit a B/H factor between 0.05 and 0.10, where B is the baseline and H is the distance between the scene and the camera system.

3D reconstruction under ideal conditions. However, these networks require additional effort to adjust the format of the input pairs and may produce more incomplete results in challenging or unusual scenarios (e.g., very distant acquisition dates or small baselines). It is critical that the rectified images emulate the training conditions. For optimal results, it is also best to adapt the size of the input images.

PSM provides remarkable robustness to image resolution and accuracy, especially for highly textured inputs like panchromatic images, but becomes impractical as the input size increases. Alternatively, HSM is much faster but loses detail sharpness depending on the input resolution. One network or the other may be more convenient depending on the application and type of input.

10 Conclusion

In this thesis we have addressed the problem of modeling digital surfaces from high-resolution optical satellite images. Instead of focusing on the classic stereo scenario with two input images, this thesis aims to facilitate the use of multiple satellite views, so that they can be easily plugged into novel multi-view methods or conventional stereo-based pipelines, which are currently the predominant approach and are presented in the opening chapter (Section 1.4.2).

Collections of multiple satellite images are becoming increasingly common. These collections can be composed of multi-date acquisitions; semi-simultaneous multi-image acquisitions such as those captured by the new small satellite constellations introduced in Section 1.3; or a mixture of the two cases, as small satellite constellations can provide very short revisit times. In this thesis we worked with multi-date WorldView-3 images and multi-image acquisitions from the SkySat constellation.

Part I. Geometric modeling of multi-view satellite imagery

A common problem of satellite image collections is that the geometry of the associated camera models is rarely consistent, due to the inaccuracy of internal measurements related to sensor orientation [GD03]. The first contribution of this thesis is a bundle adjustment methodology that corrects the RPC camera models typically used to represent the acquisition process of satellite optical sensors, so that consistent multi-view geometry with sub-meter accuracy is obtained. As a result, altitude estimates derived from different views are consistent and naturally registered in 3D space. In Chapter 3 we reviewed various methods for registering independent altitude models in 3D space, including bundle adjustment strategies, and highlighted that these methods usually employ a local affine camera approximation instead of working with the original RPCs. The previous implies that the corrected camera models are represented in a simplified form plus a complementary corrective function. This format does not facilitate the subsequent reuse of the corrected models for other tasks, such as dense 3D reconstruction. Motivated by this fact, our bundle adjustment methodology, proposed in Chapter 4, makes a direct use of the RPC models, and returns the corrected cameras also in RPC format. The corrected RPC cameras are obtained thanks to the direct composition of the original RPC functions with a rotation matrix, which compensates for errors related to sensor orientation. The corrected RPC model resulting from the composition is obtained by a regularized least squares fitting algorithm, which is described in Chapter 5, that can be used to encode other complex sensor models in a generic manner. For instance, we also use it to encode the physical sensor model associated with SAR images acquired by Sentinel-1.

The bundle adjustment methodology proposed in the first part of the thesis is also compared with geometry-based methods for registering independent altitude models, as done in [FdFML17]. When a sufficient number of tie points connecting all cameras is available, we show that the bundle adjustment approach for RPC correction performs better. Geometry-based registration methods are subject to erroneous altitude values produced by the dense stereo reconstruction process, whereas bundle adjustment provides robustness to inaccurate initial geometry estimates because it simultaneously optimizes camera models and scene geometry.

Part II. Applications of multi-view satellite imagery

The second part of the thesis presented different applications related to the generation of multi-view digital surface models (DSMs). The majority of the methods in this second part take advantage of the RPC bundle adjustment methodology proposed in the first part (Chapter 4) and show the practical advantages of working with multiple geometrically consistent satellite views. The open-source satellite stereo pipeline S2P [dFMLM+14a] was used in all chapters as a reference for dense stereo reconstruction.

In Chapter 6 we used the RPC bundle adjustment to correct the camera models of a time series of SkySat multi-image acquisitions observing the storage area of a coal terminal. The resulting camera models are then used with S2P to obtain highly consistent multi-view stereo DSMs at each date of the time series, where coal stockpile volume can be accurately measured by integrating altitude values above a terrain model. We validated our estimates with on-site measurements from the same time period, which reveal a high correlation between the two distributions.

In Chapter 7 we presented an image stitching method to create a single large-scale product from a fragmented multi-image SkySat acquisition, more precisely, from a strip of small images overlapping along the direction of satellite motion. The RPC bundle adjustment is used to ensure geometric consistency between the input camera models, that are then used with the RPC fitting algorithm detailed in Chapter 5 to construct a geometrically consistent RPC camera model (or perfect sensor geometry model [BLK09]) for the output mosaics. As a result, the large-scale products can be used to reconstruct large areas of interest with a single stereo pair of mosaics instead of multiple pairs of smaller images, as we demonstrated in practice with S2P.

In Chapter 8 we introduced a real multi-view method, the Sat-NeRF neural network, that uses self-supervised neural rendering techniques [MST+20] to learn a continuous representation of the geometry and appearance of a satellite scene. We trained the network with multi-date collections of WorldView-3 images, so both the architecture and cost function were adapted to distinguish shadows and transient objects from permanent structures; and the input sampling strategy was also tailored to work with RPC camera models previously corrected by bundle adjustment. Our experiments show that Sat-NeRF outperforms previous state-of-the-art neural rendering models for satellite images [DI21] and is able to produce DSMs of similar accuracy with respect to classic multi-view stereo reconstruction with S2P, with average numerical differences of the order of centimeters. Visual inspection of the DSMs revealed that there is still room for improvement, as Sat-NeRF is able to achieve greater level of geometry detail, but is affected by local irregularities.

Lastly, in Chapter 9 we continued investigating neural networks for DSM generation, focusing this time on the stereo matching step. In particular, we reviewed the PSM [CC18]

and HSM [YMHR19] networks and used them to replace the dense stereo matching of S2P, which by default uses the MGM variant [FdFM15] of the classic SGM algorithm [Hir07]. For both networks we used fine-tuned weights based on aerial images, which are semantically similar to high-resolution satellite images. We tested all methods with S2P and WorldView-3 image pairs as input, and verified that the neural networks achieved higher numerical accuracy than MGM. PSM provides the best level of detail, but is impractical for large input images, for which HSM is more efficient. Despite the promising results, we cannot conclude yet that neural networks are entirely ready to replace classic correlators because they are less generic, as they are often limited to predefined disparity ranges and highly sensitive to the rectification of input pairs.

Final remarks

Improving digital surface modeling from satellite imagery is a problem of general interest, affecting from image providers to geoanalytics companies, government space agencies or end customers. The research presented in this thesis follows a similar direction to that of the space agencies of the United States, France or Germany, which have published a large part of the cited literature and have their own satellite stereo pipelines [BAM18, MSY+20, KdSG13]. In August 2020, during the development of this thesis, the European Space Agency (ESA) started a project on multi-view satellite photogrammetry with neural radiance fields [DI21], which have been gaining enormous momentum in the field of 3D vision since then [MST+20] and motivated the Sat-NeRF methodology presented in Chapter 8. The French geoanalytics company Kayrros has also been actively involved in most of the presented methods, to provide input data and real use cases [MdFML+21, MEA+22, AMdF+21]. This context shows that the ideas discussed in this dissertation have the potential to make an immediate impact on industry and society. Handling multiple remote sensing images efficiently is set to be key in the present and future era of Earth observation, with the entire surface of our planet imaged several times a day and large-scale continuous monitoring becoming a reality.

Part III

Appendix

A Fundamentals of bundle adjustment optimization

This appendix overviews the fundamental concepts of numerical analysis behind the optimization algorithms employed to solve bundle adjustment problems [CCW19, TMHF00].

Bundle adjustment as nonlinear weighted least squares

Given a set of K 3D points $\boldsymbol{X}_k$ seen across M cameras $\mathcal{P}_m$ and denoting $\boldsymbol{x}_{mk}$ the 2D observation of $\boldsymbol{X}_k$ in the m-th image. The bundle adjustment problem aims at minimizing the reprojection distance $r_i = \|\boldsymbol{x}_{mk} - \mathcal{P}_m(\boldsymbol{X}_k)\|$, of all points in all cameras by jointly optimizing the parameters of the camera projection functions $\mathcal{P}_m$ and the positions of the 3D points $\boldsymbol{X}_k$

$$\sum_{k=1}^{K}\sum_{m=1}^{M} \|\boldsymbol{x}_{mk} - \mathcal{P}_m(\boldsymbol{X}_k)\|^2. \tag{A.1}$$

If each 3D point is seen by all cameras, the total number of 2D observations $\boldsymbol{x}_{mk}$ is $N = K \cdot M$.

Equation (A.1) defines a nonlinear weighted least squares problem. For the following discussion we rewrite it as a generic problem of the form

$$\min_{\mathbf{x}} \frac{1}{2}\sum_{i=1}^{N} w_i r_i^2(\mathbf{x}), \tag{A.2}$$

where r_i is a set of N nonlinear functions, $\mathbf{x} = (x_1, \ldots, x_L)$ is a vector of length L with all the parameters to estimate, and w_i are the weights. Denoting W the diagonal matrix with the weights w_i, $\mathbf{r} = (r_1, \ldots, r_N)$ the vector of nonlinear functions and omitting the dependency on $(\mathbf{x})$, the above expression rewrites as

$$\min_{\mathbf{x}} \frac{1}{2}\mathbf{r}^T W \mathbf{r}. \tag{A.3}$$

We will detail the fundamentals behind the Gauss–Newton algorithm, which iteratively finds the value of the variables $\mathbf{x}$ starting from an initial guess $\mathbf{x}^0$ by iterating

$$\mathbf{x}^{t+1} = \mathbf{x}^t - (J^T W J)^{-1} J^T W \mathbf{r}(\mathbf{x}^t), \tag{A.4}$$

where $J = (J)_{ij} = \frac{\partial r_i(\mathbf{x}^t)}{\partial x_j}$ is the Jacobian matrix of $\mathbf{r}(\mathbf{x}^t)$.

Newton's method

Given a twice differentiable function $f : \mathbb{R}^N \to \mathbb{R}$, we seek to solve the optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^N} f(\mathbf{x}). \tag{A.5}$$

This can be solved by gradient descent, which constructs a sequence $\{\mathbf{x}^t\}$ from an initial guess $\mathbf{x}^0$ by taking small steps $\delta\mathbf{x} = -\mu\mathbf{g}$ in the direction of the gradient $\mathbf{g}$,

$$\mathbf{x}^{t+1} = \mathbf{x}^t + \delta\mathbf{x}. \tag{A.6}$$

The step $\delta\mathbf{x}$ must be set small enough (by controlling μ) to assure that $f(\mathbf{x}^{t+1}) < f(\mathbf{x}^t)$ at each iteration until convergence.

Newton's method improves the gradient descent strategy by determining the step at each iteration $\mathbf{x}^t$ by solving a second-order Taylor approximation of f

$$f(\mathbf{x}+\delta\mathbf{x}) \simeq f(\mathbf{x}) + \mathbf{g}^T\delta\mathbf{x} + \frac{1}{2}\delta\mathbf{x}^T H\delta\mathbf{x}, \tag{A.7}$$

where $\mathbf{g}$ and H represent the gradient and Hessian of f, respectively. The local minimum of (A.7) is obtained by setting its differential to 0. This yields the optimal step

$$\delta\mathbf{x} = -H^{-1}\mathbf{g}, \tag{A.8}$$

which is known as the Newton step. Based on this definition, Newton's method iterates according to (A.6) and estimating $\delta\mathbf{x}$ at each iteration as defined in (A.8).

Newton's method is faster than a classic gradient descent algorithm. Unfortunately it can fail in many ways: it may converge to a saddle point rather than a minimum or it may be unstable if the Hessian is close to a singular matrix. To enforce stability, a *damped* version of (A.8) can be used, i.e.

$$\delta\mathbf{x} = -(H+\lambda I)^{-1}\mathbf{g}, \tag{A.9}$$

where λ is some weighting factor and I is the identity.

Gauss-Newton method

When applying Newton's method to a nonlinear weighted least squares problem as (A.3), the gradient and Hessian of f can be expressed in terms of the Jacobian matrix of $\mathbf{r}(\mathbf{x})$. Given

$$J = (J)_{ij} = \frac{\partial r_i(\mathbf{x})}{\partial x_j}, \tag{A.10}$$

the gradient writes as

$$\mathbf{g} = J^T W\mathbf{r}, \tag{A.11}$$

and the Hessian as

$$H = J^T WJ + S, \tag{A.12}$$

where $(S)_{jk} = \sum_{i=1}^{N} w_i r_i(\mathbf{x}) \frac{\partial^2 r_i(\mathbf{x})}{\partial x_j \partial x_k}$. The S term is usually very small compared to $J^T WJ$ if the current $\mathbf{x}$ is not too far from a minimum, so it can be omitted leading to the Gauss-Newton approximation of the problem, which consequently defines $\delta\mathbf{x}$ as

$$\delta\mathbf{x} = -(J^T WJ)^{-1} J^T W\mathbf{r}. \tag{A.13}$$

In the same way as (A.9), a dampened Gauss-Newton step can also be defined

$$\delta\mathbf{x} = -(J^T WJ + \lambda I)^{-1} J^T W\mathbf{r}. \tag{A.14}$$

Equation (A.14) is used by the Levenberg-Marquardt algorithm, which is a classic and well-known iterative algorithm to solve bundle adjustment problems.

Damped Newton methods use λ to regulate a trade-off between the Newton and gradient descent directions, i.e. if $\lambda \to \infty$ they behave as a gradient descent ($\delta\mathbf{x} \propto -\mathbf{g}$). λ can be chosen to limit the step to a dynamically chosen maximum size (Trust Region methods), or manipulated more heuristically, to shorten the step if the prediction is poor (Levenberg-Marquardt methods).

Construction of the Jacobian matrix

The Jacobian J of a bundle adjustment problem corresponds to a large sparse matrix [CCW19]. This can be easily illustrated with a toy example. Suppose we have $M = 3$ cameras and $K = 4$ 3D points. For simplicity we assume that each point is seen across all cameras. J can be expressed as

$$J = \begin{array}{c} \\ \text{number of 2D observations} \\ \downarrow \end{array} \begin{array}{c} \overset{\text{number of parameters to estimate}}{\longrightarrow} \\ \begin{array}{cccccccc} & c_1 & c_2 & c_3 & p_1 & p_2 & p_3 & p_4 \end{array} \\ \begin{array}{c} (p_1, c_1) \\ (p_1, c_2) \\ (p_1, c_3) \\ (p_2, c_1) \\ (p_2, c_2) \\ (p_2, c_3) \\ (p_3, c_1) \\ (p_3, c_2) \\ (p_3, c_3) \\ (p_4, c_1) \\ (p_4, c_2) \\ (p_4, c_3) \end{array} \begin{pmatrix} A_1 & 0 & 0 & B_1 & 0 & 0 & 0 \\ 0 & A_2 & 0 & B_2 & 0 & 0 & 0 \\ 0 & 0 & A_3 & B_3 & 0 & 0 & 0 \\ A_4 & 0 & 0 & 0 & B_4 & 0 & 0 \\ 0 & A_5 & 0 & 0 & B_5 & 0 & 0 \\ 0 & 0 & A_6 & 0 & B_6 & 0 & 0 \\ A_7 & 0 & 0 & 0 & 0 & B_7 & 0 \\ 0 & A_8 & 0 & 0 & 0 & B_8 & 0 \\ 0 & 0 & A_9 & 0 & 0 & B_9 & 0 \\ A_{10} & 0 & 0 & 0 & 0 & 0 & B_{10} \\ 0 & A_{11} & 0 & 0 & 0 & 0 & B_{11} \\ 0 & 0 & A_{12} & 0 & 0 & 0 & B_{12} \end{pmatrix} \end{array}, \tag{A.15}$$

where the i-th row contains the partial derivatives of the reprojection error r_i related to the i-th feature observation, which corresponds to the k-th point seen in the m-th camera, i.e. (p_k, c_m).

The vector of parameters to optimize $\mathbf{x}$ usually follows a camera-structure order, $\mathbf{x} = [c, p]$, where c and p are the sets of camera parameters and 3D point coordinates to estimate, respectively. Note that since r_i is a norm of a residual we can write $r_i^2 = rx_i^2 + ry_i^2$. Thus the first block of columns in J contains the partial derivatives with respect to the camera parameters c_m for each dimension of the image plane, i.e. $A_i = \frac{\partial r_i}{\partial c_m} = (\frac{\partial rx_i}{\partial c_m}; \frac{\partial ry_i}{\partial c_m})$, while the second block of columns contains the partial derivatives with respect to the coordinates of the 3D points p_k, i.e. $B_i = \frac{\partial r_i}{\partial p_k} = (\frac{\partial rx_i}{\partial p_k}; \frac{\partial ry_i}{\partial p_k})$. Therefore J has a total size equal to

$$(2 \cdot N) \times (q_c \cdot M + q_p \cdot K), \tag{A.16}$$

where

- K is the number of 3D points, 4 in this example.
- M is the number of cameras, 3 in this example.
- N is the number of 2D observations. $N = K \cdot M = 12$ in this example, since we assume that each point is observed in all cameras.
- q_c is the number of parameters to optimize per camera and defines the number of columns of each block A_i in (A.15). E.g. $q_c = 3$ if the camera parameters to estimate consist of 3 Euler angles related to the camera orientation, as in Chapter 4.
- q_p is the number of parameters to optimize per 3D point and defines the number of columns of each block B_i in (A.15). E.g. $q_p = 3$ if all 3D coordinates of each point are optimized.

On the finite differences approximation and sparsity structure

Computing the Jacobian matrix J of a cost function $f(\mathbf{x})$ can be very difficult for large sparse problems as in bundle adjustment. Finite differences are a popular approach to approximate J in an simpler and faster manner. Instead of explicitly computing the partial derivatives, these are approximated as a difference between the values of $f(\mathbf{x})$, e.g. $f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x})$ for a simple forward difference with step h. Under this approximation, finite differences offer the advantage that the cost function can be treated as a black box in the optimization process.

The task can be further simplified by providing a *sparsity structure* that indicates where to place each approximated partial derivative in J [CPR74, AMB+94]. The sparsity structure of J, denoted as J_S, corresponds to the binarization of J into zeros and non-zeros. J_S can be used to minimize the number of function evaluations that are necessary to compute J. Instead of constructing each column of J individually, J_S allows to group the columns beforehand into structurally orthogonal groups, where the columns in each group do not have non-zeros in the same row position. This is useful to speed up the process, since it allows the calculation of the partial derivatives of each group at once.

B Disparity estimation networks: Extras

B.1 List of DFC2019 stereo pairs

This section lists the image pairs of the DFC2019 dataset [BFC+19] used in Chapter 9, ordered by the heuristic criterion of Section 9.5.2 [FdFML17]. Distances between acquisition dates are in days modulo one year and Θ is the intersection angle between views.

JAX 004

pair	image id 1	image id 2	angle Θ [deg]	date dist. [days]
01	21JAN15WV031100015JAN21161253	21JAN15WV031100015JAN21161308	8.3	0.00
02	02MAY15WV031100015MAY02161943	01MAY15WV031200015MAY01160357	43.1	1.01
03	02MAY15WV031100015MAY02161943	26APR15WV031200015APR26162435	15.3	6.00
04	19APR15WV031100015APR19161439	26APR15WV031200015APR26162435	24.9	7.01
05	19APR15WV031100015APR19161439	01MAY15WV031200015MAY01160357	36.1	11.99
06	19APR15WV031100015APR19161439	02MAY15WV031100015MAY02161943	10.4	13.00
07	27DEC14WV031100014DEC27161109	14DEC14WV031100014DEC14160402	28.7	13.00
08	15FEB15WV031200015FEB15161208	21JAN15WV031100015JAN21161253	40.9	25.00
09	27DEC14WV031100014DEC27161109	21JAN15WV031100015JAN21161253	11.8	25.00
10	27DEC14WV031100014DEC27161109	21JAN15WV031100015JAN21161308	20.1	25.00
11	14DEC14WV031100014DEC14160402	21JAN15WV031100015JAN21161253	39.4	38.01
12	15JUN15WV031100015JUN15161248	02MAY15WV031100015MAY02161943	26.6	44.00
13	15JUN15WV031100015JUN15161248	01MAY15WV031200015MAY01160357	16.5	45.01
14	15JUN15WV031100015JUN15161248	26APR15WV031200015APR26162435	34.0	49.99
15	27DEC14WV031100014DEC27161109	15FEB15WV031200015FEB15161208	29.1	50.00
16	01NOV15WV031100015NOV01161954	27DEC14WV031100014DEC27161109	25.2	55.99
17	15JUN15WV031100015JUN15161248	19APR15WV031100015APR19161439	20.5	57.00
18	15FEB15WV031200015FEB15161208	19APR15WV031100015APR19161439	37.2	63.00
19	14DEC14WV031100014DEC14160402	15FEB15WV031200015FEB15161208	11.1	63.01
20	15FEB15WV031200015FEB15161208	26APR15WV031200015APR26162435	31.8	70.01
21	15FEB15WV031200015FEB15161208	01MAY15WV031200015MAY01160357	29.4	74.99
22	15FEB15WV031200015FEB15161208	02MAY15WV031100015MAY02161943	36.1	76.01
23	01NOV15WV031100015NOV01161954	21JAN15WV031100015JAN21161253	35.5	81.00
24	01NOV15WV031100015NOV01161954	21JAN15WV031100015JAN21161308	43.2	81.00
25	19APR15WV031100015APR19161439	21JAN15WV031100015JAN21161308	13.9	88.00
26	19APR15WV031100015APR19161439	21JAN15WV031100015JAN21161253	7.9	88.00
27	26APR15WV031200015APR26162435	21JAN15WV031100015JAN21161308	27.3	95.01
28	26APR15WV031200015APR26162435	21JAN15WV031100015JAN21161253	21.1	95.01
29	21JAN15WV031100015JAN21161253	01MAY15WV031200015MAY01160357	43.7	99.99
30	02MAY15WV031100015MAY02161943	21JAN15WV031100015JAN21161308	13.8	101.00

Table B.1: List of 30 stereo pairs, DFC2019 Jacksonville area 004.

JAX 068

pair	image id 1	image id 2	angle Θ [deg]	date dist. [days]
01	05OCT14WV031100014OCT05160149	05OCT14WV031100014OCT05160138	7.2	0.00
02	21JAN15WV031100015JAN21161253	21JAN15WV031100015JAN21161308	8.3	0.00
03	02MAY15WV031100015MAY02161943	01MAY15WV031200015MAY01160357	43.1	1.01
04	01NOV15WV031100015NOV01161954	30OCT14WV031100014OCT30155732	15.3	2.02
05	02MAY15WV031100015MAY02161943	26APR15WV031200015APR26162435	15.3	6.00
06	05OCT14WV031100014OCT05160149	11OCT14WV031100014OCT11155720	27.5	6.00
07	27JAN15WV031100015JAN27160845	21JAN15WV031100015JAN21161308	29.6	6.00
08	11OCT14WV031100014OCT11155720	05OCT14WV031100014OCT05160138	34.0	6.00
09	27JAN15WV031100015JAN27160845	21JAN15WV031100015JAN21161253	21.9	6.00
10	19APR15WV031100015APR19161439	26APR15WV031200015APR26162435	24.9	7.01
11	18OCT14WV031100014OCT18160722	11OCT14WV031100014OCT11155720	25.1	7.01
12	19APR15WV031100015APR19161439	01MAY15WV031200015MAY01160357	36.1	11.99
13	18OCT14WV031100014OCT18160722	30OCT14WV031100014OCT30155732	25.3	11.99
14	19APR15WV031100015APR19161439	02MAY15WV031100015MAY02161943	10.4	13.00
15	18OCT14WV031100014OCT18160722	05OCT14WV031100014OCT05160149	18.6	13.00
16	18OCT14WV031100014OCT18160722	05OCT14WV031100014OCT05160138	24.8	13.00
17	27DEC14WV031100014DEC27161109	14DEC14WV031100014DEC14160402	28.7	13.00
18	18OCT14WV031100014OCT18160722	01NOV15WV031100015NOV01161954	19.6	14.01
19	21MAY15WV031200015MAY21161849	02MAY15WV031100015MAY02161943	15.0	19.00
20	30OCT14WV031100014OCT30155732	11OCT14WV031100014OCT11155720	8.4	19.00
21	27JAN15WV031100015JAN27160845	15FEB15WV031200015FEB15161208	22.0	19.00
22	21MAY15WV031200015MAY21161849	01MAY15WV031200015MAY01160357	33.0	20.01
23	15JUN15WV031100015JUN15161248	05JUL15WV031100015JUL05162954	37.8	20.01
24	01NOV15WV031100015NOV01161954	11OCT14WV031100014OCT11155720	21.8	21.02
25	21MAY15WV031200015MAY21161849	15JUN15WV031100015JUN15161248	17.6	25.00
26	21MAY15WV031200015MAY21161849	26APR15WV031200015APR26162435	16.6	25.00
27	05OCT14WV031100014OCT05160149	30OCT14WV031100014OCT30155732	21.8	25.00
28	30OCT14WV031100014OCT30155732	05OCT14WV031100014OCT05160138	27.4	25.00
29	15FEB15WV031200015FEB15161208	21JAN15WV031100015JAN21161253	40.9	25.00
30	27DEC14WV031100014DEC27161109	21JAN15WV031100015JAN21161253	11.8	25.00

Table B.2: List of 30 stereo pairs, DFC2019 Jacksonville area 068.

JAX 214

pair	image id 1	image id 2	angle Θ [deg]	date dist. [days]
01	21JAN15WV031100015JAN21161243	21JAN15WV031100015JAN21161253	8.2	0.00
02	05OCT14WV031100014OCT05160149	05OCT14WV031100014OCT05160138	7.2	0.00
03	21JAN15WV031100015JAN21161253	21JAN15WV031100015JAN21161308	8.3	0.00
04	21JAN15WV031100015JAN21161243	21JAN15WV031100015JAN21161308	16.5	0.00
05	01NOV15WV031100015NOV01161954	01NOV15WV031100015NOV01162034	27.1	0.00
06	02MAY15WV031100015MAY02161943	01MAY15WV031200015MAY01160357	43.1	1.01
07	01NOV15WV031100015NOV01161954	30OCT14WV031100014OCT30155732	15.3	2.02
08	30OCT14WV031100014OCT30155732	01NOV15WV031100015NOV01162034	17.4	2.02
09	15FEB15WV031200015FEB15161208	18FEB16WV031200016FEB18164007	35.5	3.02
10	15FEB15WV031200015FEB15161208	11FEB16WV031100016FEB11163042	38.7	3.99
11	02MAY15WV031100015MAY02161943	26APR15WV031200015APR26162435	15.3	6.00
12	05OCT14WV031100014OCT05160149	11OCT14WV031100014OCT11155720	27.5	6.00
13	27JAN15WV031100015JAN27160845	21JAN15WV031100015JAN21161308	29.6	6.00
14	11OCT14WV031100014OCT11155720	05OCT14WV031100014OCT05160138	34.0	6.00
15	27JAN15WV031100015JAN27160845	21JAN15WV031100015JAN21161253	21.9	6.00
16	27JAN15WV031100015JAN27160845	21JAN15WV031100015JAN21161243	15.1	6.00
17	11FEB16WV031100016FEB11163042	18FEB16WV031200016FEB18164007	24.5	7.01
18	19APR15WV031100015APR19161439	26APR15WV031200015APR26162435	24.9	7.01
19	18OCT14WV031100014OCT18160722	11OCT14WV031100014OCT11155720	25.1	7.01
20	05OCT14WV031100014OCT05160149	25SEP15WV031100015SEP25163525	42.2	9.98
21	19APR15WV031100015APR19161439	01MAY15WV031200015MAY01160357	36.1	11.99
22	18OCT14WV031100014OCT18160722	30OCT14WV031100014OCT30155732	25.3	11.99
23	19APR15WV031100015APR19161439	02MAY15WV031100015MAY02161943	10.4	13.00
24	18OCT14WV031100014OCT18160722	05OCT14WV031100014OCT05160149	18.6	13.00
25	18OCT14WV031100014OCT18160722	05OCT14WV031100014OCT05160138	24.8	13.00
26	27DEC14WV031100014DEC27161109	14DEC14WV031100014DEC14160402	28.7	13.00
27	18OCT14WV031100014OCT18160722	01NOV15WV031100015NOV01161954	19.6	14.01
28	18OCT14WV031100014OCT18160722	01NOV15WV031100015NOV01162034	22.3	14.01
29	27JAN15WV031100015JAN27160845	11FEB16WV031100016FEB11163042	21.0	15.02
30	11OCT14WV031100014OCT11155720	25SEP15WV031100015SEP25163525	43.9	15.97

Table B.3: List of 30 stereo pairs, DFC2019 Jacksonville area 214.

JAX 260

pair	image id 1	image id 2	angle Θ [deg]	date dist. [days]
01	21JAN15WV031100015JAN21161243	21JAN15WV031100015JAN21161253	8.2	0.00
02	02MAY15WV031100015MAY02161943	01MAY15WV031200015MAY01160357	43.1	1.01
03	01NOV15WV031100015NOV01161954	30OCT14WV031100014OCT30155732	15.3	2.02
04	15FEB15WV031200015FEB15161208	18FEB16WV031200016FEB18164007	35.5	3.02
05	15FEB15WV031200015FEB15161208	11FEB16WV031100016FEB11163042	38.7	3.99
06	02MAY15WV031100015MAY02161943	26APR15WV031200015APR26162435	15.3	6.00
07	27JAN15WV031100015JAN27160845	21JAN15WV031100015JAN21161253	21.9	6.00
08	27JAN15WV031100015JAN27160845	21JAN15WV031100015JAN21161243	15.1	6.00
09	11FEB16WV031100016FEB11163042	18FEB16WV031200016FEB18164007	24.5	7.01
10	19APR15WV031100015APR19161439	26APR15WV031200015APR26162435	24.9	7.01
11	19APR15WV031100015APR19161439	01MAY15WV031200015MAY01160357	36.1	11.99
12	18OCT14WV031100014OCT18160722	30OCT14WV031100014OCT30155732	25.3	11.99
13	19APR15WV031100015APR19161439	02MAY15WV031100015MAY02161943	10.4	13.00
14	18OCT14WV031100014OCT18160722	05OCT14WV031100014OCT05160138	24.8	13.00
15	27DEC14WV031100014DEC27161109	14DEC14WV031100014DEC14160402	28.7	13.00
16	18OCT14WV031100014OCT18160722	01NOV15WV031100015NOV01161954	19.6	14.01
17	27JAN15WV031100015JAN27160845	11FEB16WV031100016FEB11163042	21.0	15.02
18	21MAY15WV031200015MAY21161849	02MAY15WV031100015MAY02161943	15.0	19.00
19	27JAN15WV031100015JAN27160845	15FEB15WV031200015FEB15161208	22.0	19.00
20	21MAY15WV031200015MAY21161849	01MAY15WV031200015MAY01160357	33.0	20.01
21	21JAN15WV031100015JAN21161243	11FEB16WV031100016FEB11163042	6.1	21.01
22	27JAN15WV031100015JAN27160845	18FEB16WV031200016FEB18164007	32.9	22.02
23	21MAY15WV031200015MAY21161849	26APR15WV031200015APR26162435	16.6	25.00
24	30OCT14WV031100014OCT30155732	05OCT14WV031100014OCT05160138	27.4	25.00
25	15FEB15WV031200015FEB15161208	21JAN15WV031100015JAN21161253	40.9	25.00
26	21JAN15WV031100015JAN21161243	15FEB15WV031200015FEB15161208	32.7	25.00
27	27DEC14WV031100014DEC27161109	21JAN15WV031100015JAN21161253	11.8	25.00
28	01NOV15WV031100015NOV01161954	05OCT14WV031100014OCT05160138	12.3	27.01
29	21JAN15WV031100015JAN21161253	18FEB16WV031200016FEB18164007	27.6	28.02
30	21JAN15WV031100015JAN21161243	18FEB16WV031200016FEB18164007	24.1	28.02

Table B.4: List of 30 stereo pairs, DFC2019 Jacksonville area 260.

B.2 Completeness values

This section provides the completeness percentage associated with the experiments of Section 9.5.4 (panchromatic inputs) and 9.5.4 (RGB inputs), shown in Table B.5 and Table B.6 respectively. The definition of completeness is given in Section 9.5.3.

Panchromatic inputs

		Single pair completeness [%] / NaNs [%] / Successful pairs			
Area index		004	068	214	260
MGM	$s = 1, lr = 1$	43.62 / 32.01 / 24	61.64 / 23.63 / 29	53.67 / 31.23 / 24	41.03 / 33.30 / 24
MGM multi	$s = 1, lr = 1$	44.64 / 34.63 / 18	60.33 / 25.74 / 28	53.61 / 31.95 / 23	39.75 / 34.48 / 21
PSM	$s = 1, lr = 1$	51.38 / 33.17 / 12	63.41 / 25.00 / 28	53.13 / 34.23 / 23	42.95 / 33.94 / 16
HSM	$s = 1, lr = 1$	39.78 / 34.64 / 20	52.93 / 28.90 / 28	43.42 / 34.89 / 23	35.87 / 31.72 / 24
PSM	$s = 2, lr = 2$	50.51 / 38.61 / 12	64.30 / 28.34 / 28	55.01 / 36.88 / 23	45.43 / 38.33 / 14
HSM	$s = 2, lr = 2$	45.06 / 37.63 / 16	61.79 / 28.09 / 28	52.98 / 35.40 / 23	43.97 / 32.98 / 19
PSM	$s = 3, lr = 3$	54.09 / 37.87 / 12	61.10 / 30.33 / 26	54.54 / 36.89 / 18	45.14 / 38.37 / 11
HSM	$s = 3, lr = 3$	48.83 / 37.15 / 12	61.32 / 28.88 / 28	52.90 / 36.14 / 23	45.02 / 34.49 / 18

		Multi-pair completeness [%] / NaNs [%]			
Area index		004	068	214	260
MGM	$s = 1, lr = 1$	67.06 / 0.65	78.65 / 0.62	74.28 / 0.39	64.59 / 0.69
MGM multi	$s = 1, lr = 1$	67.11 / 0.90	78.47 / 0.68	74.43 / 0.55	64.23 / 0.78
PSM	$s = 1, lr = 1$	67.71 / 2.59	82.14 / 0.06	74.00 / 0.28	65.38 / 1.08
HSM	$s = 1, lr = 1$	54.39 / 1.38	73.64 / 0.11	61.56 / 0.34	50.19 / 1.04
PSM	$s = 2, lr = 2$	65.30 / 6.13	83.50 / 0.25	77.95 / 1.10	68.52 / 3.39
HSM	$s = 2, lr = 2$	67.53 / 1.52	81.89 / 0.13	73.39 / 0.48	65.53 / 1.16
PSM	$s = 3, lr = 3$	70.15 / 7.08	83.11 / 0.22	76.50 / 1.26	67.26 / 4.72
HSM	$s = 3, lr = 3$	68.36 / 3.13	81.01 / 0.48	74.62 / 0.65	68.02 / 0.77

Table B.5: Quantitative results using panchromatic images as input. Equivalent to Table 9.1, with completeness in substitution of altitude MAE. The best completeness values are highlighted in yellow.

RGB inputs

		Single pair completeness (diff. w.r.t. PAN) [%] / NaNs [%]			
Area index		004	068	214	260
MGM multi	$s = 1, lr = 1$	58.11 (+13.47) / 11.19	68.34 (+8.01) / 14.11	56.90 (+3.29) / 20.04	48.61 (+8.86) / 15.85
PSM	$s = 2, lr = 2$	53.32 (+2.81) / 27.69	64.89 (+0.59) / 26.25	52.50 (−2.50) / 36.12	47.20 (+1.77) / 30.81
HSM	$s = 2, lr = 2$	46.48 (+1.42) / 31.66	61.50 (−0.29) / 28.42	51.38 (−1.60) / 35.67	43.23 (−0.74) / 32.10

		Multi-pair completeness (diff. w.r.t. PAN) [%] / NaNs [%]			
Area index		004	068	214	260
MGM multi	$s = 1, lr = 1$	66.27 (−0.84) / 0.68	78.71 (+0.24) / 0.58	70.25 (−4.18) / 0.21	62.67 (−1.56) / 0.59
PSM	$s = 2, lr = 2$	65.06 (−0.24) / 1.36	82.55 (−0.95) / 0.33	72.69 (−5.26) / 1.22	65.93 (−2.59) / 1.76
HSM	$s = 2, lr = 2$	66.82 (−0.71) / 0.71	81.96 (+0.07) / 0.11	71.82 (−1.57) / 0.62	62.21 (−3.32) / 0.88

Table B.6: Quantitative results using RGB images as input. Equivalent to Table 9.2, with completeness in substitution of altitude MAE. The best completeness values are highlighted in yellow. The percentage difference with respect to the equivalent PAN experiment is shown in parentheses.

Bibliography

[AA20] Saif Aati and Jean-Philippe Avouac. Optimization of optical image geometric modeling, application to topography extraction and topographic change measurements using PlanetScope and SkySat imagery. *Remote Sensing*, 12(20):3418, 2020. https://doi.org/10.3390/rs12203418.

[AEdFF20] Jérémy Anger, Thibaud Ehret, Carlo de Franchis, and Gabriele Facciolo. Fast and accurate multi-frame super-resolution of satellite images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 5-1-2020:57–64, 2020. https://doi.org/10.5194/isprs-annals-V-1-2020-57-2020.

[AGA08] Vicente Arévalo, Javier González, and Gregorio Ambrosio. Shadow detection in colour high-resolution satellite images. *International Journal of Remote Sensing*, 29(7):1945–1963, 2008. https://doi.org/10.1080/01431160701395302.

[AM15] Cesar Arango and Andrés C. Morales. Comparison between multicopter UAV and total station for estimating stockpile volumes. *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40-1/W4:131–135, 2015. https://doi.org/10.5194/isprsarchives-XL-1-W4-131-2015.

[AMB+94] Brett M Averick, Jorge J Moré, Christian H Bischof, Alan Carle, and Andreas Griewank. Computing large sparse Jacobian matrices using automatic differentiation. *SIAM Journal on Scientific Computing*, 15(2):285–294, 1994. https://doi.org/10.1137/0915020.

[AMdF+21] Roland Akiki, Roger Marí, Carlo de Franchis, Jean-Michel Morel, and Gabriele Facciolo. Robust rational polynomial camera modelling for SAR and pushbroom imaging. In *2021 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 7908–7911, 2021. https://doi.org/10.1109/IGARSS47720.2021.9554583.

[AQ22] Hessah Albanwan and Rongjun Qin. Fine-tuning deep learning models for stereo matching using results from semi-global matching. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 5-2-2022:39–46, 2022. https://doi.org/10.5194/isprs-annals-V-2-2022-39-2022.

[ASM+12] Aleksandr Aravkin, Michael Styer, Zachary Moratto, Ara Nefian, and Michael Broxton. Student's T robust bundle adjustment algorithm. In *2012 19th IEEE International Conference on Image Processing*, pages 1757–1760, 2012. https://doi.org/10.1109/ICIP.2012.6467220.

[BAM18] Ross A Beyer, Oleg Alexandrov, and Scott McMichael. The Ames Stereo Pipeline: NASA's open source software for deriving and processing terrain data. *Earth and Space Science*, 5(9):537–548, 2018. https://doi.org/10.1029/2018EA000409.

[BCL99] Mary Ann Branch, Thomas F Coleman, and Yuying Li. A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. *SIAM Journal on Scientific Computing*, 21(1):1–23, 1999. https://doi.org/10.1137/S1064827595289108.

[BdKR18] Ksenia Bittner, Pablo d'Angelo, Marco Körner, and Peter Reinartz. DSM-to-LoD2: Spaceborne stereo digital surface model refinement. *Remote Sensing*, 10(12):1926, 2018. https://doi.org/10.3390/rs10121926.

[Ber04] Frank Vanden Berghen. Levenberg-Marquardt algorithms vs Trust Region algorithms, 2004. http://www.applied-mathematics.net/LMvsTR/LMvsTR.pdf.

[BFC+19] Marc Bosch, Kevin Foster, Gordon Christie, Sean Wang, Gregory D Hager, and Myron Brown. Semantic stereo for incidental satellite images. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1524–1532, 2019. https://doi.org/10.1109/WACV.2019.00167.

[BFS18] Thibaud Briand, Gabriele Facciolo, and Javier Sánchez. Improvements of the inverse compositional algorithm for parametric motion estimation. *Image Processing On Line*, 8:435–464, 2018. https://doi.org/10.5201/ipol.2018.222.

[BJIC17] Sungho Byun, Keonhwa Jung, Sungbin Im, and Minho Chang. Registration of 3D scan data using image reprojection. *International Journal of Precision Engineering and Manufacturing*, 18(9):1221–1229, 2017. https://doi.org/10.1007/s12541-017-0143-z.

[BK00] Gary Bradski and Adrian Kaehler. OpenCV. *Dr. Dobb's journal of software tools*, 2000.

[BKFR19] Ksenia Bittner, Marco Körner, Friedrich Fraundorfer, and Peter Reinartz. Multi-task cGAN for simultaneous spaceborne DSM refinement and roof-type classification. *Remote Sensing*, 11(11):1262, 2019. https://doi.org/10.3390/rs11111262.

[BKHB16] Marc Bosch, Zachary Kurtz, Shea Hagstrom, and Myron Brown. A multiple view stereo benchmark for satellite imagery. In *2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–9, 2016. https://doi.org/10.1109/AIPR.2016.8010543.

[BLK09] Simon Baillarin, Laurent Lebegue, and Philippe Kubik. Pleiades-HR system qualification: A focus on ground processing and image products performances, a few months before launch. 1:I–76–I–79, 2009. https://doi.org/10.1109/IGARSS.2009.5416940.

[BM92] Paul J Besl and Neil D McKay. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992. https://doi.org/10.1109/34.121791.

[BM01] Simon Baker and Iain Matthews. Equivalence and efficiency of image alignment algorithms. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages 1090–1097, 2001. https://doi.org/10.1109/CVPR.2001.990652.

[BMT+21] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5835–5844, 2021. https://doi.org/10.1109/ICCV48922.2021.00580.

[BNM+09] Michael J Broxton, Ara V Nefian, Zachary Moratto, Taemin Kim, Michael Lundy, and Aleksandr V Segal. 3D lunar terrain reconstruction from Apollo images. In *Advances in Visual Computing*, pages 710–719, 2009. https://doi.org/10.1007/978-3-642-10331-5_66.

[BRK19] Ksenia Bittner, Peter Reinartz, and Marco Körner. Late or earlier information fusion from depth and spectral data? Large-scale digital surface model refinement by hybrid-cGAN. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1471–1478, 2019. https://doi.org/10.1109/CVPRW.2019.00188.

[BSAH21] Shashank Bhushan, David Shean, Oleg Alexandrov, and Scott Henderson. Automated digital elevation model (DEM) generation from very-high-resolution Planet SkySat triplet stereo and video imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 173:151–165, 2021. https://doi.org/10.1016/j.isprsjprs.2020.12.012.

[BXS+20] Sai Bi, Zexiang Xu, Pratul Srinivasan, Ben Mildenhall, Kalyan Sunkavalli, Miloš Hašan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. Neural reflectance fields for appearance acquisition. *arXiv preprint arXiv:2008.03824*, 2020. https://arxiv.org/abs/2008.03824.

[CC18] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5410–5418, 2018. https://doi.org/10.1109/CVPR.2018.00567.

[CC20] CMLA and CNES. S2P - Satellite Stereo Pipeline, 2020. https://github.com/cmla/s2p.

[CCW19] Yu Chen, Yisong Chen, and Guoping Wang. Bundle adjustment revisited. *arXiv preprint arXiv:1912.03858*, 2019. https://arxiv.org/abs/1912.03858.

[CMT20] Giacomo Curzi, Dario Modenini, and Paolo Tortora. Large constellations of small satellites: A survey of near future challenges and missions. *Aerospace*, 7(9):133, 2020. https://doi.org/10.3390/aerospace7090133.

[CPR74] Alan R Curtis, Michael JD Powell, and John K Reid. On the estimation of sparse Jacobian matrices. *IMA Journal of Applied Mathematics*, 13(1):117–119, 1974. https://doi.org/10.1093/imamat/13.1.117.

[CSC21] Anthony F Cannistra, David E Shean, and Nicoleta C Cristea. High-resolution CubeSat imagery and machine learning for detailed snow-covered area. *Remote Sensing of Environment*, 258:112399, 2021. https://doi.org/10.1016/j.rse.2021.112399.

[CSH17] Hainan Cui, Shuhan Shen, and Zhanyi Hu. Tracks selection for robust, efficient and scalable large-scale structure from motion. *Pattern Recognition*, 72:341–354, 2017. https://doi.org/10.1016/j.patcog.2017.08.002.

[Cur82] John C. Curlander. Location of spaceborne SAR imagery. *IEEE Transactions on Geoscience and Remote Sensing*, GE-20(3):359–364, 1982. https://doi.org/10.1109/TGRS.1982.350455.

[DDSS22] Loïc Dumas, Véronique Defonte, Yoann Steux, and Emmanuelle Sarrazin. Improving pairwise DSM with 3SGM: A semantic segmentation for SGM using an automatically refined neural network. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 5-2-2022:167–175, 2022. https://doi.org/10.5194/isprs-annals-V-2-2022-167-2022.

[dF15] Carlo de Franchis. *Earth Observation and Stereo Vision*. PhD thesis, Université Paris-Saclay (ComUE), 2015. http://www.theses.fr/2015SACLN002.

[dFMLGF15] Carlo de Franchis, Enric Meinhardt-Llopis, Daniel Greslou, and Gabriele Facciolo. Attitude refinement for orbiting pushbroom cameras: A simple polynomial fitting method. *Image Processing On Line*, 5:328–361, 2015. https://doi.org/10.5201/ipol.2015.146.

[dFMLM+14a] Carlo de Franchis, Enric Meinhardt-Llopis, Julien Michel, Jean-Michel Morel, and Gabriele Facciolo. An automatic and modular stereo pipeline for pushbroom images. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2-3:49–56, 2014. https://doi.org/10.5194/isprsannals-II-3-49-2014.

[dFMLM+14b] Carlo de Franchis, Enric Meinhardt-Llopis, Julien Michel, Jean-Michel Morel, and Gabriele Facciolo. On stereo-rectification of pushbroom images. In *2014 IEEE International Conference on Image Processing*

(ICIP), pages 5447–5451, 2014. https://doi.org/10.1109/ICIP.2014.7026102.

[DI21] Dawa Derksen and Dario Izzo. Shadow neural radiance fields for multi-view satellite photogrammetry. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1152–1161, 2021. https://doi.org/10.1109/CVPRW53098.2021.00126.

[dK12] Pablo d'Angelo and Georg Kuschk. Dense multi-view stereo from satellite imagery. In *2012 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 6944–6947, 2012. https://doi.org/10.1109/IGARSS.2012.6352565.

[dKR14] Pablo d'Angelo, Georg Kuschk, and Peter Reinartz. Evaluation of Skybox video and still image products. *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40-1:95–99, 2014. https://doi.org/10.5194/isprsarchives-XL-1-95-2014.

[DL16] Liuyun Duan and Florent Lafarge. Towards large-scale city reconstruction from satellites. In *Computer Vision – ECCV 2016*, pages 89–104, 2016. https://doi.org/10.1007/978-3-319-46454-1_6.

[DLZR22] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12872–12881, 2022. https://doi.org/10.1109/CVPR52688.2022.01254.

[dPM+20] Marie d'Autume, Alexandra Perry, Jean-Michel Morel, Enric Meinhardt-Llopis, and Gabriele Facciolo. Stockpile monitoring using linear shape-from-shading on PlanetScope imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 5-2-2020:427–434, 2020. https://doi.org/10.5194/isprs-annals-V-2-2020-427-2020.

[DR07] Julie Delon and Bernard Rougé. Small baseline stereovision. *Journal of Mathematical Imaging and Vision*, 28(3):209–223, 2007. https://doi.org/10.1007/s10851-007-0001-1.

[dR11] Pablo d'Angelo and Peter Reinartz. Semiglobal matching results on the ISPRS stereo matching benchmark. *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 38-4/W19:79–84, 2011. https://doi.org/10.5194/isprsarchives-XXXVIII-4-W19-79-2011.

[dR12] Pablo d'Angelo and Peter Reinartz. DSM based orientation of large stereo satellite image blocks. *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 39-B1:209–214, 2012. https://doi.org/10.5194/isprsarchives-XXXIX-B1-209-2012.

[dR21] Pablo d'Angelo and Peter Reinartz. Digital elevation models from stereo, video and multi-view imagery captured by small satellites. *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43-B2-2021:77–82, 2021. https://doi.org/10.5194/isprs-archives-XLIII-B2-2021-77-2021.

[dRM+14] Pablo d'Angelo, Cristian Rossi, Christian Minet, Michael Eineder, Michael Flory, and Irmgard Niemeyer. High resolution 3D earth observation data analysis for safeguards activities. In *Symposium on International Safeguards*, pages 1–8, 2014.

[DWM+] Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. DeepPruner: Learning efficient stereo matching via differentiable PatchMatch. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4383–4392. https://doi.org/10.1109/ICCV.2019.00448.

[EMF22] Thibaud Ehret, Roger Marí, and Gabriele Facciolo. NeRF, meet differential geometry! *arXiv preprint arXiv:2206.14938*, 2022. https://arxiv.org/abs/2206.14938.

[FdFM15] Gabriele Facciolo, Carlo de Franchis, and Enric Meinhardt. MGM: A significantly more global matching for stereovision. In *Proceedings of the British Machine Vision Conference (BMVC)*, number 90, pages 1–12, 2015. https://doi.org/10.5244/C.29.90.

[FdFML17] Gabriele Facciolo, Carlo de Franchis, and Enric Meinhardt-Llopis. Automatic 3D reconstruction from multi-date satellite images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1542–1551, 2017. https://doi.org/10.1109/CVPRW.2017.198.

[FDG06] Clive S Fraser, Gene Dial, and Jacek Grodecki. Sensor orientation via RPCs. *ISPRS Journal of Photogrammetry and Remote Sensing*, 60(3):182–194, 2006. https://doi.org/10.1016/j.isprsjprs.2005.11.001.

[FH05] Clive S Fraser and Harry B Hanley. Bias-compensated rpcs for sensor orientation of high-resolution satellite imagery. *Photogrammetric Engineering & Remote Sensing*, 71(8):909–915, 2005. https://doi.org/10.14358/PERS.71.8.909.

[FH15] Yasutaka Furukawa and Carlos Hernández. Multi-view stereo: A tutorial. *Foundations and Trends in Computer Graphics and Vision*, 9(1-2):1–148, 2015. https://doi.org/10.1561/0600000052.

[fPS12] International Society for Photogrammetry and Remote Sensing. ISPRS test project on urban classification, 3D building reconstruction and semantic labeling, 2012. https://www.isprs.org/education/benchmarks/UrbanSemLab/default.aspx.

[FRC+07] Tom Farr, Paul Rosen, Edward Caro, Robert Crippen, Riley Duren, Scott Hensley, Michael Kobrick, Mimi Paller, Ernesto Rodriguez,

Ladislav Roth, et al. The shuttle radar topography mission. *Reviews of geophysics*, 45(2), 2007. https://doi.org/10.1029/2005RG000183.

[FS11a] David Chin-Lung Fong and Michael Saunders. LSMR: An iterative algorithm for sparse least-squares problems. *SIAM Journal on Scientific Computing*, 33(5):2950–2971, 2011. https://doi.org/10.1137/10079687X.

[FS11b] David Chin-Lung Fong and Michael Saunders. LSMR software, 2011. https://web.stanford.edu/group/SOL/software/lsmr.

[FY04] Clive S Fraser and Takeshi Yamakawa. Insights into the affine model for high-resolution satellite sensor orientation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 58(5):275–288, 2004. https://doi.org/10.1016/j.isprsjprs.2003.10.001.

[GD03] Jacek Grodecki and Gene Dial. Block adjustment of high-resolution satellite images described by rational polynomials. *Photogrammetric Engineering & Remote Sensing*, 69(1):59–68, 2003. https://doi.org/10.14358/PERS.69.1.59.

[GF16] Ke Gong and Dieter Fritsch. A detailed study about digital surface model generation using high resolution satellite stereo imagery. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3-1:69–76, 2016. https://doi.org/10.5194/isprs-annals-III-1-69-2016.

[GF17] Ke Gong and Dieter Fritsch. Relative orientation and modified piecewise epipolar resampling for high resolution satellite images. *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42-1/W1:579–586, 2017. https://doi.org/10.5194/isprs-archives-XLII-1-W1-579-2017.

[GF18] Ke Gong and Dieter Fritsch. Point cloud and digital surface model generation from high resolution multiple view stereo satellite imagery. *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42-2:363–370, 2018. https://doi.org/10.5194/isprs-archives-XLII-2-363-2018.

[GF19] Ke Gong and Dieter Fritsch. DSM generation from high resolution multi-view stereo satellite imagery. *Photogrammetric Engineering & Remote Sensing*, 85(5):379–387, 2019. https://doi.org/10.14358/PERS.85.5.379.

[GFZ+20] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2492–2501, 2020. https://doi.org/10.1109/CVPR42600.2020.00257.

[GLJ21] Jian Gao, Jin Liu, and Shunping Ji. Rational polynomial camera model warping for deep learning based satellite multi-view stereo matching. In *2021 IEEE/CVF International Conference on Computer*

Vision (ICCV), pages 6128–6137, 2021. https://doi.org/10.1109/ICCV48922.2021.00609.

[GLU12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012. https://doi.org/10.1109/CVPR.2012.6248074.

[Gon21] Ke Gong. *New methods for 3D reconstructions using high resolution satellite data*. PhD thesis, University of Stuttgart, 2021. http://dx.doi.org/10.18419/opus-11453.

[GRFvGG22] Alvaro Gómez, Gregory Randall, Gabriele Facciolo, and Rafael von Gioi Grompone. An experimental comparison of multi-view stereo approaches on satellite images. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 707–716, 2022. https://doi.org/10.1109/WACV51458.2022.00078.

[Gro01] Jacek Grodecki. IKONOS stereo feature extraction - RPC approach. In *ASPRS Annual Conference*, 2001.

[HCZH19] Haiqing He, Ting Chen, Huaien Zeng, and Shengxiang Huang. Ground control point-free unmanned aerial vehicle-based photogrammetry for volume estimation of stockpiles carried on barges. *Sensors*, 19(16):3534, 2019. https://doi.org/10.3390/s19163534.

[Hir07] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, 2007. https://doi.org/10.1109/TPAMI.2007.1166.

[HO93] Per Christian Hansen and Dianne Prost O'Leary. The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM Journal on Scientific Computing*, 14(6):1487–1503, 1993. https://doi.org/10.1137/0914086.

[HPSM+00] Hank Heidt, Jordi Puig-Suari, Augustus Moore, Shinichi Nakasuka, and Robert Twiggs. CubeSat: A new generation of picosatellite for education and industry low-cost space experimentation. 2000.

[HSM+21] Peter Hedman, Pratul P Srinivasan, Ben Mildenhall, Jonathan T Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5855–5864, 2021. https://doi.org/10.1109/ICCV48922.2021.00582.

[HWG+20] Yilong Han, Shugen Wang, Danchao Gong, Yue Wang, Yuan Wang, and Xiaoliang Ma. State of the art in digital surface modelling from multi-view high-resolution satellite images. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 5-2-2020:351–356, 2020. https://doi.org/10.5194/isprs-annals-V-2-2020-351-2020.

[HZ04] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, second ed., 2004. https://doi.org/10.1017/CBO9780511811685.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. https://doi.org/10.1109/CVPR.2016.90.

[IPG+07] Laura Igual, Javier Preciozzi, Luis Garrido, Andrés Almansa, Vicent Caselles, and Bernard Rougé. Automatic low baseline stereo in urban areas. *Inverse Problems and Imaging*, 1(2):319–348, 2007. https://doi.org/10.3934/ipi.2007.1.319.

[JTA21] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting NeRF on a diet: Semantically consistent few-shot view synthesis. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5865–5874, 2021. https://doi.org/10.1109/ICCV48922.2021.00583.

[KdG+17] Georg Kuschk, Pablo d'Angelo, David Gaudrie, Peter Reinartz, and Daniel Cremers. Spatially regularized fusion of multiresolution digital surface models. *IEEE Transactions on Geoscience and Remote Sensing*, 55(3):1477–1488, 2017. https://doi.org/10.1109/TGRS.2016.2625040.

[KdSG13] Thomas Krauß, Pablo d'Angelo, Mathias Schneider, and Veronika Gstaiger. The fully automatic optical processing system CATENA at DLR. *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40-1/W1:177–181, 2013. https://doi.org/10.5194/isprsarchives-XL-1-W1-177-2013.

[KGAL20] Chiman Kwan, David Gribben, Bulent Ayhan, and Jude Larkin. Practical digital terrain model extraction using image inpainting techniques. In *Recent Advances in Image Restoration with Applications to Real World Problems*. IntechOpen, 2020. https://doi.org/10.5772/intechopen.93184.

[KGC18] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7482–7491, 2018. https://doi.org/10.1109/CVPR.2018.00781.

[KMD+17] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 66–75, 2017. https://doi.org/10.1109/ICCV.2017.17.

[Kus13] Georg Kuschk. Large scale urban reconstruction from remote sensing imagery. *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40-5/W1:139–146,

2013. https://doi.org/10.5194/isprsarchives-XL-5-W1-139-2013.

[LA05] Manolis L.A. Lourakis and Antonis A. Argyros. Is Levenberg-Marquardt the most efficient optimization algorithm for implementing bundle adjustment? In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1526–1531, 2005. https://doi.org/10.1109/ICCV.2005.128.

[LBK20] Lukas Liebel, Kesnia Bittner, and Marco Körner. A generalized multi-task learning approach to stereo DSM filtering in urban areas. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:213–227, 2020. https://doi.org/10.1016/j.isprsjprs.2020.03.005.

[LJBB20] Hamid Laga, Laurent Valentin Jospin, Farid Boussaid, and Mohammed Bennamoun. A survey on deep learning techniques for stereo-based depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):1738–1764, 2020. https://doi.org/10.1109/TPAMI.2020.3032602.

[LJH15] Tengfei Long, Weili Jiao, and Guojin He. RPC estimation via ℓ_1-norm regularized least squares (L1LS). *IEEE Transactions on Geoscience and Remote Sensing*, 53(8):4554–4567, 2015. https://doi.org/10.1109/TGRS.2015.2401602.

[LKC15] Thomas Lillesand, Ralph W Kiefer, and Jonathan Chipman. *Remote sensing and image interpretation*. John Wiley & Sons, 2015.

[LLJ+19] Matthew J Leotta, Chengjiang Long, Bastien Jacquet, Matthieu Zins, Dan Lipsa, Jie Shan, Bo Xu, Zhixin Li, Xu Zhang, Shih-Fu Chang, et al. Urban semantic 3D reconstruction from multiview satellite imagery. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1451–1460, 2019. https://doi.org/10.1109/CVPRW.2019.00186.

[LMG+19] Manuel Lopez, Roger Marí, Pau Gargallo, Yubin Kuang, Javier Gonzalez-Jimenez, and Gloria Haro. Deep single image camera calibration with radial distortion. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11809–11817, 2019. https://doi.org/10.1109/CVPR.2019.01209.

[Low04] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. https://doi.org/10.1023/B:VISI.0000029664.99615.94.

[LRF+19] Lorenzo Lastilla, Roberta Ravanelli, Francesca Fratarcangeli, Martina Di Rita, Andrea Nascetti, and Mattia Crespi. Foss4g date for dsm generation: Sensitivity analysis of the semi-global block matching parameters. *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42-2/W13, 2019. https://doi.org/10.5194/isprs-archives-XLII-2-W13-67-2019.

[LSYH+19] Bertrand Le Saux, Naoto Yokoya, Ronny Hansch, Myron Brown, and Greg Hager. 2019 data fusion contest [technical committees]. *IEEE*

Geoscience and Remote Sensing Magazine, 7(1):103–105, 2019. https://doi.org/10.1109/MGRS.2019.2893783.

[MBRS+21] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the wild: Neural radiance fields for unconstrained photo collections. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7206–7215, 2021. https://doi.org/10.1109/CVPR46437.2021.00713.

[MdFML+21] Roger Marí, Carlo de Franchis, Enric Meinhardt-Llopis, Jérémy Anger, and Gabriele Facciolo. A generic bundle adjustment methodology for indirect RPC model refinement of satellite imagery. *Image Processing On Line*, 11:344–373, 2021. https://doi.org/10.5201/ipol.2021.352.

[MdFMLF19] Roger Marí, Carlo de Franchis, Enric Meinhardt-Llopis, and Gabriele Facciolo. To bundle adjust or not: A comparison of relative geolocation correction strategies for satellite multi-view stereo. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 2188–2196, 2019. https://doi.org/10.1109/ICCVW.2019.00274.

[MdFMLF21] Roger Marí, Carlo de Franchis, Enric Meinhardt-Llopis, and Gabriele Facciolo. Automatic stockpile volume monitoring using multi-view stereo from SkySat imagery. In *2021 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 4384–4387, 2021. https://doi.org/10.1109/IGARSS47720.2021.9554482.

[MEA+22] Roger Marí, Thibaud Ehret, Jérémy Anger, Carlo de Franchis, and Gabriele Facciolo. L1B+: A perfect sensor localization model for simple satellite stereo reconstruction from push-frame image strips. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 5-1-2022:137–143, 2022. https://doi.org/10.5194/isprs-annals-V-1-2022-137-2022.

[MEF22] Roger Marí, Thibaud Ehret, and Gabriele Facciolo. Disparity estimation networks for aerial and high-resolution satellite images: A review. *Image Processing On Line*, 12:501–526, 2022. https://doi.org/10.5201/ipol.2022.435.

[MESK22] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4), 2022. https://doi.org/10.1145/3528223.3530127.

[MFE22] Roger Marí, Gabriele Facciolo, and Thibaud Ehret. Sat-NeRF: Learning multi-view satellite photogrammetry with transient objects and shadow modeling using RPC cameras. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1310–1320, 2022. https://doi.org/10.1109/CVPRW56347.2022.00137.

[MG15] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3061–3070, 2015. https://doi.org/10.1109/CVPR.2015.7298925.

[MHMB+22] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P Srinivasan, and Jonathan T Barron. NeRF in the dark: High dynamic range view synthesis from noisy raw images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16169–16178, 2022. https://doi.org/10.1109/CVPR52688.2022.01571.

[MIH+16] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4040–4048, 2016. https://doi.org/10.1109/CVPR.2016.438.

[ML14] Marius Muja and David G Lowe. Scalable nearest neighbor algorithms for high dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2227–2240, 2014. https://doi.org/10.1109/TPAMI.2014.2321376.

[MLd19] Enric Meinhardt-Llopis and Marie d'Autume. 3D texturing from multi-date satellite images. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 2197–2205, 2019. https://doi.org/10.1109/ICCVW.2019.00275.

[MM12] Pierre Moulon and Pascal Monasse. Unordered feature tracking made fast and easy. In *European Conference on Visual Media Production (CVMP)*, 2012. https://doi.org/10.13140/2.1.3743.3287.

[MMPM17] Pierre Moulon, Pascal Monasse, Romuald Perrot, and Renaud Marlet. OpenMVG: Open multiple view geometry. In *Reproducible Research in Pattern Recognition*, pages 60–74. Springer, 2017. https://doi.org/10.1007/978-3-319-56414-2_5.

[MQS08] Haijian Ma, Qiming Qin, and Xinyi Shen. Shadow segmentation and compensation in high resolution satellite images. In *2008 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, volume 2, pages 1036–1039, 2008. https://doi.org/10.1109/IGARSS.2008.4779175.

[MSS+14] Kiran Murthy, Michael Shearn, Byron D Smiley, Alexandra H Chau, Josh Levine, and M Dirk Robinson. SkySat-1: Very high-resolution imagery from a small satellite. In *Sensors, Systems, and Next-Generation Satellites XVIII*, volume 9241, page 92411E. International Society for Optics and Photonics, 2014. https://doi.org/10.1117/12.2074163.

[MST+20] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Computer Vision –*

ECCV 2020, pages 405–421, 2020. https://doi.org/10.1007/978-3-030-58452-8_24.

[MSY+20] Julien Michel, Emmanuelle Sarrazin, David Youssefi, Myriam Cournet, Fabrice Buffe, Jean-Marc Delvit, Aurélie Emilien, Julien Bosman, Olivier Melet, and Céline L'Helguen. A new satellite imagery stereo pipeline designed for scalability, robustness and performance. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 5-2-2020:171–178, 2020. https://doi.org/10.5194/isprs-annals-V-2-2020-171-2020.

[NAD+21] Ngoc Long Nguyen, Jérémy Anger, Axel Davy, Pablo Arias, and Gabriele Facciolo. Self-supervised multi-image super-resolution for push-frame satellite images. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1121–1131, 2021. https://doi.org/10.1109/CVPRW53098.2021.00123.

[NBM+22] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. RegNeRF: Regularizing neural radiance fields for view synthesis from sparse inputs. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5470–5480, 2022. https://doi.org/10.1109/CVPR52688.2022.00540.

[NK11] Christopher Nuth and Andreas Kääb. Co-registration and bias corrections of satellite elevation data sets for quantifying glacier thickness change. *The Cryosphere*, 5(1):271–290, 2011. https://doi.org/10.5194/tc-5-271-2011.

[NYD16] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Computer Vision – ECCV 2016*, pages 483–499, 2016. https://doi.org/10.1007/978-3-319-46484-8_29.

[OAH93] Atsushi Okamoto, Sin-ichi Akamatu, and Hiroyuki Hasegawa. Orientation theory for satellite CCD line-scanner imageries of hilly terrains. *ISPRS International Archives of Photogrammetry and Remote Sensing*, 29:217–217, 1993.

[ODM+14] Ozge C Ozcanli, Yi Dong, Joseph L Mundy, Helen Webb, Riad Hammoud, and Tom Victor. Automatic geo-location correction of satellite imagery. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 307–314, 2014. https://doi.org/10.1109/CVPRW.2014.54.

[ODM+15] Ozge C Ozcanli, Yi Dong, Joseph L Mundy, Helen Webb, Riad Hammoud, and Victor Tom. A comparison of stereo and multiview 3-D reconstruction using cross-sensor satellite imagery. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 17–25, 2015. https://doi.org/10.1109/CVPRW.2015.7301292.

[ODO16] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016. https://doi.org/10.23915/distill.00003.

[OLT+10] Jaehong Oh, Won Hee Lee, Charles K Toth, Dorota A Grejner-Brzezinska, and Changno Lee. A piecewise approach to epipolar resampling of pushbroom satellite images based on rpc. *Photogrammetric Engineering & Remote Sensing*, 76(12):1353–1363, 2010. https://doi.org/10.14358/PERS.76.12.1353.

[PA17] Ravi Ancil Persad and Costas Armenakis. Automatic co-registration of 3D multi-sensor point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 130:162–186, 2017. https://doi.org/10.1016/j.isprsjprs.2017.05.014.

[PCPMMN21] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural radiance fields for dynamic scenes. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10313–10322, 2021. https://doi.org/10.1109/CVPR46437.2021.01018.

[PCSM13] François Pomerleau, Francis Colas, Roland Siegwart, and Stéphane Magnenat. Comparing ICP variants on real-world data sets. *Autonomous Robots*, 34(3):133–148, 2013. https://doi.org/10.1007/s10514-013-9327-2.

[PDJB+] Marc Pierrot-Deseilligny, Denis Jouin, Jérémie Belvaux, Grégoire Maillet, Luc Girod, Ewelina Rupnik, Jean-Michael Muller, Mehdi Daakir, Gérald Choqueux, and Matthieu Deveau. MicMac, Apero, Pastis and other beverages in a nutshell. *Institut Géographique National*. https://github.com/micmacIGN/Documentation.

[PHZC21] Hongbo Pan, Tao Huang, Ping Zhou, and Zehua Cui. Self-calibration dense bundle adjustment of multi-view Worldview-3 basic images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 176:127–138, 2021. https://doi.org/10.1016/j.isprsjprs.2021.04.013.

[Pla19] Planet. Planet imagery product specifications, May 2019.

[Pla22] Planet. Planet imagery product specifications, May 2022.

[PSB+21] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5845–5854, 2021. https://doi.org/10.1109/ICCV48922.2021.00581.

[PSH+21] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. HyperNeRF: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6), 2021. https://doi.org/10.1145/3478513.3480487.

[PSR+17] Jiahao Pang, Wenxiu Sun, Jimmy SJ Ren, Chengxi Yang, and Qiong Yan. Cascade residual learning: A two-stage convolutional neural

network for stereo matching. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 878–886, 2017. https://doi.org/10.1109/ICCVW.2017.108.

[PT12] Daniela Poli and Thierry Toutin. Review of developments in geometric modelling for high resolution satellite pushbroom sensors. *The Photogrammetric Record*, 27(137):58–73, 2012. https://doi.org/10.1111/j.1477-9730.2011.00665.x.

[PY09] Charalambos Poullis and Suya You. Automatic reconstruction of cities from remote sensor data. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2775–2782, 2009. https://doi.org/10.1109/CVPR.2009.5206562.

[Qin17] Rongjun Qin. Automated 3D recovery from very high resolution multi-view images. In *ASPRS Annual Conference*, 2017.

[QLFR22] Rongjun Qin, Xiao Ling, Elisa Mariarosaria Farella, and Fabio Remondino. Uncertainty-guided depth fusion from multi-view satellite images to improve the accuracy in large-scale DSM generation. *Remote Sensing*, 14(6):1309, 2022. https://doi.org/10.3390/rs14061309.

[Rag06] Hannes Raggam. Surface mapping using image triplets. *Photogrammetric Engineering & Remote Sensing*, 72(5):551–563, 2006. https://doi.org/10.14358/PERS.72.5.551.

[RBM+22] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12882–12891, 2022. https://doi.org/10.1109/CVPR52688.2022.01255.

[RDPD17] Ewelina Rupnik, Mehdi Daakir, and Marc Pierrot-Deseilligny. MicMac–a free, open-source solution for photogrammetry. *Open Geospatial Data, Software and Standards*, 2(14), 2017. https://doi.org/10.1186/s40965-017-0027-2.

[RFF16] P.L. Raeva, S.L. Filipova, and D.G. Filipov. Volume computation of a stockpile - A study case comparing GPS and UAV measurements in an open pit quarry. *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 41-B1:999–1004, 2016. https://doi.org/10.5194/isprs-archives-XLI-B1-999-2016.

[ROD14] Ives Rey Otero and Mauricio Delbracio. Anatomy of the SIFT method. *Image Processing On Line*, 4:370–396, 2014. https://doi.org/10.5201/ipol.2014.82.

[RPDD18] Ewelina Rupnik, Marc Pierrot-Deseilligny, and Arthur Delorme. 3D reconstruction from multi-view VHR-satellite images in MicMac. *ISPRS Journal of Photogrammetry and Remote Sensing*, 139:201–211, 2018. https://doi.org/10.1016/j.isprsjprs.2018.03.016.

[RSJ+12] Franz Rottensteiner, Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, Sebastien Benitez, and Uwe Breitkopf. The IS-

PRS benchmark on urban object classification and 3D building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 1-3:293–298, 2012. https://doi.org/10.5194/isprsannals-I-3-293-2012.

[RWFH12] Mathias Rothermel, Konrad Wenzel, Dieter Fritsch, and Norbert Haala. SURE: Photogrammetric surface reconstruction from imagery. In *Proceedings LC3D Workshop, Berlin*, volume 8, 2012.

[SAM+16] David E. Shean, Oleg Alexandrov, Zachary M. Moratto, Benjamin E. Smith, Ian R. Joughin, Claire Porter, and Paul Morin. An automated, open-source pipeline for mass production of digital elevation models (DEMs) from very-high-resolution commercial stereo satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 116:101–117, 2016. https://doi.org/10.1016/j.isprsjprs.2016.03.012.

[SBD10] Rainer Sandau, Klaus Brieß, and Marco D'Errico. Small satellites for global coverage: Potential and limits. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(6):492–504, 2010. https://doi.org/10.1016/j.isprsjprs.2010.09.003.

[Sci20] SciPy. Generic interface for least-squares minimization, 2020. https://github.com/scipy/scipy/blob/master/scipy/optimize/_lsq/least_squares.py.

[SCM10] Anuj Sehgal, Daniel Cernea, and Milena Makaveeva. Real-time scale invariant 3D range point cloud registration. In *Image Analysis and Recognition*, pages 220–229. Springer, 2010. https://doi.org/10.1007/978-3-642-13772-3_23.

[SdKR10] Beril Sirmacek, Pablo d'Angelo, Thomas Krauß, and Peter Reinartz. Enhancing urban digital elevation models using automated computer vision techniques. *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 38, 2010.

[SDZ+21] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. NeRV: Neural reflectance and visibility fields for relighting and view synthesis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7491–7500, 2021. https://doi.org/10.1109/CVPR46437.2021.00741.

[SF16] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. https://doi.org/10.1109/CVPR.2016.445.

[SJS+17] Pauziyah Mohammad Salim, Muhammad Anas Md Jais, Noorita Sahriman, Abd Manan Samad, Mohd Azwan Abbas, Ismail Maarof, and Nazirah Md Tarmizi. Monitoring quarry areas using remote sensing techniques. In *2017 IEEE 13th International Colloquium on Signal Processing & its Applications (CSPA)*, pages 323–328, 2017. https://doi.org/10.1109/CSPA.2017.8064974.

[SLB09] Dale E Schinstock, Chris Lewis, and Craig Buckley. An alternative cost function to bundle adjustment used for aerial photography from UAVs. In *ASPRS Annual Conference*, 2009.

[SMB+20] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020.

[SMTR15] Benjamin Schmidt, Matt Malgesini, Jim Turner, and Jeff Reinson. Satellite monitoring of a large tailings storage facility. In *Tailings and Mine Waste Conference*, 2015. https://doi.org/10.14288/1.0314311.

[SS02] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1):7–42, 2002. https://doi.org/10.1023/A:1014573219977.

[SS20] Corinne Stucker and Konrad Schindler. ResDepth: Learned residual stereo reconstruction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 707–716, 2020. https://doi.org/10.1109/CVPRW50498.2020.00100.

[SSK18] Yusuke Sugawara, Sayaka Shiota, and Hitoshi Kiya. Super-resolution using convolutional neural networks without any checkerboard artifacts. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 66–70, 2018. https://doi.org/10.1109/ICIP.2018.8451141.

[SSS06] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: Exploring photo collections in 3D. *ACM Trans. Graph.*, 25(3):835–846, 2006. https://doi.org/10.1145/1141911.1141964.

[Sze10] Richard Szeliski. *Computer vision: Algorithms and applications*. Springer Science & Business Media, 2010. https://doi.org/10.1007/978-1-84882-935-0.

[SZFP16] Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision – ECCV 2016*, pages 501–518, 2016. https://doi.org/10.1007/978-3-319-46487-9_31.

[TFT+20] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. State of the art on neural rendering. In *Computer Graphics Forum*, volume 39, pages 701–727, 2020. https://doi.org/10.1111/cgf.14022.

[TGC+19] Grazia Tucci, Antonio Gebbia, Alessandro Conti, Lidia Fiorini, and Claudio Lubello. Monitoring and computation of the volumes of stockpiles of bulk material by means of UAV photogrammetric surveying. *Remote Sensing*, 11(12):1471, 2019. https://doi.org/10.3390/rs11121471.

[TGH08] Andrew J Tatem, Scott J Goetz, and Simon I Hay. Fifty years of earth-observation satellites: Views from space have led to countless advances on the ground in both scientific knowledge and daily life. *American scientist*, 96(5):390–398, 2008.

[TH01] C. Vincent Tao and Yong Hu. A comprehensive study of the rational function model for photogrammetric processing. *Photogrammetric Engineering & Remote Sensing*, 67(12):1347–1357, 2001.

[TH02] C Vincent Tao and Yong Hu. 3D reconstruction methods based on the rational function model. *Photogrammetric Engineering & Remote Sensing*, 68(7):705–714, 2002.

[TJ16] Charles Toth and Grzegorz Jóźków. Remote sensing platforms and sensors: A survey. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115:22–36, 2016. https://doi.org/10.1016/j.isprsjprs.2015.10.004.

[TLW10] Xiaohua Tong, Shijie Liu, and Qihao Weng. Bias-corrected rational polynomial coefficients for high accuracy geo-positioning of QuickBird stereo imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(2):218–226, 2010. https://doi.org/10.1016/j.isprsjprs.2009.12.004.

[TMHF00] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment — A modern synthesis. In *Vision Algorithms: Theory and Practice*, pages 298–372. Springer, 2000. https://doi.org/10.1007/3-540-44480-7_21.

[Tod77] William J. Todd. Urban and regional land use change detected by using Landsat data. *Journal of Research of the U.S. Geological Survey*, 5(5):529–534, 1977. https://pubs.er.usgs.gov/publication/70112341.

[VI18] Styliani Verykokou and Charalabos Ioannidis. A photogrammetry-based structure from motion algorithm using robust iterative bundle adjustment techniques. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4-4/W6:73–80, 2018. https://doi.org/10.5194/isprs-annals-IV-4-W6-73-2018.

[WBPVLdCGT17] Timo Wekerle, José Bezerra Pessoa, Luís E. Vergueiro Loures da Costa, and Luís Gonzaga Trabasso. Status and trends of small-sats and their launch vehicles—an up-to-date review. *Journal of Aerospace Technology and Management*, 9:269–286, 2017. https://doi.org/10.5028/jatm.v9i3.853.

[WGA06] Darrel L. Williams, Samuel Goward, and Terry Arvidson. Landsat: Yesterday, today, and tomorrow. *Photogrammetric Engineering & Remote Sensing*, 72:1171–1178, 10 2006. https://doi.org/10.14358/PERS.72.10.1171.

[WHL10] Mi Wang, Fen Hu, and Jonathan Li. Epipolar arrangement of satellite imagery by projection trajectory simplification. *The Photogrammetric*

Record, 25(132):422–436, 2010. Epipolararrangementofsatelliteimagerybyprojectiontrajectorysimplification.

[WHP+12] Jurgen Wohlfeil, Heiko Hirschmuller, Björn Piltz, Anko Börner, and Michael Suppa. Fully automated generation of accurate digital surface models with sub-meter resolution from satellite imagery. *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 39-B3:75–80, 2012. https://doi.org/10.5194/isprsarchives-XXXIX-B3-75-2012.

[WVPDR21] Teng Wu, Bruno Vallet, Marc Pierrot-Deseilligny, and Ewelina Rupnik. A new stereo dense matching benchmark dataset for deep learning. *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43-B2-2021:405–412, 2021. https://doi.org/10.5194/isprs-archives-XLIII-B2-2021-405-2021.

[WW06] Jeffrey P Walker and Garry R Willgoose. A comparative study of Australian cartometric and photogrammetric digital elevation model accuracy. *Photogrammetric Engineering & Remote Sensing*, 72(7):771–779, 2006. https://doi.org/10.14358/PERS.72.7.771.

[WZB21] Yi Wang, Stefano Zorzi, and Ksenia Bittner. Machine-learned 3D building vectorization from satellite imagery. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1072–1081, 2021. https://doi.org/10.1109/CVPRW53098.2021.00118.

[WZS16] Yiran Wang, Ye Zhang, and Nan Su. RPC estimation via feature points for urban areas. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 6684–6687, 2016. https://doi.org/10.1109/IGARSS.2016.7730745.

[XAS21] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-NeRF: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *Advances in Neural Information Processing Systems*, 34:14955–14966, 2021.

[XdF+22] Yuanxin Xia, Pablo d'Angelo, Friedrich Fraundorfer, Jiaojiao Tian, Mario Fuentes Reyes, and Peter Reinartz. GA-Net-Pyramid: An efficient end-to-end network for dense matching. *Remote Sensing*, 14(8):1942, 2022. https://doi.org/10.3390/rs14081942.

[XLG+08] Yong Xue, Yingjie Li, Jie Guang, Xiaoye Zhang, and Jianping Guo. Small satellite remote sensing and applications–history, current and future. *International Journal of Remote Sensing*, 29(15):4339–4372, 2008. https://doi.org/10.1080/01431160801914945.

[XZ09] Zhen Xiong and Yun Zhang. A generic method for RPC refinement using ground control information. *Photogrammetric Engineering & Remote Sensing*, 75(9):1083–1092, 2009. https://doi.org/10.14358/PERS.75.9.1083.

[YLL+18] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. In *Computer Vi-*

sion – ECCV 2018, pages 785–801, 2018. https://doi.org/10.1007/978-3-030-01237-3_47.

[YLT+21] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. PlenOctrees for real-time rendering of neural radiance fields. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5732–5741, 2021. https://doi.org/10.1109/ICCV48922.2021.00570.

[YMHR19] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5510–5519, 2019. https://doi.org/10.1109/CVPR.2019.00566.

[YYTK21] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4576–4585, 2021. https://doi.org/10.1109/CVPR46437.2021.00455.

[YZP+22] Guo-Wei Yang, Wen-Yang Zhou, Hao-Yang Peng, Dun Liang, Tai-Jiang Mu, and Shi-Min Hu. Recursive-NeRF: An efficient and dynamically growing NeRF. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–14, 2022. https://doi.org/10.1109/TVCG.2022.3204608.

[Zac14] Christopher Zach. Robust bundle adjustment revisited. In *Computer Vision – ECCV 2014*, pages 772–787. Springer, 2014. https://doi.org/10.1007/978-3-319-10602-1_50.

[ZFL+10] Guo Zhang, Wen-bo Fei, Zhen Li, Xiaoyong Zhu, and De-ren Li. Evaluation of the RPC model for spaceborne SAR imagery. *Photogrammetric Engineering & Remote Sensing*, 76(6):727–733, 2010. https://doi.org/10.14358/PERS.76.6.727.

[ZHB+11] Lu Zhang, Xueyan He, Timo Balz, Xiaohong Wei, and Mingsheng Liao. Rational function modeling for spaceborne SAR datasets. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(1):133–145, 2011. https://doi.org/10.1016/j.isprsjprs.2010.10.007.

[ZPYT19] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip H.S. Torr. GA-Net: Guided aggregation net for end-to-end stereo matching. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 185–194, 2019. https://doi.org/10.1109/CVPR.2019.00027.

[ZQW+16] Wuming Zhang, Jianbo Qi, Peng Wan, Hongtao Wang, Donghui Xie, Xiaoyan Wang, and Guangjian Yan. An easy-to-use airborne LiDAR data filtering method based on cloth simulation. *Remote Sensing*, 8(6):501, 2016. https://doi.org/10.3390/rs8060501.

[ZSL+17] Lingli Zhu, Juha Suomalainen, Jingbin Liu, Juha Hyyppä, Harri Kaartinen, and Henrik Haggren. A review: Remote sensing sensors.

Multi-purposeful Application of Geospatial Data, pages 19–42, 2017. https://doi.org/10.5772/intechopen.71049.

[ZSQ+17] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6230–6239, 2017. https://doi.org/10.1109/CVPR.2017.660.

[ZSS19] Kai Zhang, Noah Snavely, and Jin Sun. Leveraging vision reconstruction pipelines for satellite imagery. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 2139–2148, 2019. https://doi.org/10.1109/ICCVW.2019.00269.

[ZW94] Ramin Zabih and John Woodfill. Non-parametric local transforms for computing visual correspondence. In *Computer Vision — ECCV '94*, pages 151–158, 1994. https://doi.org/10.1007/BFb0028345.

[ZWDF15] Enliang Zheng, Ke Wang, Enrique Dunn, and Jan-Michael Frahm. Minimal solvers for 3D geometry from satellite imagery. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 738–746, 2015. https://doi.org/10.1109/ICCV.2015.91.

[ZY19] Wang Zhang and Deshan Yang. Lidar-based fast 3D stockpile modeling. In *2019 International Conference on Intelligent Computing, Automation and Systems (ICICAS)*, pages 703–707, 2019. https://doi.org/10.1109/ICICAS48597.2019.00152.

[ZYF+18] Jianjun Zhu, Jian Yang, Jingfan Fan, Danni Ai, Yurong Jiang, Hong Song, and Yongtian Wang. Accurate measurement of granary stockpile volume based on fast registration of multi-station scans. *Remote Sensing Letters*, 9(6):569–577, 2018. https://doi.org/10.1080/2150704X.2018.1452060.