Towards scalable, multi-view urban modeling using structure priors

Amine Bourki

Université Paris-Est – MSTIC

LIGM, Ecole des Ponts ParisTech

Specialty: Signal, Image, Automation

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science.

Towards Scalable, Multi-View Urban Modeling using Structure Priors

---

Tahar Amine Bourki
Paris, France, 2017

Jury Members

President	Pr. Valérie Gouet-Brunet	LaSTIG, IGN
Reviewer	Dr. Renaud Keriven	Acute3D, Bentley Systems
Reviewer	Pr. David Fofi	Le2i, Université de Bourgogne
Advisor	Pr. Renaud Marlet	LIGM, Ecole des Ponts ParisTech

To my family...

Abstract

In this thesis, we address the problem of 3D reconstruction from a sequence of calibrated street-level photographs with a simultaneous focus on scalability and the use of structure priors in Multi-View Stereo (MVS).

While both aspects have been studied broadly, existing scalable MVS approaches do not handle well the ubiquitous structural regularities, yet simple, of man-made environments. On the other hand, structure-aware 3D reconstruction methods are slow and scale poorly with the size of the input sequences and/or may even require additional restrictive information. The goal of this thesis is to reconcile scalability and structure-awareness within common MVS grounds using soft, generic priors which encourage: (i) piecewise planarity, (ii) alignment of objects boundaries with image gradients and (iii) with vanishing directions (VDs), and (iv) objects co-planarity. To do so, we present the novel "Patchwork Stereo" framework which integrates photometric stereo from a handful of wide-baseline views and a sparse 3D point cloud combining robust 3D plane extraction and top-down image partitioning from a unified 2D-3D analysis in a principled Markov Random Field energy minimization.

We evaluate our contributions quantitatively and qualitatively on challenging urban datasets and illustrate results which are at least on par with state-of-the-art methods in terms of geometric structure, but achieved in several orders of magnitude faster paving the way for photo-realistic city-scale modeling.

Keywords: Multi-View Stereo, Structure Priors, 3D Reconstruction, Image-Based Modeling, Scalability, Top-Down Image Segmentation, Urban Modeling.

Résumé étendu de la Thèse

Résumé général.
Nous étudions dans cette thèse le problème de reconstruction 3D multi-vue à partir d'une séquence d'images au sol acquises dans des environnements urbains ainsi que la prise en compte d'a priori permettant la preservation de la structure sous-jacente de la géométrie 3D observée, ainsi que le passage à l'échelle de tels processus de reconstruction qui est intrinsèquement délicat dans le contexte de l'imagerie urbaine.

Bien que ces deux axes aient été traités de manière extensive dans la littérature, les méthodes de reconstruction 3D structurée souffrent d'une complexité en temps de calculs restreignant significativement leur intérêt. D'autre part, les méthodes de reconstruction 3D large échelle produisent généralement une géométrie approchée, omettant ainsi des éléments de structure qui sont importants dans le contexte urbain. L'objectif de cette thèse est de concilier les avantages des techniques de reconstruction 3D structurée à ceux des méthodes passant à l'échelle. Pour ce faire, nous présentons "Patchwork Stereo", un système qui combine stéréoscopie photométrique à partir d'une poignée d'images issues de points de vue éloignés et d'un nuage de points épars. Notre méthode intègre une analyse simultanée 2D-3D réalisant une extraction robuste de plans 3D ainsi qu'une segmentation d'images top-down structurée et repose sur une optimisation par champs de Markov aléatoires.

Les contributions présentées sont évaluées via des expériences quantitatives et qualitatives sur des données d'imagerie urbaine complexes illustrant des performances tant quant à la fidélité structurelle des reconstructions 3D que du passage à l'échelle.

Mots clés: Stéréoscopie Multi-Vue, A priori de Structure, Reconstruction 3D, Modélisation Basée Image, Passage à l'échelle, Segmentation Top-Down d'Images, Modélisation Urbaine.

Contexte.
Face à la demande grandissante de modèles 3D d'environnements créés par l'Homme, telles que les scènes d'extérieurs ou d'intérieurs de bâtiments, de nombreux efforts ont été réalisés afin de générer des modèles réalistes ou de restituer le plus fidèlement possible des scènes existantes.

Cet intérêt est illustré par de très nombreuses applications comme par exemple venant des industries du jeu vidéo et du cinéma pour ce qui est de la génération de modèles réalistes, où l'enjeu principal est d'obtenir des représentations visuellement crédibles, avec un niveau de détails adapté aux contraintes éventuelles de stockage en mémoire ou de temps de calculs des rendus des scènes. Une autre catégorie d'applications s'intérêsse à la numérisation de villes ou de bâtiments existants afin de reconstruire une représentation digitale servant de support pour des processus d'aide à la décision comme l'analyse de performances énergétiques de bâtiments ou d'autres calculs liés à la plannification, la vie ou la destruction d'un ou de plusieurs bâtiments.

Pour résoudre le problème de numérisation de scènes régulières existantes, l'intérêt des méthodes de reconstruction 3D traditionnelles est limité en raison de leur coût en temps de calculs et stockage mémoire ainsi que pour la complexité des maillages qu'elles produisent contrastant paradoxalement avec la simplicité structurelle des scènes urbaines ou d'intérieurs. Par ailleurs, les méthodes passant à l'échelle génèrent typiquement une géométrie approchée qui ne prend pas en compte des aspects structurels qui sont indispensables dans le cas de nombreux scénarii applicatifs. Nous proposons de concilier les avantages des méthodes de reconstruction approchée, à base de patches, ainsi que les méthodes intégrant des a priori de structure qui raisonnent au niveau du pixel et dont le passage à l'échelle est donc initialement limité. Les deux axes d'intérêt de notre étude concernent donc d'une part (i) la prise en compte d'a priori de structure preservant la régularité ainsi que la simplicité des scènes observées et (ii) le passage à l'échelle.

Aperçu des contributions.
Nous proposons un système de reconstruction 3D basé images prenant en entrées une séquence d'images au sol calibrées décrivant une scène urbaine (des bâtiments) ainsi qu'un nuage de points épars et bruité obtenu – par exemple – à partir de la procédure de calibrage des caméras. L'une des images est considérée comme référence, le reste comme étant des images de reprojection. La sortie de notre système est une reconstruc-

tion 3D dans le repère de l'image de référence sous forme d'une carte de profondeur et d'un maillage associé, en favorisant des principes de structure géométrique que nous définissons comme suit : (i) planarité par morceaux alignement des contours des principaux éléments de la scène selon (ii) les principaux gradients dans le domaine image ainsi qu'avec (iii) les lignes de fuite dominantes de la scène et (iv) la co-planarité des éléments considérés (simplicité géométrique).

Le système de reconstruction que nous proposons dans cette thèse, appelé "Patchwork Stereo" s'articule selon les étapes suivantes. Dans un premier temps, l'image de référence est segmentée et des hypothèses de plans 3D dominants sont extraits à partir du nuage de points épars associé à la scène.

Notre méthode passe d'abord par la détection de directions principales (points de fuites) présentes dans la scène en exploitant des indices visuels basés images (segments) via une approche gloutonne et s'en suit une détection des lignes de fuite principales par une approche de balayage de faisceaux de lignes à travers les pixels de l'image de référence issus de chacun des points de fuite détectés. L'arrangement complets de ces lignes de fuite génère un partitionnement de l'image en superpixels top-down (car exploitant des attributs structurels de haut niveau, par opposition à des superpixels bottom-up engendrés par un assemblage de pixels voisins partageant des similarités d'apparence ou de textures). Intuitivement, l'attrait des lignes de fuite dans le partitionnement de l'image réside principalement dans le fait que ces dernières sont adaptées à la structure géométrique de scènes régulières telles que les façades de bâtiments et les scènes d'intérieurs.

Afin de consolider la qualité des détections de lignes de fuites structurelles ainsi que des hypothèses de plans 3D, mais également afin d'établir une compatibilité entre les hypothèses 2D (superpixels) et 3D (plans), nous proposons une phase d'analyse conjointement en 2D/3D en procédant par balayage de faisceaux dans le domaine image, en construisant une fonction de score pour extraire des lignes et plans 3D supplémentaires et compatibles.

Enfin, nous proposons une énergie globale par champs de Markov sur la topologie induite par le partitionnement top-down de l'image de référence en formalisant le problème de reconstruction 3D comme un problème d'étiquetage discret de chaque superpixel par une hypothèse planaire, en encourageant des combinaisons compatibles entre des patches voisins dans le plan image. Ces relations de compatibilités binaires

favorisent la continuité planaire et les jonctions le long de lignes de fuite structurelles au détriment d'autres configurations beaucoup moins probables en pratique.

En somme, les contributions que nous présentons dans cette thèse peuvent se résumer de la manière suivante :

1. Une extraction robuste d'hypothèses de plans 3D à partir d'un nuage de point épars et bruité, typiquement acquis lors d'une phrase de Structure-from-Motion (SfM).

2. Une analyse conjointe 2D/3D afin d'établir un partitionnement d'image top-down, respectant la structure globale de scènes créées par l'Homme, tels que les environnements urbains.

3. Un schéma de reconstruction 3D par champs de Markov combinant les éléments sus-mentionnés dans une énergie globale et résolue via graph-cuts.

4. Les principales contributions avancées dans cette thèse ont fait l'objet d'une communication internationale en Vision par Ordinateur (WACV 2017).

Conclusions et perspectives.
Nous avons proposé des solutions qui s'inscrivent dans le traitement de la problématique de reconstruction 3D multi-vue (Multi-View Stereo, MVS) de scènes "créées par l'Homme", qui sont typiquement régulières en termes de structure géométrique. Nous avons simultanément orienté notre étude selon deux axes que sont la préservation de la régularité structurelle des objets et scènes observés, ainsi que le passage à l'échelle du processus de reconstruction 3D.

En termes de limites et éléments perfectibles des travaux proposés, nous évoquons les points suivants :

(i) Expériences et données.
La première limite du travail présenté dans ce manuscrit tient dans le manque de richesse en termes de variété et du nombre de scènes considérées dans nos expériences afin d'illustrer les performances de nos méthodes. En particulier nous n'avons pas évalué nos méthodes sur des scènes créées par l'Homme comprenant plus de 3 points de fuite, bien que le système que nous proposons soit tout à fait capable de traiter des scènes plus complexes, sans que cela ne nécessite de réaliser de modifications.

Trouver de telles scènes est particulièrement difficile dans la mesure où la vaste majorité des scènes urbaines comprennent des bâtiments dont la structure ne présentant le plus souvent que 3 points de fuite.

Une solution possible serait de considérer des scènes synthétiques ou encore des scènes d'intérieurs qui dérogent plus souvent à une logique structurelle à 3 points de fuite.

La majorité des travaux traitant de notre problème ne proposent pas de validation expérimentale de la précision géométrique des modèles 3D reconstruits en raison du faible nombre de datasets mis à disposition de la communauté scientifique comprenant à la fois des photos de scènes structurellement régulières, les propriétés de calibrage des points de vue considérés, un nuage de points épars correspondant par scène, ainsi qu'un modèle 3D de référence vis à vis duquel la précision géométrique serait mesurée de manière globale, ou pour un point de vue donnée. Nous avons fait le choix de pallier ce manque en construisant des datasets à partir d'un nombre très important d'images par scène décrivant des bâtiments, produisant ainsi des reconstructions 3D sous forme de maillages fins en utilisant une méthode générique de MVS. Nous confrontons ainsi la précision de nos reconstructions qui n'utilisent qu'un faible sous-ensemble des images disponibles (jursqu'à 9 fois moins) au modèle 3D de référence sur les portions de géométries jugées pertinantes en termes de structure géométrique.

(ii) Applicabilité et robustesse des méthodes proposées.
Notre méthode de reconstruction 3D repose sur une segmentation top-down d'une image de référence qui elle même, dépend de la détection préalable des principaux points de fuite de la scène dans l'image en question en combinant des indices visuels du domaine image ainsi que d'informations 3D éparse et bruitées issues d'un nuage de point SfM. Notre reconstruction mêle ainsi ces indices dans un schéma d'optimisation mathématique globale pour produire des reconstructions structurées en sortie. L'architecture séquentielle de notre système induit donc intrinsèquement une relative fragilité dans la mesure où chaque étape dépend de la qualité de celles qui la précèdent. Bien que cette fragilité relative ne soit que théorique, une piste d'amélioration pourrait être de combiner toutes les données dans une unique optimisation et qui améliorerait chaque étape du système en fonction de tous les indices en présence, en les liens de précédences entre ces étapes intermédiaires.

(iii) Complétude vs. Structure.
Notre méthode de reconstruction 3D reposant sur un raisonnement dans le repère image, le maillage généré peut potentiellement présenter des artefacts liés aux points de vue, menant à des trous dans les modèles 3D (par exemple en raison d'auto-occultations). Bien qu'il existe des techniques afin de fusionner des cartes de profondeur en un modèle 3D global, ces approches ne tiennent pas en compte les régularités structurelles par vue. Ainsi, une stratégie spécifique pourrait être mise en oeuvre à cet effet.

Nous proposons également trois principales perspectives d'extensions de ce travail :
(i) Selection automatique de vues et segmentation conjointe de nuage de points.
Cette première piste vise à permettre un traitement totalement automatisé du début (acquisition des données) à la fin de la chaîne de traitement (production d'un maillage 3D structuré). Les enjeux sont de sélectionner les images pertinentes pour la reconstruction (images de références ainsi que les images de reprojection associées), en minimisant (voire en supprimant) le recouvrement des vues dans l'espace modèle. Pour ce faire, la prise en compte de la géométrie 3D (points et/ou triangles) dans le processus de sélection doit se faire simultanément.

(ii) Analyse de régularités 2D/3D.
La tâche la plus coûteuse en termes de temps de calcul dans notre système "Patchwork Stereo" réside dans les calculs de coûts de photo-consistence. Afin de soustraire cet élément à notre système ainsi que pour consolider la régularité structurelle (notamment en prenant en compte des co-planarités non locales de patches alignés selon des lignes de fuites communes), l'analyse conjointe de régularités en 2D et en 3D sont des notions complémentaires et mutuellement informatives qui pourraient refondre notre approche actuelle en un processus de reconstruction qui ne nécessiterait qu'une seule image et une source d'informations 3D éparse (ou dense) où les régularités 2D/3D.

(iii) Raisonnement sémantique.
L'intérêt d'une telle perspective est double. D'une part, pour améliorer la qualité structurelle des reconstructions 3D en utilisant le fait que sémantique et géométrie soient des notions mutuellement informatives. D'autre part, de nombreuses applications nécessitent la présence de l'information sémantique en plus d'une

géométrie 3D structurée, dans l'optique de produire une maquette numérique complète de bâtiment.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my thesis advisor, Professor Renaud Marlet for his unwavering support and guidance throughout the years I have spent on this work. Thank you for all your precious advice and mentorship.

Besides my advisor, I would like to thank the other members of my thesis committee: Professor Valérie Gouet-Brunet, Doctor Renaud Keriven and Professor David Fofi for their insightful comments, encouragements, as well as for the fruitful remarks and questions which allowed me to perceive my research from new perspectives.

I also thank my supervisors who contributed to my graduate experience: Stéphanie Derouineau and Mireille Jandon from CSTB for trusting me and allowing me to start this wonderful experience, François Guéna from MAP-MAACC for providing his vast knowledge on building architecture, Olivier Tournaire from CSTB for his technical help and advice. Pascal Monasse from IMAGINE, your office door was always wide open to me whenever I needed help on various 3D geometry, math-related or technical topics. You have also provided me the opportunity to teach C++ programming in the course you were giving at Ecole des Ponts ParisTech. Thank you for entrusting me with this great responsibility and for all your valuable help.

Big thanks to all the IMAGINE team at Ecole des Ponts ParisTech for the great moments spent at the lab. Martin De La Gorce, thanks for all your valuable help and advice as well as for the fruitful discussions on state-of-the-art. Nikos Komodakis, thank you for taking the time to answer the questions I could have regarding MRF optimization and research in general. Thank you Guillaume Obozinski for your contagious enthusiasm and the interesting discussions on machine learning, sci-fi and many other topics; Gabriele Facciolo for the fruitful discussions and brainstorming on shape-from-texture

and 3D reconstruction. Thank you Nikos Paragios for your numerous advice, Arnak Dalayan, Mathieu Aubry, Bertrand Neuveu for your help with accessing documentation and with preparing my defense, and Brigitte Mondou for making our life easier at the lab on a day-to-day basis.

I am grateful to all the fellow Ph.D. candidates for the scientific and non-scientific interactions we have shared at some point. David Ok, Pierre Moulon (for anything SfM-related), Victoria Rudakova, Alexandre Boulc'h. Thanks also to Mateusz Kozinski (façade parsing and optimization), Raghudeep and Francisco (litterature discussions), Yohann Salaun (SfM and line detection), Zhe (calibration and matching); gracias to Marina, Laura and Maria, Shell (graphical models, J-culture and basket-ball), Praveer (Deep Learning and GPUs), Spyros, Sergey, Martin Simonovsky, and Benjamin Dubois (it was a pleasure sharing the same office).

Je souhaite également remercier le CSTB et l'équipe AGE en particulier pour m'avoir accueilli durant mes premières années de thèse. Je garderai d'excellents souvenir de ce passage et de l'environnement qui régnait au sein de l'équipe. Merci Stéphanie et Mireille, Benjamin Haas, Pascal Schetelat, Guillaume Ansanay-Alex, Jean-Marie Alessandrini, Pierre, et Chantal Bodeau pour ton aide précieuse tout au long de mon passage au CSTB Champs-sur-Marne.

I also acknowledge that a significant part of this research has been funded by the CSTB, then by the french ANR project ANR-13-CORD-0003 Semapolis.

Thank you Mohamed Kherat, Ahmed, Mohamed Lamine, David, Matthieu, Paul, and all my friends in Kuala Lumpur, Tok Abah, May, M. Akhirudin and all Kampung Baru: Terima Kasih!

Last but not least, a very special thanks to my family and relatives. My mother, my father and my sister for their boundless love and all their countless sacrifices. Thanks to my aunts, uncles and cousins. A special thanks to my wife for supporting me through this adventure and beyond.

Contents

List of Figures

List of Tables

1 Introduction

Contents

In this introductory chapter, we first examine the importance of 3D models in urban environments by discussing a range of related applications in section 1.1. Next, we present the different 3D representations which can be produced as abstraction models for buildings in section 1.2 and highlight the main challenges involved in the context of urban modeling in section 1.3. We then define the scope of our work in section 1.4 and provide an overview of our main contributions in section 1.5. Last, we provide a succinct overview of how the remainder of this thesis is organized in section 1.6.

1.1 The Need of 3D Models for Urban Scenes

In the ever growing pursuit of digitizing the world in 3D, the automatic generation of 3D building mock-ups is receiving more and more attention from the scientific and industrial worlds. The study of various methods to produce 3D representations of buildings is a particularly active topic in the fields of building architecture, Computer Graphics, Photogrammetry and Computer Vision with a wide range of mushrooming applications. To mention only the most common ones, we will develop how 3D building models have become a pivotal notion in the following contexts, to only name a few: the entertainment industry, building construction & simulation, navigation & mapping, and advertising.

1.1.1 Applications for 3D Building Models

1.1.1.1 Entertainment

One of the most popular and mainstream applications which require 3D models in general, particularly of urban environments, come from the entertainment industry through movies and video games.

With the democratization of powerful hardware in gaming consoles and computers (especially in terms of CPUs and graphics cards), consumer applications and video games produce more and more impressive 3D representations of the world, reaching unprecedented levels of visual and structural realism, to the point where it is hard to distinguish between synthetic images and photographs (Figure 1.1).

In order to achieve such a prowess, studios can spend up to several man-years in manual 3D design, animation and rendering in order to produce 1 to 2 hours of pure synthesis footage. To this end, most of the efforts aim at recovering the maximum detail-level of the depicted parts of the scene with a potential trade-off when run-time or mem-

Figure 1.1: Example illustrations from the video game *Final Fantasy XV*, currently the latest iteration of the *japanese role-playing game* franchise, which has been famous over the years for its fine 3D graphics. Top: The imaginary city of "Insomnia", from a pre-computed 3D footage. Bottom: An example of in-game, gameplay animation, including a dynamic seamless environment [22]. Images are courtesy of *Square Enix Co.*

ory storage constraints are involved. To limit the tedious manual creation of 3D contents and in order to generate realistic and faithful reproduction of existing buildings and cities, the use of automatic Procedural Modeling softwares (PM), e.g., *Esri CityEngine*, has been increasingly used in the entertainment industry. In a nutshell, PM is a tool of massive generation of plausible 3D models that combines highly-parametrizable predefined elements (e.g., buildings, architectural styles, objects) under user-defined rules as illustrated in Figure 1.2. PM for generating plausible cities in 3D has not only successfully been applied in game development (Figure 1.3) and films, but also in other visualization applications, urban planning, Geographic Information Systems (GIS), archeology or even cultural heritage [105].

For such applications, the structure of 3D buildings is a key factor for visual realism, while the simplicity of their underlying geometry is required to minimize the memory storage and allow reasonable (even feasible) rendering time.

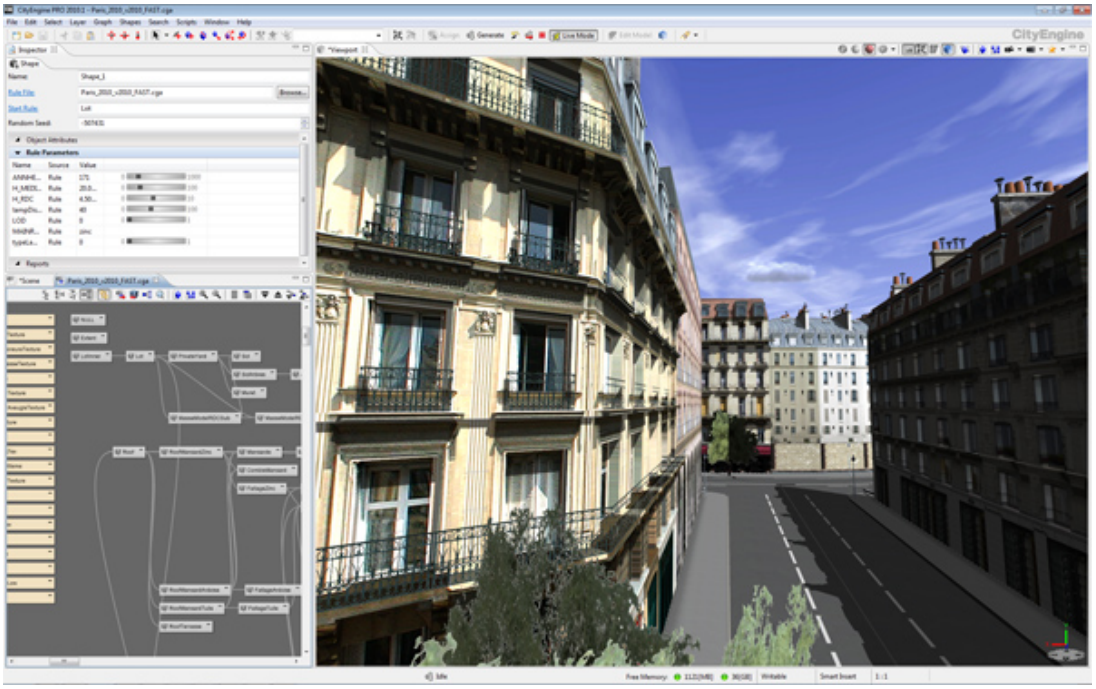

Figure 1.2: Example of a procedural generation of 3D buildings with the *Esri CityEngine* software using default Parisian looks and feels by specifying only a few parameters.

Figure 1.3: *Need for Speed* is a car racing game developed by *Electronic Arts* studios, taking place in 3D urban environments which are generated automatically using Procedural Modeling [105].

1.1.1.2 Building Construction & Simulation

Major application domains which use extensively the notion of virtual representations of buildings are building architecture, construction and the several engineering branches whose aim is to support decisions made on virtual simulations for performance enhancement, management and planning.

Several dedicated formalisms have been developed to unify the different levels of information which are required for such purposes, allowing compact and centralized representations of their 3D geometry and semantic description. For the ends of our study, we first describe the two dominant, standardized approaches which – depending on the application context – can either be seen as alternative or complementary tools.

Additional information on these dedicated formalisms for urban modeling can be found in the comprehensive Ph.D. thesis of Filip Biljecki on this subject [6].

- **Building Information Modeling (BIM)**
  Building Information Modeling (BIM) is a process which aims at modeling, improving and centralizing information on buildings in terms of fine grained 3D geometry and semantics which encompasses contextual and technical character-

istics such as the name, dimensions and functional properties of every elements (e.g., acoustics, energy performance or life-cycle analyses) of a building or its direct environment (e.g., trees, furniture).
The core objective of the BIM philosophy is to allow for multiple views which describe the same building to co-exist within a common data representation in order to improve collaboration, sharing, factorization of efforts, and minimize the loss and redundancy of data, costs and time involved in the design. It has an extremely rich database of element structural and functional properties.Its open file exchange format is the Industry Foundation Classes (IFC) which has been standardized by BuildingSmart (the former International Alliance for Interoperability).

- **3D GIS – OGC CityGML Level of Detail (LoD)**
  The City Geography Markup Language (CityGML, currently in version 2.0) is an open data model for the exchange of virtual 3D city models, standardized by the Open Geospatial Consortium (OGC) with the initial intent of serving the Geographic Information Systems (GIS), i.e., to capture, store, manipulate, analyze and manage spatial data. It is also an information model for buildings, describing their 3D geometry and multiple levels of semantic information [36]. One of the main concepts that OGC CityGML introduces is its multi-scale "Level of Detail" (LoD), which corresponds to discrete descriptions of the quantity and richness of geometric and semantic characteristics of the model providing an adjustable granularity depending on the application needs.
  The different LoDs are specified as follows:

  LoDs 1 to 4 are commonly used for city modeling (Figure 1.4).

  - **LoD0** – 2.5D flat terrain model (typically restricted to GIS-oriented applications).
  - **LoD1** – Block model without roof structures.
  - **LoD2** – Textured, with differentiated roofs.
  - **LoD3** – Detailed envelope model of the building, with openings (e.g., windows and doors).
  - **LoD4** – Highly detailed model including interiors.

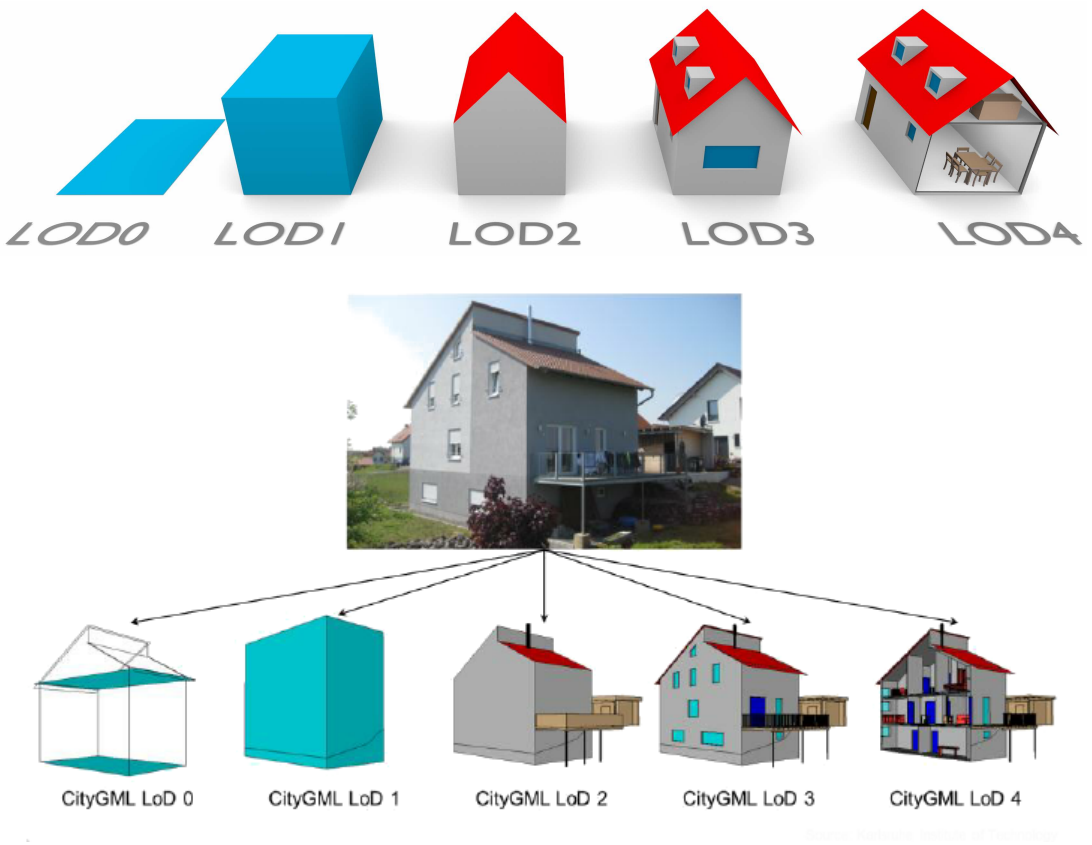

Figure 1.4: Illustration of the 5 different CityGML Levels of Detail (LoDs) from LoD0 (less detailed) to LoD4 (the more detailed). Please refer to the text for details on each level. Top image is courtesy of Filip Biljecki [6], bottom illustration is courtesy of the Karlsruhe Institute of Technology [62].

- **BIM/IFC vs. CityGML LoD**
  The way 3D geometry is stored in BIM/IFC is an element-oriented volumetric model, while CityGML/LoD stores it as a surface-oriented ensemble (Figure 1.5).

  The BIM philosophy has been thought as a bottom-up modeling process: The building is first a concept, then a set of 2D plans, then a full BIM/IFC model, then an actual constructed building until it is destroyed. On the other hand, CityGML/LoD has been intended as an implementation of GIS and is hence a top-down modeling paradigm focusing on the city-scale with a global, adjustable multi-level of granularity in geometry and semantics, i.e., "Level of Detail (LoD)".

  Despite these differences, both standards share several goals and properties, such as: unifying between indoor and outdoor modeling as well as data view in general,

cost reduction and factorization of the efforts made in design, increased analysis and decision-making at the building and urban levels.

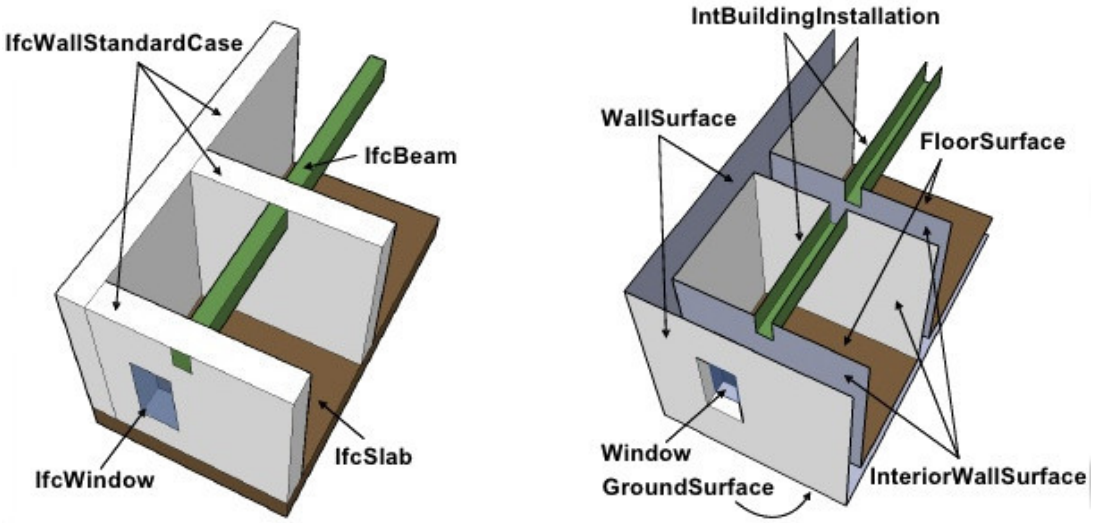

Figure 1.5: Geometric and semantic representations in BIM/IFC vs LoD – Left: In IFC, Geometry is expressed as a set of boolean operations on volumetric primitives, making it well-suited for generative design processes. On the right: In CityGML, the representation of 3D boundaries is an aggregate of observable surfaces of topographic features making it more suitable for modeling observed existing objects. Illustration is courtesy of T.H. Kolbe.

Applications of BIM and 3D GIS (CityGML).

The construction of a building is – in essence – the result of a complex, iterative, multidisciplinary collaboration between many different actors such as architects, engineers, designers and managers. The outcome of this combined effort is a unique, yet complex, building.

Throughout this process and even beyond the conception phase once the building is constructed, each actor works on a local digital representation as a tool for planning and decision-making resulting in a significant time spent in re-designing redundant information which can also be inconsistent with other versions. As a result, minimizing these additional delays and costs in the life cycle of buildings has inspired the advent of Building Information Modeling (BIM) as a common way to design and share information as well as a project management tool.

The main industrial companies who develop software technologies for architects are implementing and investing in the BIM philosophy, e.g., *Autodesk* (*Autocad* solutions,

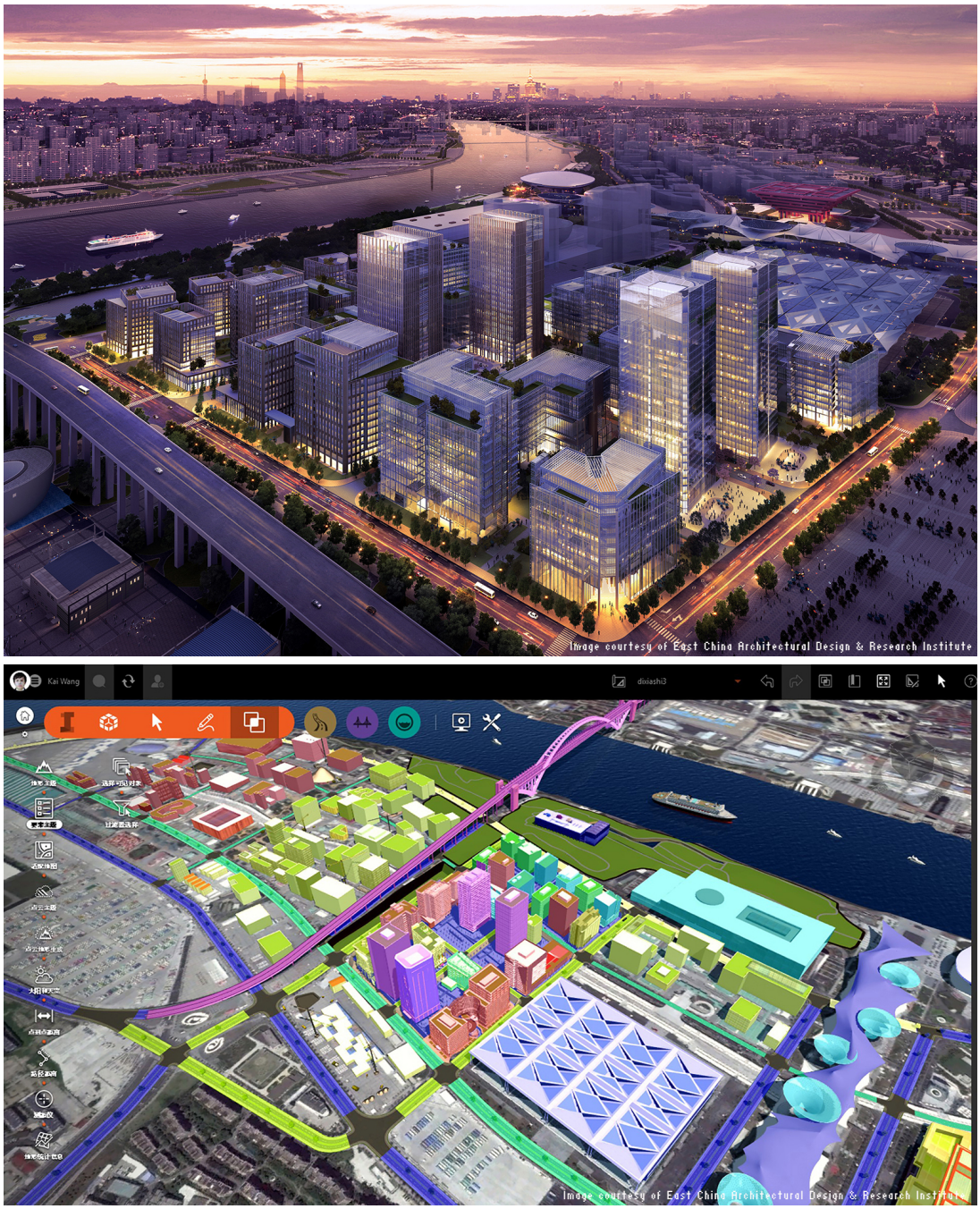

Figure 1.6: Representation of how BIM methodology allowed to a large-scale construction project of 28 buildings to achieve completion in only 4 years. Images are courtesy of East China Architectural Design & Research Institute.

Revit), *Graphisoft* (*Archicad*). Following this yet precursory initiative, a few major construction companies like *Bouygues Construction* have already endorsed this centralized tool and use it for their most emblematic projects. Figure 1.9 shows an example of rendering using a BIM model for visualization purposes.

Despite the growing interest around the world in BIM, the majority of architect agencies – who are the first chain link in the genesis of buildings – are still in the middle of what appears to be a slow transition towards the adoption of BIM-oriented tools which represents and requires a drastic cultural change in their work routines.

Another important line of applications require Information Models (BIM/IFC or 3D GIS/CityGML) for purposes which go beyond visualization use cases, for environmental simulations and decision support [7].

As an illustration of an end-to-end project using BIM in Figure 1.6, 28 buildings were delivered in only 4 years in a construction project in Shanghai thanks to BIM methodology implemented throughout the project*. This also resulted in smart design choices expected to reduce energy consumption by a spectacular 18% and global time and costs gain of about 5%. At an even larger scale, the ecoDistrict of Washington DC has been using BIM modeling extensively for various needs such as to optimize energy consumption, and urban sustainability (Figure 1.8).

Among the countless applications to simulation which use 3D urban models, e.g., CityGML LoD2 or LoD3, we can mention:

- Energy demand estimation and thermal reasoning which can help decide when and where to rehabilitate buildings for global energy performance optimization and reducing costs related to energy loss (Figure 1.7).

- A simple representation at the district-level, using LoD2 can be a meaningful support for simulation, e.g., estimation of shadow cast at a city-level, or analyzing the level of noise pollution, also shown in Figure 1.7.

1.1.1.3 Navigation & Mapping

3D urban models are an essential set of tools for building 3D cadastre databases, mapping for visualization or navigation (*Google Earth, Google Maps, Apple Maps, Microsoft BING...*).

*https://www.autodesk.com/solutions/bim/hub/2016-entry-119

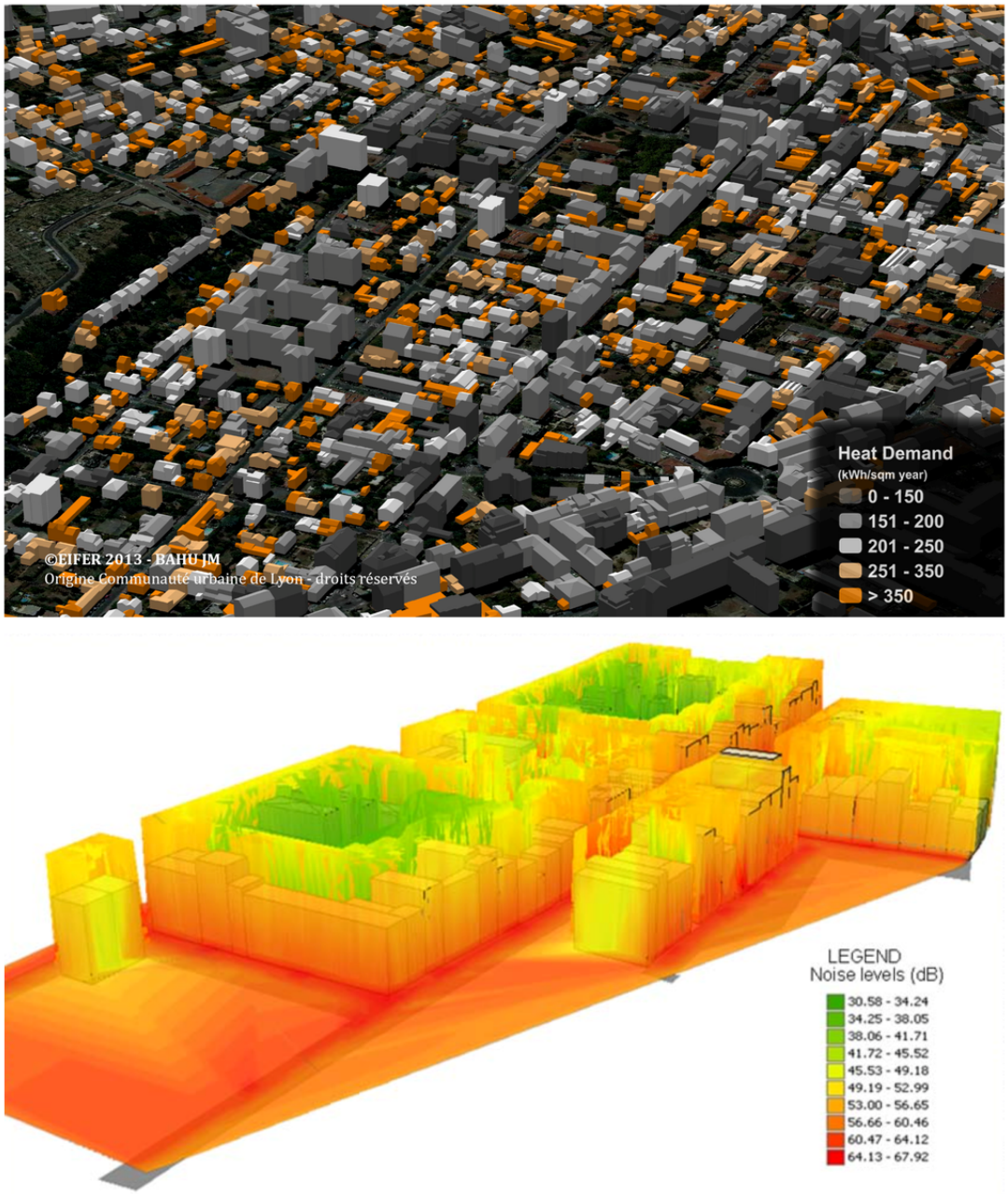

Figure 1.7: How CityGML LoD2 building models can help in energy performance analysis (top) and noise pollution management (bottom).

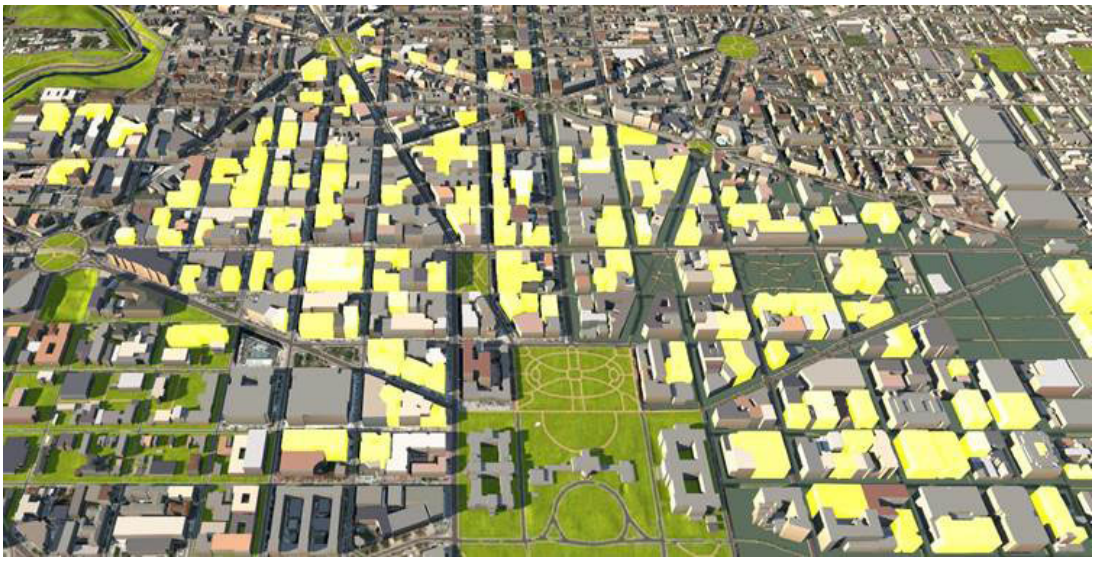

Figure 1.8: Integration of BIM for simulation – Sustainability in downtown Washington DC ecoDistrict using the BIM-compatible software solution *Autodesk InfraWorks360*.

Figure 1.9: The industrial giant *Bouygues Construction* has adopted the BIM/IFC philosophy and uses it for many of their key construction projects throughout the life-cycle of their buildings for decision-making processes as well as for advertising purposes (e.g., to sell apartments to future prospects). Image is courtesy of *Groupe Bouygues*.

The availability of highly accurate open 3D maps of cities is currently paving the way for autonomous driving by only using a bunch of cameras [87]. This technology, provided by *AutoX*, a start-up company from Silicon Valley, is in pole position to win the race of full autonomy in this field. Similarly, maps are also being used for drone autonomous flying, bringing autonomous delivery within reach. In this vein, *Amazon* has already tested their *Prime Air* service in late 2016. Other companies such as the United States Postal Service (*USPS*) are also preparing for this game-changing service. This will open unprecedented perspectives once the service will be fully deployed, where people would receive their orders and mail in less than an hour.

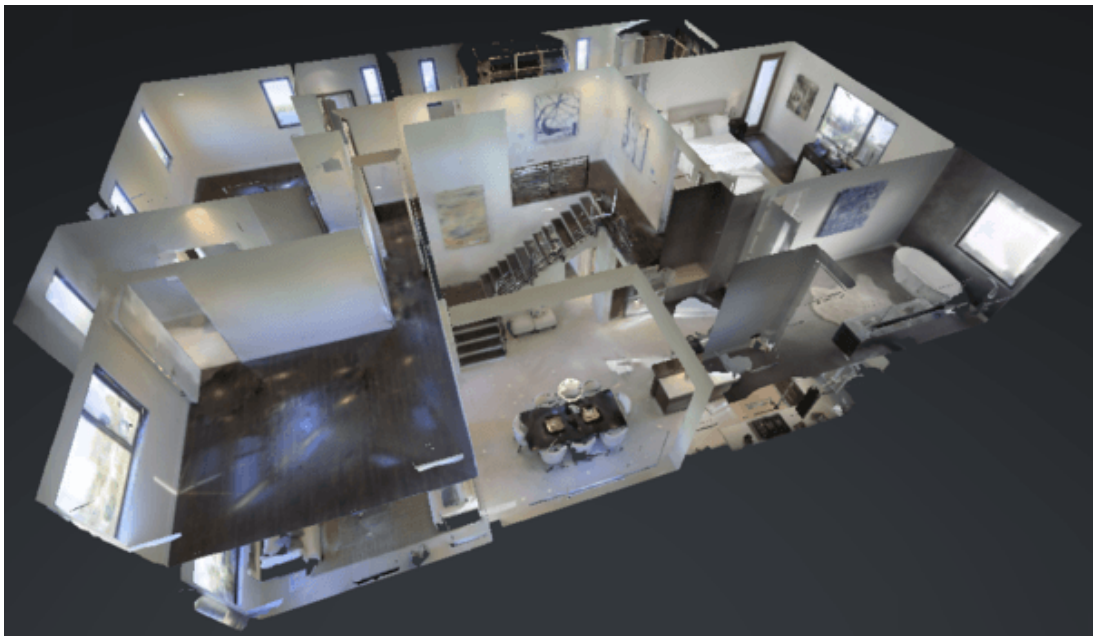

Figure 1.10: Indoor modeling for advertising – *Matterport* is a Silicon Valley start-up company which uses 3D cameras to scan virtual tours of existing indoor scenes such as real estate, hotels, retail...

1.1.1.4 Advertising

The 3D representation of buildings are also vastly utilized for advertising and marketing purposes, for printing 2D renderings of future constructions, like *Bouygues Construction* who uses their BIM models to sell apartments to future prospects (Figure 1.9). Alternatively, applications for virtual visits of already existing buildings are also very well established (Figure 1.10).

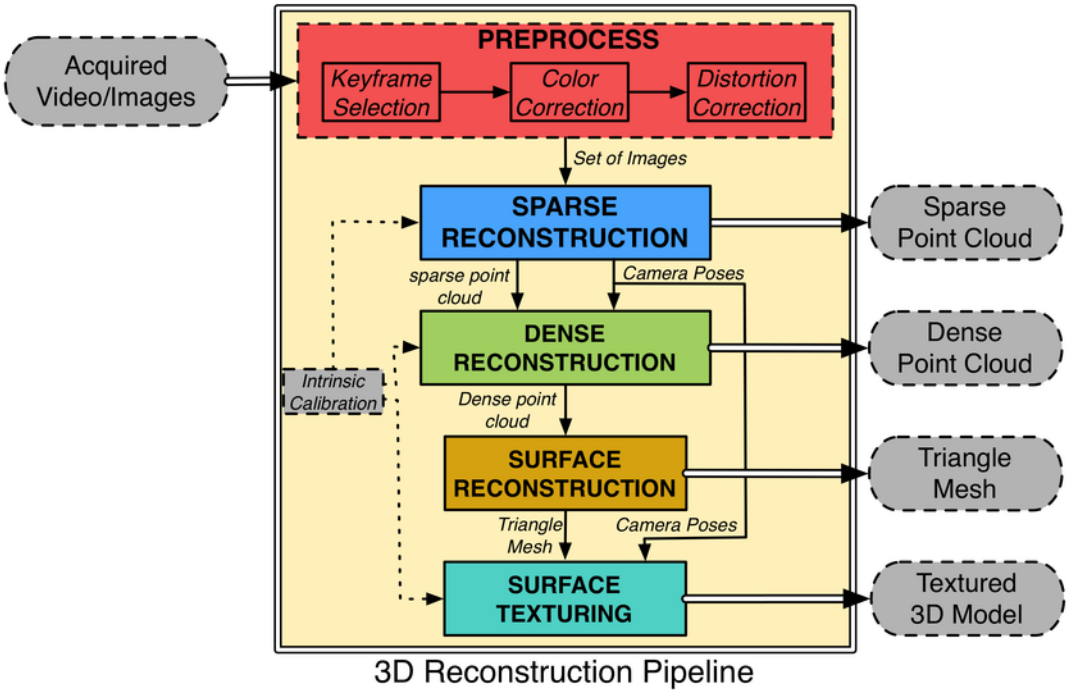

Figure 1.11: Generic 3D reconstruction workflow. From a input sequence of images, the system first estimates the camera poses in 3D and reconstructs a sparse 3D point cloud. The next step is the dense reconstruction, commonly known as Multi-View Stereo which usually produces a dense point cloud which is finally meshed and colorized. Illustration is courtesy of Hernández et al. [39].

All of the aforementioned applications which use 3D models of buildings for visualization or non-visualization purposes have in common the following needs: structurally accurate or plausible geometry, and its scalable, compact representation.

1.2 From Images to 3D Geometry

As a preamble to our discussion on the challenges in Urban Modeling in section 1.3, we first describe the structure of a typical, generic 3D reconstruction pipeline (as illustrated in Figure 1.11) and give a high-level explanation for each step involved. Next, we elaborate on the various 3D representations which can be used as an abstraction of the underlying 3D geometry of buildings.

- **Input Pre-processing.**
  The optional pre-processing may consist in various steps (see Figure 1.11).

First, the input image sequence or video stream is broken down into a subset of selected keyframes. This step not only aims at reducing the computational burden and input redundancy, but it also aims at enhancing the quality of subsequent procedures (camera pose estimations, and 3D reconstruction per se), and limitting the impact of noise [21].

Color correction can reduce the effect of drastic viewpoint and illumination changes which affect the radiometric consistency across images. Image distortion on the other hand, can be very common in urban imagery because of the use of wide viewing angle (i.e., with short focal length) settings. This is in order to capture as much information per image and reduce the scene fragmentation. Correcting such distortion artifacts improves the results in later steps but also enhances the visual structure of buildings by preserving linear features and alignments of objects.

Other frequent pre-processings in urban modeling are: image clustering which consists in the splitting of the input frames into multiple slightly overlapping clusters which can be processed in parallel and merged in order to improve the time and memory consumption [30], image masking, e.g., removing specific parts of the image from specific semantic categories like clutter objects, vegetation or sky pixels.

- **Sparse 3D Reconstruction.**
  Sparse 3D reconstruction is achieved through either Structure-from-Motion (SfM) or Visual-Simultaneous Localization and Mapping (V-SLAM). Both of these approaches basically infer the spatial poses of cameras as well as the underlying sparse 3D representation of the geometry by triangulating matched key-feature points at the image level. The main distinction between SfM and V-SLAM resides in the (near) real-time runtime constraints and potentially restricted hardware for the latter (e.g., on a mobile robot).

- **Dense 3D Reconstruction.**
  Typical methods for dense 3D reconstruction take as input a set of images, along with their corresponding poses (extrinsic calibration) and the sparse point cloud which is produced by the preceding sparse reconstruction step. Out of these input information, traditional dense reconstruction methods – a.k.a Multi-View Stereo (MVS) approaches – either do (i) densify the sparse point cloud in a global opti-

mization / reconstruction [32, 61], or (ii) create an intermediate 2.5D reconstructed depthmap for each image and then produce a global dense point cloud by applying a computationally expensive (yet highly parallelizable) depthmap fusion procedure [35, 84].
Two comprehensive comparative studies of state-of-art dense reconstruction methods have recently been published by Knapitsch et al. [44] and Schöps et al. [85], as well as a concise tutorial on mainstream MVS approaches by Furukawa et al. [31].

- **Surface Reconstruction.**
  Once a dense point cloud is obtained by MVS, a surface (e.g., polygonal mesh) can be reconstructed as a full digital representation of the observed scene by taking into account the 3D geometry of the point cloud [5] and/or information from the 2D image domain [28].

- **Surface Texturing.**
  As a final optional step, image-based texture can be applied on every entity in the reconstructed 3D surface (e.g., triangles in the case of a triangle mesh). While this step is relatively computationally expensive, it is required for visualization purposes where photorealism is as important as structural fidelity of the 3D geometry [2].

1.2.1 3D Geometry as an Abstraction Model

The following 3D representations are presented from lower to higher level of detail and abstraction.

- **Sparse Point Clouds.**
  Sparse point clouds are typically produced during a camera (extrinsic) calibration process via Structure-from-Motion (SfM) or Visual Simultaneous Localization and Mapping (V-SLAM) (which both boil down to simultaneous camera pose estimation and sparse point cloud reconstruction by triangulation).
  While they are easy to produce and scale well, their intrinsically sparse and noisy natures make them poorly suited for many applications requiring higher level of geometric abstraction, accuracy and completeness.

- **Dense Point Clouds.**
  Dense point clouds are provided by either: (i) Dense 3D photogrammetric reconstruction (a.k.a Multi-View Stereo (MVS)) which follows SfM or V-SLAM, (ii) An

active sensor such as a LIDAR scanner, or motion sensing devices (such as *Microsoft Kinect*), or (iii) by sampling 3D points along continuous surface representations. Depending on the acquisition mode, they provide a higher level of abstraction, but are sparsely structured and scale poorly because of the important number of required points in order to describe even simple, e.g., planar 3D regions which are ubiquitous in urban environments.

- **Oriented Rectangular Patches.**
  This is an intermediate representation between dense point clouds and polygonal meshes which mostly presents the same characteristics as meshes, even though the produced geometry is less complete and less smooth when using the rectangular patches produced by semi-dense stereo methods [32, 61] (illustrated in Figure 1.12).

Figure 1.12: Left: Sparse point-cloud obtained through Structure-from-Motion. Right: A dense set of oriented rectangular patches generated by a patch-based Multi-View Stereo method [61]. Illustration is courtesy of Alex Locher [61].

- **Polygonal Meshes.**
  From an initial 3D point cloud, meshing techniques such as Delaunay-Triangulation-based approaches [14] construct a set of vertices (which is a subset of the initial 3D points), edges and faces that define a polyhedral shape of the observed object.
  This representation is by far the most popular trade-off between compactness and structural geometric accuracy and allows flexible post-processing. Most of meshing techniques also have the interesting property of cleaning-up certain categories of point-cloud artifacts which are typically induced by the acquisition process (e.g., due to view redundancy, clutter and self-occluding surfaces,

texture-less areas and specularity). We refer the interested reader to the comprehensive review on surface reconstruction from point clouds by Berger et al. [5] and references therein.

- **Parametric Surfaces.**
  Parametric surfaces can be seen as an extension of polygonal meshes, by assuming a scene to be a composition of pre-defined 3D geometric primitives like cones, planes, cuboids or cylinders that are usually fitted to an intermediate 3D point cloud [5, 58].
  The advantage of such representation lies in the geometric accuracy and completeness, and it also allows a compact memory storage. Nevertheless, the primitive fitting process as well as exploiting and post-processing such models (e.g., adjusting the detail level for rendering/visualization at a large scale) are computationally expensive.

1.3 Challenges in 3D Urban Modeling

The challenges that are specifically inherent to the 3D modeling of urban scenes are of multiple natures which we summarize in this section. For a broader view on the numerous challenges in 3D urban modeling, we refer the reader to the comprehensive survey by Musialski et al. [75].

1.3.1 Acquisition Modes

Depending on the acquisition modes, the challenges as well as the benefits from using them vary.

- **Street-level vs. Airborne Acquisitions.**
  Taking images from street-level viewpoints (e.g., from standing height, or using a mounted vehicle with sensors) is the better option for capturing the maximum detail-level of building façades, as well as the low-altitude objects in the urban environments (e.g., urban furniture, vehicles). However, this does not allow to capture mid-to-high elevation details in the case to skyscrapers as well as the buildings' roofs which are meaningful LoD3 features depending on the applications. Street-level imagery (whether it is acquired from an active or passive sensor) is typically noisy because of omnipresent clutter objects, cars, pedestrians or vegetation.

On the other hand, aerial imagery (e.g., acquired by drones) allows to retrieve important LoD2-4 details like building roofs, higher parts of tall buildings, and large terrain information which can not be seen from the ground and inaccessible courtyards. Façade details however, are difficult to depict due to challenging viewing angles.

Both of these modes can hence be combined in order to achieve a full acquisition of an urban scenery by incorporation top views and all the sides of visible buildings.

- **Passive vs. Active Sensors.**
  Active sensors such as LIDAR scanners give very high-resolution 3D point clouds by emitting a pulsed laser light and measuring the reflected pulses with a sensor at one or multiple wavelength(s). This 3D scan can also be coupled with color information and/or images. Nonetheless, LIDARs are ordinarily slower, less flexible (in terms of mobility) to use and extremely expensive w.r.t using passive imagery. On the other hand, such active sensors provide a much more accurate estimate of depth.

  In contrast to active sensors, passive sensors (i.e., consumer cameras) are a cheap, yet reliable means to do 3D reconstruction by using one of the rich available photogrammetry softwares[†], but the results are more prone to noise and less complete w.r.t using active depth sensors.

1.3.2 Images & Acquisition Process

Exploiting pictures of urban scenery is prone to many challenges. Illumination conditions can influence drastically the quality of photographs which can translate into blur, noise, and other detrimental artifacts. Changes of illumination conditions from one view to another can also confuse the early steps of camera calibration (e.g., detection and matching of key features points). This can be due to many uncontrollable factors such as the time of the day, weather, outdoor artificial illumination, dynamic lighting (e.g., from vehicles).

The drastic changes in viewpoint and/or lack of image overlap (whether the images are taken from the ground or from an aerial viewpoint) can cause local or complete failures at different steps of the 3D reconstruction process. Also, the camera trajectory can have a negative influence on the process, e.g., motion blur due to fast camera

[†]Such as *Pix4D*, *Acute3D ContextCapture* or *Agisoft PhotoScan*.

displacements, capturing dynamic changes (e.g., moving objects), viewpoint-dependent self-occlusions.

Wide textureless and homogeneous areas, like walls, can result in a lack of geometric cues which are necessary to estimate spatial poses of cameras. Additionally, repetitive patterns or objects and specular surfaces (e.g., windows or glass structures) are ubiquitous in man-made environments and are very challenging for the feature matching procedure as they typically can induce false correspondences.

In such environments, wide angle cameras (i.e., with short focal length) are commonly used to capture the maximum amount of information per image. However, this choice translates into potentially significant image distortion artifacts.

1.3.3 Structure and Appearance of Buildings

Buildings come into a multitude of very different sizes, number of stories, shapes, colors, and architectural styles and composing elements. A fully automatic strategy which would cope with all the possible variations of a building's appearance is not realistic (Figure 1.13) and would require (i) focusing on a restricted subset of buildings (e.g., Haussmannian architecture), or (ii) adopting a generic set of priors with the objective to address correctly a significant proportion of input buildings.

1.3.4 Level of Detail vs. Scale

Urban places are very dense and very large by design. As a consequence, the quantity (complexity) of 3D details for modeling a small number of buildings (and hence memory and time consumption to compute and store them) is a trade-off to put into perspective of the problem scale, and how many buildings to address, from a single one to a full city scale.

1.3.5 Full Automation

3D urban modeling tools which are used in the industry and bureaus nowadays all use some sort of manual user intervention in their workflows. This is mainly due to the complex and rich variety of parameters which influence the quality of the input data and the lack of reliable priors to describe the specificities of the depicted urban scene. To cope with the unpractical variability, manual intervention is used to adapt the tools and settings to each scene or each portion of it. Fully automatic 3D urban modeling

Figure 1.13: A *unique* architectural style... – "Building 32" at *MIT*, Boston, MA. It is also known as the Stata Center designed by Frank Gehry, world-renowned architect.

pipelines are still a very active research topic which already provide solutions to real-life problems.

1.4 Scope – Towards Structured, Scalable 3D Urban Modeling

In section 1.1.1, we have enumerated examples of industrial applications which require 3D models of buildings. For the majority of such applications a tedious manual modeling step is necessary in order to produce such 3D models. Alternatively, automatic approaches can output plausible city-scale 3D models from simple parametric rules. Yet, these methods suffer from a crucial drawback. Their expressive power is limited by the need of pre-designed rules and by hard-coded libraries of atomic elements which are combined and parametrized. They can not reproduce any building with any architectural style, shape, material or color.

Additionally, as previously discussed, many applications need information models such as BIM or CityGML of already existing/constructed buildings in order to run analyses and for decision support (e.g., for rehabilitation and/or energy performance optimization by measuring the proportion of glass surface per building façade, which is a widely used quantitative indicator). The automatic acquisition of such models (especially with LoD3-equivalent details) from existing buildings will help realize the full potential of the initial promise of Information Models (BIM/IFC and OGC CityGML LoD, especially LoD3). As a common bottleneck to the aforementioned needs in the industry, we propose to address the automatic Urban Modeling from street-level imagery, aiming at a geometric granularity comparable to LoD3. In this section, we specify the scope of our work in terms of contextual use-case scenarios, building architectural styles, and terminology.

Use-case Scenarios.

We will consider two typical use-case scenarios in the context of which we will provide algorithmic solutions.

In both scenarios, we propose to use the following inputs:

- a sequence of street-level photographs for which the camera poses are supposed a given,
- a sparse point cloud which is typically reconstructed along with the camera estimation of poses during a camera calibration step through Structure-from-Motion.

Such a point cloud can either be reconstructed by using the complete sequence of images, or only a local subset of considered support for structured 3D reconstruction.

Architectural Structure.
The main objective of our work is to preserve the structure of buildings through the 3D reconstruction process without emphasizing a specific architectural style, using soft, generic priors/assumptions. However, not all building structures can reasonably be addressed with weak (generic) a priori knowledge.

We consider highly regular buildings, i.e., which are made of a composition of multiple Manhattan World[‡] models. This suggests that a finite set of dominant Vanishing Directions (VDs) can be extracted in images and describe the overall building layout. As an example, such architectural regularity can be found in the vast majority of constructions in France during the "Thirty Glorious" ("Les Trente Glorieuses" in french). This epoch refers to the thirty years between 1945 and 1975 following the second World War in France when constructed buildings were mostly slabs and tour blocks.

Goal of this Thesis.
We propose to investigate scalable, automatic ways to reconstruct the structured 3D geometry of the envelopes (outside parts, equivalent to CityGML LoD3) of existing buildings from street-level, calibrated photographs using only simple, generic structural priors (which are not limited to a particular architectural style).

Specific Terminology.
Urban Modeling. Through the remainder of this manuscript, we will refer interchangeably to "3D Urban Modeling" or "Urban Modeling", for "3D reconstruction of the envelope of buildings from multiple images" as a shortcut terminology. This term will only include 3D geometry and we will leave the incorporation of semantic information as future work, beyond the scope of this manuscript.

Structure. We define the notion of 3D geometric structure as follows:
(i) Piecewise-planarity, (ii) alignment of the boundaries of the 3D elements with their

[‡]The Manhattan World Assumption states that a scene is made by a composition of boxes where 3D plane orientations are pairwisely, mutually orthogonal or parallel. Hence, in such a context, 3 mutually orthogonal normal orientations of all boxes suffice to "explain" the scene as a whole.

corresponding 2D image gradients and (iii) with principal vanishing directions (VDs), and (iv) co-planarity of elements, and (v) global geometric simplicity.

1.5 Overview of the Main Contributions

- **Robust Extraction of 3D Planar Hypotheses from a Sparse and Noisy Point Cloud**
  Piecewise-planarity is an essential structural trait in urban modeling as well as in man-made environments in general. In order to encode this as an assumption in a reconstruction process, we consider the prior detection of relevant 3D planar hypotheses from available data is necessary.

  Extracting dominant planes can be done very robustly using straight-forward methods by analyzing a dense 3D point cloud which is generally acquired using an active sensor (e.g., LIDAR scanner which is expensive and not flexible), or through an MVS reconstruction (which is time and memory consuming). However, extracting such information from a sparse 3D point cloud (e.g., acquired through Structure-from-Motion (SfM)), or from the image domain is significantly more challenging.

  In this work, our first contribution lies in a robust method which detects 3D planes in a sparse 3D point cloud which is typically obtained during a pre-processing SfM step. To do so, we simultaneously take into consideration information from the image domain: dominant contours as well as dominant Vanishing Directions (VDs) which are strong structural cues in urban scenes. The resulting approach is fast, scalable, and combines information from the mutually informative 2D and 3D domains without additional restrictive assumptions or inputs.

- **Joint 2D/3D Reasoning for Top-down Image Partitioning**
  Image segmentation has been used in the past for scalability as piecewise-planarity priors in MVS. Methods which made such assumptions for the purposes of 3D reasoning and reconstruction typically use bottom-up, unsupervised partitioning of pixels in the image domain. While this allows to handle bigger scenes (in both image resolution and number of considered views) and also, for the contours of the reconstructed objects, to follow dominant image gradients, such segmentation approaches are completely agnostic of the scene's structure. This translates into

blatant visual artifacts, noise, and overly complex 3D surfaces in the final reconstructed models.

As a second contribution to our work, we address this issue by introducing a robust joint 2D/3D reasoning which generates a top-down, structured image partitioning into an irregular lattice topology. This is achieved by combining a set of 3D planar hypotheses which are likely to explain the underlying geometry of a man-made environment, as well as image contours and VDs. The output of the method is a top-down image partitioning and an enriched set of planar hypotheses which are mutually consistent with respect to a given reference camera viewpoint.

- **The Patchwork Stereo Framework**
  Next, we introduce a novel energy formulation in order to reconstruct a piecewise-planar, compact depth map and a mesh which are aligned with the scene's dominant structure using only a handful of wide-baseline views. The method leverages our first two contributions and addresses the problem as a revisit of patch-based stereo reconstruction by using top-down image partition priors. Experiments show that the approach not only reaches similar levels of accuracy with respect to state-of-the-art pixel-based methods while using much fewer images, but also produces much more compact, structure-aware depth map and mesh in a considerably shorter runtime by several of orders of magnitude.

- **Publication**
  The main contributions we propose in this manuscript have been published and presented at an international conference in Computer Vision and Machine Learning.

1.6 Structure of the Thesis

The remainder of this manuscript is organized as follows.

- **In chapter** 2, we discuss the most related lines of work on automatic 3D reconstruction of urban scenes from a sequence of street-level images with a specific focus on structure priors and scalability.

- **In chapter** 3, we present our "Patchwork Stereo", which gathers our main contributions.

- **In chapter** 4, we conclude by giving a summary of our work and discuss its main limitations, perspectives and the potential future lines of research.

2

Survey of Multi-View Urban Modeling

Contents

In this chapter, we present an overview of the most related work in the literature which address the problem of Multi-View Urban Modeling from street-level images. To do so, we first briefly introduce the fundamental topics involved in Urban Modeling, the inputs and outputs of such systems, and will also expose the methodology on which we will structure our survey in section 2.1. In section 2.2, we initiate the discussion by first addressing the general-purpose Multi-View Stereo (MVS) methods. Next, in section 2.3 we turn our focus to the axes of work dealing with ways to cope with scalability which is an intrinsically inherent component of urban environments. We then discuss in section 2.4, how priors are integrated in MVS in order to retrieve the structure of buildings such as the alignment of objects' boundaries and their 3D planar support, and surface compactness. We conclude the chapter with section 2.5 by summarizing the positioning of prior work w.r.t Urban Modeling and the breaches it leaves open for the contributions we propose in this thesis on the aforementioned topics.

2.1 Introduction

Throughout this literature review on Urban Modeling, we will focus on the most related techniques which are used in Multi-View Stereo to produce a dense 3D representation of a scene from a sequence of input images or a video footage with known camera spatial poses. This dense representation, which mainly takes the form of point clouds, polygonal meshes or depthmaps, allows to capture a high-level abstraction of the geometric structure of buildings.

We structure our discussion around three principal axes. Here are the dominant methodological choices we make for each of the three main parts of our literature review.

General-purpose MVS.
We organize this section by presenting a brief overview of the existing groups of approaches and their usability in the urban context by considering them by output and scene representation. The methods we consider in this category are scene and structure agnostic, and are applicable to a wide range of objects and environments beyond urban modeling and street-level imagery. Our objective is to discuss only the most related research in this category – which is extremely vast – and refer the interested reader to the broad overviews of general-purpose MVS available in [31, 88], and the recent comparative analyses of state-of-the-art pipelines in [44, 85].

Scalable MVS.
Next, we select the various strategies which have been used in order to cope with larger inputs in terms of image number, resolution, and output size and resolution. This includes methods which divide or cluster the input images and/or geometry, but also approximate modeling techniques.

Structure-preserving MVS.
Last, we span the dominant works which take into account the scene's structure in MVS. To this end, we include the approaches which leverage structure at the image level, i.e., by integrating piecewise-planarity, alignments of objects boundaries with dominant image gradient and with linear features or main Vanishing Directions of the scene (VDs), co-planarity of visually similar regions, and top-down / procedural / grammar-based methods.

2.2 General-purpose Multi-View Stereo (MVS)

The early stages of Multi-View Stereo (MVS) can be related to the pioneering work of Marr et al. [64] in the 70's, marking the first attempt at formalizing a computational approach for modeling the human stereo vision. This seminal work has paved the way for what has become one of the fundamental problems of modern Computer Vision. Stereo-vision through two-view, or its natural multi-view extension is still one of the most active research topics to this day [88], along with semantic recognition and segmentation [63]*. Yet, the most rudimentary form of general-purpose MVS is achieved by inferring pixel correspondences across images by comparing pixel appearances through photometric consistency measures (photo-consistency in short) [41].

From images to depthmaps.
Photometric pixel matching has been leveraged by a first series of approaches that work on pairs of images which are first rectified, i.e., re-projected on a common image plane. The stereo reconstruction task is then posed as an optimization problem where each pixel from the left image is labeled with a discrete disparity value (inversely proportional to the depth from the optical center) which associates it with a pixel from the second image. The most basic method to retrieve such a disparity-map for all pixels is through the "winner-takes-all" strategy, i.e., by computing all the possible disparities

*Or, as they are respectively refered to by Malik et al., "Reconstruction, Recognition and Reorganization".

along a horizontal scan-line (making use of the image rectification) at each pixel (or by considering a small neighborhood centered on the pixel of interest, for robustness) and assigning the one with the lowest matching cost. In practice, this only gives a coarse geometry because such costs are usually relatively noisy due to several challenging factors, such as: inaccuracies in camera pose estimations, wide baseline between views, illumination changes or occlusions; hence, the uniqueness of matching points in a pair of images is not guaranteed. More sophisticated variants are more robust to these sources of noise by taking into account a first order [46], or second order [106] smoothness on pixel neighbors in an MRF discrete labeling.

The produced depthmaps are view-dependent representations of the 3D reconstruction and require multiple neighboring views to compute a single depthmap. Given, the pixel-based nature of the photoconsistency computation, even by considering a small square window around each pixel for an increased robustness, such methods are limited to relatively narrow baselines between views. Even though dedicated descriptors have been proposed to limit the sensitivity to wide baseline in matching [96], their applicability remains relatively limited in the context of street-level imagery where changes in viewpoints are typically very strong [75].

Reconstruction of point clouds and oriented patches.
PMVS [32] is one of the most prominent MVS method and among the most popular ones. The method takes as input a set of calibrated images and produces a set of oriented rectangular patches in three steps. A first sparse set of patches is extracted by leveraging 2D features correspondences between views. Then, an expansion step densifies the sparse cloud by iteratively estimating the patch geometry by optimizing a photometric score. The final step filters out outliers. The main limitations of the approach lies in the computationally expensive expansion step which relies on photo-consistency and its high sensitivity to texture-less areas and specular surfaces. Its applicability to urban scenes remains restricted to moderate sized scenes with sufficient texture.

Other state-of-the-art methods first compute dense depthmaps and fuse them to a unified, global 3D point-cloud reconstruction [35, 84]. Even though these methods yield state-of-the-art pixelwise accuracy [44, 85], they scale poorly despite the use of GPU acceleration [84] and efficient parallelization [35]. Additionally, point clouds are not suited for urban modeling for a range of applications, and the fusion strategy which allows to obtain them does not handle structure [18].

Volumetric MVS.

Volumetric methods either reason in terms of voxels, or cells in a cell complex. Voxels are entities which represent a value in a three-dimensional regular grid. On the other hand, a cell complex is produced by the full arrangement of 3D primitives, e.g., planes.

The Shape-from-Silhouette framework assumes a 3D parametrization into a voxel space [26, 54, 55]. Each contributing image is segmented into a binary front-ground/back-ground mask indicating the "silhouette" of the object of interest. Then, the 3D volumetric space lying at the intersection of all the silhouette-induced visual cones from each image is considered for the final reconstruction. Iteratively, for each image, every 3D voxel (i.e., the spatial coordinate of its centroid) not reprojecting into the silhouette of a given view is carved away. The final reconstruction lies in the sett of all the remaining points and the volumetric representation allows to further generate a point cloud and/or a mesh from the output geometry [18]. Since this seminal work, many extension have been published using, MRF graph-cut resolution using photoconsistency [101], or using the visual hull as a constraint in a deformable model formulation [23].

Shape-from-Silhouette-based methods have several limitations though, with respect to our purpose of modeling urban scenes. First, they require a large density of cameras spreaded around the object of interest in order to yield sufficient visual constraints. In an outdoor scene, this would limit their applicability to isolated buildings surrounded by narrow-baseline views with narrow fields of view. In a street-level scenario though, these approaches have limited suitability. Next, they are not suited to preserve geometrically concave details and they are sensitive to the accuracy of the silhouette extractions as well as to the resolution of the voxel space. And last, the method would require an additional post-processing in order to implement our desiderata in terms of structure, which would be to the expense of an additional computational burden.

Using a volumetric representation on a Delaunay Tetrahedralization (DT) computed on a quasi-dense point cloud, Labatut et al. [51] leverage the volumetric arrangement into an MRF graph topology and formulate the surface reconstruction problem as an energy function based on the surface parameters and visibility information. The global energy is solved using graph-cuts [47]. The method takes a few minutes to compute 300 input images but requires a computationally expensive quasi-dense point cloud and it still produces an overly complex geometry for man-made scenes. Similarly, Chauve et al. [15] exploit an MRF topology on the full arrangement (i.e., the cell complex) made of

3D planar hypotheses which are extracted from an input dense point cloud and retrieve a piecewise-planar geometry. However, the method requires a dense point cloud which is in itself, already the result of an end-to-end MVS pipeline. Consequently, the quality of the reconstruction highly depends on the quality of the input point cloud which production has its own limitations.

2.3 Scalability

Urban scenes are partly characterized by the fact that they are large and dense. In order to compute, store and represent the 3D geometry of buildings at a street, district or even city scale, standard and straightforward 3D reconstruction methods quickly become intractable and require specific attention [75]. In this section, we focus on the strategies that have been utilized in the literature to make efficient and compact MVS reconstruction feasible, for the purpose of modeling urban scenes.

Scalable MVS methods primarily aim at reducing the computational burden and memory consumption implied during the reconstruction process which, as a by product, allows to handle larger inputs in terms of number of considered views, or even in image resolution [40]. Such strategies can be roughly categorized into three groups of methods.

A first group of works tackle the efficiency aspects in standard MVS methods. In terms of parametrization of the 3D model space, adopting view-dependent representation, i.e., depthmaps is much less computationally expensive alternative to volumetric representation, e.g., into voxels (the 3D extension of pixels), or using a 3D cell complex (which also model 3D volumes through cells formed by the intersection of a full arrangement of 3D primitives). Reasoning on depthmap, even as an intermediate step, allows a straight-forward parallel computation. On the other hand, several state-of-the-art pipelines [35, 84] separate the global MVS task as a sequence of view-dependent depthmap representations and then, apply a fusion strategy (e.g., TSDF-like fusion [18] which merges depthmaps into an intermediate voxel volume which can be further used to produce detailed 3D surface meshes and point clouds). This procedure aims at minimizing the geometric inconsistencies between independent views but also to reduce the impact of noise and clutter.

A specific attention has also been given to efficiency in large-scale optimization techniques which are commonly used in MVS [86].

A very recent work focuses on "progressive MVS" [60, 61]. By building upon oriented patch-based MVS [32], the geometry gradually expands and its completeness progressively increases, the longer the algorithm is given time to compute by imposing a priority per-patch, hence, inducing an algorithmic trade-off between structure completeness and computational runtime.

The second group clusters the sequence of input views into smaller overlapping batches, yielding smaller manageable, independent sub-problems [3, 30, 66]. These methods however, exploit the overlap between views which results into geometric redundancy. This geometry redundancy is a means to increase the quality of reconstruction in both accuracy and robustness around the cluster junctions, but also produces geometric redundancy which spoils the structural aspect of the merged models. To address this particular issue, other works jointly address the camera clustering problem along the one of geometry clustering. Zhang et al. [113] first reconstruct a coarse mesh from an input SfM point cloud and jointly cluster the input views and the corresponding mesh in a constrained energy minimization by optimizing per mesh-face criteria such as: smoothness, size, and coverage in terms of camera visibility.

The third category of approaches improves scalability to the expense of the structural accuracy of the produced geometry. Such approximate modeling techniques include model-based methods which represent building façades as a composition of planes: one per façade [4], 2.5D heightmaps [79], n-layermaps [34] or by exploiting the orientation of buildings with respect to the ground plane [16, 78]. Another popular method to produce an approximate geometry in man-made environments is through superpixel-modeling techniques [11, 68, 69]. These approximate modeling methods are intrisically linked to structure priors, hence we discuss such works in more details in the next section.

2.4 Structure Priors

In this section, we review the most prominent lines of research that address the priors which encourage or enforce the following notions of structure in MVS, i.e., (i) piecewise-planarity, (ii) alignment of the boundaries of the 3D elements with their corresponding 2D image gradients and (iii) with principal vanishing directions (VDs),

(iv) co-planarity of elements, and (v) global geometric simplicity.

The assumption of a piecewise-planar 3D geometry has become a very popular prior in the MVS literature [29, 33, 50, 68, 69, 91] for several reasons. First, it encodes a significant part of the structure present in man-made environments such as, for example, indoor and urban scenes which are mainly composed of planar elements. Secondly, this simplifying assumption on the scene's geometry allows to reconstruct regions where data is either missing or noisy, by propagating the existing reliable information along planar structures, making the prior applicable even to non-planar scenes [50]. Thirdly, the local smoothness in pixel assignment to similar planar structure enforces the global simplicity of the underlying geometry.

Since the seminal work of Wang et al. [104] on layered motion models which have laid the foundation of piecewise-planarity in stereo vision, several authors have further generalized their model to rigid MVS [8, 94]. In this trend, the scene is modeled as a collection of primitives across views in the presence of discontinuities (i.e., occlusion boundaries) by iterating between an image partitioning step and the assignment of each segment with a refined planar hypothesis. Nevertheless, the pairwise relationship between spatial neighboring entities is not taken into account, thus limiting their applicability to very simple scenes.

MRF pixel-based modeling.
Markov Random Field (MRF) optimization is an elegant and theoretically principled tool to model the local spatial relationships between objects in the image domain (e.g., between pixels or superpixels). Additionally, many top-performing methods in the Middlebury stereo challenge are based on MRF optimization [83, 88]. However, initial methods like [46] consider a first-order smoothness prior between pairs of neighboring pixels in an image by assuming fronto-parallel surfaces in the final depthmaps, hence, limiting the quality of piecewise-planar geometric transitions.

Woodford et al. [106] propose to integrate a second order smoothness prior, i.e., modeling the interaction between triplets of pixels instead of pairs as in traditional methods, in order to overcome the limiting assumption of fronto-parallel surfaces. However, the second order smoothness leads to a more challenging optimization and the authors propose a sophisticated inference scheme to reconstruct the final piecewise-planar depthmap based on fusion moves [56].

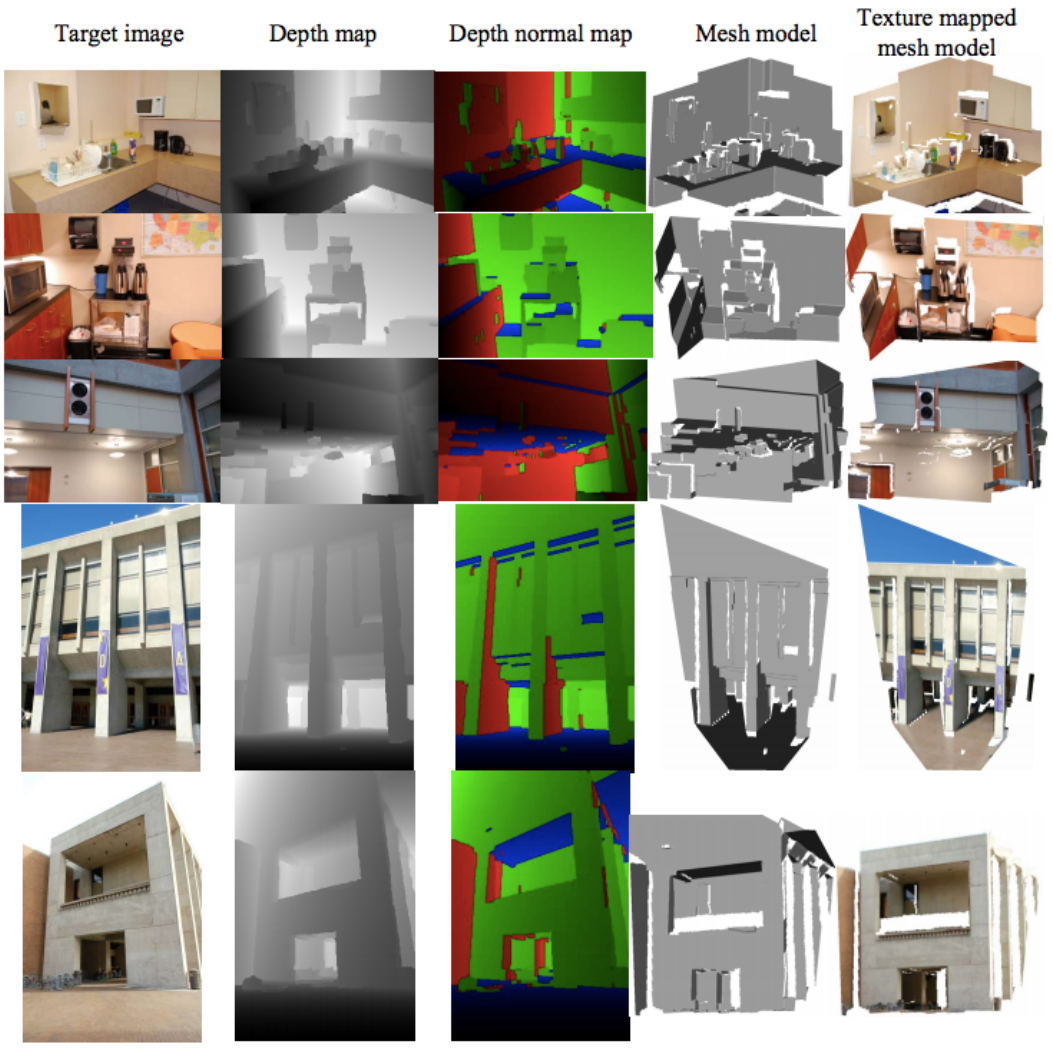

Figure 2.1: Manhattan World Stereo [29] – Their pixel-based method takes a dense point cloud [32] and a collection of calibrated images and reconstruct a structured, piecewise-planar depthmap using an energy minimization which favors geometric planar transitions which are aligned with Manhattan frames. In the reconstructed meshes, each pixel is split into 2 triangles.

Furukawa et al. [29] assume a Manhattan World Scene[†] and greedily extract a set of 3D planes (oriented along the considered Manhattan VDs) from a dense point cloud acquired by a general-purpose MVS method [32]. Then, they assign a planar hypothesis to each image pixel by encouraging Manhattan transitions along strong image-based gradients and edges pointing towards one of the three dominant Manhattan VDs (Fig. 2.1).

[†]i.e., that the environment can be fully explained geometrically using only 3 mutually orthogonal normal orientations.

The method is a, effective tool to "inpaint" the missing 3D information of the input dense point clouds in flat, textureless or specular areas by propagating the available evidence along the Manhattan directions.

Sinha et al. [91] extend this reasoning beyond the limiting Manhattan World Assumption and only require a sparse SfM point cloud alongside the input images to operate (Fig. 2.2). This is done by first extracting and fusing multiple VDs from nearby views (w.r.t the given reference view for which the depthmap is reconstructed) and recovering 3D planar hypotheses by conjointly using reconstructed 3D vanishing lines, plane fitting to the sparse SfM data and additionally, by creating hypotheses which form crease junctions along dominant VD-alined edges from the reference image's viewpoint. Their optimization combines pixel-wise photoconsistency, and the available sparse 3D information through geometric and visibility consistencies. The smoothness term they propose favors plane continuity and crease edges allowing discontinuities along strong line segments and vanishing lines.

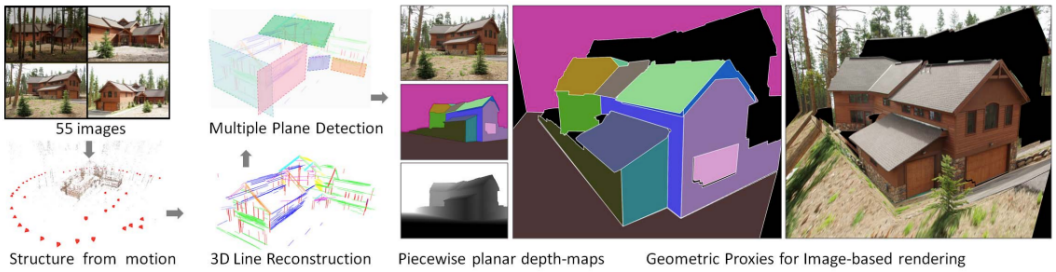

Figure 2.2: Sinha et al. [91] – Overview of their pixel-based approach. Multiple planar hypotheses are extracted from SfM points and lines and piecewise-planar depthmaps are reconstructed by encouraging planar transitions to lie along dominant image gradients and VDs.

In order to seamlessly handle piecewise-planar and non-planar geometry, Gallup et al. [33] leverage the classification of pixel appearance and pixel depth from dense depthmaps acquired from temporal stereo in order to label image pixels in street-level imagery of residential areas into planar and non-planar (which mostly consists in vegetation). This labeling is then incorporated as a binary prior to respectively switch between a piecewise-planar reconstruction by approximating such regions by planes, and the rough output from the initial depthmaps. The method works well when sufficient narrow-baseline views are available and when the parsed scene has low appearance variation.

In summary, the structure-aware pixel-based methods we discussed have successfully been used to model structure. They are mostly bottom-up methods, and the addition of top-down primitives such as lines and VDs [29, 91] allow to model the alignment of objects boundaries in the image plane as well as with VDs. However, reasoning at the pixel level is computationally expensive and unscalable, lacks robustness against strong viewpoint and illumination changes and the absence of texture. For these reasons, such methods rely on an overwhelming regularization in their global energy minimization.

Model-based reconstruction.
In a different vein, several works which focus on the reconstruction of buildings have leveraged the 3D orientation of building façades and the supporting ground. Pollefeys et al. [78] first detect the up gravity vector of the scene and project SfM points on the ground plane and estimate the 2D rotation parameters around the up vector. Once the two main orientations characterizing the building are retrieved, they compute dense depthmaps using a "Plane Sweep Stereo" approach on GPU, i.e., by computing dense photoconsistency on all the pixels by assuming plane-induced homographies by varying 3D planes (i.e., that are "swept") along the discretized set of normals through a discrete range of plane offsets which is set using the sparse SfM information. They finally fuse the depthmaps of nearby views using visibility constraints. Cornelis et al. [16] assume a canyon-like urban representation from street-side imagery with a planar ground and vertical surfaces for façades, allowing real-time modeling. Also assuming one vertical plane per building façade for fast modeling, Barinova et al. [4] consider vertical vanishing lines in a single image as candidates for façade/planar junctions. Similarly, Gallup et al. [34] assume a n-layer heightmap to model buildings, whereas Pylvanainen et al. [79] simplify the geometry even more, to a 2.5D heightmap.

These methods provide an interesting speed-up w.r.t purely pixel-based methods, but the over-simplification in terms of geometric structure, discards significant structural details in the final reconstructions, limiting their suitability to very simple scenes and/or when running nearly in real time is a requirement.

Superpixel Modeling.
Superpixel modeling techniques not only speed up the reconstruction process by considering fewer entities per image than pixels, but they also offer the benefits of an in-

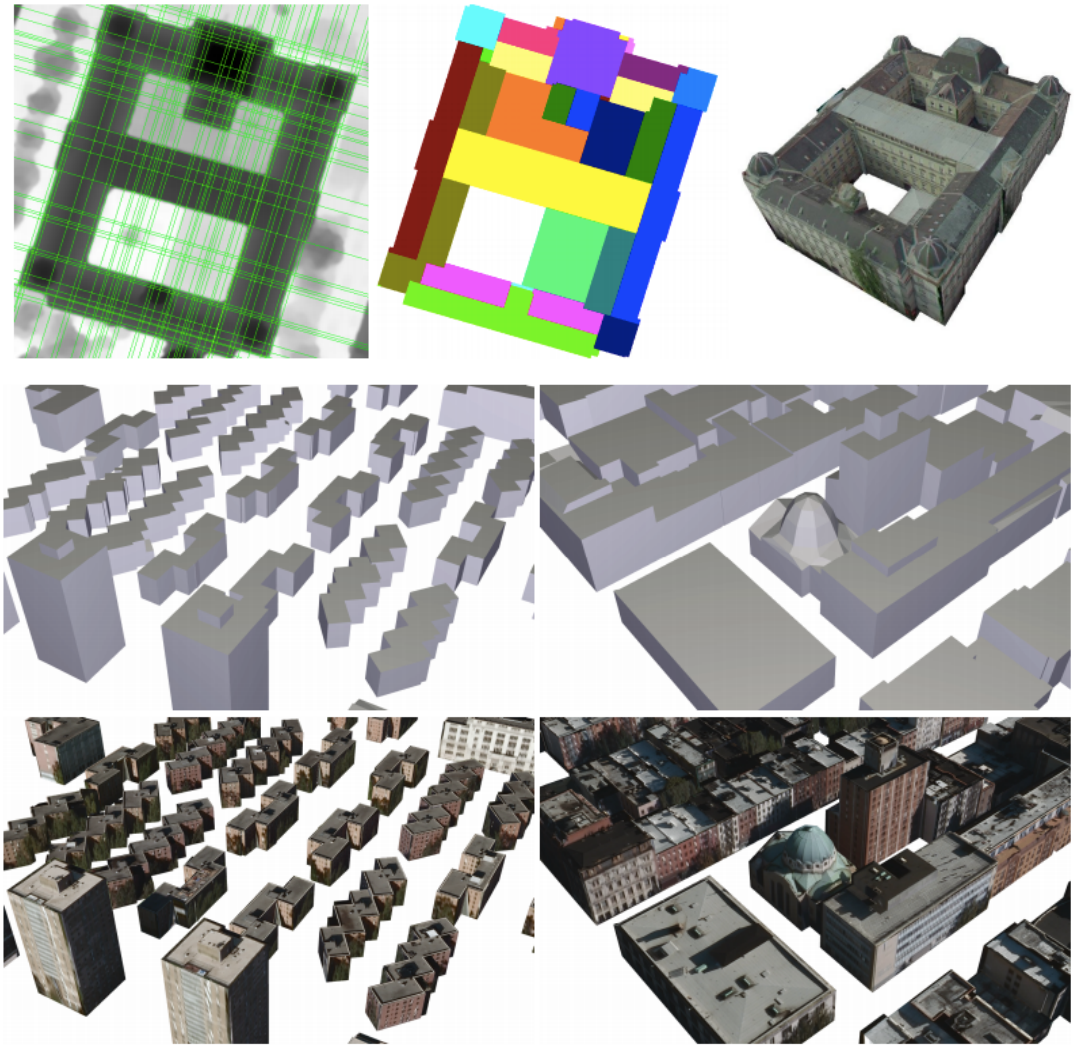

Figure 2.3: Zebedin et al. [111] – The method takes a binary mask indicating the building shape, a dense depthmap, and an image and segments the latter into a 2D rectangular grid using an arrangement of structural lines and produces a structured depthmap by fitting planes and surfaces of revolution to the image segments. Top row, from left to right: the segmented input depthmap; the region labeling after complete inference of the model; final textured result. Middle and bottom rows: additional results.

crease in robustness in the challenging urban context [75], i.e., due to strong changes in viewpoint and illumination, lack or even absence of image texture, or in presence of repetitive patterns (e.g., bricks or windows, which confuse conventional low-level feature matching).

In their seminal work, Birchfield and Tomasi [8] assume the scene to present slanted surfaces and alternate greedily between image partitioning and an affine parameter fitting step on each image segment. After convergence of the algorithm, the result is a segmented piecewise-planar depthmap, explaining the scene with a low number of planar elements.

More recently, Zebedin et al. [111] have successfully combined dense depthmaps and an image partitioning which leverages 3D line matches into an irregular 2D grid, and assign 3D primitives to the induced 2D superpixels (i.e., planar primitives and surfaces of revolution, Fig. 2.3). Their global energy formulation produces impressive digital elevation models of buildings from aerial images, but restrictively requires inputs like an accurate delineation mask for each considered individual building as well as dense depthmaps.

Mičušík and Košecká [68, 69] introduce the "Superpixel Stereo" framework which uses a conventional image partitioning into bottom-up superpixels using a graph-based segmentation method [24]. Then, they design an energy formulation to reconstruct each of such superpixels in 3D using plane-sweeping stereo along with a 3D orientation prior (assuming a Manhattan World Scene) which reasons on the 2D shape of superpixels w.r.t vanishing points [17] (illustrated in Fig. 2.4). The final optimization uses a first order smoothness between nearby superpixels by encouraging neighboring superpixels to touch in 3D and to share a similar surface orientation. Even though the method allows to cope with large-scale urban scenes by producing a coarse, piecewise-planar geometry which can be sufficient for fast approximate modeling for visualization purposes, the initial over-segmentation is agnostic of essential structural alignments such as vanishing directions, which are ubiquitous in the urban environment.

Bódis-Szomorú, Riemenschneider and Van Gool [10] extend this principle by also using bottom-up segmentations [1], but through a multi-image model where all the considered views are segmented in 2D, and produce a dense, piecewise-planar approximation of street-level scenes. They do so by propagating sparse visibility information in a simultaneous multi-view plane assignment problem where they solve jointly for the superpixels across all views, avoiding expensive photoconsistency computations.

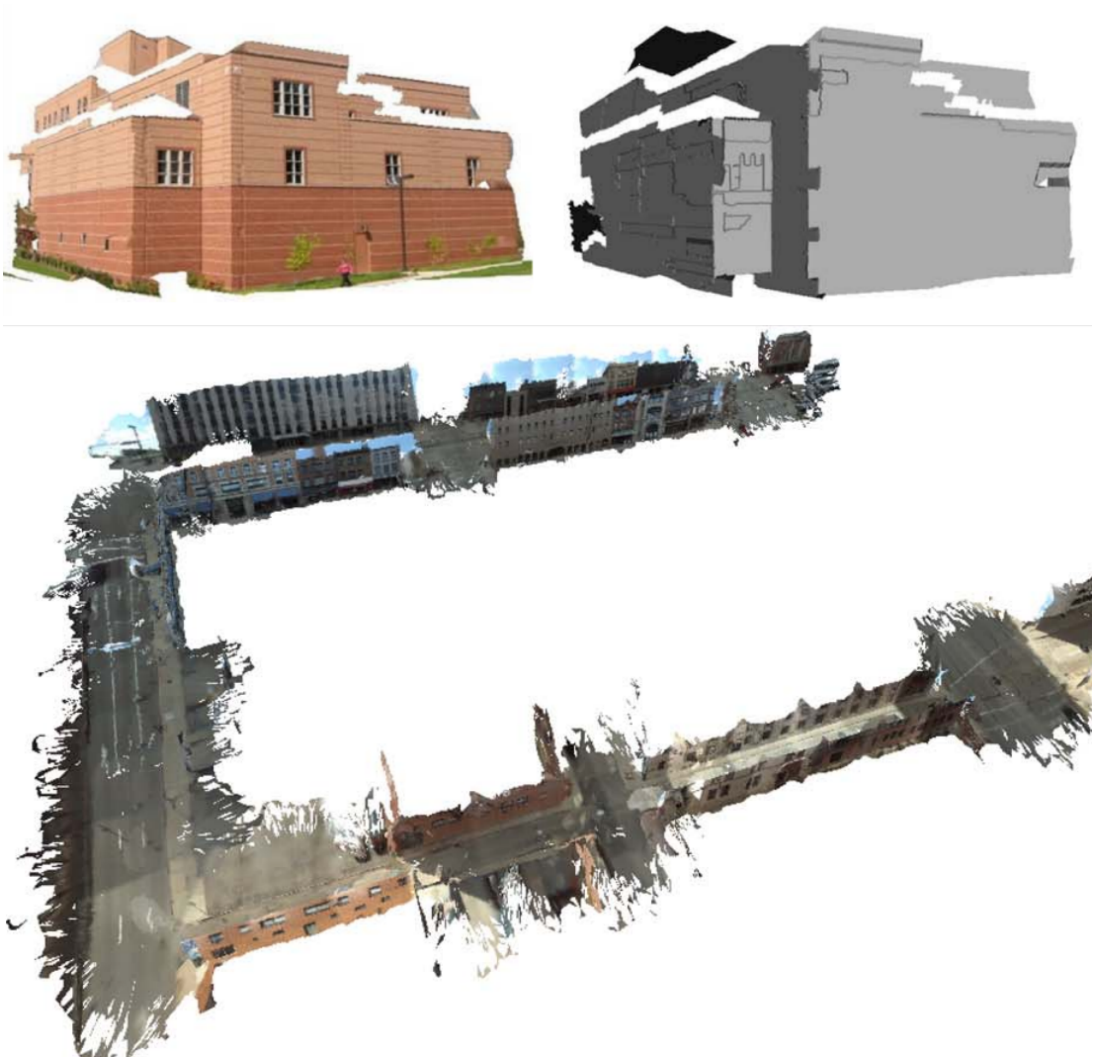

Figure 2.4: Superpixel Stereo [68, 69] – The method combines a computationally expensive plane-sweep stereo, constrained by the Manhattan World Assumption with a regularization which encourages superpixels to touch in 3D and the share the same orientation. Top row: reconstruction of the GMU-building dataset [67] with sky pixels manually masked out by the authors. The method relies on a bottom-up superpixel segmentation [24] which is detrimental to the building's alignment with VDs. Bottom row: a large-scale approximate modeling of streets.

However, computing correspondences between superpixels, and the use of bottom-up superpixels [1] which tend to produce hexagonal shaped regions in textureless areas, are detrimental to the final structure.

Figure 2.5: Left column: The method [11] takes a single view and a corresponding SfM point cloud, segments the image into superpixels [24] and adapt them to the sparse point cloud by penalizing surface curvature. Top row: results of [11], bottom row: reconstruction by PMVS-2 [32].

The same authors [11] introduce an alternative to plane-sweeps [69], multi-view plane fitting [10], and to the use of dense (or semi-dense) 3D inputs [29, 100, 111], in superpixel modeling. They approach the problem by using an unsupervised image partitioning (e.g., [24]) and treat the reconstruction problem as a joint single-view segmentation and a plane fitting one over SfM points and adapt the 3D shape of the superpixels by penalizing surface curvature of the reconstructed regions(as illustrated in Fig. 2.5). The method is very fast and the bottom-up superpixels are mostly aligned with dominant image gradients, generating a compact geometry. Nevertheless, important features such as the planarity of superpixels as well as their co-planarity and alignments, and the notion of vanishing directions are not taken into account.

In a very recent work, Verleysen and De Vleeschouwer [100] use a single pair of wide-baseline images and a dense, yet unsructured and noisy depthmap computed from the initial pair. One view serves as a reference image for the structured depthmap computation and is segmented using a color-based over-segmentation method [98, 99]. Next, the authors pose the problem of piecewise-planar reconstruction with a segmentation prior as a multi-model fitting in the iterative PEaRL MRF-based framework [19]. The method iteratively solves multi-label planar assignments with an explicit MDL prior (i.e., minimum-description length, penalizing the complexity of the inferred solution) and updates the pool of considered planar hypotheses by re-estimating them on the set of the superpixels which were labeled as co-planar. The approach is well suited for applications such as image-based rendering and works with a single pair of wide-baseline views but requires a dense depth-map and uses of a regularization which only encourages planar continuity, ignoring important structural features, such as crease transitions.

All of the aforementioned superpixel modeling techniques provide an interesting speed-up in the reconstruction process, a global increase of robustness and favor piecewise-planarity and geometrically simple solutions. However, they suffer from a two-fold drawback regarding structure: (i) the intra-superpixel planar homogeneity assumption is often broken in practice, and (ii) the alignment of boundaries with structurally meaningful contours such as VDs is totally absent from the segmentation criteria in unsupervised bottom-up methods (e.g., [1, 24]) which are widely used by superpixel modeling methods [68, 68, 100].

Procedural rules and grammars.

Another trend in the literature uses a set of hard-coded rules or grammars to process the input by successively applying corresponding procedures in a top-down fashion.

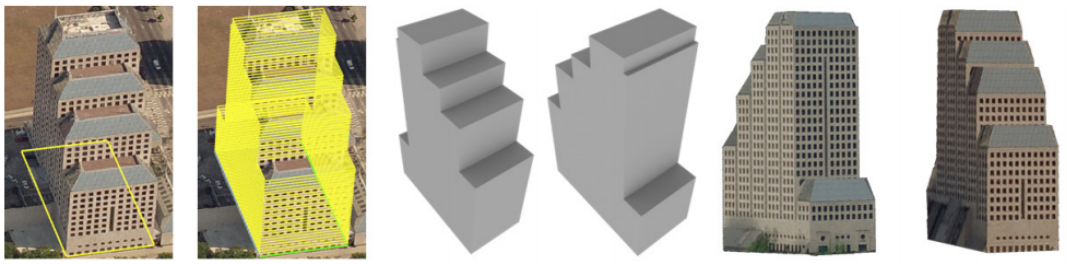

Figure 2.6: Vanegas et al. [97] – Left: one of the input images and the polygonal footprint of the building of interest along with the footprint sweeping. Middle: final volumetric, watertight reconstruction views. Right: Textured results.

Vanegas et al. [97] propose to reconstruct skylines from aerial oblique imagery using three simple, hard-coded Manhattan rewriting grammar rules, encoding: L-shape, U-shape and push-back geometric transitions. They leverage the appropriate rule while sweeping the 2D polygon which represents the building footprint from the ground up, and analyzing the changes in the 1D image-based signal. The final reconstruction is a closed-surface, watertight geometry.

Addressing street-side imagery using top-down image segmentation into irregular grids, Xiao et al. [110] exploit contours and line segments in the image domain to recursively subdivide every façade into rectangular units from street-level imagery and optimize the depth of every cell through an MRF formulation using SfM cues. However, in order to cope with robustness issues, the authors make use of manual intervention during the segmentation process.

Müller et al. [74] also perform a top-down image partitioning but use it on a single rectified façade image. This is done in three steps by first detecting the dominant split lines which, once combined, yield a top-down partitioning of the façade into rectangular irreducible tiles. Next, they group tile elements by symmetry and further subdivide the tiles similar to [73]. The methods generates impressive, structured façades but is limited in terms of inputs, to highly regular, mono-planar façades which present a single dominant grid of aligned architectural elements. In turn, the seminal work of Müller et al. [74] has been followed by many other extensions, e.g., to cope with more complex façades by splitting the input into several layer-maps where symmetry is maximized per layer [112]; or to handle even non-planar façades and architectural elements [42].

Semantic inverse procedural modeling.

Inverse procedural modeling techniques assume an ortho-rectified façade image as input and instantiate the parameters of grammar rules which best suit the data to retrieve the full structure as well as semantic labels of a façade [49, 95]. Simon et al. [90] extend this to multi-view by introducing a 3D grammar for Haussmannian building architecture. This grammar-based inference typically leads to complex and computationally expensive optimization and the required hand-written rules are hard-coded for a specific architectural style (mostly Haussmannian architecture) and it is not trivially extensible to any other building architecture. Alternative methods, e.g., [65] avoid the use of explicit grammar rules in a more bottom-up fashion by using generic architectural

principles which still limit the categories of building structure it can handle, e.g., mostly flat façades for which plausible 3D parameters can easily be suggested.

2.5 Conclusion

In this chapter we have presented the most prominent lines of work which address the Multi-View Stereo problem with a specific interest on structure-awareness and scalability with an emphasis on urban modeling applications. From this discussion, a few conclusions arise.

First, the MRF framework and view-dependent modeling (i.e., representing 3D through depthmaps) have been widely and successfully applied to structure-aware reconstruction, even though the merging strategy between per-view reconstructions is not trivial, as standard procedures such as standard TSDF-based depthmap fusion [18] are agnostic of the scene's structure. However, most of such structure-aware methods are pixel-based [29, 33, 91, 106] and hence, suffer from a lack of robustness to wide-baseline set-ups, strong illumination changes, surface specularity, and hence, they rely on an overwhelming regularization. Additionally, they scale poorly in image number and size.

Superpixel modeling techniques are – on the other hand – very scalable and robust to the inherant challenges in street-level views [10, 68]. They allow to produce a scalable, piecewise-planar approximate geometry but lack some structural features which are key in man-made scenes such as the alignment of objects' boundaries with each other and with dominant VDs. Additionally, such methods mostly rely on bottom-up over-segmentation methods which are structure-agnostic [1, 24] and as a consequence, the planar homogeneity assumption per superpixel often breaks using such approaches.

Procedural methods and grammar-based approaches are very well suited for generating a plausible structured representation of buildings, but such methods either consider only simple, mostly flat, building façades [74], or they make other very specific assumptions on the scenes they address [97], making them unsuitable for most street-level scenarios. Other grammar-based approaches rely on hard-coded grammars [90] and can only address a small fraction of existing building structures.

In the next chapter, we introduce the main contributions of this manuscript by studying how to combine the advantages of MRF structure-aware pixel-based methods such as, e.g, [29, 91] and superpixel modeling, e.g., [11, 69, 100, 111], while considering a

more suitable top-down image partitioning than bottom-up segmentations [1, 24] that are used in such methods to model structured, man-made scenes.

3

Patchwork Stereo - Scalable, Structure-aware 3D Reconstruction in Man-made Environments

Contents

Figure 3.1: Our method takes a few calibrated images and an SfM point cloud to reconstruct a compact, piecewise-planar mesh aligned with the dominant structure of the scene.

In this chapter, we address the problem of Multi-View Stereo (MVS) reconstruction of highly regular man-made scenes from calibrated, wide-baseline views and a sparse Structure-from-Motion (SfM) point cloud. We introduce a novel patch-based formulation via energy minimization which combines top-down segmentation hypotheses using appearance and vanishing line detections, as well as an arrangement of creased planar structures which are extracted automatically through a robust analysis of available SfM points and image features. The method produces a compact piecewise-planar depth map and a mesh which are aligned with the scene's structure. Experiments show that our approach not only reaches similar levels of accuracy w.r.t state-of-the-art pixel-based methods while using much fewer images, but also produces a much more compact, structure-aware mesh in a considerably shorter runtime by several of orders of magnitude.

3.1 Introduction

Over the last decade, structure-from-motion (SfM) and dense multi-view stereo (MVS) reconstruction have benefited from constant progress in feature detection and matching, and camera calibration, leading to mature systems, e.g, Bundler [92, 93], VisualSfM [107, 108], openMVG [70–72], PMVS-2 [32], CMP-MVS [43], including consumer products such as Acute3D ContextCapture and Agisoft PhotoScan.

Current state-of-the-art methods are now able to produce impressive 3D reconstructions for many scene categories with a rich level of detail, assuming there are enough input images and the scene is sufficiently textured.

However in highly-regular environments such as indoor and outdoor man-made scenes, the complexity of the produced geometry (dense point clouds or meshes) is often detrimental to the structure of reconstructed objects. In such scenes the geometry ubiquitously presents: (i) piecewise planarity, (ii) alignment of objects boundaries with image gradients and (iii) with vanishing directions (VDs), and (iv) surface simplicity, which globally induces planar alignments. This structure is even more difficult to retrieve when only few wide-apart views are considered or available, with broad textureless and specular areas which, altogether, form the typical use-case scenario in urban street-level imagery.

Moreover, the usability of traditional MVS approaches is also limited due to their insufficient computational-and-storage scalability as they consider exhaustive or significant multi-view photoconsistency at the pixel level. Typical runtimes can reach several hours to model a single street, resulting in several millions of polygons and contradicting the paradoxical simplicity of the depicted scenes.

Alternative approaches tackle these issues separately. Superpixel modeling techniques first establish an image partitioning using unsupervised methods [10, 11, 68, 69] to address the problems of robustness and scalability, but fail at respecting structure. Structure-aware reconstruction methods [29, 91] on the other hand propagate sparse 2D dominant edge detections and 3D information under heavy regularization and expensive pixelwise computations. A number of restrictive assumptions have been used to simplify the problem, such as a Manhattan-world assumption (MWA) [29, 97], semantic information [52], building footprints [97], hard-coded grammar rules [97] or the additional availability of dense point clouds from laser scans [59, 89].

In this chapter, we address the multi-view reconstruction of structured depth maps from a few images (typically 2-5 wide-baseline images with one reference view) and

a sparse SfM point cloud (typically obtained together with image calibration) using a scalable, region-based formulation. In contrast to existing region-based stereo methods, ours does not rely on a bottom-up image partitioning. Rather, we combine vanishing directions, image contours and sparse 3D data to generate top-down segmentation hypotheses, on which we define a Markov Random Field (MRF) topology. The final, structured depth map is retrieved by minimizing a global energy which groups neighboring image patches by enforcing plausible structure-aware connectivities, resulting in a "patchwork" solution.

We demonstrate pixelwise accuracy results on par with state-of-the-art dense MVS pipelines [43] while utilizing much fewer reprojection images and gaining several orders of magnitude in runtime and memory consumption. These improvements are achieved thanks to both our patch-based representation and our robust hypothesis extraction from already-available SfM data. The resulting mesh is compact, and aligned with scenes' structure and image gradients by design, which is achieved with no need of later 3D geometry simplification [80], nor additional complex mesh refinement [103], or tedious primitive fitting steps [53].

Our main contributions are as follows:

- We propose a novel region-based stereo formulation which incorporates structure priors in a principled MRF energy minimization framework where the global energy is amenable to graph-cut inference [13].
- We define a robust joint 2D-3D method for extracting structurally-relevant 2D line and 3D plane hypotheses from principal VDs, image contours and already-available sparse SfM data. It generates top-down superpixels whose boundaries are aligned with VDs.
- We present an end-to-end pipeline which treats high-resolution images (16MP) within a few seconds or minutes per building with Matlab code, paving the way for large-scale, compact, structure-aware urban modeling.

3.2 Related Work

Pixel-level MVS. A number of top-performing general MVS algorithms assume a Delaunay tetrahedralization of an initial 3D point cloud, whose cells are labeled with a discrete occupancy state according to visibility and photometric constraints; the reconstructed surface lies at the interface between empty and non-empty cells [43, 103].

Despite mesh refinement, the resulting surface remains a jagged approximation of a locally-smooth geometry, which may then require expensive post-processing to achieve a compact representation, e.g., by fitting 3D geometric primitives [45, 53]. The situation is even worse with voxel-based approaches [37, 81]. Pixel-based stereo techniques, which build disparity maps, have seen a tremendous increase in performance since early approaches [46] and their later extensions using second order smoothness priors [76, 106], color models [9] or semantic classification [52]. This category of approaches has been well established for narrow-baseline stereo problems as reported in the Middlebury challenge [83], but it scales poorly in image number and image size; besides, it is sensitive to wider baselines.

Superpixel modeling. Patch-based stereo approaches, e.g., [10, 68, 69], infer piecewise-planar depth maps for superpixels whose surface is assumed uniform. These superpixels are obtained with unsupervised bottom-up methods, that tend to randomly oversegment highly-textured regions [24] or to produce hexagonal shapes in large homogeneous areas [1]. These methods, in comparison to pixel-based and volumetric approaches, are more scalable and are less sensitive to appearance, viewpoint changes and textureless areas. They are however completely agnostic of the structure of the scene beyond the simple alignment of objects boundaries with image gradients, which translates into many blatant visual artifacts. Bodis-Szomoru et al. [10] build a multi-image graph over superpixels and reconstruct a approximate model which is very well suited for large-scale modeling. However, patch-to-patch stereo matching adds up to the lack of structured boundaries and alignments. It also assumes there are enough SfM points, even in visually homogeneous patches, which often does not hold.

Structure priors. Another line of work models weak structure priors [29, 91] by enforcing piecewise-planarity transitions to lie at both strong image gradients and along edges aligned with vanishing directions. However, these are pixelwise approaches and suffer from robustness and scalability issues which restricts their usage to scenes of low complexity and low image resolution ($\leq$ 3MP). In contrast, our patch-based formulation allows to handle 16MP images with a much lower runtime by several orders of magnitude, without assuming Manhattan scenes [29].

Top-down superpixels. Fouhey et al. [27] use a scene representation relying on multiple top-down partitions of an image. They intersect sets of 2D rays cast from pairs of vanishing points, defining projective rectilinear superpixels/patches whose boundaries reflects their 3D orientation. The authors use this intermediate

representation to estimate the orientation membership of each pixel in a monocular indoor Manhattan-world scene, as well as inter-patch spatial relationships. In contrast, our approach makes use simultaneously (vs. sequentially) of image edge detections, vanishing directions and 3D cues from sparse SfM data to help extract more subtle lines in a robust line-sweep stage.

Mesh alignment. Yet another line of work constructs a mesh in the image domain, and then reconstructs vertices in 3D. Saxena et al. [82] use supervised learning to correlate image region appearance with depth information and are able to retrieve a plausible 3D mesh from a single calibrated image for scenes that present a low variation of aspect and structure. Bodis-Szomoru et al. [11] address the problem of 3D reconstruction from a single image with sparse SfM data by triangulating superpixels [24] in the image domain, and then fitting triangles onto SfM points by penalizing surface curvature. The depth information of triangles with no sparse 3D information is linearly interpolated. This simplifying assumption is made at the expense of geometric accuracy. The rendered reconstructions can be visually satisfactory at a coarse level for nearly flat objects and buildings (e.g., Haussmannian architecture), but cannot model more complex yet ubiquitous elements such as protruding balconies and loggia recesses, especially for patches with low point density. In contrast, our method benefits from sparse SfM cues (where available) and multi-view photoconsistency; it propagates structurally plausible surface associations by favoring planar continuity and crease junctions.

3.3 Overview

Inputs/Outputs. Our method takes a collection of unordered calibrated images (one serving as reference, $\mathcal{I}$, the others for reprojection) and a sparse SfM point cloud $\mathcal{S}$ (given together with calibration information). It produces a structured depth map and a corresponding structured mesh for each reference image. Our notion of structure refers to the following properties w.r.t. the expected output geometry: (i) piecewise-planarity, (ii)+(iii) alignment of object boundaries with strong image gradients and main vanishing directions, (iv) non-local planar and boundary alignments.

Top-down segmentation and 3D plane hypotheses. Our method first computes the dominant VDs visible in $\mathcal{I}$ via a greedy procedure. Top-down superpixels are then generated by creating in $\mathcal{I}$ an arrangement of dominant vanishing lines (VLs). Intuitively, VLs play a key role in capturing the layout of a regular scene as they are plausible in-

dicators of geometric transitions. In order to extract plane candidates consistent with patch boundaries, i.e., to favor crease planar transitions in 3D, VLs and dominant planes must be mutually consistent and aligned. To this end, we extract the 3D hypotheses in a robust vanishing-line-sweeping stage which simultaneously takes into account image features along VLs and sparse 3D data (cf. Section 3.4).

MRF-Energy minimization. Our energy combines all patches in 3D by enforcing structurally-sound associations in accordance with multi-view patch-wise photoconsistency and SfM cues. It is minimized efficiently (cf. Section 3.5).

Compact, structured mesh generation. Once the final depth map is recovered, we generate a polygonal mesh for each planar region. This is carried out in the image domain with a 2D Constrained Delaunay Triangulation (CDT) which is then reprojected to 3D (cf. Section 3.6).

3.4 2D Segmentation and 3D Plane Hypotheses

In this section, we describe in detail the different elements of our pre-processing.

3.4.1 Estimating Vanishing Directions

As a first step, we extract dominant VDs visible in reference view $\mathcal{I}$. Contrary to [91], we do not merge or cluster them from different images as it would introduce inaccuracies due to calibration imprecision. It could also introduce directions which are irrelevant in the image of interest. We proceed as follows, without MWA, as opposed to [29, 69]:

First, we detect line segments, using LSD [102], and keep the segments with the best scores (lowest $-\log(\text{NFA})$). In our experiments, by keeping the top 2500 segments of sufficient length (40 pixels), we get enough cues for detecting vanishing points (VPs) with negligible outliers.

Second, we estimate VDs. We use the VP detector of Lezama et al. [57], which handles both Manhattan and non-Manhattan cases. As most non-Manhattan architectures may also include 3 Manhattan directions, we first use the Manhattan prior and seek 3 initial Manhattan VDs. We then greedily detect new VDs without the Manhattan prior, putting aside associated lines at each iteration and discarding VDs too close from previous ones (≤ 5 deg), until no more VD is detected. This strategy allows to better retrieve VDs that have subtle sets of supporting evidence. It may yield more than 3 VDs, which may or may not be orthogonal.

3.4.2 Dominant Planes

We extract plane hypotheses in two stages. First, dominant planes are detected from both the VPs and the point cloud $\mathcal{S}$. Next, more subtle planes associated to creases and fine structural details are detected (e.g., window frames).

Concretely, we first discretize the set of plane orientations by considering VP pairs $\vec{v}_i, \vec{v}_j$ and the associated plane normal $\vec{n}_{ij}$, given the intrinsic calibration matrix K [38]:

$$\vec{n}_{ij} = \frac{K^\top \vec{v}_i \times \vec{v}_j}{||K^\top \vec{v}_i \times \vec{v}_j||} \tag{3.1}$$

Then, for each $\vec{n}_{ij}$, we look for associated plane offsets (signed distance to the camera) that correspond to dominant planes. For this, each point $s \in \mathcal{S}$ votes in a 1D weighted histogram (specific to $\vec{n}_{ij}$) in the bin associated to its offset. The weight is $|\vec{n}_{ij}.\vec{n}_s|$ where $\vec{n}_s$ is the normal of a plane estimated by PCA analysis from points in a local neighborhood $N(s)$. To limit quantization issues in presence of sparse regions in $\mathcal{S}$, we define $N(s)$ as the ball whose radius is half the distance to the k-th nearest neighbor of s [77]. (In our experiments, $k = 50$.) The size of a bin is defined as:

$$g = \min_{ij}(median_{s \in \mathcal{S}}(m_{ij}(s))) \tag{3.2}$$

where $m_{ij}(s)$ is the median of the offsets of points in $N(s)$ along the normal $\vec{n}_{ij}$. In our experience, g provides a stable granularity scale throughout all the considered datasets; all dominant planes are retrieved as the maxima of the histogram, unless data is missing, e.g., due to the lack of texture.

3.4.3 Dominant Vanishing Lines

We extract dominant VLs in $\mathcal{I}$ as lines with strong and consistent edge information, in the following way.

We first reduce texture sensitivity by applying a bilateral filter (with a range parameter $\sigma_r = 130$, and a spatial parameter $\sigma_d = 3$ in all experiments). We then filter the image using a Canny-Deriche edge detector [20] with double hysteresis thresholding (with fixed thresholds 0.05 and 0.15 throughout experiments), resulting in a binary image Γ. To retrieve more subtle contours, we actually extract edges at multiple image scales (0.5, 0.75, 1 in our implementation) and merge in Γ the resulting edge maps with a logical-or.

Then, for each VP, we sweep a VL l on the binary edge map Γ through every pixel x along l within the image frame. The fixed angular deviation between two successive VLs is the smallest angle among the 4 angles corresponding to 1 pixel of deviation at

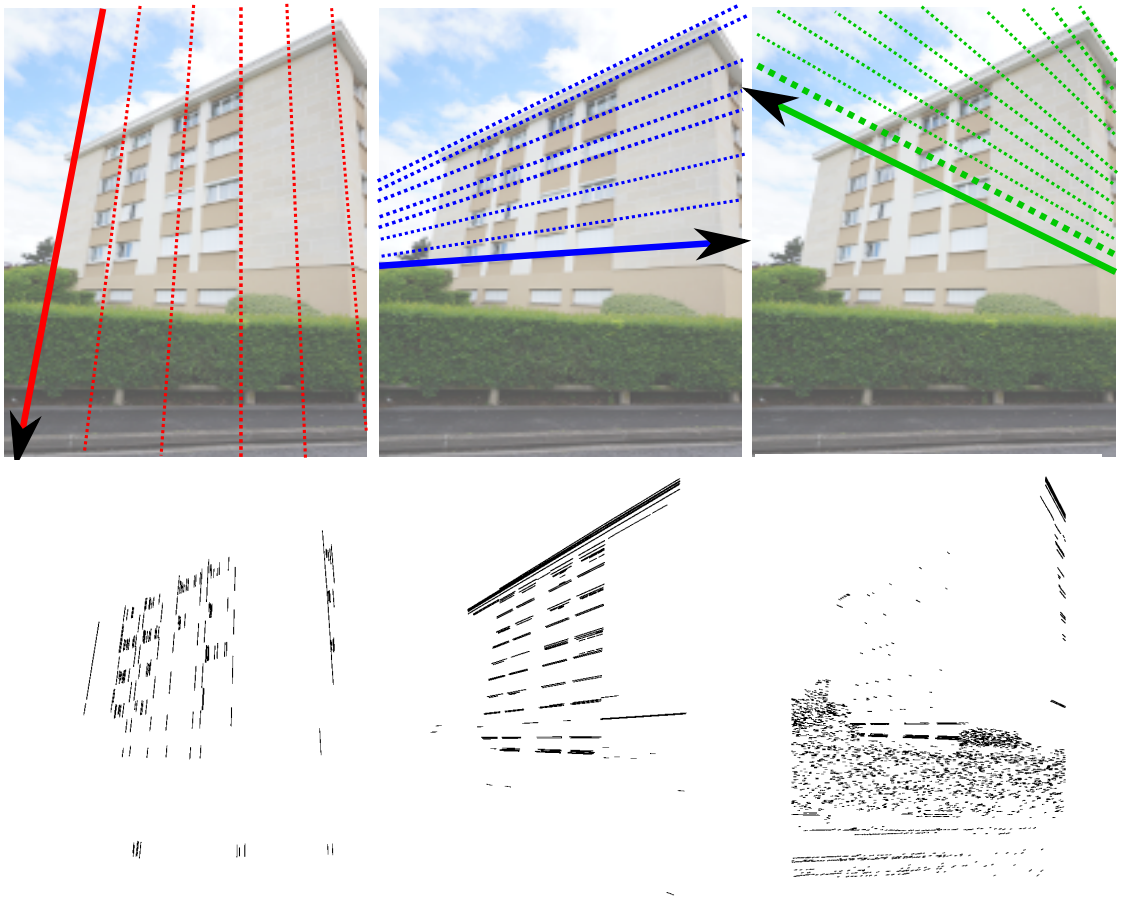

Figure 3.2: VLs swept from each VP (top row). Pixels sup-porting dominant VLs (bottom row), based on gradient features.

the 4 image corners. This ensures an adaptive and high density sweeping throughout the image. For each swept VL l, we consider the rasterized chain of binary pixels $\Gamma(x, l)$ it contains. For robustness, we initially apply a 1D Gaussian (with $\sigma = 1$), re-binarizing the line (with threshold 0.8). For consistency, we only consider as meaningful in Γ, continuous chains of pixels that are long enough (of length of at least 40 in our experiments). Resulting segments are illustrated on Fig. 3.2. Finally, dominant VLs are defined as the local maxima of the following score when l varies along the swept lines:

$$domVL(l) = \frac{1}{|l|} \sum_{x \in l} \Gamma(x, l) \tag{3.3}$$

3.4.4 Secondary Lines and Planes

Leveraging on dominant planes and VL information, we extract more subtle lines and planes. We consider the following three additional cues, based on creaseness.

For each dominant plane Π_{ij} with normal $\vec{n}_{ij}$, for each VP $\vec{v}_k$ other than $\vec{v}_i$, $\vec{v}_j$, and for each VL l swept from $\vec{v}_i$ (then symmetrically from $\vec{v}_j$), we consider a hypothetical plane Π_{ikl} defined by the normal $\vec{n}_{ik}$ and the offset s.t. Π_{ikl} and Π_{ij} intersect in 3D on a line L which reprojects as l. To assess this hypothesis, we measure the following cues:

- $ridge_{ijk}(l)$ is the number of points in $\mathcal{S}$ that lie in the slice of space at distance at most g of Π_{ikl}. It is illustrated as the stripe between the green lines in Fig. 3.3. As we only want to assess the crease hypothesis at l, each point in the slice contributes to the global score (denoted $crease_{ijk}(l)$ below) according to its 1D distance d to L, with weight $\exp(-d(s,L)/(40g))$. Formally:

$$ridje_{ijk}(l) = \sum_{s\in\mathcal{S},\ d(s,\ \Pi_{ikl})\leq g} \exp(-d(s,L)/(40g)) \tag{3.4}$$

- $volum_{ij}(l)$ is the number of "volumic" points in $\mathcal{S}$ that lie in a cylinder at distance at most g of L. It is illustrated as the disk inside the red circle in Fig. 3.3. "Volumic" points are considered not to lie on a line or plane, which would not correspond to a crease. The dimensionality of a point $s \in \mathcal{S}$ is given by PCA analysis of neighborhood $N(s)$. It is "volumic" if the 3 largest eigenvalues e_1, e_2, e_3 ($e_1 \geq e_2 \geq e_3$) are comparable: $0.35\, e_1 \leq e_2, e_3$.

- $junct_{ijk}(l)$ is the number of points lying in a rectangular cuboid centered on L with length $8g$ along $\vec{v}_j$ and width $2g$ along $\vec{v}_k$. It is illustrated as the area inside the purple rectangle in Fig. 3.3. It tells whether dominant plane Π_{ij} could have a junction with Π_{ikl} at L.

Last, if $junct_{ijk}(l) \geq 2$, we consider the following score:

$$crease_{ijk}(l) = domVL(l)\, ridge_{ijk}(l)\, volum_{ij}(l) \tag{3.5}$$

The local maxima of $crease_{ijk}(l)$ indicate secondary planes Π_{ijk} and vanishing lines l.

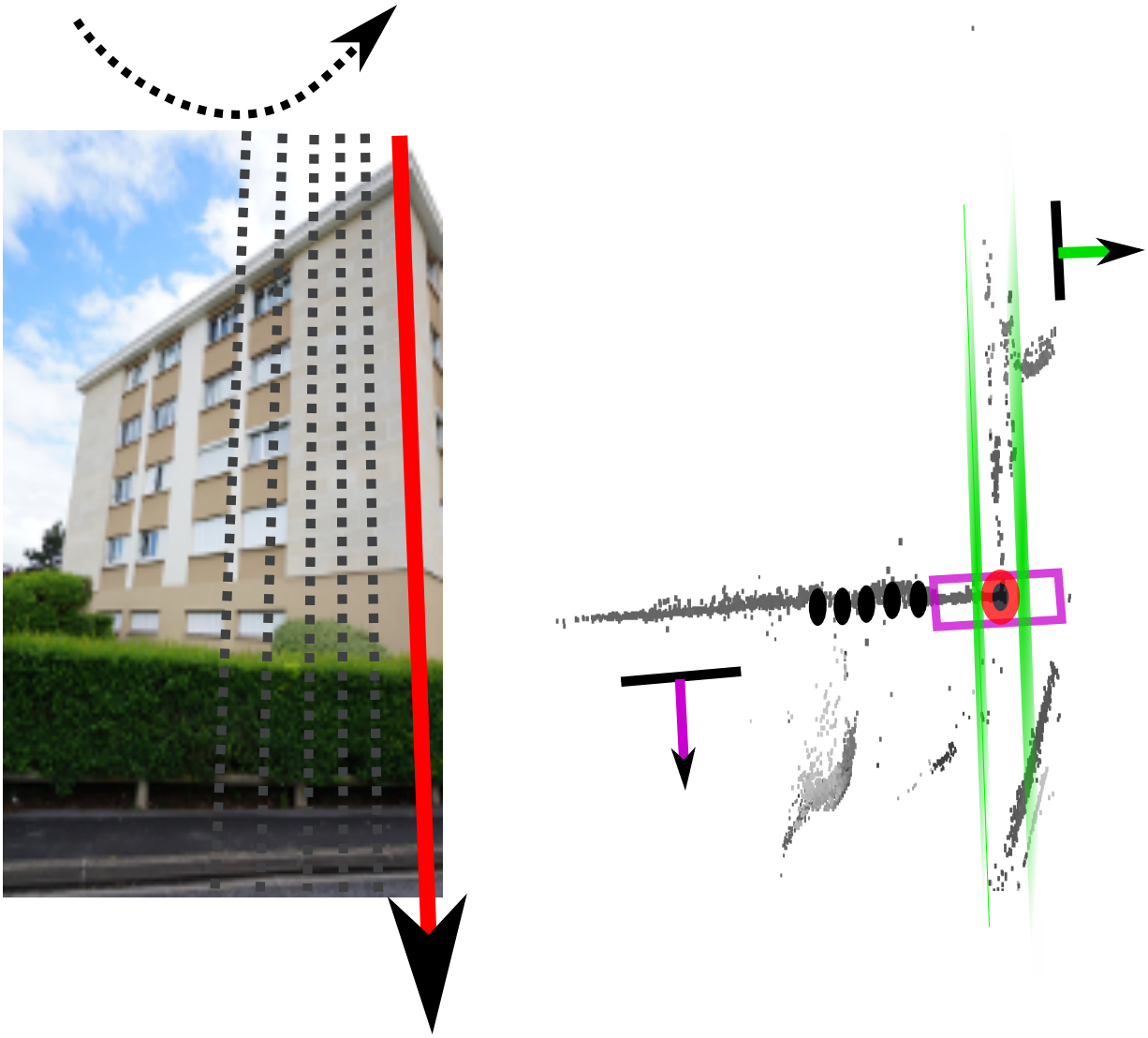

Figure 3.3: 2D-3D VL sweeping to extract secondary lines and planes (left). Top view of SfM point cloud (right) with regions to measure ridge cues (green), volumic points (red) and plane junctions (purple). Please see text for details.

3.4.5 Segmentation into Patches

The "patchwork", i.e., the final top-down segmentation into patches $p \in \mathcal{P}$, is the 2D arrangement made from dominant and secondary VLs, from which we discard peripheral patches. We only keep patches in the intersection of regions inside the two extreme VLs extracted for each VP. The fact is that peripheral patches often consist of sky, vegetation, ground or clutter pixels, which are not planar. Besides, as not all vanishing orientations

are represented at the periphery (in terms of patch boundaries), it could disfavor certain planes during inference, which could propagate by local regularization, altering proper plane assignment.

This simple region clipping automatically restrains the focus of the reconstruction on the main objects of interest (e.g., Fig. 3.5, top right). It generally defines a convex hull (unless a VP lies in the image). When a piecewise-planar structure with a convex silouhette is observed, this strategy yields a meaningful segmentation, not requiring manual masking [69] nor semantic or planarity classifiers [33]. When it forms a concave region, our assumption still restricts possible detrimental behaviors to the patches that constitute the concave fraction. Our method is however little sensitive to noise and outliers.

3.5 Patch-based Stereo Revisited

We define a pairwise MRF over the graph $\mathcal{G} = (\mathcal{P}, \mathcal{N})$ where $\mathcal{P}$ is the set of patches in Sect. 3.4.5 and $\mathcal{N}$ is the neighborhood system of pairs of patches sharing a boundary. Let $\mathcal{L} = \{(\vec{n}_1, d_1), \ldots, (\vec{n}_N, d_N)\}$ be the label space of random variables $(y_p)_{p\in\mathcal{P}} = \mathbf{y}$; $(\vec{n}_p, d_p)$ represents a plane, uniquely characterized by its normal $\vec{n}_p$ and signed offset d_p to the main camera center, i.e., camera of the reference image $\mathcal{I}$.

Our goal is to infer for all patches $p \in \mathcal{P}$ the plane assignment y_p with the lowest energy. The energy $E(\mathbf{y})$ encourages planar continuity and crease junctions, over structure disruptions and implausible planar compositions (regularization). It also favors photo-consistency between views at patch level and adherence to the sparse SfM points (data terms). It is defined as follows:

$$E(\mathbf{y}) = \underbrace{\sum_{p\in\mathcal{P}} w_p \, (\Phi_p^{\text{Photo}}(y_p) + \Phi_p^{\text{3D}}(y_p))}_{\textbf{Data terms}} + \lambda \underbrace{\sum_{(p,q)\in\mathcal{N}} w_{pq} \, \Psi_{pq}^{\text{Connectivity}}(y_p, y_q)}_{\textbf{Regularization term}} \tag{3.6}$$

where λ balances the contribution of the unary and pairwise potentials, and w_p, w_{pq} are adaptive normalizing weights respectively proportional to the patch area and the length of the common linear boundary between neighboring patches; both expressed in pixels. This allows to adaptively scale the relative contribution between unary and pairwise terms in the global energy, reducing the sensitivity of the parameter λ. The adaptive weights are defined as follows:

$$w_p = \text{area}_{\mathcal{I}}(p). \exp\left(-\frac{\sigma(\mathcal{S}_p)}{0.1}\right) \tag{3.7}$$

where $\text{area}_{\mathcal{I}}(p)$ is the area of patch p, and $\sigma(\mathcal{S}_p)$ is the *surface variation* of the 3D points reprojecting in p, as defined in [77]. This value ranges between 0 (totally planar) and 1/3 (isotropically distributed points) and plays a role of indicating whether the point distribution within a patch p is likely to be planar or not.

$$w_{pq} = |p \sqcap q|.\max\left(0.01, \frac{1}{|p \sqcap q|}\sum_{x \in p \sqcap q} \mu(x)\right) \tag{3.8}$$

where $|p \sqcap q|$ is the length of the common edge boundary between p and q, and $\mu(x)$ is the edge magnitude at pixel x (i.e., the pixel intensity, between 0 and 1). We cap the pairwise regularization in the definition of w_{pq} by allowing a minimum weight of 0.01 for robustness. This is to avoid a complete disconnection of neighboring nodes in the graph topology along strong edge boundaries. The different potential functions are detailed below.

3.5.1 Data Terms

Multi-View photoconsistency. $\Phi_p^{\text{Photo}}(y_p)$ penalizes appearance dissimilarities between a patch p and its reprojection $\pi_v(p)$ in other views $v \in \mathcal{V}$, assuming plane-induced homographies [38]. For regions not reprojecting entirely within v, the penalty is a constant. This function is subdivided into an intra-patch photoconsistency and a boundary edge consistency operating on patch boundary pixels $\mathcal{B}_p$ and their reprojection $\pi_v(\mathcal{B}_p)$.

$$\Phi_p^{\text{Photo}}(y_p) = \tfrac{1}{|\mathcal{V}|}\sum_{v \in \mathcal{V}}\{\alpha\Delta(p,\ \pi_v(p)) + \beta\mathcal{A}(\mathcal{B}_p, \pi_v(\mathcal{B}_p))\} \tag{3.9}$$

where α, β are model parameters, and $\mathcal{A}(.,.)$ measures the proportion of boundary pixels agreeing on the presence of image gradient across views. $\Delta(.,.)$ is a dissimilarity function between two image regions related by homography. We consider the zero-mean normalized cross-correlation *zncc* with exponential normalization for robustness:

$$\Delta(p, \pi_v(p)) = 1 - \exp\{\frac{-\delta^2}{0.8}\} \tag{3.10}$$

where

$$\delta = 1 - \max\{0,\ zncc(p,\ \pi_v(p))\} \tag{3.11}$$

3D point consistency. We use the sparse 3D cues to encourage surfaces to fit onto SfM points that reproject within p:

$$\Phi_p^{\text{3D}}(y_p) = 1 - \exp\{\frac{-\phi^2}{0.3}\} \tag{3.12}$$

where

$$\phi = \frac{\gamma}{\tau.|\mathcal{S}_p|}\sum_{s \in \mathcal{S}_p}\min(\tau,\ \frac{\mathcal{D}(s, y_p)}{g}) \tag{3.13}$$

where γ is a model parameter, $\mathcal{S}_p$ is the subset of SfM points reprojecting within p, τ is a distance threshold (measured in g units), and $\mathcal{D}(s, y_p)$ is the point-to-plane 3D distance.

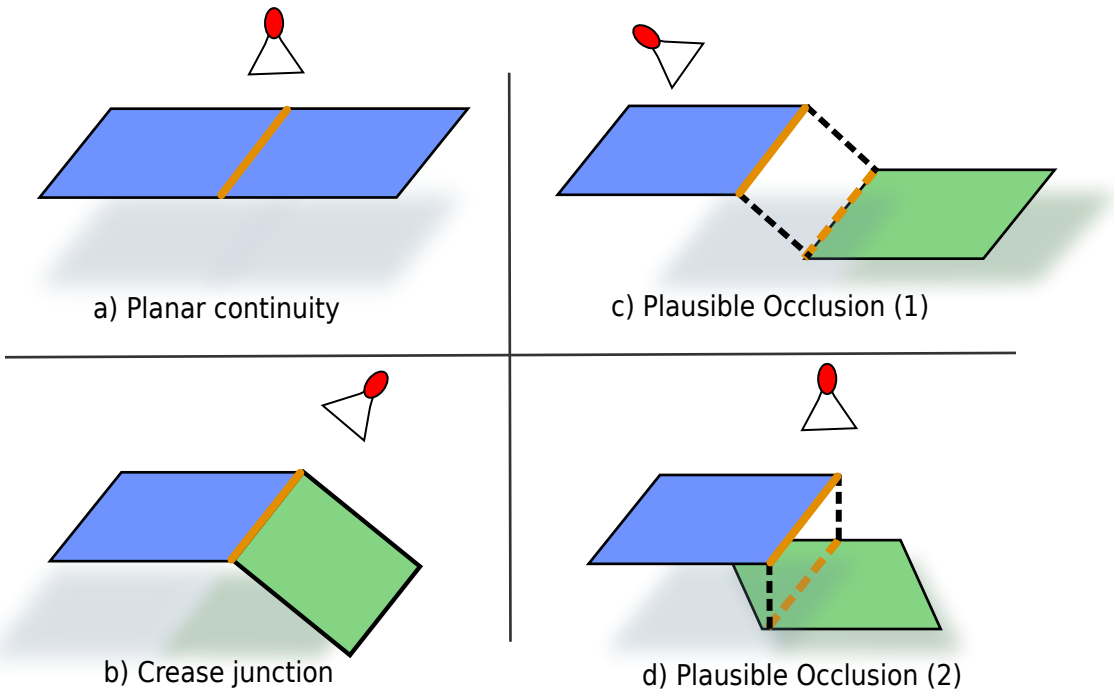

Figure 3.4: The four pairwise associations modeled by our regularization term. Surface hypotheses are represented with boundaries aligned with vanishing directions defining their 3D orientation. **Best viewed in color.**

3.5.2 Regularization

Representing 3D orientations by using vanishing points (Eq. (3.1)) suggests that two planar surfaces oriented resp. towards $\vec{n}_{ij}$ and $\vec{n}_{ij'}$ are likely to intersect in the image plane at a crease edge $\vec{e}_{pq}$ (in orange in Figure 3.4) aligned with the common vanishing direction $\vec{v}_i$. Our pairwise regularization prior $\Psi_{pq}^{\text{Connectivity}}(y_p, y_q)$ relies on this assump-

tion by reasoning on the connectivity of neighboring patches and imposing a preference over the possible configurations:

$$\Psi_{pq}^{\text{Connectivity}}(y_p, y_q) = \begin{cases} 0 & : \text{if } (y_p, y_q) \in \mathcal{T}_{continuity} \text{ else} \\ \lambda_1 & : \text{if } (y_p, y_q) \in \mathcal{T}_{crease} \text{ else} \\ \lambda_2 & : \text{if } (y_p, y_q) \in \mathcal{T}_{occlusion_1} \text{ else} \\ \lambda_3 & : \text{if } (y_p, y_q) \in \mathcal{T}_{occlusion_2} \text{ else} \\ \lambda_4 & : \text{otherwise} \end{cases} \tag{3.14}$$

where $0 \leq \lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \lambda_4$ are the respective costs for neighboring patches, and $(y_p, y_q) \in \mathcal{T}_{continuity}$ lie on the same plane, i.e., $y_p = y_q$ (case (a) in Fig. 3.4), $(y_p, y_q) \in \mathcal{T}_{crease}$ form a crease junction (case (b) in Fig 3.4), $(y_p, y_q) \in \mathcal{T}_{occlusion_1}$ lie at a depth discontinuity where $\vec{e}_{pq}$ is consistent with the orientations of both p and q (case (c) in Fig 3.4), and $(y_p, y_q) \in \mathcal{T}_{occlusion_2}$ are such that $\vec{e}_{pq}$ is consistent only with the occluding (fronting) patch (case (d) in Fig 3.4). All other configurations are given a prohibitive penalty λ_4.

3.5.3 Inference and Theoretical Details

Depending on how the penalties $\lambda_{1..4}$ are set in Eq. 3.14, the smoothness function can either be a metric, or a semi-metric. The metric case allows a more efficient inference as it guarantees the solution to be at a known factor from the global optimum, but it is more restrictive in its expressive power [48]. In all our experiments, we adopt the semi-metric case by setting the parameters to respectively $\{\alpha, \beta, \gamma, \lambda, \lambda_1, \lambda_2, \lambda_3, \lambda_4\} = \{1, 0.5, 0.4, 30, 0, 0.6, 3.8, 50\}$. The final energy can hence be optimized using, e.g., swap-based graph-cut moves [48]. In practice, we found the alpha-expansion [47] inference to give better results even in the semi-metric case although there is no theoretical guarantee to be close the optimum, and adopt it throughout our experiments.

We now provide additional details on the connectivity term in Eq. 3.14. Formally, the connectivity term can be defined as follows:

$$\Psi_{pq}^{\text{Connectivity}}(y_p, y_q) = \begin{cases} 0 & : \text{if } y_p = y_q \text{ else} \\ \lambda_1 & : \text{if } \theta_{pq}(y_p, y_q) \ \wedge \ \chi_{pq}^{3D}(y_p, y_q) \ \wedge \ \overrightarrow{n_p} \neq \overrightarrow{n_q} \text{ else} \\ \lambda_2 & : \text{if } \overline{\theta_{pq}(y_p, y_q)} \ \wedge \ \chi_{pq}^{3D}(y_p, y_q) \text{ else} \\ \lambda_3 & : \text{if } \overline{\theta_{pq}(y_p, y_q)} \ \wedge \ \overline{\chi_{pq}^{3D}(y_p, y_q)} \ \wedge \ \phi_{pq}(y_p, y_q) \text{ else} \\ \lambda_4 & : \text{otherwise} \end{cases} \tag{3.15}$$

where θ_{pq} is a 3D tightness predicate which is true when the patches touch in 3D along the whole common linear boundary, up to an $\epsilon = 10^{-5}$.
χ_{pq}^{3D} means $\overrightarrow{n_p}$ and $\overrightarrow{n_q}$ share a common vanishing point (i.e., relate to some $\overrightarrow{n_{ij}}$ and $\overrightarrow{n_{ij'}}$, hence such oriented surfaces could form a junction pointing towards $\vec{v}_i$ which they have in common). ϕ_{pq} indicates whether the orientation of the common boundary, $\overrightarrow{e_{pq}}$, belongs to the hypothesis of the fronting reconstructed 3D patch which is a case of plausible occlusion.
The $\overline{\text{top bar}}$ notation designates predicate negation.

$$\theta_{pq}(y_p, y_q) = [\![\ max\{ \ \rho(\overrightarrow{e_{pq}^1}, y_p, y_q), \ \rho(\overrightarrow{e_{pq}^2}, y_p, y_q) \ \} \leq \varepsilon \ ]\!] \tag{3.16}$$

where $\overrightarrow{e_{pq}^1}$ and $\overrightarrow{e_{pq}^2}$ refer to the two vertices at both ends of the common edge boundary between neighbor patches*. ρ is the relative 3D reconstruction error of a pixel x w.r.t planar hypotheses y_p and y_q:

$$\rho(x, y_p, y_q) = \frac{||\mathbf{X}(x, \ y_p) - \ \mathbf{X}(x, \ y_q)||}{2 \max\{ \ ||\mathbf{X}(x, \ y_p)||, \ ||\mathbf{X}(x, \ y_q)|| \ \}} \tag{3.17}$$

where $\mathbf{X}(x, \ y_p)$ (resp. $\mathbf{X}(x, \ y_q)$) is the reconstructed 3D point lying at the intersection of the infinite ray going from the camera center of $\mathcal{I}$, through pixel x and the 3D plane defined by the plane label y_p (resp. y_q).

$$\begin{aligned} \phi_{pq}(y_p, y_q) = [\![\ & (\overrightarrow{e_{pq}} \in \{i, j\} \ \wedge \ ||X_{p_1}|| < ||X_{q_1}|| \ \wedge \ ||X_{p_2}|| < ||X_{q_2}||) \\ & \vee (\overrightarrow{e_{pq}} \in \{i', j'\} \ \wedge \ ||X_{p_1}|| > ||X_{q_1}|| \ \wedge \ ||X_{p_2}|| > ||X_{q_2}||) \ ]\!] \end{aligned} \tag{3.18}$$

where X_{p_1} (resp. X_{p_2}, X_{q_1}, X_{q_2}) is a shortcut notation to designate the 3D reconstruction of pixel $\overrightarrow{e_{pq}^1}$ (resp. $\overrightarrow{e_{pq}^2}$) via y_p (resp. y_q)), and where i, j, (resp. i', j') are line directions corresponding to vanishing points $\vec{v}_i$, $\vec{v}_j$ (resp. $\vec{v}_{i'}$, $\vec{v}_{j'}$).

*[\![ ]\!] stand for the *Iverson* bracket.

3.6 Structure-aware Mesh Generation

After inferring a plane for each patch, our structured planemap representation contains a number of polygons per reconstructed plane. For each plane, we merge all associated polygons, producing larger but fewer polygons, possibly with holes. By construction, patches are either adjacent one to another or disjoint, which simplifies merging. By construction also, the polygon boundaries are aligned with VDs and image gradients. A 2D triangle mesh for these merged polygons can be then produced using a constrained Delaunay Triangulation, and then lifted to 3D.

3.7 Evaluation

We evaluate our approach on 4 challenging datasets of individual buildings presenting textureless areas and repetitive patterns, for which we use only a few wide-baseline images. Statistics for each dataset are given in Table 3.1. All experiments use the same parameters: $\alpha = 1$, $\beta = 0.4$, $\gamma = 0.5$, $\lambda = 30$, $\lambda_1 = 0$, $\lambda_2 = 0.6$, $\lambda_3 = 3.8$, $\lambda_4 = 50$.

Quantitative results. We quantify pixelwise accuracy of our reconstructions w.r.t. a reference mesh built with CMP-MVS [43] and two point clouds built using PMVS-2 with and without Poisson surface reconstruction [32]. For these baselines, we use all of the available images of each scene.

Fig. 3.5, 3.6, 3.7 and 3.8 show, for each dataset, the reference image of each dataset, the corresponding top-down segmentation, a qualitative view of the output 3D model and the corresponding quantitative results per scene. For each method, in the right column, we vary the tolerated error as a fraction of the scene's depth range and accumulate the proportion of correctly reconstructed pixels (up to the given tolerance) w.r.t. the reference mesh; the higher the curve, the better the performance. We compare our results against the reference mesh only on manually annotated regions of interest per view, i.e., on a mask that specifies the building pixels in the image. The figures show the following: (i) The sparse PMVS-2 method has poor overall accuracy due to the lack of reconstructed points in wide textureless areas. (ii) Its dense counterpart (PMVS-2+Poisson) performs better than our method (PWS) and its ablated versions for AugusteC and Hameau, which is explained by the significant amount of additional images. (iii) For the GMU dataset, PWS has a higher curve, which is due to the lack of images for CMP-MVS and PMVS-2+Poisson (only 5). (iv) In Bry2, the performance of

				PMVS-2 [32]	CMP-MVS[43]	Patchwork Stereo (PWS = ours)			
Scenes	#Img	#SfM / $\mathcal{I}$	Resol.	#Points	#Triangles	#Reproj.views	#VDs	#Normals	#Planes	#Patches
Bry2	31	2081	16MP	470185	587220	2	3	3	45	2003
AugusteC	11	2980	16MP	498270	653228	3	3	3	78	2980
GMU [69]	5	578	2MP	25750	23057	3	3	3	28	1416
Hameau	36	21824	16MP	766744	1143405	3	3	3	89	5351

Table 3.1: Datasets and comparative reconstruction characteristics. #Img: total number of available images in the corresponding dataset, #SfM/$\mathcal{I}$: number of SfM points reprojecting on $\mathcal{I}$, Resol: image resolution, #Points: number of reconstructed points by PMVS-2 [32], #Triangles: number of produced triangles by CMP-MVS[43], #Reproj.views: number of views considered by our method for reprojection and photometric comparison of patches seen in the reference image (cf. photo-consistency term in section 3.5.1), #VDs, #Normals, #Planes, #Patches: resp. number of VDs, 3D normals, planes, and image patches retrieved by our approach, PWS.

PWS is on par with the baseline.

> Although using only a small subset of wide-baseline views, our method (PWS) achieves comparable accuracy results while providing a much more compact geometry which respects the structural regularity of the scene in a fraction of the runtime (as discussed below).

Ablative study. Fig. 3.5, 3.6, 3.7 and 3.8 also show results with ablated variants of our data terms, to assess their importance. (When canceling a term, we make sure the relative weights of the data and regularization terms stay the same.)

Keeping only the SfM term $\Phi_p^{3D}(y_p)$ sometimes leads to severe errors. This robustness issue corresponds to a few anomalous planes due to point could sparsity. Apart from SfM, PWS is comparable to its ablated models, sometimes slightly better in terms of pixelwise accuracy. However, it is difficult to see quantitatively the difference because of the relative lack of accuracy of the CMP-MVS reference. Still, a qualitative analysis, as illustrated in Fig. 3.9, shows that the full PWS model presents a much more regular, structured appearance and is visually more pleasing. This also shows the limits of using CMP-MVS [43] as a reference to quantitatively assess the quality of the reconstructions.

Indicative runtime. Our CPU implementation is a mixture of pure vectorized Matlab / Mex / C++. The two main computational bottlenecks of our method are the multi-view photoconsistency, which is computed for all patches through all planar hypotheses, and the pairwise costs. Both of these tasks are written in vectorized pure Matlab, and the photoconsistency could benefit from significant speed-ups.

Photoconsistency runs in roughly 1s per 16MP image per plane candidate on a modest laptop with an Intel Core2Duo 2.40Ghz, 4GB RAM. Other running times are negligible.

Comparison to related work. [11] provides quantitative results on scenes for which our VD-based segmentation does not make sense, e.g. arches and columns of Herz-Jesu. Only scenes of streets M, P, Z of Mirbel (low-resolution, <1MP images) are relevant to us, but are unknown subsets of the ETHZ RueMonge 2014 dataset. The reference (high-resolution) mesh is unavailable anyway. Still, we ran our method, with only 2 reprojection views, on a RueMonge facade looking like Fig. 1,3 and 6 in [11]. Our reconstruction (cf. Fig. 3.11) is better aligned with the structure: window and balcony edges are straighter and sharper. Besides, we have much less triangles per image (<680 vs

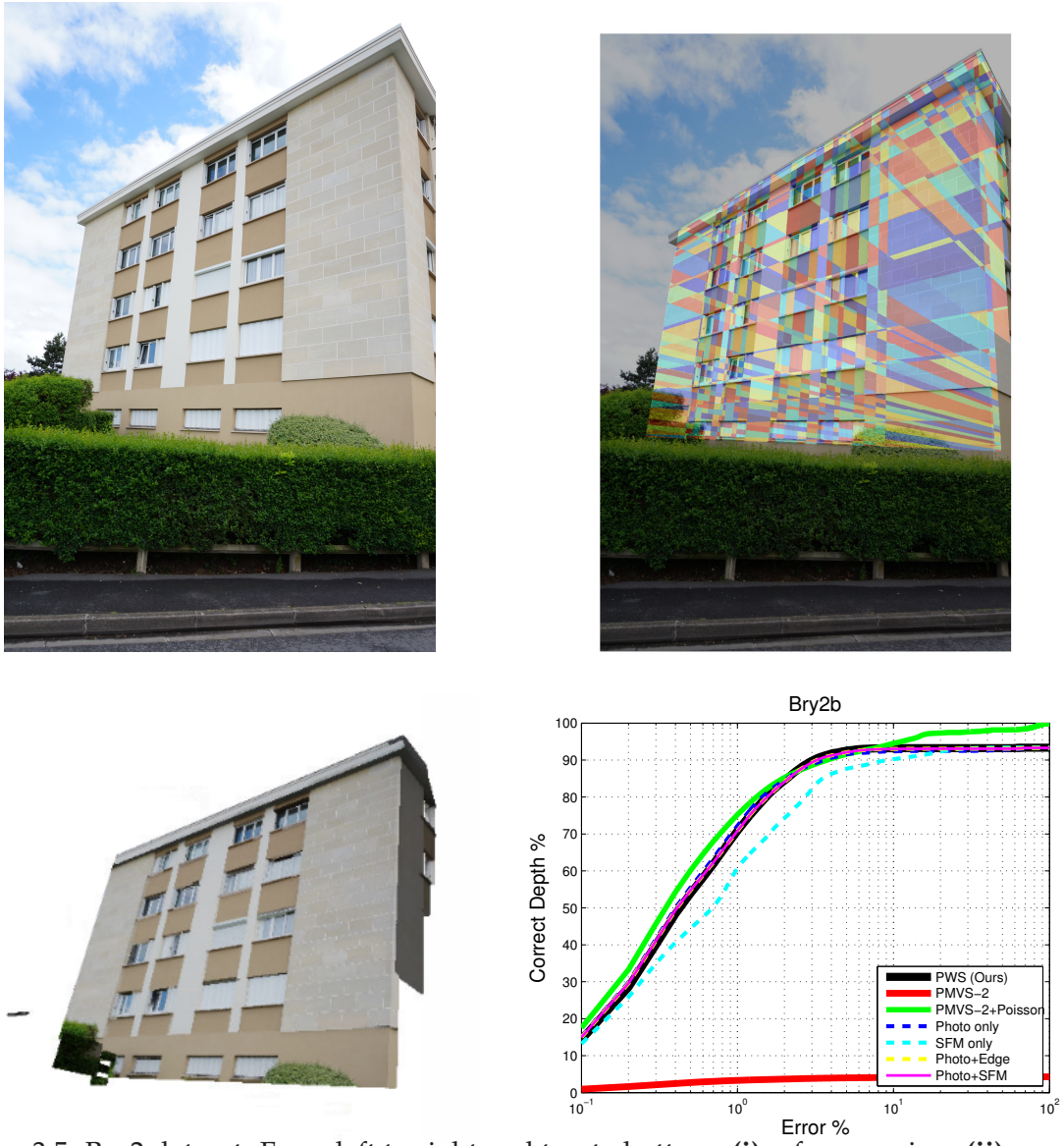

Figure 3.5: Bry2 dataset. From left to right and top to bottom: **(i)** reference view, **(ii)** our segmentation, **(iii)** our 3D reconstruction, **(iv)** semi-log-scale accuracy w.r.t CMP-MVS reference mesh. We plot the proportion of pixels whose depth is correct up to a given error tolerance, expressed as a fraction of the scene's thickness (labeled Error %). We compare with PMVS-2 [32], with and without poisson surface completion, and different ablations of our data terms. PWS: our complete model (PWS), then using different data terms in the energy, SfM only: using only the SfM 3D point consistency from Eq. 3.12, Photo only: using the photo-consistency part from Eq. 3.9), Photo+Edge: using photo and edge consistency (Eq. 3.9), and Photo+SfM: Eq. 3.9+Eq. 3.12. **Best viewed in color.**

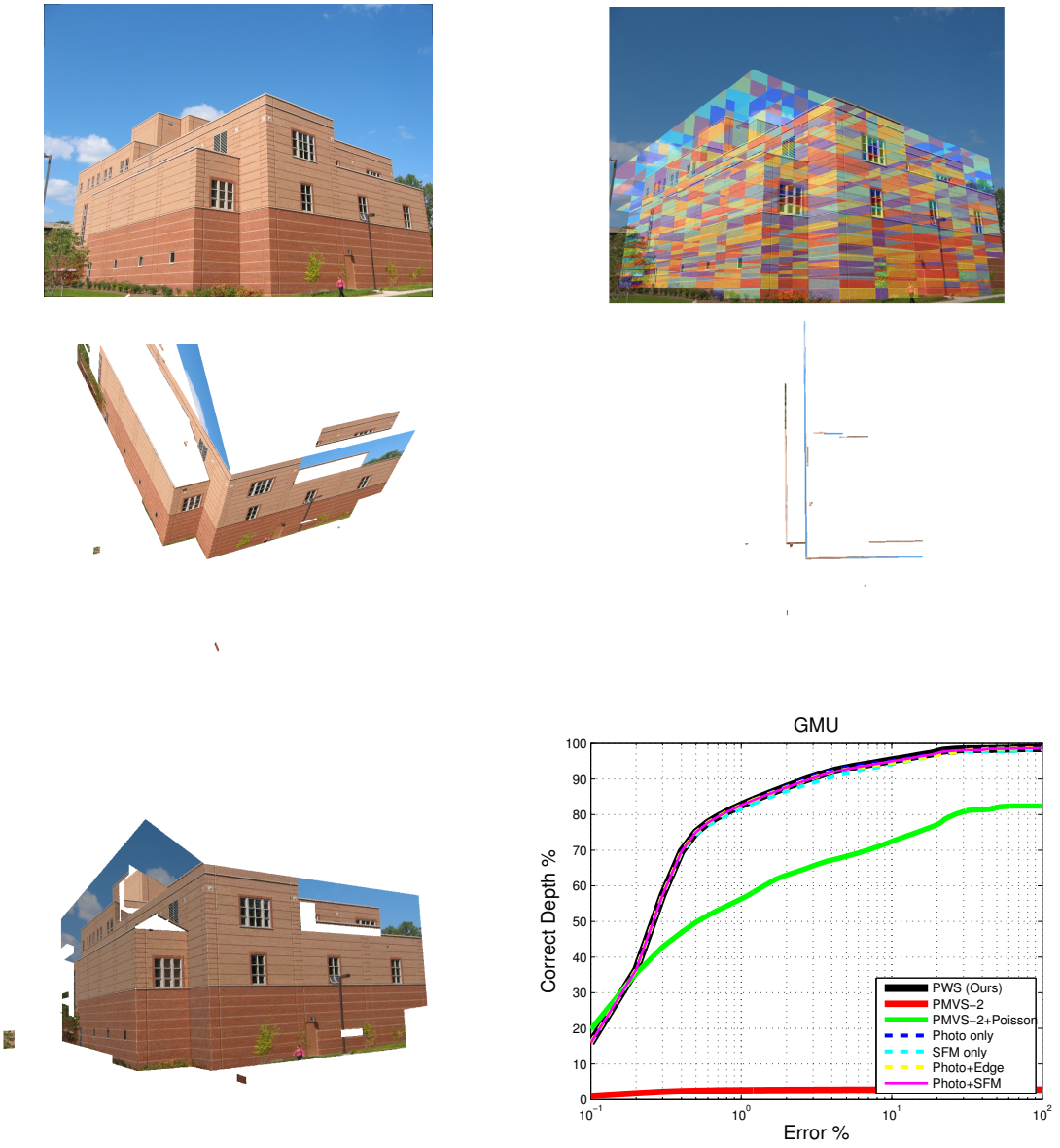

Figure 3.6: GMU [68] dataset. From left to right and top to bottom: **(i)** reference view, **(ii)** our segmentation, **(iii) & (v)** views of our 3D reconstruction, **(iv)**, top view showing the compactness of the model and its alignments to VDs, **(vi)** semi-log-scale accuracy w.r.t CMP-MVS reference mesh. We plot the proportion of pixels whose depth is correct up to a given error tolerance, expressed as a fraction of the scene's thickness (labeled Error %). We compare with PMVS-2 [32], with and without poisson surface completion, and different ablations of our data terms. PWS: our complete model (PWS), then using different data terms in the energy, SfM only: using only the SfM 3D point consistency from Eq. 3.12, Photo only: using the photo-consistency part from Eq. 3.9), Photo+Edge: using photo and edge consistency (Eq. 3.9), and Photo+SfM: Eq. 3.9+Eq. 3.12. **Best viewed in color.**

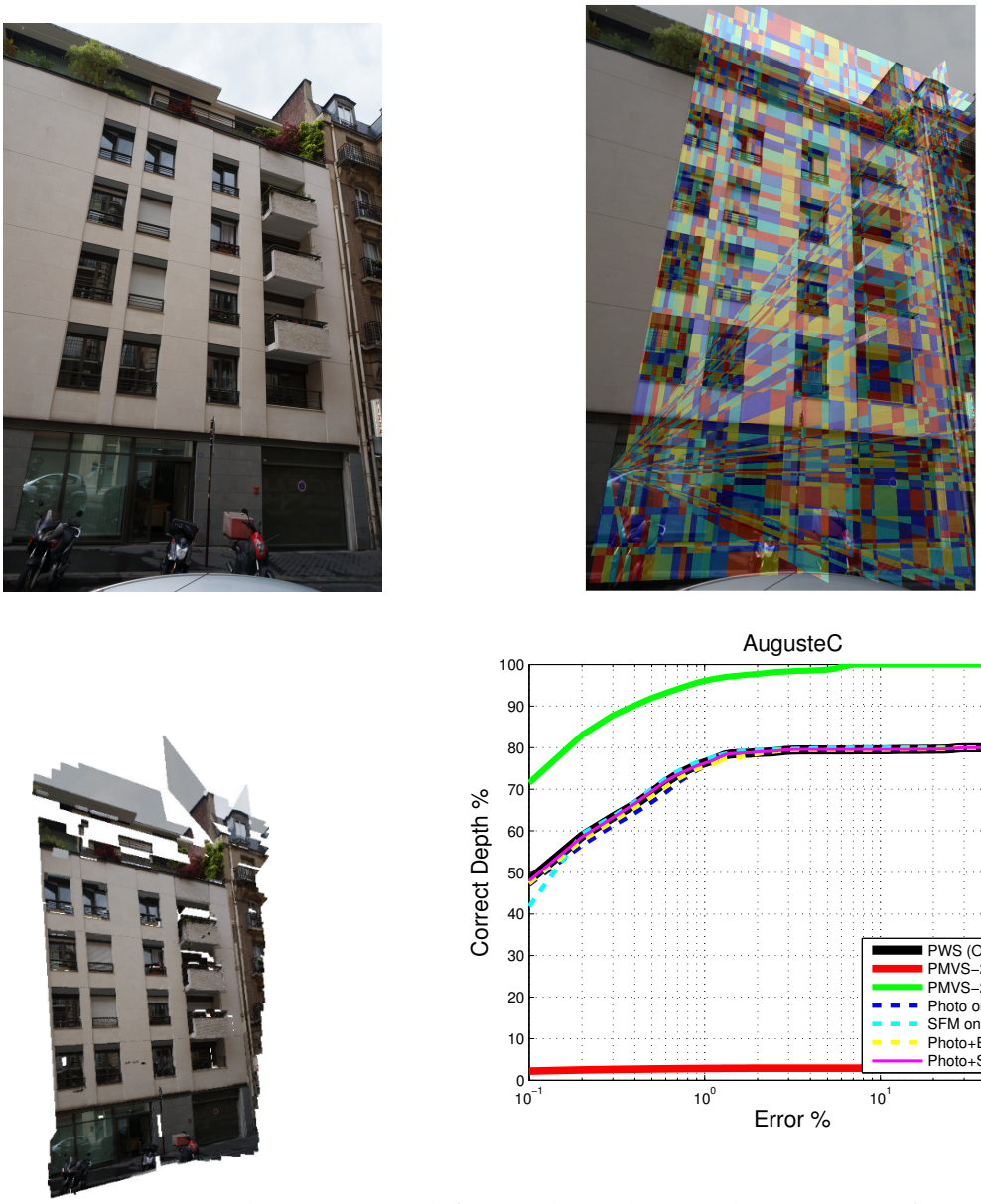

Figure 3.7: AugusteC dataset. From left to right and top to bottom: **(i)** reference view, **(ii)** our segmentation, **(iii)** our 3D reconstruction, **(iv)** semi-log-scale accuracy w.r.t CMP-MVS reference mesh. We plot the proportion of pixels whose depth is correct up to a given error tolerance, expressed as a fraction of the scene's thickness (labeled Error %). We compare with PMVS-2 [32], with and without poisson surface completion, and different ablations of our data terms. PWS: our complete model (PWS), then using different data terms in the energy, SfM only: using only the SfM 3D point consistency from Eq. 3.12, Photo only: using the photo-consistency part from Eq. 3.9), Photo+Edge: using photo and edge consistency (Eq. 3.9), and Photo+SfM: Eq. 3.9+Eq. 3.12. **Best viewed in color.**

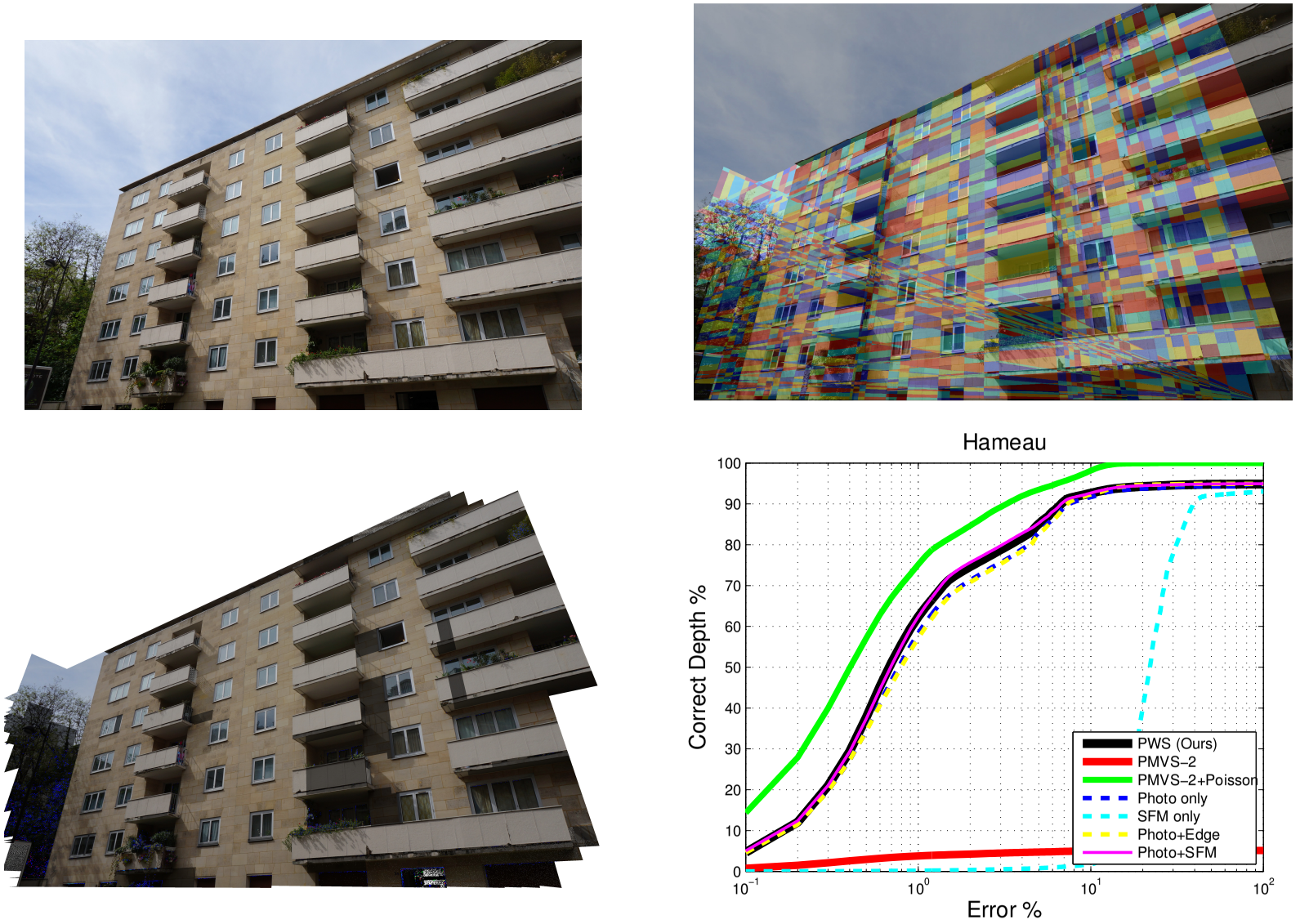

Figure 3.8: Hameau dataset. From left to right and top to bottom: **(i)** reference view, **(ii)** our segmentation, **(iii)** our 3D reconstruction, **(iv)** semi-log-scale accuracy w.r.t CMP-MVS reference mesh. We plot the proportion of pixels whose depth is correct up to a given error tolerance, expressed as a fraction of the scene's thickness (labeled Error %). We compare with PMVS-2 [32], with and without poisson surface completion, and different ablations of our data terms. PWS: our complete model (PWS), then using different data terms in the energy, SfM only: using only the SfM 3D point consistency from Eq. 3.12, Photo only: using the photo-consistency part from Eq. 3.9), Photo+Edge: using photo and edge consistency (Eq. 3.9), and Photo+SfM: Eq. 3.9+Eq. 3.12. **Best viewed in color.**

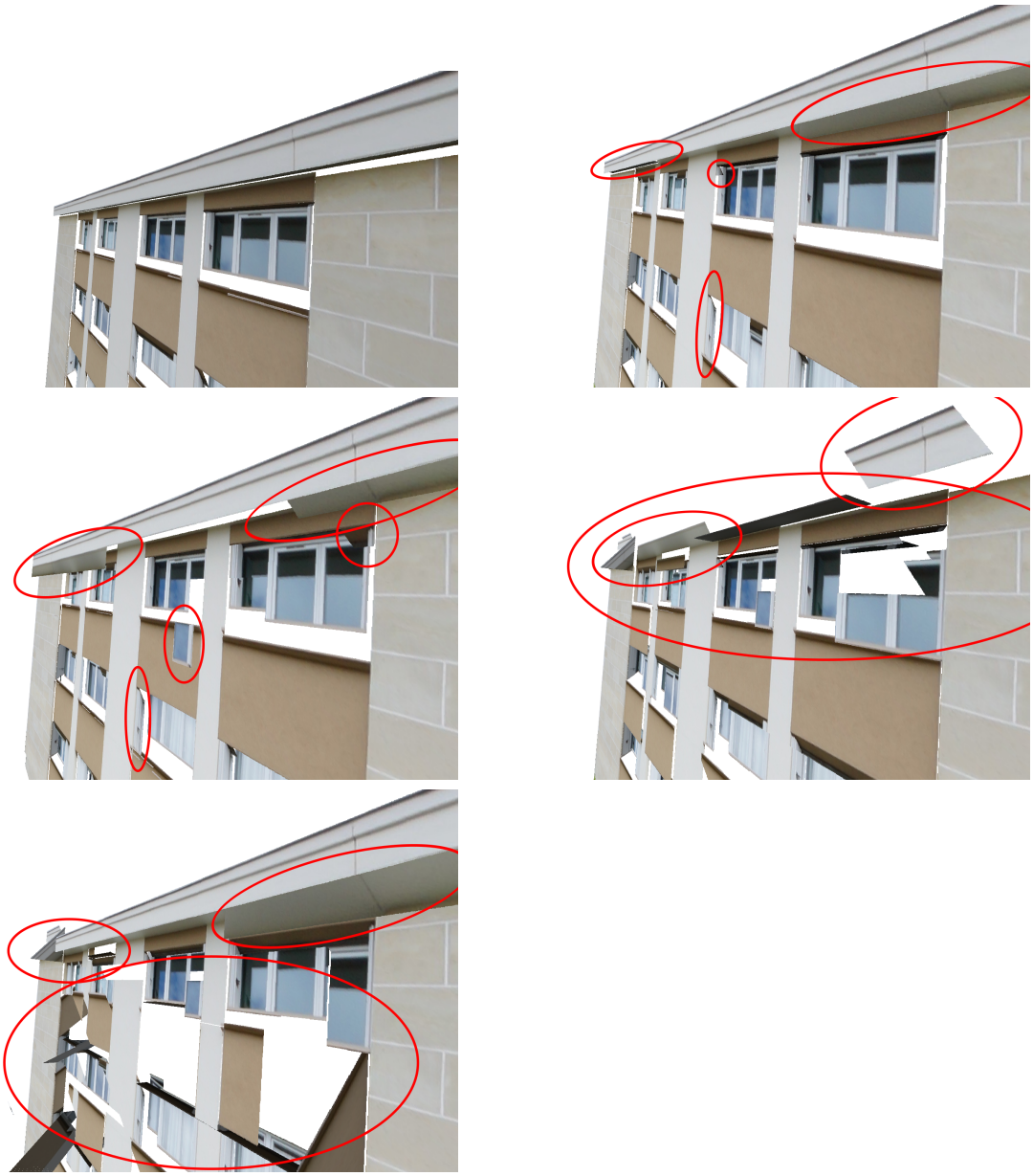

Figure 3.9: Qualitative comparison of different ablations of our data terms. From left to right and top to bottom: **(i)** our full model, **(ii)** Photo+SfM, **(iii)** Photo only, **(iv)** Photo+Edge, **(v)** SfM only. Even though the global pixelwise accuracy may be comparable between different truncated versions of our model (cf. Fig. 3.5, 3.6, 3.7 and 3.8), removing data terms translates into noticeable artifacts which degrade the 3D structure through erroneous depth or even surface orientations. **Best viewed in color.**

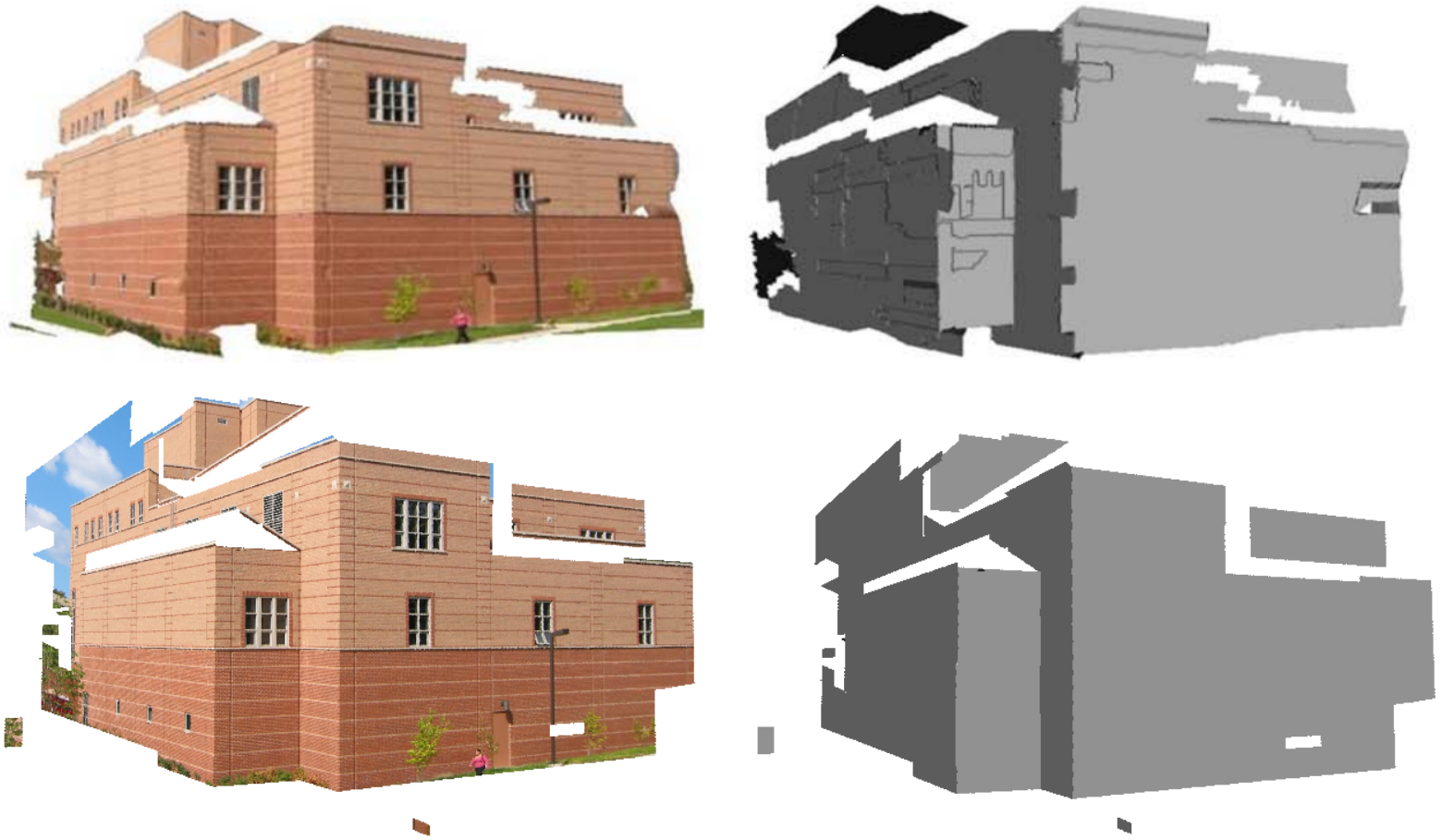

Figure 3.10: Side-by-side comparison with prior work Superpixel Stereo (SPS) [69]. In the first row, SPS [69] presents an uneven planar geometry along flat surfaces and patches tend to straddle between different plane orientation at crease transitions, as they are agnostic of VDs (sky pixels were manually masked out by the authors in [69]). In the second row, our reconstruction presents sharp edges and perfect crease transitions, seamless plane continuity and the alignment of surface boundaries with VDs and image contours (to facilitate the visual comparison, we manually mask out sky pixels from our reconstruction somilarly to [69]). **Best viewed in color.**

15k). [29, 69, 91] do not provide any quantitative evaluation of accuracy; in any case, they do not address both structure and scalability, as we do. Comparing with [69], our result is much better, as illustrated on Fig. 3.10. Our junctions form perfect creases. Our misreconstructed patches correspond either to the sky or to regions occluded in other views. All our patches are perfectly aligned with VDs in contrast to patches in [69] which form arbitrary shapes and do not touch in 3D.

As for speed, [11] processes on average 1 view of 1MP per 2s and a facade in Rue-Monge is seen by about 10 views, yielding a rate of about 20s/MP/facade. With Matlab, we process 1 plane hypothesis for a 16MP image in about 1s; assuming 80-plane scenes with 3 reprojection views per facade, our rate is 80*3*1/16 = 15s/MP/facade, comparable to [11]. Likewise, [29] takes more than 300s/MP/facade and [91] takes 60s/MP/image for scenes with 11-61 images. [69] does not provide complete time information.

3.8 Conclusion

In this chapter, we have presented a novel approach for automatic multi-view reconstruction of structured depth maps from only a few, wide-baseline high-resolution photographs. Our method produces compact meshes which are aligned with the dominant structural traits of the scene (vanishing directions and edges). We have shown how top-down segmentation hypotheses and sparse 3D data can capture most of non-local planar alignments which are typical of man-made scenes. Working at the patch-level allows significant improvements in robustness and scalability without any loss of information w.r.t working on individual pixels. Regarding pixel-wise accuracy, we are on par with dense reconstruction methods, although we use up to 9 times less images. This paves the way for large-scale structure-aware urban modeling with plausible, visually pleasing digital rendering.

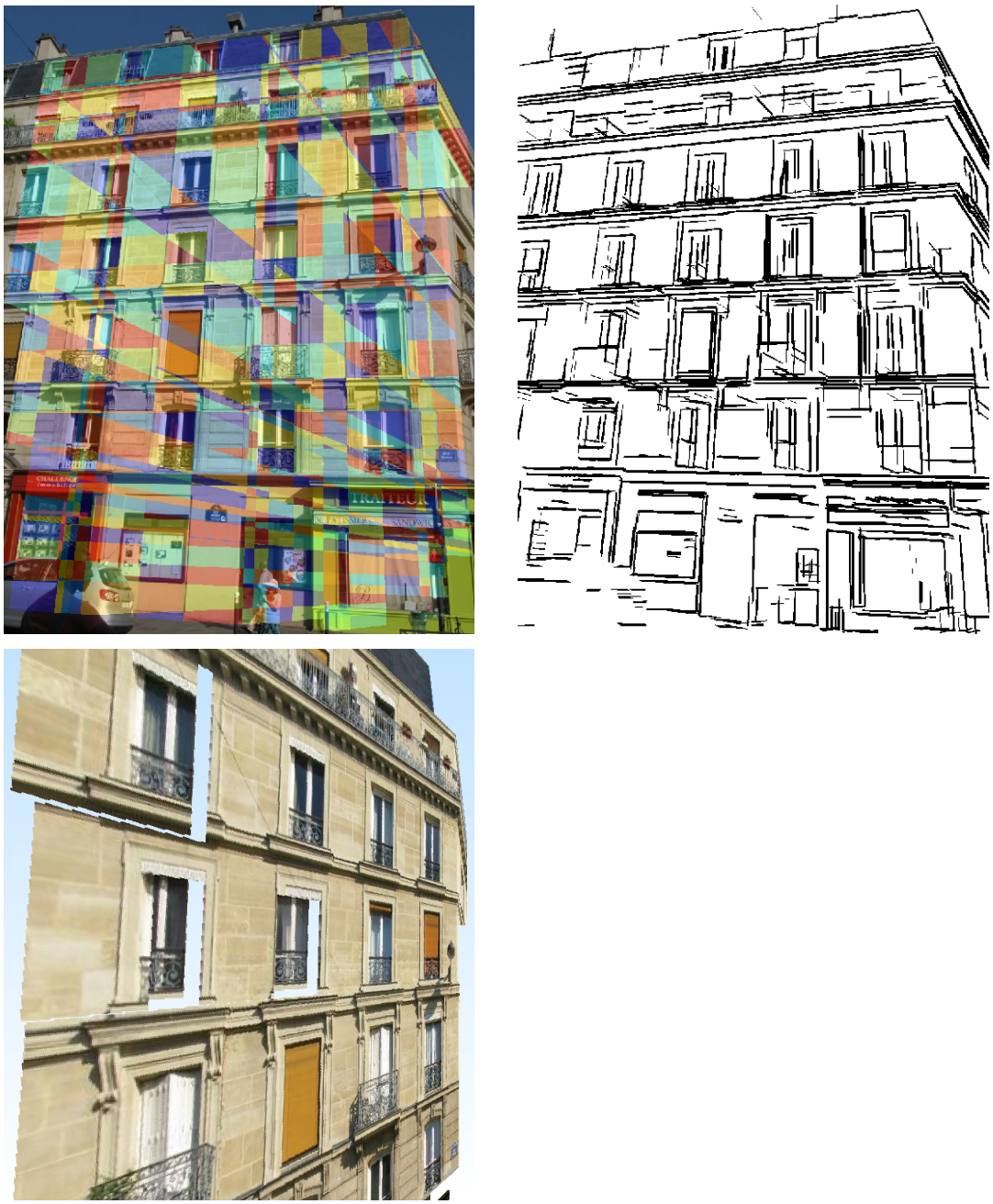

Figure 3.11: Top row: Segments and corresponding edge-map of Haussmanian façade, considering 3 VDs. Bottom row: reconstructed planes with our method.

Figure 3.12: Additional qualitative results. Comparison between the baseline PMVS-2 [32] in the first row and our method in the second row (coloured) and in third row (uncoloured) on the Bry2 dataset.

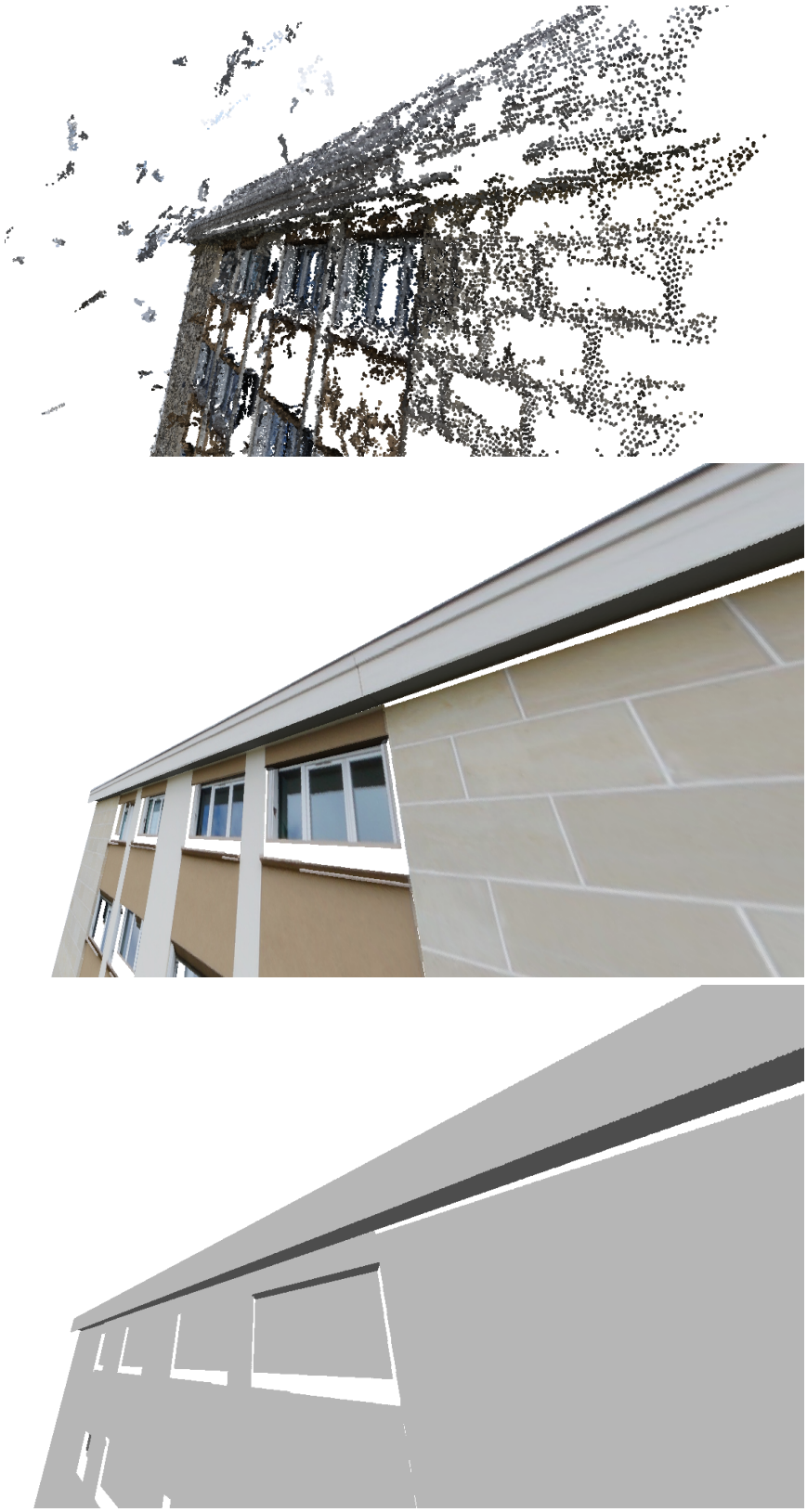

Figure 3.13: Additional qualitative results. Comparison between the baseline PMVS-2 [32] in the first row and our method in the second row (coloured) and in third row (uncoloured) on the Bry2 dataset.

Figure 3.14: Additional qualitative results. Comparison between the baseline PMVS-2 [32] in the first row and our method in the second row (coloured) and in third row (uncoloured) on the Bry2 dataset.

4

Conclusion

Contents

4.1 Summary of the Thesis and Contributions

In this thesis, we have studied the challenging problem of Urban Modeling, i.e., Image-based Modeling applied to street-level imagery, assuming the camera poses and a corresponding sparse 3D point cloud to be available. We have focused our study on two aspects in Multi-View Stereo (MVS) reconstruction, scalability and structure priors. By "structure", we intend the following principles regarding the produced 3D geometry: (i) piecewise-planarity, (ii) alignment of the boundaries of the 3D elements with their corresponding 2D image gradients and (iii) with principal vanishing directions (VDs), (iv) co-planarity of elements, and (v) global geometric simplicity.

The main contributions of this thesis can be summarized as follows.

- **Robust Extraction of 3D Planar Hypotheses from a Sparse and Noisy Point Cloud.**
  3D planes are key in the piecewise-planar representation of man-made scenes. While 3D planar hypotheses can be detected from a dense point cloud using standard robust techniques [25], extracting them from a sparse and noisy point cloud, typically acquired through SfM is a much more challenging task. To address this task, in section 3.4.2, we jointly consider information from the image domain, i.e., dominant contours as well as dominant Vanishing Directions (VDs) which are strong structural cues in man-made scenes. The resulting approach is fast and robust, scalable, and combines information from the mutually informative 2D and 3D domains without any additional restrictive assumptions or inputs.

- **Joint 2D/3D Reasoning for Top-down Image Partitioning.**
  In sections 3.4.3 and 3.4.4, we also propose a method that produces a top-down image partitioning by exploiting VDs and strong gradient information in the image domain, using a robust, adaptive line sweeping algorithm.

  The method jointly reasons in 3D on the basis of an SfM point cloud as well as in the 2D image domain and produces an image segmentation into polygonal patches and a set of planar hypotheses which are consistent with the segmentation. Dominant Vanishing Lines (VLs) and strong planar supports in the point cloud are jointly leveraged to generate mutually consistent 3D planar crease hypotheses and 2D VLs which could not have been detected using a single support (2D or 3D) alone.

The final combination of the retained VLs in a 2D line arrangement constitutes a structurally principled partitioning for top-down superpixel-based stereo, and the supporting 3D planar hypotheses for this segmentation presents a significant number of planar crease candidates which are compatible with the extracted VLs in the final image over-segmentation.

- **The Patchwork Stereo Framework.**
  We introduce a novel energy formulation which leverages the top-down over-segmentation we have proposed as well as the robust extraction of planar hypotheses, and reconstructs a piecewise-planar, compact depth map and a mesh which are aligned with the scene's dominant structure using only a handful of wide-baseline views.

  The method poses the problem as an efficient and robust revisit of patch-based stereo reconstruction, e.g., [67–69], by using top-down image partition priors in contrast to bottom-up, structure-agnostic superpixels, e.g., [24]. We show through qualitative and quantitative experiments that our approach not only reaches comparable levels of pixel-wise accuracy with respect to state-of-the-art pixel-based methods, but also produces much more compact, structure-aware depthmaps and meshes in a considerably shorter run-time by several of orders of magnitude and by using up to 9 times fewer images.

- **Publication.**
  A part of the work which is presented in chapter 3 has been published and presented at an international conference in Computer Vision [12].

4.2 Shortcomings and Limitations

Experimentation and Datasets.

A first shortcoming of the work we present in this manuscript comes from the lack of variety and quantity in the datasets we consider in our experiments. In particular, we did not evaluate our approach on man-made scenes comprising more than 3 VDs even though our method can handle more complex scenes in this regard. Finding such scenes in outdoor environments is challenging, but this could be addressed by considering indoor scenes depicting a composition of Manhattan frames, as this set-up is much more commonly found indoors.

Despite our direct comparisons to state-of-the-art baselines such as Superpixel Stereo [67–69] in terms of piecewise-planar and patch-based MVS and to competitive pixel-based approaches [32], there are other very related works, particularly in handling structure. Those approaches could not be considered as baselines in our experiments due to the lack of publicly available implementations of Manhattan-World Stereo [29], the work of Sinha et al. [91], and both approaches of Bódis-Szomorú, Riemenschneider and Van Gool [10, 11]. To the best of our knowledge, each of these papers lack significant details in their descriptions to allow us to reproduce them and support a fair experimental comparison with our results.

Even though many top-performing patch-based and structure-aware MVS methods do not publish quantitative evaluation on pixel-based accuracy performances [10, 29, 69, 91], we have done so in section 3.7. To this end, our strategy was the following. We have used considerably more images per dataset than we have considered to run our method (up to 9 times less), and built a reference mesh using a state-of-the-art mesh reconstruction pipeline [43]. We have also used a manually-edited binary mask to delimit the regions of interest (which mainly depict buildings) in measuring the quantitative performance scores for each of our baselines. However, the reference mesh is not an ideal ground truth, as the produced geometry can still be overly complex, non-planar and relatively noisy despite the important number of considered views. There is room for improvement in using a better suited ground truth or datasets. Such datasets should be compatible with our use-case scenarios and allow, for example, to retrieve dominant VDs.

Applicability and Robustness of our Methods.

Our Patchwork Stereo framework relies on a top-down segmentation of reference images, which relies on a vanishing point detector, e.g., [57]. While such state-of-the-art approaches can consistently provide robust vanishing point detections, this still limits the use case scenarios of our framework to scenes that present strong visual cues (linear features) to support the detection of vanishing points.

Additionally, our optimization combines 2D and 3D cues to produce the final reconstruction results in a principled energy-driven inference. However, as our approach implements a pipeline, some steps are done by making early decisions which – in theory – can not be recovered in case of failure, even though we have not experienced many such cases in practice. This is the case for our segmentation which is fixed once and for all, as well as for the extraction of planar hypotheses at the time of inference, once all

information is put together. Segmentation, plane detection and MVS could be solved jointly, by iterating between these steps or in a unified global energy.

Scene Completeness vs. Structure.

The 3D that we generate is view-dependent and the meshes we produce are computed in 2D and lifted to 3D in order to allow a fast and scalable computation. As a result, our final geometry is prone to holes in the scene parts which are not depicted in reference images. A straightforward approach to handle this problem in the context of using unstructured depthmaps would be to compute the depthmaps with overlaps between reference images and then exploit this redundancy in geometry to cope with the missing parts using a volumetric fusion schemes [18]. However, since we produce structured reconstructions, such fusion strategies are not seamlessly applicable.

4.3 Future Work

Simultaneous Automatic View Selection and Point Cloud Segmentation.

One of the key aspects to address in order to design an end-to-end fully automatic pipeline for reconstructing street-level scenes with a view-dependent reconstruction approach goes through the automatic view selection scheme which would minimize the number of necessary views to explain the scene, and forbid or minimize overlap between cameras in terms of 3D geometry. To jointly consider these criteria, such a method must simultaneously cluster the 3D model which here, is the sparse SfM point cloud or alternatively, its coarse mesh in the vein of [113].

Joint 2D/3D Regularity Mining for Structured Reconstruction.

In terms of scalability, the main computational bottlenecks of our Patchwork Stereo lie in the photo-consistency reprojection costs which is however linear in the number of patches and planar hypotheses. The problem could be posed as a single-view 2D patch to 3D plane fitting using our top-down segmentation and an SfM point cloud where the sparse and noisy natures of such point clouds, in addition to the potential absence of points reprojecting within textureless superpixels. Hence, to adopt such an approach and remove altogether photoconsistency considerations from our model, one should exploit the available sparse 3D cues differently by encoding a behavior for point-less patches. One way to do so elegantly is by leveraging regularities in 2D (patch

appearance in the image domain) as well as in 3D (patch co-planarity). Leveraging 2D regularities in reconstruction has already been considered successfully [109] but at the pixel level. By exploiting the observation that, on a building façade, rectilinear patches which are aligned along one VD are likely to be co-planar the more they are photometrically similar. This simple assumption can help in propagating the existing sparse 3D information in a pure single-view reconstruction scenario using superpixels and sparse SfM.

Extension to Joint Semantic Modeling.

The natural extension of our contributions in the context of urban scenes is to integrate semantic reasoning. A first straightforward way to do so would be by adding a unary term to the energy which would account for the semantic part of the updated label space, and by changing the definition of our connectivity pairwise regularization to allow only certain plane-and-semantic transitions (e.g., favoring the co-planarity of semantic labels which are naturally lying on a common plane such as “wall” and “window” classes, or reasoning on crease transitions). The label-space in such variant would be the cartesian product between planar labels and object classes.

A

Publications

- Amine Bourki, Martin de La Gorce, Renaud Marlet, and Nikos Komodakis. "Patchwork Stereo: Scalable, Structure-Aware 3D Reconstruction in Man-Made Environments." In Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on, pp. 292-301. IEEE, 2017.

B

Bibliography

[1] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Susstrunk, S. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 34(11):2274–2282.

[2] Allène, C., Pons, J.-P., and Keriven, R. (2008). Seamless image-based texture atlases using multi-band blending. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE.

[3] Bailer, C., Finckh, M., and Lensch, H. P. (2012). Scale robust multi view stereo. In *European Conference on Computer Vision*, pages 398–411. Springer.

[4] Barinova, O., Konushin, V., Yakubenko, A., Lee, K., Lim, H., and Konushin, A. (2008). Fast automatic single-view 3-d reconstruction of urban scenes. In *European Conference on Computer Vision*, pages 100–113. Springer.

[5] Berger, M., Tagliasacchi, A., Seversky, L., Alliez, P., Levine, J., Sharf, A., and Silva, C. (2014). State of the art in surface reconstruction from point clouds. In *EUROGRAPH-ICS star reports*, volume 1, pages 161–185.

[6] Biljecki, F. (2017). *Level of detail in 3D city models*. PhD thesis, Delft University of Technology, Delft, the Netherlands.

[7] Biljecki, F., Stoter, J., Ledoux, H., Zlatanova, S., and Çöltekin, A. (2015). Applications of 3d city models: State of the art review. *ISPRS International Journal of Geo-Information*, 4(4):2842–2889.

[8] Birchfield, S. and Tomasi, C. (1999). Multiway cut for stereo and motion with slanted surfaces. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 1, pages 489–495. IEEE.

[9] Bleyer, M., Rother, C., Kohli, P., Scharstein, D., and Sinha, S. (2011). Object stereo – joint stereo matching and object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3081–3088.

[10] Bodis-Szomoru, A., Riemenschneider, H., and Gool, L. V. (2014). Fast, approximate piecewise-planar modeling based on sparse structure-from-motion and superpixels. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 469–476.

[11] Bódis-Szomorú, A., Riemenschneider, H., and Van Gool, L. (2015). Superpixel meshes for fast edge-preserving surface reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[12] Bourki, A., de La Gorce, M., Marlet, R., and Komodakis, N. (2017). Patchwork stereo: Scalable, structure-aware 3d reconstruction in man-made environments. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 292–301. IEEE.

[13] Boykov, Y., Veksler, O., and Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 23(11):1222–1239.

[14] Cazals, F. and Giesen, J. (2004). *Delaunay triangulation based surface reconstruction: ideas and algorithms*. PhD thesis, INRIA.

[15] Chauve, A.-L., Labatut, P., and Pons, J.-P. (2010). Robust piecewise-planar 3d reconstruction and completion from large-scale unstructured point data. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1261–1268. IEEE.

[16] Cornelis, N., Cornelis, K., and Van Gool, L. (2006). Fast compact city modeling for navigation pre-visualization. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1339–1344. IEEE.

[17] Coughlan, J. M. and Yuille, A. L. (2003). Manhattan world: Orientation and outlier detection by bayesian inference. *Neural Computation*, 15(5):1063–1088.

[18] Curless, B. and Levoy, M. (1996). A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312. ACM.

[19] Delong, A., Osokin, A., Isack, H. N., and Boykov, Y. (2012). Fast approximate energy minimization with label costs. *International journal of computer vision*, 96(1):1–27.

[20] Deriche, R. (1987). Using Canny's criteria to derive a recursively implemented optimal edge detector. *International Journal of Computer Vision (IJCV)*, 1(2):167–187.

[21] Dunn, E. and Frahm, J.-M. (2009). Next best view planning for active model improvement. In *BMVC*, pages 1–11.

[22] Elcott, S., Chang, K., Miyamoto, M., and Metaaphanon, N. (2016). Rendering techniques of final fantasy xv. In *ACM SIGGRAPH 2016 Talks*, page 48. ACM.

[23] Esteban, C. H. and Schmitt, F. (2004). Silhouette and stereo fusion for 3d object modeling. *Computer Vision and Image Understanding*, 96(3):367–392.

[24] Felzenszwalb, P. F. and Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *International Journal of Computer Vision (IJCV)*, 59(2):167–181.

[25] Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395.

[26] Fitzgibbon, A., Cross, G., and Zisserman, A. (1998). Automatic 3d model construction for turn-table sequences. *3D Structure from Multiple Images of Large-Scale Environments*, pages 155–170.

[27] Fouhey, D. F., Gupta, A., and Hebert, M. (2014). Unfolding an indoor origami world. In *European Conference on Computer Vision (ECCV)*, pages 687–702. Springer.

[28] Fuhrmann, S. and Goesele, M. (2014). Floating scale surface reconstruction. *ACM Transactions on Graphics (TOG)*, 33(4):46.

[29] Furukawa, Y., Curless, B., Seitz, S. M., and Szeliski, R. (2009). Manhattan-world stereo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1422–1429.

[30] Furukawa, Y., Curless, B., Seitz, S. M., and Szeliski, R. (2010). Towards internet-scale multi-view stereo. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1434–1441. IEEE.

[31] Furukawa, Y. and Hernández, C. (2015). Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148.

[32] Furukawa, Y. and Ponce, J. (2010). Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32(8):1362–1376.

[33] Gallup, D., Frahm, J.-M., and Pollefeys, M. (2010a). Piecewise planar and non-planar stereo for urban scene reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1418–1425.

[34] Gallup, D., Pollefeys, M., and Frahm, J.-M. (2010b). 3d reconstruction using an n-layer heightmap. *Pattern Recognition*, pages 1–10.

[35] Goesele, M., Snavely, N., Curless, B., Hoppe, H., and Seitz, S. M. (2007). Multi-view stereo for community photo collections. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE.

[36] Gröger, G., Kolbe, T., and Czerwinski, A. (2007). Candidate opengis citygml implementation specification (city geography markup language). *Open Geospatial Consortium Inc, OGC*.

[37] Häne, C., Zach, C., Cohen, A., Angst, R., and Pollefeys, M. (2013). Joint 3D scene reconstruction and class segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[38] Hartley, R. and Zisserman, A. (2000). *Multiple view geometry in computer vision*, volume 2. Cambridge Univ Press.

[39] Hernández, J. D., Istenič, K., Gracias, N., Palomeras, N., Campos, R., Vidal, E., García, R., and Carreras, M. (2016). Autonomous underwater navigation and optical mapping in unknown natural environments. *Sensors*, 16(8):1174.

[40] Hiep, V. H., Keriven, R., Labatut, P., and Pons, J.-P. (2009). Towards high-resolution large-scale multi-view stereo. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1430–1437. IEEE.

[41] Hirschmuller, H. and Scharstein, D. (2009). Evaluation of stereo matching costs on images with radiometric differences. *IEEE transactions on pattern analysis and machine intelligence*, 31(9):1582–1599.

[42] Hou, F., Qin, H., and Qi, Y. (2016). Procedure-based component and architecture modeling from a single image. *The Visual Computer*, 32(2):151–166.

[43] Jancosek, M. and Pajdla, T. (2011). Multi-view reconstruction preserving weakly-supported surfaces. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3121–3128.

[44] Knapitsch, A., Park, J., Zhou, Q.-Y., and Koltun, V. (2017). Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (TOG)*, 36(4):78.

[45] Kobyshev, N., Riemenschneider, H., Bódis-Szomorú, A., and Van Gool, L. (2016). Architectural decomposition for 3D landmark building understanding. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*.

[46] Kolmogorov, V. and Zabih, R. (2001a). Computing visual correspondence with occlusions using graph cuts. In *IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 508–515.

[47] Kolmogorov, V. and Zabih, R. (2001b). Computing visual correspondence with occlusions using graph cuts. In *IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 508–515.

[48] Kolmogorov, V. and Zabin, R. (2004). What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 26(2):147–159.

[49] Koutsourakis, P., Simon, L., Teboul, O., Tziritas, G., and Paragios, N. (2009). Single view reconstruction using shape grammars for urban environments. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1795–1802. IEEE.

[50] Kowdle, A., Sinha, S. N., and Szeliski, R. (2012). Multiple view object cosegmentation using appearance and stereo cues. In *European Conference on Computer Vision (ECCV)*, pages 789–803. Springer.

[51] Labatut, P., Pons, J.-P., and Keriven, R. (2007). Efficient multi-view reconstruction of large-scale scenes using interest points, delaunay triangulation and graph cuts. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE.

[52] Ladickỳ, L., Sturgess, P., Russell, C., Sengupta, S., Bastanlar, Y., Clocksin, W., and Torr, P. H. (2012). Joint optimization for object class segmentation and dense stereo reconstruction. *International Journal of Computer Vision (IJCV)*, 100(2):122–133.

[53] Lafarge, F., Keriven, R., Brédif, M., and Vu, H.-H. (2013). A hybrid multiview stereo algorithm for modeling urban scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, pages 5–17.

[54] Laurentini, A. (1994). The visual hull concept for silhouette-based image understanding. *IEEE Transactions on pattern analysis and machine intelligence*, 16(2):150–162.

[55] Lazebnik, S., Boyer, E., and Ponce, J. (2001). On computing exact visual hulls of solids bounded by smooth surfaces. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE.

[56] Lempitsky, V., Rother, C., Roth, S., and Blake, A. (2010). Fusion moves for markov random field optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32(8):1392–1405.

[57] Lezama, J., Gioi, R. G. v., Randall, G., and Morel, J.-M. (2014). Finding vanishing points via point alignments in image primal and dual domains. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 509–515.

[58] Li, Y., Wu, X., Chrysathou, Y., Sharf, A., Cohen-Or, D., and Mitra, N. J. (2011a). Globfit: Consistently fitting primitives by discovering global relations. In *ACM Transactions on Graphics (TOG)*, volume 30, page 52. ACM.

[59] Li, Y., Zheng, Q., Sharf, A., Cohen-Or, D., Chen, B., and Mitra, N. J. (2011b). 2D-3D fusion for layer decomposition of urban facades. In *IEEE International Conference on Computer Vision (ICCV)*, pages 882–889.

[60] Locher, A., Havlena, M., and Van Gool, L. (2016a). Progressive 3d modeling all the way. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 11–18. IEEE.

[61] Locher, A., Perdoch, M., and Van Gool, L. (2016b). Progressive prioritized multi-view stereo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3244–3252.

[62] Löwner, M.-O., Benner, J., Gröger, G., and Häfele, K.-H. (2013). New concepts for structuring 3d city models–an extended level of detail concept for citygml buildings. In *International Conference on Computational Science and Its Applications*, pages 466–480. Springer.

[63] Malik, J., Arbeláez, P., Carreira, J., Fragkiadaki, K., Girshick, R., Gkioxari, G., Gupta, S., Hariharan, B., Kar, A., and Tulsiani, S. (2016). The three r's of computer vision: Recognition, reconstruction and reorganization. *Pattern Recognition Letters*, 72:4–14.

[64] Marr, D. and Poggio, T. (1979). A computational theory of human stereo vision. *Proceedings of the Royal Society of London B: Biological Sciences*, 204(1156):301–328.

[65] Martinović, A., Mathias, M., Weissenberg, J., and Van Gool, L. (2012). A three-layered approach to facade parsing. In *European Conference on Computer Vision (ECCV)*, pages 416–429. Springer.

[66] Mauro, M., Riemenschneider, H., Van Gool, L., and Leonardi, R. (2013). Overlapping camera clustering through dominant sets for scalable 3d reconstruction. In *Proceedings BMVC 2013*, pages 1–11.

[67] Mičušík, B. and Košecká, J. (2008). Multi-view superpixel stereo in man-made environments. Technical report, Technical Report GMU-CS-TR-2008-1, George Mason University, USA.

[68] Mičušík, B. and Košecká, J. (2009). Piecewise planar city 3D modeling from street view panoramic sequences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2906–2912.

[69] Mičušík, B. and Košecká, J. (2010). Multi-view superpixel stereo in urban environments. *International Journal of Computer Vision (IJCV)*, 89(1):106–119.

[70] Moulon, P., Monasse, P., and Marlet, R. (2012a). Adaptive structure from motion with a contrario model estimation. In *Asian Conference on Computer Vision (ACCV)*, pages 257–270.

[71] Moulon, P., Monasse, P., and Marlet, R. (2012b). OpenMVG (open multiple view geometry). `https://github.com/openMVG`.

[72] Moulon, P., Monasse, P., and Marlet, R. (2013). Global fusion of relative motions for robust, accurate and scalable structure from motion. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3248–3255.

[73] Müller, P., Wonka, P., Haegler, S., Ulmer, A., and Van Gool, L. (2006). Procedural modeling of buildings. In *Acm Transactions On Graphics (Tog)*, volume 25, pages 614–623. ACM.

[74] Müller, P., Zeng, G., Wonka, P., and Van Gool, L. (2007). Image-based procedural modeling of facades. *ACM Transactions on Graphics (TOG)*, 26(3):85.

[75] Musialski, P., Wonka, P., Aliaga, D. G., Wimmer, M., Gool, L., and Purgathofer, W. (2013). A survey of urban reconstruction. In *Computer Graphics Forum*, volume 32, pages 146–177. Wiley Online Library.

[76] Olsson, C., Ulén, J., and Boykov, Y. (2013). In defense of 3D-label stereo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1730–1737.

[77] Pauly, M., Gross, M., and Kobbelt, L. P. (2002). Efficient simplification of point-sampled surfaces. In *Proc. of the conference on Visualization*, pages 163–170. IEEE Computer Society.

[78] Pollefeys, M., Nistér, D., Frahm, J.-M., Akbarzadeh, A., Mordohai, P., Clipp, B., Engels, C., Gallup, D., Kim, S.-J., Merrell, P., et al. (2008). Detailed real-time urban 3d reconstruction from video. *International Journal of Computer Vision*, 78(2):143–167.

[79] Pylvänäinen, T., Berclaz, J., Korah, T., Hedau, V., Aanjaneya, M., and Grzeszczuk, R. (2012). 3d city modeling from street-level data for augmented reality applications. In *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2012 Second International Conference on*, pages 238–245. IEEE.

[80] Salinas, D., Lafarge, F., and Alliez, P. (2015). Structure-aware mesh decimation. *Computer Graphics Forum*, 34(6):211–227.

[81] Savinov, N., Hane, C., Pollefeys, M., et al. (2015). Discrete optimization of ray potentials for semantic 3D reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5511–5518.

[82] Saxena, A., Sun, M., and Ng, A. Y. (2009). Make3D: Learning 3D scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 31(5):824–840.

[83] Scharstein, D. and Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision (IJCV)*, 47(1-3):7–42.

[84] Schönberger, J. L., Zheng, E., Frahm, J.-M., and Pollefeys, M. (2016). Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, pages 501–518. Springer.

[85] Schöps, T., Schönberger, J. L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., and Geiger, A. (2017). A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proc. CVPR*.

[86] Schwing, A., Hazan, T., Pollefeys, M., and Urtasun, R. (2011). Distributed message passing for large scale graphical models. In *Computer vision and pattern recognition (CVPR), 2011 IEEE conference on*, pages 1833–1840. IEEE.

[87] Seff, A. and Xiao, J. (2016). Learning from maps: Visual common sense for autonomous driving. *arXiv preprint arXiv:1611.08583*.

[88] Seitz, S. M., Curless, B., Diebel, J., Scharstein, D., and Szeliski, R. (2006). A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Computer vision and pattern recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 519–528. IEEE.

[89] Shen, C.-H., Huang, S.-S., Fu, H., and Hu, S.-M. (2011). Adaptive partitioning of urban facades. *ACM Transactions on Graphics (TOG)*, 30(6):184.

[90] Simon, L., Teboul, O., Koutsourakis, P., Van Gool, L., and Paragios, N. (2012). Parameter-free/pareto-driven procedural 3d reconstruction of buildings from ground-level sequences. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 518–525. IEEE.

[91] Sinha, S. N., Steedly, D., and Szeliski, R. (2009). Piecewise planar stereo for image-based rendering. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1881–1888.

[92] Snavely, N. (2010). Bundler: Structure from motion (SfM) for unordered image collections. v0.4, `http://www.cs.cornell.edu/~snavely/bundler`.

[93] Snavely, N., Seitz, S., and Szeliski, R. (2006). Photo tourism: Exploring image collections in 3D. *ACM Transactions on Graphics (TOG)*, 25(3):835–846.

[94] Tao, H., Sawhney, H. S., and Kumar, R. (2001). A global matching framework for stereo computation. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 1, pages 532–539. IEEE.

[95] Teboul, O., Kokkinos, I., Simon, L., Koutsourakis, P., and Paragios, N. (2011). Shape grammar parsing via reinforcement learning. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2273–2280. IEEE.

[96] Tola, E., V.Lepetit, and Fua, P. (2008). A Fast Local Descriptor for Dense Matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Alaska, USA.

[97] Vanegas, C. A., Aliaga, D. G., and Benevs, B. (2010). Building reconstruction using Manhattan-world grammars. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 0:358–365.

[98] Verleysen, C. and De Vleeschouwer, C. (2012). Recognition of sport players' numbers using fast-color segmentation. In *Visual Information Processing and Communication*, page 83050R.

[99] Verleysen, C. and De Vleeschouwer, C. (2013). Learning and propagation of dominant colors for fast video segmentation. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 657–668. Springer.

[100] Verleysen, C. and De Vleeschouwer, C. (2016). Piecewise-planar 3d approximation from wide-baseline stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3327–3336.

[101] Vogiatzis, G., Torr, P. H., and Cipolla, R. (2005). Multi-view stereo via volumetric graph-cuts. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 391–398. IEEE.

[102] von Gioi, R. G., Jakubowicz, J., Morel, J.-M., and Randall, G. (2008). LSD: a fast line segment detector with a false detection control. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32(4):722–732.

[103] Vu, H.-H., Labatut, P., Pons, J.-P., and Keriven, R. (2012). High accuracy and visibility-consistent dense multiview stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 34(5):889–901.

[104] Wang, J. Y. and Adelson, E. H. (1994). Representing moving images with layers. *IEEE Transactions on Image Processing*, 3(5):625–638.

[105] Watson, B., Müller, P., Veryovka, O., Fuller, A., Wonka, P., and Sexton, C. (2008). Procedural urban modeling in practice. *IEEE Computer Graphics and Applications*, 28(3).

[106] Woodford, O., Torr, P., Reid, I., and Fitzgibbon, A. (2009). Global stereo reconstruction under second-order smoothness priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 31(12):2115–2128.

[107] Wu, C. (2011). VisualSFM: A visual structure from motion system. `http://ccwu.me/vsfm`.

[108] Wu, C. (2013). Towards linear-time incremental structure from motion. In *IEEE International Conference on 3D Vision (3DV)*, pages 127–134.

[109] Wu, C., Frahm, J.-M., and Pollefeys, M. (2011). Repetition-based dense single-view reconstruction. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3113–3120. IEEE.

[110] Xiao, J., Fang, T., Tan, P., Zhao, P., Ofek, E., and Quan, L. (2008). Image-based façade modeling. In *ACM transactions on graphics (TOG)*, volume 27, page 161. ACM.

[111] Zebedin, L., Bauer, J., Karner, K., and Bischof, H. (2008). Fusion of feature-and area-based information for urban buildings modeling from aerial imagery. In *European Conference on Computer Vision (ECCV)*, pages 873–886. Springer.

[112] Zhang, H., Xu, K., Jiang, W., Lin, J., Cohen-Or, D., and Chen, B. (2013). Layered analysis of irregular facades via symmetry maximization. *ACM Trans. Graph.*, 32(4):121–1.

[113] Zhang, R., Li, S., Fang, T., Zhu, S., and Quan, L. (2015). Joint camera clustering and surface segmentation for large-scale multi-view stereo. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2084–2092.