MINISTÉRIO DA CIÊNCIA, TECNOLOGIA, INOVAÇÕES E COMUNICAÇÕES
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m21c/2018/08.13.13.05-TDI

3-DIMENSIONAL (3D) URBAN MAPPING: A STUDY OF DETECTION AND RECONSTRUCTION OF BUILDING'S FACADE THROUGH STRUCTURE-FROM-MOTION (SFM) AND CONVOLUTIONAL NEURAL NETWORK (CNN)

Rodolfo Georjute Lotte

Doctorate Thesis of the Graduate Course in Remote Sensing, guided by Drs. Luiz Eduardo de Oliveira e Cruz de Aragão, and Yosio Edemir Shimabukuro, approved in August 24, 2018.

URL of the original document:
<http://urlib.net/8JMKD3MGP3W34R/3RKQRSE>

INPE
São José dos Campos
2018

PUBLISHED BY:

Instituto Nacional de Pesquisas Espaciais - INPE
Gabinete do Diretor (GBDIR)
Serviço de Informação e Documentação (SESID)
CEP 12.227-010
São José dos Campos - SP - Brasil
Tel.:(012) 3208-6923/7348
E-mail: pubtc@inpe.br

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA, INOVAÇÕES E COMUNICAÇÕES
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m21c/2018/08.13.13.05-TDI

3-DIMENSIONAL (3D) URBAN MAPPING: A STUDY OF DETECTION AND RECONSTRUCTION OF BUILDING'S FACADE THROUGH STRUCTURE-FROM-MOTION (SFM) AND CONVOLUTIONAL NEURAL NETWORK (CNN)

Rodolfo Georjute Lotte

Doctorate Thesis of the Graduate Course in Remote Sensing, guided by Drs. Luiz Eduardo de Oliveira e Cruz de Aragão, and Yosio Edemir Shimabukuro, approved in August 24, 2018.

URL of the original document:
<http://urlib.net/8JMKD3MGP3W34R/3RKQRSE>

INPE
São José dos Campos
2018

Cataloging in Publication Data

Lotte, Rodolfo Georjute.

L917t 3-dimensional (3D) urban mapping: A study of detection and reconstruction of building's facade through Structure-from-Motion (SfM) and Convolutional Neural Network (CNN) / Rodolfo Georjute Lotte. – São José dos Campos : INPE, 2018.

xxvii + 109 p. ; (sid.inpe.br/mtc-m21c/2018/08.13.13.05-TDI)

Thesis (Doctorate in Remote Sensing) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2018.

Guiding : Drs. Luiz Eduardo de Oliveira e Cruz de Aragão, and Yosio Edemir Shimabukuro.

1. 3D urban maping. 2. Facade features. 3. Deep learning. 4. Convolutional Neural Network. 5. Structure-from-Motion. I.Title.

CDU 911.375:004.032.26

Aluno (a): *Rodolfo Georjute Lotte*

Título: "3-DIMENSIONAL (3D) URBAN MAPPING: A STUDY OF DETECTION AND RECONSTRUCTION OF BUILDING'S FACADE THROUGH STRUCTURE-FROM-MOTION (SFM) AND CONVOLUTIONAL NEURAL NETWORK (CNN)"

Aprovado (a) pela Banca Examinadora em cumprimento ao requisito exigido para obtenção do Título de *Doutor(a)* em

Sensoriamento Remoto

Dr. Thales Sehn Körting

Presidente / INPE / São José dos Campos - SP

( ) Participação por Vídeo - Conferência

(X) Aprovado *( ) Reprovado*

Dr. Luiz Eduardo Oliveira e Cruz de Aragão

Orientador(a) / INPE / São José dos Campos - SP

( ) Participação por Vídeo - Conferência

(X) Aprovado *( ) Reprovado*

Dr. Yosio Edemir Shimabukuro

Orientador(a) / INPE / São José dos Campos - SP

( ) Participação por Vídeo - Conferência

(X) Aprovado *( ) Reprovado*

Dr. Fabien Hubert Wagner

Membro da Banca / INPE / São José dos Campos - SP

( ) Participação por Vídeo - Conferência

(X) Aprovado *( ) Reprovado*

Este trabalho foi aprovado por:

( ) maioria simples

(X) unanimidade

São José dos Campos, 24 de agosto de 2018

Aprovado (a) pela Banca Examinadora em cumprimento ao requisito exigido para obtenção do Título de ***Doutor(a)*** em

Sensoriamento Remoto

Dr. Edson Aparecido Mitishita

Convidado(a) / UFPR / Curitiba - PR

( ) Participação por Vídeo - Conferência

(X) Aprovado ( ) Reprovado

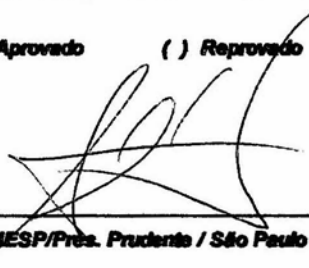
Dr. Norbert Haala

Convidado(a) / IFP / Stuttgart - GER

(X) Participação por Vídeo - Conferência

(X) Aprovado ( ) Reprovado

Dr. Antonio Maria Garcia Tommaselli



Convidado(a) / UNESP/Pres. Prudente / São Paulo - SP

( ) Participação por Vídeo - Conferência

(X) Aprovado ( ) Reprovado

Este trabalho foi aprovado por:

( ) maioria simples

(X) unanimidade

São José dos Campos, 24 de agosto de 2018

"At first I thought I was fighting to save rubber trees, then I thought I was fighting to save the Amazon rainforest. Now I realize I am fighting for humanity". Chico Mendes

"Smart-cities are only fully smart when respect the environment". Rodolfo G. Lotte

To my parents Raul and Fátia, and my sister, Cláudia

ACKNOWLEDGEMENTS

We create an imaginary show inside our heads, a show that involves more frustrations than the hope of reaching our goal. They are daily struggles, daily failures, so many blows that at first we thought we had lost all this time, but actually, it has been taking you to the highest point, to the point of seeing that small failures were all part of the same gear, the conquest. The energy that always moved me was the disbelief in my work, the lack of support, the disloyalty from part of this department, and the fear of not being able to conclude. But that gear was also "lubricated" so many times by faith, family, innumerable friends, places, travels, and so many other events that made things a little less painful. And it is for these people that I toast and dedicate this work today. Thank you to look with me this brand new horizon.

During my PhD, I had the chance to experience the guidance, friendship, support, criticism, and love of incredible people. Everyone who shared this energy played key roles in my evolution and what I am today. Above all, God and Nossa Senhora de Aparecida. Faith and science, the most famous paradox. Lord did not give me the thesis or health, but faith on you will always be the fuel of my determination and hope.

A special thanks to the great sponsors of this work, my parents, Raul and Zélia, and my sister, Claudia. I can not imagine this work, or anything else, being done without your cooperation or presence. Thanks for everything.

To the great INPE, which has hosted me since the master's degree in Applied Computing (2010), to its inspiring infrastructure and environment. To the eternal friends of the postgraduate class in Remote Sensing of 2014, with affection to my friends Aline Jacon, Ana Pessoa, Bruna Pechini, Cesare Neto, David França, Denis Mariano, Henrique Cassol, Jaidson Becker, João Felipe , João Bosco, Ranieli dos Anjos, Sacha Siani and all others who were with me. My friends from Hannover, in special Alexander Schunert, Lukas Schack and Joachim Niemeyer. To my dear friends and advisors Dr. Luiz Aragão and Dr. Yosio Shimabukuro, for having taken this work already underway, for believing, for cooperation and friendship until the end. Dr. Elisabete Moraes, my academic mother and great friend. Thank you for embracing the cause so many times, for understanding and for struggling with me. This work is a special dedication to you.

To the dear friends of the Institute of Photogrammetry (IfP), University of Stuttgart, Germany: Gottfried Mandlburger, Luka Jurjevic, Mateusz Karpina, Patrick Tutza-

uer, Stefan Schmohl, Dominik Laupheimer, Martina Kroma, Lavinia Runceanu, Chia-Hsiang, Ke Gong, Prof. Dr. Michael Cramer and, in particular, Prof. Dr. Uwe Sörgel and Prof. Dr. Norbert Haala, for believing in my work, for the support, reception and follow-up of all my work during the time I was in Stuttgart.

My co-work friends, at Brazilian Space Weather Monitoring Program (EMBRACE/INPE): Dr. Cristiano Wrasse, Dr. Clezio De Nardin, Fauéz and Débora, for the enormous collaboration, infrastructure and support in this journey. Undoubtedly, the valuable lessons and experience as an Analyst and Developer in this department have brought great professional projections and a new insight into practical aspects in problem solving.

Finally, to the Coordination for the Improvement of Higher Education Personnel (CAPES) and the National Council for Scientific and Technological Development (CNPq), thank you for the financial support in my first year of doctorate, and for the great opportunity and teachings I had at the University of Stuttgart (IfP-Stuttgart) - grant PDSE, Process No. 88881.132115/2016-01. This great opportunity was undoubtedly a very important piece in the conclusion of this work. The Army Geographic Service Directorate (DSG), for their support for cartographic specifications regarding 3D mapping in Brazil. To the friend George Longhitano, G-Drones, for the cooperation in acquiring and providing the data for our experiments.

Nobody inside this institute was a motivation for me more than myself. No one will motivate you more than your own determination and courage. This was the biggest win of my life, which will be small compared to my brand new challenges.

ABSTRACT

Urban environments are regions in which spectral and spatial variability are extremely high, with a huge range of shapes and sizes, they also demand high resolution images for applications involving their study. These environments can grow over time, applications related to their large-scale monitoring tend to rely on autonomous intelligent systems that, along with high-resolution images, can help and even predict everyday situations. In addition to the detection of these features, 3D representations of these environments have also been object of study to assist in the investigation of the environmental quality of very dense areas, occupational socioeconomic patterns, the construction of urban landscape models, building demolitions or flood simulations for evacuation plans and strategic delimitation, among countless others. The main objective of this study was to explore the advantages of such technologies, in order to present an automatic methodology for the detection and reconstruction of urban elements, and also to understand the difficulties that still surround the automatic mapping of these environments. Specifically we aimed: (i) To develop a routine of automatic classification of facade features in 2D domain, using a Convolutional Neural Network (CNN); (ii) Using the same images, obtain the facade geometry using Structure-from-Motion (SfM) and Multi-View Stereo (MVS) techniques; (iii) Evaluate the performance of the CNN for different urban scenarios and architectural styles; (iv) Evaluate the performance of the CNN in a real application in Brazil, whose architecture differs from the datasets used in the neural model training; and (v) Classify the 3D model of the extracted facade using images segmented in 2D domain by the Ray-Tracing (RT) technique. In order to atempt that, the methodology was splited into 2D analysis (detection) and 3D (reconstruction). So in the first, a supervised CNN is used to segment terrestrial optical images of facades into six classes: roof, window, wall, door, balcony and shops. At the same time, the facade is reconstructed using the SfM/MVS technique, obtaining the geometry of the scene. Finally, the results of segmentation in both domains, 2D and 3D, are then merged by the Ray-Tracing technique, finally obtaining the 3D model classified. It is demonstrated that the proposed methodology is robust toward complex scenarios. The inferences made with the CNN reached up to 93% accuracy, and 90% F1-score for most of the datasets used. For scenarios not used for training, the neural model reached lower accuracy indexes, justified by the high differentiation of architectural styles. However, the use of deep neural models gives chances for new configurations and use with other deep architectures to improve results, especially for unsupervised models. Finally, the work demonstrated the autonomous capacity of a CNN against the complexity of urban environments, in order to diversify between different styles of facades. Although there are improvements to be made regarding 3D classification, the methodology is consistent and allowed to combine state-of-the-art methods in the detection and reconstruction of urban elements, as well as providing support for new studies and projections on even more distinct scenarios.

Keywords: 3D urban mapping. facade features. deep-learning. convolutional neural network. structure-from-motion.

MAPEAMENTO URBANO TRIDIMENSIONAL (3D): UM ESTUDO SOBRE DETECÇÃO E RECONSTRUÇÃO DE FACHADAS DE EDIFICAÇÕES POR ESTRUTURA-POR-MOVIMENTO (SFM) E REDES NEURAIS CONVOLUTIVAS (CNN)

RESUMO

Ambientes urbanos são regiões cuja variabilidade espectral e espacial é extremamente alta, com uma enorme variedade de formas e tamanhos que remetem igualmente ao sensoriamento remoto de alta resolução em aplicações envolvendo seus estudos. Devido ao fato de que esses ambientes podem crescer ainda mais, as aplicações relacionadas ao seu monitoramento em larga escala tendem a recorrer a sistemas autônomos que, juntamente com imagens de alta resolução, podem ajudar e até predizer situações cotidianas. Aliado à detecção inteligente dessas feições, representações 3D desses ambientes têm sido também objeto de estudo ao auxiliar na investigação da qualidade ambiental de áreas muito densas, padrões socioeconômicos de ocupação, na construção de modelos de paisagem urbanos, avaliação de efeitos de ilhas de calor, demolições de edifícios ou simulações de inundações para planos de evacuação e delimitação estratégica, entre inúmeros outros. Por estes aspectos, o objetivo desta pesquisa de doutorado foi explorar as vantagens de tais tecnologias, de forma a apresentar não só uma metodologia automática para detecção e reconstrução de elementos urbanos, como também compreender as dificuldades que ainda cercam o mapeamento automático desses ambientes. Como objetivos específicos: (i) Desenvolver uma rotina de classificação automática de feições de fachadas no domínio 2D, utilizando-se de uma Rede Neural Convolutiva (CNN). (ii) Com as mesmas imagens, obter a geometria da fachada pelas técnicas de Estrutura por Movimento (em inglês, *Structure-from-Motion* (SfM)) e Estéreo por Multi-Visadas (em inglês, *Multi-View Stereo* (MVS)). (iii) Avaliar o desempenho do modelo neural para diferentes cenários urbanos e estilos arquitetônicos. (iv) Avaliar o desempenho do modelo neural em uma aplicação real no Brasil, cuja arquitetura diferencia-se dos dados utilizados no treinamento do modelo neural. (v) Classificar o modelo 3D da fachada extraída utilizando-se das imagens segmentadas no domínio 2D pela técnica de *Ray-Tracing* (RT). Para tanto, a metodologia do trabalho foi dividida em análise 2D (detecção) e 3D (reconstrução). De forma que no primeiro, uma CNN supervisionada é utilizada para segmentar imagens ópticas terrestres de fachadas em seis classes: telhado, janela, parede, porta, sacada e lojas. Simultaneamente, a fachada é reconstruída pelo uso do *pipeline* SfM/MVS, obtendo-se a geometria da cena. Por fim, os resultados da segmentação no domínio 2D, juntamente com 3D, são então vinculados pela técnica de RT, obtendo-se finalmente o modelo 3D classificado. É demonstrado que a metodologia proposta é robusta em relação a cenários complexos. As inferências realizadas com o modelo neural CNN alcançou até 93% de acurácia, e 90% de F1-score para maioria dos conjuntos de dados utilizados. Para cenários desconhecidos, o modelo neural atingiu índices de acurácia inferiores, justificado pela elevada diferenciação de estilos arquitetônicos. Contudo, a utilização de modelos neurais *deep*, dão margem à novas configurações e uso conjunto com outras arquiteturas *deep* para a melhoria dos resultados, sobretudo, aos modelos não-supervisionados. Por fim, o trabalho demonstrou a capacidade autônoma de

uma Rede Neural Convolutiva frente a complexidade dos ambientes urbanos, de modo a diversificar entre diferentes estilos de fachadas. Embora haja melhorias a serem realizadas quanto à classificação 3D, a metodologia é consistente e permitiu aliar métodos de última geração na detecção e reconstrução de fachadas, além de fornecer suporte à novos estudos e projeções sobre cenários ainda mais distintos.

Keywords: mapeamento 3D urbano. feições de fachadas. *deep-learning*. redes neurais convolutivas. *structure-from-motion*.

LIST OF FIGURES

Page

LIST OF TABLES

LIST OF ABBREVIATIONS

2D	–	Bi-Dimensional
3D	–	Three-Dimensional
AC	–	Auto-Context
ACon	–	Active Contour
AE	–	Autoencoder
ANAC	–	National Civil Aviation Agency
ANN	–	Artificial Neural Network
BA	–	Bundle Adjustment
CASA	–	Civil Aviation Safety Authority
CNN	–	Convolutional Neural Network
COLLADA	–	COLLAborative Design Activity
CONCAR	–	National Commission of Cartography
CPU	–	Central Processing Unit
CVMS	–	Clustering Views for Multi-View Stereo
DL	–	Deep Learning
DOBSS	–	Distinctive Order Based Self-Similarity
DSG	–	Geographic Service Directorate
DSM	–	Digital Surface Model
DTM	–	Digital Terrain Model
EASA	–	European Aviation Safety Agency
FP	–	False Positive
FN	–	False Negative
FC	–	Fully Connected
FCN	–	Fully Connected Network
GAP	–	Gradient Accumulation Profile
GPU	–	Graphical Processing Unit
HOG	–	Histogram of Oriented Gradient
IaaS	–	Infrastructure as a Service
IM	–	Image-Matching
KML	–	Keyhole Markup Language
Laser	–	Light Amplification by Stimulated Emission of Radiation
LiDAR	–	Light Detection And Ranging
LoD	–	Level of Detail
MBA	–	Multicore Bundle Adjustment
MGCM	–	Multi-Photo Geometrically Constrained Matching
ML	–	Machine Learning
MLP	–	Multi-Layer Perceptron
MRF	–	Markov Random Field
MS	–	Mean-Shift
MVS	–	Multi-View Stereo
NC	–	Normalized Cuts

NCC – Normalized Cross Correlation
OGC – Open Geospatial Consortium
PMVS – Path-based Multi-View Stereo
PC – Point Cloud
PSI – Persistent Scatterer Interferometry
Radar – Radio Detection And Ranging
ReLU – Rectified Linear Unit
RR – Recurrent Network
RT – Ray-Tracing
RVR – Reducing View Redundancy
SAR – Synthetic Aperture Radar
SfM – Structure from Motion
SGM – Semi-Global Matching
SIFT – Scale-Invariant Feature Transform
SJC – São José dos Campos
SLAM – Simultaneous Localization and Mapping
TP – True Positive
TN – True Negative
UAV – Unmanned Aerial Vehicle

LIST OF SYMBOLS

f	–	Estimated camera focal length
f_x	–	Estimated camera focal length coordinate in x
$_y$	–	Estimated camera focal length coordinate in y
c_x	–	Principal point offset coordinate in x
c_y	–	Principal point offset coordinate in y
b	–	Estimated skew coefficient
s_x	–	Pixel size coordinate in x
s_y	–	Pixel size coordinate in y
u	–	Pixel coordinate axis u at image coordinate system
v	–	Pixel coordinate axis v at image coordinate system
K	–	Camera intrinsic parameters
R	–	Rotation matrix
ω	–	R_X
ϕ	–	R_Y
κ	–	R_Z
T	–	Translation matrix
P	–	Camera projection matrix
C	–	Center of projection (camera location)
ζ	–	Radial distortion coefficient
ξ	–	Tangential distortion coefficient
Θ	–	Tangential and radial distortion contribution

CONTENTS

Page

1 INTRODUCTION

A 3-dimensional (3D) representation of cities became a common term in the last decade (DEMIR; BALTSAVIAS, 2012). What was once considered an alternative for visualization and entertainment, has become a powerful instrument of urban planning (KOLBE, 2009; STOTER et al., 2011). The technology is now well-known in most of the countries on the European continent, such as Switzerland (AGENCY, 2010), England (ACCUCITIES LTD, 2017; VERTEX MODELLING, 2015), and Germany (virtualcitySYSTEMS GmbH, 2016; ARINGER; ROSCHLAUB, 2014; KRÜGER; KOLBE, 2012; DÖLLNER et al., 2006), also being commercially popular in North America, where many leading companies and precursor institutions reside. However, the semantic 3D mapping with features and applicability that go beyond the visual scope, is still considered a novelty in many other countries. In Brazil, according to the present study (see Section 2.4), the use of volumetric information as a resource for strategic and management planning is reduced to few cities (see Table 3.1).

According to a recent survey (BILJECKI et al., 2015b), approximately thirty real applications with the use of 3D urban models have been reported, ranging from environmental simulations, support of planning, cost reduction in modeling and decision making (YANG et al., 2016; TRUONG-HONG; LAEFER, 2014). Understanding the principles that establish the organization of such environment, as well as its dynamism, requires a structural analysis between its objects and geometry (LAFARGE,). Therefore, reproducing the maximum of its geometry and volume allows studies such as the estimation of solar irradiance on rooftops (BILJECKI et al., 2015a), as well as the determination of occluded areas (EICKER et al., 2015; JOCHEM et al., 2009), in analyzing hotspots for surveillance cameras (YAAGOUBI et al., 2015), Wi-Fi coverage (LEE, 2015), in the urbanization and planning of green areas (TOOKE et al., 2011; AHMAD; GADI, 2003), in evacuation plans in case of disasters (KWAN; LEE, 2005), among others.

Representing cities digitally exactly as they look like in the real world, was considered, for many years, mostly an entertainment application, rather than Cartography. With the appearance of LiDAR (Light Detection and Ranging) (VOSSELMAN et al., 2001) and the Structure-from-Motion (SfM) and Multi-View Stereo (MVS) workflows (SNAVELY et al., 2006), brought to real the structural urban mapping. Even though the data was extremely accurate, the campaigns in mid-2010 were mostly made by airplanes, which fostered large-scale 3D reconstructions, in which buildings can be accurately represented with their rooftops, occupation, area, height or

volume characteristics (SALEHI; MOHAMMADZADEH, 2017). With this remarkable stage, today, new branches of research try not to faithfully represent the scene, but to add knowledge to it, increasingly toward semantic cities[1], where the nature of object is known and the relationship among them could be investigated.

In this sense, acquiring knowledge from remotely sensed data was always a permanent problem for Computer Vision and Pattern Recognition communities, which basically have the mission of interpreting huge amounts of data automatically. Until mid-2012, extracting any kind of information from images would require methodologies that would certainly not fully solve the problem, in many cases, only part of it. However, with the resurgence of Machine Learning (ML) technique in 2012 (KRIZHEVSKY et al., 2012), built on top of the original concept from 1989 (LECUN et al., 1989), has changed the way to interpret images due its high accuracy and robustness on complex scenarios. The respectively ML concept called Convolutional Neural Network (CNN) has enormous potential for interpretation, specially when dealing with large amount of data. In Remote Sensing, has also been used to detect urban objects (ZHENG et al., 2015; BADRINARAYANAN et al., 2015; TEICHMANN et al., 2016) with high quality inferences.

During many years, however, the use of artificial neural models to solve complex problems such as automatic image classification was common, getting popular specially to adapt in complex situations with no human interferences. Still, the methods or workflows available that time were too difficult to use considering the results it could provide, making its popularity lower in mid-2000. Recently, new neural networks architectures brought back the interest by autonomous classifiers, especially those for image classification. CNN has been a trend for pixelwise image segmentation, showing an extreme flexibility to detect and classify any kind of object even in scenarios where humans could not perceive.

Identifying simple facade features such as doors, windows, balconies, and roofs might be a tough task due to the infinite variations in shape, material compositions, and an unpredictable possibility of occlusions. That means not only a good method would be required, but the addition of another variable, such as geometry, should be used to improve class separability. A new demand in the areas of Photogrammetry and Remote Sensing is leading the research to further analysis of these urban objects, in

[1]The term "semantic", in this case, consists in the interpretation of a certain unknown information in something readable. Once this set of unknown infomation is interpreted, it could be confront to another set of interpreted information, studying their relationship and, then, make any kind of decision.

which takes advantage of the aforementioned optical campaigns, such as in Bódis-Szomorú et al. (2017), Riemenschneider et al. (2014), Teboul et al. (2011), Gadde et al. (2017), to acquire the object geometries on a low-cost and simple-to-use manner, which is the use of common cameras and minimal knowledge for acquiring.

Urban environments used to have high spectral and spatial variability, because they are dynamic scenarios, which means that not only the presence of cars, vegetation, vehicles and pedestrians are aggravating the extraction of information, but also the constant actions of man on urban elements. But not all cities are that complex. One city could present a better geometry when compared to another in terms of architectural styles, for example, the streets of Manhattan (wide), in United States, and the streets of Hong Kong (narrow), in China. In addition, suburbs use to have less traffic than city-centers, and that also affects the extraction. The term "complex" in this study refers to images where no preprocessing is performed, no cars are removed, no trees are cut off to benefit the imaging, no house or street was chosen beforehand, only images representing the perfect register of a real chaotic scenario were taken.

Considering these difficulties and the fact that today only a few experiments with complex scenarios have been carried out, a methodology using ground images[2] is used, since they can provide all the facade details that aerial imagery might not be able to (MUSIALSKI et al., 2013). The purpose here, is to delineate interest regions of facade images and assign each of them to a particular semantic label: roof, wall, window, balcony, door, and shop. Right after, these features are used to associate them onto their respective geometries. In order to detect these features, 6 datasets with distinct architectural styles are used as training samples to a CNN model. Once trained, the artificial knowledge generated for each dataset is tested to an unknown scene in Brazil. The facade geometry is then extracted through the use of a SfM/MVS pipeline, which is finally labeled by ray-tracing analysis according to each segmented image.

1.1 Hypotheses

This work is based on the following hypotheses:

- The volumetry of buildings, as well as their facade features (e.g. roof, windows, balconies, wall and doors), can be accurately extracted through optical images and SfM/MVS technique;

[2]Also referenced here as street-side images.

- Facade features can be automatically detected with CNN even under complex scenarios with no preprocessing need;
- The geometric quality of the 3D model, as well as the quality of the 3D labeling, is a direct function of the point cloud density;
- The geometric quality point cloud by SfM/MVS technique depends on the camera parameter estimation, image spectral and spatial characteristics. Therefore, the targets geometry and texture are fundamental in the process of reconstruction and classification.

1.2 Objectives

1.2.1 Main

From structural data campaigns (image-based point cloud, see Section 2.1), the main objective of this research is to explore the extraction of geometric information of buildings, simultaneously, to detect their facade features and, finally, relate both information in one single 3D labeled model.

1.2.2 Specific

a) Develop a routine to classify facade elements in 2-dimensional (2D) images using a CNN architecture;

b) Using the same images, obtain the facade geometry using SfM/MVS;

c) Evaluate the performance of the neural model for different urban scenarios and architectural styles;

d) Evaluate a case study with an application in Brazil, whose architecture differs from the datasets used during the neural model training;

e) Classify the 3D model of the extracted facade using the images previously segmented in the 2D domain by the technique of *Ray-Tracing*.

1.2.3 Thesis's structure

Based on this workflow, in Section 2 it is highlighted the essential urban characteristics for the extraction of information through remote sensing, its challenges and the evolution of techniques. In Section 3, the details of the methodology and data adopted for the study are presented. In Section 4, it is analyzed the results in both

categories: on 2-dimension (quality of detection) and 3-dimension (quality of 3D labeling), as well as the training effects between the architectural style and the inference quality under an unknown one. Finally, in Section 5, our main conclusions and future prospects.

2 LITERATURE REVIEW

Essentially, the results in the range of alternatives to reconstruct cities vary according to the definition of three main phases: (i) sensors and an appropriate measurement of the targets, (ii) processing and classification, according to a desired level of detail and (iii) standardization in well-established formats, such as CityGML (KOLBE et al., 2005). The following sections introduce the main works and methodologies in 3D reconstruction of buildings and facades, as well as the spatial and spectral characteristics usually found in these environments, which represent the great challenges of the area.

2.1 Structural data

Structural data is considered here as all data that carries with it geometric information of the scene, that after being assigned to a valid coordinate system, can represent geometric values very close to the reality. In the following section, it is presented a summary of current platforms and sensors for acquiring this information.

2.1.1 Common alternatives for structural data acquirement

Urban environments are easily recognized in images by its linear aspects, repetitive patterns, with structures frequently devoid of natural elements, with a wide variety of sizes, shapes, compositions and arrangements. Applications involving studies of these environments equally demands a high resolution acquisition, usually performed on a large or small scale. In Table 2.1, is shown a schema regarding conventional types of structural data acquisition and the maximum Level of Details (LoD) that each configuration could reach in terms of urban objects. The list makes a balance between their respective costs, knowledge required to manage the equipment, accuracy, spatial and spectral resolutions, among others. It is important to note the respective table is a non-exhaustive list, which highlights only common forms of acquirement. Other forms of acquisition, such as civil construction cranes, onboard balloons, helicopters, and any other resources are not listed.

Aerial platforms, are categorized here into three segments: long, medium and close-range. Surveys whose sensor is onboard airplanes, flying over altitudes around 1,000 meters, are usually considered as long-range surveys. They are, in fact, the most used category over the years, allowing sophisticate imaging, sometimes, with hybrid sensors and configurations that adequately serve the innumerable urban activities. Although many of these products still present high costs, aerial images are still the

Table 2.1 - Maximum Level of Detail (LoD) and quality of 3D urban models according to the acquisition and sensor characteristics.

Plataform			Spectral region	Spatial resol.	Sensor view	Building parts			Max. LoD	Costs	Operated	PC	Software	Accuracy
						Roof	Facade	Indoor						
Orbital		Satellite	Optical	• • • •	Multi-view	✓	×	×	LoD2	$$$	🎓	⋆⋆⋆	-	• • • •
		Satellite	Laser	• • • •	Multi-view	✓	×	×	LoD2	$$$	🎓	✓	📥	• • • •
		Satellite	Hybrid	• • • •	Multi-view	✓	×	×	LoD2	$$$	🎓	✓	📥	• • • •
Aerial (multiple ranges)	long	Airplane	Optical	• • • •	Nadir	✓	✓ ⋆	×	LoD2⋆	$$$	🎓	⋆⋆⋆	-	• • • •
		Airplane	Laser	• • • •	Nadir	✓	✓ ⋆	×	LoD2⋆	$$$	🎓	✓	📥	• • • •
		Airplane	Hybrid	• • • •	Multi-view	✓	✓	×	LoD3	$$$	🎓	✓	-	• • • •
		Airplane	Optical	• • • •	Multi-view	✓	✓	×	LoD3	$$	🎓	⋆⋆⋆	-	• • • •
	medium	UAV	Optical	• • • •	Multi-view	✓	✓	×	LoD3	$	🎓	⋆⋆⋆	-	• • • •
		UAV	Laser	• • • •	Nadir	✓	×	×	LoD2	$	🎓	✓	📥	• • • •
		UAV	Hybrid	• • • •	Multi-view	✓	✓	×	LoD3	$	🎓	✓	📥	• • • •
	close	UAV	Optical	• • • •	Multi-view	✓	✓	×	LoD3	-	👤	⋆⋆⋆	-	• • • •
		UAV	Laser	• • • •	Multi-view	✓	✓	×	LoD3	$	🎓	✓	📥	• • • •
Terrestrial		User	Optical	• • • •	Multi-view	×	✓ ⋆⋆	×	LoD3	-	👤	⋆⋆⋆	-	• • • •
		User	Laser	• • • •	Multi-view	×	✓ ⋆⋆	×	LoD3	$	🎓	✓	📥	• • • •
		User	Optical	• • • •	Multi-view	×	×	✓	LoD4	-	👤	⋆⋆⋆	-	• • • •
		User	Laser	• • • •	Multi-view	×	×	✓	LoD4	$	🎓	✓	📥	• • • •

Low, **Medium**, **High**, and **Very high**, respectively **(• • • •)**;
Estimated cost ($) - 0 to **3**; **Specialist (🎓)**; **Common user (👤)**; **Embedded (📥)**.

⋆ Parameter is a function of the sensor's Field of View (FOV). In cases of comprehensive FOVs, it is possible to observe not only the characteristics of roofs, but also of facades, allowing an acquisition of values above LoD2. In cases of FOVs with narrow angles, only the building footprint are observable, consequently, only LoD1 is reachable.
⋆⋆ The quality of facades images by terrestrial platforms, depends directly on the height of the building. The imaging of buildings with a height greater than 50 meters, for instance, is affected by the acquisition geometry in its upper part, as well as inner structures.
⋆⋆⋆ Control points are required.

SOURCE: Author's production.

most widely used in large-scale studies, since it does not require a detailed analysis.

Different imaging requires different costs, efforts and technical infrastructure. The estimated cost shown in Table 2.1 is relative and may vary according to area coverage. For example, the cost of orbital laser imaging or high-resolution optical sensors requires a high cost, but may benefit from covering an area whose terrestrial imaging would cover only in parts. The estimated cost in table, therefore, is the absolute cost.

On medium-range, the surveys are normally carried out by fixed-wing Unmanned Aerial Vehicles (UAV) or even by airplanes, whose altitude does not exceed 1,000 meters. In this category, remote data naturally has a gain in resolution, in addition to enabling large-scale imaging, e.g. farms, forests, neighborhoods and others. In facade analysis, only sensors with wide FOV or multi-views gives guarantees of imaging these areas. In Brazil, however, this category still goes through standardization

and other bureaucratic procedures for practice, although it is regularly offered by numerous companies (for details, see Section 2.4 and 3.1).

Finally, close-range surveys consists of small and medium-sized platforms, with target's distance around 200 meters. This specific configuration has gained interest in areas such as Agriculture and Cartography. Close-range devices, such as the UAV (commonly called drones), are usually used for its flexibility, low cost and stability in flying over narrow paths. These qualities, makes this a right alternative to image all the faces of a particular building (e.g. roof, facades, inner gardens). Contrary to the lasers sensors, the UAVs are not only accessible to experts, but also to common users. These small devices are easily purchased on the market for recreational use, but depending on their physical characteristics, they can also be adapted for scientific studies. The handling of laser scanners, however, requires a certain expertise, not just onboard UAVs, but on any other platform.

The terrestrial platforms complete the range of alternatives in close-range imaging, where three main platforms are adopted: total-station, vehicles or even carried by the own specialist. Although the ground survey does not require the use of an aerial platform, this type of imaging has almost the same properties as the close-range aerial survey. In this category, the images are taken with lateral geometry, at the lower sight level, interesting configuration for facade analysis, but not feasible when the purpose requires the complete coverage of the building. For example, indoor gardens, roof, backyards among other structures are hardly observed.

The green color in table, shows the features benefited by the respective configuration, while the yellow determines the dependence of some technical factor. In blue, the maximum LoD reachable by the respective configuration, where the lighters denote non-detailed 3D models, and darkers, detailed ones. In indoor surveys, measures are taken as a complement to lower levels, for instance, LoD2 and LoD3. Of course, the LoD4 model shown would be obtained in addition to an existing model. Measures indoor by itself, do not support the generation of LoD1, LoD2 or LoD3.

2.1.2 Other alternatives for structural data

The generation of 3D urban models are mainly archived using remote systems, either by active or passive sensors, embedded on a long or close-range acquisition platforms, which is determined often by the application purposes, as presented in the previous section. In addition, complete cities may also be reconstructed from instruments other than those previously presented, for example, by radar (Radio

Detection and Ranging), which by its imaging properties, normally allows studies such as terrain variations using SAR (Synthetic Aperture Radar) Interferometry, topography analysis under forest regions, and also cities reconstruction by the technique known as Persistent Scattered Interferometry (PSI) (FERRETTI et al., 2001), that despite presenting accurate results, it is still expensive technology for requiring systematic radar surveys (SCHUNERT; SOERGEL, 2016; SCHACK et al., 2012).

Beyond Photogrammetry, laser, and radar reconstruction techniques, cities can also be generated by procedural models or even manually[1]. The concept of virtual and smart cities has become even more common over the years, and has stimulated the continuous development of standalone applications independent of the imaging devices or platform.

2.2 Urban environments and their representations

The different categories of artificial coverage (man-made) constantly change in small fractions of space and time, often altered by humans as well. Understanding the aspects of texture, geometry, material, architectural styles, coverage, among other physical properties helps define the level of abstraction in the method to be developed (GOOL et al., 2013). The following sections briefly discuss some of the geometric aspects of buildings and their features, in order to contextualize the main factors for 3D modeling and reconstruction of these environments. For a more comprehensive reading, it is recommended the work of (MUSIALSKI et al., 2013).

2.2.1 Buildings

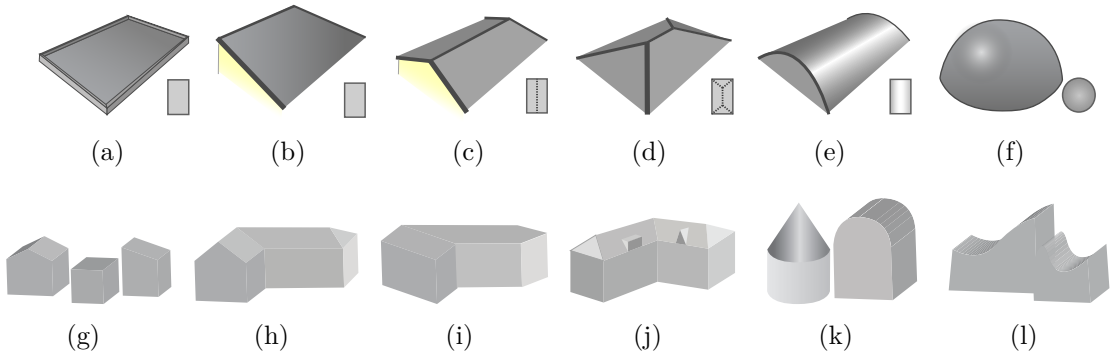
Regardless the type of imaging, artificial structures are easily distinguished from natural mostly by their linear characteristics. Areas of vegetation are generally identified by their homogeneous texture and typical spectral properties, for example, the numerous vegetation indices, which often allow to separate adequately between vegetation and artificial coverage (GERKE; XIAO, 2014; DEMIR; BALTSAVIAS, 2012). However, in some cases, vegetation mixes with urban elements, such as terrace gardens or vertical gardens on balconies, aspects that make it difficult to classify them at the spectral level, but could be easily categorized as being part of a building with the volumetry complement.

[1]Reconstruction is the process of creating digital representations as close as possible to the measurement in terms of accuracy (MUSIALSKI et al., 2013). While generation consists of artificially creating realistic representations given a set of rules or procedural mechanisms (MÜLLER et al., 2006).

A 3D urban model is a representation of an urban environment with 3D geometries of common urban objects and structures, with buildings as the most prominent feature (BILLEN, 2014; LANCELLE; FELLNER, 2010). The perception of a building through the human visual system provides innumerable premises on which characteristics should be considered at first. Although the methodology presented in this study does not use spectral or geometrical properties of buildings to extract their features, some of their characteristics are discussed in this section.

A building consists of a cubic element formed by a flat roof with one or more faces (TUTZAUER; HAALA, 2015; HAALA; BRENNER, 1999). Some of them have dome or arched characteristics, and may overlap each other (American type) or present spectral variations due to their composition. In Figure 2.1, typical roofing examples are illustrated, where it is possible to find from the simplest form (Figure 2.2(a)), to more complex ones (Figure 2.2(e) and 2.2(f)). These characteristics are essential when the extraction of information is carried out essentially by aerial images.

Figure 2.1 - Typical representations of roofs and building topologies. (a) Plane roof. (b) One-face roof. (c) Two-faces roof. (d) Four-faces roof. (e) Arched and (f) domed. (g) and (h) Parametric shape. (i) Prismatic. (j) Polyhedral. (l) Curved and (k) free-form.



SOURCE: Adapted from Haala and Brenner (1999) and Brenner (2000).

Brenner (2000) considers six different models of building classes: parametric, combined, prismatic, polyhedral, curved and free-form (Figures 2.2(g) to 2.2(l)). Parametric buildings are the simplest forms, usually referring to houses or large sheds, sometimes combining one another (Figure 2.2(h)), forming large complexes, usually

present in industrial regions, in which buildings tend to have a sparse disposition. The prismatic model, frequent in commercial areas, represents buildings whose roof is flat and not necessarily in rectangular form. Polyhedral models expose the geometric details, such as chimneys, eaves, different level of roofs, among others. The curved model, is commonly observed in churches or religious buildings. The other buildings, whose shape does not have a defined pattern, are characterized as free forms, such as the Gherkin building in London, England, or the famous Copan building, in São Paulo, Brazil.

In real world, buildings are complex structures with different orientations, slopes, shapes, textures and compositions. In addition, tasks involving the extraction of facades features can be impaired by the presence of other objects surrounding, such as cars, lighting poles, pedestrians, and trees (VERDIE et al., 2015; CHENG et al., 2011). They are located too close to the walls, which could be treat either as an occluder object or could be simply modeled as being part of the wall.

Besides to the physical properties of buildings, choosing a specific platform and sensor would define, for example, which areas of the building are better imaged than the others. The Figure 2.2 illustrated a close-range acquisition through multiple views. The faces in gray represent the regions not observed by the respective type of acquisition, while red consists the ones that totally imaged. Then, the best configuration for buildings, according to the picture, is the multi-view and terrestrial close-range acquisition (Figure 2.3(d)).

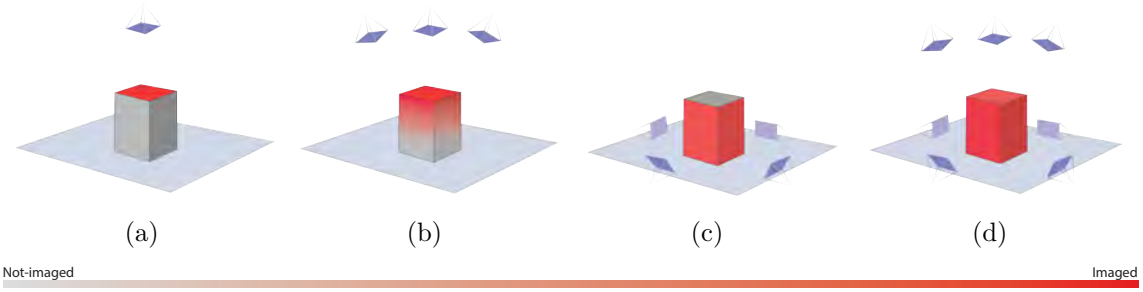
2.2.2 Facade features

The reconstruction of facades fulfills an important segment in inspecting and enforcing urban planning laws. For instance, the mapping of facade features[2] could assist whether a new building can be erected in front or at a given distance from a reference point. Burochin et al. (2014) analyzed the facade characteristics in order to validate building constructions in accordance to French planning laws. Not only the geometry of buildings is important, but their semantics as well. For the validation of urban plans, it is essential that windows and doors are not only geometrically represented but also explicitly labeled as such (TUTZAUER; HAALA, 2015).

In this study, we focus only on the details of the building facades and, so far, there is no better way than close-range imaging to observe those details (as exemplified

[2]Sometimes called openings and referenced here as facade features, such as doors, windows, balconies, gates, etc.

Figure 2.2 - Building's faces benefited according to the platform. (a) Sensors with geometry to nadir, allow to observe only roofs (some sensors, however, have wide FOVs that allow the analysis of adjacent facades). (b) Sensors with multi-view. A trend in urban mapping by aerial platforms. Oblique images provide wide imaging coverage. (c) Ground sensors on board mobile platforms. They allow the observation of details in high resolution. (d) Urban hybrid imaging by terrestrial sensors and close-range UAVs. They allow the observation of all the faces of buildings.

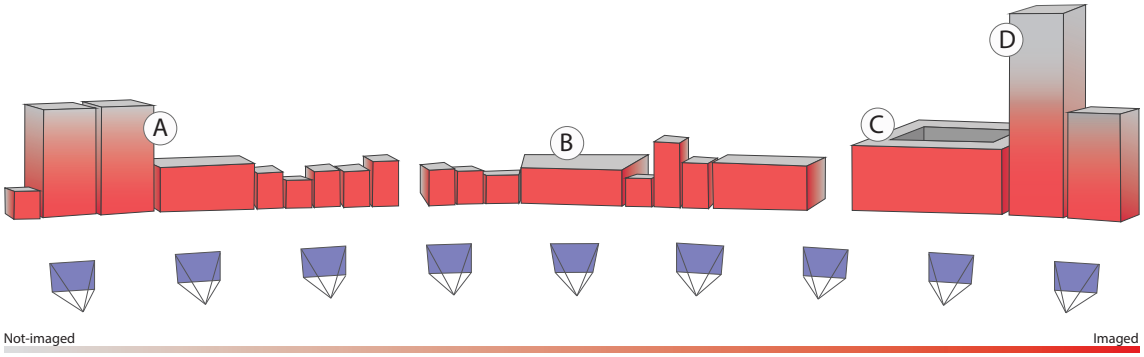


SOURCE: Author's production.

in Figure 2.2). The campaigns of urban data via terrestrial platforms benefits from the rich information collection, but it is a disadvantage as it does not allow a wide observation of the structure, such as their internal architecture, roof and, depending to their height, the upper part. To illustrate this issue, the diagram in the Figure 2.3 shows each of these non-imaged features. The points A and D, the high structures that, due to the limited FOV, might only be partially observed. The point B, the roof characteristics, which are hardly observed by terrestrial configurations. Finally, the point C, another example of structures that could not be observed by such imaging.

The characteristics of doors and windows are pretty much distinguishable, a rectangular geometric pattern, sometimes, occluded by cars, poles, or other objects. The structure's texture uniformity and repeatability of such openings could be verified by the use of radiometric statistics, Histogram of Oriented Gradient (HOG), and Gradient Accumulation Profile (GAP), as in Burochin et al. (2014). However, the symmetry between the openings may not favor the respective method, unless the imaging is done by two different platforms or the architectural style be a well defined facade layout.

Figure 2.3 - Diagram representing a typical terrestrial campaign. In particular, points A and D denote areas that were negatively affected. High buildings are only partially observed in this type of acquisition. B and C, show details of characteristics that can not be observed either by the type of imaging (only street-side view), or because they are internal structures, such as winter gardens - highlighted in C.



SOURCE: Author's production.

In terms of architectural style classification, recent works such as (MATHIAS et al., 2011), (HENN et al., 2012), and (WEISSENBERG, 2014), has addressed the problem to the first stage in a 3D reconstruction methodology: first, to identify what to deal with - Is it a residential area? Is it industrial? The success of any method depends solely on the distribution, testing and execution of micro tasks, which means in dismembering the global task in smaller tasks, in the end, all micro tasks performed and properly tested, will lead to a more consistent method. In this case, the classification of the facade geometry is a micro task of 3D reconstruction, which even though is not approached in this study, it is important to emphasize the need of this stage as a pre-step in 3D reconstruction issue.

The methodology presented here does not include facade layouts categorization. Instead, it is intended to segment facade features by using Machine Learning (ML) techniques, which has proven they are robust under complex scenarios, such as undefined architectural styles or areas occluded by obstacles. The method considers not only multi-scale analysis, but it is also sensitive towards context, when objects obstructing doors or walls, for example, are easily ignored by the neural model when the obstruction does not cover the facade entirely.

2.2.3 Architectural styles

In fact, the number of possibilities to get rid from these scenarios is wide. The range of alternatives, however, can grow even more according to the regions in which it is mapped in the city. Industrial sectors have more differentiated structures than residential or commercial sectors. For example, bigger buildings, few windows, the space in between is also bigger, straight lines geometries for walls, windows, doors, paths, so on. These, however, are more critical because they are sectors of higher traffic, in general, with dynamic structures, greater density that, in the end, make difficult for both to access (at acquisition) and to apply any autonomously method to extract the features.

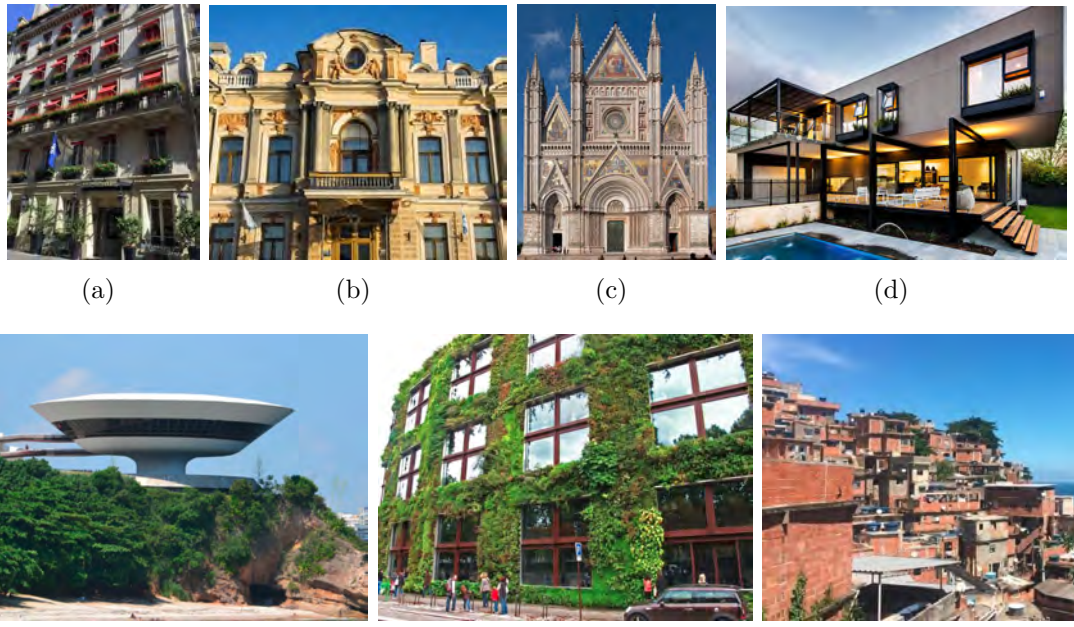
Looking at the architectural styles shown in Figure 2.4[3], the layout of facade features for these styles differ drastically and, in some cases, would be hardly detected by an automatic method due to the variety of shapes, textures, non-symmetry, and other.

The Haussmanian style, in Figure 2.5(a), was the result of a French renovation during the 1870s and represents the predominant style in France, with straight facades, with around 4 to 5 floors, windows often together with balconies, ground floor with the presence of shops, with clear walls and well defined textures. This style, as well as Classicism and Historicism (Figure 2.5(b) and 2.5(c)) also present in many European countries (e.g. Austria, Germany and Spain), is also characterized by the symmetry among the facade features, such that by the layout of the windows and roof, would be possible to estimate the number of floors or amount of internal luminosity of each sector. These styles make up part of the datasets used in this study and in the literature regarding facade feature detection and 3D reconstruction.

Figures 2.5(d) to 2.5(g), on the other hand, present examples of complex architectural geometries that, in terms of autonomous methods of detection and reconstruction, require approaches different from those treated in this study. That is, it is evident that methodologies dealing with the extraction and reconstruction of facade features in Hausmaniann architectures would certainly fail on modern architectures. However, these discussions are difficult to find in the literature that, in certain moments discard this question. Modern architectures have glass in abundance, not only in windows, but also on every surface of their walls. Reflexive surfaces, for example,

[3]Each of them can be access in: Figure 2.5(a): RueMonge2014 dataset; Figure 2.5(b): Link; Figure 2.5(c): Link; Figure 2.5(d): Link; Figure 2.5(e): Link; Figure 2.5(f): Link; Figure 2.5(g): Link.

Figure 2.4 - Examples of architectural styles according to different countries and regions in the city. (a) Hausmaniann. (b) Neoclassic. (c) Religious. (d) Modern. (e) Free-form. (f) Vertical-gardens. (g) *Aglomerados subnormais - favelas.*



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

(e) (f) (g)

SOURCE: Author's production. The images was taken from multiple sources in internet, which authors were unknown.

have features that render the SfM/MVS method useless.

Modern architectures, as in Figure 2.5(e), demarcate another stage of reconstruction. These structures, because they are atypical, do not require automatic methods of reconstruction. They are usually mapped or generated artistically.

Not only vertical gardens, in Figure 2.5(f), but also terraced gardens, are becoming common throughout the years. A sustainable practice that tries to recover the space occupied by constructions, but which represent a difficulty in 3D reconstruction. Its texture is confused with the undergrowth or arboreal vegetation which, for methods that use this metric, can be easily confused. Separate one facade from another, only using spectral information, is still a remaining problem. Although works such as the proposed by Martinovic et al. (2015) still get good separability, the case does not look on real situations, with the presence of any kind of object in front of it, or even under the facade surface.

Finally, although in Figure 2.5(g) the building does not have a defined architecture, it represents one of the most important structures that, once reconstructed, would greatly benefit 3D mapping in Brazil: the subnormal clusters (in Portuguese, *aglomerados subnormais*), popularly called "favelas". These buildings are difficult to access, usually accessible only by walking. By aerial campaigns, they are difficult to map by their spectral characteristics, they do not present any symmetry or pattern. However, they have cultural centers, a large part of the population lives in these areas, it is impossible to measure issues such as sanitation or disasters, attacks on public safety, among others.

It is necessary to attack smaller causes, and science, as in any area, will merge each of these branches according to their context. Although this study has approached the analysis of different scenarios, including complex ones, it is stated that the 3D reconstruction issue must be treated according to the city or region in which intends to study. Urban environments are extremely dense and complex environments. The conception of a generic model that draws attention to all these disparities is impractical. However, constant observation of these structures narrows down the range of options and encourages future studies. Bringing a natural evolution of such routines.

2.2.4 Standardization - Level of Details (LoDs)

The Open Geospatial Consortium (OGC)[4] with the purpose of standardizing the file formats and quantifying different levels of details of 3D models, established the CityGML language, which provides support for the declaration of physical characteristics and relations between urban elements.

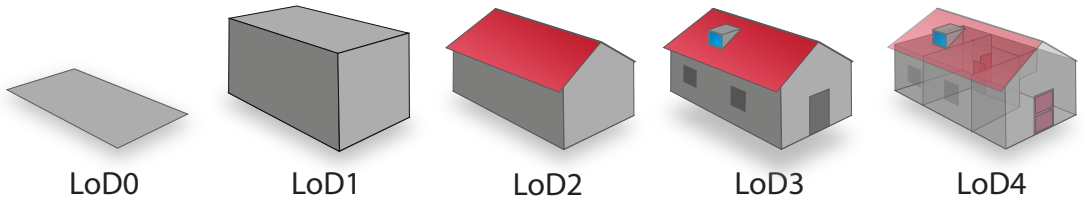
The language consists of a common semantic model for 3D representations of urban environments. Developed as open-source in XML (eXtensible Markup Language) format, the standard adopts the Geography Markup Language 3 (GML), which represents the international standard for spatial data exchange and coding, idealized by OGC and ISO TC211 (KRESSE; FADAIE, 2004). The standard establishes LoDs consisting of degrees that each 3D object is represented. Objects in the same scene may have higher or lower degrees than others.

LoDs are classified into five levels: first, LoD0, corresponds to the basic level of detail, comprising the planimetric information. The second level, LoD1, corresponds

[4] Association of companies, universities and government agencies for the development of publicly available geospatial interface standards to support interoperable solutions, making geospatial services accessible and useful (OGC, 2014).

to the 3D representation, simple extrusions, where the buildings are represented not only by its spatial location, but also by its height. At the LoD2 level, simple structural features such as external columns or garages are aggregated. The LoD3, presents more refined external structural details. At this level, texture components and openings are added to their facades, such as doors, windows and balconies. Finally, the LoD4 represents the highest level, with information on the internal structure, usually, acquired by indoor instruments (KOLBE et al., 2005) (Figure 2.5).

Figure 2.5 - Levels of detail (LoDs) of the standard *CityGML* for 3D buildings models.



SOURCE: Adapted from Kolbe et al. (2005).

Even though there are other standards such as COLLADA (COLLAborative Design Activity) and KML (Keyhole Markup Language), the OCG CityGML standard is particularly the most adopted. CityGML mainly describes the geometry, attributes and semantics of different kinds of 3D urban elements. These can be supplemented with textures or colours in order to give a better impression of their appearance. In addition, semantic information can also be provided, where a description of the geometry, attributes and relationship between different objects can be specified, for example, a building can be decomposed into three parts, or has a safe or a balcony.

2.2.5 Advances in 3D urban reconstruction

The use of high resolution images has been the most effective method for systematic monitoring in all contexts, be it forest, ocean or city. Understanding patterns of changes over time is, however, a task that requires enormous effort when executed manually. In addition, a human interpreter is susceptible to failure and could require prior experience in target perception. As matter of fact, so far few works have

been carried out in the field of architectural style identification or facade feature extraction and reconstruction (MARTINOVIC et al., 2015). Research has reached a certain level of maturity today, with a variety of technologies for acquisition, high graphics processing, storage and dissemination of information that is available for the automatic operators to make use of.

The development of intelligent operators for image labeling can be categorized into model-free, model-based, and procedural-models. The first, classical segmentation methods such as Normalized Cuts (NC) (SHI; MALIK, 2000), Markov Random Fields (MRF) (KOLMOGOROV; ZABIN, 2004), Mean Shift (MS) (COMANICIU; MEER, 2002), Superpixel (ACHANTA et al., 2012), and Active Contours (ACon) (KASS et al., 1987), do not consider the shape of objects or their spectral characteristics, which, in practice, means they always fail in regions where the elements of a same facade do not share the same spectral attributes. Model-based or parametric-model operators, use *a priori* knowledge base, which together with segmentation procedures, provide more consistent results on a given region. However, this knowledge is finite and opens up new possibilities for failures when applied in regions with different characteristics. The third and last, procedural-models like Grammar Shape (STINY; GIPS, 1971), comprise the group of rule-based methods, in which algorithms are applied to the production of geometric forms (TEBOUL et al., 2011).

Image segmenters that carry some intelligence usually carry issues with them as well. First, it is necessary to configure and train the model, then, make inferences in what each pixel or region corresponds to. Teboul et al. (2011) proposed a Grammar-based procedure to segment building facades, where a finite number of architectural styles are considered. The proposed method is able to classify a wide variety of facade layouts and its features using a Tree-based classifier, which improves the detection with only a small percentage of false negatives.

The Grammar-based approaches, however, are normally formulated by rules which follow common characteristics, such as the sequentiality, which is normally present on "Manhattan-world" or European styles (WENZEL; FÖRSTNER, 2008), where the shapes found seem to have lower geometric accuracy since the 3D model is generated, not reconstructed. Still, the outcomes of Grammar-based approaches provide simplicity and perfectly resume the real scene. Other similar works, such as (BECKER, 2009), (NAN et al., 2010), (WAN; SHARF, 2012), and (BOULCH et al., 2013), have proposed equivalents solutions to the problem, but still lies on the same deficiencies mentioned above.

In addition to procedural modeling, urban 3D reconstruction incorporates a new class of research, one based on physical (structural) measurements, which comprise measurements by laser scanners and MVS workflows (mainly). As far as it is known, in this category lie the works that present the most consistent methodologies and results, which take into account the geometric accuracy, where the classification of objects can also be explored by their shapes or volume, in addition to their spectral information.

Jampani et al. (2015) and Gadde et al. (2017), respectively proposed a 2D and 3D segmentation based on Auto-Context (AC) classifier (TU; BAI, 2010). The facade features are explored by their spectral attributes and then, iteratively refined until results are acceptable. The AC classifier applied in a urban environment is a good choice since it considers the vicinity contribution, which is essential in this particular scenario. The downside, however, is that in this case the AC only succeeds when the feature detection (based on spectral attributes) is good enough, otherwise, it could demand many AC stages to get an acceptable output.

Unlike the approaches mentioned above, there are also lines of research that address the problem of 3D reconstruction over the mesh itself, for this reason, other structural data can be explored, such as LiDAR. Lafarge and Mallet (2011), Verdie et al. (2015), Oesau et al. (2016) present different contributions, however, with strictly related focuses. It should be noted, therefore, that the approaches in this line of research demand complex geometric operations, for instance, regularities using parallelism, coplanarity or orthogonality. These operations usually have refinement purposes, and also give the 3D modeling an alternative to acquire more consistent and simplified models.

Automatic 3D reconstruction from images using SfM/MVS workflows is challenging due to the non-uniformity of the point cloud, and it might contains higher levels of noise when compared to laser scanners. In addition to that, missing data is an unavoidable problem during data acquisition due to occlusions, lighting conditions, and the trajectory planning (LI et al., 2016). The papers mentioned in the last paragraph, explore what it is understood as the best methodologies in 3D reconstruction, according to our established goals: exploring the texture first, and then acquiring the semantic 3D model (BROSTOW et al., 2008).

Martinovic et al. (2015) proposed an end-to-end facade modeling technique by combining image classification and semi-dense point clouds. As seen in Riemenschneider et al. (2014) and Bódis-Szomorú et al. (2017), the facade features are extracted and

semantized by the analysis of its texture, not only, the informations extracted are also associated with the geometry, extracted by the use of SfM. Both approached have motivated this thesis in the sense that facade 2D information can be explored more thoroughly in order to improve its volumetry reconstruction. Martinovic et al. (2015), for instance, uses the extracted facade features to analyze the alignment among them, where a simple discontinuity shows the boundaries between different facades. Thereby, it could be used to pre-classify subareas, such as residential, commercial, industrial, and others. Sengupta et al. (2013), Riemenschneider et al. (2014) and Bódis-Szomorú et al. (2017), similarly, explored different spectral attributes in order to discriminate the facade features as well as possible. Moreover, the 3D modeling is later supported by these outcomes by performing a complex regularization and refinements over the mesh faces.

2.3 3D mapping around the World

Investigating the exact number of cities that actually use 3D urban models as strategic tool in their daily lives can be a difficult task. However, Biljecki et al. (2015b) presented a consistent review of entities (industry, government agencies, schools and others) that make or made use of 3D maps beyond the visual purpose. Hence, only applications supported by the respective technology are, in fact, listed. Examples of such applications are the visibility analysis for security cameras installation (MING et al., 2002; YAAGOUBI et al., 2015), urban planning (SABRI et al., 2015; LESZEK, 2015), air quality analysis (AMORIM et al., 2012; JANSSEN et al., 2013), evacuation plans in emergency situations (KWAN; LEE, 2005), urban inventories with cadastral updating (QIN, 2014), among others.

Although the number of cities adopting this tool is uncertain, some of these are known for their technological advances and social development, an important indicator in the implementation of innovative projects. Countries such as the United States, Canada, France, Germany, Switzerland, England, China and Japan are among the leading suppliers of Earth Observation equipment, for example, Leica Geosystems™ (Switzerland) laser systems, FARO™ (USA), Zoller-Fröhlich™ (Germany), RIEGL Laser Measurement Systems™ (Austria), Trimble Inc.™ (USA), TOPCON™ (Japan), and countless optical sensors used in ground and airborne surveys. It is natural, therefore, that these great providers also become reference in conducting research in the sector.

In Germany, the so-called "city-models" were built with the basic purpose of assisting and visualizing simple scenarios or critical situations. At that time, these models did

not have sufficient quality for certain analyzes or permanent updating, making use of the old 2D registers for queries. In the end, the 3D models never became part of the register. The concept of urban 3D reconstruction has become, due to demand, a scientific trend in Cartographic, Photogrammetry and Remote Sensing (RS) almost everywhere in the world, especially in the aforementioned countries.

Naturally, new questions arose. How to merge information already available in 2D databases with the ones in 3D? In certain circumstances, what is the limit on the use of 3D information? When the 2D is already enough and when not? Biljecki et al. (2015b) argue that all applications that require 2D information can be solved with 3D, but that does not make it a unique feature, but an optional one. For example, Kluijver and Stoter (2003) carried out a study of the propagation of noise in urban environments from 2D data. Years later, in Stoter et al. (2008), the study was complemented with 3D information, showing considerable improvement in the estimation.

2.4 3D mapping in Brazil

In Brazil, the Geographic Service Directorate (in Portuguese, *Diretoria de Serviço Geográfico* (DSG))[5] is the unit of the Brazilian Army responsible for establishing Brazilian cartographic standards for the 1: 250,000 and larger scales, which implies standardizing the representation of urban space for basic reference mapping. Recently, the National Commission of Cartography (CONCAR) has put forward the new version of the Technical Specification for Structuring of Vector Geospatial Data (ET-EDGV) (COMISSÃO NACIONAL DE CARTOGRAFIA, 2016), which standardizes reference geoinformation structures from the 1:1000 scale. The data on this scale serve as a basis for the planning and management of the urban geographic Brazilian space. In Brazil, the demand for 3D urban mapping is still low and faces challenges that go beyond its standardization.

As documented in this section, the state-of-art in automatic 3D urban reconstruction covers areas with moderate modeling, whose architectural styles are very specific, streets with large spacing, or symmetry between facade elements, which makes the creation of automatic methods a bit more feasible (GOOL et al., 2013). In countries where there is a high density of buildings, such as Brazil, India, or China, this factor is aggravated by the urban geometry. Many of the Brazilian cities do not have a specific style. In suburbs, for example, this factor can prove even more aggravating,

[5] Available at http://www.dsg.eb.mil.br. Accessed November 2, 2018.

where settlement areas or subnormal settlements (in Portuguese, *favelas* – see Figure 2.5(g)) are all built under these circumstances, with high density and sometimes, erected irregularly or over risky areas, which makes them even more prominent in mapping.

Initiatives such as the TáNoMapa, by Grupo Cultural AfroReggae[6], together with the North American company Google™, consist of mapping hard-to-reach areas such as streets with narrow paths, cliff, among others, by the own local residents. Such areas, in addition to being geometrically complex, require not only cooperation from the government but also from the community who lives there, which by social or security reasons, may require some consent.

Even though it faces many obstacles, Brazilian urban mapping is moving towards more sophisticated levels. In 2016, the National Civil Aviation Agency (in Portuguese, *Agência Nacional de Aviação Civil* (ANAC)) regulated the use of UAVs for recreational, corporate, commercial or experimental use (Brazilian Civil Aviation Regulation, in Portuguese, *Regulamentos Brasileiros da Aviação Civil* (RBAC), "portaria" E nº 94 (ANAC, 2015)). The regulation, widely discussed with society, associations, companies and public agencies, establishes limits that still follow the definitions established by other civil aviation entities such as the Federal Aviation Administration (FAA), the Civil Aviation Safety Authority (CASA) and the European Aviation Safety Agency (EASA), regulators from the United States, Australia and the European Union, respectively (ANAC, 2017). Thus, close-range acquisitions through the use of UAVs became feasible and have fostered research in these fields.

2.5 Geometry extraction

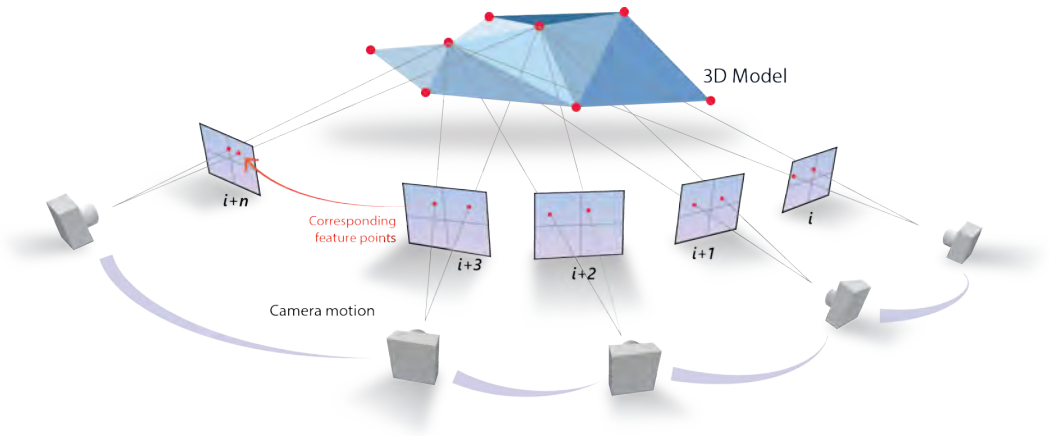
2.5.1 Structure-from-Motion (SfM)

The drawbacks of 3D reconstruction by Photogrammetry are not so many when taking in account its advantages (REMONDINO et al., 2005). In the past decades, close and long-range Photogrammetry have become a widely used tool in 3D city modeling (REMONDINO; EL-HAKIM, 2006; FRASER; CRONK, 2009). Its low costs and availability in remote areas increased its demand, especially in the analysis of such environments. The SfM technique is based on the same physical principles of Stereoscopy, in which a structure can be reconstructed from a series of images taken at small positional variations and overlapping (Figure 2.6). In conventional Photogrammetry, however, both the position and orientation of the sensor must be known. By

[6]Available at https://www.afroreggae.org/ta-no-mapa/. Accessed November 2, 2018.

the SfM technique, the positional parameters can be estimated. In contrast, point clouds generated by SfM have coordinate system in 3D local reference, or image space, which must later be aligned to the coordinate system of the object (WESTOBY et al., 2012).

Figure 2.6 - Structure-from-Motion and camera projection. Instead of a single stereo pair, the SfM technique requires multiple, overlapping images as input to feature extraction and 3D reconstruction algorithms.



SOURCE: Adapted from Westoby et al. (2012).

The workflow to obtain the point cloud using SfM begins with the detection of corners in all input images. For every point, a descriptor based on Histogram Oriented Gradient (HOG) is attributed. Then, a set of descriptors from one image are correlate to another one. The ones remaining are called keypoints. The keypoints are used as descriptors in a correspondence analysis among other images, also known as Image-Matching (IM), where pairs are formed according to the level of correlation between each descriptor (i.e., which feature points in several different images depict the same 3D point on the original object). Once the matching has been concluded, the camera orientation and position estimation is started. With the estimated parameters, the matching points now have depth, which together provides detailed structural information of the entire imaged environment.

The camera position and orientation are estimated iteratively in a process called

Bundle Adjustment (BA) (TRIGGS et al., 1999), which in turn, requires a set of overlapping images sharing a set of measurements of point images at "von Gruber" positions (AGOURIS et al., 2004). The procedure consists of refining the 3D coordinates based on the number of overlapping images and the intersection of the light rays from each point image and camera projection center. The level of overlap and alignment, therefore, determine the quality of estimation (geometric accuracy) of the point cloud in any SfM methodology (SNAVELY et al., 2008; HOFER et al., 2016).

Different approaches for SfM pipelines have change the way how to 3D reconstruct cities. As summarized in the Table 2.1, the main contribution for 3D reconstruction by SfM reside in the fact that geometry can be acquired only using optical images, specialists on image acquisitions are no more required since the pictures available on internet could provide sufficient dataset for 3D modeling. For instance, the Bundler (SNAVELY et al., 2008) takes an unorganized image collection as input, then it is able to reorganize the collection in a way that each pair correspond to an overlapped image.

This approach solves, for example, the question of the physical presence of the human operator in the places of interest. Although it only solves the problem mostly at touristic sites, the idea of using images from ordinary users has stimulated areas such as Cartography, Photogrammetry and Remote Sensing since then. Another example, the VarCity project[7](RIEMENSCHNEIDER et al., 2014), which uses information from a variety of internet sources and integrates them to reconstruct every scenario in 3D.

Keypoints are simply distinctive points denoting the image identity, areas with optimal characteristics for correlating two images, free of spectral and spatial variations, for example, sharp edges and corners. Fundamental in the areas of Computer Vision, Photogrammetry and Remote Sensing, the concept of corner detection is old and is based on the analysis of images in differential space. The issue, in this case, is how to detect corners in one image and at the same time correlate it with another one, which in turn can present variations in scale, texture or orientation? First, Harris and Stephens (1988) proposed to find keypoints using measures based on eigenvalues of smoothed gradients, which allowed the detection of corners independently of rotation, translation and brightness, but not in depth (scale). Then, in 2004, Lowe (2004) proposed the Scale Invariant Feature Transform (SIFT) method, solving the scale problem and also adding the image matching resolution.

[7]Details about the project are available at https://varcity.ethz.ch/. Accessed November 2, 2018.

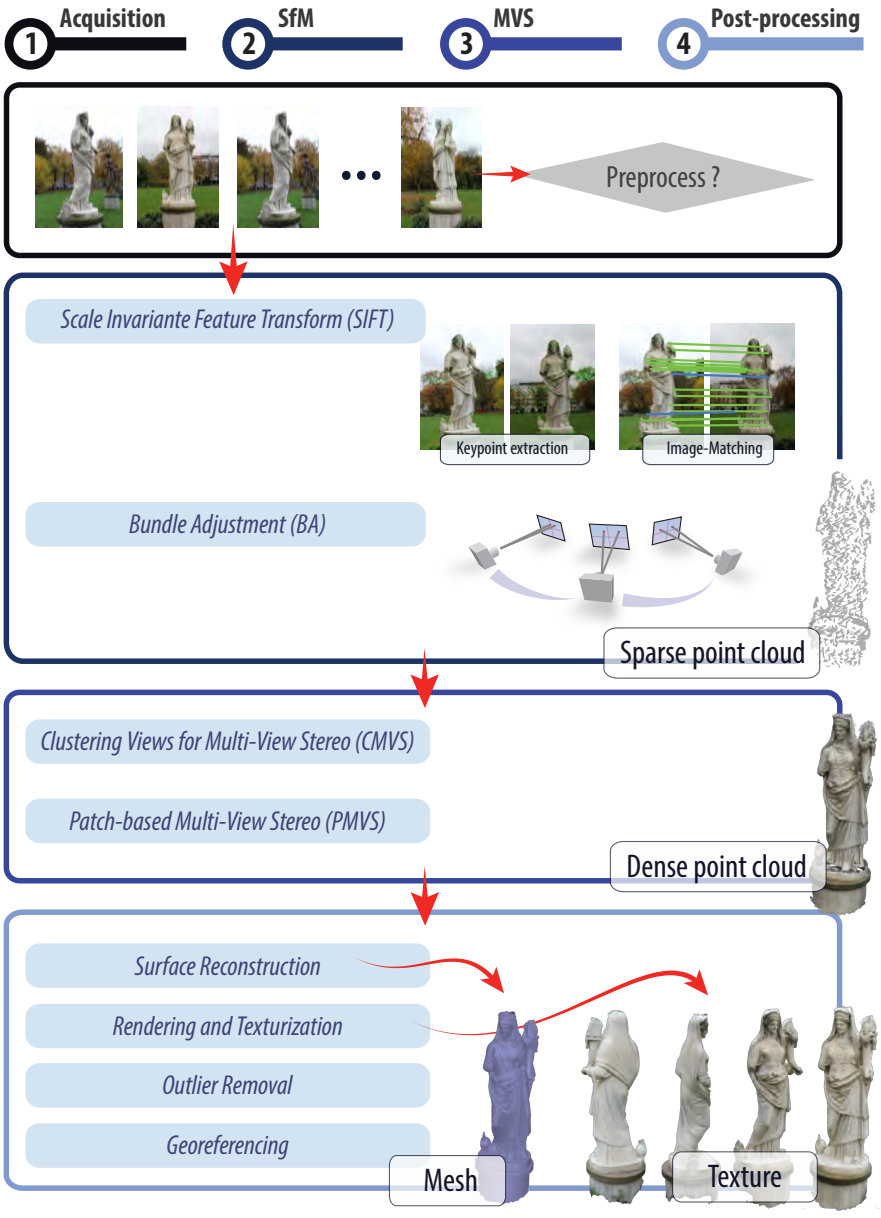
Yuan et al. (2017) categorizes the matching algorithms in two segments, the ones based only in radiometric information and those based either on radiometric and geometric information. Respectively examples of approaches are the Normalized Cross Correlation (NCC) (GONZALEZ; WOODS, 1992), Scale Invariant Feature Transform (SIFT) (LOWE, 2004), and the Distinctive Order Based Self-Similarity (DOBSS) (SEDAGHAT; EBADI, 2015), which stand in the high quality of the images. The second group of algorithms include the Semi-global Matching (SGM) (HIRSCHMULLER, 2008), Path-based Multi-View Stereos (PMVS) (FURUKAWA; PONCE, 2007), and Multi-photo Geometrically Constrained Matching (MGCM) (ZHANG; GRUEN, 2006). These last algorithms take advantage not only from fine radiometric information, but also from the predetermined geometric information, functions as a supplementary matching constraint.

Among these, SIFT has become the most popular. SIFT is independent of scales, variations of brightness, contrast and orientation. After identified, the matching points are determined in pairs of images according to the correlation of local descriptors in each point. A common descriptor is the use of the HOG, which considers the magnitude and orientation of its neighbors as an identifier in the matching process. In the case of low correlation, these are discarded by the algorithm. The size and quantity of images are also part of the problem in the identification of keypoints, depending, demanding days of processing. More sophisticated procedures have improved SIFT by the use of GPUs (Graphic Processing Unit)[8], for example, SiftGPU (WU, 2007).

Some of the SfM stages have been improved over the years. Examples of these transformations are the recent use of BA for the iterative adjustment of camera positioning, the use of SiftGPU (WU, 2007) for computational optimization for image-matching and Multicore Bundle Adjustment (MBA) (WU et al., 2011a), the technique Clustering Views for Multi-View Stereo (CMVS) (FURUKAWA et al., 2010), which the entries are decomposed into smaller processing groups in order to enable parallel processes, PMVS-2 (FURUKAWA; PONCE, 2010), an improved version of PMVS, responsible for the automatic reconstruction of only rigid objects, pedestrians, moving cars and others. In Figure 2.7, SfM is summarized. The post-processing phase is composed of a dense point cloud refinement, where points can be either simplified (points reducing), filtered (noise removal), or used as input for surface reconstruction (e.g. Poisson (HOPPE, 2008)) and texturing.

[8]Electronic circuit dedicated to high-performance graphics processing.

Figure 2.7 - Most recent Structure-from-Motion pipelines. From photographs to sparse point-clouds.



SOURCE: Adapted from Yuan et al. (2017), Westoby et al. (2012).

2.6 Deep Learning (DL)

In terms of image labeling, years of advances have brought what is now considered a gold-standard in segmentation and classification: the use of DL. The technology

is the new way to solve old problems in Remote Sensing (AUDEBERT et al., 2017). It is one of the branches of ML that allows computational models with multiple processing layers to learn representations in multiple levels of abstraction. The term "deep", refers to the amount of processing layers that is usually used during training.

Mitchell et al. (1997) succinctly defines the learning from a ML as: "A computer program learns from a given experience E relative to application classes T with a performance P, if the performance on task T is better than in E ". Widely categorized as unsupervised or supervised, ML is defined by the type of experience that can be applied during the learning process. Shortly, unsupervised learning does not require samples during the training process (learning occurs with the dataset itself), while the supervised assume its behavior according to a reference set. LeCun et al. (2015) believe that unsupervised learning will become far more important in the longer term, since machine learning tends to look more like human learning, which in turn has its learning largely unsupervised: "*they discover the structure of the world by observing it, not by being told the name of every object*".

These models have made remarkable advances in the state-of-art of pattern recognition, speech recognition, detection of objects, faces, and others. Shortly, DL architectures are trained to recognize structures in a massive amount of data using, for example, supervised learning with the concept of backpropagation[9], where the model is indicated the proportion of what must be changed in each of its layers until "learning" occurs (error decay) (LECUN et al., 2015).

Unsupervised learning had a catalystic effect in reviving interest in DL, since then, it has been overshadowed by the successes of supervised learning. LeCun et al. (2015), on the other hand, believes that unsupervised learning will become far more important in the longer term, as the technological tendency is to get closer and closer to human behavior.

2.6.1 Convolutional Neural Network (CNN)

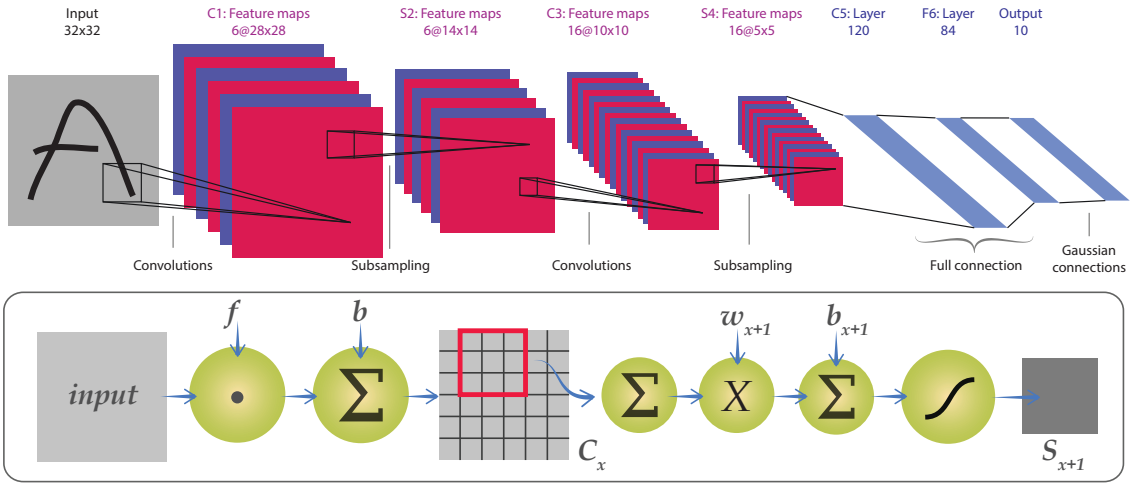
In 1943, the first Artificial Neural Network (ANN) appeared (MCCULLOCH; PITTS, 1943). With only a few connections, the authors were able to demonstrate how a computer could simulate the human learning process. In 1968, (HUBEL; WIESEL, 1968) proposed an explanation for the way in which mammals visually perceive the world using a layered architecture of neurons in their brain. Then, in 1989, a

[9]Backpropagation is a method commonly used in ML to calculate the error contribution of each neuron after each training iteration.

neural model started to get attention not only because of its results, but also for its similarities with the biological visual system, with processing and sensation modules.

LeCun et al. (1989) presented a sophisticated neural model to recognize handwritten characters, named first LeNet, then later, Convolutional Neural Network (CNN). Precisely, the reason for this name is due the successive mathematical operations of image convolutions. Since then, many engineers have been inspired by the development of similar algorithms for pattern recognition in Computer Vision. Different models have emerged and contributed to the evolved state of neural networks in the present days. In Figure 2.8, the LeNet neural model, developed by Yann Lecun in 1989 (LECUN et al., 1989).

Figure 2.8 - LeNet: The first version of a CNN, projected by Yann Lecun in 1989.



SOURCE: Adapted from LeCun et al. (1989).

The CNN is structured in stages, the first ones are composed of two types of layers: convolution and pooling (in the figure, in blue and red). Units in a convolution layer are organized into feature maps or filters, where each unit is connected to a window (also called patch C_x - details in the bottom figure) in the feature map of the previous layer. The connection between the window and the feature map is given by weights (w) and biases (b). The weighted sum of the convolution operations is followed by a nonlinear activation function, called Rectified Linear Unit (ReLU). For many years,

activations in neural networks were composed of smoother functions, such as the $tanh(x)$ or $1/(1+e^{-x})$ sigmoids, but recent study has shown ReLU to be faster on learning in multilayered architectures (LECUN et al., 2015). The fully-connected layer corresponds to a traditional Multi-Layer Perceptron (MLP), with hidden layer and logistic regression. Then, the input to the MLP layer is the set of all features maps at the previous one.

CNN have made advances in image, video, speech and audio analysis, while Recurrent Networks (RR) have supported the path to sequential data such as text and speech (LECUN et al., 2015). It was designed to recognize visual patterns directly from pixel images with minimal preprocessing. Due to convolution over images and its reduced versions (downsampling), they can recognize patterns with extreme variability, scales, and with robustness to distortions and simple geometric transformations such as handwritten characters.

Because of its robustness in image classification and segmentation over complex environments, CNN has been a trend in many Computer Vision applications. For example, the approach is considered gold-standard in applications such as face recognition (LAWRENCE et al., 1997), speech recognition (ABDEL-HAMID et al., 2013), natural language (KALCHBRENNER et al., 2014; CIRESAN et al., 2011), and others. In image analysis, not the first, but the reference on the use of CNNs for images can be named to the model AlexNet (KRIZHEVSKY et al., 2012), when the technique began to be exhaustively tested and became a practical and fast way in object classification.

With the success on ordinary image classification, neural models came to be used also in numerous applications involving Remote Sensing (ZHU et al., 2017). Examples of that are the analysis of orbital images (CASTELLUCCIO et al., 2015; MARMANIS et al., 2016), radar (CHEN; WANG, 2014; CHEN et al., 2016; SUN et al., 2012), hyperspectral (CHEN et al., 2014; ROMERO et al., 2014) and in urban 3D reconstruction (HÄNE et al., 2013; BLAHA et al., 2016; BLÁHA et al., 2016). Although not focused specifically on the analysis of facades, excellent results have been reported in the classification of urban elements through the use of DL.

An example of this evolution can be observed in the annual PASCAL VOC (EVERINGHAM et al., 2015) challenge that, brings together experts to solve classical tasks in Computer Vision and related areas. The applications range from recognition (KRIZHEVSKY et al., 2012) to environment understanding, where the analysis is focused on the relationship between the object itself. Therefore, certain constraints could be imposed to the relation, for example, between a pedestrian and street or

vehicles in applications involving self-driving (BADRINARAYANAN et al., 2015; TEICHMANN et al., 2016), such that a distance and speed constraints could be imposed between these detected objects. Lettry et al. (2017) used CNN to detect repeating features in rectified facade images, wherein the repeated patterns are verified on a projected grid, then, it is used as a device to detect those regular characteristics and reconstruct the scene.

2.6.1.1 Autoencoder

How humans perceive the environment around them is still a topic of recurrent research and, so far, not fully understood. The human visual system consists basically of two processes: radiation capturing and visual perception. The first, numerous chemical and optical processes take place in the eyes of the observer. All visible radiation captured is transformed into synaptic signals by photo-receptor cells, called cones and rods. From this point, these signals representing this lapse moment are then directed to specific regions of the brain, which through chemical reactions and many others to be understood, allow to connect these signals to the perceptual neural senses, which then allows to see and perceive the environment, completing the basic mechanism of the human visual system.

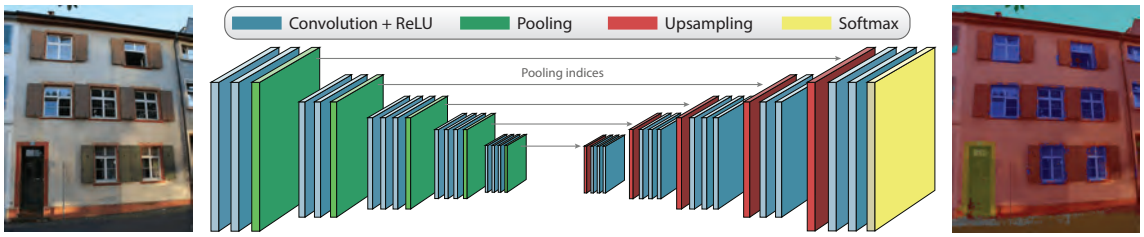
CNNs can assume different processing mechanisms, different numbers of layers and how they are connected. The architecture called Autoencoders or encoder-decoder (BADRINARAYANAN et al., 2015; TEICHMANN et al., 2016), presents similarity not only in architecture but also in the behavior of human visual system. As described in the previous paragraph, two basic mechanisms can be noticed in the biological system: one for processing the radiation in synaptic signals by photo-receptor neurons; and another, for environmental perception, mainly performed by the visual cortex.

Similarly, the encoder and decoder mechanism respectively simulate this two process. First, the input image is subjected to numerous convolution and pooling operations. For each pooling, the downsampling occurs, where the image resolution is reduced and passed on as input to the subsequent convolution layer. Each convolutional layer has a set of convolutional filters, also called filter bank, each of these looks for specific features on image, such as horizontal lines, vertical, curved ones, faces, cars, and others. Consequently, the filter bank learns different patterns in different scales, orientation and location (functions similar to those of the biological lens and retina) as it goes down on deep layers.

Signals in small dimensions are then transmitted to the decoder, which from three

convolution layers performs the upsampling, which is simply deconvolutional operations, the transformation of signals from smaller resolutions to larger resolutions. The result of this transformation is therefore a map of "perceptions", the segmented image (Figure 2.9).

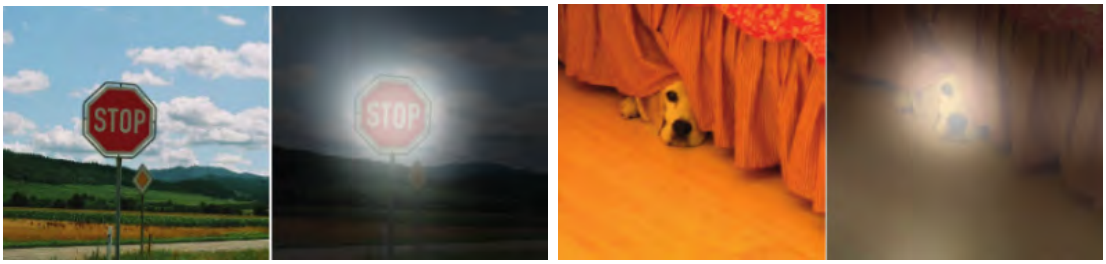
Figure 2.9 - Autoencoder architecture to segment images.



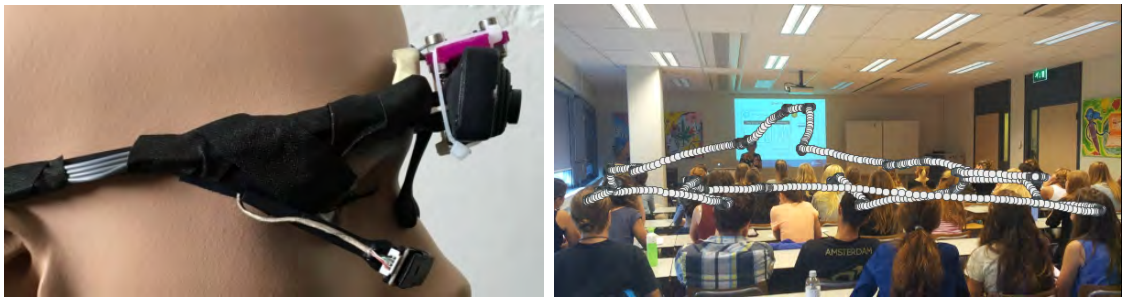
SOURCE: Adapted from Badrinarayanan et al. (2015).

Another feature of these systems and also idealized observing human behavior is the way humans perceive space and objects in their environment. In a lapse of time, innumerable processes are triggered instantly in the brain. In one of these, it is possible to note the human capacity to identify distinct features of all the rest. It is like the human visual system "would not worry" at capturing all the information on that environment, but rather, only those that might be interesting. Permanently, the human eyes scan the environment in very fast movements, process called Saccade (KO et al., 2016; PIRKL et al., 2016). These very fast movements, allow the eyes to capture fractions of images step-by-step. In the Figure 2.10, for instance, the features highlighted by the CNN and the path of the iris when exposing a random image is accurately close.

The semantic segmentation of images by the use of neural networks has gained adherents in a wide range of applications, ranging from the recognition of objects of different natures (KRIZHEVSKY et al., 2012), as well as the understanding of the environment, in which the analysis is also applied on the relation between the objects themselves. Thus, certain constraints could be easily applied, for example, a pedestrian with the object street or automobiles in an application involving autonomous driving (BADRINARAYANAN et al., 2015; TEICHMANN et al., 2016; LETTRY et al., 2017).



(a)



(b)

Figure 2.10 - Analogy to the computational neural system and biological. (a) Output of a Recurrent Neural Network (RNR), a special type of *deep* network. The network is trained to translate high levels of representations into texts. In the figure, the network's ability to focus its attention on specific sections of the image; (b) Eye-tracking device and the capture (gray points) of eyeball movements - Saccade effect.

SOURCE: Adapted from LeCun et al. (2015), Pirkl et al. (2016), Dalmaijer et al. (2014).

As emphasized by Zhu et al. (2017), Remote Sensing data represent a new challenge for DL. Their imaging nature can often be differentiated. For example, optical, laser and radar data provide different representations, which would require analysis that has not yet been so explored with the use of DL.

The DL as an automatic extractor of urban features is a scientific question of great interest to the community and also covers a limited number of works. The efforts so far, show that there is progress in identifying facade features in specific architectural layouts, with well-defined, symmetrical, and accessible modeling facades. The use of benchmark datasets, such as the ones used in this study, is common and provide a wide overview about the extraction algorithms available today.

In Liu et al. (2017), for example, DL is used for the identification of facade features on two online datasets: eTRIMS and ECP datasets, presenting similar results to

those shown in this study. The VarCity project (RIEMENSCHNEIDER et al., 2014)[10], provides accurate perspective on 3D cities and image-based reconstruction. The research involves not only studies in "how" to reconstruct, but how these semantic models could automatically assist to the daily life events (e.g. traffic, pedestrians, vehicles, green areas, among others).

[10] Available at https://varcity.ethz.ch/index.html. Accessed November 2, 2018.

3 MATERIAL AND METHOD

3.1 The current brazilian 3D urban maps

The content presented in this section, represents a complementary study to the Section 2.4, driven by the need to understand the current state of Brazilian 3D urban mapping. The goal in this respective study was to understand more the local mapping infrastructure among the Brazilian capitals, the availability of resources for their urban planning, and in how an accurate 3D urban model would help in the management of the city.

For this task, a poll was elaborated on the use of 3D maps for the strategic planning of municipalities and sent to the secretaries of each Brazilian capital. The capitals were used as reference for the research, therefore, when reporting as being a "non-user" of any 3D information, it was considered that all cities referring to the respective state were also considered as non-users. The questions of the poll were:

a) Does the infrastructure/planning department have urban 3D city maps?
b) If so, how did this benefit the management and urban planning service?
c) If yes, what applications are used today?
d) If not, how is urban planning currently done?
e) How could 3D urban maps contribute?

Among the 26 Brazilian capitals and 1 federal district, only 8 of them answered the poll (Table 3.1). The details of each answer can be seen in Table B.1, in Appendix B.

Even though only a few secretaries have replied, the answers from the largest capitals have been obtained. Despite the different responses, it is believed that the vast majority still adopt the same technologies for 3D surveys: long-range aerial surveys through LiDAR, to roughly observe the urban structuring. Still, the datasets acquired by some of these cities are out of operation for urban planning purposes. The names and emails of those responsible for the answers can be seen in Appendix B.

3.2 Study areas and datasets

To train supervised DL methods with a good generalization, a large dataset is always required. This would contribute to both fine-tuning models and training small networks from scratch (ZHU et al., 2017). In order to vary the building architectural styles, the experiments in this work cover different areas. In recent years, several

Table 3.1 - Poll answers from each Brazilian Capital regarding the use of 3D maps on urban planning.

City	Question 1	Released or Planned?	Products	Sensor	Platform	Scale	Coverage
Fortaleza	×	Planned	–	Laser	✈	Large	–
Vitória	×	Planned	–	Laser	✈	Large	–
Porto Alegre	✓	Planned	DSM, DTM	Laser	✈	Large	Full
São Paulo	✓	Released	DSM, DTM	Laser	✈	Large	Full
Belo Horizonte	✓	Released	–	Laser	✈	Large	Medium
Rio de Janeiro	✓	Released	–	Laser	✈	Large	Medium
Curitiba	✓	Released	–	Laser	✈	Large	Medium
Recife	✓	Released	–	Laser	✈	Large	Full

(✈) onboard of airplane

SOURCE: Author's production.

datasets have been shared in order to contribute in new studies, such as to train deep neural networks, for benchmarks, or simply both.

In Table 3.2, seven different datasets are listed. The first 6 rows are online shared datasets, mainly used for evaluation and to perform benchmarks over different extraction models. They are then used in this study as diversified inputs, since each of them presents different facade characteristics. The last row is a dataset obtained exclusively for this work and used as test images. Eight semantic classes were defined: roof, wall, window, balcony, door, shop, and finally two more, but unrelated to the facade: sky and background. Some of the datasets listed did not provide all eight classes and, in some cases, their annotations had to be adapted.

RueMonge2014: The RueMonge2014 dataset was acquired to provide a benchmark for 2D and 3D facade segmentation, and inverse procedural modeling. It consists of 428 high resolution images, with street-side view (overlapped) of the facade, with Haussmanian architecture, a street in Paris, Rue Monge. Together with the 428 images, a set of 219 annotated images with seven semantic classes are also provided. Due to the geometry of acquisition, the dataset offers the possibility to generate a 3D reconstruction of the entire street scene. Into its evaluation framework, three tasks are proposed: (i) image segmentation; (ii) mesh labeling; and (iii) point cloud labeling.

Center for Machine Perception (CMP): The CMP consists of 378 rectified

Table 3.2 - Datasets for facade analysis and benchmarks. RueMonge2014 and Graz, ETH Zürich; CMP, Center for Machine Perception; eTRIMS, University of Bonn; ECP, Ecole Centrale Paris; SJC, São José dos Campos.

Name	Location	Architecture	Images	Labels	Rectified	PC Generation	Reference
RueMonge2014	France	*Hausmaniann*	428	219	×	✓	(RIEMENSCHNEIDER et al., 2014)
CMP	Multiple	Multiple	378	378	✓	×	(TYLEČEK; ŠÁRA, 2013)
eTRIMS	Multiple	Multiple	60	60	×	×	(KORC; FÖRSTNER, 2009)
ENPC	France	*Hausmaniann*	79	79	✓	×	(GADDE et al., 2017)
ECP	France	*Hausmaniann*	104	104	✓	×	(TEBOUL et al., 2010)
Graz	Austria	Classicism, *Biedermeier*, Historicism, Art Nouveau	50	50	✓	×	(RIEMENSCHNEIDER et al., 2012)
SJC	Brazil	Multiple	175	-	×	✓	-

SOURCE: Author's production.

facade images of multiple architectural styles. Here, the annotated images have 12 semantic classes, among them, some facade features such as pillars, decoration, and window-doors were considered as being part of the wall (for pillars and decoration) and window (window-doors). Then, we have adapted the CMP dataset by unifying its classes and their respective colors.

eTRIMS: The facades in this set do not have a specific architecture style and sequence, as well as in the previous dataset. The eTRIMS provides 60 images, with two sets of annotated images, one with 4 semantic classes (wall, sky, pavement, and vegetation), and another with 8 (window, wall, door, sky, pavement, vegetation, car, and road). For our project, we chose the last, but adapted it to window, wall, and door features only. The other classes were considered as background.

ENPC Art Deco: The ENPC dataset provides 79 rectified and cropped facades in Hausmaniann style. The annotations, however, are shared not in image format, but in text, which also had to be adapted to the 7 classes and colors defined in this work.

Ecole Centrale Paris (ECP): Just like RueMonge2014, the 104 facade images provided by ECP are in Hausmaniann style, but the images are rectified, with cropped facades. In some cases, the classes windows, roof and walls were not perfectly delineated, which may be considered noise by supervised neural models. The same issue can also be found in ENPC dataset. Even though we noticed the problem, no adaptation was performed.

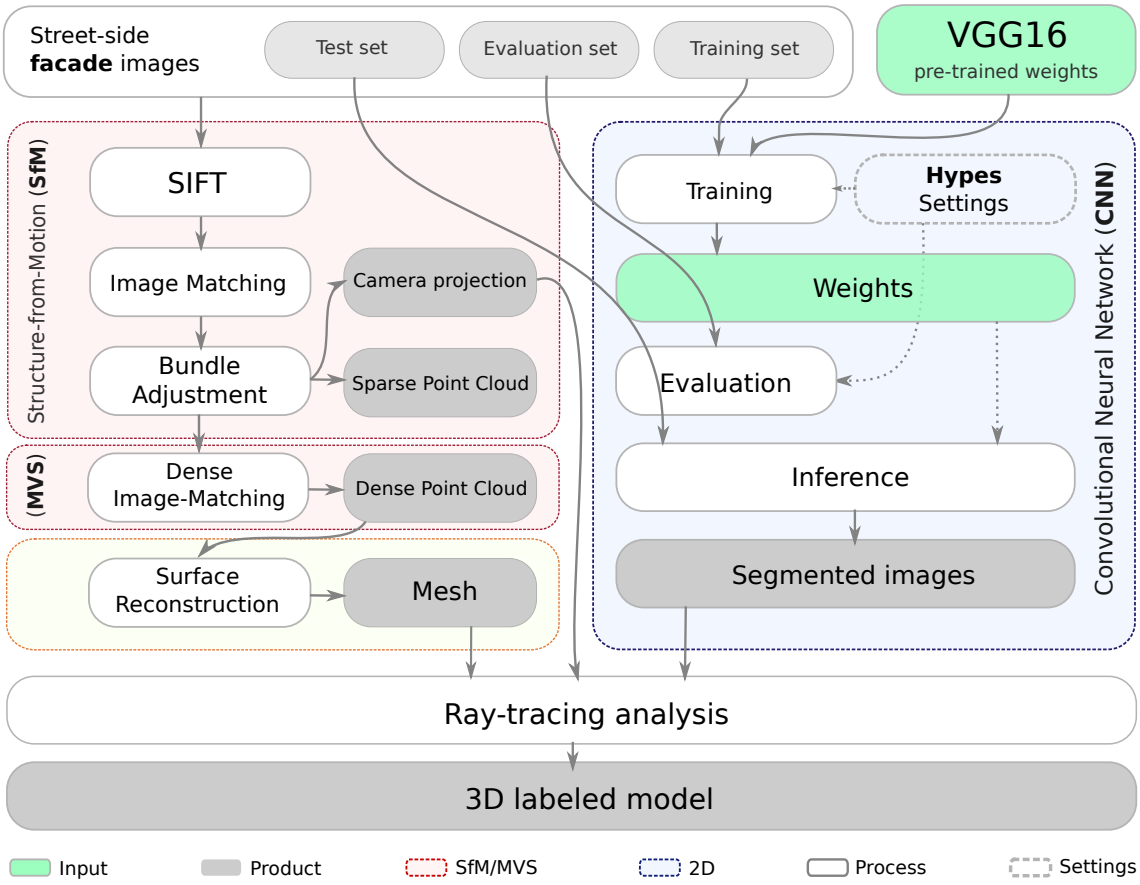
Graz: The Graz dataset consists of multiple architectural styles, selected from the streets in Graz (Austria), rectified, with the same 7 semantic classes defined in RueMonge2014.

São José dos Campos (SJC): The SJC dataset consists of buildings at a residential area in São José dos Campos, São Paulo, Brazil. Like most of the country, the architectural style throughout this city is not unique, often diverging between free-form and modern styles. This set consists of 175 sequential images, overlapped, and taken at the same moment.

3.3 Method

The complete methodology of this case study, as shown in Figure 3.1, consists of three stages: A supervised CNN model for semantic segmentation (blue); Scene geometry acquirement (3D reconstruction) through SfM and MVS pipeline (red); Post-processing procedures (yellow); and 3D labeling through ray-tracing analysis (white). The boxes in gray represent the products, delivered in different steps of the workflow. The following sections, therefore, are presented according to this sequence.

Figure 3.1 - 3D facade model: Facade feature extraction and reconstruction workflow.



SOURCE: Author's production.

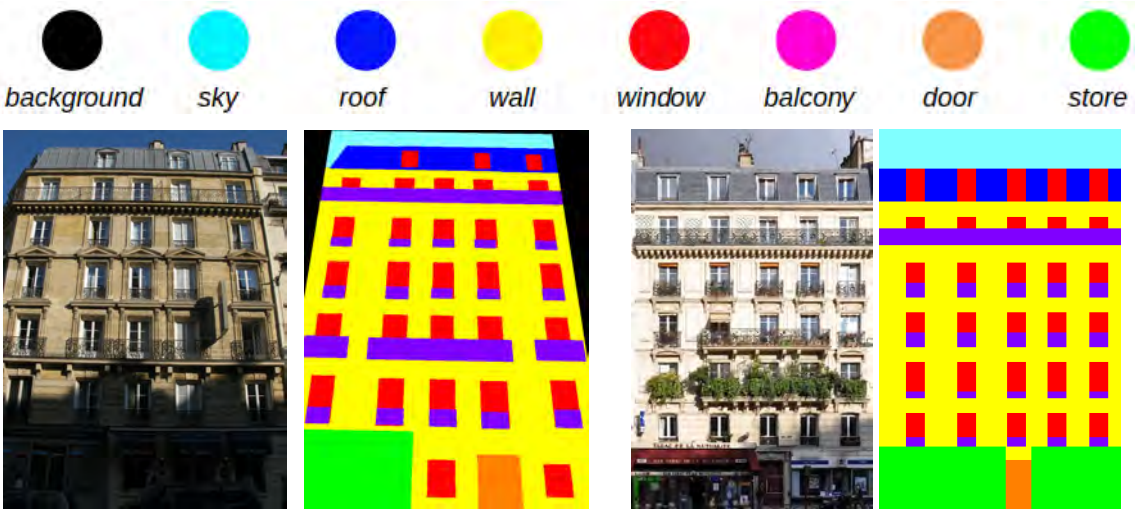
3.3.1 Facade feature detection

3.3.1.1 Training dataset

Each of the 6 datasets has been divided in three different subsets: training, validation, and tests. 80% of the annotated images were used for training, and 20% for validation. Only RueMonge2014 had a non-annotated set of images (209), which was used as test. Due to the small number of training samples, no set of test images was used for the other group of data. Instead, a new acquisition with similar geometry as RueMonge2014 was performed in the city of São José dos Campos (SJC), São Paulo, Brazil. The images will be used only for test, where each of the mentioned

datasets are used for training.

Figure 3.2 - Example of input images for training. (a) Pair of non-rectified images (original and annotated) from RueMonge2014 dataset. (b) Pair of rectified images (original and annotated) from ECP dataset.



(a) (b)

SOURCE: Author's production.

3.3.1.2 Neural model

The classic DL architectures used in visual data processing can be categorized in Autoencoders (AE) and CNN architectures (ZHU et al., 2017) (as discussed in Section 2.6). An AE is a neural network that is trained to reconstruct its own input as an output. It consists of three layers: input, hidden, and output. The hidden layer takes care of all operations behind this model. The weights are iteratively adjusted to become more and more sensitive to the input (GOODFELLOW et al., 2016).

The CNNs, on the other hand, take advantage of performing numerous convolution operations in the image domain, where a finite number of filters are repetitively applied in a downsampling image strategy, which allows the analysis of the scene at different orientations and scales. The convolution is one of the most important operations in signal and image processing, and can be applied in different domains: 1D, in speech processing (SWIETOJANSKI et al., 2014), 2D, in image pro-

cessing (KRIZHEVSKY et al., 2012; AUDEBERT et al., 2017; LONG et al., 2015) or 3D, in video processing (KARPATHY et al., 2014).

Although it has been categorized by Zhu et al. (2017) as being different neural architectures, AE and CNN can be at the same architecture, so called Deep Convolutional Encoder-Decoder, but it has only variations on the layers arrangement, not in the concept, then, considered here simply as CNN.

The deep neural architecture adopted in this study consists of two main components: encoder and decoder. Most recent deep architectures for segmentation have identical encoder networks, i.e. VGG16 (SIMONYAN; ZISSERMAN, 2014), but differ on their decoder, training and inference (BADRINARAYANAN et al., 2015). As the goal in this study is not to test variations of neural architectures, the encoder here corresponds to the same topological structure of the convolution layers of VGG16 (SIMONYAN; ZISSERMAN, 2014), initialized using pretrained VGG weights on ImageNet (KRIZHEVSKY et al., 2012), randomly initialized using a uniform distribution in the range $(-0.1, 0.1)$.

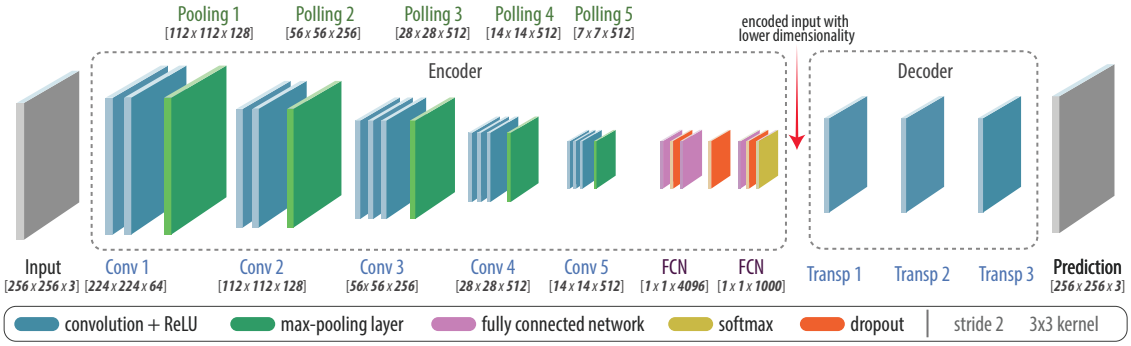
The encoder is originally composed by 13 convolution layers, followed by their respective Pooling and Fully Connected (FC) layers, which in total gives VGG the name of 16. The FC layer, however, is replaced by a 1×1 convolution layer, which takes the output from the last pooling, and generates a low resolution segmentation Teichmann et al. (2016), with dimension of $1x1x4096$ (encoded input, see Figure 3.3). This makes the network smaller and easier to train (BADRINARAYANAN et al., 2015).

The decoder is seen as a component that interprets chaotic signals into something intelligible, similar to the human senses. For example, it would be like the equivalence between a noise (signal) and a person talking in a known language (interpreted signals by our brain), radiation (signal) and the perception of being under a garden with flowers and animals (environment perception, which is the brain interpretation of this same radiation), etc. The neural architecture presented by (TEICHMANN et al., 2016), for instance, used 3 different decoders with 3 different tasks in a way that real-time application could be performed. Multiple decoders mimics the human multiple sense being performed simultaneously by the brain. Once the purpose here is only the visual sense, the use of multiple decoders is not interesting due to useless computational demand, and unnecessary processing.

Using the previous analogy, the input signals to the decoder is the final matrix

from the encoder, which performs the upsampling operation resulting in a pixelwise prediction. The operations performed by this component is composed by 3 deconvolutional layers, so the predictions result in upsampled in the same size as the input image. To better understand the fully path, in the Figure 3.3 is illustrated the details of the final CNN used in this study, as just described.

Figure 3.3 - Convolutional Neural Network details used in this study. Encoder-Decoder convolutional architecture.



SOURCE: Author's production.

The convolution layers sequentially perform the linear operations regarding a certain kernel size and stride (the number of pixel jumping over the kernel slide). Each convolution layer has also a pre-determined number of filter, their functionality is to learn local features, such as horizontal lines, vertical, curves, so on. As soon as the convolution layers are going further, these filter become more and more, since the dimensionality get decreased. In green, the pooling layers take the convolution result and downsampled it to the next convolution layers. In this transition, the dimension get half of its original size.

When the encoded input finally encounter the softmax layer (resulting in the probabilities performed by the cross-entropy function), 3 transposed layers decode the signals (deconvolution). In other words, the decoder performs the upsampling based on pooling indexes (from the encoder polling), finally getting the predicted map. The dropout layer is a technique for reducing overfitting, which avoids some redundant or complex adaptations on training data.

3.3.2 Multi-View surface reconstruction

The input is a set of images which are initially fed to standard SfM/MVS algorithms to produce a 3D model. To produce the 3D model using the standard SfM/MVS algorithms, the set of optical images to be presented must respect some specifications. For example, not all datasets listed in Table 3.2 have properties that could allow the application of SfM/MVS. Random, rectified, and cropped images are not overlapped or, at least, were not taken in the same moment. Only with RueMonge2014 it was possible to run this experiment. A case where random images are taken at different time was proposed by Snavely (2009), but not used in this study.

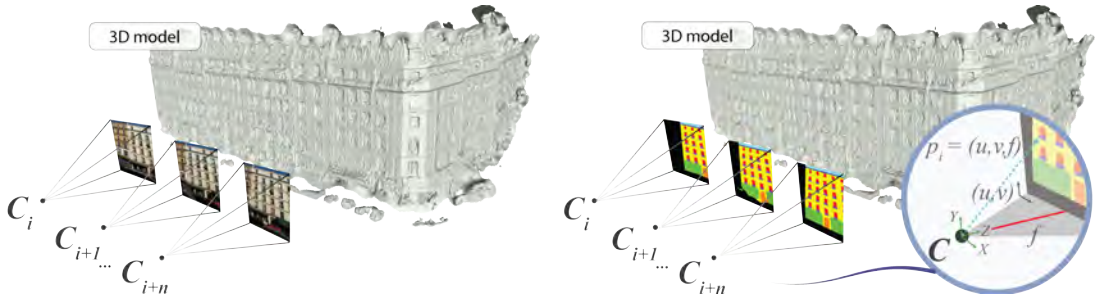
Every measure made by any photogrammetric technique would require the camera calibration in order to obtain high accurate metric information, such as depth and dimensional measurements from 2D domain (HE et al., 2006). Camera calibration involves the estimation of either internal (intrinsic) and external (extrinsic) parameters to a certain camera. These determine the relation between the scene and the instrument itself. The first, consists in the parameters particular to the camera, where is determined how the image coordinate of a point is derived, given the spatial position of the point with respect to the instrument (camera positioning). The external, on the other hand, determines the geometrical relation between the camera and the scene, or even different cameras (LLC, 2018).

The SfM, however, is a fully automatic and iteratively way to solve this need. During the process, the intrinsic and extrinsic parameters are estimated and the outcome point cloud is then densified by the MVS technique. In this study, the estimated internal and external parameters are also used to ray-trace the segmented image toward the mesh (see Section 3.3.3), where they are preserved from the reconstruct operation to ray-tracing (Figure 3.4). The internal parameters are then given by the distance between the lens and image sensor, called focal length (f), the principal point offset (c_x, c_y), the skew coefficient (b), and the pixel size (s_x, s_y).

The lens distortions parameters not considered

The external are given by two parameters, R and T. They are the coordinate system transformations from 3D world coordinates to 3D local camera coordinates system, where $R = (R_X(\omega), R_Y(\phi), R_Z(\kappa))$ is the rotation matrix, and $T = (T_X, T_Y, T_Z)$, the translation matrix, consisting in the local camera coordinate system with origin at the camera projection center. The Z axis points towards the viewing direction, X axis to the right, and Y axis points down. The image coordinate system has origin

Figure 3.4 - Facade geometry obtained from photos. (a) Street-side images of facades and its reconstructed surface after the SfM/MVS technique. (b) The camera parameters are kept and the photos are replaced by the CNN facade features predictions.



(a) (b)

SOURCE: Author's production.

at the top left image pixel, with the center of the top left pixel having coordinates $(0.5, 0.5)$, u_0 and v_0, respectively (bottom-right in Figure 3.5(b)).

Summarizing, the camera can be described by a matrix notation, which is the cumulative effect (Π) of all parameters (HARTLEY; ZISSERMAN, 2004):

$$\Pi = K \; P \; [R \; T] \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} , \tag{3.1}$$

where K corresponds to the intrinsic parameters:

$$K = \begin{bmatrix} f_x & b & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} , \tag{3.2}$$

which can be decomposed on the three principal 2D units: principal point offset, focal length, and skew:

$$K = \begin{bmatrix} 1 & 0 & u_0 \\ 0 & 1 & v_0 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} f_x & 0 & 0 \\ 0 & f_y & 0 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & \frac{b}{f} & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} , \tag{3.3}$$

where f_x and f_y correspond to f, with same value. P denote a 4×3 matrix, which corresponds to the projection:

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} . \tag{3.4}$$

The rotation and translation matrices are then given by:

$$R = \begin{bmatrix} cos\kappa\ cos\phi & -sin\kappa\ cos\phi & sin\phi & 0 \\ cos\kappa\ sin\omega + sin\kappa\ cos\omega & cos\kappa\ cos\omega - sin\kappa\ sin\omega\ sin\phi & -sin\omega\ cos\phi & 0 \\ sin\kappa\ sin\omega - cos\kappa\ cos\omega\ sin\phi & sin\kappa\ cos\omega\ sin\phi + cos\kappa\ sin\omega & cos\omega\ cos\phi & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} , \tag{3.5}$$

and:

$$T = \begin{bmatrix} 1 & 0 & 0 & T_X \\ 0 & 1 & 0 & T_Y \\ 0 & 0 & 1 & T_Z \\ 0 & 0 & 0 & 1 \end{bmatrix} , \tag{3.6}$$

and the camera position C given by:

$$C = -R^{-1}T . \tag{3.7}$$

All the procedures in this stage were performed through Agisoft™ PhotoScan®. Very similar results could be reached by the use of VisualSfM (WU et al., 2011b) or COLMAP (SCHöNBERGER; FRAHM, 2016), for instance. The respective softwares have more flexible licenses and preserve most of the guarantees proposed by Photoscan. However, it still needs some expertise to install and to use. PhotoScan incorporates an improved SIFT algorithm for Feature Matching across the photos; for camera intrinsic and extrinsic orientation parameters, the software uses a greedy algorithm to find approximate camera locations and refines them later using BA.

In addition to the camera parameters estimation, operations involving the use of SfM require good photo-taking practices. One of these is the texture conditions of the targets. Targets whose surface is too homogeneous or with specular properties, will harm the method. Any instrument configuration (e.g. zoom, lens or distortions), once configured, must be preserved until the end of the acquisition. In case of facades, the motion between one photo and another must have orientation as constant as possible and always perpendicular to the target. Therefore, these condictions were

considered during the acquisition of SJC images.

The geometric accuracy in this study corresponds to the proximity between the reconstructed model and the point cloud, not necessarily to the positional part. In this case, it was assumed that the point cloud had previously proven positional accuracy. It was beyond the scope of this study to analyze adjustments or positioning issues. Then, no geometrical accuracy has been reported due to the fact that no 3D references were modeled to compare against the 3D models reconstructed.

3.3.3 3D labeling by ray-tracing analysis

At this methodological point, two products were archived: the classified facade features (2D image segmentation) and their respective geometry (mesh). The idea here is to merge each feature to its respective geometry, and that can be done by analysing the ray-tracing of each image respect to their camera projection (estimated during the SfM pipeline) onto the mesh.

Often used in computer graphics for rendering real-world scenarios, such as lighting and reflections, ray-tracing analysis mimics real physical processes that happen in nature. A energy source emits radiation at different frequencies of the electromagnetic spectrum. The small portion visible to human eyes, called the visible region, travels straight in wave forms and it is only intercepted when it encounters a surface in its trajectory. Such surface has specific physical, chemical and biological properties, responsible to define the radiation behavior under this specific structure. A surface whose absorption characteristics is high, the tendency is that this object has a dark appearance, as the radiation incident on its surface is absorbed, and a minimal portion is reflected to the eyes.

Each facade image, in essence, is the record of the reflection of electromagnetic waves in a tiny interval of time, captured by a sensor at a certain distance and orientation. Once the camera's projection parameters are known, the original images used for its estimation during SfM are replaced by the CNN predictions. Thus, the "reverse" ray-tracing process can then be performed.

The ray-tracing consists in the transformation from point coordinates in the local camera coordinate system to the pixel coordinates in the image. It is essential to consider in which context the photos were taken in order to use the right transformation from the center of projection toward the mesh. For example, if a fish-eye model transformation is used in camera frame model, the result would certainly not

be the right one.

Semi-professional or even professional cameras are usually used to make these campaigns. Some cameras provide several categories of parametric lens distortions, some of them, by default, also include non-distorted configurations (RAW images), which does not apply any transformation to the image. Therefore, both RueMonge2014 and SJC are images obtained in the camera frame model, which pixel coordinate (u, v) from the 3D point projection are given by the following transformation. Let:

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{X}{Z} \\ \frac{Y}{Z} \end{pmatrix} \tag{3.8}$$

be the homogeneous point r the squared 2D radius from the optical center:

$$r = \sqrt{x^2 + y^2} \ , \tag{3.9}$$

then, all the coefficients of distortions have to be considered. Then, the pixel coordinate (u, v) of the 3D point projection with distortion model is given by:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} (w * 0.5) + c_x + \Theta_x f_x + \Theta_x b_1 + \Theta_y b_2 \\ (h * 0.5) + c_y + \Theta_y f_y \end{pmatrix} \ , \tag{3.10}$$

where b_1 and b_2 are the skew coefficients, w and h the image width and height. The parameter Θ denotes the radial and tangential distortions:

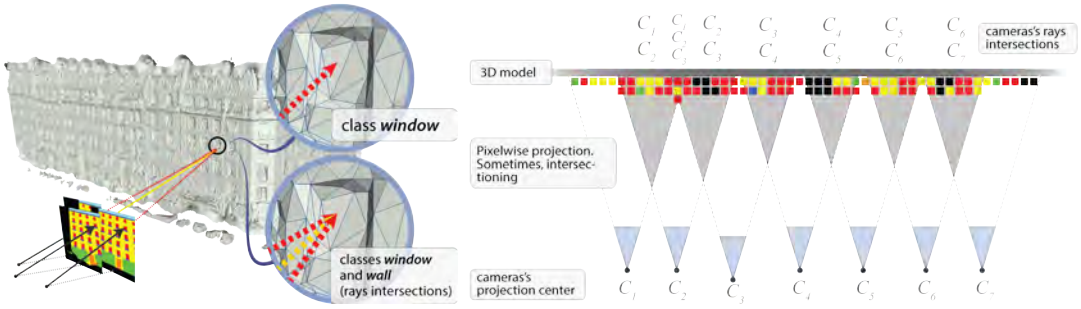
$$\begin{pmatrix} \Theta_x \\ \Theta_y \end{pmatrix} = \begin{pmatrix} x * (1 + \xi_1 r^2 + \xi_2 r^4 + \xi_3 r^6 + \xi_4 r^8) + (1 + \zeta_3 r^2 + \zeta_4 r^4) * (\zeta_1 (r^2 + 2x^2) + 2\zeta_2 xy) \\ y * (1 + \xi_1 r^2 + \xi_2 r^4 + \xi_3 r^6 + \xi_4 r^8) + (1 + \zeta_3 r^2 + \zeta_4 r^4) * (\zeta_2 (r^2 + 2y^2) + 2\zeta_1 xy) \end{pmatrix} \ , \tag{3.11}$$

where ξ_1, ξ_2, ξ_3, and ξ_4 are the radial and, ζ_1, ζ_2, ζ_3, and ζ_4, the tangential distortion coefficients. As the images used have no radial and tangential distortions, they are null and turn Θ as being the own axis x and y.

Now, knowing the pixel class from the incident ray (Figure 3.6(a)), the mesh triangle is finally labeled as such. It is evident, therefore, that the rays trajectory from the images can intersect one another. Because of that, different rays can reach an identical point on the mesh, what in fact creates questions such as "which class should be assigned to each individual mesh facet?". (RIEMENSCHNEIDER et al., 2014) proposed the Reducing View Redundancy (RVR) technique, where the number of overlapped images is reduced, which does not fit to our purpose, since the greater the number of overlaps, better the labeling (more classes to choose).

If more than one ray reach the same triangle (detail in Figure 3.6(a)), then, it is labeled by the most frequent class from the C_n rays (Figure 3.6(b)).

Figure 3.5 - Ray-tracing analysis: This diagram shows the intersections of rays between the overlapped images, where the class assignment is made by choosing the most frequent class (mode) at the intersections. The colors on the right side of the picture, correspond to the pixels from different images, overlapping the same region on the mesh. To decide which class to assign, a simple mode (most frequent class) operation is used. The labeling legend can be seen in Figure 3.2.



(a) (b)

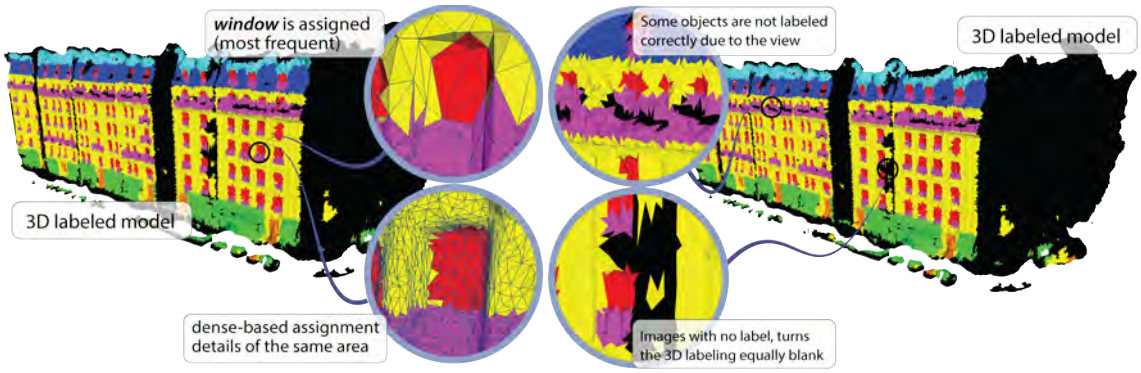
SOURCE: Author's production.

The final result of 3D labeling by ray-tracing is then acquired (Figure 3.6). Regions that have not been segmented (such as facades not perpendicular to acquisition - adjacent streets) are also treated as such during the process (dark regions in the figure - class background). Other regions, although they have been labeled as background, they were a consequence of the acquisition and angle of incidence of the rays.

3.4 Experiments

Aligning with the hypotheses of this study, in this section we present details about the analysis strategy and the validation metrics adopted throughout the tests. The experiments were elaborated as much as possible on the conditions that were also imposed on the objectives.

Figure 3.6 - Ray-tracing analysis: Details of the final ray-tracing result. (Center) The most common labels from 2D images are given to the geometric feature, which are not always correctly label due to the acquisition view point.



SOURCE: Author's production.

3.4.1 Strategy of analysis

The experiments in this study were divided in 2D and 3D domains. To mitigate the influences of each dataset, or the influences of the model under each specific architectural style, it was split the 2D experiments in three different CNN trainings. First, all the six online datasets listed in Table 3.2 were trained and inferred independently. Second, each knowledge reached from the respective datasets is used to test under SJC, which has a complete undefined architectural style. Third, all datasets were then put together and a new training was made under SJC (see Table 3.3). The term "independent" means the inference was made using only one dataset for training or prediction. In 3D analysis, the experiments consist of permuting the point cloud density, allowing us to know how the number of points that affects the 3D labeling, and how many is actually necessary to acquire reliable geometry.

3.4.2 Evaluation

Only two validation metrics were adopted to measure the quality of the predictions, accuracy and F1-score. In addition, this section also explains the objective function defined to the neural model, Cross-Entropy.

Table 3.3 - Experiments performed in this study.

#	Domain	Dataset	Inferences	Goal
1	2D	Independent	Independent	Evaluate the performance of the neural model according to each dataset
2	2D	Independent	SJC	Evaluate the performance of the neural model according to SJC dataset, where the inferences are made six times, using each dataset knowledge separately
3	2D	All-together	SJC	Evaluate the performance of the neural model according to the SJC dataset, where only one inference is made, using all-together knowledge
4	3D	RueMonge2014	-	Evaluate how accurate the 3D labeling is according to the point cloud density (sparse and dense), under a known dataset
5	3D	SJC	-	Evaluate how accurate the 3D labeling is according to the point cloud density (sparse and dense), under an unknown dataset

SOURCE: Author's production.

3.4.2.1 Accuracy

The accuracy is calculated according to a pixelwise analysis, in which the success and error rates are measured through values of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN), then, placed in a confusion matrix[1] M (more details in Annex B). Finally, the accuracy is given by:

$$Accuracy = \frac{\sum_{i=1,j=i}^{n} M_{ij}}{N} \ , \tag{3.12}$$

where n, the number of classes, and N, the number of samples used. The numerator consists in the sum of all elements in diagonal.

3.4.2.2 Objective function

An optimization problem seeks to minimize a loss-function. The weight-loss consists of the error levels in which the neural network has according to an ideal of prediction that is given by the reference images. This ideal is an optimization problem which aims to be minimal, called then loss-function. In this case, the loss-function is given by the cross-entropy (LONG et al., 2015), a common and effective way to calculate the distance between multiple predicted and ground-truth classes, also known as

[1]Considering a matrix which ground truth is in x and predicted in y, the TP is given by elements in diagonal, TN, sum of all elements except the diagonal, FP, the sum of elements in x minus diagonal and, FN, the sum of elements in y minus diagonal.

multinominal logistic classification, given by:

$$Loss_{y'}(y) = -\sum_{i} y'_i \ \log\ (y_i)\ , \tag{3.13}$$

where y_i is the predicted probability value for class i, and y'_i is the ground-truth probability for that class (more about cross-entropy calculation and meaning can be found in Annex A).

3.4.2.3 F1-Score

F1-Score expresses the harmonic mean of precision and recall. These are calculated values to understand how aligned the prediction is in relation to the reference object. F1-score is given by:

$$F1 = 2 * \frac{precision * recall}{precision + recall}\ , \tag{3.14}$$

where $precision = TP/(TP+FP)$, and $recall = TP/(TP+FN)$. Then, both values reveal how good the segmentation was according to the correct object delineation (more details in Annex B).

4 RESULTS AND DISCUSSION

4.1 Image segmentation

4.1.1 CNN inputs

As mentioned in Section 3.3.1.1, 20% of annotated images from each dataset were used to evaluate the model, the other 80%, for training. The set consists of pairs of original and ground-truth images, which are not used during the training. The experiments carried out in this study were done individually – where, the quality of the segmentation through the use of CNN for each dataset (following the sequence according to Table 3.2) was firstly discussed, followed by the impressions on the detection of objects, and in which situations it might have fail or still need attention. Then, the following section present an analysis of the geometry extraction and the quality of the 3D labeled model.

4.1.2 CNN performance

As a supervised methodology, the DL requires reference images[1]. Which means the methodology is extensible for images of any kind, but it will always require their respective reference. On the other hand, the same neural model could fit to any other detection issue, for instance, in the segmentation of specific tree species in a vast forest image, as soon as a sufficient amount of training samples are presented.

All the source-code regarding DL procedures was prepared to support GPU processing. Unfortunately, the server used during all the experiments was not equipped with such technology, increasing training time significantly (Table 4.1).

The DL source-code was mainly developed under the Tensorflow™ library[2] and adjusted to the problem together with other Python libraries. Except for the 3D tasks in Agisoft® Photoscan™, the source-code are freely available in a public platform, and can be easily extended. For training and inferences, it was used an Intel® Xeon® CPU E5-2630 v3 @ 2.40GHz. For SfM/MVS and 3D labeling, respect to RueMonge2014 and SJC datasets, it was used an Intel® Core™ i7-2600 CPU @ 3.40GHz. Both attended our expectations, but it is strongly recommended machines with GPU support or alternatives such as IaaS (Infrastructure as a Service).

In the Figure 4.1 is shown the neural network training results. Each line represent

[1]Referenced also as annotation, labels or ground-truth images.
[2]Available at https://www.tensorflow.org/. Accessed November 2, 2018.

Table 4.1 - Training attributes and performance. In bold, values which have reached the lowest performance. RueMonge2014 and SJC have larger dimensions and demanded a bit more time processing.

Dataset	Num. iterations	Resolution (pixels)	Training (hours)	Inference (secs per image)
RueMonge2014	50k	800x1067	**172.46**	5.4
CMP	50k	550x1024	135.25	4.45
eTRIMS	50k	500x780	83.57	3.52
ENPC	50k	570x720	53.47	2.01
ECP	50k	400x640	38.32	3.13
Graz	50k	450x370	29.27	2.05
SJC	-	1037x691	-	**6.2**

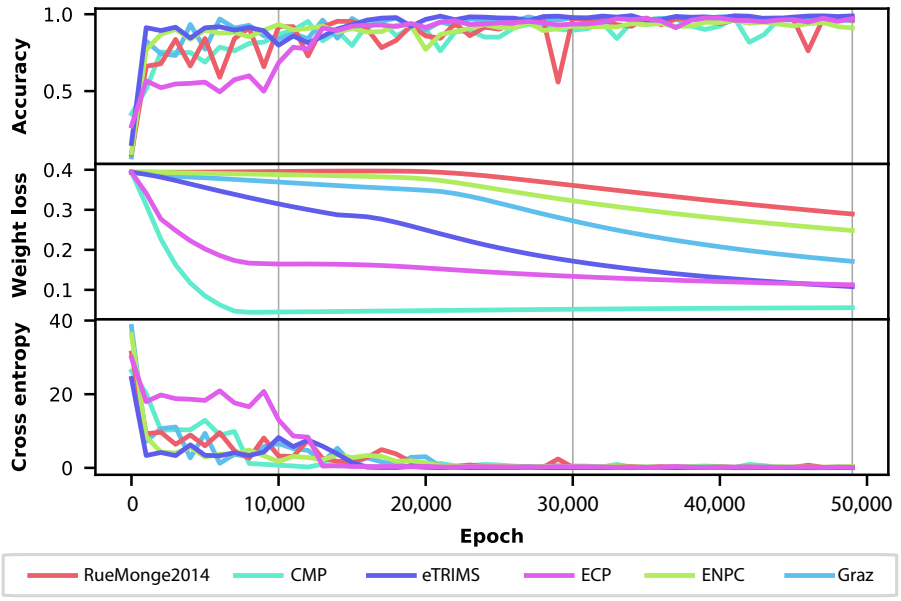
SOURCE: Author's production.

an online dataset, conducted in individual training processes, with 50 thousand (50k) iterations each. Each graph's row represents the metric used to analyze the CNN performance: accuracy, weight-loss and cross-entropy. The accuracy allows to measure how good the segmentation is according to the training progress. While the weight-loss is the CNN's error rate against reference images, and cross-entropy, the objective function defined.

The evaluation is performed repeatedly, according to a certain number of iterations, where partial inferences (prediction) is obtained and the measures are calculated by comparing the result against the ground-truth (references). These metrics are commonly adopted in the literature about classification using neural models.

A similar behavior for all training dataset is observed, except for the weight-loss. The weight-loss decay is strongly related to the image dimension, which of course requires more iterations to learn the features. The demand for the learning of all features (generalization) is greater and varies among them. Accuracy and cross-entropy, on the other hand, had progressed mostly from 0 to 10k iterations, stabilizing near 90% and 0.1 thereafter, respectively. 30k iterations were sufficient to reach similar results for all datasets (as shown later in the visual inspection). However, RueMonge2014, ENPC, and Graz still had high error rates, which means that not all classes could be detected or clearly delineated, even with 50k iterations.

Figure 4.1 - Training performance for all online datasets.



SOURCE: Author's production.

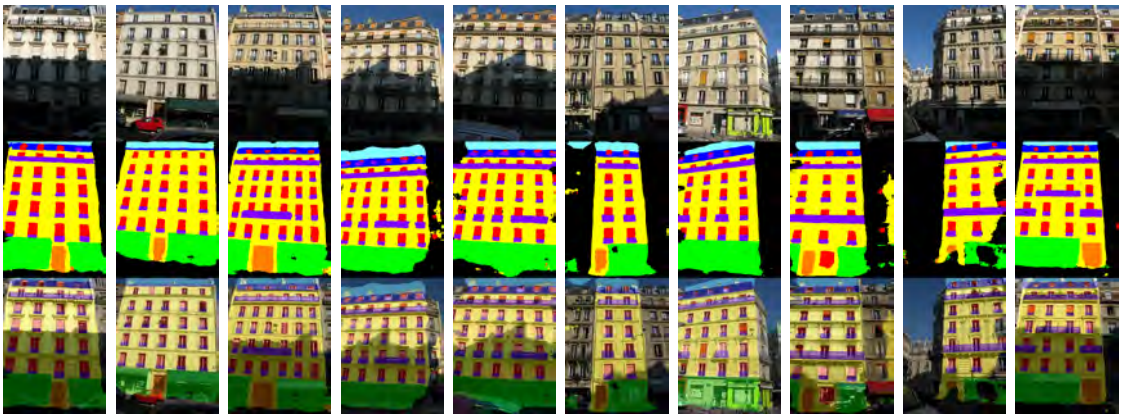
4.1.3 Inference over the online datasets (Experiment 1 - Table 3.3)

In this section, each of the online facade sets will be evaluated in detail, according to the Experiment 1 in Table 3.3. In this evaluation was adopted the same metric used during the training (accuracy) with the addition of the F1-score, also described in Section 3.4.2.

4.1.3.1 RueMonge2014 dataset classification results

The Figure 4.2 shows the inferences from RueMonge2014 over the validation set. Instead of showing only a few example results, they were exposed as much as possible to allow the reader to better understand how the neural model behaves according to different situations. Here, it is positively highlighted two aspects. First, the robustness of the neural model in the detection of facade features even under shadow or occluded areas, such as in the presence of pedestrians or cars. This aspect has been one of the most difficult issue to overcome due to the respective obstacles being dynamic and difficult to deal with, especially by the use of pixelwise segmenters. The second aspect is that at 50 thousand (k) iterations, all images presented fine class delineation. Only in a few situations the inferences were not satisfactory.

Figure 4.2 - Results over RueMonge2014 dataset. The rows are splited respectively in original, segmented image, and both. These segmented images are the inferences under evaluation sets only. (a)–(j) Example of RueMonge2014 images, segmented by the neural model presented in Section 3.3.1.2. In the first line, the original image, the second line, the result of the inference (segmentation), and the third and last line, the overlapping images.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

SOURCE: Author's production.

During manual annotation production for RueMonge2014, the labels did not cover the entire scene. For example, sky, street intersections or background buildings (adjacent to the main facades), were partially annotated as background, sometimes, completely. This means that when presented to the CNN during training, all those features (sky, street intersection, etc.), are going to be trained as background as well. Therefore, whenever an intersection or sky appears, the neural model treats it as being background. The problem is that only half of the feature will be assigned as background, which is not the case with the other half. The same behavior was visible in other classes. For instance, when an facade is fully annotated and appears only partially in the validation set, the neural model will act as it was not presented in the image, only part of it (clear on Figures 4.3(f), 4.3(h), and 4.3(i)). Supervised neural model is strongly related to the context it has been trained. If a feature appears in the image, but only part it is detected, the segmentation will fail because of the incomplete context.

The confusion matrix over all validation set is presented in Table 4.2. Pretty close to the visual inspection, the RueMonge2014 confusion matrix shows prediction over

82% for all classes, with a tiny confusion between background and wall against the other classes (first and forth columns). It is explained by the fact that the respective classes are the overall classes in the entire prediction, which means it shares boundaries with all other classes. As the object delineation is not always clear, the evaluated pixels on the edge is predicted as been background or wall instead the right class. Although it is minimal (0.8624 as F1-score), the prediction over edges seems to be the bottleneck of the adopted CNN architecture for all online datasets, as shown in the following sections. The labels in the bottom of the table, express the classes presented in the respective dataset.

Table 4.2 - Normalized confusion matrix for RueMonge2014 predictions.

	Classes	Predicted 1	2	3	4	5	6	7	8	Scale	Evaluation
Ground-Truth 1	Background	0.842	0.004	0.008	0.079	0.014	0.015		0.035	Success (1.0)	
2	Roof	0.056	0.826	0.058	0.021	0.015	0.025				
3	Sky	0.061	0.013	0.924	0.002						
4	Wall	0.042			0.919	0.010	0.015	0.005	0.007		
5	Balcony	0.044	0.011		0.039	0.892	0.013				
6	Window	0.031	0.009		0.072	0.023	0.863				
7	Door	0.047			0.073			0.875	0.005		
8	Shop	0.111			0.014				0.873	Error (0.0)	
	Rates:									**Accuracy**:	**0.9563**
										F1-Score:	**0.8624**

SOURCE: Author's production.

4.1.3.2 CMP dataset classification results

The CMP annotation set present labels beyond those already analyzed in this study. In the original CMP dataset, for example, there are annotations for decoration and pillars, which are sub-elements of the class wall and, for that reason, out of the scope in this study. Even so, still considered as being a single class: wall. For other sub-elements, all the labels not related to this study were equally ignored and had their annotations adapted to the problem.

The results presented in Figure 4.3, represent the automatic segmentation reached at 50k iterations of training. Once the architectural style has straight lines and facades whose texture is homogeneous, as presented in the CMP, the results of segmentation

Figure 4.3 - Results over CMP dataset. (a)–(g) Example of CMP images, segmented by the neural model presented in Section 3.3.1.2. The three different rows correspond to the same description as in Figure 4.2.

(a) (b) (c) (d) (e) (f) (g)

SOURCE: Author's production.

tend to be better precise due to the disturbance factors are minimal. CMP does not have annotations for background, sky, roof, and shop, and for these classes, therefore, the prediction is replaced by the labels wall, window, balcony or door, for example, in Figure 4.4(d), there are stores predicted as wall. As good results, the Figures 4.4(a), 4.4(c), 4.4(e) and 4.4(g) highlight the predictions that did not get much confusion. Facades that share different texture, as in Figure 4.4(f), presented a lot of confusion between the classes balcony and wall, specially under building decoration regions.

Among the four classes, only wall and window got good scores (TP), reaching up accuracy of 0.9357 for the entire prediction. Balcony and door were often confused with wall, specially around the boundaries of the classes, what also explains the F1-score of 0.7418.

4.1.3.3 eTRIMS dataset classification results

As in CMP, eTRIMS had more annotations than the ones adopted in this study[3]. In addition to the classes not approached in this study, there were facade features where the annotation belonged to only one class, e.g. the roof in eTRIMS is annotated as being wall. For that reason, images with roof had it assigned as wall

[3]The annotations provide by online datasets might vary on number of labels and colors. If there are more labels (or classes) than is needed, then, it is necessary to adapt to the same colors in Figure 3.2 or reduce the number of classes, such as in eTRIMS.

Table 4.3 - Normalized confusion matrix for CMP predictions.

	Classes	Predicted 1	2	3	4	5	6	7	8	Scale	Evaluation
Ground-Truth	1 Background									Success (1.0)	
	2 Roof										
	3 Sky										
	4 Wall				0.937	0.022	0.022	0.019			
	5 Balcony				0.390	0.508	0.085	0.017			
	6 Window				0.109	0.013	0.825	0.052			
	7 Door				0.268		0.037	0.694			
	8 Shop									Error (0.0)	
	Rates:									**Accuracy**:	**0.9357**
										F1-Score:	**0.7418**

SOURCE: Author's production.

and, consequently, assumed as True Positive (TP). Among the 6 facade features of interest, only three in eTRIMS were considered: window, door and wall.

As can be seen in Figure 4.4, the predictions over eTRIMS dataset reached the second best score among the online datasets, good accuracy (object location), 0.9632, and F1-score (object delineation), 0.8291. The eTRIMS consist of facade pictures of different styles, "non-patterned", non-rectified.

Figure 4.4 - Results over eTRIMS dataset. (a)–(h) Example of eTRIMS images, segmented by the neural model presented in Section 3.3.1.2. The three different rows correspond to the same description as in Figure 4.2.

(a) (b) (c) (d) (e) (f) (g) (h)

SOURCE: Author's production.

Considering a system of mapping of walls, windows and doors, the predictions made on this dataset show the ability of the neural network to detect the urban elements accurately (Figure 4.5(b)). The eTRIMS was distinguished compare to others because fewer classes were used (neural network learns more), also, the quality of annotations is better. In any case, objects obstructing the facade are ignored, precisely because training data consider these objects as being another class. Here, considered as background and therefore not part of the facade. This lack of information, for example in Figures 4.5(c), 4.5(e) and 4.5(f), is the best scenario in the detection of features, still, it will need an algorithm to regularize these missing informations.

The confusion matrix of eTRIMS shows a tiny confusion between window, door and wall, with only 20% of pixels . Once again, the FN rates regarding to this prediction are mainly related to the boundaries between one class and another.

Table 4.4 - Normalized confusion matrix for eTRIMS predictions.

		Classes	Predicted								Scale	Evaluation
			1	2	3	4	5	6	7	8		
Ground-Truth	1	Background	0.956			0.037		0.006			Success (1.0) / Error (0.0)	
	2	Roof										
	3	Sky										
	4	Wall	0.024			0.947		0.020	0.009	0.007		
	5	Balcony										
	6	Window				0.206		0.773	0.017			
	7	Door	0.049			0.204		0.078	0.669			
	8	Shop										
		Rates:									**Accuracy**:	**0.9632**
											F1-Score:	**0.8291**

SOURCE: Author's production.

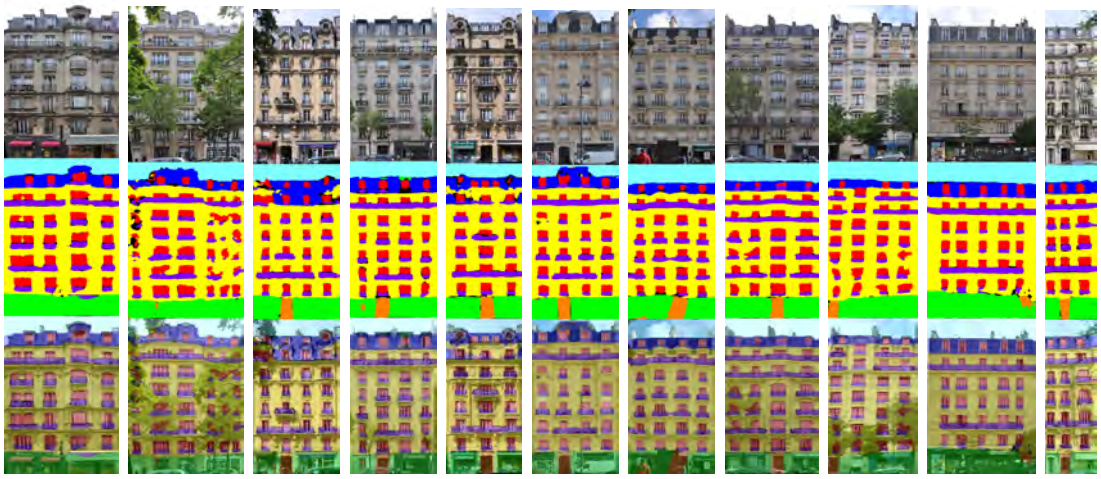
4.1.3.4 ENPC dataset classification results

The level of accuracy for all datasets made the use of CNN the best of all alternatives. However, the quality of the annotations is essential to perform a good segmentation. As noticed in eTRIMS, when a tree or car obstruct the facade, they also add disturbances in the training phase. In case of a tree, it could be annotated as either vegetation or part of the facade itself. For example, note the differences between Figures 4.5(b) and 4.6(b). The tree in eTRIMS is ignored in their annotation, but labeled as being facade in ENPC, which leads to different results.

It is understood that the lack of information in the first figure is the best inference,

and in this case, the neural model is actually right: there is a facade with unknown object in front of it. But in cases such as in Figure 4.6(b), the facade inference is noisy or unreadable, which is not the case in Figure 4.6(h), where the disturbance is minimal.

Figure 4.5 - Results over ENPC dataset. (a)–(k) Example of ENPC images, segmented by the neural model presented in Section 3.3.1.2. The three different rows correspond to the same description as in Figure 4.2.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k)

SOURCE: Author's production.

Table 4.5 - Normalized confusion matrix for ENPC predictions.

		Classes	Predicted 1	2	3	4	5	6	7	8	Scale	Evaluation
Ground-Truth	1	Background	0.105	0.060	0.033	0.315	0.235	0.190	0.013	0.049	Success (1.0) / Error (0.0)	
	2	Roof	0.041	0.783	0.055	0.072	0.015	0.034				
	3	Sky	0.006	0.015	0.977							
	4	Wall	0.011	0.007		0.916	0.028	0.034		0.004		
	5	Balcony	0.034	0.004		0.125	0.794	0.042				
	6	Window	0.020	0.023		0.113	0.023	0.817				
	7	Door	0.022			0.033			0718	0.227		
	8	Shop	0.007			0.038	0.004		0.017	0.935		
		Rates:									**Accuracy**:	**0.9636**
											F1-Score:	**0.7655**

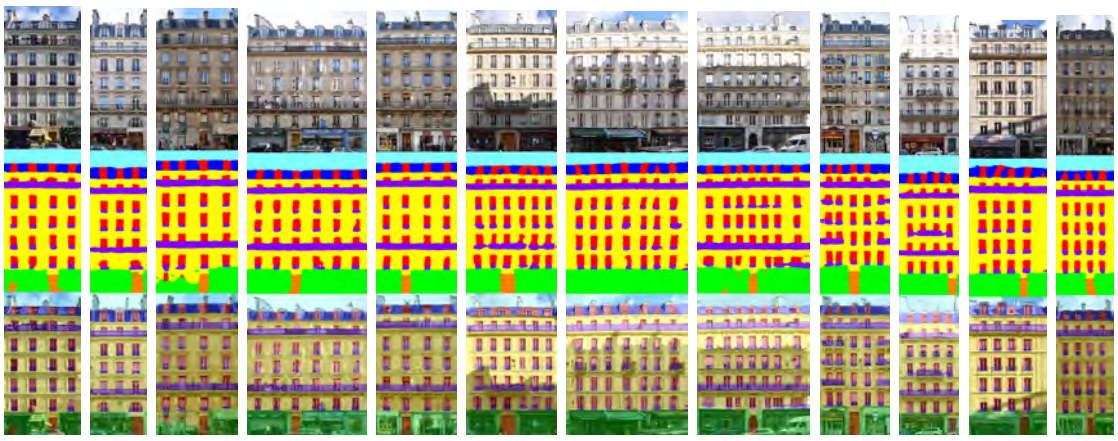
SOURCE: Author's production.

The confusion matrix for ENPC predictions, revealed that all classes has been assigned correctly (accuracy 0.9636 and F1-score 0.7655), except for background. Once this dataset only have rectified images (facade covering the entire image), the class background actually should not exist. Then, the respective class have not so much presence, when predicted as such, it was incorrect. For example, in Figures 4.6(b) and 4.6(c), is visible that the pixels in black actually corresponds to the classes wall (yellow) and roof (blue), respectively. The high accuracy, however, is explained by the fact all classes have been assigned correctly in general, getting 7% of errors due to the classes boundaries.

4.1.3.5 ECP dataset classification results

All online datasets does not have any certificate of quality. When checking the annotated images of some of them, there is a high degree of inconsistency between the annotations. This implies incorrect segmentation (see overlapping images - detail on the roofs) according to the real scenario – visual inspection, but not to the validation set. It means the validation metrics might present some inconsistency, since they are calculated according to the validation (annotated) images.

Figure 4.6 - Results over ECP dataset. (a)–(l) Example of ECP images, segmented by the neural model presented in Section 3.3.1.2. The three different rows correspond to the same description as in Figure 4.2.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l)

SOURCE: Author's production.

ECP also presented inconsistencies in some of its annotations. The missing roof-parts in Figures 4.7(a) to 4.7(f) are expected behaviors since the annotations from the training sets do not consider these objects as being part of the roof. However, the learning happens for most of the features and should not be a problem since the neural model will identify the main content in the image.

According to the accuracy and F1-Score metrics, ECP got the best scores among the others, with 0.9762 and 0.8946, respectively. The lower TP was 0.841 for class window, which is already considered a good mark since the other classes got better scores. A tiny confusion between balcony, window, and wall, is highlighted, as well as for door and shop.

Table 4.6 - Normalized confusion matrix for ECP predictions.

		Classes	Predicted								Scale	Evaluation
			1	2	3	4	5	6	7	8		
Ground-Truth	1	Background									Success (1.0) / Error (0.0)	
	2	Roof		0.844	0.043	0.030		0.083				
	3	Sky		0.009	0.986							
	4	Wall			0.390	0.938	0.030	0.025				
	5	Balcony				0.109	0.854	0.032		0.004		
	6	Window		0.030		0.081	0.041	0.841				
	7	Door				0.014			0.873	0.113		
	8	Shop			0.006	0.023			0.013	0.955		
		Rates:									**Accuracy**:	**0.9762**
											F1-Score:	**0.8946**

SOURCE: Author's production.

4.1.3.6 Graz dataset classification results

Among the inputs, Graz has the smallest number of images (50), but the spectral variability is clearly greater when compared to the others. The symmetry between windows, however, was pretty much the same in CMP, ECP, and ENPC. It is noticed, then, that the results for Graz (Figure 4.7) did not change much from what was seen in the other datasets.

Even the predictions has reached accuracy of 0.9368 and F1-score of 0.7698, a lot of confusion involving the class wall can be notice in Table 4.7, specially for balcony and sky, where the predictions were almost null.

Figure 4.7 - Results over Graz dataset. (a)–(h) Example of Graz images, segmented by the neural model presented in Section 3.3.1.2. The three different rows correspond to the same description as in Figure 4.2.

(a) (b) (c) (d) (e) (f) (g) (h)

SOURCE: Author's production.

Table 4.7 - Normalized confusion matrix for Graz predictions.

		Classes	Predicted								Scale	Evaluation
			1	2	3	4	5	6	7	8		
Ground-Truth	1	Background				0.846		0.013	0.029	0.112	Success (1.0)	
	2	Roof		0.039	0.171	0.733		0.056				
	3	Sky	0.059	0.004	0.757	0.171		0.005				
	4	Wall				0.941		0.040		0.012		
	5	Balcony				0.836	0.051	0.102		0.008		
	6	Window				0.169		0.816		0.009		
	7	Door				0.242		0.045	0.661	0.052		
	8	Shop				0.624		0.068	0.056	0.245	Error (0.0)	
		Rates:									**Accuracy**:	**0.9368**
											F1-Score:	**0.7698**

SOURCE: Author's production.

In Table 4.8, it is summarized the accuracies and F1-scores for each online dataset. ECP presented the best results among the datasets, not only the accuracy, but also in precision (F1-score). Datasets where the accuracy is high but F1-score inferior, demonstrated an excellent inference in which region the object was found on the image, but not equally efficient regarding its delineation, that are the cases of CMP, ENPC and Graz. The predictions on RueMonge2014 and eTRIMS datasets presented better quality in both metrics.

As the evaluation was performed under multiple validation images, the columns in red and green represent the variance and standard deviation, respectively. None of them have reached significant variations.

Table 4.8 - Inference accuracy over the online datasets. Var. (Variance = σ^2) and StD. (Standard Deviation = σ) stands for the inferences over different images. The values in bold, expose the best datasets according to the Accuracy and F1-Score metrics.

Dataset	**Accuracy**	**Var.**	**StD.**	**F1-score**	**Var.**	**StD.**
RueMonge2014	0.9563	0.008	0.090	0.8624	0.000	0.027
CMP	0.9357	0.005	0.073	0.7418	0.001	0.043
eTRIMS	0.9632	0.000	0.027	0.8291	0.000	0.017
ENPC	0.9636	0.001	0.031	0.7655	0.000	0.009
ECP	**0.9762**	0.000	0.021	**0.8946**	0.000	0.014
Graz	0.9368	0.014	0.117	0.7698	0.000	0.023

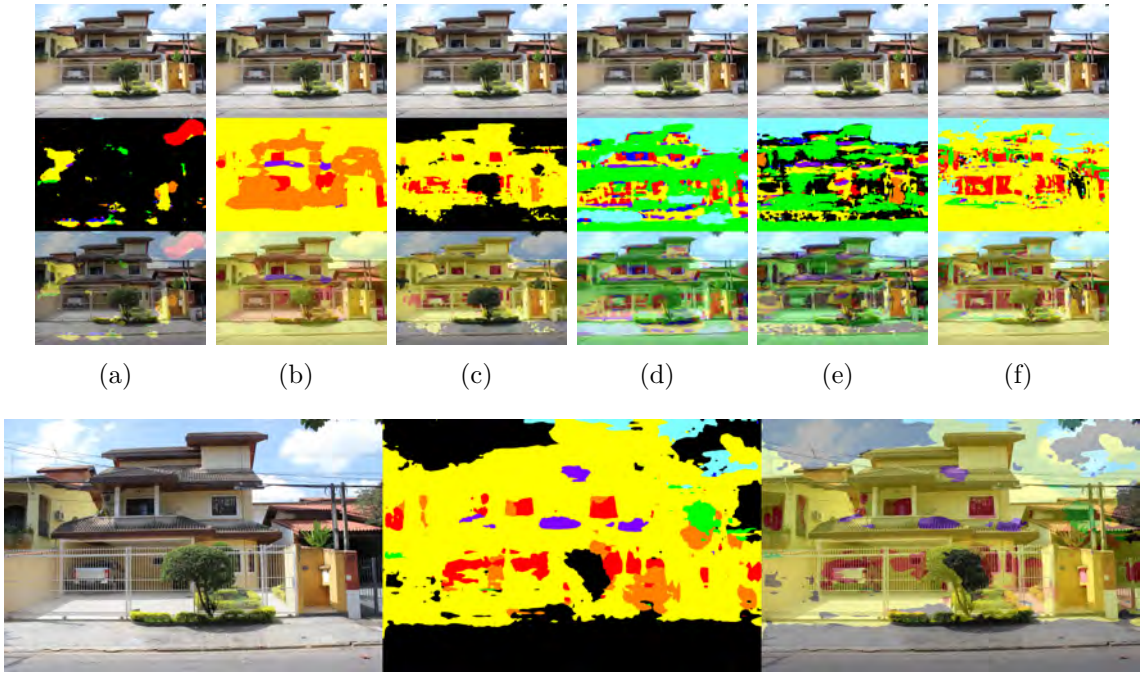
SOURCE: Author's production.

4.1.4 Inference over the SJC dataset (Experiments 2 and 3 - Table 3.3)

The idea behind the use of SJC dataset was simple: to observe how the neural network reacts to an unknown architectural style after being trained with different ones. The outcome could then provide insights on how the training set should look like for the detection of facades of any kind. Figure 4.8 shows the results after presenting SJC images to a different version of training data (knowledge).

The Figures 4.9(a) to 4.9(f) are the respective results from datasets listed in Table 3.2 (online). When looking at these results as seen in the figures, it is safely concluded that these are incorrect and inaccurate segmentations (accuracies mostly lower than 24%). The fact is that in environments whose diversity of objects, any

Figure 4.8 - Segmented image from SJC dataset. Inferences between individual training knowledges. (a) Result using RueMonge2014 knowledge, (b) CMP, (c) eTRIMS, (d) ECP, (e) ENPC, (f) and Graz. (g) result using all knowledge.



(a) (b) (c) (d) (e) (f)

(g)

SOURCE: Author's production.

other segmentation and classification methods would add a certain imprecision on it. The operation of a CNN is not to perfectly delineate an object, but to provide hints (as close as possible) to where a given object is located in the image. This shows us that in order to extract precise parameters, such as height and area of a feature, a post-processing phase should certainly be conducted on the CNN beforehand.

Going through one problem at a time, it is noticed that, firstly, there is the need to define a background class in supervised approaches. Using RueMonge2014 knowledge, the inference process was not able to segment properly even under the most common feature: wall. Unlikely, some knowledge generalized it to sky, sidewalk and street, especially when there is no general class that represents too many objects in the scene (Figure 4.9(b)). When the class is annotated correctly such as sky, a proper segmentation can be seen: that is the case with ECP (Figure 4.9(d)), ENPC (Figure

4.9(e)), and Graz (Figure 4.9(f)). When it is not, the inference is poor or average: which was the case with RueMonge2014 (Figure 4.9(a)). Features in RueMonge2014 were pretty much dependent on local architectural style. In general, eTRIMS is the dataset with the most similar features for SJC. Despite being trained with only 4 classes, including background, the results have shown a certain level of intelligence in detecting sidewalks and street as being part of the background, as well as for sky and vegetation.

Therefore, when using a supervised neural network, it is evident that the arrangement of the annotations can affect the inference, either positively or negatively. For instance, annotations for all the sky coverage instead of only part of it, or background annotations for sidewalks and street, in cases that it is not a desired feature. The confusion matrix of all-together knowledge applied in SJC data is shown in Table 4.9.

The confusion can be mainly described between the classes roof, sky, balcony and door. Only classes background, wall and window were reasonably assigned, most influenced by the knowledge from eTRIMS and Graz (Figures 4.9(c) and 4.9(f)).

Table 4.9 - Normalized confusion matrix for SJC (all-together) predictions.

	Classes	Predicted 1	2	3	4	5	6	7	8	Scale	Evaluation
Ground-Truth 1	Background	0.811			0.158		0.006	0.020	0.004	Success (1.0)	
2	Roof	0.035	0.002		0.870	0.042	0.014	0.009	0.028		
3	Sky	0.297	0.001	0.211	0.487				0.003		
4	Wall	0.113			0.833	0.003	0.011	0.020	0.020		
5	Balcony	0.029			0.514	0.213	0.051	0.193			
6	Window	0.027			0.176	0.009	0.683	0.106			
7	Door	0.075			0.543	0.002	0.116	0.244	0.020		
8	Shop									Error (0.0)	
	Rates:									**Accuracy**:	**0.8591**
										F1-Score:	**0.4706**

SOURCE: Author's production.

In Figure 4.9(g), a summarized contribution of each dataset is presented. For instance, eTRIMS was the only one sensitive to sidewalk and street, balconies were only detected in CMP, even though this was incorrectly segmented. Meanwhile, the

results for unknown features were understandable and expected. With the addition of more classes (e.g. gate), improvements to the annotation process and an increase in the number of training epochs, the better the results of the inferences would be. Table 4.10 shows the accuracy overview for each individual learned feature (knowledge) over the SJC dataset.

Table 4.10 - Inference accuracy over SJC data. Last row corresponds to the accuracy with the knowledge of all training together. The values in bold, expose the best datasets according to the Accuracy and F1-Score metrics. When together, the quality metrics increased due to the better generalization of the neural network, as it has received a bigger amount of images.

Knowledge from...	**Accuracy**	**Var.**	**StD.**	**F1-score**	**Var.**	**StD.**
RueMonge2014	0.7009	0.003	0.054	0.1612	0.000	0.014
CMP	0.7907	0.006	0.080	0.3763	0.001	0.043
eTRIMS	0.7726	0.001	0.037	**0.3915**	0.000	0.017
ENPC	0.8123	0.001	0.031	0.2394	0.000	0.009
ECP	**0.8610**	0.012	0.016	0.2225	0.000	0.014
Graz	0.8011	0.006	0.078	0.2669	0.000	0.023
All together	**0.8591**	0.011	0.107	**0.4706**	0.000	0.020

SOURCE: Author's production.

Both visually (Figure 4.4) and in the table, it is evident that eTRIMS was better suited to deal with the architectural style seen in the SJC dataset. It was expected however, that the values for accuracy and precision would be low, due to the characteristics (similarities) of the SJC dataset. eTRIMS consists of non-rectified facades, lacks symmetry between doors and windows and presents specific architectural styles, characteristics that have similarity with the SJC images. Once the entire online collection has been merged (all-together dataset), the accuracy was increased, but without improvements to the correct delineation of the objects (low F1-score).

4.2 3D labeling - Experiments 4 and 5 - Table 3.3

4.2.1 RueMonge2014

The quality of the reconstructed surface (mesh) is highly dependent on the density of the point cloud and the method of reconstruction. Very sparse point clouds can

generalize feature volumetry too much, while very dense point clouds can represent it faithfully, but the associated computational cost will also increase. Therefore, there is a limit between the quality of the 3D labeled model and the point cloud density, which falls on the question: how many points it is needed to fairly represent a specific feature?

Features that are segmented in 2D domain might perfectly align with their geometry, but imprecisions between the geometric edges and the classification may occur. Despite of that, the segmentation alignment onto the mesh is also related to the estimated camera parameters, which are used during ray-tracing. These impressions are directly related to the mesh quality.

Table 4.11 shows how the ray-tracing procedure performed. It was responsible for connecting each segmented feature onto its respective geometry. The 3D reconstruction of both scene had average performance and was not critical since the density was not the highest. RueMonge2014 dense point cloud has around 9 million points, reconstructed in 46.4 minutes. Using a proper computer, equipped with GPU, this rate could certainly be optimized, as well as the other reconstructions.

Table 4.11 - Ray-tracing performance for geometry classification.

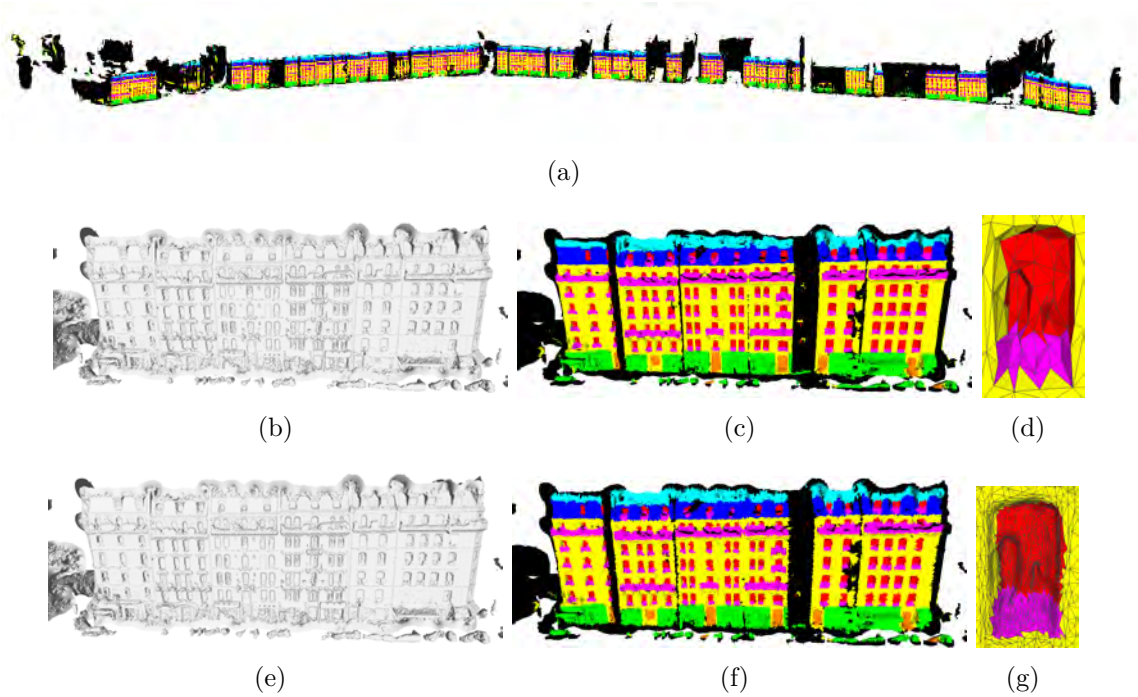
Dataset	**Point Cloud density**	**Num. faces (triangles)**	**3D reconstruction - SfM/MVS (min)**	**Ray-tracing (sec)**
RueMonge2014	Sparse	1,072,646	21.4	12.42
RueMonge2014	Dense	9,653,679	46.4	27.13
SJC	Sparse	800,000	13.6	20.89
SJC	Dense	3,058,329	35.5	41.12

SOURCE: Author's production.

In order to illustrate the influences of the point cloud density on the quality of 3D labeling, Figure 4.9 shows the result for the RueMonge2014 dataset. Only sparse and dense point clouds were tested. However, to explore the limits between the number of points and the geometric accuracy is some of the issues to approach in a future work. Also note in Figure 4.10(a) that the geometry of the whole street was reconstructed, with few deformations caused by the SfM.

In Figures 4.10(d) and 4.10(g), it is highlighted how well the point cloud density

Figure 4.9 - 3D labeled model of RueMonge2014. (a) Wide view of the street. (b) Details of facade geometry by a sparse point cloud, (c) its labels after ray-tracing analysis, and (d) close-look of 3D window labels. (e) The facade geometry by a dense point cloud and (f) its labels, and (g) close-look of 3D window labels.



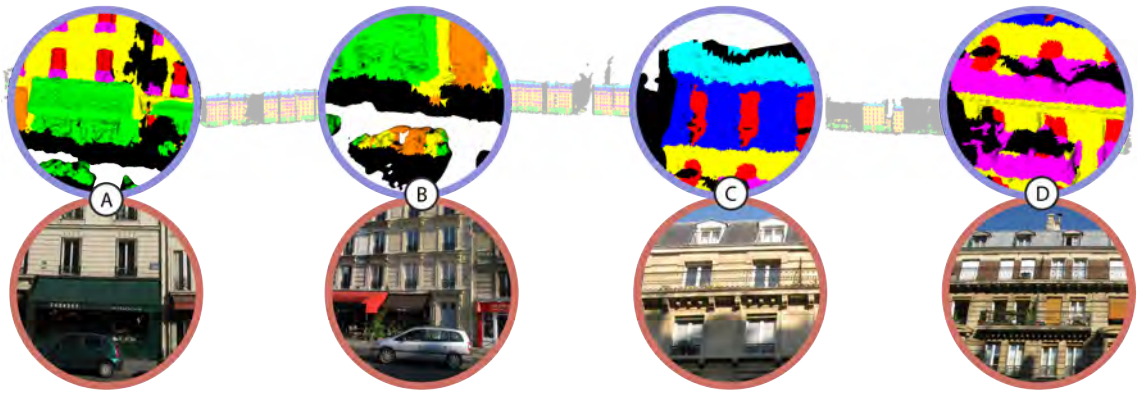
SOURCE: Author's production.

could represent a labeled 3D model. Assuming a hypothetical situation where area or window height information are required in a building inspection to estimate luminosity (indoor and outdoor), the estimation of these parameters should be as close as possible to reality. Therefore, the height and area obtained from the mesh, as in the respective figures, may be inaccurate to this kind of need. Martinovic et al. (2015) and Boulch et al. (2013) propose a post-processing procedure, in which the facade has its features simplified by the so called Parsing, where most of the time, Grammar-based approaches (STINY; GIPS, 1971) are used. Perhaps the post-processing phase is essential in applications where precise geometric information is required, but it has to ensure that the geometric accuracy does not be penalized once the Grammar can over-generalize the facade feature shape.

In the Figure 4.10, is highlighted four details about the labeling. In "A"and "B", details on the bottom region of the building, where the store and door class mainly

appear. This part of the building, as well as the top, are the critical regions during the reconstruction.

Figure 4.10 - Zoom-in details of the 3D labeled RueMonge2014 reconstruction. (A) and (B) Building's bottom-part: objects that impair the reconstruction and labeling. (C) and (D) Building's upper-part: here, the reconstruction and labeling are impaired by the view that the photos have been taken. On the bottom pictures, is shown the real details.



SOURCE: Author's production.

At the bottom, the presence of pedestrians, cars, light poles not only hinder the reconstruction, but also become part of the incorrect classification during the projection of pixels on the mesh. In detail in "A", however, it may be noted that, despite the presence of vehicles, the classification of stores (in green) is not much impaired. In the illustrations in "C" and "D", features are highlighted at the top of the building. In this case, the classification is impaired by the projection center. In some regions, as in "D"and "C", the balconies and roof are often represented either by background or partially the right class. The absence of information is a consequence to the perspective which the photos were taken, not necessarily a poor labeling. By terrestrial acquisition only, it is not possible to observe much about this structure (as seen in Figure 2.2 and 2.3, Section 2.2).

4.2.2 SJC

The 3D reconstruction performed by SfM is based on the identification of corners and image analysis, with the purpose of checking for correspondence and acquiring overlapping pairs of images. For this reason, the spectral properties from the urban elements influence the reconstruction process directly. For example, surfaces where the texture is too homogeneous or specular, these will be potential problems and will not be detected by the algorithm. Once not detected, the 3D model will end up with a rough representation, with gaps instead of mapped features. Similarly, the MVS technique (responsible for dense 3D reconstruction) is equally dependent on the homogeneity of the objects.

Unfortunately, many of these spectral properties over SJC facades can be seen. The texture related to walls are often uniform, with windows completed in glasses. Besides, the geometry of acquisition did not contribute in this case. As seen in Figures 4.12(a) to 4.12(c), all over the street, there are always gaps between the gate and the facade itself. These gaps often imposes problems during and after the reconstruction. As a consequence, a lot of them among important artifacts that could be determinant when trying to identify features in a semantic system. Hence, in order to fully map buildings through the use of SfM/MVS, the imaging of these areas, at least in Brazil, should be complemented by aerial imagery with the aim of targeting these areas (as presented and discussed in Section 2.2.2, Figure 2.3). The final 3D reconstruction, however, was moderate as the segmentation in 2D domain.

Figure 4.11 - 3D model of SJC. (a), (b), and (c) Example of gap between the gate and the facade, often present in this specific architectural style. The picture (c) represents the point of view in (b), and vice-versa.

(a) (b) (c)

SOURCE: Author's production.

The Figures 4.13(a) (overview), 4.13(c) (reconstruction from sparse point cloud), and 4.13(d) (reconstruction from dense point cloud), are the same residential building as the previous picture, whose geometry is characterized by high walls and gates. Trees and cars appear in most of the images. These objects act as obstacles, especially in terrestrial and optical campaigns. Of course, this depends solely on the imaged region. In case of RueMonge2014, for example, while pedestrians, cars and vegetation act negatively in the reconstruction, the architectural style contributes positively. This makes the final 3D model be penalized, but still, it is an acceptable product. As can be seen in Figures 4.11 and 4.12, however, not only the texture, but also the houses geometry and the frequent presence of obstructing objects, negatively affected the reconstruction. In Figure 4.13(d), a different area of very poor labeling. Although the gate has been assigned partially correct, the features here are mostly unreadable. Two blocks of SJC were imaged, each with approximately 100 meters. The degree of overlap between the images was about 86%, allowing a detailed reconstruction of the frontal face of the houses.

Figure 4.12 - 3D labeled model of SJC. On top, a wide view of the street. (a) Same view of Figure 4.12(b), after 3D labeling procedure. (b) Region with spurious labeling - most of the 3D street model was spurious due to the image segmentation quality. (c) Close-look at features details reconstructed by sparse point cloud. (d) Example of the same area using a dense point cloud reconstruction. The labeling legend can be seen in Figure 3.2.

(a) (b)

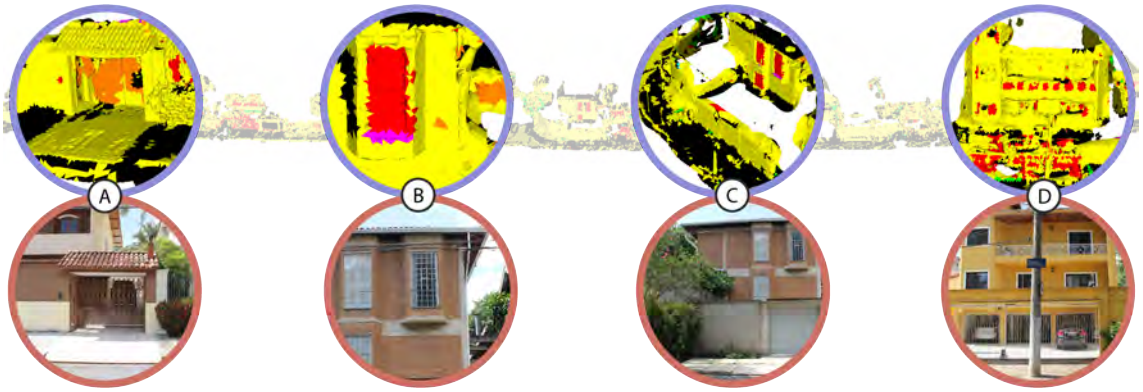
(c) (d)

SOURCE: Author's production.

In Figure 4.13, some peculiarities are highlighted in the reconstruction and classification of the SJC dataset. In the illustration, the letter "A"highlights details of the roof geometry (high resolution reconstruction - dense cloud), however, the classification of the roof is labeled as a wall (in yellow), and partially correct in the region of the gate (in orange). In "B", as previously highlighted, windows with features very similar to the online datasets were correctly identified. However, there were very few occurrences. The architectural style of SJC has its own style and very different from the styles used during CNN training (In "C", details of the gap between the gate

and the wall itself). Even so, it is concluded that it is possible to identify features of interest using generic training data. The highlight in D, is shown the poor classification in regions whose gates with "grid" aspects, confused the prediction with the window feature.

Figure 4.13 - Zoom-in details of the 3D labeled SJC reconstruction. (A) Detail of the gate partially segmented. (B) Few windows and balconies were detected, this is an example of a properly detected feature. (C) Gaps between the gate and house (facade). (D) Confusion under gates with "grid" characteristics. On the bottom pictures, is shown the real details.



SOURCE: Author's production.

5 CONCLUSION

5.1 General conclusions

Increasingly, the research regarding the facade feature extraction from complex structures, under a dynamic and hard-to-work environment (crowded cities) represent a new branch of research, with perspectives to the areas of technology, such as the concept of smart cities, as well as the areas of Cartography, toward to more detailed maps and semantized systems. In this study, an overview of the most common techniques was presented, as well as an introduction of instruments and ways of observing structural information through remote sensing data. Besides, was also presented a methodology to detect facade features by the use of a CNN, incorporating this detection to its respective geometry through the application of a SfM pipeline and ray-tracing analysis.

The experiments are mainly focused in aspects such those aforementioned techniques and their computational capability in detecting facade features, regardless of architectural style, location, scale, orientation or color variation. All the images used in the training procedures underwent no preprocessing whatsoever, keeping the study area as close as possible as to what would be a common-user dataset (photos taken from the streets).

In this sense, the edges of the acquired delineated features show the robustness of the CNN technique in segmenting any kind of material, in any level of brightness (shadow and occluded areas), orientation, or presence of pedestrians and cars. Considering that the values achieved for the individual datasets were above 90%, it is concluded that the CNNs can provide good results for image segmentation in many situations. However, being a supervised architecture, the network has to pass through a huge training set, with no guarantees of good inputs, in order to get reliable inferences. When applied over unknown data, such as the experiment on the SJC data, it was noted that the neural network failed, except in regions where the facade features share similar characteristics, though such occasions were rare.

5.1.1 Brazilian architectural styles and unreachable areas

In Section 2.4, it is shown some of the difficulties encountered specifically in Brazil. In addition, in Chapter 4, experiments were carried out on the extraction and reconstruction of features on Brazilian facades. As mentioned before, factors such as local culture, number of inhabitants and economy are determinant in local geometry

issues, especially in poor regions whose civil construction is often made irregularly, over equally irregular and difficult access areas. Not only that, but the factors can also determine, for example, the geometry of city-centers, such as the presence of skyscrapers (e.g. São Paulo, Brazil; Changhai, China; and New York, United States) or commercial areas with buildings with a maximum of four floors (e.g. London, Germany, and France). Thus, studies involving the measurement of these factors could contribute as input for the development of 3D reconstruction techniques as presented in this document.

5.2 Source-code: Usability, Licenses and Extension

References to the codes and experiments carried out in this study can be found in the Table A.1, in Appendix A. Since the beginning of this work, libraries such as CGAL, PCL, VisualSfM, COLMAP, OpenCV, among others, have been used for tests and experiments throughout development. Of course, some of these libraries and software were discarded despite having a key role in refining the final methodology. Therefore, listing the technologies used, their sources, and other characteristics, is a way of guiding other works and contributing with related researches.

5.2.1 Difficulties

This work was entirely carried out with the aid of Open-Source tools, for the most part, under MIT and GPL licenses[1], which includes the non-commercial extension and free-to-use code. In the table A.1, in the Appendix A, the tools used and their relation with the respective methodological step are presented.

In addition to the results archived in this study, every implementation was carried out in order to be expandable, interoperable, understandable, cohesive, public and easily accessible. Therefore, all methodological steps are available for consultation and public use at the GitHub™ (CHARLES, 2013), where the links and respective explanations are available in Table A.2. All the datasets used for this research were provided with any cost. In the same way, private tools that were essential for the development of the research were difficult to access or too expensive to acquire, most of the technologies used in this study were free, for example, the tools mentioned in Section 5.2.

Due to the technology approached in this study covering the state-of-the-art of classification and urban mapping solutions, the supply of training and courses is scarce.

[1]Licenses of a public domain that could be extended and, in case of GPL, must be shared.

Likewise, in Brazil, there was a great difficulty in learning how to process LiDAR data during the first phases of this study. As we progressed, the operations involving SfM/MVS became more consistent, mainly, with the involvement and cooperation of multiple institutions (academic, military - national and international). For example, the DSG, in the revision of text regarding Brazilian standards and specifications for 3D geographic mapping in Brazil. The Department of Photogrammetry (IfP), at the University of Stuttgart, Germany, by offering disciplines of Computer Vision and Pattern Recognition. The Federal University of Paraná (UFPR), by their methodic assistance regarding the use of LiDAR in urban areas, among others.

5.3 Future prospects

Identifying facade features under a great variety and arrangement make up these tasks still a scientific challenge whose tendency is to expand. The technologies to observe cities, such as sophisticated sensors, reconstruction and classification techniques, evolve as the numerous architectural styles change according to local culture and way of life. Moreover, it is essential to think that the multiplicity of architectural styles is not the only problem. Studies, such as those carried out at the MIT Center for Art, Science and Technology (CAST), Massachusetts Institute of Technology (MIT)[2] and at Eidgenössische Technische Hochschule (ETH) Zürich (ADRIAENSSENS et al., 2016), show, for example, that materials used in construction might become dynamic and therefore do not present a single static structure of a building. Urban occupation tends to evolve, which also demands that mapping techniques must both to follow the current architectural structures, as well as their eminent evolution.

As future prospects, to explore aspects such as the use of non-supervised models, separate tasks such as pre-classification of architectural styles, and mix different DL techniques to deal with specific scenarios, such as the chaotic arrangement of urban elements. Even it is the first study case, the methodology presented is highly dependent of the quality and number of images for training. The power of generalization in a neural network occurs as soon as the training set is large enough, as well as their resolutions. Besides, once this study has shown the robustness of CNN over complicated situations, we believe that efforts directed towards post-processing techniques could make the final 3D labeled model even more accurate.

[2]Video available at https://www.youtube.com/watch?v=vRfNbhyPPKs. Accessed November 2, 2018.

REFERENCES

ABDEL-HAMID, O.; DENG, L.; YU, D. Exploring convolutional neural network structures and optimization techniques for speech recognition. In: **Proceedings...** [S.l.]: ISCA, 2013. p. 3366–3370. 30

ACCUCITIES LTD. **3D Model of London & 3D City Models**. 2017. Available from: <http://www.accucities.com>. 1

ACHANTA, R.; SHAJI, A.; SMITH, K.; LUCCHI, A.; FUA, P.; SÜSSTRUNK, S. Slic superpixels compared to state-of-the-art superpixel methods. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 34, n. 11, p. 2274–2282, 2012. 19

ADRIAENSSENS, S.; GRAMAZIO, F.; KOHLER, M.; MENGES, A.; PAULY, M. **Advances in Architectural Geometry 2016**. [S.l.]: vdf Hochschulverlag AG, 2016. 79

AGENCY, S. N. M. **swissSURFACE3D**. 2010. Available from: <https://shop.swisstopo.admin.ch/en/products/height_models/surface3d>. 1

AGOURIS, P.; DOUCETTE, P.; STEFANIDIS, A. Automation and digital photogrammetric workstations. **Manual of photogrammetry**, Bethesda, MD: American Society for Photogrammetry and Remote Sensing, p. 949–981, 2004. 25

AHMAD, A.; GADI, M. Simulation of solar radiation received by curved roof in hot-arid regions. In: EIGHTH INTERNATIONAL IBPSA CONFERENCE. **Proceedings...** [S.l.]: IBPSA, 2003. p. 11–14. 1

AMORIM, J.; VALENTE, J.; PIMENTEL, C.; MIRANDA, A.; BORREGO, C. Detailed modelling of the wind comfort in a city avenue at the pedestrian level. In: USAGE, USABILITY, AND UTILITY OF 3D CITY MODELS. **Proceedings...** EDP Sciences, 2012. p. 03008. Available from: <https://doi.org/10.1051/3u3d/201203008>. 21

ANAC. **Regulamento Brasileiro de Aviação Civil Especial - RBAC - E no. 94**. 2015. Available from: <http://www.anac.gov.br/assuntos/legislacao/legislacao-1/rbha-e-rbac/rbac/rbac-e-94-emd-00/@@display-file/arquivo_norma/RBACE94EMD00.pdf>. 23

_____. **Agência Nacional de Aviação Civil**. 2017. Available from: <www.anac.gov.br>. 23

ARINGER, K.; ROSCHLAUB, R. Bavarian 3D building model and update concept based on lidar, image matching and cadastre information. In: **ISIKDAG, U. (Ed.). Innovations in 3D Geo-Information Sciences**. [S.l.]: Springer, 2014. p. 143–157. 1

AUDEBERT, N.; BOULCH, A.; RANDRIANARIVO, H.; SAUX, B. L.; FERECATU, M.; LEFÉVRE, S.; MARLET, R. Deep learning for urban remote sensing. In: JOINT URBAN REMOTE SENSING EVENT. **Proceedings...** Dubai, United Arab Emirates: JURSE, 2017. p. 1–4. 28, 41

BADRINARAYANAN, V.; KENDALL, A.; CIPOLLA, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. **arXiv preprint arXiv:1511.00561**, 2015. 2, 31, 32, 41

BECKER, S. Generation and application of rules for quality dependent façcade reconstruction. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 64, n. 6, p. 640–653, 2009. 19

BILJECKI, F.; HEUVELINK, G. B.; LEDOUX, H.; STOTER, J. Propagation of positional error in 3D gis: estimation of the solar irradiation of building roofs. **International Journal of Geographical Information Science**, v. 29, n. 12, p. 2269–2294, 2015. 1

BILJECKI, F.; STOTER, J.; LEDOUX, H.; ZLATANOVA, S.; çCÖLTEKIN, A. Applications of 3D city models: State of the art review. **ISPRS International Journal of Geo-Information**, v. 4, n. 4, p. 2842–2889, 2015. 1, 21, 22

BILLEN, R. e. a. **3D City Models and urban information: Current issues and perspectives**. [S.l.]: edp Sciences Les Ulis, France, 2014. 11

BLAHA, M.; VOGEL, C.; RICHARD, A.; WEGNER, J. D.; POCK, T.; SCHINDLER, K. Large-scale semantic 3D reconstruction: an adaptive multi-resolution model for multi-class volumetric labeling. In: CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION. **Proceedings...** [S.l.]: CVPR, 2016. p. 3176–3184. ISSN 1063-6919. 30

BLÁHA, M.; VOGEL, C.; RICHARD, A.; WEGNER, J. D.; POCK, T.; SCHINDLER, K. Towards integrated 3D reconstruction and semantic interpretation of urban scenes. In: DREILÄNDERTAGUNG DER SGPF, DGPF

UND OVG: LÖSUNGEN FÜR EINE WELT IM WANDEL: VORTRÄGE. **Proceedings...** Zurich, Switzerland: DGPF, 2016. p. 44–53. 30

BÓDIS-SZOMORÚ, A.; RIEMENSCHNEIDER, H.; GOOL, L. V. Efficient edge-aware surface mesh reconstruction for urban scenes. **Computer Vision and Image Understanding**, v. 157, p. 3–24, 2017. 3, 20, 21

BOULCH, A.; HOULLIER, S.; MARLET, R.; TOURNAIRE, O. Semantizing complex 3D scenes using constrained attribute grammars. In: PROCEEDINGS OF 11TH EUROGRAPHICS/ACMSIGGRAPH SYMPOSIUM ON GEOMETRY PROCESSING. **Proceedings...** Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2013. (SGP '13), p. 33–42. Available from: <http://dx.doi.org/10.1111/cgf.12170>. 19, 70

BRENNER, C. Towards fully automatic generation of city models. **International Archives of Photogrammetry and Remote Sensing**, v. 33, n. B3/1; pt.3, p. 84–92, 2000. 11

BROSTOW, G. J.; SHOTTON, J.; FAUQUEUR, J.; CIPOLLA, R. Segmentation and recognition using structure from motion point clouds. In: EUROPEAN CONFERENCE ON COMPUTER VISION. **Proceedings...** Berlin, Heidelberg: ECCV, 2008. v. 5302, p. 44–57. 20

BUROCHIN, J.-P.; VALLET, B.; BRÉDIF, M.; MALLET, C.; BROSSET, T.; PAPARODITIS, N. Detecting blind building faç cades from highly overlapping wide angle aerial imagery. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 96, p. 193–209, 2014. 12, 13

CASTELLUCCIO, M.; POGGI, G.; SANSONE, C.; VERDOLIVA, L. Land use classification in remote sensing images by convolutional neural networks. **arXiv preprint arXiv:1508.00092**, 2015. 30

CHARLES, P. **Project Title**. GitHub, 2013. Available from: <https://github.com/charlespwd/project-title>. 78

CHEN, S.; WANG, H. SAR target recognition based on deep learning. In: INTERNATIONAL CONFERENCE ON DATA SCIENCE AND ADVANCED ANALYTICS. **Proceedings...** Santa Rosa, CA, USA, 2014. p. 541–547. 30

CHEN, S.; WANG, H.; XU, F.; JIN, Y.-Q. Target classification using the deep convolutional networks for sar images. **IEEE Transactions on Geoscience and Remote Sensing**, v. 54, n. 8, p. 4806–4817, 2016. 30

CHEN, Y.; LIN, Z.; ZHAO, X.; WANG, G.; GU, Y. Deep learning-based classification of hyperspectral data. **IEEE Journal of Selected topics in applied earth observations and remote sensing**, v. 7, n. 6, p. 2094–2107, 2014. 30

CHENG, L.; GONG, J.; LI, M.; LIU, Y. 3d building model reconstruction from multi-view aerial imagery and LiDAR data. **Photogrammetric Engineering & Remote Sensing**, v. 77, n. 2, p. 125–139, 2011. 12

CIGNONI, P.; CALLIERI, M.; CORSINI, M.; DELLEPIANE, M.; GANOVELLI, F.; RANZUGLIA, G. Meshlab: an open-source mesh processing tool. In: SCARANO, V.; CHIARA, R. D.; ERRA, U. (Ed.). **Eurographics Italian Chapter Conference**. [S.l.]: The Eurographics Association, 2008. ISBN 978-3-905673-68-5. 97

CIRESAN, D. C.; MEIER, U.; GAMBARDELLA, L. M.; SCHMIDHUBER, J. Convolutional neural network committees for handwritten character classification. In: 2011 INTERNATIONAL CONFERENCE ON DOCUMENT ANALYSIS AND RECOGNITION. **Proceedings...** [S.l.]: IEEE, 2011. p. 1135–1139. ISSN 2379-2140. 30

COMANICIU, D.; MEER, P. Mean shift: a robust approach toward feature space analysis. **IEEE Transactions on pattern analysis and machine intelligence**, v. 24, n. 5, p. 603–619, 2002. 19

COMISSÃO NACIONAL DE CARTOGRAFIA. **Especificação técnica para estruturação de dados geoespaciais vetoriais de defesa da força terrestre (ET-EDGV 3.0)**. 2016. CONCAR. 22

DALMAIJER, E. S.; MATHÔT, S.; STIGCHEL, S. Van der. Pygaze: an open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. **Behavior research methods**, v. 46, n. 4, p. 913–921, 2014. 33

DEMIR, N.; BALTSAVIAS, E. Automated modeling of 3D building roofs using image and LiDAR data. In: XXII CONGRESS OF THE INTERNATIONAL SOCIETY FOR PHOTOGRAMMETRY, REMOTE SENSING. **Proceedings...** Melbourne, Australia, 2012. v. 25, p. 35–40. 1, 10

DÖLLNER, J. et al. The virtual 3D city model of berlin-managing, integrating, and communicating complex urban information. In: PROCEEDINGS OF THE 25TH URBAN DATA MANAGEMENT SYMPOSIUM. **Proceedings...** Aalborg, Denmark: UDMS, 2006. v. 2006, p. 15–17. 1

EICKER, U.; MONIEN, D.; DUMINIL, É.; NOUVEL, R. Energy performance assessment in urban planning competitions. **Applied Energy**, v. 155, p. 323–333, 2015. 1

EVERINGHAM, M.; ESLAMI, S. M. A.; GOOL, L. V.; WILLIAMS, C. K. I.; WINN, J.; ZISSERMAN, A. The pascal visual object classes challenge: a retrospective. **International Journal of Computer Vision**, v. 111, n. 1, p. 98–136, jan. 2015. 30

FERRETTI, A.; PRATI, C.; ROCCA, F. Permanent scatterers in SAR interferometry. **IEEE Transactions on Geoscience and Remote Sensing**, v. 39, n. 1, p. 8 –20, jan 2001. ISSN 0196-2892. 10

FRASER, C. S.; CRONK, S. A hybrid measurement approach for close-range photogrammetry. **ISPRS journal of photogrammetry and remote sensing**, v. 64, n. 3, p. 328–333, 2009. 23

FURUKAWA, Y.; CURLESS, B.; SEITZ, S. M.; SZELISKI, R. Clustering views for multi-view stereo. In: COMPUTER VISION AND PATTERN RECOGNITION (CVPR). **Proceedings...** San Francisco, CA, USA: IEEE, 2010. v. 13, p. e18. 26

FURUKAWA, Y.; PONCE, J. Accurate, dense, and robust multi-view stereopsis. In: COMPUTER VISION AND PATTERN RECOGNITION. **Proceedings...** Minneapolis, MN, USA: IEEE, 2007. p. 1–8. ISSN 1063-6919. 26

_____. **Patch-based multi-view stereo software**. 2010. Available from: <https://www.di.ens.fr/pmvs/>. 26

GADDE, R.; JAMPANI, V.; MARLET, R.; GEHLER, P. Efficient 2d and 3D facade segmentation using auto-context. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 2017. 3, 20, 37

GERKE, M.; XIAO, J. Fusion of airborne laserscanning point clouds and images for supervised and unsupervised scene classification. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 87, p. 78–92, 2014. 10

GONZALEZ, R.; WOODS, R. E. **Digital image processing**. [S.l.]: Addison-wesley Reading, 1992. 26

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. [S.l.]: MIT Press, 2016. 40

GOOL, L. V.; MARTINOVIC, A.; MATHIAS, M. Towards semantic city models. In: PHOTOGRAMMETRIC WEEK'13. **Proceedings...** [S.l.], 2013. p. 217–232. 10, 22

HAALA, N.; BRENNER, C. Extraction of buildings and trees in urban environments. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 54, n. 2, p. 130–137, 1999. 11

HÄNE, C.; ZACH, C.; COHEN, A.; ANGST, R.; POLLEFEYS, M. Joint 3D scene reconstruction and class segmentation. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION. **Proceedings...** Portland, Oregon, USA: CVPR, 2013. p. 97–104. 30

HARRIS, C.; STEPHENS, M. A combined corner and edge detector. In: ALVEY VISION CONFERENCE. **Proceedings...** [S.l.], 1988. v. 15, n. 50, p. 10–5244. 25

HARTLEY, R. I.; ZISSERMAN, A. **Multiple view geometry in computer vision**. 2.ed. cambridge. ed. [S.l.]: Cambridge University Press, 2004. ISBN 0521540518. 44

HE, X.; ZHANG, H.; HUR, N.; KIM, J.; WU, Q.; KIM, T. Estimation of internal and external parameters for camera calibration using 1d pattern. In: IEEE INTERNATIONAL CONFERENCE ON VIDEO AND SIGNAL BASED SURVEILLANCE. **Proceedings...** [S.l.]: AVSS, 2006. p. 93–93. 43

HENN, A.; RÖMER, C.; GRÖGER, G.; PLÜMER, L. Automatic classification of building types in 3D city models. **GeoInformatica**, v. 16, n. 2, p. 281–306, 2012. 14

HIRSCHMULLER, H. Stereo processing by semiglobal matching and mutual information. **IEEE Transactions on pattern analysis and machine intelligence**, v. 30, n. 2, p. 328–341, 2008. 26

HOFER, M.; MAURER, M.; BISCHOF, H. Efficient 3D scene abstraction using line segments. **Computer Vision and Image Understanding**, Elsevier, 2016. 25

HOPPE, H. Poisson surface reconstruction and its applications. In: SYMPOSIUM ON SOLID AND PHYSICAL MODELING. **Proceedings...** New York, USA: SSPM, 2008. p. 10–10. 26

HUBEL, D. H.; WIESEL, T. N. Receptive fields and functional architecture of monkey striate cortex. **The Journal of physiology**, v. 195, n. 1, p. 215–243, 1968. 28

HUNTER, J. D. Matplotlib: a 2d graphics environment. **Computing in Science & Engineering**, v. 9, n. 3, p. 90–95, May-Jun 2007. 97

JACOBSON, A. et al. **libigl: a simple C++ geometry processing library**. 2016. Available from: <http://libigl.github.io/libigl/>. 97

JAMPANI, V.; GADDE, R.; GEHLER, P. V. Efficient facade segmentation using auto-context. In: APPLICATIONS OF COMPUTER VISION. **Proceedings...** [S.l.]: WACV, 2015. p. 1038–1045. 20

JANSSEN, W.; BLOCKEN, B.; HOOFF, T. van. Pedestrian wind comfort around buildings: comparison of wind comfort criteria based on whole-flow field data for a complex case study. **Building and Environment**, v. 59, p. 547–562, 2013. 21

JOCHEM, A.; HÖFLE, B.; RUTZINGER, M.; PFEIFER, N. Automatic roof plane detection and analysis in airborne lidar point clouds for solar potential assessment. **Sensors**, v. 9, n. 7, p. 5241–5262, 2009. 1

KALCHBRENNER, N.; GREFENSTETTE, E.; BLUNSOM, P. A convolutional neural network for modelling sentences. **arXiv preprint arXiv:1404.2188**, 2014. 30

KARPATHY, A.; TODERICI, G.; SHETTY, S.; LEUNG, T.; SUKTHANKAR, R.; FEI-FEI, L. Large-scale video classification with convolutional neural networks. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION. **Proceedings...** [S.l.]: CVPR, 2014. 41

KASS, M.; WITKIN, A.; TERZOPOULOS, D. Snakes: active contour models. **International Journal of Computer Vision**, v. 1, n. 4, p. 321–331, 1987. 19

KLUIJVER, H. de; STOTER, J. Noise mapping and gis: optimising quality and efficiency of noise effect studies. **Computers, Environment and Urban Systems**, v. 27, n. 1, p. 85–102, 2003. 22

KO, H.-k.; SNODDERLY, D. M.; POLETTI, M. Eye movements between saccades: measuring ocular drift and tremor. **Vision Research**, v. 122, p. 93–104, 2016. 32

KOLBE, T. H. Representing and exchanging 3D city models with citygml. In: **3D geo-information sciences**. Seoul, Korea: Springer, 2009. chapter 2, p. 15–31. 1

KOLBE, T. H.; GRÖGER, G.; PLÜMER, L. CityGML: interoperable access to 3D city models. In: **Geo-information for disaster management**. [S.l.]: Springer, 2005. p. 883–899. 7, 18

KOLMOGOROV, V.; ZABIN, R. What energy functions can be minimized via graph cuts? **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 26, n. 2, p. 147–159, 2004. 19

KORC, F.; FÖRSTNER, W. eTRIMS image database for interpreting images of man-made scenes. **Bonn: University of Bonn**, 2009. 37

KRESSE, W.; FADAIE, K. **ISO standards for geographic information**. [S.l.]: Springer Science & Business Media, 2004. 17

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS. **Proceedings...** [S.l.]: ANIPS, 2012. p. 1097–1105. 2, 30, 32, 41

KRÜGER, A.; KOLBE, T. H. Building analysis for urban energy planning using key indicators on virtual 3D city modelsï¿½the energy atlas of Berlin. **International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences**, v. 39, n. B2, p. 145–150, 2012. 1

KWAN, M.-P.; LEE, J. Emergency response after 9/11: the potential of real-time 3D gis for quick emergency response in micro-spatial environments. **Computers, Environment and Urban Systems**, v. 29, n. 2, p. 93–113, 2005. 1, 21

LAFARGE, F. Some new research directions to explore in urban reconstruction. In: **Proceedings...** [S.l.: s.n.]. ISSN 2334-0932. 1

LAFARGE, F.; MALLET, C. Building large urban environments from unstructured point data. In: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION. **Proceedings...** [S.l.]: ICCV, 2011. p. 1068–1075. 20

LANCELLE, M.; FELLNER, D. W. Current issues on 3D city models. **Proceedings of the 25th International Conference in Image and Vision Computing**, p. 363–369, 2010. 11

LAStools. **Efficient LiDAR Processing Software**. Germany, 2014. Available from: <http://rapidlasso.com/LAStools>. 97

LAWRENCE, S.; GILES, C. L.; TSOI, A. C.; BACK, A. D. Face recognition: A convolutional neural-network approach. **IEEE Transactions on Neural Networks**, v. 8, n. 1, p. 98–113, 1997. 30

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, v. 521, n. 7553, p. 436–444, 2015. 28, 30, 33

LECUN, Y.; BOSER, B.; DENKER, J. S.; HENDERSON, D.; HOWARD, R. E.; HUBBARD, W.; JACKEL, L. D. Backpropagation applied to handwritten zip code recognition. **Neural computation**, v. 1, n. 4, p. 541–551, 1989. 2, 29

LEE, G. 3D coverage location modeling of wi-fi access point placement in indoor environment. **Computers, Environment and Urban Systems**, v. 54, p. 326–335, 2015. 1

LESZEK, K. Environmental and urban spatial analysis based on a 3D city model. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL SCIENCE AND ITS APPLICATIONS. **Proceedings...** [S.l.]: Springer, 2015. p. 633–645. 21

LETTRY, L.; PERDOCH, M.; VANHOEY, K.; GOOL, L. V. Repeated pattern detection using cnn activations. In: IEEE WINTER CONFERENCE ON APPLICATIONS OF COMPUTER VISION. **Proceedings...** [S.l.]: WACV, 2017. p. 47–55. 31, 32

LI, M.; NAN, L.; SMITH, N.; WONKA, P. Reconstructing building mass models from uav images. **Computers & Graphics**, v. 54, p. 84–93, 2016. 20

LIU, H.; ZHANG, J.; ZHU, J.; HOI, S. C. DeepFacade: a deep learning approach to facade parsing. p. 93, 2017. 33

LLC, A. **Agisoft Photoscan User Manual**. [S.l.], 2018. Available from: <http://www.agisoft.com>. 43

LONG, J.; SHELHAMER, E.; DARRELL, T. Fully convolutional networks for semantic segmentation. In: CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION. **Proceedings...** [S.l.]: CVPR, 2015. p. 3431–3440. 41, 50

LOWE, D. G. Distinctive image features from scale-invariant keypoints. **International Journal of Computer Vision**, v. 60, n. 2, p. 91–110, 2004. 25, 26

MARMANIS, D.; DATCU, M.; ESCH, T.; STILLA, U. Deep learning earth observation classification using imagenet pretrained networks. **IEEE Geoscience and Remote Sensing Letters**, v. 13, n. 1, p. 105–109, 2016. 30

MARTINOVIC, A.; KNOPP, J.; RIEMENSCHNEIDER, H.; GOOL, L. V. 3D all the way: semantic segmentation of urban scenes from start to end in 3d. In: CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION. **Proceedings...** Boston, Massachusetts, USA: CVPR, 2015. p. 4456–4465. 16, 19, 20, 21, 70

MATHIAS, M.; MARTINOVIC, A.; WEISSENBERG, J.; HAEGLER, S.; GOOL, L. V. Automatic architectural style recognition. **ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences**, v. 3816, p. 171–176, 2011. 14

MCCULLOCH, W.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **Bulletin of Mathematical Biophysics**, v. 5, p. 115–133, 1943. 28

MING, Y.; JIANG, J.; BIAN, F. 3d-city model supporting for cctv monitoring system. **International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences**, v. 34, n. 4, p. 456–459, 2002. 21

MITCHELL, T. M. et al. **Machine learning. WCB**. [S.l.]: Boston: McGraw-Hill, 1997. 28

MÜLLER, P.; WONKA, P.; HAEGLER, S.; ULMER, A.; GOOL, L. V. Procedural modeling of buildings. In: ACM TRANSACTIONS ON GRAPHICS. **Proceedings...** [S.l.], 2006. v. 25, n. 3, p. 614–623. 10

MUSIALSKI, P.; WONKA, P.; ALIAGA, D. G.; WIMMER, M.; GOOL, L.; PURGATHOFER, W. A survey of urban reconstruction. In: COMPUTER GRAPHICS FORUM. **Proceedings...** [S.l.], 2013. v. 32, n. 6, p. 146–177. 3, 10

NAN, L.; SHARF, A.; ZHANG, H.; COHEN-OR, D.; CHEN, B. Smartboxes for interactive urban reconstruction. In: ACM TRANSACTIONS ON GRAPHICS. **Proceedings...** [S.l.], 2010. v. 29, n. 4, p. 93. 19

OESAU, S.; LAFARGE, F.; ALLIEZ, P. Planar shape detection and regularization in tandem. In: COMPUTER GRAPHICS FORUM. **Proceedings...** [S.l.], 2016. v. 35, n. 1, p. 203–215. 20

OGC. **OGC Standards**. Feb. 2014. Available from: <http://www.opengeospatial.org/>. 17

PIRKL, G.; HEVESI, P.; LUKOWICZ, P.; KLEIN, P.; HEISEL, C.; GRÖBER, S.; KUHN, J.; SICK, B. Any problems? a wearable sensor-based platform for representational learning-analytics. In: NTERNATIONAL JOINT CONFERENCE ON PERVASIVE AND UBIQUITOUS COMPUTING: ADJUNCT. **Proceedings...** [S.l.]: ACM, 2016. p. 353–356. 32, 33

PROJECT, T. C. CGAL, computational geometry algorithms library. In: . 4.10. ed. [s.n.], 2017. Available from: <http://doc.cgal.org/4.10/Manual/packages.html>. 97

QIN, R. Change detection on lod 2 building models with very high resolution spaceborne stereo imagery. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 96, p. 179–192, 2014. 21

REMONDINO, F.; EL-HAKIM, S. Image-based 3D modelling: A review. **The Photogrammetric Record**, v. 21, n. 115, p. 269–291, 2006. 23

REMONDINO, F.; GUARNIERI, A.; VETTORE, A. 3d modeling of close-range objects: photogrammetry or laser scanning. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. **Proceedings...** [S.l.]: SPIE, 2005. v. 5665, p. 216–225. 23

RIEMENSCHNEIDER, H.; BÓDIS-SZOMORÚ, A.; WEISSENBERG, J.; GOOL, L. V. Learning where to classify in multi-view semantic segmentation. In: EUROPEAN CONFERENCE ON COMPUTER VISION. **Proceedings...** Zurich, Switzerland: ECCV, 2014. p. 516–532. 3, 20, 21, 25, 34, 37, 47

RIEMENSCHNEIDER, H.; KRISPEL, U.; THALLER, W.; DONOSER, M.; HAVEMANN, S.; FELLNER, D.; BISCHOF, H. Irregular lattices for complex shape grammar facade parsing. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION. **Proceedings...** Province, Rhode Island: CVPR, 2012. p. 1640–1647. 37

ROMERO, A.; GATTA, C.; CAMPS-VALLS, G. Unsupervised deep feature extraction of hyperspectral images. In: PROC. 6TH WORKSHOP HYPERSPECTRAL IMAGE SIGNAL PROCESS. EVOL. REMOTE SENS.(WHISPERS). **Proceedings...** [S.l.], 2014. 30

RUSU, R. B.; COUSINS, S. 3D is here: Point Cloud Library (PCL). In: IEEE INTERNATIONAL CONFERENCE ON ROBOTICS AND AUTOMATION. **Proceedings...** Shanghai, China: ICRA, 2011. 97

SABRI, S.; PETTIT, C. J.; KALANTARI, M.; RAJABIFARD, A.; WHITE, M.; LADE, O.; NGO, T. What are essential requirements in planning for future cities using open data infrastructures and 3D data models. In: 14TH INTERNATIONAL CONFERENCE ON COMPUTERS IN URBAN PLANNING AND URBAN MANAGEMENT. **Proceedings...** Cambridge, MA, USA, 2015. p. 7–10. 21

SALEHI, A.; MOHAMMADZADEH, A. Building roof reconstruction based on residue anomaly analysis and shape descriptors from lidar and optical data. **Photogrammetric Engineering & Remote Sensing**, v. 83, n. 4, p. 281–291, 2017. 2

SCHACK, L.; SCHUNERT, A.; SOERGEL, U. Lattice detection in persistent scatterer point clouds and oblique aerial imagery. In: IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM. **Proceedings...** [S.l.]: IGARSS, 2012. p. 451–454. 10

SCHöNBERGER, J. L.; FRAHM, J.-M. Structure-from-motion revisited. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION. **Proceedings...** [S.l.]: CVPR, 2016. 45

SCHUNERT, A.; SOERGEL, U. Assignment of persistent scatterers to buildings. **IEEE Transactions on Geoscience and Remote Sensing**, IEEE, v. 54, n. 6, p. 3116–3127, 2016. 10

SEDAGHAT, A.; EBADI, H. Distinctive order based self-similarity descriptor for multi-sensor remote sensing image matching. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 108, p. 62–71, 2015. 26

SENGUPTA, S.; VALENTIN, J.; WARRELL, J.; SHAHROKNI, A.; TORR, P. Mesh based semantic modelling for indoor and outdoor scenes. In: COMPUTER VISION AND PATTERN RECOGNITION. **Proceedings...** [S.l.]: CVPR, 2013. v. 6, n. 6.2, p. 2067–2074. ISSN 1063-6919. 21

SHI, J.; MALIK, J. Normalized cuts and image segmentation. **IEEE Transactions on pattern analysis and machine intelligence**, Ieee, v. 22, n. 8, p. 888–905, 2000. 19

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. **arXiv preprint arXiv:1409.1556**, 2014. 41

SNAVELY, K. N. Scene reconstruction and visualization from internet photo collections. 2009. 43

SNAVELY, N.; SEITZ, S.; SZELISKI, R. Photo tourism: exploring image collections in 3d. **ACM Transactions on Graphics**, 2006. 1

SNAVELY, N.; SEITZ, S. M.; SZELISKI, R. Modeling the world from internet photo collections. **International Journal of Computer Vision**, v. 80, n. 2, p. 189–210, 2008. 25

STINY, G.; GIPS, J. Shape grammars and the generative specification of painting and sculpture. In: IFIP CONGRESS (2). **Proceedings...** [S.l.], 1971. v. 2, n. 3. 19, 70

STOTER, J.; KLUIJVER, H. D.; KURAKULA, V. 3d noise mapping in urban areas. **International Journal of Geographical Information Science**, v. 22, n. 8, p. 907–924, 2008. 22

STOTER, J.; VOSSELMAN, G.; GOOS, J.; ZLATANOVA, S.; VERBREE, E.; KLOOSTER, R.; REUVERS, M. Towards a national 3D spatial data infrastructure: case of the netherlands. **Photogrammetrie-Fernerkundung-Geoinformation**, v. 2011, n. 6, p. 405–420, 2011. 1

SUN, Z.; XUE, L.; XU, Y.; WANG, H. Marginal fisher analysis feature extraction algorithm based on multilayer auto-encoder. **JOURNAL OF INFORMATION &COMPUTATIONAL SCIENCE**, v. 9, n. 18, p. 5897–5906, 2012. 30

SWIETOJANSKI, P.; GHOSHAL, A.; RENALS, S. Convolutional neural networks for distant speech recognition. **IEEE Signal Processing Letters**, v. 21, n. 9, p. 1120–1124, 2014. 40

TEBOUL, O.; KOKKINOS, I.; SIMON, L.; KOUTSOURAKIS, P.; PARAGIOS, N. Shape grammar parsing via reinforcement learning. In: COMPUTER VISION AND PATTERN RECOGNITION. **Proceedings...** [S.l.]: CVPR, 2011. p. 2273–2280. 3, 19

TEBOUL, O.; SIMON, L.; KOUTSOURAKIS, P.; PARAGIOS, N. Segmentation of building facades using procedural shape priors. In: COMPUTER VISION AND PATTERN RECOGNITION. **Proceedings...** [S.l.]: CVPR, 2010. p. 3105–3112. 37

TEICHMANN, M.; WEBER, M.; ZOELLNER, M.; CIPOLLA, R.; URTASUN, R. Multinet: Real-time joint semantic reasoning for autonomous driving. **arXiv preprint arXiv:1612.07695**, 2016. 2, 31, 32, 41

TOOKE, T. R.; COOPS, N. C.; VOOGT, J. A.; MEITNER, M. J. Tree structure influences on rooftop-received solar radiation. **Landscape and Urban Planning**, v. 102, n. 2, p. 73–81, 2011. 1

TRIGGS, B.; MCLAUCHLAN, P. F.; HARTLEY, R. I.; FITZGIBBON, A. W. Bundle adjustment: a modern synthesis. In: INTERNATIONAL WORKSHOP ON VISION ALGORITHMS. **Proceedings...** [S.l.]: Springer, 1999. p. 298–372. 25

TRUONG-HONG, L.; LAEFER, D. F. Octree-based, automatic building facade generation from lidar data. **Computer-Aided Design**, v. 53, p. 46–61, 2014. 1

TU, Z.; BAI, X. Auto-context and its application to high-level vision tasks and 3D brain image segmentation. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 32, n. 10, p. 1744–1757, 2010. 20

TUTZAUER, P.; HAALA, N. Facade reconstruction using geometric and radiometric point cloud information. **The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences**, v. 40, n. 3, p. 247, 2015. 11, 12

TYLEČEK, R.; ŠÁRA, R. Spatial pattern templates for recognition of objects with regular structure. In: GERMAN CONFERENCE ON PATTERN RECOGNITION. **Proceedings...** [S.l.]: Springer, 2013. p. 364–374. 37

VEDALDI, A.; SOATTO, S. Quick shift and kernel methods for mode seeking. **Computer vision–ECCV 2008**, Springer, p. 705–718, 2008. 97

VERDIE, Y.; LAFARGE, F.; ALLIEZ, P. **LOD Generation for urban scenes**. [S.l.], 2015. 12, 20

VERTEX MODELLING. **Vertex Modelling: 3D Model of London**. 2015. Available from: <http://vertexmodelling.co.uk>. 1

virtualcitySYSTEMS GmbH. **virtualcitySYSTEMS**. 2016. Available from: <http://www.virtualcitysystems.de>. 1

VOSSELMAN, G.; DIJKMAN, S. et al. 3d building model reconstruction from point clouds and ground plans. **International Archives of Photogrammetry**

Remote Sensing and Spatial Information Sciences, v. 34, n. 3/W4, p. 37–44, 2001. 1

WAN, G.; SHARF, A. Grammar-based 3D facade segmentation and reconstruction. **Computers & Graphics**, v. 36, n. 4, p. 216–223, 2012. 19

WEISSENBERG, J. **Inverse Procedural Modelling and Applications**. PhD Thesis (PhD) — ETH-Zürich, 2014. 14

WENZEL, S.; FÖRSTNER, W. Semi-supervised incremental learning of hierarchical appearance models. In: 21ST CONGRESS OF THE INTERNATIONAL SOCIETY FOR PHOTOGRAMMETRY AND REMOTE SENSING. **Proceedings...** [S.l.]: ISPRS, 2008. v. 3, p. 399–405. 19

WESTOBY, M.; BRASINGTON, J.; GLASSER, N.; HAMBREY, M.; REYNOLDS, J. Structure-from-Motion photogrammetry: a low-cost, effective tool for geoscience applications. **Geomorphology**, Elsevier, v. 179, p. 300–314, 2012. 24, 27

WU, C. Siftgpu: a GPU implementation of scale invariant feature transform (SIFT). 2007. Available from: <http://cs.unc.edu/~{}ccwu/siftgpu>. 26

WU, C.; AGARWAL, S.; CURLESS, B.; SEITZ, S. M. Multicore bundle adjustment. In: COMPUTER VISION AND PATTERN RECOGNITION. **Proceedings...** Colorado, USA: CVPR, 2011. p. 3057–3064. 26

WU, C. et al. VisualSFM: a visual structure from motion system. 2011. 45

YAAGOUBI, R.; YARMANI, M. E.; KAMEL, A.; KHEMIRI, W. HybVOR: a voronoi-based 3D gis approach for camera surveillance network placement. **ISPRS International Journal of Geo-Information**, v. 4, n. 2, p. 754–782, 2015. 1, 21

YANG, L.; SHENG, Y.; WANG, B. 3d reconstruction of building facade with fused data of terrestrial lidar data and optical image. **Optik-International Journal for Light and Electron Optics**, v. 127, n. 4, p. 2165–2168, 2016. 1

YUAN, X.; CHEN, S.; YUAN, W.; CAI, Y. Poor textural image tie point matching via graph theory. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 129, p. 21–31, 2017. 26, 27

ZHANG, L.; GRUEN, A. Multi-image matching for DSM generation from IKONOS imagery. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 60, n. 3, p. 195–211, 2006. 26

ZHENG, S.; JAYASUMANA, S.; ROMERA-PAREDES, B.; VINEET, V.; SU, Z.; DU, D.; HUANG, C.; TORR, P. H. S. Conditional random fields as recurrent neural networks. In: INTERNATIONAL CONFERENCE ON COMPUTER VISION. **Proceedings...** [S.l.]: ICCV, 2015. 2

ZHU, X. X.; TUIA, D.; MOU, L.; XIA, G.-S.; ZHANG, L.; XU, F.; FRAUNDORFER, F. Deep learning in remote sensing: a review. **arXiv preprint arXiv:1710.03959**, 2017. 30, 33, 35, 40, 41

APPENDIX A - MATERIAL DETAILS

A.1 Tools and modules used along the study

Table A.1 - Sources-code, softwares, libraries and their public links on Github.

Name	Description	Type	Version	Purpose	URL	Reference
scipy	Mathematics, science, and engineering	Python module		Image Processing		-
numpy	Scientific computing	SciPy module		Image Processing		-
matplotlib	Comprehensive 2D Plotting	SciPy module		Image Processing		(HUNTER, 2007)
sklearn	Machine Learning	Python module	0.18.1	Image Processing		-
pylsd	Line Segment Detector (LSD)	Python Library	-	Image Processing		-
skimage	Image processing	Python module	0.14	Image Processing		-
quickshift	Image segmentation	SkImage module	-	Image Processing		(VEDALDI; SOATTO, 2008)
cv2	OpenCV Image Processing	Python Library	3.2.0.8	Image Processing		-
tensorflow	Tensorflow	Python Library	1.5	Image Processing		-
embree	High Performance Ray Tracing Kernels	C++ Library	2.0	Ray-tracing		-
vtk	The Visualization Toolkit	C++ Library	7.1.1	Ray-tracing		-
cgal	Computational Geometry Algorithms (CGAL)	C++ Library	4.10	Point cloud processing		(PROJECT, 2017)
pcl	Point Cloud Library (PCL)	C++ Library	4.7	Point cloud processing		(RUSU; COUSINS, 2011)
lastools	LAStools	C++ Module	-	Point cloud processing		(LAStools, 2014)
libigl	C++ geometry processing	C++ Library	1.0	Visualizatior		(JACOBSON et al., 2016)
CloudCompare	3D point cloud and mesh processing	GUI	2.9	Visualizatior		-
MeshLab	3D point cloud and mesh processing	GUI	2016.12	Visualizatior		(CIGNONI et al., 2008)

SOURCE: Author's production.

A.2 Author's production along this study

Table A.2 - Sources-code produced by the author throughout the study. The usability and further explanations, can be found at the respective Github links.

Name	Description	Methodological stage	Language	Github	Type	URL
Image processing	Preprocessing on images to apply the Superpixel routine. The preprocessing procedures here include filtering, colormap conversion (i.e. HSV, Lab, Local Entropy), etc	Detection	C++	rodolfolotte/image	-	
Superpixel	Segmentation routine with numerous versions of Superpixels, for example, SLIC, SLICO, LSD and SEEDS. The routine was programmed to perform experiments involving classification of facades	Detection	C++	rodolfolotte/gdal-segment		
Point cloud preprocessing	Includes point cloud preprocessing routines, such as simplification, filtering, CRF classification, etc	Reconstruction	C++	rodolfolotte/point-cloud	-	
Inputs preparation	Preparation code for neural network inputs. The code includes preparation of annotated images and creation of text files containing the paths of the images	Detection	Python	rodolfolotte/deep-learning	-	
CNN	Routines responsible for the Convolutional Neural Network training, as well as the routines for the inferences	Detection	Python	rodolfolotte/deep-learning		
Ray-tracing	Read the camera parameters, and by the configuration of image blocks, project the segmented image onto the mesh, in a way the triangles mesh are assigned according to each class	Ray-Tracing	C++	rodolfolotte/3d-reconstruction	-	
Evaluation	Evaluates the segmented images against ground-truth. Here, all evaluation metrics and plots can be setted up	Detection	Python	rodolfolotte/image	-	
PC vizualization	Routines to visualize meshes and customized point of views, colors, etc	Reconstruction	C++	rodolfolotte/3d-reconstruction	-	
Thesis	LaTeXversion of this thesis	Documentation	LaTeX	rodolfolotte/doc	-	

SOURCE: Author's production.

APPENDIX B - 3D MAPPING IN BRAZIL

B.1 Poll about the use of 3D maps in Brazil

The questionary was made by email, in some cases, was not possible to reach some of the capitals either by email or phone. The detailed answers from Table 3.1 can be seen in Table B.1.

The contact of each professional responsible for the answers, can be found in Table B.2.

Table B.1 - Poll answers from Table 3.1 in details.

Capital	State	Answer
Fortaleza	CE	**a)** The City Hall of Fortaleza is acquiring the 3D mapping by laser profiling. Currently, department does not use this data for studies. **d)** Planning is performed without any 3D map. **e)** Studies using the Z component.
Vitória	ES	**a)** No. **d)** By 2D thematic maps, with additional database information, such as occupation, use, height, feedback, distribution of activities, commerce and urban services, public equipment, socioeconomic characteristics, among others. **e)** For example, Vitória has a map on the protection of the natural heritage of the municipality and the landscape. A 3D map would aid in reconciling territory occupation policies with the protection policy of this heritage, as well as advanced studies such as volumetric tests, among others. It can be an important tool in the studies of protection of the environmental patrimony, mangroves, hills and conservation units, allowing a reading of the constructed object of the city and its relation with the environmentally protected areas.
Porto Alegre	RS	**a)** Yes, Digital Surface Model (DSM) and Digital Terrain Model (DTM). **b)** At the moment, in the identification of areas susceptible to geological, geotechnical and hydrological risks; in the estimation of the height of the buildings, for tax purposes. GIS system in Web platform, for internal queries. **e)** Support for decision making, confronting the planned environment with the built environment, impact and shading simulations, volumetric simulations. Volumetric studies.
São Paulo	SP	**a)** Yes. DSM and DTM. In 2017, a new campaign using LiDAR sensors was performed and new 3D models will be produced. **b)** Workings such as the Strategic Master Plan, Zoning, Drainage Plan, among others. **c)** The DTM is used as a reference for other surveys, including planning and execution of some current aerial imaging stages. However, in projects and infrastructure, the effective use of these 3D digital technologies is still in the process of being disseminated to the technical staff. **d)** Although the 3D models are available, 2D is still predominantly used in urban planning activities. **e)** They incorporate technical features that are extremely useful for urban management.
Belo Horizonte	MG	**a)** Yes. The execution of 3D modeling is carried out on demand in projects or urban plans that require analysis of modifications in the urban landscape. **b)** The 3D maps help in the visualization of impacts, in the increase of the constructive density, in the generation of obstructions of sightings, in the relations between equipment and urban infrastructures, among others. **c)** Volumetry studies are carried out for specific projects, such as regions of Urban Operations, Plans in Areas of Social Interest, Plans of Cultural Regions. **d)** Planning is already done with the eventual use of 3D modeling of the terrain and buildings. **e)** The 3D maps help in diagnosing the situation before and after the implementation of projects to evaluate the urban impacts of density and volumetry in parameters established in the plans.
Rio de Janeiro	RJ	**a)** The buildings are represented in two dimensions with possibility of elevation in 3D. Calculated density calculations for proposed zoning alterations, urban parameters and studies of impacts on the landscape. **b)** 3D simulations allow for technical discussion, improvement of proposals and better communication with managers, city councilors and the general population. **c)** Most of urban law plans made in the last five years have included 3D maps on strategic areas. **d)** In addition to the 3D studies, the urban planning uses information produced by both federal and state bodies such as IBGE, ISPE and also municipal, produced by the municipal authorities. **e)** Associating the relief with the buildings, it contributes to studies of insolation, heat islands, aimed at the natural and cultural landscape, simulations and urban parameters such as building heights and their distances to the other buildings and impacts on the environment.
Curitiba	PR	**a)** Yes. Partial maps, that is, not of the whole city, but of parts of it. **b)** Providing visualization of occupancy scenarios (for example, using all basic building potential or with extra potential acquisition); in studies of insolation and microclimate. Answered in the previous question. **e)** In addition to the current applications, these maps could be used, for example, in Neighborhood Impact Studies and project detailing.
Recife	PE	**a)** Yes, the Urban Planning Dept. has georeferenced databased with vector and raster data from LiDAR campaings. Another tool that it is used is Google Earth Pro, which has the 3D model of the entire municipality of Recife. **b)** A 3D model allows us to manipulate, take measurements, visualize any interventions already built, and also simulate transformations from anywhere. **c)** Any analysis of morphology and occupation of the territory. These activities include analyzes of impact projects, simulation of changes in legislation and simulations, and studies of interventions.

SOURCE: Author's production.

Table B.2 - Responsibles for the 3D mapping poll answers, sent to the Brazilian capitals infraestructure department (Table B.1, in Section 3.1).

Capital	**State**	**Name**	**Email**
Fortaleza	CE	Ouvidoria	ouvidoria.seuma@fortaleza.ce.gov.br
Vitória	ES	Clivia Leite Mendonça	clivia.leite@correio1.vitoria.es.gov.br
Porto Alegre	RS	Rodrigo Marsillac Linn	rodrigoml@smurb.prefpoa.com.br
São Paulo	SP	Ana Maria A. Laurenza and Silvio C. L. Ribeiro	amlaurenza@prefeitura.sp.gov.br
Belo Horizonte	MG	Guilherme Pereira de Vargas	guilherme.vargas@pbh.gov.br
Rio de Janeiro	RJ	Valéria Hazan	valeriahazan.pcrj@gmail.com
Curitiba	PR	Oscar Ricardo M. Schmeiske and Alessandro Dias	geoprocessamento@ippuc.org.br
Recife	PE	Tiago Henrique	icps@recife.pe.gov.br

SOURCE: Author's production.

GLOSSARY

3D representation A digital 3D version of a real life object.

3D reconstruction A real life object that has been created digitally through an automatic or semi-automatic computational procedure.

3D generation A real life object that has been created digitally through a manual edition, or specialist in 3D designs.

3D model Same as 3D representation.

3D labeling The procedure of classifying a 3D representation.

Image annotation Images that carries the classified features, or, what each pixel corresponds to. These images are used as ground-truth during CNN training.

Street-side images Photos taken from the streets, from the ground and pointed toward the facade (perpendicular).

Facade features Individual elements essential to urban mapping. These elements could be, for example, decisive in the application and execution of urban and supervisory laws.

Openings Same as Facade Features. Openings are all those related to window, door, gates, entrances, etc.

ANNEX A - CROSS-ENTROPY

A.1 Example of cross-entropy calculation

The main goal when developing a probabilistic classificator is to map inputs to probabilistic predictions, incrementally adjusting the model's parameters as the amount of erros are observed. Thus, that prediction get closer and closer to ground-truth probabilities. One way to interpret cross-entropy is to see it as a negative log-likelihood for the ground-truth (y'_i), under their prediction (y_i). Then, "get closer" means that when distance between y'_i and y_i is minimal.

Taking an example from internet[1], supposing a fixed model (hypothesis) that predicts for n classes $1, 2, ..., n$ their hypothetical occurrence probabilities $y_1, y_2, ..., y_n$. Suppose that now observing (in reality) k_1 instances of class 1, k_2 instances of class 2, k_n instances of class n, so on. According to this model, the likelihood of this happening is:

$$P[data|model] = y_1^{k_1}\ y_2^{k_2}\ ...\ y_n^{k_n}\ , \tag{A.1}$$

taking the logarithm and changing the sign:

$$\begin{aligned} -log\ P[data|model] &= -k_1\ \log\ y_1\ - k_2\ \log\ y_2\ ...\ - k_n\ \log\ y_n & \text{(A.2)} \\ &= -\sum_i k_i\ log\ y_i\ , & \text{(A.3)} \end{aligned}$$

dividing the right-hand sum by the number of observations $N = k_1 + k_2 + ... + k_n$, and denote the empirical probabilities as $y'_i = k_i/N$, the cross-entropy is then obtained:

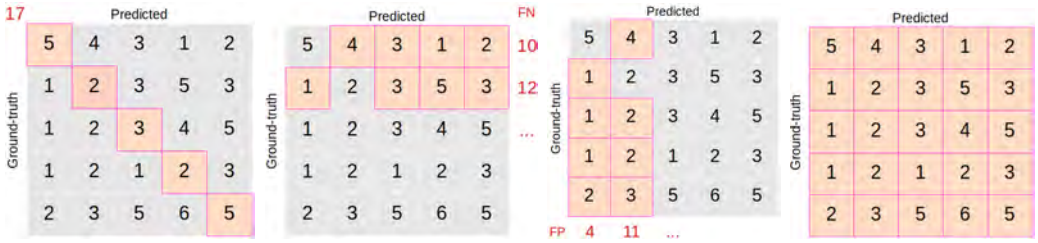
$$\begin{aligned} -\frac{1}{N} log\ P[data|model] &= -\frac{1}{N}\sum_i k_i\ \log\ y_i & \text{(A.4)} \\ &= -\sum_i y'_i\ \log\ y_i & \text{(A.5)} \\ &= Loss_{y'}(y)\ . & \text{(A.6)} \end{aligned}$$

[1] Available at https://rdipietro.github.io/friendly-intro-to-cross-entropy-loss/#cross-entropy. Accessed November 2, 2018.

ANNEX B - 2D EVALUATION

For validation, reference and prediction images (of the same dimension and color map) are essentially required. Once in agreement, the images are confronted so that an N number of randomly selected samples are compared according to their color properties. The resulting success and error count is acquired in the form of a matrix, the confusion matrix. One of the axes denotes the reference classes, and another, the predicted classes. The resulting matrix as well as the success and error metrics (TP, TN, FP, and FN), therefore, is illustrated in Figure B.1.

Figure B.1 - Multiclass confusion matrix and the respective success and error rates.



(a) TP (b) FN (c) FP (d) TN

SOURCE: Author's production.

Then, considering M as the confusion matrix, C number of classes, i (predicted) and j (reference) as the axes, the metrics are given by:

$$M_{ij} = \begin{cases} TP = \forall\ m\ \in\ M:\ i = j, m\ \Re > 0\ , \\ FP = \sum_{j=1}^{C} M_{ij} - TP:\ j = 1, ..., C\ , \\ FN = \sum_{i=1}^{C} M_{ij} - TP:\ i = 1, ..., C\ , \\ TN = \sum_{i=1,j=1}^{C} M_{ij} - TP - FP - FN:\ i = 1, ..., C, j = 1, ..., C\ , \end{cases} \tag{B.1}$$

where, in summary, TP is the diagonal elements, FP is the sum of elements in column, except TP, FN is the sum of elements in row, except TP, and TN is the sum of all elements execpt TP, FP, and FN. Then, the accuracy and F1-score is

finally calculated:

$$Accuracy = \frac{TP + TN}{n} \ , \tag{B.2}$$

where n is the number of random samples used,

$$Precision = TP/(TP + FP) \ , \tag{B.3}$$

$$Recall = TP/(TP + FN) \ , \tag{B.4}$$

finally:

$$F1 = 2 * \frac{precision * recall}{precision + recall} \ , \tag{B.5}$$

The evaluation's source-code described here, is shared and can be found in Github (link available in Table A.2).

B.1 Practical example

Using synthetic images (Figure B.2), this practical example demonstrates the influence of boundary delineation over F1-score metric.

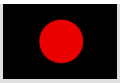
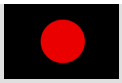
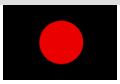
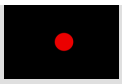
Figure B.2 - Evaluation over synthetic data. A practical example using (a) reference. (b) object segmented out of the reference boundary (outer). (c) object segmented out of the reference boundary (inner). (d) object partially segmented.

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

SOURCE: Author's production.

In the Table B.1 is shown the values for the synthetic evaluation. When test over the same reference image, the accuracy and F1-score seems correct, as there is no changing, but when the predicted object is bigger than in reference, then the recall is high but not the precision, leading to a low F1-score. When the predicted object is smaller it should be, it inverts, recall is lower than precision, also leading to a low F1-score. Finally, a predicted object intersecting only part of the reference, both values are low, equally to F1-score.

Table B.1 - Demonstration of the Accuracy and F1-score changing according to the target boundary and location.

Reference	Predicted	Accuracy	Precision	Recall	F1-Score
		1.0	1.0	1.0	1.0
		0.8903	0.8101	0.9345	0.8478
		0.8539	0.9252	0.5871	0.6080
		0.6162	0.5134	0.5212	0.4969

SOURCE: Author's production.